Un état de l'art du traitement automatique du dialecte tunisien

Jihene Younes[1]* — **Emna Souissi**[2]** — **Hadhemi Achour**[3]* — **Ahmed Ferchichi**[4]*

* *Université de Tunis, ISGT, LR99ES04 BESTMOD, 2000, Le Bardo, Tunisia*
[1] *jihene.younes@gmail.com,*
[3] *Hadhemi_Achour@yahoo.fr*
[4] *Ahmed.Ferchichi@planet.tn*
** *Université de Tunis, ENSIT, 1008, Montfleury, Tunisia*
[2] *emna.souissi@ensit.rnu.tn*

RÉSUMÉ. *Dans le domaine du traitement automatique de la langue arabe, la majorité des recherches menées et des réalisations accomplies ont porté principalement sur l'arabe standard moderne (ASM). Les divers dialectes arabes (DA) comptent encore parmi les langues sous-dotées. Ce n'est que depuis une dizaine d'années que ces dialectes ont commencé à susciter un intérêt accru au sein de la communauté TAL, notamment compte tenu de leur utilisation de plus en plus importante sur le Web social. Dans ce travail, nous nous focalisons sur le dialecte tunisien (DT), et proposons de fournir un état de l'art sur le traitement automatique de ce dialecte. Une revue des travaux accomplis à ce jour, ainsi qu'un inventaire détaillé des divers outils TAL et ressources linguistiques disponibles pour le DT sont présentés puis discutés.*

ABSTRACT. *In the area of Arabic Natural Language Processing, most of the undertaken research and achievements have mainly involved Modern Standard Arabic (MSA). The various Arabic dialects (AD) are still considered to be among under-resourced languages. It's only in the last decade that these dialects began to arouse the interest of NLP researchers, especially given their increasing use on the social web. In this work, we focus on the Tunisian dialect (TD), and propose to provide a state of the art on the automatic processing of this dialect. A review of the works carried out to date and a detailed inventory of the various NLP tools and language resources available for the TD are presented and discussed.*

MOTS-CLÉS : *dialecte tunisien*$_1$, *ressources linguistiques*$_2$, *traitement automatique des langues*$_3$.

KEYWORDS: *Tunisian dialect*$_1$, *Language resources*$_2$, *Natural language processing*$_3$.

1. Introduction

La diglossie est l'une des principales caractéristiques de la langue arabe. Dans les pays arabes, il existe au moins deux formes d'arabe qui coexistent : l'arabe standard moderne (ASM) d'une part, utilisé comme langue officielle, et l'arabe dialectal d'autre part, recouvrant divers dialectes régionaux. Les dialectes arabes (DA) sont des variantes régionales de la langue arabe, naturellement parlées par les populations arabes et utilisées dans leur communication quotidienne. Ils diffèrent d'une région à une autre et d'un pays à un autre et parfois même, différents dialectes peuvent exister dans un même pays.

Les DA sont souvent considérés par la communauté TAL comme faisant partie des langues sous-dotées (Novotney *et al.*, 2016 ; Harrat *et al.*, 2017) compte tenu de la rareté des ressources linguistiques et outils TAL disponibles associés à ces dialectes. En effet, dans le domaine du traitement automatique de la langue arabe, la majorité des recherches menées et des réalisations accomplies ont porté principalement sur l'ASM. Étant la langue utilisée dans la presse, documents administratifs et officiels et enseignée dans les écoles, les ressources linguistiques électroniques en ASM sont largement répandues et disponibles, contrairement aux DA qui sont des langues essentiellement parlées et très rarement écrites. La rareté de ressources volumineuses écrites en arabe dialectal a constitué probablement l'un des freins majeurs à son étude et à son traitement automatique.

Par ailleurs, les DA commencent depuis une dizaine d'années à susciter un intérêt croissant au sein de la communauté TAL, s'expliquant, entre autres, par la prolifération de divers contenus textuels informels sur le Web et plus particulièrement sur le Web social, et qui sont de plus en plus riches en arabe dialectal. Le Web devient, ainsi, une source privilégiée par plusieurs chercheurs, pour l'extraction de contenus dialectaux et la construction de diverses ressources linguistiques. En parallèle, plusieurs travaux portant sur les DA visant, entre autres, leur analyse morphosyntaxique, le traitement de la parole ou encore l'analyse des sentiments ont vu le jour ces dernières années.

Dans ce travail, nous nous penchons sur le dialecte tunisien en particulier, et nous proposons d'élaborer un état de l'art sur son traitement automatique. Dans cette étude, une revue des travaux accomplis à ce jour avec des tableaux synthétiques indiquant les tailles des ressources, les approches utilisées et les performances, ainsi qu'un inventaire détaillé des divers outils TAL et des ressources linguistiques disponibles pour le DT sont présentés puis discutés. Notons, à ce niveau, que notre investigation a porté sur tous les travaux publiés avant la fin du mois de mars 2018. Nous ne rapportons, pour les travaux qui ont porté sur un groupe de dialectes arabes comprenant le tunisien[1], que ceux qui mentionnent les tailles des données utilisées et les performances

1. Parmi les travaux ayant porté sur un groupe de dialectes comprenant le tunisien, sans préciser les tailles de données ni les performances obtenues spécifiques à ce dialecte, nous pouvons citer (Suwaileh *et al.*, 2016 ; Eldesouki *et al.*, 2017).

obtenues explicitement pour le DT. Ce travail pourrait ainsi constituer un point de départ pour tout chercheur souhaitant travailler sur le traitement automatique du DT.

2. Dialecte tunisien (DT)

La Tunisie compte plus de 11 millions d'habitants[2] dans la région du Maghreb située en Afrique du Nord, avec une diaspora de plus de 1,3 million de personnes[3]. La langue maternelle des Tunisiens est le dialecte tunisien, souvent appelé par ses locuteurs « darija » (en arabe, « دارجة », qui est le féminin de la forme « دارج » qui signifie « familier, commun, populaire, utilisé »[4]) . Il est aussi courant de se référer au DT directement en tant que langue nommée « tounsi » (tunisien) .

Le DT est principalement une langue parlée, spontanément utilisée par les Tunisiens pour leur communication de tous les jours, mais il est parfois aussi un langage littéraire au moyen duquel on dit des proverbes, des comptines, des contes, des devinettes et des poèmes et une langue d'écriture lorsque par exemple les paroles de chansons et de pièces de théâtre sont écrites en DT (Pereira, 2005). Aujourd'hui, il est largement utilisé à la radio, à la télévision et dans la publicité. Étant une variante de la langue arabe, le DT est en général, et depuis longtemps, écrit à l'aide de l'alphabet arabe, mais il peut aussi être écrit en utilisant l'alphabet latin. Durant ces dernières années, l'adoption massive de nouveaux modes de communication (SMS, e-mails, Facebook, Twitter, etc.) a grandement renforcé l'écriture de messages et de divers contenus textuels en dialecte, notamment dans le script latin.

2.1. *Spécificités linguistiques*

Le dialecte tunisien est un dialecte arabe maghrébin qui présente d'importantes dissimilarités avec l'ASM d'une part, et avec les autres dialectes arabes d'autre part, surtout avec les dialectes orientaux (arabe Mashriqi). Le DT est en effet, très peu compris par les arabophones orientaux (d'Égypte, du Soudan, du Levant, d'Irak et de la péninsule Arabique) car il dérive de différents substrats et d'un mélange de plusieurs langues (Sayahi, 2014 ; Mohand, 1999 ; Elimam, 2009 ; Elimam, 2012 ; Mzoughi, 2015) : punique, berbère, arabe, turc, français, espagnol et italien. Les différentes civilisations qui ont transité par le pays, ont en effet, laissé leurs empreintes dans le patrimoine linguistique tunisien et le dialecte parlé par les Tunisiens est aujourd'hui bien riche en emprunts lexicaux d'origines multiples. Nous pouvons citer, à titre d'exemples les mots du DT: « كرموس - karmous », qui signifie « figue » et a pour origine le mot berbère « takarmoust », « فرفطو - farfatttou », qui signifie « papillon » et

2. Selon http://countrymeters.info, consulté le 10 janvier 2018.
3. Selon http://ote.nat.tn/wp-content/uploads/2018/05/Repartition_de_la_communaute_tunisienne_a_l_etranger__2012.pdf, consulté en juillet 2018.
4. Dans le dictionnaire « معجم المعاني عربي عربي » : www.almaany.com

a pour origine le mot italien « farfalla », « برجوازي - borjouazi », qui signifie « bourgeois » et a pour origine le mot français « bourgeois », « صباط - sabbat », qui signifie « chaussure » et a pour origine le mot espagnol « zapato », etc. De plus, le DT est une langue en continuelle évolution dans la mesure où de nouveaux mots étrangers sont fréquemment intégrés dans le dialecte et souvent conjugués selon les règles de l'arabe (Sayahi, 2014 ; Elimam, 2009). Baccouche (1994) appelle ce phénomène, pour le cas du français, la « tendance de *tunisifier* le français ». Les exemples suivants montrent la dynamique du système linguistique du dialecte tunisien : « ريفز - rivez » (« réviser »), « يريفز - yrivez » (« il révise »), « مريفز - mrivez » (« en situation de réviser »), etc.

Il est aussi à noter que sur le plan lexical, le DT comprend des mots qui lui sont bien spécifiques et qui le distinguent des autres dialectes maghrébins et orientaux, tels que : « باهي - bahi » (« d'accord »), « برشة - barcha » (« beaucoup »), « فيسع - fissaa » (« vite »), « عالسلامة - asslama » (« bonjour »), « بالسلامة - bislama » (« au revoir »), « ينجّم - ynajjim » (« il peut »), « يزّي - yezzi » (« Arrête, ça suffit ! »), etc.

