TRAITEMENT AUTOMATIQUE DU DIALECTE TUNISIEN : CONSTRUCTION DE RESSOURCES LINGUISTIQUES

Inès Zribi

République Tunisienne

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université de Sfax

Faculté des Sciences économiques et de Gestion de Sfax

Thèse pour l'obtention du titre de docteur en :

Informatique

TRAITEMENT AUTOMATIQUE DU DIALECTE TUNISIEN : CONSTRUCTION DE RESSOURCES LINGUISTIQUES

Présentée et soutenue publiquement le 05 décembre 2016 par :

Inès Zribi

Membres du jury :

Mr. Ahmed HADJ KACEM	Professeur d'Ens. Sup., FSEG – Sfax	**Président**
Mme. Lamia HADRICH BELGUITH	Professeur d'Ens. Sup., FSEG – Sfax	**Directeur de thèse**
Mr. Philippe BLACHE	Professeur d'Ens. Sup., Aix-Marseille université – France	**Co-Directeur de thèse**
Mr. Nabil HATHOUT	Maître de conférences, Université de Toulouse Jean Jaurès – France	**Rapporteur**
Mme. Rim FAIZ	Professeur d'Ens. Sup., IHEC – Carthage	**Rapporteur**
Mr. Mohamed HAMMAMI	Maître de conférences. FS – Sfax	**Membre**

Année universitaire : 2016-2017

Dédicace

À mes chers parents

À ma adorable nièce

À mes frères

À mes belles-soeurs

À mes amies

À ma famille...

Remerciements

J'ai une vive dette envers tous ceux qui m'ont aidé à rassembler les faits qui constituent l'indispensable fondation de ce travail.

Mes respectueux remerciements s'adressent à ma directrice de thèse Mme. Lamia HADRICH BELGUITH, professeur à la faculté des sciences économiques et de gestion de Sfax, et mon codirecteur de thèse Mr. Philippe BLACHE, directeur de recherche au CNRS-LPL, pour avoir accepté d'être mes directeurs de thèse et pour m'avoir fait profiter de leurs expériences respectives. Je tiens à les remercier pour leur encadrement patient et exigeant, pour leur rigueur et leur compréhension tout au long de ces années de travail.

Je tiens aussi à exprimer ma profonde gratitude envers Mme. Mariem ELLOUZE Maître assistante à l'école supérieure de commerce de Sfax, pour ses conseils, son aide et ses encouragements, ainsi que l'intérêt continu qu'il a porté à mon travail. Les discussions que j'ai eues avec elle, ses remarques et ses relectures ont beaucoup contribué ce travail de recherche.

Je remercie le président de jury Mr. Ahmed HADJ KACEM, professeur à la faculté des sciences économiques et de gestion de Sfax, pour l'honneur qu'il m'accorde en jugeant ce travail.

Ainsi, je remercie vivement mes rapporteurs Mme. Rim FAIZ, professeur à l'Institut des hautes études commerciales de Carthage, et Mr. Nabil HATHOUT, directeur de recherche au CNRS CLLE/ERSS, qui m'ont fait le plaisir d'accepter la charge de rapporteur ainsi que pour leurs remarques constructives sur mon travail.

Par la même occasion, je remercie Mr. Mohamed HAMMAMI qui me fait l'honneur d'être l'examinateur de ce manuscrit.

Enfin, je remercie toute ma famille pour leur encouragement et incessant soutien moral tout au long de la préparation de ce travail et plus particulièrement mes parents qui m'ont guidé sur le chemin de la recherche.

Table des matières

Table des figures

Liste des tableaux

Introduction générale

Contexte et motivations

L'être humain exprime et communique ses idées, ses sentiments, ses croyances et ses valeurs à travers l'écriture, la langue des signes, les gestes et la voix. Cependant, la forme verbale orale reste la modalité de communication la plus utilisée et la plus facile à réaliser par rapport à celle écrite. L'Homme peut décrire ses pensées sans trop réfléchir aux bonnes expressions et aux règles de grammaire et de conjugaison. La production de l'oral est plus rapide ainsi qu'elle est plus spontanée que celle de l'écrit étant donnée qu'elle ne nécessite pas d'apprentissage particulier. Ainsi, on voit intéressant voir même fructueux de mettre l'accent sur la forme orale de communication.

De nos jours, les travaux de recherche en Traitement Automatique de Langue (TAL) et en Traitement Automatique de Parole (TAP) se sont orientés vers le traitement de cette forme de communication. En effet, beaucoup d'applications interactives utilisant la parole comme moyen de communication, ont fait leur apparition dans notre vie courante (serveurs de dialogue oral Homme-machine, systèmes d'extraction d'information à partir d'enregistrements audio, applications de téléphone mobile (Siri [1]), etc.). La plupart de ces applications ont été conçues pour interagir avec des locuteurs parlant les langues indo-européennes comme l'anglais et le français. La langue arabe, classée la sixième langue la plus parlée dans le monde, reste peu abordée au niveau de TAP notamment sa forme dialectale. Depuis 2011, les événements du « Printemps arabe » que les pays arabes ont témoigné, à savoir la Tunisie, ont su exercer une attraction irrésistible. Le **Dialecte Tunisien** (DT) est devenu de plus en plus présent et utilisé dans les réseaux sociaux, les blogs et les médias vu que les Arabes utilisent leur langue maternelle pour exprimer leur rage et leur mécontentement. Le besoin de le comprendre et de l'analyser est devenu progressivement impérieux. Par l'évolution des technologies de la parole et des applications en téléphonie, le traitement automatique de la forme dialectale de la langue arabe se dévoile d'une portée capitale.

La conception des systèmes capables d'analyser l'**Arabe Dialectal** (AD) parlé s'appuie sur les outils de traitement de signal en les combinant avec un ensemble de ressources (corpus et lexique) et d'outils de TAL (analyseur morphologique, syntaxique, sémantique, etc.).

Dans le cadre du traitement automatique de l'AD parlé s'inscrit ce sujet de thèse. Il s'agit de contribuer au traitement automatique du DT, en proposant des méthodes adéquates pour créer des ressources linguistiques (corpus et outils) en faveur de ce dialecte.

1. http ://www.apple.com/fr/ios/siri/

Problématiques

L'enjeu aujourd'hui est de proposer des systèmes capables de comprendre la forme orale spontanée d'une langue peu dotée mais très répandue, en l'occurrence du DT. Le traitement automatique du DT parlé est confronté à de nombreux défis. Certains sont partagés avec toutes les langues parlées. D'autres sont spécifiques aux caractéristiques de la langue traitée [Zouaghi *et al.* 2008].

- Problèmes liés au caractère spontané des productions orales : en effet, l'efficacité et la performance d'un système de traitement de la langue parlée dépendent de sa puissance à surmonter les difficultés liées au caractère spontané de l'oral qui rendent son traitement automatique plus délicat. De nombreuses hésitations, amorces, répétitions, allongements vocaliques et autres disfluences affectent la production, et, éventuellement, la compréhension de la production orale [Bouraoui 2008]. Ces disfluences nécessitent des traitements particuliers qui alourdissent le traitement automatique de la langue parlée [Bouraoui 2008].
- Problèmes liés aux caractéristiques du dialecte tunisien : la langue arabe se caractérise par des propriétés morphologiques et syntaxiques qui rendent son traitement automatique très complexe. La complexité de traitement de cette langue est encore plus sensible lorsqu'on parle du DT qui montre des différences phonologiques, morphologiques, lexicologiques, et syntaxiques par rapport à l'**Arabe Standard Moderne** (ASM) et à d'autres dialectes arabes [Bahou *et al.* 2008]. Les dialectes arabes ne sont pas officiellement écrits et ils n'ont pas d'orthographe standard ce qui rend la tâche de son traitement automatique plus difficile que l'ASM.

Objectifs et contributions

L'originalité de notre travail par rapport à l'existant vient de la volonté de traiter automatiquement un dialecte peu doté qui se caractérise, d'une part, par le manque de ressources, et d'autre part, par des différences significatives avec sa langue source (l'ASM). Cette thèse se donne deux principaux objectifs.

Le premier est d'entamer la tâche de création des ressources textuelles (corpus) pour le DT oral. Afin d'atteindre cet objectif, une première contribution consiste à aborder la tâche de *transcription orthographique* d'une langue non normalisée. Nous proposons deux conventions pour transcrire orthographiquement le DT parlé et annoter ses caractéristiques pour pallier le manque de règles orthographiques. La construction d'un corpus pour le DT parlé sera notre deuxième contribution. Ce corpus est le résultat d'un ensemble de transcriptions manuelles.

Le second objectif est le développement des outils linguistiques (analyseur morphologique et étiqueteur morphosyntaxique) pour la forme orale du DT en exploitant plusieurs méthodes de TAL et de TAP permettant de surmonter les problèmes rencontrés lors du traitement de l'écrit et de l'oral.

Dans le but de réaliser cet objectif, trois objectifs intermédiaires ont été identifiés :

- Étude de différentes approches du traitement automatique de l'AD et de l'ASM en vu de proposer des méthodes adéquates pour la forme orale du DT : nous proposons d'étudier la possibilité d'adapter et d'étendre les méthodes et techniques d'analyse de TAL de l'ASM et de l'AD pour le traitement du DT.
- Exploitation des ressources de l'ASM afin de créer un analyseur morphologique : notons que le traitement automatique de l'ASM a été étudié depuis plus de 30 ans. Un certain nombre d'outils ([Chaâben & Belguith 2004], [Aloulou *et al.* 2003], [Baccour *et al.* 2003]) et de ressources ([Trigui *et al.* 2010], [Hammami *et al.* 2009], [Bayoudhi *et al.* 2014]) ont ainsi été développés pour son traitement. L'utilisation et/ou l'adaptation de ressources et de données existantes peuvent être considérées comme un facteur facilitant le traitement de l'AD [Duh & Kirchhoff 2005], plutôt que de développer des outils spécifiques à un dialecte particulier. Dans ce cadre, nous visons à adopter cette approche pour la création d'un lexique pour le DT qui sera notre troisième contribution. Ainsi, nous dévoilons une méthode, à base de règles, pour créer ce lexique en DT, en partant d'un lexique en ASM. Nous testons une méthode peu supervisée pour enrichir ce lexique. Nous préconisons l'approche d'adaptation des ressources pour créer un *analyseur morphologique* en intégrant le lexique DT à un analyseur morphologique de l'ASM. La création d'un analyseur morphologique pour le DT sera notre quatrième contribution.
- Annotation morphosyntaxique du DT parlé : vu que la majorité des méthodes proposées dans la littérature ne prennent pas en considération les spécificités de l'oral de l'AD, nous avons fixé comme objectif de proposer une approche permettant, dans un premier temps, d'étiqueter morpho-syntaxiquement le DT et de tenir compte, dans un deuxième temps, de la présence du caractère spontané de l'oral tels que les mots incomplets, les pauses remplies, etc. La désambiguïsation morphosyntaxique du DT parlé sera notre cinquième contribution. Nous proposons des techniques statistiques pour désambiguïser le résultat de l'analyse morphologique. Étant donné que l'identification des frontières des phrases est une tâche non triviale, nous proposons un ensemble de méthodes (linguistique, statistique et hybride) pour détecter les frontières de phrases des transcriptions orales.

Structure du document

Cette thèse est organisée en trois parties.

Dans **la première partie**, nous proposons un état de l'art du traitement automatique des dialectes arabes, en général, et du DT, en particulier. Ainsi, nous présentons un aperçu des différentes méthodes utilisées dans le traitement automatique des dialectes arabes.

Le **chapitre 1** est consacré à une présentation des dialectes arabes et les principaux défis de son traitement automatique. Nous exposons, ensuite, un survol des travaux réalisés pour

la construction de corpus (écrit et parlé) et de ressources lexicales de l'AD. Nous clôturons ce chapitre par une étude des méthodes proposées pour développer les outils de traitement automatique de l'AD.

Le **chapitre 2** se concentre sur le dialecte tunisien dont nous présentons les principales caractéristiques phonologiques, morphologiques, lexicales et syntaxiques. Nos motivations pour le traitement de ce dialecte sont, ensuite, décrites. Nous terminons ce chapitre par une présentation des méthodes proposées pour la construction de ressources textuelles et pour le développement d'outils de traitement automatique du DT.

La création de ressources textuelles pour le DT fait l'objet de **la deuxième partie** de cette thèse.

Nous abordons dans **le chapitre 3** la problématique de la transcription orthographique du DT. Cette langue est peu dotée. Elle présente plusieurs difficultés lors de sa transcription orthographique. Ces dernières sont présentées dans la première section de ce chapitre. Ensuite, nous exposons deux conventions de transcription et d'annotation pour surmonter ces embarras. Nous concluons ce chapitre par une comparaison avec d'autres travaux réalisés pour d'autres dialectes arabes.

Le chapitre 4 décrit la création de notre corpus STAC « **Spoken Tunisian Arabic Corpus** ». Nous détaillons d'abord la collection des données en montrant l'application du schéma d'annotation et de transcription que nous utilisons. La validation de la qualité de notre corpus est effectuée par le calcul d'accord inter-annotateurs et intra-annotateur. Finalement, nous présentons les enrichissements que nous proposons pour améliorer la qualité de notre corpus pour le DT.

Dans **la troisième partie**, nous abordons la création d'outils linguistiques pour le traitement du DT.

Nous traitons le problème de l'analyse morphologique du DT et montrons que l'adaptation des ressources et des outils de l'ASM permet d'aboutir à des résultats encourageants. Nous commençons, ainsi, **le chapitre 5** par une étude du lexique du DT. Nous proposons, ensuite, une méthode d'adaptation d'un analyseur morphologique de l'ASM vers le tunisien. Nous détaillons par la suite les étapes de cette méthode. Enfin, nous concluons ce chapitre par la présentation de quelques expérimentations en discutant les résultats obtenus.

Nous entamons au niveau du **chapitre 6**, l'étiquetage morphosyntaxique du DT. Tout d'abord, nous exposons les problèmes rencontrés lors de l'étiquetage morphosyntaxique. Puis, une méthode pour la segmentation des transcriptions, qui représente l'un des problèmes à surmonter lors de l'étiquetage morphosyntaxique du DT, est présentée. Nous montrons, par la suite, que l'étiquetage morphosyntaxique est réalisé en désambiguïsant les résultats de l'analyse morphologique par différentes techniques d'apprentissage. Enfin, nous concluons ce chapitre par la présentation de résultats des expérimentations et la discussion de ses résultats.

Cette thèse est clôturée par une synthèse des principaux points traités dans ce travail. Nous faisons un bilan général des résultats de notre recherche et des problèmes actuels du traitement automatique d'un dialecte peu doté avec peu de ressources. Enfin, nous dégageons

quelques perspectives comme la détection automatique des disfluences et le développement d'autres ressources et outils pour le DT.

