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

Rahma Boujelbane, Mariem Ellouze, Frédéric Béchet, Lamia Belguith

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

Rahma Boujelbane[1,2] — Mariem Ellouze[1] — Frédéric Béchet [2] — Lamia Belguith[1]

[1] *Multimedia, InfoRmation Systems and Advanced Computing Laboratory, Sfax 3021, TUNISIE*
rahma.boujelbane@gmail.com ; mariem.ellouze@planet.tn ; l.belguith@fsegs.rnu.tn
[2] *Laboratoire d'Informatique Fondamentale de Marseille- CNRS - UMR 7279 Université Aix-Marseille*
prenom.nom@lif.univ-mrs.fr

RÉSUMÉ. *Dans ce travail, nous nous intéressons aux problèmes liés au traitement automatique de l'oral parlé dans les médias tunisiens. Cet oral se caractérise par l'emploi de l'alternance codique entre l'arabe standard moderne (MSA) et le dialecte tunisien (DT). L'objectif consiste à construire des ressources utiles pour apprendre des modèles de langage dédiés à des applications de reconnaissance automatique de la parole. Comme il s'agit d'une variante du MSA, nous décrivons dans cet article une démarche d'adaptation des ressources MSA vers le DT. Une première évaluation en termes de couverture lexicale et de perplexité est présentée.*

ABSTRACT. *In this work, we focus on the problems of the automatic treatment of oral spoken in the Tunisian media. This oral is marked by the use of code-switching between the Modern Standard Arabic (MSA) and the Tunisian dialect (TD). Our goal is to build useful resources to learn language models that can be used in automatic speech recognition applications. As it is a variant of MSA, we describe in this paper an adjustment process of the MSA resources to the TD. A first evaluation in terms of lexical coverage and perplexity is presented.*

MOTS-CLÉS : *corpus oral, dialecte tunisien, modèle de langue, ressources.*

KEYWORDS: *oral corpus, Tunisian Dialect, Language model, resources.*

1. Introduction

Le terme *langue arabe* est aujourd'hui utilisé à la fois pour désigner une norme utilisée dans les milieux de l'éducation connue sous le nom de *Modern Standard Arabic* (MSA) et un certain nombre de langues vernaculaires parlées connues sous le nom de dialectes arabes (DA). Pendant longtemps, la seule forme connue de ces DA était la forme orale familière, ils étaient absents à la fois de tout document écrit, mais aussi des médias officiels où les locuteurs professionnels étaient tenus de s'exprimer en MSA. De nos jours, les DA sont représentés, à la fois sous forme de textes dans les réseaux sociaux, les textes en ligne sur Internet, mais aussi dans les médias où les émissions de débat et d'interview font intervenir des locuteurs non professionnels s'exprimant dans leur langue naturelle. Les différences entre les DA et le MSA vont au-delà des différences de registre existant dans d'autres langues (officiel *vs* informel). Les deux variétés de la langue arabe, le MSA et les DA, coexistent dans un état de diglossie (Fishman, 1967) : *situation où sont en usage deux langues apparentées génétiquement et structurellement et dont les distributions fonctionnelles sont complémentaires*. La plupart des ressources existantes pour la langue arabe se limitent au MSA, conduisant à une abondance d'outils pour le traitement automatiques de cette variété. Étant donné les différences significatives entre le MSA et les DA, les performances de ces outils s'écroulent lors du traitement des DA par des outils MSA. Les différences se retrouvent notamment au niveau lexical où plusieurs formes de mots graphiquement similaires, surtout en l'absence des voyelles courtes, ne sont pas apparentées sémantiquement. Ce qui conduit à une augmentation notable de l'ambiguïté dans les approches computationnelles des DA. Par conséquent, la création de ressources telles que des lexiques spécifiques pour chaque dialecte est cruciale. L'étude linguistique des différents dialectes, notamment à travers les relations de chacun d'eux avec l'arabe standard peut permettre d'améliorer leur traitement automatique. Dans ce contexte quelques DA ont commencé à être étudiés pour la traduction automatique (Salloum et Habash, 2013), (Zbib *et al.*, 2012) et la reconnaissance de la parole (Soltau *et al.*, 2011a), en traitant particulièrement les dialectes du Moyen-Orient. Les travaux décrits dans cette étude s'inscrivent dans ce cadre à travers la modélisation de la langue parlée dans les médias tunisiens. Cette source de données contient une quantité importante d' *alternance codique* (AC) entre la langue normative MSA et la langue parlée. Les ressources nécessaires pour modéliser le dialecte tunisien étant quasiment inexistantes, nous proposons une méthode permettant de développer des ressources à partir du langage MSA pour le traitement automatique du dialecte tunisien (DT). Pour ce faire, nous avons adopté une approche qui consiste à adapter les ressources MSA au DT. Cette approche comporte trois phases à savoir : la phase du construction de lexique, la phase de génération de corpus en DT et la phase d'évaluation de ressources. Concernant la première phase, nous avons étudié tout d'abord les différences entre les unités lexicales MSA et DT. Ensuite, nous avons essayé de construire pour les unités lexicales du DT des représentations similaires à celles du MSA. Enfin, nous avons traduit ces correspondances dans des dictionnaires bilingues MSA-DT. Dans la deuxième phase, nous avons proposé une méthode automatique de conversion de corpus MSA au DT. La troisième phase consiste à évaluer la qualité des

ressources produites. Nous proposons d'évaluer le corpus obtenu dans le contexte de la reconnaissance automatique de la parole (RAP) en mesurant l'impact de la couverture lexicale et de la perplexité d'un modèle de langage appris sur un tel corpus et testé sur des transcriptions d'émissions de télévision tunisiennes contenant à la fois du MSA et du DT. Le plan de cet article est le suivant : la section 2 décrit les spécificités du corpus de médias tunisiens collecté, transcrit et annoté. La section 3 présente d'abord une étude succincte sur les travaux antérieurs traitant le traitement automatique des DA en général et le DT en particulier. Elle finit par présenter l'approche proposée pour la création de ressources dédiées à la construction d'un modèle de langage pour l'oral parlé dans les médias tunisiens. Les sections 4 et 5 détaillent les étapes de cette approche. Enfin, la section 6 présente une évaluation du corpus produit.

2. Diglossie et alternance codique dans les médias tunisiens

La situation linguistique en Tunisie est caractérisée par une diglossie entre la langue normative (le MSA) et la langue usuelle (le DT) (Baccouche, 1974). Cette situation se retrouve dans tous les pays arabes. D'un côté il y a le MSA qui est la langue de la littérature et des journaux, elle n'est parlée que dans des contextes particuliers tels que l'enseignement ou les déclarations officielles. D'un autre côté, il y a l'arabe dialectal qui est la langue pratiquée par tous les tunisiens. Elle présente quelques variantes régionales aux niveaux phonologique et lexical sans poser aucun obstacle à l'intercompréhension entre variantes.

La langue dialectale, de par son caractère utilitaire, a évolué beaucoup plus rapidement que la langue classique. On peut considérer maintenant qu'il s'agit de deux langues, bien qu'elles soient clairement apparentées. Comme il est précisé dans (Boukadida, 2008), l'arabe dialectal se distingue de l'arabe classique par une syntaxe simplifiée, un lexique plus riche en vocables étrangers et une phonologie altérée.

Baccouche (1974) distingue deux niveaux dans les registres de MSA :

– l'arabe littéral classique utilisé dans les écrits religieux et certains recueils littéraires de haute tenue stylistique ;

– l'arabe littéral moderne représenté par la langue journalistique, les livres scientifiques. Il est le plus utilisé dans l'enseignement.

Il distingue aussi deux niveaux dans les registres de DT :

– le dialectal populaire (familier) qui véhicule les besoins quotidiens ;

– le dialectal intellectualisé, auquel nous nous intéressons dans ce travail et qu'on retrouve dans les conversations des lettrés dans les émissions radiophoniques et télévisées (Boukadida, 2008). Ce dialecte se présente comme un mélange entre le MSA et le DT. Ce dernier, quoiqu'il soit énormément stigmatisé et dévalorisé, est bien présent dans les émissions tunisiennes.

L'usage du dialecte dans les réseaux sociaux sur Internet est également en train de le modifier, en passant d'une langue purement orale à une langue écrite, sans normalisation ni standard d'orthographe bien établis.