Sur le plan phonologique, certains phonèmes non arabes peuvent être utilisés dans le DT, tels que / g / ڤ, / p / پ et / v / ڤ. Les voyelles longues sont généralement raccourcies et les trois voyelles courtes sont souvent réduites à deux. Dans de nombreux cas, CvCC est changé en CCvC (par exemple, « سقف / saqf » (« toit ») en ASM, est dit « سقف / sqaf » en DT). En outre, la préférence est pour le stress de syllabe finale, en particulier avec la réduction des voyelles courtes non stressées (par exemple, « كتاب / kitaab » (« livre ») en ASM, est dit « كتاب / ktaab » en DT).

Sur le plan morphologique, plusieurs aspects distinguent le DT de l'arabe standard. En effet, dans le DT, la catégorie du nombre duel de l'ASM n'existe pas et les marques casuelles nominales sont aussi supprimées. La conjugaison des verbes en DT présente aussi des dissimilarités avec l'ASM, en utilisant des affixes différents. Le tableau 1 montre quelques exemples de caractéristiques spécifiques au DT concernant la conjugaison des verbes.

Conjugaison des verbes en DT	Exemple
Le préfixe n-ن est utilisé avec la première personne du singulier	نـ كتب / **n**iktib (j'écris)
Le préfixe n-ن avec le suffixe u-و sont utilisés pour le pluriel	نـ كتب ـو / **n**iktb**ou** (nous écrivons)
Le futur utilise le préfixe verbal (باش, بش) plus le verbe	**باش** نكتب / **bech** niktib (je vais écrire)

Tableau 1. *Conjugaison des verbes en DT*

Au niveau syntaxique, la structure de la phrase en DT présente des différences par rapport à l'ASM et parfois, les caractéristiques syntaxiques du système dialectal

tunisien sont en rupture complète avec l'ASM. Par exemple, la structure la plus dominante pour une phrase en DT est (SVO : Sujet Verbe Objet) tandis que pour l'ASM la structure la plus dominante est (VSO) (Saidi, 2014). En DT, un seul pronom relatif (« اللّي-elli », « qui ») remplace tous les autres pronoms relatifs de l'ASM. De même, les pronoms personnels sont réduits à 7 pronoms par opposition aux douze pour l'ASM, la forme duelle des pronoms démonstratifs est abandonnée, etc. (Mejri *et al.*, 2009).

2.2. *Présence sur les réseaux sociaux*

La Tunisie a connu ces dernières années une utilisation accrue des médias sociaux. Si l'on prend comme exemple Facebook, qui constitue aujourd'hui la plateforme de médias sociaux la plus populaire en Tunisie, le nombre total d'utilisateurs au cours des trois dernières années est passé de 4,6 en 2014 à 6,1 millions d'utilisateurs en 2017 (Mourtada *et al.*, 2014 ; Salem, 2017), ce qui représente plus de 53 % de la population tunisienne.

Pour ce qui est des langues utilisées sur le Web social, il est important de noter que les contenus générés par les utilisateurs tunisiens sont fortement multilingues, comprenant principalement les trois langues : arabe, français et anglais. Selon les mêmes études (celles de Mourtada *et al.* (2014), Salem (2017)), l'utilisation de la langue arabe sur le réseau Facebook en Tunisie a connu une croissance significative. Le taux d'utilisateurs du réseau utilisant la langue arabe est passé en effet de 18 % en 2016 à 68,6 % en 2017, de 15 % à 19,5 % pour l'anglais et de 91 % à 93,1 % pour le français. Remarquons ici, que malgré la forte croissance de l'utilisation de la langue arabe, le français reste une langue dominante sur le réseau social Facebook. La langue arabe utilisée sur le Web social ne se limite pas à l'ASM uniquement, mais les contenus textuels exprimés en arabe dialectal prolifèrent également sur les plateformes sociales. De plus, les productions dialectales peuvent être écrites en utilisant à la fois l'alphabet arabe et l'alphabet latin. Dans (Younes *et al.*, 2015), une étude menée sur différentes pages Facebook tunisiennes (politique, média, sport, etc.), a montré que 58 % du contenu total étudié est en dialectal et que 81,3 % des productions en DT sont transcrites en alphabet latin. Cela est expliqué dans (Younes *et al.*, 2015) par plusieurs facteurs comme la rareté des claviers arabes au début des années du Web et de la technologie mobile ainsi que le phénomène du multilinguisme caractérisant la population tunisienne.

2.3. *Difficultés et défis*

Le traitement automatique du DT est une tâche non triviale présentant des difficultés particulières. Ces difficultés sont principalement dues au manque de conventions de normes orthographiques, à l'évolution continue de la langue avec l'emprunt de nouveaux mots, aux variétés du DT pouvant différer d'une région à une autre à travers le pays et présenter des dissemblances à différents niveaux linguistiques (Baccouche et Mejri, 2004), ainsi qu'à ses importantes dissimilarités avec l'ASM. Ces défis rendent

difficile l'utilisation directe des outils de TAL disponibles pour l'ASM. Notons aussi que les phénomènes de diglossie et alternance de code (*code switching*) posent, notamment, le problème de l'identification du DT. Cette identification constitue une tâche complexe qui doit résoudre de nombreux problèmes d'ambiguïté. L'ambiguïté concerne les mots pouvant être à la fois des mots de l'ASM et du DT lorsqu'ils sont transcrits en alphabet arabe, ou encore les mots pouvant être à la fois des mots étrangers (français ou anglais) et du DT, lorsqu'ils sont transcrits en alphabet latin. Par exemple, le mot « خاطر » signifie « parce que » en DT et « esprit » en ASM, le mot « bard » signifie « froid » en DT et « poète » en anglais, le mot « flous » signifie « argent » en DT et « flous » en français, etc. (Younes *et al.*, 2015). Ces divers problèmes accentuent la difficulté du traitement automatique du DT, qui ne pourra pas se développer sans la disponibilité de larges ressources linguistiques et le développement de divers outils de TAL appropriés. Dans la suite de cet article, nous proposons une revue des travaux de TAL portant sur le DT et dressons un inventaire des outils et ressources actuellement disponibles pour son traitement automatique.

3. Traitement automatique du DT

3.1. *Construction de ressources linguistiques*

Dans cet article, nous nommons « ressource linguistique » (RL), toute collection de données relatives à la langue telles que les corpus oraux ou écrits, lexiques, dictionnaires, ontologies, etc. Ce type de ressources représente, en effet, un matériau essentiel pour l'étude des langues et le développement d'outils et applications de TAL. Pour ce qui concerne le DT, et vu le manque de ressources disponibles dans un format électronique pour ce dialecte, plusieurs travaux ont porté sur la construction de divers types de collections de données dialectales. Certains d'entre eux sont partis d'enregistrements de dialogues, conversations, radios et télédiffusions comme (Belgacem, 2009 ; Graja *et al.*, 2010 ; Masmoudi *et al.*, 2014a ; Masmoudi *et al.*, 2014b ; Masmoudi *et al.*, 2014c) et ont procédé à leur transcription en corpus écrits. D'autres chercheurs ont eu recours, entre autres, au Web et aux médias sociaux afin d'extraire des données en dialectal et construire divers types de ressources. Nous pouvons citer parmi ceux-ci, McNeil et Faiza (2011) qui ont construit un corpus appelé TAC (*Tunisian Arabic Corpus*) dans le cadre d'un projet de création d'un dictionnaire DT-anglais. Ce corpus a ensuite été organisé dans une application Web permettant un traitement linguistique de base. Younes *et al.* (2014 ; 2015) ont eu recours au Web social pour construire diverses ressources pour le DT. Ils ont développé un module qui extrait automatiquement les commentaires des pages Facebook tunisiennes et filtre les spams et les messages écrits en langues étrangères en s'appuyant sur des lexiques.

Mise à part la génération de corpus bruts pour le DT, il convient de mentionner les travaux qui ont visé la création de corpus parallèles et divers corpus annotés. Nous citons à cet effet, Bouamor *et al.* (2014) et Harrat *et al.* (2017). Par ailleurs, divers

corpus annotés ont été construits, notamment par Graja *et al.* (2013) et Zribi *et al.* (2015) et décrits dans les tableaux 2 et 3.

Les travaux de Boujelbane (2013) et Boujelbane *et al.* (2013a ; 2013b ; 2014) ont conduit à la construction de lexiques bilingues ASM-DT, en utilisant le Penn Arabic Treebank (Maamouri *et al.*, 2014) et des règles de conversion fondées sur les différences entre l'ASM et le DT. Hamdi *et al.* (2014) ont construit un lexique pour les noms déverbaux en DT. Masmoudi *et al.* (2014a ; 2014b ; 2014c), qui ont construit un corpus nommé TARIC (corpus arabe d'interaction du chemin de fer tunisien) en transcrivant manuellement des enregistrements audio, ont généré automatiquement un dictionnaire de prononciation du DT nommé TunDPDic.

Nous avons pu identifier, par ailleurs, des ontologies de domaine construites dans le cadre de travaux sur le traitement du DT dans les systèmes de dialogue (dans les gares) par Graja *et al.* (2011a ; 2011b), Karoui *et al.* (2013a ; 2013b) et Graja *et al.* (2015). En effet, Graja *et al.* (2011a ; 2011b) ont utilisé des ontologies pour couvrir le lexique utilisé dans les gares en utilisant le corpus TuDiCoI construit par Graja *et al.* (2010). Karoui *et al.* (2013a ; 2013b) ont proposé une méthode hybride (combinant une approche statistique et une approche linguistique) pour la construction semi-automatique d'une ontologie de domaine, appelée Railway Information Ontology (RIO), à partir du corpus TuDiCoI (Graja *et al.*, 2010). Des travaux sur la construction d'ontologies en DT comprennent également le WordNet (TunDiaWN) proposé par Bouchlaghem *et al.* (2014) et qui peut être considéré comme une ressource parallèle DT-ASM puisqu'il préserve le contenu AWN (Arabic WordNet de Elkateb *et al.* (2006)). À partir d'un corpus nommé MultiTD (Multi-source Tunisian Dialect corpus), Bouchlagem *et al.* (2014) ont utilisé une méthode permettant de regrouper les mots en groupes significatifs, fondée sur l'algorithme de K-modes (Huang, 1998) pour enrichir le WordNet. Citons enfin, l'ontologie « aebWordNet », proposée par Ben Moussa Karmani *et al.* (2014 ; 2015) et Ben Moussa Karmani et Alimi (2016), qui a été modélisée à partir du dictionnaire bilingue anglais-arabe tunisien « *Peace corps dictionary* » de Abdelkader (1977). Une synthèse de ces travaux, résumant les caractéristiques des ressources construites, est présentée dans les tableaux 2 et 3.