Première partie

Etat de l’art

CHAPITRE 1

Traitement automatique de l'arabe dialectal

Sommaire

1.1 Introduction

La langue arabe fait partie de la famille des langues sémitiques. Elle se caractérise par la présence de trois variétés (l'arabe classique, l'arabe standard moderne et l'arabe dialectal) dont l'arabe dialectal est la variété la plus utilisée et parlée au quotidien. L'évolution des médias électroniques et les technologies de communication a engendré une évolution de l'arabe dialectal au niveau de son utilisation [Belguith 2009]. Il devient de plus en plus une nécessité de l'employer dans les nouvelles technologies. De même, les événements géopolitiques et les évolutions dans les pays arabes (comme le printemps arabe) ont augmenté le besoin de traiter automatiquement les dialectes arabes, qui s'impose désormais comme une nécessité.

Dès le début des années 2000, le traitement automatique des dialectes arabes est devenu l'objet d'étude de plusieurs chercheurs en TAL. Mais, il reste encore aujourd'hui dans un état préliminaire. En particulier, l'existence de plusieurs dialectes dans le monde arabe et les différences entre eux constituent des défis qui compliquent son traitement automatique.

Dans ce chapitre, nous proposons une présentation des dialectes arabes et de leur traitement automatique. Nous présentons, tout d'abord, l'arabe dialectal en exposant une classification des dialectes et les principales différences et similitudes avec l'arabe standard. Ensuite, nous exposons dans la troisième section une description des défis posés par le traitement automatique de l'arabe dialectal. Enfin, dans les sections 1.4 et 1.5, nous abordons les ressources de l'arabe dialectal ainsi que les outils développés pour son traitement.

1.2 Aperçu sur les dialectes arabes

La langue arabe (العربية, Alɕrbyħ[1]) est originaire de la péninsule Arabique. Elle fait partie de la famille des langues sémitiques, telles que l'hébreu, l'amharique, le phénicien, le syriaque, etc. L'expansion territoriale au Moyen Âge et les conquêtes islamiques ont rendu l'arabe déployé géographiquement sur plusieurs continents, en Asie (le Moyen-Orient et le Golfe) et en Afrique (les pays du Maghreb et quelques pays d'Afrique comme la Somalie, Djibouti, etc.). L'arabe est aujourd'hui la langue maternelle de plus de 400 millions de personnes ainsi que la langue liturgique pour plus d'un milliard de musulmans dans le monde entier [Al-Kabi *et al.* 2016].

Le vaste patrimoine littéraire de l'arabe datant de l'ère préislamique (V^{e} et VI^{e} siècles) et la vaste zone géographique (du Moyen-Orient à l'Afrique du Nord) que couvre la langue arabe, ont engendré un caractère diglossique pour cette langue. Elle se distingue par un certain nombre de variétés à savoir ; l'arabe vernaculaire et l'arabe littéraire. Ce dernier comprend l'Arabe Classique (AC) et l'Arabe Standard Moderne (ASM). L'arabe dialectal (AD) ou vernaculaire comprend toutes les variétés des dialectes arabes régionaux avec des caractéristiques singulières ([Hamdi 2007] ; [Baccouche 2009] ; [Jalloh 2006] ; [Al-Saidat & Al-Momani 2010])

1. La translitération arabe utilisée dans ce rapport est la translittération définie par Habash-Soudi-Buckwalter schème [Habash *et al.* 2007].

sur lesquelles nous allons revenir.

1.2.1 L'arabe classique

L'Arabe Classique (AC) est le type d'arabe le plus ancien. Il s'agit d'une forme linguistique dont la grammaire a été fixée entre le VIII^{e} et le X^{e} siècle. Il regroupe l'arabe ancien de la phase pré-coranique (de l'antiquité jusqu'au début du Moyen Âge), l'arabe du Coran et l'arabe de la phase post-coranique (depuis l'avènement de l'Islam jusqu'au XVIII^{e} siècle). L'AC n'est plus que la langue du patrimoine culturel passé avec ses oeuvres classiques et son livre sacré : le Coran, les textes classiques de l'Empire islamique (le Hadith [2], la poésie ancienne, l'histoire, la grammaire, la médecine, la littérature : chroniques, proverbes, etc.). Cette variété de l'arabe est considérablement enrichie grâce à la traduction des langues anciennes (le grec, le persan, l'araméen, etc.).

L'AC n'est pas une langue de conversation courante, mais, il est appris dans les établissements d'enseignement à travers la littérature arabe classique et les cours de théologie [Hamdi 2007]. L'AC est encore utilisé aujourd'hui, mais il est limité à des contextes religieux et très formels [Al-Saidat & Al-Momani 2010]. Il est différent de l'arabe standard principalement dans le style et le vocabulaire, dont une partie est archaïque [Jalloh 2006].

1.2.2 L'arabe standard moderne

L'Arabe Standard Moderne (ASM) est la forme moderne de l'arabe depuis la renaissance arabe au XIX^{e} siècle jusqu'à aujourd'hui. C'est la langue officielle du monde arabe. Elle est comprise par la majorité des arabophones. L'ASM n'acquiert pas comme une langue maternelle, mais elle est apprise comme une langue seconde. La littérature identifie également la prédominance de cette forme standard dans les médias, particulièrement, les médias écrits. L'ASM est la langue de toutes les publications écrites : les livres, les journaux, les rapports de journal, etc. La plupart des documents sont imprimés en ASM. De même, c'est la langue utilisée dans les médias de diffusion y compris la radio et la télévision.

La plupart des locuteurs natifs de la langue arabe sont incapables de produire l'ASM de façon spontanée. Mais, ils peuvent le comprendre à cause de l'imbrication entre ses différentes variétés.L'ASM n'a pas de pratique spontanée et n'est utilisé que pour des actes formels particuliers. Cependant, étant commune à tous les pays arabes, l'ASM est la seule langue de communication inter-arabe. Il reste largement uniforme dans le monde arabe [Al-Saidat & Al-Momani 2010].

L'ASM est dérivé directement de l'AC. Il est essentiellement basé sur la syntaxe, la morphologie et la phonologie de l'AC. Cependant, au niveau lexical, l'ASM est plus moderne. Il est le résultat d'une évolution de l'AC avec une interaction entre ce dernier et les autres langues européennes grâce à une ouverture sur la culture européenne par la traduction, l'emprunt et les calques. [Jalloh 2006].

2. Les paroles du prophète Mohamed (صلى الله عليه وسلم).

1.2.3 L'arabe dialectal

L'Arabe Dialectal (AD) réfère à la langue arabe régionale utilisée généralement au niveau de la communication informelle quotidienne. Il recouvre les dialectes arabes, résultant d'une interférence linguistique entre la langue arabe et les langues locales ou voisines, à l'issue d'un processus d'arabisation ou d'une influence culturelle quelconque due principalement à la colonisation, aux mouvements migratoires, au commerce, et plus récemment aux médias [Bahloul 2009]. Par exemple, le dialecte algérien a de nombreuses influences issues de la langue berbère (la langue maternelle des pays du grand Maghreb) ainsi que de la langue française.

L'AD n'est pas enseigné dans les écoles ou encore normalisé bien qu'il y ait une culture dialectale populaire riche de contes, d'actes théâtraux, de poèmes, de chansons, de films et d'émissions télévisées. Les dialectes sont souvent utilisés dans les médias parlés informellement, tels que les feuilletons et les débats. Ils sont essentiellement parlés, donc, souvent, on ne les trouve pas écrits [Al-Saidat & Al-Momani 2010].

Les dialectes arabes sont les langues maternelles des Arabes de différents pays. Ses formes linguistiques sont généralement très différentes d'un pays à un autre [Hamdi 2007]. Il y a aussi des différences entre les dialectes des régions d'un même pays.

1.2.3.1 Classification des dialectes arabes

Principalement, les dialectes arabes varient selon deux dimensions : géographique et sociologique. Selon une classification géographique, les dialectes arabes peuvent être à leur tour subdivisés en deux groupes : les dialectes de l'Est (les dialectes de Moyen-Orient : le dialecte levantin, le dialecte du Golfe, le dialecte égyptien, etc.) et les dialectes de l'Ouest du monde arabe (l'arabe du Maghreb ou les dialectes d'Afrique du Nord) [Alorifi 2008]. La frontière naturelle entre ces deux zones est le Nil. La propagation de l'arabe avec l'Islam au nord, à l'Est et à l'Ouest a créé des dialectes orientaux et occidentaux qui sont semblables dans des niveaux, mais ils restent uniques dans leurs propres moyens. Les dialectes orientaux et occidentaux diffèrent morphologiquement, syntaxiquement et lexicalement. Les locuteurs de certains de ces dialectes sont incapables de dialoguer avec ceux d'un autre dialecte arabe. Les locuteurs du Moyen-Orient peuvent généralement se comprendre les uns et les autres, mais, ils ont souvent du mal à comprendre les locuteurs Nord-Africains (même si l'inverse n'est pas vrai, en raison de la popularité des films en dialecte égyptien et d'autres médias du Moyen-Orient) [Jalloh 2006].

Ces deux zones géographiques principales se décomposent en cinq groupes arabophones [Hamdi 2007].

— Les dialectes de la péninsule arabique : cette zone couvre les pays du Golfe (le Koweït, l'Arabie saoudite, le Bahreïn, le Qatar, les Émirats Arabes Unis), Oman et le Yémen. Les dialectes de la péninsule arabique se composent de quatre sous-groupes : Yéménite, Omani, Saoudite et le Golfe [Hamdi 2007].

— Les dialectes du nord : ils regroupent les dialectes levantins ou syro-libanais et le dialecte irakien. Le dialecte du Levant est parlé par les Arabes près de la côte de la Méditerranée, y compris les pays comme le Liban, la Syrie, la Palestine et la Jordanie [Hamdi 2007].
— Les dialectes égyptiens : ils regroupent les dialectes parlés autour de la vallée du Nil (Égypte et Soudan).
— Les dialectes maghrébins : ils sont parlés dans la région géographique arabe qui couvre le Maroc, la Tunisie, l'Algérie et l'Ouest de la Libye [Hamdi 2007].
— Il existe aussi deux autres groupes de dialectes arabes qui ne sont pas bien reconnus : le dialecte Saharan qui regroupe les dialectes Chadian et Hassaniya (parlé en Mauritanie et le Sahara du West) et le dialecte Tajiki, Uzbeki et Khorasan [Diab & Habash 2007].

Plusieurs linguistes ont proposé une classification sociolinguistique pour les dialectes arabes. Les dialectes de chaque pays arabe peuvent être classés selon des différences sociologiques et régionales en deux sous dialectes : les dialectes des citadins et les dialectes des paysans/agriculteurs ou des Bédouins [Diab & Habash 2007]. Ces dialectes présentent des différences phonologiques et lexicales. Des lexèmes identiques peuvent avoir des significations entièrement différentes entre les dialectes bédouins et les dialectes urbains.

1.2.3.2 Ressemblances et différences entre l'arabe standard et les dialectes arabes

Les dialectes arabes sont plus ou moins liés à l'ASM. Ils sont en évolution constante. Ils intègrent constamment (comme toutes les langues vivantes) de nouveaux mots et phrases, tirés la plupart du temps de langues occidentales comme le français, le turc, l'espagnol ou l'anglais. Les dialectes arabes sont le résultat de l'interaction entre l'ASM et les différentes langues originaires qui existent dans le monde arabe avec l'influence de la colonisation et l'interaction avec les langues des pays voisins. Les dialectes arabes ont de nombreux points de similitude, mais également de divergence. Ils partagent, parfois, la même grammaire, la même phonologie et le même lexique, mais, ils restent des dialectes très distincts. Cependant, les Arabes ne les considèrent pas comme deux langues distinctes.

À première vue, les différences qui sont les plus apparentes sont [Belguith *et al.* 2014] :

— Lexicales : plusieurs mots de l'AD n'appartiennent pas au lexique de l'ASM et vice-versa ;
— Phonologiques : il existe des mots en commun pour les deux variétés de l'arabe (dialectal et standard), mais, la prononciation de ces mots est généralement différente.
— Morphologiques : la conjugaison des verbes dans l'AD est relativement différente de celle de l'ASM.
— Syntaxiques : les structures des phrases présentent plusieurs différences et les règles syntaxiques ne sont pas toujours similaires.

Dans cette section, nous présentons brièvement les principales différences et similarités entre l'ASM et l'AD aux niveaux de la phonologie, la morphologie, le lexique et la syntaxe.

Phonologie. Il existe de nombreuses différences entre les dialectes du Maghreb et les dialectes orientaux. La première différence concerne le *système vocalique* des dialectes : les voyelles courtes et longues. En effet, les locuteurs du Maghreb ont tendance à abandonner les voyelles courtes et réduire la longueur des voyelles longues [Alorifi 2008]. Les voyelles /a :/, /i :/ et /u :/ se transforment respectivement en /a/, /i/ et /u/. En revanche, les locuteurs de dialectes de l'Orient ont tendance à maintenir les voyelles classiques de l'ASM. Il existe aussi des dialectes qui utilisent d'autres voyelles non définies pour l'ASM. Par exemple, l'irakien utilise la voyelle courte /o/ et le DT utilise la voyelle longue /e :/ ([Alorifi 2008] ; [Zribi *et al.* 2014]). L'Égyptien utilise la voyelle /a/ pour désigner la voyelle /i/. En outre, les diphtongues de l'ASM /aw/ et /ay/ se transforment généralement en /o :/ et /e :/, respectivement. Les changements vocaliques affectent la structure syllabique de l'ASM [Biadsy *et al.* 2009]. De plus, les dialectes arabes utilisent régulièrement des phonèmes non arabes issus d'autres langues latines.

La deuxième différence concerne le système consonantique. Le tableau 1.1 résume les principales différences par rapport à l'ASM. Notons que les changements de prononciation n'affectent pas les textes religieux.

Phonèmes de l'ASM		**Phonèmes correspondants en AD**
ق	/q/	La consonne (ق, q) se prononce /g/ en dialecte irakien et en dialecte du Golfe. Les dialectes du Maghreb conservent la prononciation de l'ASM. En égyptien et en levantin, /q/ se prononce /a/.
ك	/k/	Les dialectes du Maghreb, levantins, et égyptiens conservent la prononciation de l'ASM /k/, par contre, les dialectes de Golfe et irakien changent la phonétique de (ك, k) en /tš/.
ج	/j/	Les dialectes irakiens, levantins, et maghrébins conservent la prononciation de l'ASM, alors que les dialectes du Golfe et le dialecte égyptien changent la phonétique de (ج, j) respectivement en /y/ et /g/.
ذ	/ð/	En dialecte levantin et dialecte égyptien, la consonne (ذ, ð) se prononce /d/ dans des cas et dans d'autres /z/. Cependant, les dialectes d'Irak et du Golfe conservent la prononciation de l'ASM.
ث	θ	La consonne (ث, θ) est accordée aux deux phonèmes (/t/ ou /s/) selon la position de la lettre dans le mot contrairement aux dialectes d'Irak et du Golfe qui conservent la prononciation de l'ASM.
ض et ظ	/D/ et /Ď/	Les consonnes (ض, D) et (ظ, Ď) sont prononcées comme /D/ dans les dialectes égyptiens et levantins. Parfois, (ض, D) et (ظ, Ď) changent de phonétique. Elles sont prononcées /z/. En revanche, elles se prononcent /Ď/ en dialectes du Golfe et d'Irak.
ص	/S/	La consonne (ص, S) est prononcée dans quelques dialectes comme /s/.