2.1. *Corpus d'étude : description*

Dans le contexte de la recherche sur les dialectes arabes, les données orales recueillies par les chercheurs ne sont pas toujours librement accessibles et à la disposition de l'ensemble de la communauté scientifique. Des corpus de dialectes levantin ou égyptien sont disponibles auprès d'agences de création de ressources linguistiques comme le LDC (Language Data Consortium) ou ELRA (European Language Resources Association) mais, à notre connaissance, il n'existe aucun corpus en dialecte tunisien transcrit et annoté fourni par ces organismes. Certains travaux sur le DT familier (Graja *et al.*, 2010 ; Masmoudi *et al.*, 2014) ont permis de collecter un corpus de conversations dans des situations agent et client sur les renseignements (les tarifs des billets, réservations, etc.) dans des gares tunisiennes. Ce corpus est à notre connaissance le seul exemple de corpus en DT. Cependant, la petite taille du vocabulaire employé et le champ sémantique très limité des conversations enregistrées en font un corpus inadéquat pour modéliser l'oral des médias et représenter ses spécificités. Par conséquent, la construction d'un corpus oral de type DT intellectualisé s'est avéré indispensable pour cette étude. Aujourd'hui, il n'existe pas de normes ni d'outils pour la transcription automatique du DT. La tâche de transcription manuelle est d'autant plus difficile qu'il n'y a pas de conventions de transcription admises par la communauté scientifique. De fait, avant de commencer la transcription, nous avons développé une convention d'écriture nommée CODA (Zribi *et al.*, 2014). Puis, nous avons adopté cette convention pour transcrire cinq heures et vingt minutes d'enregistrements recueillis principalement depuis une chaîne télévisée tunisienne. Le logiciel que nous avons utilisé pour la transcription est *Transcriber* [1]. La thématique principale de ces enregistrements est la politique. Il s'agit soit de journaux télévisés, soit d'émissions de débat politique. Les journaux sont animés par un présentateur unique, qui introduit des reportages ou des séquences sur des sujets divers et invite quelquefois une personne liée à l'actualité. Les émissions de débat rassemblent un groupe de personnes discutant du sujet à l'ordre du jour. Les locuteurs dans ces émissions sont tous des locuteurs natifs de DT. Dans ces programmes, nous distinguons l'usage simultané de deux langues dans le même énoncé, la même proposition et parfois le même syntagme. L'emploi du DT dépend du type d'émission : nous avons remarqué que dans les journaux les mots dialectaux apparaissent beaucoup plus chez les intervenants dans les interviews ou les invités que chez les présentateurs ; en revanche, dans les émissions de débat, il n'y a pas d'habitude langagière, chacun veut défendre son idée en mélangeant les langues. Le tableau 1 montre quelques statistiques sur le corpus transcrit. Comme nous pouvons le voir le pourcentage de mots en DT est bien

1. http ://transcriber.softonic.fr/

plus important dans les débats que dans les journaux télévisés (37,2 % contre 21,4 %).

Type d'émission	Nombre d'heures	Nombre de mots (occurrence)	Nombre de mots (types)	Mots DT
Journaux télévisés	1 h 42 min 52 s	12 207	4504	21,4 %
Émissions de débat	3 h 40 min	25 757	6110	37,2 %

Tableau 1. *Statistiques sur le corpus transcrit*

2.2. *Voyellation et alternance codique*

La norme orthographique que nous avons développée (Zribi *et al.*, 2014) n'impose aucune contrainte sur la voyellation des textes : chacun a la liberté de choisir, selon ses besoins, s'il voyelle ou pas les transcriptions. Usuellement, les textes écrits en MSA ne sont pas voyellés, ce qui ajoute de l'ambiguïté dans le traitement automatique car une même forme sans voyelles peut correspondre à plusieurs mots voyellés. Pour traiter cette ambiguïté, beaucoup de travaux de recherche ont été proposés dans la communauté du TALN pour pallier ce manque de voyelles tels que les travaux de diacritisation de textes MSA (Roth *et al.*, 2008 ; Elshafei *et al.*, 2006). À l'oral, cette information est disponible car tous les locuteurs prononcent les voyelles. C'est pour cela que nous avons choisi de voyeller les mots lors de la transcription manuelle de notre corpus afin de décrire le plus précisément possible, sans ambiguïté de voyellation, la langue dans les médias tunisiens. Cette étude nous a permis de distinguer quatre types de mots : les mots (avec voyelles) en arabe standard (MSA) ; les mots en dialecte (DT) ; les mots en MSA contenant des affixes en dialecte (DT*) et enfin les mots MSA dont la voyellation suit les règles de la langue dialectale (MSA*). Le texte suivant présente un extrait de transcriptions avec voyelles où nous distinguons plusieurs niveaux de variations entre le MSA et le DT que nous avons annotés comme suit :

– DT : mot DT
– DT* : mot MSA avec des affixes DT (AC)
– MSA : mot MSA avec des voyelles MSA
– MSA* : mot MSA avec des voyelles DT

Présentateur :
لَكِن *lakin* : MSA/(mais) انتِ *ānti* : MSA* /(tu)تَعرَف *taʿaraf* : MSA*/(sais) مَا *mā* : MSA/(ne) يخفَاكِش *yẖfākiš* : DT*/(te cache pas) الكَلَام *ālklām* : MSA*/(le discours) اللّي *āl-ly* : DT/(qui) قَاعَد *qāʿad* : DT/(est en train) يِتقَال *yitqāl* : DT* /(d'être dit) هُنَا

hunā : MSA/(ici) وهُنَاك *whunāk* : MSA/(et là-bas) فِي *fiy* : MSA/(dans) **السفَارَة** *ālsfārah* : MSA*/(l'ambassade) **ألغَربِيّة** *alġarbiyah* : MSA/(étrangère) ولَهنَا *wlahnā* : DT/(et ici). **وتصير** *wtṣyr* : MSA*/(et elle se déroule) **لِقَاءَات** *liqāʾāt* : MSA/(des rencontres) **مَع** *mʿa* : MSA*/ (avec) كُبرَى *kubraā* : MSA/ (les grandes) **الشَخصِيَات** *ālšaḥṣiyāt* : MSA/(personnages) **الوَطنيّة** *ālwaṭanyah* : MSA/(nationales) **لِطَبخَة** *liṭabḥah* : MSA/(pour une cuisine) **مُعَينَة** *muʿaynah* : MSA*/(particulière) شنُوَّة *šnuwwah* : DT /(qu'est ce que) **رَايِك** *rāyik* : MSA* / (tu penses) ?

Invité :
برشَا *bršā* : DT/beaucoup de **كَلَام** *klām* : MSA/(discours) غَالط *ġālṭ* : DT /(faux) شوف *šwf* : DT/(regarde) **انَا** *ānā* : DT /(moi) بَاش *bāš* : DT /(je vais) نفَسر *nfasr* : DT*/(expliquer) **لِك** *lik* : MSA*/(à toi).

Ainsi les informations sur les voyelles sont particulièrement importantes pour détecter l'alternance codique entre le MSA et le DT. Cependant, la majorité des outils de traitement automatiques tels que les analyseurs morphosyntaxiques de l'arabe comme celui de Buckwalter Buckwalter (2004) Beesley (1998) ou l'analyseur MADA (Habash *et al.*, 2009) n'analysent que des textes non diacritisés à cause du manque de ressources arabes voyellées. Par conséquent, si l'entrée est partiellement voyellée, ces analyseurs commencent par éliminer tous les diacritiques avant d'effectuer l'analyse. Les analyseurs morphosyntaxiques de l'arabe ne profitent donc pas des diacritiques présents dans les textes pour désambiguïser les mots. La nécessité de disposer de très grands corpus de textes pour l'apprentissage de modèles de langage a également contraint les systèmes de reconnaissance automatique de la parole pour le MSA à utiliser des textes non voyellés pour apprendre les modèles. Malgré l'ambiguïté engendrée, de bons résultats ont pu être obtenus sous réserve de disposer d'assez de ressources (Mangu *et al.*, 2011).

Cette étude ayant comme objectif l'adaptation d'un système de RAP appris sur des corpus MSA pour le traitement du DT, nous proposons également d'omettre les voyelles de tous les mots du corpus transcrit. Ce prétraitement a cependant pour conséquence une augmentation de l'ambiguïté pour notre corpus par rapport aux corpus en MSA uniquement. En effet, la suppression des voyelles entraîne la transformation de plusieurs mots MSA* en des mots MSA. Certains mots ayant une étiquette DT* peuvent également se transformer en mots MSA en engendrant des ambiguïtés morphologiques et sémantiques. Par exemple, le mot : **نفَسَر** *nfasar* : DT*/ (j'explique) est un mot qui a une racine en MSA et les affixes de la première personne du singulier en DT. En éliminant les voyelles, ce mot pourra être analysé comme un mot MSA ayant une racine en MSA et les affixes de la première personne du pluriel en MSA qui signifie (*nous expliquons*). Nous présentons dans l'exemple ci-dessous, les transformations subies sur chaque mot après omission de voyelles.