3.2. *Traitement de la parole*

Parmi les travaux de recherche menés sur le dialecte tunisien, nous pouvons distinguer ceux qui ont porté sur le traitement de la parole. Certains d'entre eux ont visé l'identification de différents dialectes arabes (comprenant le tunisien). Dans cette catégorie, nous citons Belgacem *et al.* (2010) qui ont utilisé le modèle de mélange gaussien (GMM) de Reynolds *et al.* (2000), qui permet de reconnaître les similitudes et les différences entre chaque dialecte. Les travaux de Lachachi et Adla (2015 ; 2016a ; 2016b) ont aussi porté sur deux systèmes d'identification automatique de 5 dialectes arabes du Maghreb : marocain, tunisien et trois dialectes algériens. Le premier système est basé sur les modèles de mélange de lois gaussiennes (GMM) et le second est fondé sur une combinaison d'un modèle du monde et des modèles de mélanges de lois gaussiennes

Auteurs	ADA	Script	Ressources construites
Belgacem (2009)	✓	A	- Corpus arabe multidialectal, de 10 h de parole de 8 DA dont 37 % transcrites (90 min dont 5 % transcrites pour le DT). - Source : discours enregistrés, journaux radio ou télédiffusés + transcription
Graja *et al.* (2010)		A	- Corpus TuDiCoI : 127 dialogues ; 893 segments ; 3 403 mots. - Source : conversations enregistrées dans la gare de la SNCFT + transcription
McNeil et Faiza (2011) ; McNeil (2015)		A	- Corpus TAC 2011 : 400 k mots ; Corpus TAC 2015 : 820 k mots - Source : écrits traditionnels + blogs + e-mails + Facebook + audio transcription
Graja *et al.* (2011a ; 2011b)		A	- Ontologie de 15 concepts - Source : corpus TuDiCoI
Karoui *et al.* (2013a ; 2013b)		A	- RIO : 14 concepts, 25 relations, 387 instances - Source : corpus TuDiCoI
Graja *et al.* (2013)		A	- Corpus TuDiCoI : 1 825 dialogues ; 12 182 segments ; 21 682 mots dont 7 814 mots sont annotés (normalisation lexicale, analyse morphologique, lemmatisation et attribution des synonymes) - Source : conversations enregistrées dans la gare de la SNCFT + transcription
Boujelbane *et al.* (2013a ; 2013b ; 2014b) ; Boujelbane (2013)		A	- Corpus : 5 h 20 min de paroles transcrites ; 37 964 mots dont 12 149 en DT - Corpus : 12 k mots - Lexique bilingue ASM-DT - Source: Arabic Tree Bank (ATB) + paroles transcrites + numérisation de la constitution tunisienne
Bouamor *et al.* (2014)	✓	A	- Corpus parallèle multidialectal MPCA : 2 000 phrases pour chaque langue dont 10 896 mots pour le DT (5 DA + ASM + anglais). - Source : 2 000 phrases du corpus égyptien-anglais de Zbib et Callison-Burch (2012)

Tableau 2. *Construction de ressources linguistiques en DT (1) (ADA : avec d'autres dialectes arabes : nous marquons dans cette colonne les travaux qui ne sont pas spécifiques au DT, mais qui l'ont traité comme faisant partie d'un ensemble de dialectes arabes (DA), script : le système d'écriture des ressources (A : arabe / L : latin))*

(UBM-GMM). Ils ont collecté un corpus de dialectes parlés à partir d'émissions télévisées. Pour le tunisien, le corpus comprend 53,37 heures avec 130 locuteurs.

D'autres travaux ayant porté sur la reconnaissance automatique de la parole en DT comprennent ceux de Neifar *et al.* (2014), fondés sur l'adaptation d'un système élaboré pour l'ASM, les travaux de Hassine *et al.* (2016) qui ont développé un système de reconnaissance automatique de la parole en arabe afin de reconnaître 10 chiffres arabes (de 0 à 9) parlés en dialectes marocain et tunisien ainsi que ceux de (Hassine *et al.*, 2018) portant sur la reconnaissance de la parole en DT. La conversion de Graphème-en-Phonème (G2P) pour le DT, consistant à convertir une séquence de graphèmes en une séquence de symboles phonétiques, a fait l'objet des travaux de Masmoudi *et al.* (2016) ainsi que Masmoudi *et al.* (2017). Tous ces travaux sont décrits dans le tableau 4.

Auteurs	ADA	Script	Ressources construites
Younes et Souissi (2014)		L	- Corpus de 43 222 messages en DT - Lexique DT de 19 763 mots - Source : SMS + Chat + forum + Facebook
Hamdi *et al.* (2014)		A	- Lexique bilingue ASM-DT de 39 793 entrées (14 804 - 5 017 lemmes) - Source : Arabic Tree Bank (ATB) (Maamouri *et al.*, 2014)
Masmoudi *et al.* (2014a ; 2014b ; 2014c)		A	- Corpus TARIC : 20 h de paroles transcrites ; 71 684 mots. - Dictionnaire phonétique TunDPDic : 18 k mots - Source: discours enregistré avec transcription manuelle
Bouchlaghem *et al.* (2014)		AL	- Corpus MultiTD de 32 848 mots - WordNet TunDiaWN - Types d'entités: *synset*, *word*, *form*, *word relations*, *annotator* - Source: corpus multiTD issu de réseaux sociaux (Twitter, Facebook, etc.), pièces de théâtre écrites, dictionnaires, discours enregistré, etc.
Ben Moussa Karmani *et al.* (2014 ; 2015) Ben Moussa Karmani et Alimi (2016)		AL	- WordNet (aebWordNet) - Synset : 18 209 entrées (8 279 lemmes et 25 748 mot-sens) - Source : dictionnaire bilingue anglais-DT de « *Peace corpus dictionary* »
Zribi *et al.* (2015)		A	- Corpus annoté STAC : 4 h 50 min - 42 388 mots – 7 788 phrases - Corpus écrit : annoté par segmentation en phrases, segmentation des mots en affixes et clitiques, lemme, genre, nombre, personne, voix, étiquettes, etc. - Corpus oral : annoté par frontières des phrases et disfluences - Source : chaînes de télévision et stations de radio + 30 min de TuDiCoI
Younes *et al.* (2015)		AL	- Corpus de 31 158 messages – 420 897mots en LDT et 7 145 messages – 160 418 mots en ADT - Lexique L→A 19 763 entrées - Lexique A→L 18 153 entrées - Source : SMS + Chat + forum + Facebook
Graja *et al.* (2015)		A	- Ontologie de 18 concepts - Source : corpus TuDiCol
Harrat *et al.* (2017)	✓	A	- Corpus parallèle multidialectal PADIC : 6 400 phrases pour chaque langue dont 36 648 mots pour le DT (5 dialectes + ASM). - Source : 6 400 phrases du dialecte algérien issues de discours et émissions télévisés algériens enregistrés et transcrits

Tableau 3. *Construction de ressources linguistiques en DT (2) (ADA : avec d'autres dialectes arabes, script : le système d'écriture des ressources (A : arabe / L : latin))*

3.3. *Analyse morphosyntaxique*

L'étude de la littérature montre que plusieurs travaux ont été menés sur l'analyse morphosyntaxique du DT. Dans (McNeil, 2012), un analyseur du DT est présenté. Il permet la segmentation des mots en suivant un ensemble de règles relatives aux suffixes et aux préfixes après une étape de prétraitement consistant à filtrer la ponctuation et les mots étrangers, et à translittérer l'écriture arabe en écriture latine en adaptant la translittération de Buckwalter au DT (Buckwalter, 2004). Zribi *et al.* (2016) se sont, pour leur part, concentrés sur la détection des limites des phrases dans le dialecte tunisien transcrit en utilisant trois approches : une approche fondée sur des règles faites à la main sur la base d'éléments lexicaux et caractéristiques prosodiques ; une approche statistique de classification de mots fondée sur PART (Mohamed *et al.*, 2012) ; une approche hybride combinant les deux approches précédentes. D'autres travaux sur l'analyse morphologique du DT comprennent, notamment, celui de Zribi *et al.* (2013)

Auteurs	ADA	Type de traitement et approche	Évaluation
Belgacem *et al.* (2010)	✓	- Identification de dialectes arabes (9 DA) - GMM	TR : 73,33 %
Neifar *et al.* (2014)		- Système de compréhension de la parole Comp-Dial System (*literal understanding of the Tunisian dialect*) - Adaptation d'un système d'ASM pour DT SARF (*arabic vocal server of railway information*) (Bahou, 2014)	TE : 20,34 %
Lachachi et Adla (2015 ; 2016a ; 2016b)	✓	- Identification de 5 dialectes maghrébins - Modèles de mélange de lois gaussiennes (GMM, UBM-GMM)	Précision : 80,49 %
Hassine *et al.* (2016)	✓	- Reconnaissance de la parole des dialectes maghrébins - FFBPNN et SVM	TR pour FFBPNN : 98,3 % TR pour SVM : 97,5 %
Masmoudi *et al.* (2016)		- Conversion Graphème-en-Phonème - CRF	TE phonétique : 14,09 %. Rappel : 91,41 % Précision : 87,3 %
Masmoudi *et al.* (2017)		- Conversion Graphème-en-Phonème (G2P) et reconnaissance de la parole (RP) - G2P : à base de règles - RP : modèle acoustique de PLP (perceptual linear predictive) et modèle de langage trigramme	G2P : TR niveau phonème : 99,6 % TE niveau mot : 22,6 % RP : TE : 22.6 %
Hassine *et al.* (2018)		- Reconnaissance de la parole du DT - FFBPNN	TR : 98,5 %