TABLE 1.1 – Les différences phonologiques entre l'ASM et les dialectes arabes

Morphologie. Les différences morphologiques entre l'AD et l'ASM sont le résultat d'une simplification des paradigmes de l'ASM, dans certains cas, et dans d'autres cas, le résultat de la définition de nouvelles structures qui rendent la morphologie de l'AD plus complexe en la comparant à celle de l'ASM [Habash *et al.* 2012a]. Parmi ces différences, nous citons la disparition des marques casuelles nominales, la disparition de la voix et le mode des verbes dans

certains dialectes et la conservation de ces marques dans d'autres. D'autres phénomènes de simplification sont remarqués au niveau l'AD tels que la disparition du duel et la consolidation du féminin et du masculin au pluriel au niveau de la conjugaison des verbes dialectaux [Habash *et al.* 2012a].

Parmi les différences morphologiques, nous citons la définition de la particule de progression verbale utilisée dans plusieurs dialectes sous forme de clitiques ((+ب, bi+) pour l'égyptien (EGY) et levantin (LEV), (+د, da+) pour l'irakien (IRQ) et (+ك, ka+) pour le marocain (MOR)) et de noms comme (قاعد, qAʕd) pour le DT où cette forme verbale n'a pas de correspondant pour l'ASM. La particule de future (+س, sa+) de l'ASM est remplacée par les clitiques (+ح, Ha+) (ou (+ه, ha+)) et (+غ, γa) respectivement pour le LEV, l'EGY et le MOR et par une particule non attachée (باش, bAš) pour le DT. De la même façon, le proclitique démonstratif (+ه, ha+), qui précède l'article défini (+ال, Al+) dans plusieurs dialectes arabes, n'a pas de correspondant en ASM [Diab & Habash 2008]. Plusieurs dialectes comprennent le proclitique (+ع, ʕa+) : une forme réduite de la préposition (على, ʕlỳ). En outre, ils comprennent la circumclitique de négation non utilisée en ASM (ما, mA + + ش, š). La forme de certains clitiques pronominaux et d'affixes est également changée. Par exemple, les clitiques (+تم, +tum) / (+كم, +kum) de l'ASM deviennent en EGY (+تو, +tuwa) / (+كو, +kuw). De même, nous remarquons qu'il existe des affixes des langues étrangères qui sont utilisés lors de la conjugaison des verbes ou même dans le processus de formation des mots en arabe dialectal [Diab & Habash 2008]. Les parlers arabes utilisent ces affixes avec des mots arabes. Citons par exemple, le suffixe turc (+جي, jiy) [Baccouche 2009].

Les locuteurs des dialectes occidentaux ont tendance à utiliser l'affixe (+ن, n) dans le cas du pluriel et du singulier, par contre, les dialectes de l'Est utilisent l'affixe (+ن, n) uniquement avec le cas du pluriel. Par exemple, les deux verbes « *J'écris / nous écrivons* » sont traduits en dialectes occidentaux par (نكتب, niktib / نكتبوا, niktbwAu). Tandis que dans les dialectes de l'orient, ils sont traduits par (اكتب, Aaktib) / نكتب, niktib) [Alorifi 2008].

Orthographe. Contrairement à l'ASM et l'AC, les dialectes arabes n'ont pas une orthographe standard. Les quelques transcriptions de l'AD utilisent différentes formes orthographiques pour un seul mot. Par exemple, le mot « *beaucoup* » en DT peut avoir au moins trois formes orthographiques possibles : (برشة, brš*ħ*), (برشه, brš*h*) et (برشا, brš*A*). En plus, on trouve, parfois, plusieurs formes orthographiques pour le même mot en passant d'un dialecte arabe à un autre. Les dialectes arabes peuvent être transcrits en utilisant des scriptes romaines, des symboles (langage SMS), des lettres arabes et même des mélanges entre ces symboles.

Lexique. Le lexique de l'ASM est très différent de celui des dialectes arabes malgré l'existence de plusieurs intersections entre les deux variétés de la langue arabe. Les différences sont dues à l'impact des langues étrangères sur les dialectes suite à des faits de la colonisation et du commerce.

Les locuteurs d'un dialecte utilisent, généralement, le même lexique avec certaines spécificités soit au niveau de la phonologie soit au niveau du sens des mots. Par exemple, le mot (قدام, qudaAm) dans le DT désigne « *devant* » alors qu'en dialecte algérien désigne « *à côté* ». Le tableau 1.2 montre quelques exemples de mots de dialectes arabes.

ASM	AD					Traduction
	DT	ALG	EGY	LEV	IRQ	
قارورة qArwrħ	دبوسة dbwsħ	قرعة qrς	قنينة qnynħ	قنينة qnynħ	بطل buTl	*Une bouteille*
الجزر Aljzr	سفنارية sfnAryħ	زرودية zrwdyħ	جزر jzr	جزر jzr	جزر jzr	*Les carottes*

TABLE 1.2 – Exemples de mots dans les dialectes de la Tunisie, l'Algérie, l'Égypte, le Liban et l'Irak

Syntaxe. Les différences syntaxiques sont très limitées entre l'AD et l'ASM. La syntaxe des dialectes est affectée par l'influence des langues étrangères et par l'alternance codique entre l'AD et l'ASM et même avec des langues étrangères. Parmi les différences syntaxiques, nous citons la forme négative des verbes qui est différente de celle de l'ASM [Bouamor *et al.* 2014].

1.3 Les défis du traitement automatique des dialectes arabes

Avec la présence croissante de l'AD dans le Web et le développement continu des technologies linguistiques pour de nombreuses langues, le traitement automatique des dialectes arabes est devenu une nécessité afin de développer les ressources nécessaires pour des technologies telles que la reconnaissance vocale, la synthèse vocale, les applications de téléphonie mobile et la traduction automatique [Zribi *et al.* 2014]. Cependant, le traitement de l'AD est confronté à plusieurs défis.

Complexité du traitement de la langue arabe. La langue arabe est une langue morphologiquement riche. Elle combine une morphologie flexionnelle riche avec une orthographe très ambiguë. Ces caractéristiques posent de nombreux problèmes pour son traitement automatique [Habash 2010]. La complexité de la tâche se développe lorsqu'on parle de l'AD qui présente un ensemble de langues à tradition orale. En outre, l'AD est la langue du quotidien ; son lexique évolue dans le temps, donnant lieu à l'apparition de nouveaux mots et entrainant des structures morphologiques et syntaxiques complexes. Ces nouveaux mots peuvent être le résultat d'un mariage entre des langues étrangères avec la langue arabe et ses variétés.

L'absence de standards orthographiques. L'AD est la langue maternelle des Arabes. Il est souvent parlé au quotidien de façon spontanée, sans tradition écrites. L'absence de standards orthographiques est la principale cause de rareté (voire d'absence) de ressources écrites pour l'AD. L'orthographe est une spécification de la façon dont les mots d'une langue sont mappés

vers et à partir d'un script particulier (l'écriture arabe) [Habash 2010]. L'absence d'une orthographe pour les dialectes arabes a donné naissance à quelques textes en AD (*e.g.*, les blogs, les commentaires des réseaux sociaux, etc.) avec une mauvaise qualité où chaque mot possède plusieurs formes orthographiques avec différents symboles (des lettres arabes, des lettres latines, combinaison entre des lettres latines et des chiffres comme dans le langage SMS, etc.). Ceci rend le traitement automatique des dialectes arabes ne disposant que rarement de ressources écrites, difficile.

L'absence d'outils de transcription de l'oral pour les dialectes arabes. L'AD est une langue parlée qui apparaît rarement sous forme du support écrit. Les textes de l'AD sont obtenus par une transcription des enregistrements audio. L'absence d'outils de transcription automatique de l'oral pour les dialectes engendre la rareté et l'absence de ressources textuelles. La transcription manuelle aussi est très coûteuse en matière de temps et d'argent. De même, la transcription manuelle peut ne pas être fidèle à ce qui est dit au niveau de l'enregistrement audio. Ceci diminue donc la qualité des corpus et par la suite dégrade le résultat des outils d'analyse de la langue dialectale.

L'absence des ressources pour l'AD. En raison de ses caractéristiques sociopolitiques, sociolinguistes et géographiques, la langue arabe se caractérise par un nombre important de dialectes avec des différences phonologiques, morphologiques, lexicologiques et syntaxiques par rapport à d'autres dialectes et à l'ASM. Cette multitude a des conséquences négatives sur le traitement automatique de l'AD. En effet, les dialectes arabes sont considérés comme des langues peu dotées [Belguith *et al.* 2014]. Il est difficile et très coûteux de développer des ressources adéquates pour chaque dialecte arabe. De plus, les rares ressources développées pour un dialecte ne sont pas exploitables pour traiter un autre sans recourir à un ensemble de traitements qui augmentent la complexité de la tâche. Il est donc très onéreux d'obtenir un corpus adéquat utilisable pour la création et l'apprentissage des outils de TAL.

Contrairement à l'ASM, les dialectes arabes sont généralement des langues parlées qui apparaissent rarement sous une forme écrite. Les données textuelles pour l'AD peuvent être obtenues à partir d'une transcription manuelle orthographique. Le difficulté de cette tâche prévient la collecte de corpus de taille importante. En outre, l'absence des standards orthographique engendre des incohérences dans l'orthographe qui réduisent la valeur de ces corpus.

Problèmes dus au caractère spontané des dialectes arabes. L'utilisation de la transcription des enregistrements audio pour la construction des corpus pour l'AD engendre la présence de nouvelles difficultés pour le traitement de l'AD. Parmi ces difficultés, on peut citer la présence des disfluences (les erreurs et les autocorrections, etc.) au niveau des corpus qui exigent des traitements particuliers pour qu'on puisse utiliser ces ressources pour le développement des outils de traitement de l'AD.

1.4 Construction de corpus pour l'arabe dialectal

1.4.1 Conventions de transcription et d'annotation pour les dialectes arabes

Dans la littérature, peu de travaux ont défini des conventions de transcription pour l'AD. Ces dernières ont été proposées pour un nombre très réduit de dialectes, tels que le dialecte levantin (LEV) et égyptien (EGY).

Les conventions de transcription des dialectes arabes suivent généralement deux approches. La première est orthographique basée sur l'ASM : l'orthographe choisie est très proche de celle de l'ASM en suivant typiquement les mêmes règles de transcription à l'exception de quelques-unes. L'objectif de cette approche est de réduire les différences entre l'ASM et l'AD afin d'exploiter les ressources développées pour l'ASM [Maamouri *et al.* 2004a]. En effet, le coût de création de ressources (la rapidité et la facilité de création) pour une langue donnée présente un défi pour les chercheurs en TAL. Étant donné l'absence des conventions d'écriture pour les dialectes arabes et dans la mesure où les annotateurs peuvent utiliser leurs connaissances en ASM, il sera facile de construire un corpus pour l'AD en le transcrivant suivant les règles de l'ASM au lieu d'apprendre et d'utiliser les symboles phonétiques.

La deuxième approche de transcription est phonétique qui se base sur une transcription reflétant la prononciation de l'AD puisque la phonologie présente la principale différence entre la forme dialectale et standard de la langue arabe. Généralement, on utilise les scripts de l'ASM et/ou les symboles phonétiques.

1.4.1.1 Approche orthographique à base de l'ASM

En suivant une approche orthographique à base de l'ASM, [Zawaydeh *et al.* 2003] et [Maamouri *et al.* 2004a] ont développé un ensemble de conventions permettant la transcription des dialectes levantins. Le principe de ces conventions exige la transcription de LEV en utilisant les caractères arabes sans voyelles courtes (sauf la nunation (تنوين)) en respectant les règles orthographiques ainsi que les règles de segmentation des mots de l'ASM.

Prenons l'exemple du mot levantin (أَلتلك, Âltlk, « *je t'ai dit* »). Ce mot sera transcrit comme deux mots séparés. Cette segmentation est le résultat de l'application de la règle de l'ASM qui exige la séparation entre le verbe et le groupe prépositionnel d'objet. Étant donné que la lettre vélaire (ق, q) en dialecte LEV se prononce /a/, on transforme la lettre (أ, Â) en (ق, q). Ainsi, le mot levantin (أَلتلك, Âltlk, « *je t'ai dit* ») sera transcrit en (قلت لك, qlt lk).

Dans le cadre de la même approche, [Habash *et al.* 2012a] ont proposé la convention CODA (*Conventional Orthography of Dialectal Arabic*) pour la transcription des dialectes arabes. Cette convention a pour objectif de faciliter le développement des outils du traitement automatique des dialectes arabes. [Habash *et al.* 2012a] se sont basés sur les différences et les similitudes entre les dialectes arabes et l'ASM pour définir les règles de la convention CODA. Cette dernière a été développée, en premier lieu, pour la transcription du dialecte EGY, le dialecte palestinien (PAL) [Jarrar *et al.* 2014] et le dialecte algérien (ALG)

[Saadane & Habash 2015] plus tard. Plus de détails sur la convention CODA seront présentés dans le chapitre 3.

1.4.1.2 Approche phonétique

Une convention fondée sur les principes de l'approche phonétique a été présentée par [Diab *et al.* 2010] pour la transcription orthographique du dialecte LEV. Elle a été proposée dans le cadre du projet COLABA [Diab *et al.* 2010] qui vise à collecter et annoter les données de médias sociaux arabes comme les blogs, les forums de discussions, les chats, etc. [Diab *et al.* 2010] ont décrit une méthode de transcription phonologique nommée « *COLABA Conventional Orthography* » (CCO). Cette méthode de transcription est fidèle à la prononciation et indépendante de la façon dont un mot est typiquement écrit [Diab *et al.* 2010]. Elle préserve et représente explicitement tous les phonèmes du mot, y compris les voyelles.

La CCO n'est pas définie comme un standard orthographique pour l'écriture des dialectes arabes. Elle n'utilise pas les scripts arabes, mais elle propose d'utiliser un ensemble de symboles qui regroupent des lettres latines et d'autres symboles pour représenter les mots en AD. Il s'agit d'une représentation qui se situe entre les représentations morpho-phonémiques et phonétiques.

Parmi les règles de transcriptions : la COO exige une représentation explicite de toutes les voyelles courtes prononcées dans les mots et même les consonnes doublées. Elle utilise le symbole « ˆ » pour indiquer la présence du « Ta Marbuta » (la marque du féminin) ou le Tanween (nunation) (la marque de l'indéfini). Le signe « + » est utilisé, aussi, pour définir les frontières des clitiques dans le but de combler l'écart entre la phonologie et la morphologie. Par exemple, les mots المَكتَبَة Almaktabaħ « la bibliothèque » et عَمَلِيًا ʕmaliyaAã « pratiquement » sont rendus selon la COO respectivement en Al+maktabaˆ et 3amaliyyanˆ.

Le tableau 1.3 résume les conventions de transcription pour l'AD.