Présentateur :
لكن *lkn* : MSA/(mais) **انتي** *ānty* : MSA/(tu) **تعرف** *tʕrf* : MSA/(sais) **مَا** *mā* : MSA/(ne) **يخفَاكش** *yẖfākš* : DT*/(te cache pas) **الكلَام** *ālklām* : MSA/(le discours) **اللّي** *āl-ly* : DT/(qui) **قَاعَد** *qāʕad* : DT/(est en train) **يتقَال** *ytqāl* : DT*/ (d'être dit) **هنَا** *hnā* : MSA /(ici) **وهنَاك** *whnāk* : MSA /(et là-bas) **في** *fy* : MSA/(dans) **السفَارة** *ālsfārh* : MSA/(l' ambassade) **الغربية** *ālġrbyh* : MSA/(étrangère) **ولهنَا** *wlhnā-* : MSA /(et ici) **وتصير** *wtṣyr* : MSA/(et elle se déroule) **لقَاءَات** *lqāʔāt* : MSA/(des rencontres) **مع** *mʿ* : MSA/(avec) **كبرَى** *kbrā* : MSA/(les grandes) **الشخصيَات** *ālšẖṣyāt* : MSA/(personnages) **ألوطنية** *alwṭnyh* :MSA/(nationales) **لطبخة** *lṭbẖh* : MSA/(pour une cuisine) **معينة** *mʕynh* : MSA/(particulière) **شنوّة** *šnwwh* : DT/(qu'est ce que) **رَايك** *rāyk* : MSA/(tu penses) ?
Invité :
برشَا *bršā* : DT/beaucoup de **كلَام** *klām* : MSA/(discours) **غَالط** *ġālṭ* : DT /(faux) **شوف** *šwf* : DT/(regarde) **انَا** *ānā* :DT /(moi) **بَاش** *bāš* :DT /(je vais) **نفسر** *nfsr* : MSA/(expliquer) **لك** *lk* : MSA/(à toi).

2.3. *Agglutination et alternance codique*

En complément des ambiguïtés supplémentaires dues à l'AC dans la voyellation des mots, le phénomène d'agglutination, caractéristique de la langue arabe, est fortement touché par cette dualité MSA-DT.

La langue arabe est fortement agglutinante : des articles, des conjonctions, des prépositions, matérialisés par des clitiques, se rattachent aux formes fléchies. On distingue généralement les proclitiques qui se situent avant la forme fléchie et les enclitiques qui se situent après. Contrairement à la plupart des langues latines, les articles, les prépositions ou encore les pronoms se collent aux adjectifs, noms, verbes et particules auxquels ils se rapportent. Comparé au français, un mot arabe peut parfois correspondre à toute une phrase. Par exemple, le verbe MSA **أَتَتذكّرُونَنَا** *ʔatḏkkruwnanā* correspond en français à la phrase : *Est-ce que vous vous souvenez de nous ?* et le verbe DT **وشرِيتُوهَاشي** *wšriytuwhāšy* correspond en français à la phrase : *et est-ce que vous l'avez acheté*. Cette caractéristique engendre des ambiguïtés morphologiques au cours de l'analyse (Belguith *et al.*, 2007). À cette morphologie riche des mots arabes, s'ajoute le problème d'AC intra-mot. Plusieurs natures de mots peuvent exister dans le corpus impliquant des ambiguïtés dans l'analyse des mots. Le tableau 2.3 illustre six différents types de mots présents dans le corpus DT :

– les types 1 et 2 décrivent l'AC intra-mot que peut porter les formes nominales ;

– le type 3 représente les verbes qui ont un préfixe DT et un lemme ayant une racine commune entre le MSA et le DT. Cependant, en l'absence de contexte ce mot peut être un mot MSA ;

– le type 4 présente une alternance codique intra-mot que peut porter une forme verbale : la racine est en MSA alors que le suffixe est en DT ;

– le corpus contient aussi des mots entièrement MSA ou DT pouvant être synonymes (types 5 et 6).

1) Préfixe MSA + racine DT	بالتصَاوِر *biāltṣāwir*	avec les photos
2) Préfixe DT + racine MSA	هَالأَسعَار *haālʿasʿaār*	ces prix
3) Préfixe DT + racine MSA DT	نَعرَف *naʿraf*	je sais
4) Préfixe MSA + racine MSA + enclitique DT	يَقتَنِيهَاش *yaqtanyhāš*	il ne l'achète pas
5) Mot MSA	أَسعَار *ʿasʿaār*	prix
6) Mot DT	أَسوَام *ʿaswām*	prix

Tableau 2. *Les différents types de mots dans le corpus des médias tunisiens*

3. Traitement automatique du dialecte intellectualisé

3.1. *Travaux en cours sur le dialecte tunisien*

Plusieurs travaux en cours visent le traitement automatique du DT. Les travaux présentés dans (Graja *et al.*, 2011) consistent à construire un système capable de comprendre les énoncés oraux en dialecte tunisien de voyageurs dans une gare ferroviaire. Pour cela, les auteurs ont proposé de construire une ontologie du domaine et de la projeter sur les énoncés oraux afin de les annoter sémantiquement. Toutefois, le domaine applicatif reste assez limité avec un vocabulaire utilisé de petite taille. Par conséquent, les ressources en cours de développement sont restreintes à leur application. Par ailleurs, les travaux de Zribi *et al.* (2013) visent à adapter l'analyseur morphologique alkhalil (Boudlal *et al.*, 2010) au DT. L'analyseur ne fonctionne actuellement que pour les verbes DT. Dans une approche complémentaire, l'objectif des travaux de Hamdi *et al.* (2013a) est l'analyse syntaxique du DT en utilisant un parseur conçu pour le MSA. Pour cela, l'étude vise à adapter MAGEAD (*Morphological analyzer and generator of arabic dialect*) (Habash et Rambow, 2006) au DT afin de convertir automatiquement des textes DT au MSA. Une fois les textes convertis en MSA, un analyseur standard peut être appliqué. Cette dernière étude est très proche de nos objectifs étant donné que MAGEAD peut également fonctionner dans le sens inverse (convertir des textes MSA vers le DT). Mais dans son état actuel, MAGEAD ne peut convertir que

les verbes, ce qui est insuffisant dans notre volonté de produire un modèle de langage pouvant être utilisé dans un système de RAP. Nous pouvons donc confirmer que ni les ressources ni les approches proposées pour le DT ne sont, à ce jour, suffisantes pour créer des corpus de dialecte intellectualisé pouvant servir à apprendre des modèles de langage probabilistes. C'est pourquoi, nous nous sommes focalisés sur l'étude des travaux pour le développement des ressources pour les langues peu dotées étant donné que le DT peut être classé parmi ces langues.

3.2. *Création de ressources pour le traitement des langues peu dotées*

Plusieurs travaux ont tenté de pallier les problèmes liés à l'informatisation des langues peu dotées. Scherrer (2012), dans le but d'informatiser le dialecte existant en Suisse, a développé un système de traduction allemand standard et suisse allemand. Le système développé traduit, en se fondant sur un lexique bilingue, l'allemand standard vers n'importe quelle variété du continuum dialectal de la Suisse alémanique. Par ailleurs, les auteurs dans (Shaalan *et al.*, 2007) ont proposé un système de traduction du dialecte égyptien pour la construction d'un corpus parallèle EGY-MSA, et ce, en s'appuyant sur des règles de correspondance EGY-MSA. Récemment, les travaux sur l'adaptation des systèmes de RAP MSA au dialecte arabe ont commencé à émerger. Par exemple Kirchhoff et Vergyri (2005) utilise des transcriptions de conversations téléphoniques effectuées par le LDC (Language Data Consurtium) pour construire un système de reconnaissance automatique de la parole pour un domaine limité. La même démarche a été faite pour le levantin dans (Vergyri *et al.*, 2005). Aussi, dans le cadre du projet Gale, Soltau *et al.* (2011b) ont présenté le développement d'un système de RAP à grand vocabulaire pour le levantin. Ils ont bénéficié d'une grande quantité de données audio, les transcriptions associées et un grand corpus textuel en arabe. En effet, pour construire un corpus dialectal, ils ont développé un identificateur de dialectes pour sélectionner parmi l'ensemble des données celles qui sont en levantin. À part les dialectes, il existe plusieurs langues parmi les langues peu dotées qui n'ont pas de relation avec une langue bien dotée comme le somalien, le khmer, etc. Par exemple, pour pallier le manque de ressources en khmer, Seng (2010) a choisi les sites de nouvelles en khmer pour collecter les corpus textuels. La situation de l'arabe est particulière dans la mesure où les différentes variétés de l'arabe entretiennent une relation privilégiée avec le MSA pour lequel nous disposons de ressources importantes. Pour doter le DT en ressources, nous avons suivi une approche consistant à exploiter une langue apparentée et bien dotée en ressources (le MSA) afin de l'adapter à l'oral tunisien.