Tableau 4. *Traitement de la parole (ADA : avec d'autres dialectes arabes, TR : taux de reconnaissance, TE : taux d'erreur)*

qui a consisté en une adaptation de l'analyseur morphologique existant de l'arabe standard (Al-khalil (Boudlal *et al.*, 2010)) à l'aide d'un ensemble de transformations des patrons verbaux et nominaux de l'ASM en des patrons adaptés au DT, ainsi que la définition d'un ensemble d'affixes et mots-outils du DT. Dans un travail ultérieur, Zribi *et al.* (2017) ont proposé une méthode de désambiguïsation pour l'analyse morphologique du DT et ont testé, pour cela, diverses techniques d'apprentissage automatique (RIPPER (Cohen, 1995), PART (Mohamed *et al.*, 2012) et SVM (Vapnik, 1995)). Ben Moussa Karmani et Alimi (Karmani et Alimi, 2016) ont développé un analyseur morphologique tenant compte des spécificités du DT, et en s'appuyant sur des règles, le WordNet « aebWordNet », un dictionnaire lexical et un système expert linguistique. Par ailleurs, de récents travaux portant sur l'étiquetage grammatical et l'analyse syntaxique du DT commencent à voir le jour. Nous citons en particulier, Boujelbane *et al.* (2014) qui ont eu recours à des ressources existantes pour l'arabe standard, à savoir l'étiqueteur ASM de Stanford entraîné sur une version DT du corpus ATB (Arabic Tree Bank). Cette version a été générée en utilisant un lexique bilingue ASM-DT ainsi qu'un outil de traduction en DT développé par Boujelbane *et al.* (2013a ; 2013b). Dans le même esprit, Hamdi *et al.* (Hamdi *et al.*, 2015) ont également développé un étiqueteur grammatical pour le DT en exploitant sa proximité avec l'ASM. Le processus d'étiquetage est fondé sur des modèles HMM d'ordres différents, entraînés sur 24 k de phrases ASM obtenues du corpus ATB. Pour leur part, Mekki *et al.* (2017) ont travaillé sur la création d'un Treebank arabe tunisien et son utilisation pour effectuer

une analyse syntaxique du DT. Le corpus est la version de la constitution tunisienne écrite en DT qui comprend 12 k mots et 492 phrases. Ils ont utilisé l'analyseur syntaxique de Stanford, principalement dédié à l'ASM, qui reçoit en entrée les phrases tunisiennes normalisées et fournit l'arbre syntaxique de chaque phrase. Ben Moussa Karmani et Alimi. (2016) ont développé un analyseur morphologique tenant compte des spécificités du DT en utilisant une approche fondée sur des règles, le WordNet « aebWordNet », un dictionnaire lexical et un système expert linguistique. Ces divers travaux sont brièvement décrits dans le tableau 5.

Auteur	ADA	Script	Type de traitement et approche	Évaluation
McNeil (2012)		L	- Segmentation des mots - À base de règles	Exactitude : 89,2 %
Zribi *et al.* (2013b)		A	- Analyse morphologique - Adaptation de l'analyseur morphologique Al-Khalil du ASM (Boudlal *et al.*, 2010)	Précision : 80 % F-mesure : 88,86 %
Boujelbane *et al.* (2014a)		A	- Étiquetage grammatical - Adaptation de l'étiqueteur ASM de Stanford	Exactitude : 78,5 %
Hamdi *et al.* (2015)		A	- Étiquetage morphosyntaxique - Adaptation de l'analyseur morphologique MAGEAD (Habash et Rambow, 2006) et étiqueteur HMM	Exactitude : 89 %
Zribi *et al.* (2016)		A	- Segmentation en phrases - À base de règles, statistique, hybride	Précision : 94,8 %
Ben Moussa karmani et Alimi (2016)		A	- Décomposition en morphèmes et analyse morphologique - À base de règles	Précision décomposition : 58,94 % Précision étiquetage : 84,41 %
Zribi *et al.* (2017)		A	- TAMDAS : Système d'étiquetage grammatical - Classification à base de règles PART (Mohamed *et al.*, 2012) et RIPPER (Cohen, 1995), SVM et classifieur bigramme	Exactitude : 87,32 %
Mekki *et al.* (2017)		A	- Analyse syntaxique - Adaptation de l'analyseur syntaxique ASM de Stanford	Précision : 64,43 % F-mesure : 65,58 %

Tableau 5. *Analyse morphosyntaxique (ADA : avec d'autres dialectes arabes, script : le système d'écriture des ressources (A : arabe / L : latin))*

3.4. *Identification de langue sur l'écrit*

L'identification automatique du DT a fait l'objet d'un nombre relativement réduit de travaux. Il s'agit d'identifier le DT dans des écrits pouvant être aussi bien dans le script arabe ou latin. Notons, cependant, que la plupart des travaux recensés ne sont pas vraiment spécifiques à la détection du DT en particulier, mais visent essentiellement à reconnaître l'arabe dialectal en général. Nous avons pu recenser un seul travail sur le dialecte tunisien en particulier, qui est celui de Aridhi *et al.* (2017), sur l'identification des mots du DT transcrits en alphabet latin et dans lequel deux approches ont été expérimentées. La première est fondée sur la méthode N-Gram CSIF (N-Gram Cumulative Sum of Internal Frequencies) de Ahmed *et al.* (2004), et la seconde sur une classification SVM (Support Vector Machines). D'autres travaux ont visé l'identification de l'arabe dialectal en général, y compris le tunisien. Nous citons

principalement ceux de Sadat *et al.* (2014a ; 2014b) qui ont travaillé sur la détection de l'arabe dialectal (couvrant 18 dialectes dont le DT) dans les médias sociaux, en utilisant un classificateur NB (Naive Bayes) fondé sur un modèle de bigrammes de caractères. Le travail de Harrat *et al.* (2015) a porté sur l'identification de 5 dialectes et de l'ASM au niveau des phrases, en utilisant une classification basée sur NB avec les trigrammes de caractères comme caractéristiques. Malmasi *et al.* (2015) ont aussi traité l'identification, au niveau des phrases, de 5 variétés de dialectes avec ASM. Ils ont utilisé le corpus MPCA et ont mené plusieurs expériences avec un classificateur SVM linéaire et un métaclassificateur, en utilisant diverses caractéristiques de surface fondées sur des caractères et des mots. Adouane *et al.* (2016) ont pour leur part, utilisé les SVM pour identifier l'arabe dialectal comprenant plusieurs DA dont le DT. Nous citons enfin, Saadane *et al.* (2017) qui ont comparé une approche linguistique exploitant des dictionnaires avec une approche statistique à base de n-grammes, pour la détection automatique de DA dont le maghrébin (tunisien, algérien et marocain) et l'égyptien. Ces différents travaux sont décrits dans le tableau 6.

Auteur	ADA	Script	Type de traitement et approche	Évaluation
Sadat *et al.* (2014a ; 2014b)	✓	A	- Identification des dialectes arabes (18 DA) - Modèles de langage n-gramme de Markov fondés sur des caractères et classification Naive Bayes	Exactitude : 98 %
Harrat *et al.* (2015)	✓	A	- Identification des dialectes arabes au niveau phrase (5 DA + ASM) - Classification par Naive Bayes basée sur un modèle trigramme de caractères	Précision pour le DT : 68 %
Malmasi *et al.* (2015)	✓	A	- Identification des dialectes arabes au niveau phrase (5 DA + ASM) - Classification SVM et SGM	Exactitude : 74 %
Adouane *et al.* (2016)	✓	A	- Identification des dialectes arabes (8 DA + berbère) - Classification de Cavnar, SVM et PPM fondée sur des modèles n-grammes de caractères et de mots.	F-mesure : 92,94 %
Saadane *et al.* (2017)	✓	AL	- Identification des dialectes arabes au niveau mot (4 DA + ASM) - Approche linguistique à base de dictionnaires + aproche statistique	TE : 14,5 %
Aridhi *et al.* (2017)		L	- Identification du DT au niveau mot - Méthode CSIF + classification SVM	F-mesure - CSIF : 87 % F-mesure - SVM : 90 %

Tableau 6. *Identification (ADA : avec d'autres dialectes arabes, script : le système d'écriture des ressources (A : arabe / L : latin))*

3.5. *Translittération*

La translittération consiste à transformer un mot d'un système d'écriture en un autre tout en préservant sa prononciation. Comme le DT peut être écrit à la fois en arabe et en latin, en particulier sur les réseaux sociaux, les blogs et les forums, des chercheurs se sont intéressés à la tâche de translittération du DT, qui peut être particulièrement utile pour le processus de construction de ressources dialectales et pour de nombreuses autres applications telles que la traduction automatique (traitement des

noms propres), la recherche d'informations, etc. Masmoudi *et al.* (2015) ont abordé la translittération du latin vers l'arabe, en utilisant des règles préétablies selon la norme CODA. Younes *et al.* (2016), se sont orientés vers des méthodes d'apprentissage automatique, en proposant un modèle de translittération fondé sur les HMM. Nous résumons ces deux travaux dans le tableau 7.