Auteur	Dialecte	Nature du dialecte	Approche
[Diab *et al.* 2010]	LEV	Écrit	Approche phonétique
[Zawaydeh *et al.* 2003]	LEV	Oral	Approche orthographique à base d'ASM
[Maamouri *et al.* 2004a]	LEV	Oral	
[Habash *et al.* 2012a]	EGY	Écrit	
[Jarrar *et al.* 2014]	PAL	Écrit	
[Saadane & Habash 2015]	ALG	Écrit	

TABLE 1.3 – Les conventions de transcription des dialectes arabes.

1.4.2 Collection des données

De nos jours, le traitement de la langue arabe et ses dialectes est devenu un nouvel axe de recherche. Plusieurs travaux de recherche ([Jarrar *et al.* 2014] ; [Al-Sabbagh & Girju 2012] ; [Salama *et al.* 2014] ; [Younes & Souissi 2014] ; [Mubarak & Darwish 2014] ; etc.) se sont intéressés au développement des corpus pour les dialectes arabes qui sont encore à un stade préliminaire. Les corpus créés en faveur de l'AD sont classés principalement en deux types :

les corpus oraux et les corpus écrits. Le premier type regroupe les corpus qui sont issus d'une transcription (manuelle ou automatique) de l'AD parlé. Le deuxième type regroupe les textes en AD qui se trouvent souvent sous forme des commentaires dans les blogs de l'Internet, les réseaux sociaux, etc. et d'autres formes écrites de l'AD.

Dans cette section, nous présentons les principales méthodes proposées pour la construction des corpus oraux et écrits pour l'AD.

1.4.2.1 Corpus oraux

Enregistrement de parole. Les dialectes égyptiens (EGY) et levantin (LEV) sont les premiers dialectes arabes qu'ont suscité l'intérêt des chercheurs en TAL. Le premier corpus est le CALLHOME développé par [Canavan *et al.* 1997] pour le dialecte EGY parlé. Ce corpus est composé de 120 appels téléphoniques entre des locuteurs natifs de l'Égypte dont chaque conversation dure environ 30 minutes. Il est composé des mots de l'ASM et du dialecte EGY. Ensuite, dans le cadre du projet de LDC « *Fisher Levantine Arabic Project* », [Maamouri *et al.* 2004a] ont collecté des enregistrements de plus de 9 400 locuteurs parlant différents dialectes levantins (le dialecte du nord, du sud et les dialectes bédouins). Ces locuteurs ont participé pour collecter 2 000 appels téléphoniques [Maamouri *et al.* 2004a] ayant une durée totale égale à 250 heures traitant des thématiques différentes. Notons que ces corpus présentés ci-dessus sont distribués sous une licence payante.

De même, le dialecte de l'Arabie Saoudite a été représenté par le corpus SAAVB « the *Saudi Accented Arabic Voice Bank* » qui est très riche en matière de contenu sonore de parole et de la diversité de locuteurs de l'Arabie Saoudite [Alghamdi *et al.* 2008]. La durée de ce corpus est environ 96 heures et 270 828 mots distribués sur 60 947 fichiers audio. Le corpus a été validé par la branche d'IBM en Égypte afin d'apprendre leur système de reconnaissance de parole. Il est distribué sous licence d'IBM.

Un corpus parallèle ASM/ALG a été construit par [Meftouh *et al.* 2012] afin de réaliser un système de traduction pour l'ASM et l'ALG. [Meftouh *et al.* 2012] ont enregistré des discussions dans différents environnements pour garantir une grande couverture des mots de lexique algérien, plus précisément le dialecte d'Annaba. Ils ont pu enregistrer 10 heures de dialogues filtrés des bruits. Une étape de transcription manuelle est suivie d'une étape d'extraction des mots à analyser. Notons que uniquement 30 % de leur corpus a été transcrit [Meftouh *et al.* 2012]. La construction de ce corpus parallèle est faite de façon manuelle, en traduisant les mots du dialecte ALG vers l'ASM. Pour améliorer la taille de ce corpus, [Meftouh *et al.* 2012] ont ajouté des phrases qui sont issues d'un remplacement de chaque mot d'une phrase par ses synonymes. Notons qu'à notre connaissance, il n'existe pas un standard orthographique utilisé pour transcrire ces dialectes.

Téléchargement à partir du Web. Dans le cadre du projet Oréodule, [Belgacem 2009] a adopté l'approche « télécharger et enregistrer » qui propose d'enregistrer un grand nombre

d'émissions de journaux radiophoniques ou télévisés diffusées à partir du Web pour créer un corpus oral multi-dialectes. Ces émissions présentent un contenu varié avec une grande diversité de locuteurs de différentes nationalités traitant différents thèmes. En se basant sur cette approche, dans une première étape, [Belgacem 2009] a enregistré 10 heures de parole regroupant les dialectes tunisiens, algériens, marocains, égyptiens, palestiniens, libanais (LIB), syriens (SYR), du Golfe (GLF), somaliens (SOM) et soudaniens (SOD). Les enregistrements collectés sont segmentés suivant les locuteurs en attribuant à chaque locuteur les informations concernant leurs noms, leurs sexes, leur origine, leurs dialectes, etc. La deuxième étape de la création du corpus est l'identification des tours de paroles et des locuteurs, l'identification des sections thématiques, la transcription orthographique, et enfin, la vérification [Belgacem 2009]. [Belgacem 2009] n'a réussi à transcrire que 37 % de leur corpus (3 heures et demie).

Dans le même cadre, [Almeman *et al.* 2013] se sont concentrés sur trois dialectes arabes (le dialecte EGY, le dialecte du GLF et les dialectes LEV) et l'ASM afin de créer un corpus oral parallèle multi-dialectes. Le corpus créé est limité un domaine des voyages et du tourisme. Ce corpus est composé de messages écrits pour l'ASM qui sont ensuite traduits vers les trois variétés des dialectes [Almeman *et al.* 2013]. Enfin, les messages sont sauvegardés sous forme d'enregistrements audio. Ce travail a donné lieu à 32 heures de dialogues transcrits. En outre, dans le cadre du programme du DARPA Transtac, un corpus de 40 heures de dialogues en dialecte irakien et en langue anglaise est collecté et transcrit [Precoda *et al.* 2007].

Ressources diverses. [Harrat *et al.* 2014] se sont intéressés au développement de ressources pour deux dialectes algériens (le dialecte d'Alger et le dialecte d'Annaba). Le corpus développé dans le cadre de ce travail est basé sur des conversations enregistrées pour le dialecte d'Annaba et des films et émissions télévisées en dialecte algérois. La transcription de ces données est faite manuellement. Pour améliorer la taille du corpus, ils ont traduit les phrases de chaque dialecte vers l'autre. Ainsi, [Harrat *et al.* 2014] ont collecté 6 415 phrases et 9 688 mots pour le dialecte d'Annaba et 6 415 phrases et 10 790 mots pour le dialecte algérois.

1.4.2.2 Corpus écrits

Les dialectes arabes sont généralement utilisés dans des contextes informels et traditionnels comme les séries télévisées, les films et aussi les réseaux sociaux sous forme de commentaires traitant différents domaines et sujets. Cette importante source de données a été utilisée par plusieurs chercheurs en TAL pour collecter et créer des corpus en dialectes arabes.

Téléchargement à partir du Web. [Diab *et al.* 2010] ont pris l'initiative pour collecter et traiter les données des médias sociales arabes comme les blogs, les forums de discussions, les chats, etc. dans le cadre du projet COLABA « Cross Lingual Arabic Blog Alerts » [Diab *et al.* 2010]. Afin de collecter les données, [Diab *et al.* 2010] ont créé un ensemble de

requêtes sur trois domaines : les problèmes sociaux, la religion et la politique pour collecter les données concernant quatre dialectes différents : le dialecte EGY, le dialecte IRQ, le dialecte LEV et le dialecte MOR. La collecte des données est faite suivant une procédure qui demande des annotateurs d'assembler des données à partir des blogs concernant les trois domaines d'intérêt en déterminant pour chaque phrase en AD sa traduction en ASM et en anglais. Une fois les données de blog sont collectées, elles sont soumises à plusieurs processus permettant le nettoyage métalinguistique, le classement de blogs, le nettoyage de contenu des blogs, le reclassement des blogs et l'annotation des blogs pour qu'elles soient prêtes à l'emploi avec les outils de TAL. Le corpus créé dans le cadre du projet COLABA a été utilisé pour tester l'analyseur morphologique « MAGEAD » [Diab *et al.* 2010] et l'outil de recherche d'informations dans les textes arabes en AD « DIRA ».

[Zaidan & Callison-Burch 2011] ont suivi la même approche pour construire un corpus pour l'AD en utilisant les ressources sur le Web. [Zaidan & Callison-Burch 2011] ont utilisé les commentaires des lecteurs sur articles publiés pendant une période de six mois de trois journaux arabophones : « Al-Ghad » de la Jordanie, « Al-Riyadh » de l'Arabie Saudia et « Al-Youm Al-Sabe' » de l'Égypte. Les principaux dialectes utilisés dans les commentaires dans ces journaux sont le LEV, le GLF et l'EGY. [Zaidan & Callison-Burch 2011] ont collecté 180 000 phrases qui contiennent des locutions en ASM et en AD. La méthode proposée par [Zaidan & Callison-Burch 2011] a permis d'identifier 44 618 phrases en AD contenant 855 000 mots dont 11 350 phrases pour le dialecte LEV, 20 741 phrases pour le GLF et 12 527 phrases pour l'EGY. Plus tard, [Cotterell & Callison-Burch 2014] ont élargi le corpus de [Zaidan & Callison-Burch 2011] par deux autres dialectes de la langue arabe qui sont le dialecte IRQ et les dialectes du Maghreb. Pour collecter ces données, les auteurs se sont basés sur les commentaires des journaux en ligne et aussi les messages de Twitter. Les phrases du corpus sont annotées manuellement via *Amazon's Mechanical Turk* [Al-Sabbagh & Girju 2012].

[Mubarak & Darwish 2014] ont collecté un corpus multi dialectal en se basant sur l'information géographique des tweets. Ils ont recueilli 123 millions de tweets par l'émission des requêtes afin d'extraire des informations concernant le profil utilisateur et des informations concernant sa location géographique. Les tweets collectés passent par une étape de normalisation et une étape de filtrage en utilisant des bigrammes de mots. À l'issue de cette étape, [Mubarak & Darwish 2014] ont pu retenir 6,5 millions de tweets qui couvrent les dialectes suivants : 3,99 millions (61 %) pour le saoudien (SAD), 880 000 (13 %) pour l'EGY, 707 000 (11 %) pour le dialecte du Kuwait (KUW), 302 000 (5 %) le dialecte des Émirats Arabie Unis (EAU), 65 000 pour le dialecte qatarien (2 %) (QTR), et le reste pour des dialectes d'autres pays arabes comme le Maroc et le Soudan (8 %).

Enfin, « YADAC » [Al-Sabbagh & Girju 2012] est un corpus arabe dialectal multigenres compilé à partir des données du Web : les micros blogs (Twitter), les blogs/les forums et les avis des utilisateurs sur les services en ligne. Il est composé de 11 millions de mots. « YOUDACC » [Salama *et al.* 2014] est, aussi, un corpus multi dialectal annoté automatiquement et collecté à partir des commentaires des utilisateurs des vidéos YouTube. Il traite différents

groupes de dialectes : EGY, GLF, IRQ, LEV et les dialectes du Maghreb. Chaque phrase de ce corpus est annotée par son correspondant dialecte. YOUDACC est composé de 2 416 105, 2 287 892, 852 438, 553 900 et 411 203 tokens respectivement pour le dialecte GLF, dialecte EGY, dialecte IRQ, les dialectes du Magreb et le dialecte LEV.

Traduction manuelle des documents. [Bouamor *et al.* 2014] ont suivi une autre méthode pour créer un corpus arabe multi dialectal parallèle composé de 2 000 phrases en ASM, EGY, DT, jordanien (JOR), PAL, SYR et aussi en anglais (ANG). Ce corpus est construit sur la base d'une traduction manuelle de 2 000 phrases pour les autres dialectes arabes. Les phrases sont sélectionnées à partir du corpus réalisé par [Zbib *et al.* 2012] : un corpus composé de phrases EGY/ANG. La tâche de traduction est réalisée par quatre traducteurs qui sont des locuteurs natifs des dialectes cibles. Un autre groupe de traducteurs ont traduit les 2 000 phrases de l'EGY vers l'ASM. La transcription orthographique du corpus ne suit pas une convention de transcription sauf pour le DT et l'EGY.

Ressources diverses. [Jarrar *et al.* 2014] ont décrit la création d'un corpus annoté pour le PAL. Ce corpus est composé de 43 000 mots collectés manuellement à partir des commentaires des blogs, des forums, Twitter et Facebook en choisissant les discussions qui contiennent plus de contenu en dialecte PAL afin de garantir la qualité de ces données. De même, [Jarrar *et al.* 2014] ont recueilli manuellement certaines histoires en PAL reflétant une diversité de sujets, de contextes et des classes sociales. Environ la moitié du corpus provient de 41 épisodes d'une émission de la télévision palestinienne où chaque épisode traite et fournit des critiques satiriques concernant différents sujets.

Le tableau 1.4 présente les principaux corpus réalisés pour les dialectes arabes. Nous signalons que la majorité des travaux cités dans cette section (à l'exception des corpus de [Canavan *et al.* 1997], [Maamouri *et al.* 2004b], [Alghamdi *et al.* 2008], [Diab *et al.* 2010] et [Jarrar *et al.* 2014]) ne donnent aucune information sur la disponibilité et la licence des corpus construits. Nous remarquons que le problème de transcription de l'AD n'a pas été évoqué et résolu par la plupart de ces travaux. En fait, il y a une absence totale des standards ou des conventions orthographiques pour la majorité des corpus créés pour la plupart des dialectes arabes (GLF, SOD, SOM, IRQ, MOR, etc.).

1.5 Création de ressources lexicales pour l'arabe dialectal

Depuis quelques années, plusieurs travaux sur le traitement des dialectes arabes se sont concentrés sur la construction des lexiques comme étant une étape préliminaire pour le développement des ressources.

Dans cette section, nous présentons les principales méthodes proposées pour la création de ces ressources lexicales.

3. Non mentionné

Auteur	Dialecte(s)	Nombre de mots/heures	Licence de distribution	Nature du corpus
[Canavan *et al.* 1997]	EGY	60 heures	Payante	Oral
[Maamouri *et al.* 2004b]	LEV	250 heures	Payante	
[Alghamdi *et al.* 2008]	SAD	96 heures/ 270 828 mots	Non libre	
[Belgacem 2009]	DT/ALG/MOR/ EGY/PAL/LIB/ GLF/SOM/SOD /SYR	10 heures dont 37 % transcrites	NM [3]	
[Meftouh *et al.* 2012]	ALG	10 heures dont 30 % transcrites	NM	
[Almeman *et al.* 2013]	IRQ	40 heures	NM	
[Harrat *et al.* 2014]	ALG	20 478	NM	
[Diab *et al.* 2010]	EGY/IRQ/LEV /MOR	NM	Libre	Écrit
[Zaidan & Callison-Burch 2011]	LEV/GLF/EGY	855 000 mots	NM	
[Al-Sabbagh & Girju 2012]	EGY/ASM	11 millions mots	NM	
[Salama *et al.* 2014]	EGY/GLF/IRQ LEV et les dialectes du Maghreb	6 521 538 mots	NM	
[Jarrar *et al.* 2014]	PAL	43 000 mots	Libre	
[Mubarak & Darwish 2014]	SAD/KUW/EGY /QAR/EAU	123 millions tweets	Libre	
[Bouamor *et al.* 2014]	EGY/DT/SYR /JOR/PAL/ANG /ASM	2 000 phrases traduites	NM	

TABLE 1.4 – Tableau récapitulatif des corpus pour l'arabe dialectal.