3.3. *Approche pour le traitement automatique du dialecte intellectualisé*

L'approche que nous proposons pour pallier le manque de ressources nécessaires au traitement automatique de l'oral intellectualisé, consiste à exploiter les ressources MSA existantes et à les adapter au DT. Pour ce faire, nous avons commencé par exploiter le corpus Arabic TreeBank (ATB MSA) pour développer des dictionnaires

MSA-DT. Puis, face à l'absence de corpus écrits en DT, nous avons tiré profit des dictionnaires construits pour générer à partir des corpus MSA disponibles, des corpus DT. Finalement, afin de tester la qualité des ressources produites, nous avons proposé de les utiliser pour adapter un modèle de langage MSA à l'oral des médias tunisiens. La figure 1 illustre cette approche.

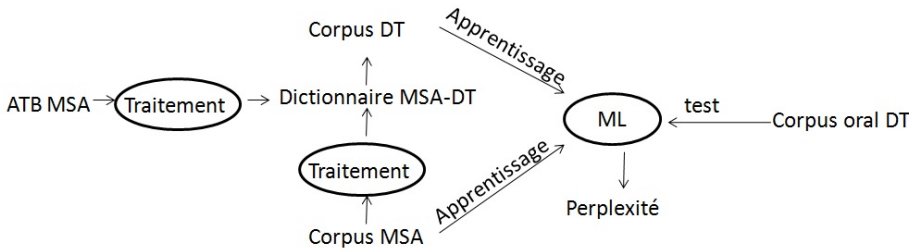

Figure 1. *Approche proposée pour le développement de ressources en DT, utilisant l'ATB MSA*

4. Dictionnaires bilingues MSA-DT

Notre approche de projection d'un corpus MSA vers un corpus DT s'appuie essentiellement sur un lexique bilingue MSA-DT. Cependant, les ressources disponibles sont rares ou presque inexistantes en dialecte et notamment en DT. Pour créer un tel lexique tunisien, nous avons adopté une méthodologie de transformation fondée sur les parties du discours des mots du corpus ATB (Maamouri et Bies, 2004). Ce corpus diffusé par l'agence LDC contient cent vingt transcriptions d'émissions d'actualité en arabe standard diffusées par différentes chaînes telles que : Abu Dhabi TV, Al Alam News Channel, Al Arabiya, Al Baghdadya TV, Al Fayha, Alhurra, Al Iraqiyah, etc. Le corpus transcrit contient 517 080 mots annotés morphosyntaxiquement et syntaxiquement. Le choix de l'ATB se justifie par la volonté de construire un lexique bilingue utilisant les structures syntaxiques en MSA pour les projeter sur le DT. Cette méthodologie a déjà été employée pour d'autres dialectes arabes tels que le levantin dans (Chiang *et al.*, 2006). De plus l'application de reconnaissance de la parole que nous souhaitons développer pour le DT concerne le traitement d'émissions d'informations similaires au niveau thématique à celles de l'ATB. Par conséquent, devant l'absence de corpus électroniques en DT, la conversion de l'ATB en DT pourra servir par défaut comme corpus d'apprentissage pour un modèle de langage DT pour la RAP.

Le but de ce travail est de projeter les entrées lexicales MSA de l'ATB vers le dialecte tunisien. Nous distinguons trois étapes : la projection des verbes, des noms, puis des mots-outils. Pour les deux premières étapes nous effectuons la projection des lemmes verbaux et nominaux sans prendre en compte le contexte, directement sur le lexique de l'ATB. Nous avons extrait les lemmes en utilisant l'analyseur morphologique ELEXIR FM(Smrž, 2007). Bien que certains lemmes puissent changer de sens

en changeant de contexte, cette méthodologie se justifie par le fait que notre principal but était de décrire des règles de transformation au niveau morphologique entre le MSA et le DT. L'adjonction du contexte pour résoudre certaines ambiguïtés pourra être l'objet d'une étude ultérieure. Le contexte est cependant pris en compte pour la projection des mots-outils dans notre dernière étape de transformation MSA-DT. La figure 2 présente le processus de construction de dictionnaires MSA-DT. Une des

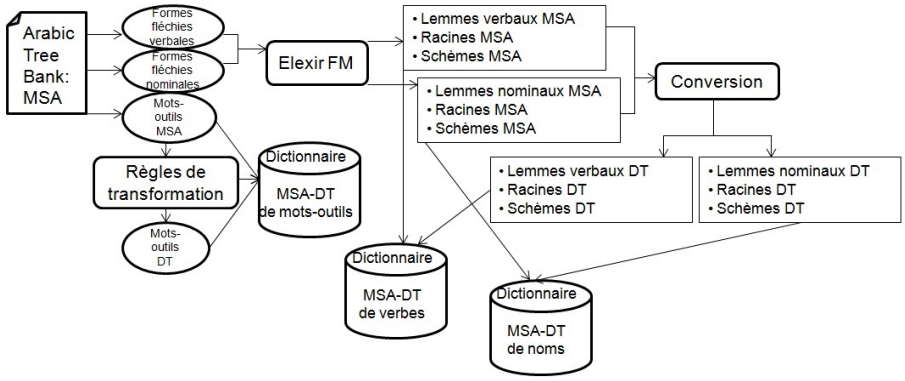

Figure 2. *Processus de construction de dictionnaires MSA-DT*

principales contributions de cette étude est l'effort de projection des structures morphologiques des noms et des verbes MSA vers le DT. En MSA, il existe trois concepts principaux pour décrire la morphologie des verbes et des noms :

– **racine** : c'est la base de toutes les formes de verbes et de certains noms arabes. La racine n'est pas un vrai mot mais plutôt une séquence de consonnes qui peuvent être trouvées dans tous les mots qui sont liés à elle. En MSA, la plupart des racines se composent de trois consonnes, très peu sont composées de quatre ou cinq consonnes ;

– **schème** : en MSA, les schèmes sont des modèles avec des structures différentes qui sont appliqués à la racine pour créer un lemme. À la même racine, nous pouvons appliquer différents schèmes pour avoir des lemmes différents avec des significations différentes, par exemple :

- Racine1 : xrj /خ ر ج/ *h* *r* *ǧ* + schème I : [FaEaL] ou [CaCaC] (F et E et L indiquent la position des consonnes de la racine) = خَرَجَ *haraǧ* / (sortir)

- Racine1 : xrj /خ ر ج/ *h* *r* *ǧ* + schème VI : [>aFEal] ou [>aCaCC] =axraj > / (faire sortir)

- Racine1 : xrj خ ر ج *h* *r* *ǧ* + schème XI : [AistafEal] ou [AistaCCaC] = إِسْتَخْرَجَ *ā'staḫraǧ* /(extraire) ;

– **lemme** : un lemme arabe peut être analysé comme une racine insérée dans un schème.

Ainsi, la connaissance du lemme ou du couple (racine, schème) permet de déduire les différentes formes fléchies d'un verbe ou d'un nom.

4.1. *Construction d'un dictionnaire MSA-DT de verbes*

Afin de définir les concepts de base pour les verbes DT, nous avons procédé à la démarche suivante :

– construire des lemmes verbaux en DT : l'ATB compte 29 911 unités lexicales de nature verbale dont 1 500 sont différentes. De ces dernières, nous avons extrait le lemme et la racine MSA. Certains des auteurs étant des locuteurs natifs du DT, nous avons construit manuellement pour les 1 500 lemmes MSA que nous avons recueillis, des lemmes en dialecte. En résultat, nous avons constaté que 60 % des verbes se comportent différemment en dialecte ce qui montre la différence lexicale entre le MSA et le DT ;

– construire des schèmes verbaux pour les verbes en DT : à l'issue de l'étape de lemmatisation, nous avons créé 1 500 lemmes en DT. Partant du principe que chaque verbe MSA appartient à un schème ou classe permettant de regrouper les verbes ayant les mêmes représentations morphologiques, nous avons essayé de construire des schèmes pour les verbes en DT. En MSA, il y a essentiellement quinze schèmes de verbes. Nous signalons que seul le premier schème (MSA-I) peut avoir six sous-schèmes différents selon la variation de la voyelle de la deuxième consonne de la racine au passé et au présent. Le tableau 3 présente la liste de ces sous-schèmes. Ainsi, MSA-I-au signifie que la voyelle de la deuxième consonne des verbes appartenant à ce sous-schème prend un a et un u au présent. Le reste des schèmes en MSA sont tous des superclasses c'est-à-dire qu'au sein d'un même schème tous les verbes ont la même voyelle de la deuxième consonne de leur racine. Nous avons présenté dans le tableau 3 le schème MSA-II. La voyelle de la deuxième consonne des verbes appartenant à ce schème, prend toujours un a au passé et un i au présent. En revanche, nous avons remarqué qu'au niveau des verbes en DT appartenant à ces schèmes, il y a variation au niveau de la voyellation de la deuxième consonne. Pour décrire cette variation, nous avons créé des sous-schèmes spécifiques au DT par exemple, DT-II-ii, DT-II-ai, etc. En outre, le MSA est caractérisé par le non-changement de la marque du présent au sein d'un même schème. Ceci n'est pas toujours vrai en DT, car cette marque peut changer toujours au sein d'un même schème. Pour cela, nous avons proposé d'étaler les sous-schèmes déjà créés et de faire d'autres sous-schèmes pour mettre en évidence la variation de la voyelle du signe du présent par exemple DT-I-aa-i et DT-I-aa-a ;