Auteur	Sens	Approche adoptée	Évaluation
Masmoudi *et al.* (2015)	L → A	À base de règles	Rappel : de 92 % à 93 %
Younes *et al.* (2016)	L → A	Étiquetage séquentiel (HMM)	Précision : 53 %

Tableau 7. *Translittération (Sens : le sens de la translittération (A : arabe / L : latin))*

3.6. *Traduction*

Les travaux réalisés sur la traduction du DT ont porté essentiellement sur la traduction du DT vers l'ASM, comme ceux réalisés par Hamdi *et al.* (2013a ; 2013b), qui ont exploité des ressources et outils ASM existants pour traduire automatiquement le DT en ASM, ou encore Sadat *et al.* (2014c) qui ont commencé par construire un lexique DT-ASM manuellement comprenant environ 1 600 entrées, ainsi qu'un ensemble de règles de conversion qu'ils ont appliquées à la transformation des verbes du DT vers l'ASM. D'autres travaux ont abordé la traduction, en ASM, de divers dialectes arabes dont le DT tels que Meftouh *et al.* (2015) et Harrat *et al.* (2015) qui ont utilisé le corpus (PADIC), formé d'une collection de 6 400 phrases de dialectes ASM parallèles pour développer un système de traduction, en ayant recours aux outils GIZA++ (Och et Ney, 2003) et SRILM (Stolcke, 2003). L'ensemble de ces travaux sont décrits dans le tableau 8.

Auteur	ADA	Script	Type de traitement et approche	Évaluation
Hamdi *et al.* (2013a ; 2013b)		A	- Traduction de verbes entre DT et ASM - Approche à base de règles	Rappel : de 80 % à 84 %
Sadat *et al.* (2014c)		A	- Traduction DT → ASM de textes - Approche à base de règles	Score BLEU : 14,32
Meftouh *et al.* (2015) Harrat *et al.* (2015)	✓	A	- Traduction entre paires de langues (5 dialectes et ASM) - Techniques de Kneser-Ney et Written-Bell *smoothing*	Score BLEU : 40,48

Tableau 8. *Traduction (ADA : avec d'autres dialectes arabes, script : le système d'écriture des ressources (A : arabe / L : latin))*

3.7. *Analyse de sentiments*

Avec l'utilisation croissante des dialectes arabes dans les médias sociaux, plusieurs travaux ont été initiés sur l'analyse de sentiments (parfois désignée aussi par détection d'opinion), au cours des dernières années. Sayadi *et al.* (2016) ont procédé à la construction de données annotées contenant à la fois du DT et de l'ASM, collectées pendant la période de l'élection de l'Assemblée nationale et de l'élection présiden-

tielle en Tunisie. Ils ont également comparé 5 classificateurs appliqués à la tâche de l'analyse des sentiments (Naive Bayes, SVM, KNN, arbres de décision, forêts d'arbres décisionnels), et mis en œuvre une méthode d'extraction et de sélection des caractéristiques. Ameur *et al.* (2016) se sont focalisés sur l'analyse des sentiments de commentaires en DT, recueillis sur des pages Facebook tunisiennes et ont proposé une méthode pour la construction de dictionnaires émotionnels en distinguant 9 classes émotionnelles (surpris, satisfait, heureux, joyeux, romantique, déçu, triste, en colère et dégoûté). Cette méthode est fondée sur la présence d'émoticônes dans le corpus et sans utiliser de ressources linguistiques externes. Dans le même contexte, Mdhaffar *et al.* (2017) ont commencé par collecter un corpus, appelé TSAC (Tunisian Sentiment Analysis Corpus) à partir de Facebook, qu'ils ont manuellement annoté par les polarités positive et négative. Ils ont utilisé des techniques d'apprentissage automatique (dont le classifieur Perceptron multicouche) pour la classification binaire des commentaires écrits en DT. Le tableau 9 présente un inventaire de ces travaux.

Auteur	ADA	Script	Type de traitement et approche	Évaluation
Sayadi *et al.* (2016)		A	- Construction de ressource pour l'analyse de sentiments (polarité « positive, négative, neutre ») des commentaires issus de Twitter et écrits en ASM et DT. - NB, SVM, KNN, DT, RF	- Exactitude : 71 % - Corpus annoté de 1 754 tweets en DT
Ameur *et al.* (2016)		AL	- Génération des dictionnaires émotionnels en indiquant la polarité : « surpris, satisfait, heureux, joyeux, romantique, déçu, triste, en colère, dégoûté » - Présence d'émoticônes dans les pages Facebook	- F-mesure : 81,01 % - Lexique de 131 937 mots avec polarité
Mdhaffar *et al.* (2017)		AL	- Construction de ressources pour l'analyse de sentiment et détermination de la polarité « positive, négative » des commentaires écrits en DT - SVM, NB et MLP	- TE : 0,22 - Corpus TSAC de 17K commentaires Facebook annotés

Tableau 9. *Analyse de sentiments (ADA : avec d'autres dialectes arabes, script : le système d'écriture des ressources (A : arabe / L : latin))*

3.8. *Codification et normalisation*

La nature informelle des dialectes et leur non-conformité à des règles orthographiques précises rendent difficile leur traitement automatique. Par conséquent, certains chercheurs ont eu recours à une étape intermédiaire susceptible de faciliter leur traitement, à savoir établir des conventions orthographiques. C'est le cas de Zribi *et al.* (2013) qui ont proposé OTTA, une convention orthographique pour la transcription de l'arabe tunisien parlé. Ils ont utilisé les règles orthographiques ASM afin d'en définir de nouvelles qui sont spécifiques au DT. Une autre convention a été développée par Zribi *et al.* (2014) en adaptant le projet CODA (une orthographe conventionnelle pour l'arabe dialectal) de Habash *et al.* (2012) au dialecte tunisien. Le CODA du DT suit les mêmes règles orthographiques que l'ASM avec quelques exceptions et extensions phonologiques, phonolexicales, morphologiques et lexicales. Le système COTA (COnventionalized Tunisian Arabic orthography : orthographe conventionnelle

de l'arabe tunisien) proposé par Boujelbane *et al.* (2016), est un système de normalisation automatique de l'orthographe du DT. Il utilise une méthode hybride combinant des méthodes fondées sur des règles CODA du DT et des méthodes statistiques.

4. Disponibilité des ressources et outils DT

La consultation des plateformes des organisations telles que le Language Data Consortium (LDC) et l'European Language Resources Association (ELRA) permet de constater qu'il existe très peu des ressources pour le traitement du DT, contrairement à l'ASM et à certains autres dialectes arabes (levantin et égyptien). En fait, une simple requête[5] pour les ressources macro-arabes disponibles dans le catalogue LDC donne 164 ressources. Seulement 11 RL multidialectales comprennent le DT. Avec le moteur de recherche ELRA, nous avons trouvé 108 ressources, dont seules 3 RL multidialectales comprennent le DT. Les divers travaux entrepris sur le DT que nous avons présentés dans la section 3 de cet article ont conduit à la construction de diverses RL du DT (de type données telles que des corpus bruts ou annotés, lexiques, dictionnaires, ontologies, etc.). Nous notons cependant que, sur l'ensemble des RL produites dans le cadre de ces travaux, seuls 24 % sont disponibles sur le Web. Le tableau 10 recense ces RL en indiquant si elles sont librement téléchargeables.

RL (auteurs)	Lien	Libre
Corpus TuDiCoI (Graja *et al*., 2010 ; 2013)	https://sites.google.com/site/marwagraja/resources	✓
Corpus TAC (McNeil, 2011 ; 2015)	http://tunisiya.org/ (consultable en ligne)	
Corpus TARIC (Masmoudi *et al.*, 2014a ; 2014b ; 2014c)	http://www-lium.univ-lemans.fr/ bougares/ressources.php https://sites.google.com/site/masmoudiabir/res	✓
Corpus STAC brut et annoté avec les disfluences : audio + transcription (Zribi *et al*., 2015)	https://sites.google.com/site/ineszribi/ressources/corpus	✓
aebWordNet (Ben Moussa Karmani *et al.* 2014, 2015)	https://github.com/nadou12/aebWordNet-Lexicon	✓
Dictionnaire lexical (étiqueté grammaticalement) (Ben Moussa Karmani et Alimi, 2016)	https://github.com/nadou12/Tunisian-Arabic-Lexical-Dictionary	✓
Corpus de 1500 mots (Ben Moussa Karmani et Alimi, 2016)	https://github.com/nadou12/Intelligent-Tunisian-Arabic-Morphological-Analyzer-evaluation-corpus	✓
Corpus TSAC (Mdhaffar *et al*., 2017)	https://github.com/fbougares/tsac	✓
Corpus PADIC (Meftouh *et al.* 2015)	https://sourceforge.net/projects/padic/files/	✓
Lexique DT (Zribi *et al*., 2013b)	https://sites.google.com/site/ineszribi/ressources/lexique	✓
RIO ontology (Karoui *et al.*, 2013a ; 2013b)	https://sites.google.com/site/marwagraja/resources	✓

Tableau 10. *Ressources linguistiques du DT disponibles*

Pour collecter des données afin de construire des lexiques et des corpus, certains chercheurs ont eu recours à différentes ressources librement disponibles sur le Web. Ces ressources sont énumérées dans le tableau 11.

5. Requêtes réalisées le 14 février 2018.

RL (auteurs)	Année	Description - Lien
Documents sélectionnés de la littérature arabe et dialectale (Mohammad Bakri)	2010	Variétés de chansons, théâtres, articles de journaux http://www.langue-arabe.fr/spip.php ?article25
Nouvelle constitution tunisienne (Klibi Salsabil, Hamraoui Salwa, Ben Abda Hana, Gaddes Chawki, Horcheni Farhat, Maalla Anouar)	2014	12 k de mots distribués sur 492 phrases https://www.babnet.net/9/destourderjaaa.pdf
Dictionnaire anglais-tunisien : « *Peace Corps dictionary* » (Rached Ben Abdelkader, Abdeljelil Ayed et Aziza Naouar)	1977	Manuscrit numérisé https://files.eric.ed.gov/fulltext/ed183017.pdf
Dictionnaire tunisien-français « *le Karmous* » (Karim Abdellatif)	2010	3 800 mots, proverbes et expressions https://www.fichier-pdf.fr/2010/08/31/m14401m/dico-karmous.pdf
Dictionnaire tunisien-français	N/A	Plus de 4 k de mots et expressions arabetunisien.com

Tableau 11. *Autres données en DT disponibles*

Nous n'avons pu recenser que deux outils de traitement du DT,(représentant 6 % des travaux présentés dans notre revue) qui sont disponibles sur le Web. Il s'agit de :

– un outil proposé par Zribi *et al.* (2013 ; 2016 ; 2017), qui permet de faire une analyse morphologique, une segmentation en phrases ainsi qu'un étiquetage morphosyntaxique. Cet outil est accessible par le lien : https://sites.google.com/site/ineszribi/ressources/outils-de-traitement-du-dialecte-tunisien ;

– un convertisseur de graphème en phonème (G2P) (Masmoudi *et al.*, 2014b) accessible par le lien : https://sites.google.com/site/masmoudiabir/res.