1.5.1 Méthodes basées sur des ressources de l'arabe dialectal et de l'arabe standard

1.5.1.1 Enrichissement par apprentissage transductif

[Duh & Kirchhoff 2006] ont présenté un lexique pour le dialecte LEV enrichi avec des étiquettes morphosyntaxiques. L'acquisition du lexique est basée sur l'extraction d'un ensemble des flexions à partir d'un corpus oral transcrit [Maamouri *et al.* 2004a] et l'attribution à ces flexions des étiquettes des catégories grammaticales en appliquant un analyseur morphologique de l'ASM. Pour étiqueter les mots non reconnus par l'analyseur, [Duh & Kirchhoff 2006] ont testé quatre méthodes d'apprentissage transductif (apprentissage transductif avec sorties structurées, regroupement transductif, SVM transductif et transducteur graphique spectral). Ces méthodes se basent sur un lexique enrichi avec les étiquettes de catégories grammaticales {Xm} pour deviner les étiquettes des mots non reconnus au niveau du lexique {Xu}. Le résultat de l'application de ces méthodes, un lexique pour le dialecte LEV ayant comme taille u+m entrées lexicales. Les expérimentations effectuées par [Duh & Kirchhoff 2006] ont montré que la méthode SVM transductif donne une meilleure exactitude (66,54 %) pour l'attribution des étiquettes aux mots. Avec cette méthode, [Duh & Kirchhoff 2006] ont pu créer un lexique d'environ 15 000 entrées.

1.5.1.2 Enrichissement par analyse morphologique

Un lexique bilingue dialecte IRQ/ANG a été développé par [Graff *et al.* 2006]. Il est extrait à partir d'un corpus oral composé d'un ensemble de conversations téléphoniques transcrites pour le dialecte IRQ. Ce lexique comprend la prononciation, la morphologie et les étiquettes de catégories grammaticales des mots (formes fléchies). La création du lexique se fait en deux étapes. La première étape consiste à segmenter les mots en morphèmes et ajouter des voyelles aux différents morphèmes pour réduire les ambiguïtés dues au nombrs important de prononciations. La deuxième étape est l'attribution des analyses morphologiques avec l'outil ABUMORPH [Graff *et al.* 2006] pour chaque morphème. [Graff *et al.* 2006] ont attribué à chaque morphème les étiquettes qui correspondent à la catégorie grammaticale et leur signification en anglais. La figure 1.1 montre un exemple d'une entrée lexicale extraite du lexique de [Graff *et al.* 2006].

bi + ha + Al + salfuwn

PREP + DEM_PRON + DET + NOUN

with + this + the + cell phone

FIGURE 1.1 – Exemple d'une entrée lexicale extraite du lexique de [Graff *et al.* 2006]

1.5.1.3 Alignement d'un corpus avec un petit dictionnaire

[Rambow *et al.* 2006] ont proposé une méthode permettant de créer un lexique bilingue ASM/LEV afin de l'utiliser au niveau l'analyse syntaxique du dialecte LEV. La méthode proposée par [Rambow *et al.* 2006] est similaire à la méthode d'alignement de corpus proposée par [Rapp 1999]. Cette méthode permet la recherche de tous les mots ayant des cooccurrences similaires à ceux existant dans un petit dictionnaire « Seed Dictionnary ». Pour appliquer la méthode de [Rapp 1999] sur la langue arabe, [Rambow *et al.* 2006] ont effectué quelques modifications. Ils ont proposé d'ajouter une étape répétitive permettant de choisir la paire des mots la plus confidente et l'ajouter au dictionnaire. [Rambow *et al.* 2006] ont utilisé deux corpus de la langue arabe : une partie du treebank de l'ATB pour l'ASM et le treebank LATB du dialecte LEV pour induire un lexique bilingue composé d'environ 100 entrées lexicales. Aucun résultat spécifique n'a été donné sur les performances de l'utilisation de l'algorithme de [Rapp 1999] sur les dialectes arabes. [Rambow *et al.* 2006] ont conclu que leur algorithme a donné des résultats nettement inférieurs à ceux atteints par [Rapp 1999].

1.5.1.4 Construction manuelle et traduction automatique

SANA est un lexique de sentiment multigenre composé de 224 564 entrées couvrant l'ASM, l'EGY, le LEV et l'ANG. Ce lexique a été développé par [Abdul-Mageed & Diab 2014].

Il est construit à partir de différentes ressources, y compris le Penn Arabic TreeBank (PATB) [Maamouri *et al.* 2004a], des conversations en dialecte EGY, des commentaires de YouTube et Twitter et le SentiWordNet version anglaise. Une partie importante des entrées SANA est listée sous forme de mots (lemmes et formes fléchies) enrichis avec les voyelles courtes et des étiquettes décrivant leurs caractéristiques morphologiques (catégorie grammaticale, le genre, le nombre, la racine, et la classe du genre). Certains composants du lexique ont été construits à la main, d'autres ont été obtenus en utilisant des méthodes de traduction automatique [Diab *et al.* 2014]. Pour l'étape manuelle, [Abdul-Mageed & Diab 2014] ont étiqueté manuellement deux listes de mots de genres différents : une liste de 3 325 adjectifs arabes extraits de PATB ([Maamouri *et al.* 2004a] ; [Abdul-Mageed & Diab 2011]) et un lexique extrait de l'ensemble des chats en dialecte EGY. La figure 1.2 montre des exemples du lexique SANA [Abdul-Mageed & Diab 2014].

TAG	ENG	MSA	EGY	LEV
Pos	intelligent	لبيب	ناصح	حربوق
	fragrance	عبير	ريحه	عطر
	contended	راض	مبسوط	مرتضي
Neg	chatterbox	ثرثار	رغاي	حكوجي
	huffy	غاضب	زعلان	معصب
	fault	وزر	غلطة	خطية

FIGURE 1.2 – Exemples du lexique SANA [Abdul-Mageed & Diab 2014]

1.5.2 Acquisition circulaire à partir du Web

[Al-Sabbagh & Girju 2010] ont proposé une approche qui utilise le même principe de l'approche associationniste utilisée par [Rapp 1999] pour induire un lexique générique pour l'ASM et le dialecte EGY. L'approche associationniste suppose que les patterns des mots co-occurrents d'une langue source sont les mêmes dans les traductions de cette langue. Pou cette raison, elle propose de chercher ces partons co-occurrents dans les traductions de la langue source. Cette approche nécessite la présence d'un grand corpus pour les deux langues pour avoir des statistiques de corrélation précises. Cependant, [Al-Sabbagh & Girju 2010] ont proposé une nouvelle technique pour l'acquisition de mots co-occurrents qui se base sur le Web en le considérant comme étant un grand corpus. Cette technique a permis d'extraire 1 000 formes fléchies du dialecte EGY avec leurs synonymes en ASM.

1.5.3 Acquisition des données à partir de ressources hétérogènes

Le lexique monolingue « ECAL » développé pour le dialecte égyptien (EGY) [Kilany *et al.* 1997] est le premier lexique de prononciation développé par le LDC. Il est composé de 51 202 entrées (formes fléchies) dont chacune est présentée dans sa forme phonologique, orthographique, son lemme et ses caractéristiques morphologiques. [Kilany *et al.* 1997]

ont exploité de diverses ressources y compris un corpus oral et deux dictionnaires manuscrits version papier pour le dialecte EGY.

« LA/MSA dictionnary » [Maamouri *et al.* 2006] est un lexique développé en faveur du dialecte LEV. Il s'agit d'un petit dictionnaire bilingue ASM/LEV qui utilise les formes fléchies des mots pour réduire au minimum les difficultés de l'analyse morphologique et de la génération nécessaire dans le processus de traduction et pour avoir un format simple de dictionnaire utilisable par toutes les approches de TAL. [Maamouri *et al.* 2006] ont testé quatre méthodes (« Pont Automatique », « Egyptian-Cognate », « The Human-Checked » et « Simple-Modification ») en parallèle, pour produire quatre sous dictionnaires. [Maamouri *et al.* 2006] ont remarqué que l'emploi de leur dictionnaire « LA/MSA dictionnary » a amélioré la précision de l'analyse syntaxique du dialecte LEV.

[Graff & Maamouri 2012] ont créé et mis-à-jour trois lexiques bilingues publiés pendant les années 1960 pour les anglophones qui veulent apprendre trois dialectes arabes : IRQ, SYR et MOR. Les objectifs de ce travail sont :

— de conserver la totalité des contenus de ces dictionnaires en corrigeant les erreurs,
— d'ajouter de nouvelles entrées pour le dictionnaire avec leurs synonymes en anglais,
— de translittérer la totalité de ces lexiques avec les alphabets arabes et aussi les alphabets phonétiques internationaux (IPA) pour la prononciation,
— de produire une structure de lexique uniforme via le biais du balisage (LMF, ISO 24 613) pour être utilisable par les linguistes et aussi dans les travaux de TAL.

[Graff & Maamouri 2012] ont enrichi le lexique IRQ avec environ 4 100 mots distincts extraits à partir d'une collection de 25 heures de paroles en IRQ transcrit [Appen 2006] et du lexique développé par [Graff *et al.* 2006]. Ils ont aussi utilisé le Web pour enrichir les dictionnaires SYR et MOR. Aucune information n'est fournie sur la taille totale de ces trois dictionnaires.

Thawra [Diab *et al.* 2014] est un lexique multilingue (EGY/ASM/ANG) basé sur le lemme enrichi avec les informations morphosyntaxiques et morphosémantiques relatives au dialecte EGY. À ce lexique s'ajoutent des informations linguistiques comme l'étiquette de catégorie grammaticale, le genre, le nombre, la rationalité, le patron morphologique et la racine morphologique. [Diab *et al.* 2014] ont utilisé plusieurs ressources hétérogènes préexistantes pour créer ce lexique. Toutes ces ressources ont passé par une étape de normalisation orthographique suivant la convention orthographique CODA [Habash *et al.* 2012a] définie pour le dialecte EGY, une étape d'ajout des traductions en ANG et en ASM pour les mots non traduits, et une étape d'enrichissement avec des informations linguistiques.

La figure 1.3 montre des exemples du lexique THARWA [Diab *et al.* 2014].

Dans le cadre du développement d'un système de traduction automatique du dialecte MOR, [Tachicart *et al.* 2014] ont développé un lexique bilingue ASM/MOR. Leur méthode de création du lexique se base sur une traduction manuelle dans les deux sens, c'est-à-dire, à partir des ressources textuelles de l'ASM (un dictionnaire électronique de l'ASM) où une traduction vers le MOR est faite. En outre, en partant des ressources du MOR (un diction-

ID	EGY		POS	Root	Pattern	MSA		ENG
25	آدي	*|diy*	dem	-	-	هذا	*h'*A*	this
3077	اتأجّل	*Aito>aj~il*	verb	*>jl*	AitoC1aC2C2iC3	تأخّر	*ta>ax~ar*	be postponed
10541	بايخ	*bAyix*	adj	*bwx*	C1AC2iC3	سخيف	*saxiyf*	silly
15539	ترباس	*tirobAs*	noun	*trbs*	C1iC2oC3AC4	مزلاج	*mizolAj*	latch
17578	جنينة	*jinaynap*	noun	*jnn*	C1iC2ayC3ap	حديقة	*Hadiyqap*	garden
19857	خلبصة	*xalobaSap*	vbn	*xlbS*	C1aC2oC3aC4ap	عربدة	*Earobadap*	raucous
20591	دكاكيني	*dakAkiyniy*	adv	*dkn*	C1aC2AC2iyC3iy	سرا	*sir~A*	secretly
21941	رقّاصة	*raq~ASap*	noun	*rqS*	C1aC2C2AC3ap	راقصة	*rAqiSap*	female dancer
23334	زرار	*zurAr*	noun	*zrr*	C1uC2AC3	زرّ	*zir~*	button
24754	سوّاق	*saw~Aq*	noun	*swq*	C1aC2C2AC3	سائق	*sA}iq*	driver
37891	مشغولات	*ma$oguwlAt*	noun	*$gl*	maC1oC2uwC3At	تحف	*tuHaf*	artifacts

FIGURE 1.3 – Exemples du lexique THARWA [Diab *et al.* 2014]

naire du MOR version papier) une traduction est faite vers l'ASM. Pour enrichir leur lexique avec d'autres termes qui n'existent pas dans les lexiques générés, [Tachicart *et al.* 2014] ont recueilli à partir du Web, les mots en MOR les plus récents et fréquemment utilisés aux commentaires des pages Web. Un certain nombre de traitements et d'éliminations nécessaires ont été effectués pour obtenir 18 000 entrées.

Le tableau 1.5 récapitule les différentes ressources lexicales développées pour l'AD. Les méthodes de construction des lexiques de l'AD sont, généralement, basées sur les ressources de l'ASM. Cette approche a montré son efficacité pour la construction des lexiques de l'AD, mais, la disponibilité et la qualité des ressources de l'ASM mettent en question l'efficacité de cette approche.

Auteurs	Dialecte(s)/langue(s)	Nombre de flexions/lemmes
[Kilany *et al.* 1997]	EGY	51 202 flexions
[Duh & Kirchhoff 2006]	LEV	15 000 flexions
[Maamouri *et al.* 2006]	LEV/ASM	1 560 flexions
[Graff *et al.* 2006]	IRQ/ANG	13 000 flexions
[Rambow *et al.* 2006]	LEV/ASM	100 flexions
[Al-Sabbagh & Girju 2010]	EGY/ASM	1 000 flexions
[Graff & Maamouri 2012]	IRQ/ANG SYR/ANG MOR/ANG	Non mentionné
[Diab *et al.* 2014]	EGY/LEV/ASM	73 000 flexions
[Abdul-Mageed & Diab 2014]	LEV/ASM/EGY/ANG	224 564 lemmes et flexions
[Tachicart *et al.* 2014]	MOR/ASM	18 000 flexions

TABLE 1.5 – Tableau récapitulatif des travaux réalisés pour la création de lexiques en dialectes arabes.