– construire des racines verbales en DT : en DT, il n'existe pas encore de standard pour la définition de la racine. Ainsi, la construction de racines en dialecte n'est pas évidente quand le verbe change complètement de racine en passant du MSA vers le DT. Dans notre cas, nous avons déjà défini le lemme (DT) et le schème (DT). On peut alors utiliser la règle MSA : racine + schème = lemme, et, en suivant une démarche déductive, en déduire une racine. En suivant cette règle, l'extraction de la racine est rendue alors facile. Par exemple, nous avons classé le lemme إِسْتَنَّى *ā'stannaā*/(attend)

Schèmes MSA	Sous-schèmes MSA	Schèmes DT	Sous-schèmes DT
MSA-I	MSA-I-aa	DT-I-aa	DT-I-aa-i DT-I-aa-a
	MSA-I-au	DT-I-au	DT-I-au-i DT-I-au-u
	MSA-I-ai	DT-I-ai	DT-I-ai-i DT-I-au-a
	MSA-I-uu	DT-I-uu	-
	MSA-I-ia	-	-
	MSA-I-ii	DT-I-ii	DT-I-ii-u DT-I-ii-a DT-I-ii-i
MSA-II	- -	DT-II-ii	-
		DT-II-ai	-
		DT-II-aa	-

Tableau 3. *Correspondance des schèmes MSA-I, MSA-II avec les schèmes DT-I, DT-II*

sous le schème إِسْتَفْعَل *ā'staf'al* ou (AistaCCaC). En respectant la règle d'extraction de racine, la racine est (ن *n* ن *n* ي *y*) ;

– définir les formes fléchies verbales en DT : à partir des racines, schèmes et lemmes obtenus pour le DT, nous pouvons obtenir maintenant notre lexique de formes fléchies. La flexion verbale du MSA est très régulière. Elle est fondée sur la concaténation d'affixes aux lemmes verbaux. La détermination des affixes repose sur les valeurs des traits morphologiques suivants : l'aspect (perfectif, imperfectif, impératif), le mode (indicatif, subjonctif, jussif, conditionnel), la personne, le genre et le nombre du sujet. En DT, la flexion des verbes tunisiens est plus pauvre que celle des verbes MSA. En particulier, le mode n'est plus marqué, les valeurs du nombre qui étaient au nombre de trois en MSA (singulier, duel et pluriel) sont réduites à deux (singulier et pluriel). Quant au genre, il n'est spécifié que lorsqu'il s'agit de la troisième personne du singulier.

À l'issue de ce processus nous avons construit un dictionnaire ayant comme entrée le couple (schème MSA, racine MSA) et son correspondant en DT (schème DT, racine DT). Pour un verbe arabe, la connaissance de ces deux caractéristiques permet de déduire les différentes formes fléchies. Pour cela, nous avons stocké dans notre dic-

tionnaire les différentes formes fléchies du verbe DT associées à leurs caractéristiques morphologiques.

4.2. *Construction d'un dictionnaire MSA-DT de noms*

Suivant le même processus que pour les verbes, nous avons tout d'abord construit manuellement une base de 1 050 lemmes traduits en DT à partir de l'ATB. Par correspondance aux noms MSA, nous avons classé ces lemmes DT en noms primitifs, nombres et noms dérivés :

– **les noms primitifs** sont des noms qui ne peuvent pas être rattachés à une racine verbale, beaucoup d'entre eux sont des noms concrets. Exemple كَلْب *kalb* / (chien), فَرَاشَة *farāšah* / (papillon). Dans cette catégorie, les noms à racine bilatérale sont aussi inclus, exemple اَبٌ *ābun* / (un père). Parmi les lemmes nominaux traduits, nous avons extrait 306 lemmes qui ne peuvent pas être rattachés à une racine verbale et nous leur avons attribué l'étiquette NTW (*Non Templatic Noun*) ;

– **les nombres** sont constitués de deux sous-catégories. La première, est elle-même constituée par les numéros simples représentant les unités : exemple صفر *ṣfr* / (zéro 0). La deuxième, regroupe les noms qui désignent la quantité des choses. exemple أَكْثَر *akṯar* / (supérieur) ;

– **les noms dérivés ou verbaux** sont les noms qui peuvent être dérivés à partir d'une racine verbale et d'un schème. Le nombre et la nature de ces formes varient selon le statut du verbe auquel ils se rattachent.

Une fois les lemmes obtenus, nous avons construit les schèmes nominaux DT. Par correspondance avec les schèmes MSA, nous avons commencé par classer les lemmes qui ont le même comportement au singulier (modèle + voyelle) dans le même schème. Puis, étant donné que chaque schème peut avoir plus qu'une forme au pluriel, nous avons créé pour chacun des sous-schèmes qui décrivent le comportement morphologique au pluriel. Ensuite, nous avons défini pour chaque schème le genre qu'il peut accepter car certains noms acceptent le féminin et le masculin à la fois alors que d'autres n'acceptent qu'un seul genre. Enfin, en se fondant sur les variations nominales, nous avons construit un dictionnaire qui a comme entrée le triplet (lemme MSA, schème MSA, racine MSA) et son correspondant en DT (lemme DT, schème DT, racine DT).

4.3. *Dictionnaires des mots-outils*

Les mots-outils ont un volume très important dans les textes arabes. Mais, leur transformation n'est pas directe : elle nécessite une étude des différents contextes de chacun d'eux. Nous avons dégagé trois types de transformations par l'étude des différents contextes des mots-outils ATB. La première est une transformation dépen-

dante du contexte : en effet, plusieurs traductions des particules MSA dépendent des contextes postérieur, antérieur ou les deux dans lesquels ils se trouvent. Par exemple pour avoir la bonne structure de négation en DT, il faut prendre en compte la structure MSA suivante : particule de négation + verbe MSA.
Le syntagme : لَمْ يَذْهَب *lam yaḏhab* / (il n'est pas parti) est nécessaire pour avoir la bonne traduction en DT : مَا مشَا ش *mā mšā š* / (il n'est pas parti). La deuxième transformation est directe ou indépendante du contexte : la traduction du mot-outil est la même quel que soit le contexte. Par exemple, la traduction de la particule لِمَاذَا *limāḏā* / (pourquoi) est quel que soit le contexte : علَاش *ʕlāš* / (pourquoi). La troisième transformation nécessite l'inversion de l'ordre de mots en passant du MSA vers le DT. Tel est le cas de certaines particules suivies d'adjectifs qui se transforment en DT en adjectif + particule. Le syntagme : كُتُبٌ كَثِيرَة *kutubun kaṯyrah* /(livres beaucoup) se transforme en DT en بَرْشَا كتُب *baršā ktub* / (beaucoup de livres).

5. Application de la conversion MSA-DT pour la modélisation de l'oral parlé dans les médias tunisiens

5.1. *Conversion MSA-DT pour la génération de corpus en dialecte intellectualisé*

En utilisant les dictionnaires bilingues MSA-DT de verbes, noms et de mots-outils décrits dans la section précédente, nous proposons une méthode de conversion de corpus MSA vers le DT. Ce processus commence par annoter morphosyntaxiquement le corpus MSA à l'aide de l'analyseur MADA (Habash *et al.*, 2009). MADA effectue à la fois la lemmatisation, l'analyse morphologique et l'étiquetage morphosyntaxique de chaque mot. Chaque couple (lemme MSA, POS) sera utilisé comme entrée dans les dictionnaires pour produire la traduction appropriée. Cette conversion est partielle étant donné que la couverture des dictionnaires que nous avons proposés est limitée. Par conséquent, la conversion peut ne pas être totale. Ceci peut être avantageux, puisque nous souhaitons avoir un corpus mixte (mot MSA, mot DT) afin de représenter l'oral parlé avec alternance codique entre les différentes langues dans les émissions. La figure 3 décrit le processus de conversion MSA-DT. Dans cette étude, l'évalua-