5. Discussion

5.1. *Portée des travaux de traitement du DT*

Il ressort de cette étude que le traitement automatique du dialecte tunisien suscite, depuis quelques années, un intérêt croissant de la part de plusieurs chercheurs en TAL. Nous avons recensé à ce jour un total de 63 travaux impliquant ce dialecte, dont 50 (79 %) sont dédiés spécifiquement au DT et 13 (21 %) ont porté sur un ensemble de dialectes arabes incluant le DT. Ce nombre reste encore très limité en comparaison de l'arabe standard ou d'autres langues.

En examinant la nature de ces travaux et les types de traitement visés, nous constatons que plus de 43 % d'entre eux ont porté sur la construction de diverses RL pour le DT à cause du manque de ressources indispensables pour travailler sur ce langage. Une part importante de ces travaux (14 %) a été consacrée au traitement de la parole, ce qui s'explique par le fait que le DT constitue une langue essentiellement parlée et très peu écrite. C'est surtout avec la récente croissance de son utilisation sur le Web social que des ressources écrites ont commencé à être collectées et traitées. L'analyse morphosyntaxique ainsi que l'identification des dialectes ont aussi constitué l'objet

de plusieurs travaux (avec respectivement 13 % et 11 % du total des travaux recensés). En revanche, des traitements comme la traduction, l'analyse des sentiments ou encore la translittération restent encore peu abordés pour le DT avec respectivement des pourcentages de 6 %, 5 % et 3 % du total des travaux effectués sur le DT.

Il est également à noter que la plupart des travaux entrepris sur le DT (76 % du total) ont concerné une seule forme écrite de ce dialecte, à savoir celle transcrite dans l'alphabet arabe, malgré la présence importante de contenus dialectaux produits sur le Web en latin. Ainsi, 15 % d'entre eux ont considéré les deux scripts et seulement 9 % d'entre eux ont considéré le DT transcrit en latin. Mis à part le fait que le DT constitue une variante de la langue arabe et est, par conséquent, naturellement écrit en arabe, d'autres raisons peuvent expliquer la focalisation de plusieurs travaux sur la transcription arabe du DT. Parmi ces raisons, l'idée d'exploiter la proximité du DT avec l'ASM afin d'utiliser des ressources et des outils disponibles dédiés à l'arabe pour générer des ressources ou faire des traitements en dialectal, et qui a constitué une approche adoptée par plusieurs chercheurs.

5.2. *Approches de traitement du DT*

Les principales approches adoptées pour le traitement du DT peuvent être classées selon la langue des ressources utilisées ainsi que les méthodes de traitement proposées. Une synthèse de ces approches est présentée dans le tableau 12. Elle distingue les approches fondées sur l'utilisation exclusive de ressources spécifiques au DT de celles utilisant d'autres ressources en ASM et autres dialectes.

Approches	
Ressources utilisées	**Méthode**
DT	Statistique (à base d'apprentissage automatique)
	Linguistique (à base de règles)
	Hybride
ASM	Par adaptation à base de règles, de ressources et d'outils
	Par traduction manuelle de ressources
ASM + autres dialectes arabes	Par traduction manuelle de ressources

Tableau 12. *Synthèse des approches adoptées dans le traitement du DT*

Une comparaison rigoureuse de l'efficacité de ces deux types d'approches ne peut être réalisée, vu le nombre réduit de travaux de part et d'autre pour chaque type d'application, ainsi que l'utilisation de métriques d'évaluation et jeux de données différents. Nous pouvons, cependant, dégager les principales limites et problèmes affectant leur efficacité. Les approches qui s'appuient sur l'utilisation de ressources existantes en ASM et autres dialectes englobent, essentiellement, celles fondées sur la traduction manuelle ou l'adaptation, moyennant des règles préétablies de RL existantes afin de générer leurs équivalents en DT. Elles englobent aussi, les approches fondées sur une

adaptation (essentiellement à base de règles) d'outils d'analyse morphosyntaxique de l'ASM, afin de traiter le DT. Bien que très utiles pour la construction de ressources et d'outils du DT, ces approches posent plusieurs problèmes dont le principal est lié aux étymologies des mots dialectaux. En fait, le DT, comme mentionné dans la section 2, est caractérisé par une interférence linguistique entre l'ASM et d'autres langues comme le français, l'espagnol, le turc, l'italien, etc. Ainsi, de nombreux mots utilisés dans le DT ne proviennent pas de l'ASM et dérivent d'une langue complètement étrangère. De plus, le multilinguisme caractérisant les Tunisiens fait que le DT subit des changements remarquables de jour en jour avec l'introduction de nombreux mots en DT possédant des racines en langues étrangères, auxquelles s'ajoutent de nouveaux suffixes et préfixes et pouvant générer diverses formes fléchies (cf. section 2). Par conséquent, l'idée de s'appuyer uniquement sur l'analogie avec l'ASM serait insuffisante et parfois inefficace. Dans (Almeman et Lee, 2012), on montre que l'analyseur morphologique de l'ASM peut analyser seulement 32 % des mots dialectaux. De plus, l'utilisation de ce type d'approche limite le traitement du DT à celui de sa forme arabe uniquement, ce qui demeure insuffisant pour traiter le dialecte tunisien tel qu'il est largement produit sur le Web. En effet, et comme nous l'avons indiqué dans la section 2.2, l'écriture en latin du dialectal est beaucoup plus utilisée par les Tunisiens dans leurs communications électroniques (SMS, e-mails, Facebook, Twitter, etc.). Les approches ayant recours à des ressources spécifiques au DT se sont principalement fondées sur des méthodes à base de règles. Elles ont été utilisées dans des travaux d'analyse morphosyntaxique du DT, en ayant recours à des ressources annotées de taille relativement réduite (telles que le corpus de McNeil (2012) comprenant uniquement 2 000 mots annotés). Les méthodes fondées sur l'apprentissage automatique ont été principalement adoptées dans des applications telles que l'identification, l'analyse des sentiments et la translittération en utilisant des corpus généralement extraits du Web. Notons cependant, qu'à la date de cette étude et à notre connaissance, des approches à base d'apprentissage profond (*deep learning*), très prisées aujourd'hui dans le domaine du TAL, n'ont pas encore été explorées dans les travaux portant sur le DT. Cela pourrait s'expliquer entre autres par la non-disponibilité de ressources volumineuses indispensables pour l'efficacité de ces méthodes.

5.3. *Ressources linguistiques du DT*

Malgré les efforts fournis dans plusieurs travaux pour la construction de diverses RL pour le DT comme les corpus, lexiques, dictionnaires, etc., celles-ci restent encore relativement limitées en taille (variant de 3 à 800 k mots pour les corpus et de 2 à 40 k entrées pour les lexiques) et en nombre (tableau 11). De plus, ces RL ont souvent concerné le vocabulaire d'un domaine précis et limité. Nous citons à cet effet, les travaux de Graja *et al.* (2010), Karoui *et al.* (Karoui *et al.*, 2013a ; Karoui *et al.*, 2013b) et Neifar *et al.* (2014), dont les données ont concerné les interactions entre le personnel et les clients dans les gares, avec un vocabulaire limité aux réservations, achat de billets, heure, prix, etc., ou encore le travail de Hassine *et al.* (2016) qui a ciblé la prononciation des chiffres de 0 à 9. La construction de ressources de taille et couverture

significatives reste donc une priorité pour pouvoir étudier et traiter le dialecte tunisien, d'autant plus que le Web (et en particulier le Web social) constitue aujourd'hui une source importante de données en dialectal.

6. Conclusion

Cet article présente un état de l'art du traitement automatique du dialecte tunisien. Il commence par une revue des principales caractéristiques linguistiques de ce dialecte ainsi que les difficultés qu'il présente quant à son traitement. Il présente ensuite une revue des principaux travaux ayant porté sur le DT et recense les principales ressources linguistiques et outils TAL actuellement disponibles sur le Web pour ce dialecte. Cette étude a clairement démontré que malgré l'intérêt croissant qu'il a suscité durant ces dernières années, chez de nombreux chercheurs, le traitement du DT est encore à ses débuts et qu'en termes de ressources et outils de TAL qui lui sont dédiés, ce dialecte demeure encore une langue peu dotée. Très peu de travaux ont été effectués sur son analyse linguistique en se limitant au niveau morphosyntaxique. Les applications TAL impliquant le DT se sont principalement limitées à l'identification, la traduction DT-ASM, l'analyse des sentiments et la translittération latin arabe du DT.

Des efforts restent encore à fournir dans la construction de larges ressources linguistiques en DT à rendre disponibles. En cela, les médias sociaux, très largement utilisés par les Tunisiens, constituent une importante source à exploiter. De telles ressources sont indispensables pour l'étude et le développement d'outils et d'applications pour le DT. Elles permettront également de mettre en œuvre des méthodes d'apprentissage profond, qui ont été jusqu'ici très peu explorées dans les travaux sur le DT.

En outre, entreprendre des travaux sur le DT transcrit en latin qui, jusque-là, a été très peu abordé, nous semble d'une grande importance si l'on veut traiter le dialecte tunisien tel qu'il est largement produit sur le Web. Dès lors, les problèmes posés par l'identification du DT et sa translittération vers l'arabe seront des problèmes principaux à résoudre lorsque l'on s'intéresse au traitement des contenus dialectaux produits par les utilisateurs des médias sociaux.