1.6 Approches de développement d'outils de l'arabe dialectal

Depuis les années 2000, le traitement automatique des dialectes arabes a pris de l'essor avec une variété de travaux traitant plusieurs dialectes arabes notamment le dialecte EGY,

LEV, IRQ, etc. Au cours des les cinq dernières années, ces travaux se sont multipliés suite aux révolutions arabes. La plupart de ces travaux se focalisent sur le développement des systèmes de traduction de l'AD vers l'ANG et l'ASM ([Zbib *et al.* 2012] ; [Salloum & Habash 2013] ; [Sajjad *et al.* 2013] ; etc.) et la création des ressources et des outils de traitement de TAL ([Habash *et al.* 2013] ; [Rambow *et al.* 2006] ; etc.). Ces travaux se basent principalement sur deux approches : la première rassemble les travaux qui adaptent les ressources de l'ASM pour créer des outils de l'AD et la deuxième regroupe les travaux qui se basent sur des petites ressources de l'ASM et l'AD pour créer les outils de l'AD.

1.6.1 Première approche : adaptation d'outils de l'arabe standard pour traiter les dialectes

1.6.1.1 Adaptation à base d'un lexique parallèle AD/ASM

[Rambow *et al.* 2006] ont présenté une méthode d'analyse syntaxique qui n'exige ni l'existence d'un corpus annoté pour l'AD (sauf pour le développement et l'essai), ni d'un corpus parallèle ASM/LEV. En revanche, elle exige l'existence d'un lexique reliant les lexèmes de l'AD aux lexèmes de l'ASM et la connaissance des différences morphologiques et syntaxiques entre l'ASM et un dialecte. Pour réaliser cette idée, [Rambow *et al.* 2006] ont eu recours aux deux principales ressources : une partie du treebank de l'ASM (ATB) de LDC [Maamouri *et al.* 2004a] et le treebank du dialecte LEV (LATB) [Maamouri *et al.* 2006]. Trois méthodes ont été proposées pour l'analyse syntaxique de l'AD [Chiang & Rambow 2006] : la transduction des phrases, la transduction du treebank et la transduction de grammaire. L'idée fondamentale de la méthode de la transduction des phrases consiste à traduire les mots d'une phrase de l'AD en un ou plusieurs mots en ASM. L'ensemble des mots résultant de la traduction sont gardés en forme de treillis. Le meilleur chemin dans le treillis est transmis à l'analyseur de l'ASM [Bikel 2002]. Enfin, ils remplacent les noeuds terminaux dans la structure d'analyse résultante par les mots originaux en dialecte LEV. La deuxième méthode est la transduction du treebank. L'idée est de convertir le treebank de l'ASM (ATB), dans une approximation, en un treebank pour l'AD en utilisant les connaissances linguistiques des variations systématiques au niveau syntaxique, lexical et morphologique entre les deux variétés de l'arabe. Sur ce nouveau treebank, l'analyseur syntaxique de [Bikel 2002] est appris et par la suite, évalué sur le LEV. Enfin, la transduction de grammaire englobe les deux autres méthodes [Chiang & Rambow 2006]. Elle utilise le mécanisme des grammaires synchrones pour générer des paires d'arbres reliant les structures syntaxiques des phrases de l'ASM et le LEV. Ces grammaires synchrones peuvent être utilisées pour analyser les nouvelles phrases dialectales. L'évaluation de ces trois méthodes a montré que la transduction de la grammaire a donné la meilleure performance. Elle a permis de réduire le taux d'erreur de 10,4 % et 15,3 % respectivement, avec et sans l'utilisation des étiquettes de catégories grammaticales.

1.6.1.2 Adaptation par une mise-à-jour d'un lexique de l'ASM

L'analyse morphologique de l'AD a fait l'objet de plusieurs travaux de recherche ([Afify *et al.* 2006] ; [Almeman & Lee 2012] ; [Harrat *et al.* 2014] ; [Salloum & Habash 2014] ; [Boujelbane *et al.* 2014a]). Leur idée de base s'articule autour de l'ajout des termes dialectaux à un lexique de l'ASM afin de l'intégrer dans des outils proposés pour l'analyse morphologique.

Ajout des affixes dialectaux. Pour la segmentation des mots en IRQ, [Afify *et al.* 2006] ont ajouté les affixes irakiens à l'analyseur morphologique Buckwalter [Buckwalter 2004] dans l'objectif d'améliorer le résultat d'un système de reconnaissance de parole. Avec cet analyseur adapté à l'IRQ, leur système de reconnaissance de parole a eu une relative réduction de 10 % du taux d'erreur des mots.

[Almeman & Lee 2012] ont décrit l'adaptation de l'analyseur morphologique de l'ASM « Al-Khalil-ASM » [Boudlal *et al.* 2010] pour l'AD. L'adaptation se limite uniquement à l'ajout des affixes des dialectes arabes à la base lexicale de « Al-Khalil-ASM ». L'analyseur morphologique résultant n'a pu analyser que 69 % des mots de l'AD. Afin de segmenter les mots non reconnus par l'analyseur adapté, [Almeman & Lee 2012] ont proposé un processus de recherche sur le Web qui permet de déterminer pour chaque mot de l'AD une des quatre segmentations possibles qui sont : un mot complet sans suffixes et préfixes, le(s) préfixe(s) + lemme + le(s) suffixe(s), le(s) préfixe(s) + lemme et le lemme + le(s) suffixe(s) en utilisant les affixes de l'AD. À ce niveau, [Almeman & Lee 2012] ne proposent pas de chercher les traits morphologiques des mots dialectaux. Ils suggèrent uniquement de détecter leurs morphèmes. [Almeman & Lee 2012] ont exploité le Web comme un corpus pour calculer la fréquence d'apparition de chaque segmentation proposée afin de choisir celle avec la plus grande fréquence. Les auteurs ont justifié leur choix d'utiliser le Web par le fait que ce dernier est une source très riche avec le contenu en AD. De même, [Almeman & Lee 2012] se sont basés sur l'idée que les mots dialectaux utilisent souvent des lemmes similaires à ceux de l'ASM et que les principales différences sont au niveau des voyelles courtes et des affixes. La méthode proposée pour la segmentation a permis de segmenter 94 % des mots d'un corpus pour les dialectes arabes.

Pour créer « ADAM » : un analyseur morphologique pour les dialectes arabes, [Salloum & Habash 2014] ont proposé une méthode qui consiste à améliorer la base des préfixes et des suffixes de l'analyseur morphologique arabe « SAMA » [Maamouri *et al.* 2009] en ajoutant des affixes spécifiques aux dialectes EGY, LEV et IRQ. Ils ont traité uniquement ces trois dialectes car ils ont un comportement morphosyntaxique similaire (les particules de future, la négation des verbes, les pronoms d'objet indirect, etc.).

[Salloum & Habash 2014] ont créé un ensemble de 1 021 règles qui permettent l'extension de la liste de clitiques et d'affixes de l'ASM par ses correspondantes en dialectes cibles. Par exemple, les préfixes dialectaux de futur (ح, H), (رح, rH-) et (ه, h) ont un comportement morphosyntaxique similaire au préfixe de futur de l'ASM (س, s). Ainsi, [Salloum & Habash 2014] ont développé des règles permettant de créer une copie de chaque occurrence du préfixe de

l'ASM et de le remplacer avec les préfixes dialectaux. Un autre ensemble de règles ont été élaborées. Elles permettent d'ajouter des affixes qui n'ont pas de correspondance avec l'ASM, tels que le préfixe démonstratif du dialecte LEV (ﺪ, h, « *ce* »). Ce lexique a été intégré dans les analyseurs morphologiques de l'ASM SAMA et l'analyseur morphologique de l'EGY « CALIMA » [Habash *et al.* 2012b]. Les systèmes $\text{ADAM}_{\text{CALIMA}}$ et $\text{ADAM}_{\text{SAMA}}$ résultants ont été testés uniquement pour les dialectes LEV et EGY.

$\text{ADAM}_{\text{SAMA}}$ a réduit le taux des mots hors-vocabulaire de 50 % pour les mots types et 66 % pour les tokens pour le dialecte LEV. Les valeurs respectives pour le dialecte EGY sont de 29 % et 50 %. En outre, $\text{ADAM}_{\text{CALIMA}}$ améliore les performances du système CALIMA.

Enrichissement du lexique de l'ASM par des termes dialectaux. [Harrat *et al.* 2014] ont adapté le lexique de l'analyseur morphologique BAMA [Buckwalter 2004] en gardant les affixes et les lemmes qui appartiennent au dialecte ALG, en supprimant ceux qui n'appartiennent pas et en ajoutant d'autres specifiques à l'ALG. Ainsi, [Harrat *et al.* 2014] ont exploité un corpus du dialecte algérois pour enrichir le lexique de l'analyseur morphologique BAMA. De même, ils ont modifié les voyelles courtes des mots suivant la prononciation en ALG pour la partie en commun avec l'ASM. L'analyseur résultant est testé sur 1 618 mots distincts extraits à partir de 600 phrases en ALG. Cet analyseur a pu analyser correctement 43,3 % des mots qui sont en dialecte ALG et 68,98 % des mots en ASM et en ALG. [Harrat *et al.* 2014] ont constaté que l'échec de l'analyse de certains mots est dû à l'absence d'un standard orthographique pour le dialecte ALG et l'appartenance d'autres mots à des langues étrangères comme le français.

1.6.1.3 Adaptation à base d'intégration de ressources dialectales

Lexique pour l'AD. [Habash *et al.* 2012b] ont proposé la transformation du lexique « ECAL » : un lexique pour le dialecte EGY, en une forme tabulaire compatible avec la structure utilisée par l'analyseur morphologique de l'ASM « SAMA » [Maamouri *et al.* 2009] afin de créer l'analyseur morphologique du dialecte EGY « CALIMA ». Cette forme tabulaire est composée de six tableaux. Les trois premiers tableaux stockent la liste des préfixes, la liste des suffixes et les lemmes. Les trois autres tableaux enregistrent les différentes combinaisons possibles entre le préfixe(s)-lemme(s), le préfixe(s)-lemme(s)-suffixe(s) et le lemme(s)-suffixe(s). [Habash *et al.* 2012b] ont étendu ces tableaux en ajoutant des nouveaux clitiques, des nouvelles formes orthographiques et de nouvelles catégories grammaticales qui ne sont pas mentionnées au niveau d'ECAL. CALIMA a été testé sur un corpus composé de 48 millions de mots en EGY. Il a pu, dans 84.1 % des cas, donner une étiquette grammaticale correcte parmi les étiquettes données. Ainsi, dans 7.9 % des cas, il a échoué de reconnaitre les mots analysés.

Lexique et analyseur pour l'AD. L'adaptation des outils existants pour l'ASM est aussi appliquée à l'étiquetage morphosyntaxique du dialecte EGY. [Habash *et al.* 2013] ont adapté un outil de désambiguïsation morphosyntaxique de l'ASM « MADA » [Habash & Rambow 2005]. Cet outil permet l'analyse morphologique et la désambiguïsation des résultats suivant le contexte

de chaque mot dans une phrase. MADA se base sur deux ressources : un analyseur morphologique et une large collection annotée manuellement avec les caractéristiques morphologiques correspondantes. Le développement de ces deux ressources pour le dialecte EGY a été respectivement l'objet du travail de [Habash *et al.* 2012b] pour le développement de l'analyseur morphologique de l'EGY « CALIMA » et du travail de [Maamouri *et al.* 2012] pour la création d'un corpus pour le dialecte EGY annoté de façon compatible avec « CALIMA ». Ainsi, [Habash *et al.* 2013] ont suivi la même démarche utilisée par « MADA » afin de créer « MADA-ARZ » (MADA version EGY). L'adaptation consiste à remplacer l'analyseur morphologique de l'ASM « SAMA » par l'analyseur morphologique de l'EGY « CALIMA » et le lexique de l'ASM par celui de l'EGY. De même, un ensemble de modifications ont été effectuées à « MADA » pour traiter l'EGY. L'évaluation menée sur cet outil montre que « MADA-ARZ » a donné des résultats meilleurs que l'application du système « MADA » version ASM. Il a donné une valeur d'exactitude égale à 75,4 % montrant que le système a choisi toutes les valeurs correctes des traits morphologique pour le mots en question. [Habash *et al.* 2013] ont montré que « MADA-ARZ » a amélioré la qualité de la traduction automatique du dialecte EGY vers l'ANG en améliorant le score BLEU de 3.1 % et en réduisant le taux des mots non reconnus de 16% (par rapport à l'application de MADA).

1.6.2 Deuxième approche : développement de nouveaux outils

1.6.2.1 Méthodes fondées sur l'apprentissage

Apprentissage peu supervisé. [Yang *et al.* 2007] et [Riesa & Yarowsky 2006] ont présenté une méthode d'apprentissage semi-supervisé pour la segmentation des morphèmes des dialectes arabes. La méthode proposée par [Yang *et al.* 2007] est une adaptation du modèle de segmentation basée sur les règles, présentée par [Riesa *et al.* 2006]. [Yang *et al.* 2007] ont incorporé divers types de ressources y compris des règles linguistiques, un petit lexique annoté, une liste précompilée des affixes et des racines et un ensemble de mots extraits depuis des corpus non annotés dans une architecture heuristique. Pour segmenter les mots, le modèle heuristique utilise un module de segmentation basé sur les règles. À chaque appel, le modèle de segmentation à base de règles essaie de segmenter le mot en question selon ces trois patrons : préfixe(s)+racine, racine+suffixe et préfixe(s)+racine+suffixe. Le modèle heuristique vérifie l'appartenance du mot à segmenter à la liste des mots segmentés « Seed Table ». Cette liste contient tous les mots qui ont pu être segmentés. Le module de segmentation les utilise pour segmenter d'autres morphèmes. Cette liste est mise à jour à chaque itération de l'application du module heuristique. Si l'appartenance à cette liste est validée, donc on prend la segmentation présente dans cette liste. Dans le cas contraire, ce mot passe par un module de segmentation basé sur les règles en utilisant la liste des racines. Si aucune segmentation n'est trouvée, alors le mot passe par le module de segmentation basé sur les règles en utilisant cette fois-ci la liste « word table ». Cette liste contient les mots que le module heuristique n'a pas réussi à les segmenter. Ces mots sont considérés comme des racines mais ils ont un

ordre de priorité différent de celles qui existent dans la liste des racines. Cette liste est mise à jour à chaque itération de l'application du module heuristique au lieu de la liste des racines. Si une segmentation est trouvée avec cette liste, alors on étendu cette liste par cette nouvelle segmentation. Cependant, si le module à base des règles n'a pas réussi à segmenter le mot, alors ce mot reste comme entrée pour la prochaine itération de cet algorithme heuristique. [Yang *et al.* 2007] ont utilisé le lexique annoté développé par LDC pour initialiser ses ressources textuelles. Ce lexique contient les racines et les patrons de 12 934 mots du dialecte IRQ. Le module de segmentation des morphèmes du dialecte IRQ a été appliqué sur un corpus d'apprentissage parallèle ANG/IRQ d'un système de traduction automatique statistique développé au SRI. [Yang *et al.* 2007] ont prouvé que l'utilisation d'un module heuristique de segmentation pour un système de traduction aide à accroître l'exactitude de la traduction et à donner de meilleurs scores BLEU.