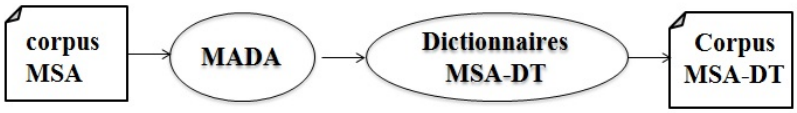

Figure 3. *Processus de conversion MSA-DT*

tion de ce processus de conversion se fait par rapport à une tâche de reconnaissance automatique de la parole. Nous construisons un modèle de langage (ML) pour l'oral

parlé dans les médias tunisiens appris sur de grands corpus MSA automatiquement projetés vers le DT grâce à notre méthode. L'impact des phénomènes d'agglutination sur la modélisation de langage en arabe peut être très important. En effet, la sortie du processus de conversion MSA-DT peut varier selon le degré de tokenisation choisi. Peu de travaux se sont intéressés aux problèmes de la reconnaissance automatique des dialectes arabes. Ce manque de données est amplifié par la morphologie très riche du MSA et de ses dialectes. En effet, en raison du grand nombre d'affixes pouvant être attachés à un même mot, la taille du vocabulaire est généralement plus importante que pour les autres langues si on considère un corpus d'apprentissage de même grandeur. Cette richesse de la morphologie de l'arabe conduit donc à des taux de mots hors vocabulaire (MHV) élevés et dégrade la qualité de l'estimation des modèles statistiques de langage pour la reconnaissance automatique de la parole ou la traduction. Un moyen pour traiter ce problème est d'utiliser un analyseur morphologique qui segmente chaque mot afin de réduire la taille du vocabulaire, et, par conséquent, la taille des modèles. La décomposition morphologique a été proposée par Lamel *et al.* (2008) pour le MSA et par Besacier (2007) pour le dialecte irakien afin de traiter le problème du lexique volumineux.

Pour le DT, le problème d'agglutination est également accentué par ceux liés à l'AC. Par exemple, en DT, à un même nom défini, vingt préfixes différents peuvent s'accoler donnant à chacun une forme différente. Ceci est problématique face à l'absence de corpus d'apprentissage volumineux et représentatif de l'oral des médias. Pour ces raisons nous avons opté pour la décomposition morphologique des mots. Dans la littérature, il existe deux types d'approches pour la décomposition morphologique : la première est fondée sur des règles linguistiques proposées par Vergyri *et al.* (2004)Xiang *et al.* (2006) et la deuxième s'appuie sur des méthodes non-supervisées (Adda-Decker, 2003 ; Goldsmith, 2001). Puisque les mots arabes ont un nombre limité d'affixes l'approche à base de règles donne généralement de bons résultats. Nous avons adopté cette approche pour proposer un modèle de langage DT.

5.2. *Tokenisation des corpus MSA-DT*

Afin de pouvoir capturer les règles de tokenisation, nous avons étudié les différentes structures des mots MSA et DT qui peuvent exister dans le corpus d'étude. Les verbes MSA admettent un seul enclitique, le pronom complément d'objet direct (PRND), qui varie en genre et en nombre et les proclitiques suivants présentés selon leur position, du plus éloigné au plus proche du verbe :

– CNJ : les conjonctions وَ *wa* / (et) et فَ *fa* / (alors) ;

– QST : la particule d'interrogation أَ *ʔa* / (est-ce que) ;

– PRP : la préposition لِ *li* / (pour) ;

– PRT : la particule de futur سَ *sa*.

Ainsi, la structure d'un verbe MSA peut être décrite par l'expression régulière suivante : CNJ ? QST ? PRP ? PRT ? Forme fléchie PRND ?

Dans les verbes DT, les proclitiques, QST, PRP, PRT disparaissent. En revanche, de nouveaux enclitiques apparaissent à savoir l'enclitique de négation **ش** *š* et d'interrogation **شي** *šy*. Ainsi, la structure d'un verbe DT peut être décrite par l'expression régulière suivante (Hamdi *et al.*, 2013b) : CNJ ?Forme fléchie PRND ? QST ? NEG ?

Par ailleurs, les noms admettent la même structure en MSA et DT qui peut être décrite par l'expression régulière suivante : CNJ ? PRT ? DEF ? Forme fléchie PRND ?. Le problème d'agglutination s'accentue avec l'apparition de nouveaux préfixes qui peuvent être agglutinés dans le corpus des médias aux noms MSA ou DT. En effet, les particules, telles que les prépositions et les pronoms démonstratifs MSA, qui étaient isolées du nom en MSA, se transforment en enclitiques collés aux noms MSA ou DT. Voici l'ensemble de préfixes qui peuvent être attachés à un nom dans notre corpus d'étude :

– 12 préfixes MSA avec l'article défini **ال** *āl/* (le, la) : **ال** *āl* , **وَال** *wāl* , **فَال** *fāl* , **بَال** *bāl* , **وبَال** *wbāl* , **فبَال** *fbāl* , **لِل** *lil* , **ولِل** *wlil* , **فلِل** *flil* , **كَال** *kāl* , **وكَال** *wkāl* , **فكَال** *fkāl* ;

– 8 préfixes DT avec l'article défini **ال** *āl/* (le, la) : **هَال** *hāl* , **وهَال** *whāl* , **عَال** *ʿaāl* , **وعَال** *wʿaāl* , **مَال** *māl* , **ومَال** *wmāl* , **بِهَال** *bihāl* , **وبِهَال** *wbihāl* ;

– 10 préfixes MSA sans l'article défini **ال** *āl/* (le, la) : **بِ** *bi*, **فَ** *fa*, **لِ** *li*, **كَ** *ka*, **وَبِ** *wabi*, **وَلِ** *wali*, **فَلِ** *fali*, **فَبِ** *fabi*, **وَكَ** *waka*, **فَكَ** *faka*.

Remarquons que les préfixes les plus fréquents sont ceux qui contiennent l'article défini **ال** *āl*. Il sont aussi les plus fréquents en occurrence dans le corpus. Pour cette raison, nous avons proposé un algorithme qui décompose chaque mot contenant dans son préfixe l'article défini **ال** *āl* ou la conjonction **و** *w*. Ainsi, chaque mot commençant par l'article **ال** *āl/* (le, la) sera décomposé en 'Al + mot'. Par exemple le mot : **الحُكُومَة** *ālḥukuwmah* / (le gouvernement) sera décomposé en : **ال** *āl* / (le) + **حُكُومَة** *ḥukuwmah* / (gouvernement). Si à l'article **ال** *āl* est agglutiné une particule telle que la conjonction **و** *w* ou autres, le mot sera décomposé en (particule + AL) + mot. Par exemple le mot DT : **وهَالحُكُومَة** *whālḥukuwmah* / (et ce gouvernement) sera décomposé en : **وَال** *wāl* / (et ce) + **حُكُومَة** *ḥukuwmah* / (gouvernement).

Par ailleurs, nous avons également remarqué que les mots (verbes, noms et particules) précédés par la conjonction 'w' sont aussi assez fréquents dans le corpus. C'est pourquoi, nous avons proposé de décomposer chaque mot préfixé par la conjonction **وَ** *wa/* (et) en 'w + mot'. Par exemple le mot : **ومشَى** *wmšaā* / (et il est parti) sera décomposé

en : و *w* / (et) + مشَى *mšaā* / (il est parti). Toutefois, en adoptant cet algorithme de décomposition un problème d'ambiguïté apparaît. En effet, il est parfois difficile de distinguer entre un proclitique ou enclitique et un caractère du mot en question. Par exemple, le caractère و *w* dans le mot وَصَل *waṣal* / (est arrivé) est un caractère qui fait partie de ce mot alors que dans le mot وفَتحَهُ *wfatḥahu* / (et a ouvert), il s'agit d'un proclitique (Belguith *et al.*, 2007). Ce problème se présente aussi en DT. Par exemple dans le le syntagme : عَلَى العَالَم *ʿalaā 'lʿaālam* / (sur le monde) où عَلَى *ʿalaā*/ (sur) : préposition et العَالَم *ālʿaālam* / (monde) : nom défini, se transforme en DT en un seul mot : عَالعَالَم *ʿaālʿaālam* / (sur le monde). Dans cet exemple, le caractère عَ *ʿa* avec lequel le nom indéfini عَالَم *ʿaālam* / (monde) commence, est un caractère qui fait partie du mot alors que dans le mot عَالعَالَم *ʿaālʿaālam*/ (sur le monde), il s'agit d'un préfixe DT.