7. Bibliographie

Abdelkader R. B., « Peace Corps English-Tunisian Arabic Dictionary », *ERIC Clearinghouse [Washington, D.C.]*, 1977.

Adouane W., Semmar N., Johansson R., Bobicev V., « Automatic Detection of Arabicized Berber and Arabic Varieties », *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects*, Osaka, Japan, p. 63-72, 2016.

Ahmed B., Cha S.-H., Tappert C., « Language Identification from Text Using N-gram Based Cumulative Frequency Addition », *Proceedings of Student/Faculty Research Day, CSIS, Pace University*, New York, US, 2004.

Almeman K., Lee M. G., « Towards Developing a Multi-Dialect Morphological Analyser for Arabic », *Proceedings of the 4th International conference on Arabic language processing*, Rabat, Morocco, 2012.

Ameur H., Jamoussi S., Hamadou A. B., « Exploiting emoticons to generate emotional dictionaries from facebook pages », *Intelligent Decision Technologies*, p. 39-49, 2016.

Aridhi C., Achour H., Souissi E., Younes J., « Word-Level Identification of Romanized Tunisian Dialect », *Proceedings of the 22nd International Conference on Applications of Natural Language to Information Systems*, Liège, Belgium, p. 170-175, 2017.

Baccouche T., « L'emprunt en arabe moderne », *Beit Al-Hikma–Carthage et I.B.L.V.– Université de Tunis I*, 1994.

Baccouche T., Mejri S., « Atlas linguistique de Tunisie : du littéral au dialectal », *Institut de recherche sur le Maghreb contemporain*, vol. , p. 387-399, 2004.

Bahou Y., « Compréhension Automatique de la Parole Arabe Spontanée : Intégration dans un Serveur Vocal Interactif », *Université de Sfax*, 2014.

Belgacem M., « Construction d'un corpus robuste de différents dialectes arabes », *Proceedings of Actes des VIII èmes RJC Parole*, Avignon, France, 2009.

Belgacem M., Antoniadis G., Besacier L., « Automatic Identification of Arabic Dialects », *Proceedings of the 7th edition of the Language Resources and Evaluation Conference*, Malta, 2010.

Bouamor H., Oflazer N. H. K., « A multidialectal parallel corpus of arabic », *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland, 2014.

Bouchlaghem R., Elkhlifi A., Faiz R., « Tunisian dialect Wordnet creation and enrichment using web resources and other Wordnets », *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, Doha, Qatar, p. 104-113, 2014.

Boudlal A., Lakhouaja A., Azzeddine M., Abdelouafi M., « Alkhalil Morpho Sys1: A Morpho-syntactic analysis System for Arabic texts », *Proceedings of the 2010 International Arab Conference on Information Technology (ACIT'2010)*, Benghazi, Libya, 2010.

Boujelbane R., « Génération de corpus en dialecte tunisien pour l'adaptation de modèles de langage », *Proceedings of Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, 2013.

Boujelbane R., Ellouze M., Béchet F., Belguith L., « De l'arabe standard vers l'arabe dialectal : projection de corpus et ressources linguistiques en vue du traitement automatique de l'oral dans les médias tunisiens », *TAL*, 2014.

Boujelbane R., Khemekhem M. E., Belguith L. H., « Mapping Rules for Building a Tunisian Dialect Lexicon and Generating Corpora », *Proceedings of the International Joint Conference on Natural Language Processing*, Nagoya, Japan, 2013a.

Boujelbane R., Khemekhem M. E., BenAyed S., Belguith L. H., « Building Bilingual Lexicon to Create Dialect Tunisian Corpora and Adapt Language Model », *Proceedings of the 2nd Workshop on Hybrid Approaches to Translation, ACL*, Sofia, Bulgaria, 2013b.

Boujelbane R., Zribi I., et Mariem Ellouze S. K., « An Automatic Process for Tunisian Arabic Orthography Normalization », *Proceedings of the tenth International Conference on Natural Language Processing (HrTAL2016)*, Dubrovnik, Croatia, 2016.

Buckwalter T., « Issues in Arabic orthography and morphology analysis », *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages*, Geneva, Switzerland, p. 31-34, 2004.

Cohen W. W., « Fast effective rule induction », *Proceedings of the 12th International Conference on Machine Learning*, Tahoe City, California, p. 115-123, 1995.

Eldesouki M., Samih Y., Abdelali A., Attia M., Mubarak H., Darwish K., Kallmeyer L., « rabic Multi-Dialect Segmentation: bi-LSTM-CRF vs. SVM », *3CoRR*, 2017.

Elimam A., « Du Punique au Maghribi Trajectoires d'une langue sémito-méditerranéenne », *Synergies Tunisie*, vol. 1, p. 25-38, 2009.

Elimam A., « Le maghribi, vernaculaire majoritaire à l'épreuve de la minoration », *ENSET – Oran*, 2012.

Elkateb S., Black B., Rodríguez H., Alkhalifa M., Vossen P., Pease A., Fellbaum C., « Building a wordnet for arabic », *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genoa, Italy, 2006.

Graja M., Jaoua M., Belguith L. H., « Lexical Study of A Spoken Dialogue Corpus in Tunisian Dialect », *Proceedings of the International Arab Conference on Information Technology*, Benghazi-Libya, 2010.

Graja M., Jaoua M., Belguith L. H., « Building Ontologies to Understand Spoken Tunisian Dialect », *International Journal of Computer Science, Engineering and Applications (IJCSEA)*, 2011a.

Graja M., Jaoua M., Belguith L. H., « Towards Understanding Spoken Tunisian Dialect », *Proceedings of the 18th International Conference on Neural Information Processing (ICONIP)*, Shanghai, China, 2011b.

Graja M., Jaoua M., Belguith L. H., « Discriminative Framework for Spoken Tunisian Dialect Understanding », *Proceedings of the 1st International Conference on Statistical Language and Speech Processing (SLSP13)*, Tarragona, Spain, 2013.

Graja M., Jaoua M., Belguith L. H., « Statistical Framework with Knowledge Base Integration for Robust Speech Understanding of the Tunisian Dialect », *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2015.

Habash N., Diab M., Rambow O., « Conventional orthography for dialectal Arabic », *Proceedings of the 8th International Conference on Language Resources and Evaluation*, Istanbul, Turkey, 2012.

Habash N., Rambow O., « MAGEAD: A Morphological Analyzer and Generator for the Arabic Dialects », *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, 2006.

Hamdi A., Boujelbane R., Habash N., Nasr A., « The Effects of Factorizing Root and Pattern Mapping in Bidirectional Tunisian - Standard Arabic Machine Translation », *Proceedings of the MT Summit 2013*, Nice, France, 2013a.

Hamdi A., Boujelbane R., Habash N., Nasr A., « Un Système de Traduction de Verbes entre Arabe Standard et Arabe Dialectal par Analyse Morphologique Profonde », *Proceedings of Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, Les Sables d'Olonne, France, 2013b.

Hamdi A., Gala N., Nasr A., « Automatically building a Tunisian Lexicon for Deverbal Nouns », *Proceedings of the 1st Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, Dublin, Ireland, p. 95-102, 2014.

Hamdi A., Nasr A., Habash N., Gala N., « POS-tagging of Tunisian Dialect Using Standard Arabic Resources and Tools », *Proceedings of the Second Workshop on Arabic Natural Language Processing*, Beijing, China, p. 59-68, 2015.

Harrat S., Meftouh K., Abbas M., Jamoussi S., Saad M., Smaili K., « Cross-Dialectal Arabic Processing », *Proceedings of the 16th International Conference on Intelligent Text Processing and Computational Linguistics*, Cairo, Egypt, 2015.

Harrat S., Meftouh K., Smaili K., « Creating Parallel Arabic Dialect Corpus: Pitfalls to Avoid », *Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing*, Budapest, Hungary, 2017.

Hassine M., Boussaid L., Massaoud H., « Tunisian Dialect Recognition Based on Hybrid Techniques », *International Arab Journal of Information Technology*, 2018.

Hassine M., Boussaid L., Messaoud H., « Maghrebian dialect recognition based on support vector machines and neural network classifiers », *International Journal of Speech Technology*, vol. 19, n° 4, p. 687—695, 2016.

Huang J. Z., « Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values », *Data Mining and Knowledge Discovery*, 1998.

Karmani N. B., Alimi A. M., « Construction d'un Wordnet standard pour l'Arabe tunisien », *Colloque pour les Etudiants Chercheurs en Traitement Automatique du Langage naturel et ses applications*, Sousse, Tunisia, 2016.

Karmani N. B., Soussou H., Alimi A. M., « Building a standardized Wordnet in the ISO LMF for aeb language », *Proceedings of the 7th Global Wordnet Conference (GWC 2014), Association for computational linguistics*, Tartu, Estonia, p. 71-77, 2014.

Karmani N. B., Soussou H., Alimi A. M., « Tunisian Arabic aebWordnet: Current state and future extensions », *Proccedings of the 2015 First International Conference on Arabic Computational Linguistics*, Cairo, Egypt, 2015.

Karoui J., Graja M., Boudabous M. M., Belguith L. H., « Domain Ontology Construction from a Tunisian Spoken Dialogue Corpus », *International Conference on Web and Information Technologies (ICWIT 2013)*, Hammamet, Tunisia, 2013a.

Karoui J., Graja M., Boudabous M. M., Belguith L. H., « Semi-automatic Domain Ontology Construction from Spoken Corpus in Tunisian Dialect: Railway Request Information », *International Journal of Recent Contributions from Engineering, Science and IT (iJES)*, vol. 1, n° 1, p. 35-38, 2013b.

Lachachi N., Adla A., « Identification Automatique des Dialectes du Maghreb », *Revue Maghrébine des Langues (RML10)*, vol. , p. 85-101, 2016a.

Lachachi N., Adla A., « Two approaches-based L2-SVMs reduced to MEB problems for dialect identification », *International Journal of Computational Vision and Robotics*, 2016b.