[Duh & Kirchhoff 2005] ont élaboré un étiqueteur morphosyntaxique pour l'AD. La méthode proposée a pour objectif de réaliser un étiqueteur de manière peu supervisée, en utilisant les ressources existantes pour l'ASM et l'AD combiné avec des techniques d'apprentissage non supervisé. [Duh & Kirchhoff 2005] ont utilisé le modèle de Markov caché (HMM) comme un modèle probabiliste pour réaliser leur étiqueteur des catégories grammaticales pour l'AD en se basant sur un apprentissage peu supervisé. Deux systèmes ont été développés pour étudier l'effet de l'apprentissage peu supervisé. Le premier système est un étiqueteur peu supervisé appris en utilisant un lexique dérivé du corpus d'apprentissage du dialecte EGY où l'étiquette correcte n'est pas connue pendant l'apprentissage. Le lexique est annoté avec un expert. Le deuxième système est appris en utilisant un corpus non annoté pour le dialecte EGY et un lexique qui est le résultat de l'application de l'analyseur morphologique de l'ASM « BAMA » [Buckwalter 2004] sur le corpus d'apprentissage du dialecte EGY et la récupération de toutes les étiquettes résultantes pour chaque mot. L'étiqueteur peu-supervisé résultat a permis d'obtenir une exactitude de 62,76 %. Pour améliorer la valeur d'exactitude, [Duh & Kirchhoff 2005] ont jugé intéressant d'ajouter au modèle d'apprentissage les caractéristiques morphologiques des mots (les affixes) et d'améliorer la qualité du lexique utilisé pour l'apprentissage peu supervisé. Pour améliorer la qualité du lexique, ils ont essayé de réduire l'ensemble d'étiquettes pour les mots non analysés appartenant au lexique. Ils ont groupé les mots analysables et les mots non analysables, et ils ont réduit l'ensemble des étiquettes possibles pour les mots non analysables en fonction de leur appartenance à ce groupe. Ainsi, avec ces améliorations, la valeur de l'exactitude a augmenté de 7,07 %. Pour remédier au problème dû au manque de données pour les dialectes arabes, [Duh & Kirchhoff 2005] ont essayé d'exploiter les points communs entre les dialectes. De ce fait, ils ont essayé d'exploiter un corpus non annoté du dialecte LEV. [Duh & Kirchhoff 2005] ont suivi deux méthodes pour le partage des données entre les dialectes. Le partage de données peut être pendant la phase d'apprentissage et au niveau de tous les composants de l'étiqueteur. L'exactitude de l'étiqueteur peu supervisé appliqué sur le corpus d'évaluation s'est amélioré en passant de 69,83 % à 70,88 %.

Apprentissage supervisé. [Riesa & Yarowsky 2006] ont proposé une méthode supervisée pour la segmentation des morphèmes des dialectes arabes. À l'aide d'un petit lexique, ils ont créé un modèle de segmentation supervisé basé sur le modèle des arbres. Pour chaque affixe dialectal, un modèle de classification est construit et appris. Chaque classificateur est entraînée avec des petites quantités de données pour effectuer la segmentation des morphèmes. Le modèle de segmentation proposé par [Riesa & Yarowsky 2006] a été testé sur deux dialectes de la langue arabe : l'IRQ et le LEV. L'application de cette approche a réduit le nombre des mots inconnus au moment de la traduction (50 %) et elle a permis d'améliorer le score BLEU pour les systèmes de traduction automatique.

1.6.2.2 Méthode à base de règles

[Habash *et al.* 2005] ont conçu « MAGEAD » un analyseur et générateur morphologique pour la famille des langues arabes (ASM et AD). « MAGEAD » est un système bidirectionnel qui fait la liaison entre un lexème et un ensemble de caractéristiques linguistiques avec une forme de surface d'un mot grâce à une séquence de transformations. Les analyses morphologiques proposées par « MAGEAD » [Habash *et al.* 2005] sont représentées en termes d'un lexème et d'un ensemble de caractéristiques. Le lexème est composé d'une racine, d'une classe de comportement morphologique (CCM) et d'un index de signification. [Habash *et al.* 2005] ont utilisé une représentation hiérarchique avec une notion d'héritage non monotone pour définir une CCM. Cette hiérarchie spécifie toutes les classes de comportement morphologique qui partagent les mêmes caractéristiques et les mêmes morphèmes. La hiérarchie de CCM est variante et indépendante de la nature de la langue arabe. Cette spécification de la hiérarchie est valable pour l'ASM et pour le dialecte LEV. La hiérarchie de CCM ne couvre que les verbes. Pour garder le critère de variance et d'indépendance de la hiérarchie de CCM, [Habash *et al.* 2005] ont proposé une représentation variante et indépendante pour les morphèmes qui représentent les hiérarchies de CCM. Ces morphèmes sont considérés comme des morphèmes abstraits (MA) qui sont classés dans l'ordre de surface des morphèmes concrets correspondants. Ces morphèmes abstraits sont ensuite transformés en morphèmes concrets qui sont concaténés dans un ordre spécifié. Pour transformer les morphèmes abstraits en morphèmes concrets, deux types de règles sont utilisés : les règles morpho-phonémiques/phonologiques qui transforment la représentation morphémique en une représentation phonologique et orthographique et les règles orthographiques qui réécrivent seulement la représentation orthographique (par exemple, des règles de l'utilisation de la gémination « šadda »).

Pour que MAGEAD accepte aussi des verbes levantins, certaines modifications ont été effectuées. [Habash *et al.* 2005] se sont concentrés sur la représentation orthographique pour simplifier la tâche. Ils ont employé l'orthographe diacritique libre développée au LDC pour le dialecte LEV [Maamouri *et al.* 2006]. [Habash *et al.* 2005] ont amélioré les CCM pour inclure deux morphèmes abstraits qui n'appartiennent pas à l'ASM. De même, une extension de la grammaire hors contexte, qui représente l'ordre des morphèmes, est effectuée pour ordonner

les deux nouveaux morphèmes abstraits. En outre, [Habash *et al.* 2005] ont effectué quatre changements au niveau du passage de morphèmes abstraits aux morphèmes concrets. De même, ils ont effectué des modifications au niveau des règles morphologiques, phonologiques, et orthographiques en ajoutant une règle et en modifiant une autre. [Habash *et al.* 2005] ont modifié la hiérarchie de classes de comportement morphologique, mais des changements mineurs étaient nécessaires. MAGEAD a été adapté pour traiter d'autres dialectes arabes en l'occurrence le DT [Hamdi 2015] [4].

[Abuata & Al-Omari 2015] ont présenté un algorithme pour l'extraction des lemmes pour les dialectes du pays du Golfe (Koweït, Bahreïn, Qatar, Émirats Arabes Unis, l'Est de l'Arabie Saudia et le Sud de l'Irak). Cet algorithme est dédié pour la segmentation des mots dialectaux modernes extraits à partir des forums d'Internet et les sites de conversation. Il permet aussi de segmenter les mots non arabes issus de différentes langues comme le français, l'indien, l'iranien, etc. en se basant sur une liste bien définie de mots. L'évaluation de cet algorithme a donné une précision de 88 % en segmentant des mots en dialectes du Golfe.

1.7 Conclusion

Dans ce chapitre, nous avons étudié l'arabe dialectal et son traitement automatique. D'abord, nous avons exposé la langue arabe en présentant ses différentes variétés. Ensuite, nous avons présenté une classification des dialectes et discuté leurs principales différences et similitudes avec l'ASM. De plus, nous avons présenté les défis du traitement automatique de l'AD, puis, nous avons terminé ce chapitre par une étude des principaux travaux réalisés pour le développement de ressources et d'outils en arabe dialectal.

Dans le chapitre suivant, nous nous intéressons au traitement automatique du dialecte tunisien en présentant et discutant les principaux travaux réalisés pour ce dialecte.

4. Plus de détails sur ce travail seront présentés dans la section 2.4.2.3 du chapitre 2

CHAPITRE 2

Traitement automatique du dialecte tunisien

Sommaire

2.1 Introduction

Le Dialecte Tunisien (DT) est une variété de dialectes arabes. Il appartient aux dialectes de l'Afrique du nord qui sont des variétés moins compréhensibles par les Arabes, notamment à cause du contact avec plusieurs langues étrangères (le berbère, le français, etc.). Depuis 2011, les événements de la révolution tunisienne ont mis la Tunisie au centre de l'attention du monde entier. Le DT est devenu de plus en plus présent et utilisé dans les réseaux sociaux, les blogs, les SMS, les débats, etc. Le besoin de le comprendre et de l'analyser est ainsi devenu progressivement une nécessité. De même, l'évolution des technologies de la parole et des applications

en téléphonie ont augmenté l'exigence de traiter la forme parlée de la langue arabe. Cependant, l'exploitation directe des ressources déjà développées pour l'ASM a donné des faibles performances pour le traitement de l'AD ([Boujelbane *et al.* 2013] ; [Diab *et al.* 2010]).

Dans ce chapitre, nous essayons de cerner l'objet de notre thèse, à savoir le traitement automatique du DT. D'abord, nous présentons le DT en détaillant les principales caractéristiques et similitudes avec l'ASM. La section 2.3 est consacrée à la présentation des motivations du traitement automatique du DT. Finalement, au niveau de la section 2.4, nous abordons les principaux travaux de traitement automatique du DT.

2.2 Le dialecte tunisien

2.2.1 Présentation

Le dialecte tunisien (DT) [Gibson 2009] est le principal dialecte parlé en Tunisie, utilisé par quelques onze millions de personnes qui vivent principalement en Tunisie. Il est, généralement, connu sous le nom de « daArijaħ » ou « ʕaAm˜iyaħ », afin de le distinguer de l'ASM, ou « tuwnsiy » ce qui signifie simplement « *tunisien* ». Le DT est considéré comme une forme de l'arabe, ni codifiée ni standardisée, tout en étant la langue maternelle de toute la population en Tunisie ([Zribi *et al.* 2014] ; [Saidi 2007]). Le DT a principalement des variétés régionales : le dialecte tunisois (le dialecte de Tunis : la capitale), le dialecte sahélien parlé dans certaines villes côtières, le dialecte sfaxien (le dialecte de Sfax), le dialecte du Nord-Ouest (près de l'Algérie), le dialecte du Sud-est (près de la Libye) et le dialecte du Sud-ouest ([Gibson 1998] ; [Khalfaoui 2009] ; [Talmoudi 1983]). Dans cette thèse, nous nous intéressons à la forme dialectale du DT utilisée par les médias. C'est la variété du DT qui est généralement la plus comprise par la majorité des tunisiens.

2.2.2 Caractéristiques

Le DT a des caractéristiques uniques qui le distinguent des dialectes voisins ainsi que des autres dialectes arabes. Comme les dialectes maghrébins, il est fortement influencé par le berbère mais aussi par d'autres langues telles que le turc, l'italien, l'espagnol et le français. La morphologie, la syntaxe, la prononciation et le vocabulaire du DT présentent des différences et des similitudes par rapport à l'ASM ([Mejri *et al.* 2009] ; [Mejri & Baccouche 2003] ; [Habash *et al.* 2012a] ; [Zribi *et al.* 2013a] ; [Zribi *et al.* 2014]).

Dans cette section, nous présentons les principales caractéristiques du DT au niveau phonologique, morphologique, lexical et syntaxique.

2.2.2.1 Phonologie

Nous faisons, dans cette section, une comparaison entre le système consonantique et vocalique des deux variétés de l'arabe.

Le système vocalique du DT. L'ASM ne comporte que trois voyelles courtes (اَ, a), (اُ, u) et (اِ, i) qui correspondent respectivement aux allophones /a/, /u/ et /i/. Ces voyelles peuvent être doublées de leurs correspondantes longues (اَ, aA, /a :/), (اُو, uw, /u :/) et (اِي, iy, /i :/). L'ASM possède aussi deux diphtongues (اَو, Aaw, /aw/) et (اَي, Aay, /ay/).

Le DT garde les mêmes voyelles courtes mais la situation est plus complexe que l'ASM. Le système vocalique du DT se caractérise, d'une part, par la transformation des voyelles longues en voyelles courtes, notamment lorsqu'elles sont situées en position finale des mots, et d'autre part par la négligence des voyelles courtes [Mejri *et al.* 2009].

En effet, les phonèmes des voyelles longues en DT ont toujours des allophones courts, mais les phonèmes des voyelles courtes n'ont pas d'allophones longs. Les voyelles longues en DT sont toutes susceptibles d'un raccourcissement comme dans le cas pour plusieurs dialectes arabes tels que les dialectes maghrébins. Si une voyelle est située à la fin d'un mot portant l'accent sur une seule syllabe (*e.g.* (جَا, jaA, /ja/, « *il est venu* »), (مشَى, mšaỳ, /mša/, « *il est parti* »)), elle sera courte.

Ainsi, comme dans la plupart des dialectes maghrébins, les voyelles courtes sont négligées en DT, notamment lorsqu'elles sont localisées à la fin d'une syllabe [Mejri *et al.* 2009]. Prenons l'exemple du verbe (شَرِبَ, /šariba/) « *il a bu* » de l'ASM. En DT, ce verbe se transforme au (شرِب, /šrib/) en remarquant la suppression de la première et la dernière voyelle (اَ, a). La suppression de la première voyelle, en général, modifie la structure syllabique des unités lexicales qui tendent vers les monosyllabes pour certains mots. Par exemple, le mot (يَدَين, /yadayn/, « *deux mains* ») en ASM et le mot (سَمَاء, samaA', /samaA'/, « *le ciel* ») de l'ASM seront transformés en DT en /ydin/ et /sma/.

La disparition de la diphtongue dans tout un paradigme d'unités au profit d'une voyelle longue est, aussi, considérée parmi les caractéristiques du système vocalique du DT ([Mejri *et al.* 2009] ; [Zribi *et al.* 2014]). En effet, la diphtongue /ay/ de l'ASM est transformée principalement /i :/ dans la majorité des dialectes tunisiens (la variété tunisoise, le sahélien, etc.). Cependant, dans d'autres dialectes tunisiens (*e.g.* le dialecte sfaxien), cette diphtongue garde sa forme de l'ASM. Par exemple, les mots de l'ASM (بَيت, /bayt/, « *la maison* ») et (لَيل, /layl/, « *une nuit* ») et la préposition (بَين, /bayn/, « *entre* ») se prononcent respectivement en DT comme (بِيت, /bi :t/), (بِين, /bi :n/) et (لِيل, /li :l/). De même, le DT connaît l'ajout d'un nouvel allophone pour la voyelle longue (ا, A). Nous signalons que cet allophone n'est pas utilisé en ASM [Zribi *et al.* 2014]. La voyelle (ا, A) peut être réalisée en /a :/ dans certains mots (*e.g.* (شاف, šAf, /ša :f/, « *il a regardé* »), (دار, daAr, /da :r/, « *une maison* »), etc.). Elle est aussi réalisée en /e :/ dans d'autres mots ((لام, lAm, /le :m/, « *il a blâmé* »), etc.). Il existe des mots en DT dont la voyelle longue (ا, A) correspond à deux prononciations différentes avec deux sens différents. Nous prenons l'exemple du mot حرام qui signifie en français « *couverture* » lorsque la voyelle (ا, A) se prononce /e :/ et « *péché* » lorsque la voyelle se prononce /a :/. Le tableau 2.1 présente l'exemple du mot حرام.