Pour ces raisons, nous devons appliquer une méthode de tokenisation sur le corpus d'apprentissage et celui de test qui prend en compte la morphologie du mot, sa nature grammaticale pour pouvoir appliquer correctement l'algorithme de décomposition proposé. Pour le corpus d'apprentissage, étant donné qu'il est obtenu par notre méthode de conversion MSA-DT s'appuyant sur une analyse linguistique, le processus consiste uniquement à ajouter les règles de tokenisation lors du processus de conversion. Pour la tokenisation du corpus de test, étant donné qu'aucune annotation n'est disponible sur ces transcriptions d'émissions mélangeant MSA et DT, nous avons dû développer des ressources spécifiques au DT pour appliquer notre processus. Nous avons développé un outil qui tokenise automatiquement les conversations orales en exploitant les ressources développées dans (Boujelbane *et al.*, 2013). Nous avons adapté le *Part Of Speech POS tagger* MSA Stanford au DT en l'entraînant sur un corpus, annoté en parties du discours, généré par notre méthode. Puis, nous avons développé des patrons d'enrichissement morphologique, en exploitant les dictionnaires bilingues pour enrichir les POS tags avec d'autres traits morphologiques. Ces patrons ont permis la tokenisation des mots du corpus de test (Boujelbane *et al.*, 2014).

6. Évaluation du modèle de langage

En l'absence d'un système de reconnaissance de la parole, la mesure la plus couramment utilisée pour évaluer un modèle de langue est la mesure de perplexité. Cette dernière consiste à mesurer la capacité de prédiction d'un modèle de langage sur un corpus de test. Nous avons également évalué le modèle de langage en mesurant la couverture lexicale sur le corpus de test ce qui permet de déterminer le taux de mots hors vocabulaire (MHV). Nous souhaitons à travers ces deux mesures comparer un modèle appris sur des données MSA seulement, et des modèles intégrant des corpus convertis en DT ainsi que des corpus en DT transcrits manuellement. Ainsi, nous avons construit trois types de corpus :

– un corpus MSA : ce corpus contient 12 M de mots collectés depuis des dépêches de l'agence France-Presse (AFP), des transcriptions d'Aljazeera ainsi que diverses sources d'actualité provenant du Web ;

– un corpus MSA-DT : le corpus MSA traduit en DT *via* notre approche de conversion MSA-DT ;

– un corpus DT : le texte de la Constitution tunisienne qui a été récemment rédigé en DT, il contient 12 k mots. Même si ce corpus ne contient pas d'oral, il est un des rares exemples de textes directement écrits en DT.

Sur trois combinaisons différentes de ces corpus à savoir APP1, APP2 et APP3, nous avons appris trois modèles de langage de type 3-gram : APP1 contient le corpus MSA ; APP2 contient APP1 ainsi que le corpus MSA-DT ; enfin, afin de couvrir davantage de mots DT, nous avons construit un troisième corpus, APP3, qui englobe APP2 et le corpus DT de la Constitution tunisienne. Nous avons testé ces différents modèles sur le corpus de test décrit dans la section 2. Les tests ont été effectués d'abord sur les corpus de journaux télévisés, puis sur le corpus de débats et enfin sur l'ensemble.

La composition du vocabulaire a une influence très forte sur les performances des systèmes de RAP et notamment sur les mesures d'évaluation. Ainsi, nous avons testé trois alternatives pour le choix du vocabulaire. La première (VOCAB1) consiste à sélectionner les n-mots les plus fréquents de chaque corpus d'apprentissage. Nous avons fixé n à 90 k car c'est la valeur qui donne un compromis de résultats acceptables entre les MHV et la valeur de perplexité (PPL). Cependant, il est difficile de comparer deux valeurs de PPL obtenues avec des lexiques différents. En effet, l'amélioration de la nature du vocabulaire permet d'avoir une meilleure couverture lexicale ce qui induit une augmentation de perplexité par incorporation de mots peu probables dans le calcul, alors que dans le même temps, l'injection d'un corpus d'apprentissage, plus pertinent, fait baisser cette même perplexité. Ces changements de perplexité, de nature différente, ne conduisent pas à la possibilité d'analyser significativement une hausse ou une baisse. Pour cette raison nous évaluons aussi nos modèles en utilisant un vocabulaire fixe (VOCAB2) pour chaque corpus d'apprentissage. Nous calculons donc une nouvelle valeur de perplexité (PPL1), pour chaque corpus d'apprentissage, avec un même lexique. Ce dernier est fixé en sélectionnant les 90 k mots les plus fréquents pris du corpus APP3. Le tableau 4 illustre les résultats obtenus.

	Débats			Journaux télévisés			Débats + journaux télévisés		
	MHV	PPL	PPL1	MHV	PPL	PPL1	MHV	PPL	PPL1
APP1	12 %	960	NS [1]	7,02 %	666	695	10,39 %	845	903
APP2	7,2 %	774	793	4,05 %	543	554	6,2 %	686	687
APP3	6,8 %	770	770	3,6 %	529	529	6,2 %	682	682

Tableau 4. *Évaluation d'un ML avec VOCAB1 et VOCAB2*

1. Non Significatif

Les résultats montrent que l'introduction des corpus traduits a permis de diminuer aussi bien la perplexité que le nombre de MHV, sur tout type de corpus de test. Toutefois, nous avons remarqué dans la liste des mots inconnus qu'il y a encore des MHV dus à certains mots présentant une alternance codique MSA-DT intra-mot. Pour cette raison, nous avons effectué une nouvelle expérience en rajoutant au corpus d'apprentissage les corpus d'émissions de télévision tunisienne transcrits manuellement. Nous signalons que, dans le cas où le corpus de test est le corpus de journaux télévisés, nous augmentons APP3 par les transcriptions de débats et inversement si les corpus de test est celui des débats. Nous remarquons que l'introduction des transcriptions dans APP3 a permis d'améliorer les valeurs de MHV et de perplexité aux corpus de journaux télévisés et de débats. En revanche, en l'absence de ce type de corpus dans l'apprentissage, comme dans la troisième colonne du tableau 4, les valeurs de MHV et de perplexité dans APP3 ne s'améliorent presque pas par rapport à APP2. Le tableau 5 montre la typologie des mots inconnus en fonction du corpus d'apprentissage. En conclusion, nous pouvons tirer de cette évaluation que tout d'abord le fait que la méthode de projection MSA-DT est bénéfique pour la modélisation de l'oral des émissions télévisées. Ensuite, nous avons constaté que l'utilisation des transcriptions dans l'apprentissage apporte un gain significatif dans les performances des modèles. Ceci nous mène à penser qu'il serait intéressant de construire le vocabulaire d'apprentissage en utilisant un corpus de développement afin d'ajuster au mieux les proportions de mots MSA, DT et MSA-DT à utiliser pour représenter nos données. Néanmoins, ceci n'est pas possible dans le cadre du présent travail face à l'absence d'une quantité suffisante de transcriptions. Enfin, nous retenons du tableau 5 que le taux des mots inconnus MSA augmente en intégrant de nouveaux corpus d'apprentissage. Ceci est dû au fait que la taille du vocabulaire est fixe et aussi que les mots MSA, qui avaient des fréquences faibles dans le lexique du APP1, vont être éliminés et remplacés par les mots les plus fréquents du nouveau corpus APP2. Mais, étant donné que la traduction MSA-DT ne couvre pas tous les mots, le vocabulaire du corpus MSA-DT peut contenir des mots non traduits du corpus MSA. Par conséquent, leur fréquence dans le corpus APP2 et APP3 va augmenter ce qui entraîne l'élimination des mots DT ayant un poids faible.

	Mots avec alternance codique	Mots en DT	Mots en MSA
APP1	22,55 %	13,55 %	60,3 %
APP2	15,80 %	9,1 %	74,01 %
APP3	13,8 %	7,03 %	79,1 %

Tableau 5. *Typologie des mots inconnus en fonction du corpus d'apprentissage*

Ainsi, afin d'évaluer l'apport du vocabulaire DT nous avons opté pour la troisième alternative (APP3). Il s'agit de prendre une liste de 90 k mots MSA et de rajouter les mots DT provenant soit d'APP2, soit d'APP3. Le tableau 6 montre les résultats obtenus. La PPL1 est calculée en utilisant un vocabulaire fixe contenant 90 k mots MSA + 80 k mots DT provenant de APP2 et APP3.

	Débats			Journaux télévisés			Débats + journaux télévisés		
	MHV	PPL	PPL1	MHV	PPL	PPL1	MHV	PPL	PPL1
APP1	12,07 %	960	NS	7,02 %	666	756	10,39 %	845	NS
APP2	5,8 %	916	916	3,05 %	617	618	4,8 %	799	800
APP3	5,02 %	927	927	2,5 %	609	609	4,5 %	817	817

Tableau 6. *Évaluation d'un ML avec VOCAB3*

Les résultats du tableau 6 montrent que le taux des MHV reste relativement élevé, même en exploitant tous les mots DT à notre disposition. Ceci peut être dû au choix du corpus AFP dont les thèmes ne sont pas assez proches de ceux présents dans les actualités tunisiennes traitées dans cette étude. De plus le corpus recueilli depuis le web englobant ces actualités est petit par rapport aux corpus généralement utilisés pour apprendre un ML. Par ailleurs, cette augmentation dans la taille du vocabulaire a induit une augmentation de la perplexité et ce par incorporation de mots peu probables dans le calcul. Ce qui renforce l'idée d'utiliser un corpus de développement pour construire un vocabulaire pour la modélisation de l'oral des émissions télévisées tunisiennes.