Lachachi N.-E., Adla A., « GMM-Based Maghreb Dialect Identification System », *Journal of Information Processing Systems*, vol. 11, n° 1, p. 22-38, 2015.

Maamouri R., Bies A., Kulick S., Ciul M., Habash N., Eskander R., « Developing an egyptian arabic treebank: Impact of dialectal morphology on annotation and tool development », *Proceedings of the 9th International Conference on Language Resources and Evaluation*, Reykjavik, Iceland, 2014.

Malmasi S., Refaee E., Dras M., « Arabic Dialect Identification using a Parallel Multidialectal Corpus », *Proceedings of the 14th Conference of the Pacific Association for Computational Linguistics (PACLING 2015)*, Bali, Indonesia, p. 209-217, 2015.

Masmoudi A., Bougares F., Ellouze M., Estève Y., Belguith L., « Automatic speech recognition system for Tunisian dialect », *Language Resources and Evaluation*, vol. 52, n° 1, p. 249-267, 2017.

Masmoudi A., Ellouze M., Bougares F., Estève Y., Belguith L., « Conditional Random Fields for the Tunisian Dialect Grapheme-to-Phoneme Conversion », *Proceedings of INTERSPEECH 2016*, San Francisco, USA, 2016.

Masmoudi A., Estève Y., Khmekhem M. E., Bougares F., Belguith L. H., « Phonetic Tool for the Tunisian Arabic », *Proceedings of the 4th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU'14)*, St. Petersburg, Russia, 2014a.

Masmoudi A., Habash N., Ellouze M., Estève Y., Belguith L. H., « Arabic Transliteration of Romanized Tunisian Dialect Text: A Preliminary Investigation », *Proceedings of the 16th International Conference on Intelligent Text Processing and Computational Linguistics*, Cairo, Egypt, 2015.

Masmoudi A., Khemakhem M. E., Estève Y., Belguith L. H., Habash N., « A Corpus and Phonetic Dictionary for Tunisian Arabic Speech Recognition », *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC 2014)*, Reykjavik, Iceland, 2014b.

Masmoudi A., Khemakhem M. E., Estève Y., Bougares F., Dabbar S., Belguith L. H., « Phonétisation automatique du dialecte tunisien », *Proceedings of the 14th Conference of the Pacific Association for Computational Linguistics (PACLING 2015)*, Le Mans, France, 2014c.

McNeil K., « Tunisian Arabic Morphological Parser », *Ling-420*, 2012.

McNeil K., « Tunisian Arabic Corpus: A written corpus of an "unwritten" language », *Proceedings of the International Symposium on Tunisian and Libyan Arabic Dialects*, Vienna, Austria, 2015.

McNeil K., Faiza M., « Tunisian Arabic Corpus: Creating a written corpus of an "unwritten" language », *Proceedings of Workshop on Arabic Corpus Linguistics*, Lancaster, UK, 2011.

Mdhaffar S., Bougares F., Estève Y., Hadrich-Belguith L., « Sentiment Analysis of Tunisian Dialect: Linguistic Resources and Experiments », *Proceedings of the 3rd Arabic Natural Language Processing Workshop (WANLP)*, Valencia, Spain, p. 55-61, 2017.

Meftouh K., Harrat S., Jamoussi S., Abbas M., Smaili K., « Machine Translation Experiments on PADIC: A Parallel Arabic DIalect Corpus », *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, shanghai, China, 2015.

Mejri S., Said M., Sfar I., « Pluringuisme et diglossie en Tunisie », *Synergies Tunisie*, vol. 1, p. 53-74, 2009.

Mekki A., Zribi I., Ellouze M., Belguith L. H., « Syntactic Analysis of the Tunisian Arabic », *Proceedings of the International Workshop on Language Processing and Knowledge Management*, Sfax, Tunisia, 2017.

Mohamed W. N. H. W., Salleh M. N. M., Omar A. H., « A comparative study of reduced error pruning method in decision tree algorithms », *Proceedings of the 2012 IEEE International Conference on Control System, Computing and Engineering*, US, 2012.

Mohand T., « Substrat et convergences: Le berbere et l'arabe nord-africain », *Estudios de Dialectologia*, vol. 4, p. 99-119, 1999.

Mourtada R., Salem F., Alshaer S., « The Arab Social Media Report 2014: Citizen Engagement and Public Services in the Arab World : The Potential of Social Medai », *MBR School of Government*, 2014.

Mzoughi I., « Intégration des emprunts lexicaux au franć cais en arabe dialectal tunisien Linguistique », *Université de Cergy Pontoise*, 2015.

Neifar W., Bahou Y., Graja M., Jaoua M., « Implementation of a Symbolic Method for the Tunisian Dialect Understanding », *Proceedings of the 5th International Conference on Arabic Language Processing (CITALA 2014)*, Oujda, Maroc, 2014.

Novotney S., Schwartz R., Khudanpurn S., « Getting more from automatic transcripts for semi-supervised language modeling », *Computer Speech and Language*, vol. 36, p. 93—109, 2016.

Och F. J., Ney H., « A systematic comparison of various statistical alignment models », *Computational Linguistics archive*, vol. 29, nº 1, p. 19-51, 2003.

Pereira C., « Arabe maghrébin », *Proceedings of Actes du Colloque International Langues d'Europe et de la Méditerranée LEM*, Nice, France, 2005.

Reynolds D. A., Quatieri T. F., Dunn R. B., « Speaker Verification Using Adapted Gaussian Mixture Models », *Digital Signal Processing*, 2000.

Saadane H., Nouvel D., Seffih H., « Une approche linguistique pour la détection des dialectes arabes », *Actes de TALN 2017*, 2017.

Sadat F., Kazemi F., Farzindar A., « Automatic identification of arabic dialects in social media », *Proceedings of the 1st international workshop on Social media retrieval and analysis*, Gold Coast, Australia, p. 35-40, 2014a.

Sadat F., Kazemi F., Farzindar A., « Automatic Identification of Arabic Language Varieties and Dialects in Social Media », *Proceedings of the 2nd Workshop on Natural Language Processing for Social Media (SocialNLP)*, Dublin, Ireland, p. 22-27, 2014b.

Sadat F., Mallek F., Sellami R., Boudabous M. M., Farzindar A., « Collaboratively Constructed Linguistic Resources for Language Variants and their Exploitation in NLP Applications-the case of Tunisian Arabic and the Social Media », *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing*, Dublin, Ireland, p. 102-110, 2014c.

Saidi D., « Développement de la compétence narrative en arabe tunisien : rapport entre formes linguistiques et fonctions discursives », *Université Lyon 2*, 2014.

Salem F., « The Arab Social Media Report 2017: Social Media and the Internet of Things: Towards Data-Driven Policymaking in the Arab World », *MBR School of Government*, 2017.

Sayadi K., Liwicki M., Ingold R., Bui M., « Tunisian Dialect and Modern Standard Arabic Dataset for Sentiment Analysis », *Proceedings of the 2nd International Conference on Arabic Computational Linguistics*, Konya, Turkey, 2016.

Sayahi L., « Diglossia and language contact: Language variation and change in North Africa », *Cambridge University Press*, 2014.

Stolcke A., « SRILM: An Extensible Language Modeling Toolkit », *Proceedings of the 8th European Conference on Speech Communication and Technology, EUROSPEECH 2003 - INTERSPEECH*, Geneva, Switzerland, 2003.

Suwaileh R., Kutlu M., Fathima N., Elsayed T., Lease M., « ArabicWeb16: A New Crawl for Today's Arabic Web », *Proceedings of the 39th annual international ACM SIGIR conference*

on Research and development in information retrieval: SIGIR '16, Pisa, Italy, p. 673-676, 2016.

Vapnik V., « The Nature of Statistical Learning Theory », *Springer New York*, 1995.

Younes J., Achour H., Souissi E., « Constructing Linguistic Resources for the Tunisian Dialect Using Textual User-Generated Contents on the Social Web », *Proceedings of the 1st International Workshop on Natural Language Processing for Informal Text (NLPIT 2015) In conjunction with The International Conference on Web Engineering (ICWE 2015)*, Rotterdam, The Netherlands, 2015.

Younes J., Souissi E., « A quantitative view of Tunisian dialect electronic writing », *Proceedings of the 5th International Conference on Arabic Language Processing*, Oujda, Morocco, p. 63-72, 2014.

Younes J., Souissi E., Achour H., « A Hidden Markov Model for Automatic Transliteration of Romanized Tunisian Dialect », *Proceedings of the 2nd International Conference on Arabic Computational Linguistics*, Konya, Turkey, 2016.

Zbib R., Malchiodi E., Devlin J., Stallard D., Matsoukas S., Schwartz R., Makhoul J., Zaidan O. F., Callison-Burch C., « Machine translation of Arabic dialects », *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Montreal, Canada, p. 49-59, 2012.

Zribi I., Boujelbane R., Masmoudi A., Khemekhem M. E., Belguith L. H., Habash N., « A Conventional Orthography for Tunisian Arabic », *Proceedings of the 9th International Conference on Language Resources and Evaluation*, Reykjavik, Iceland, 2014.

Zribi I., Boujelbane R., Masmoudi A., Khemekhem M. E., Belguith L. H., Habash N., « Spoken Tunisian Arabic Corpus STAC: Transcription and Annotation », *Research in Computing Science*, 2015.

Zribi I., Kammoun I., Khemekhem M. E., Belguith L. H., Blache P., « Sentence Boundary Detection for Transcribed Tunisian Arabic », *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, Bochum, Germany, 2016.

Zribi I., Khemekhem M. E., Belguith L. H., Blache P., « Morphological Disambiguation of Tunisian Dialect », *Journal of King Saud University - Computer and Information Sciences*, 2017.

Zribi I., Khmekhem M. E., Belguith L. H., « Morphological analysis of Tunisian Dialect », *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP 2013)*, Nagoya, Japan, 2013.