7. Conclusion

Dans cet article, nous avons proposé une démarche permettant la construction de ressources pour le traitement automatique du dialecte tunisien parlé dans les médias. Ce dialecte a une situation particulière, il est constitué d'un mélange de MSA et de DT. Du fait que le DT est une variante de l'arabe standard, nous avons proposé une approche qui consiste à exploiter les ressources MSA pour développer des ressources en DT et ce en étudiant les correspondances qui peuvent exister entre le MSA et le DT. En effet, nous avons défini une méthode qui convertit, en utilisant un dictionnaire MSA-DT, des corpus MSA vers des corpus DT. Pour évaluer la qualité des corpus traduits, nous les avons utilisés pour construire un modèle de langue pour l'oral parlé dans les médias tunisiens afin de l'intégrer dans un système de reconnaissance automatique de la parole. Lors de cette application, nous avons montré l'intérêt de la décomposition morphologique pour la modélisation du langage MSA-DT. Nous envisageons, dans un travail futur, d'intégrer ce modèle dans un système de reconnaissance et évaluer les taux d'erreur de mots (WER). Nous souhaitons également améliorer la qualité de la traduction MSA-DT en augmentant la taille du lexique de traduction. A long terme, nous désirons exploiter les ressources développées (dictionnaires, analyseurs, tokeniseurs, modèle de langue, ...) pour le traitement automatique du langage utilisé dans les réseaux sociaux tunisiens.

8. Bibliographie

Adda-Decker M., « A corpus-based decompounding algorithm for German lexical modeling in LVCSR. », *INTERSPEECH*, 2003.

Baccouche T., « Esquisse d'une étude comparative des schémas des verbes en arabe classique et en arabe tunisien », *Les cahiers de Tunisie*, vol. 22, p. 87-88, 1974.

Beesley K. R., « Arabic morphological analysis on the Internet », *Proceedings of the 6th International Conference and Exhibition on Multi-lingual Computing*, Citeseer, 1998.

Belguith L. H., Aloulou C., Hamadou A. B., « MASPAR : De la segmentation à l'analyse syntaxique de textes arabes », *CÉPADUÈS-Editions, editeur, Revue Information Interaction Intelligence I*, vol. 3, p. 9-36, 2007.

Besacier L., « De l'utilisation d'unités sous-lexicales pour la traduction automatique de la parole », *Séminaire ATALA*, 2007.

Boudlal A., Lakhouaja A., Mazroui A., Meziane A., BEBAH M. O. A. O., SHOUL M., « Alkhalil Morpho SYS1 : A Morphosyntactic Analysis System for Arabic Texts », *International Arab Conference on Information Technology*, 2010.

Boujelbane R., Khemekhem M. E., Belguith L. H., « Mapping Rules for Building a Tunisian Dialect Lexicon and Generating Corpora », *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, p. 419-428, 2013.

Boujelbane R., Mallek M., Ellouze M., Belguith L. H., « Fine-Grained POS Tagging of Spoken Tunisian Dialect Corpora », *Natural Language Processing and Information Systems*, Springer, p. 59-62, 2014.

Boukadida N., Connaissances phonologiques et morphologiques dérivationnelles et apprentissage de la lecture en arabe (Etude longitudinale), PhD thesis, Université Rennes 2, 2008.

Buckwalter T., Buckwalter arabic morphological analyzer version 2.0. LDC catalog number LDC2004L02, Technical report, ISBN 1-58563-324-0, 2004.

Chiang D., Diab M. T., Habash N., Rambow O., Shareef S., « Parsing Arabic Dialects. », *EACL*, 2006.

Elshafei M., Al-Muhtaseb H., Alghamdi M., « Statistical methods for automatic diacritization of Arabic text », *The Saudi 18th National Computer Conference. Riyadh*, vol. 18, p. 301-306, 2006.

Fishman J. A., « Bilingualism with and without diglossia ; diglossia with and without bilingualism », *Journal of social issues*, vol. 23, n° 2, p. 29-38, 1967.

Goldsmith J., « Unsupervised learning of the morphology of a natural language », *Computational linguistics*, vol. 27, n° 2, p. 153-198, 2001.

Graja M., Jaoua M., Belguith L. H., « Building Ontologies to Understand Spoken Tunisian Dialect », *arXiv preprint arXiv :1109.0624*, 2011.

Graja M., Jaoua M., Hadrich Belguith L., « Lexical Study of A Spoken Dialogue Corpus in Tunisian Dialect », *The International Arab Conference on Information Technology (ACIT), Benghazi–Libya*, 2010.

Habash N., Rambow O., « MAGEAD : a morphological analyzer and generator for the Arabic dialects », *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, p. 681-688, 2006.

Habash N., Rambow O., Roth R., « Mada+ tokan : A toolkit for arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization », *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt*, p. 102-109, 2009.

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

Hamdi A., Boujelbane R., Habash N., Nasr A. *et al.*, « Un système de traduction de verbes entre arabe standard et arabe dialectal par analyse morphologique profonde », *Traitement Automatique des Langues Naturelles*p. 396-406, 2013b.

Kirchhoff K., Vergyri D., « Cross-dialectal data sharing for acoustic modeling in Arabic speech recognition », *Speech Communication*, vol. 46, n° 1, p. 37-51, 2005.

Lamel L., Messaoudi A., Gauvain J.-L., « Investigating morphological decomposition for transcription of Arabic broadcast news and broadcast conversation data », *INTERSPEECH*, p. 1429-1432, 2008.

Maamouri M., Bies A., « Developing an Arabic treebank : methods, guidelines, procedures, and tools », *Proceedings of the Workshop on Computational Approaches to Arabic Script-based languages*, Association for Computational Linguistics, p. 2-9, 2004.

Mangu L., Kuo H.-K., Chu S., Kingsbury B., Saon G., Soltau H., Biadsy F., « The IBM 2011 GALE Arabic speech transcription system », *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, IEEE, p. 272-277, 2011.

Masmoudi A., Khmekhem M. E., Estève Y., Belguith L. H., Habash N., « A Corpus and Phonetic Dictionary for Tunisian Arabic Speech Recognition », *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014.*, p. 306-310, 2014.

Roth R., Rambow O., Habash N., Diab M. T., Rudin C., « Arabic Morphological Tagging, Diacritization, and Lemmatization Using Lexeme Models and Feature Ranking », *ACL (Short Papers)*, p. 117-120, 2008.

Salloum W., Habash N., « Dialectal Arabic to English Machine Translation : Pivoting through Modern Standard Arabic. », 2013.

Scherrer Y., Generating Swiss German sentences from Standard German : a multi-dialectal approach, PhD thesis, University of Geneva, 2012.

Seng S., Vers une modélisation statistique multi-niveau du langage, application aux langues peu dotées, PhD thesis, Université de Grenoble, 2010.

Shaalan K., Bakr H. M. A., Ziedan I., « Transferring egyptian colloquial dialect into modern standard arabic », *International Conference on Recent Advances in Natural Language Processing (RANLP–2007), Borovets, Bulgaria*, p. 525-529, 2007.

Smrž O., « Elixirfm : implementation of functional arabic morphology », *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages : Common Issues and Resources*, Association for Computational Linguistics, p. 1-8, 2007.

Soltau H., Mangu L., Biadsy F., « From Modern Standard Arabic to Levantine ASR : Leveraging GALE for dialects. », *ASRU*, p. 266-271, 2011a.

Soltau H., Mangu L., Biadsy F., « From Modern Standard Arabic to Levantine ASR : Leveraging GALE for dialects. », *ASRU*, p. 266-271, 2011b.

Vergyri D., Kirchhoff K., Duh K., Stolcke A., « Morphology-based language modeling for arabic speech recognition. », *INTERSPEECH*, vol. 4, p. 2245-2248, 2004.

Vergyri D., Kirchhoff K., Gadde V. R. R., Stolcke A., Zheng J., « Development of a conversational telephone speech recognizer for Levantine Arabic. », *INTERSPEECH*, Citeseer, p. 1613-1616, 2005.

Xiang B., Nguyen K., Nguyen L., Schwartz R., Makhoul J., « Morphological decomposition for Arabic broadcast news transcription », *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1, IEEE, p. I-I, 2006.

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*, Association for Computational Linguistics, p. 49-59, 2012.

Zribi I., Boujelbane R., Masmoudi A., Ellouze M., Belguith L. H., Habash N., « A Conventional Orthography for Tunisian Arabic », *LREC*, p. 2355-2361, 2014.

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