	ASM			DT			Signification en français
حرَام	حَرَام	HaraAm	/Hara :m/	حرَام	HraAm	/Hra :m/	*péché*
	حِرَام	HiraAm	/Hira :m/	حرَام	HraAm	/Hre :m/	*une couverture*

TABLE 2.1 – Les différences entre le phonème /e :/ et /a :/ dans le mot حرَام HrAm.

Le système consonantique. Le système consonantique du DT connaît quelques changements en le comparant à l'ASM qui se caractérise par un ensemble de 28 consonnes. Le tableau 2.2 montre la prononciation de ces consonnes en DT et en ASM. Nous signalons que l'alphabet présenté dans ce tableau est défini par Habash-Soudi-Buckwalter [Habash *et al.* 2007]. Cet alphabet est généralement utilisé par la plupart des travaux traitant la langue arabe et ses dialectes.

— Emphase, assimilation phonétique et métathèse.

La prononciation des consonnes se caractérise par trois phénomènes : l'emphase, l'assimilation phonétique et la métathèse [Mejri *et al.* 2009]. Ces phénomènes ont entrainé une variation dans les productions phonétiques relatives aux différentes consonnes.

Tout d'abord, l'emphase est un trait spécifique concernant certaines paires de consonnes en arabe comme les consonnes ((س, s) et (ص, S)) et ((ت, t) et (ط, T)). La prononciation de ces dernières est très proche ; ce qui fait que les mots contenant ces consonnes ont une double prononciation. Par exemple, les mots (صَايغِي, SaAyγiy, « *bijoutier* ») et (طَيَّارَة, Tay˜aAraħ, « *avion* ») ont deux prononciations différentes : /Sa :yγi :/ ou /sa :yγi :/ et /Taya :ra/ ou /taya :ra/. L'emphase touche quelquefois d'autres consonnes ne connaissant pas ce trait [Mejri *et al.* 2009].

L'assimilation phonétique [Mzoughi 2015] est un type très fréquent de modification phonétique subie par un son au contact d'un son voisin (contexte), qui tend à réduire les différences entre les deux. Elle provoque généralement des prononciations spécifiques lexicalement interchangeables. En effet, les mêmes consonnes peuvent être prononcées dans des mots avec des phonèmes et avec d'autres dans d'autres mots. On peut trouver des consonnes avec deux ou trois prononciations différentes, parmi lesquelles nous citons, par exemple, les consonnes (ص, S), (ج, j) et (غ, γ). Ces dernières se prononcent, respectivement, /z/ ou /S/, /j/ ou /z/ et /x/ ou /γ/. Ainsi, les mots (جليز jliyz /jliz/ « *tuiles* »), (صدَاق SdAq /Sde :q/ « *contrat de mariage* ») et (غسَالة γsAlħ /γasse :la/ « *machine à laver* ») sont dits respectivement /zliz/, /zde :q/ et /xasse :la/.

L'assimilation n'est pas le seul phonème responsable de la multitude de prononciations. La métathèse (permutation) ([Mzoughi 2015] ; [Mejri *et al.* 2009]) est, aussi, responsable de cette multitude. Par exemple, le mot (شمس, šms, « *soleil* ») se prononce de trois manières différentes /samš/, /šams/ ou même /sams/.

— Phonèmes non arabes.

Le système phonétique du DT utilise des phonèmes non définis en ASM. Le phonème /g/ est utilisé par plusieurs dialectes arabes ([Zribi *et al.* 2014] ; [Mejri *et al.* 2009]). Il présente, généralement, la prononciation de la consonne (ق, q) de l'ASM. Le plus souvent, (ق, q) est dit

Consonne	Translittération	Prononciation		Exemples	Traductions en français
		ASM	DT		
أ	’	/a/	/a/, /h/ ou omis	سَأل s’al /shal/	*Il a interrogé*
ب	b	/b/		بَاب bAb /bAb/	*Porte*
ت	t	/t/		تفَاحة tfAHaħ /tfa :Ha/	*Pomme*
ث	θ	/θ/	/θ/ ou /f/	ثمة θmħ /θama/ ; فمة fmħ /fama/	*Il y a*
ج	j	/j/	/j/ ou /z/	جزَار jaz ~ar /jazza :r/ ou /zazza :r/	*Boucher*
ح	H	/H/		حَديد Hadid /Hadid/	*Fer*
خ	x	/x/		خَالي xaAlyi /xaAli :/	*Oncle*
د	d	/d/		دَار dAr /da :r/	*Maison*
ذ	ð	/ð/		ذبَانة ðbAnaħ /ða :na/	*Mouche*
ر	r	/r/		رسول rasul /raSul/	*Prophète*
ز	z	/z/		زربية zrbyħ /zarbiya/	*Tapis*
س	s	/s/	/s/ ou /z/	رسول rasul /raSul/	*Prophète*
				فُستق fustaq /fuzdaq/	*Pistache*
ش	š	/ š/		شمس šams / šams/	*Soleil*
ص	S	/S/	/S/ ou /s/ ou /z/	زدَاق zdAq /zde :q/	*Contrat de mariage*
				صَايغي SaAygiy /SaAygiy/ ou /saAygiy/	*Un bijoutier*
ض	D	/D/	/D/ ou /Ď/	ضفيرة Dafira /Dafira/	Tresse
ط	T	/T/	/T/ ou /t/	طيَارة TayArap /Tayara/ ou /tayara/	*Avion*
ظ	Ď	/Ď/	/Ď/ ou /D/	ظلمة Ďalmaħ /Ďalma/	*Obscurité*
ع	ς	/ς/	/ς/ ou /H/	متَّاعهَا mteHhA /mteHha/	*La sienne*
غ	ɣ	/ɣ/	/ɣ/ ou /x/	غسَالة ɣasAlħ /xase :la/	*Machine à laver*
ف	f	/f/		فَار fAr /fa :r/	*Souris*
ق	q	/q/	/q/ ou /g/	بقرة baqraħ /bagra/	*Vache*
ك	k	/k/		كَاس kAs /ka :s/	*Verre*
ل	l	/l/		لعبة lςbħ /luςba/	*Jouet*
م	m	/m/		معلم mςlm /muςalim/	*Instituteur*
ن	n	/n/		نَار nAr /na :r/	*Feu*
ه	h	/h/		هلَال hlAl /hle :l/	*Croissant*
و	w	/w/		ورقة wrqħ /warqa/	*Feuille*
ي	y	/y/		يَاسمين yAsmyn /yasmi :n/	*Jasmin*

TABLE 2.2 – La prononciation des consonnes en DT et en ASM.

/q/ dans les dialectes urbains et ruraux, avec le /q/ prédominant dans les dialectes urbains et le /g/ dans les dialectes ruraux [Mejri *et al.* 2009]. Par exemple, « *il a dit* » se prononce /ga :l/ dans les dialectes ruraux. En dialectes urbains, (ق, q) est principalement prononcée /q/ (*e.g.* قَال, /qa :l/) sauf pour certains mots qui ont des origines rurales comme (بَقرَة, baqraħ, /bagra/, « *une vache* »). Parfois, la sémantique du mot change si on change la prononciation de la lettre (ق, q). Le mot (قرون, qrwn, /qru :n/) signifie en français « *siècles* ». Par contre, la

prononciation /gru :n/ signifie « *les cornes* ». Ainsi, le phonème /g/ est utilisé dans plusieurs mots en DT qui ont une origine berbère (*e.g.*, (bilqdA, /bilgda :/, « *Très bien* ») et (qurbiyTaħ, /gurbi :ta/, « *ruban* »)).

L'emprunt massif des mots d'autres langues a introduit de nouveaux phonèmes qui ne sont pas définis dans le système consonantique de l'ASM, comme /v/, /p/ et /g/ [Zribi *et al.* 2014]. Les mots (/purtaAbl/, « *portable* »), (/talvza/, « *télévision* ») et (/gazu :z/, « *soda* ») présentent des mots utilisés dans le DT contenant des phonèmes non arabes.

— La consonne « Hamza ».

Les mots de l'ASM qui contiennent la consonne « Hamza » [1] perdent cette consonne lorsqu'ils se prononcent en DT. Elle se transforme en une voyelle longue ou disparaît complètement. Généralement, la transformation est influencée par la voyelle courte située avant la consonne « Hamza ». Le tableau 2.3 illustre les transformations de « Hamza » en passant de l'ASM vers le DT.

ASM		DT		Signification en français
Phonème	**Exemples**	**Phonème**	**Exemples**	
/a+'/	كَأَس /ka'as/	/a :/	كَاس /ka :s/	*verre*
/i+'/	بِئر /bi'r/	/i :/	بير /bi :r/	*puit d'eau*
/u+'/	مؤمن /mu'min/	/u :/	مومن /mu :min/	*croyant*
/.'/	سَمَاء /sama :'/	//	سَمَا /sma :/	*ciel*

TABLE 2.3 – Quelques exemples de mots contenant la lettre Hamza.

2.2.2.2 Morphologie

Il existe des différences significatives entre la morphologie du DT et celle de l'ASM. Les principales différences sont autour de l'introduction de nouveaux clitiques et la simplification du système d'inflexion.

Flexion.

— Au niveau des verbes.

La disparition du duel, du genre féminin et du mode verbal sont les principales différences avec l'ASM. Le tableau 2.4 montre une comparaison entre les traits morphologiques verbaux de l'ASM et du DT. Notons que certains dialectes en Tunisie font la distinction entre le singulier masculin et féminin.

La conjugaison connaît des simplifications au niveau du système d'affixation [Ouerhani 2009]. Le DT a normalisé, également, les affixes pour la première et la deuxième personne du singulier et aussi pour les affixes de la deuxième et la troisième personne du singulier. Le tableau 2.5 présente les différences de conjugaison en DT et ASM pour le verbe (خرج, xrj,« *sortir* »).

1. La lettre Hamza a quatre formes différentes en ASM : أ, ئ, ء, ؤ suivant sa position dans le mot.

<table>
<tr><td colspan="2"></td><td colspan="3">ASM</td><td colspan="3">DT</td></tr>
<tr><td colspan="2" rowspan="3">Mode</td><td colspan="3">Indicatif</td><td colspan="3" rowspan="3">-</td></tr>
<tr><td colspan="3">Subjonctif</td></tr>
<tr><td colspan="3">Jussif</td></tr>
<tr><td colspan="2" rowspan="3">Aspect</td><td colspan="6">Accomplie</td></tr>
<tr><td colspan="6">Inaccomplie</td></tr>
<tr><td colspan="6">Impératif</td></tr>
<tr><td colspan="2">Personne</td><td>1</td><td>2</td><td>3</td><td>1</td><td>2</td><td>3</td></tr>
<tr><td rowspan="6">Genre et nombre</td><td rowspan="3">Masculin</td><td>Singulier</td><td>Singulier</td><td>Singulier</td><td>Singulier</td><td>Singulier</td><td>Singulier</td></tr>
<tr><td>-</td><td>Duel</td><td>Duel</td><td>-</td><td>-</td><td>-</td></tr>
<tr><td>Pluriel</td><td>Pluriel</td><td>Pluriel</td><td>Pluriel</td><td>Pluriel</td><td>Pluriel</td></tr>
<tr><td rowspan="3">Féminin</td><td rowspan="3">-</td><td>Singulier</td><td>Singulier</td><td>-</td><td>-</td><td>Singulier</td></tr>
<tr><td>Duel</td><td>Duel</td><td>-</td><td>-</td><td>-</td></tr>
<tr><td>Pluriel</td><td>Pluriel</td><td>-</td><td>-</td><td>-</td></tr>
</table>

TABLE 2.4 – Comparaison entre les traits morphologiques verbaux de l'ASM et du DT.

La forme passive des verbes présente une différence entre l'ASM et le DT. Pour transformer les verbes trilitères de l'ASM en forme passive, on utilise la méthode d'infixation. Donc, on ajoute les voyelles (اَ-اِ-اُ, u-i-a) pour dériver la forme passive. Par exemple, le verbe (كَتَبَ, kataba, « *écrire* ») doit suivre le modèle verbal (فُعِلَ, $r_1ur_2ir_3a$) pour être en forme passive et le verbe sera (كُتِبَ, kutiba, « *a été écrit* »). Par contre, en DT, on ajoute le préfixe (ت, t) au verbe pour qu'il soit en forme passive. Le verbe en DT (كتِب, ktib, « *écrire* ») est transformé en (تكتِب, tktib, « *a été écrit* ») à la forme passive [Maalej 1999].

— Au niveau des noms.

Le DT est caractérisé par l'absence des marques casuelles nominales. Il perd souvent la forme nominale duelle qui est remplacée par le numéral (زُوز, zuwz, /zu :z/, « *deux* ») suivi par le pluriel. Par exemple, (أُستَاذَين, ÂustaAðayn, « *deux professeurs* ») de l'ASM est traduit en (اَسَاتذَة زُوز, zuwz AasaAtðaħ,) pour le DT [Zribi *et al.* 2014]. Notons que certains noms comptables comme par exemple خُبزَة « un pain » et مِيَة « cent » garde la forme nominale duelle : (خُبزتين « deux pains ») et (مِيتين « deux cent »).

Agglutination. Les affixes et les clitiques en DT connaissent plusieurs différences par rapport à l'ASM. Comme plusieurs dialectes arabes, le DT introduit de nouveaux clitiques non définis en ASM. Parmi ces clitiques, nous citons le clitique de négation. Notons que la négation en ASM est exprimée en utilisant l'une de ces particules : (مَا, maA), (لَا, laA), (لَن, lan) et (لَم, lam). En DT, la négation utilise la lettre de négation (ش+, +š) qui est attachée à la position finale du verbe. Elle a la forme suivante : (ش+verbe+esp+مَا, mA+esp+verbe+ š). Un autre nouveau clitique concerne l'interrogation (شي, šy). Ce dernier remplace le clitique d'interrogation de l'ASM (أَ, Â) et la particule d'interrogation (هَل, hal).

Le DT a un ensemble de clitiques qui sont des formes réduites des particules de l'ASM. Le proclitique démonstratif (هَ, ha+) agglutiné à l'article défini (ال+, Al+) est le résultat d'une

	Accomplie		Inaccomplie		Impératif	
	ASM	DT	ASM	DT	ASM	DT
1ms	خَرَجتُ xaraj+tu	خرَجت xraj+t	اَخرُجُ Aa+xruju	نُخرُج nu+xruj		
1mp	خَرَجنَا xaraj+naA	خرَجنَا xraj+naA	نَخرُجُ na+xruju	نُخرُجوَا nu+xrj+uwA		
2ms	خَرَجتَ xaraj+ta	خرَجت xraj+t	تَخرُجُ ta+xruju	تُخرُج tu+xruj	أُخرُج Au+xruj	أُخرُج Au+xruj
2mf	خَرَجتِ xaraj+ti		تَخرُجِينَ ta+xruj+iyna		أُخرُجِي Au+xruj+iy	
2md	خَرَجتُمَا xaraj+tumaA	خرَجتُوَا xraj+tuwA	تَخرُجَانِ ta+xruj+aAni	تُخرُجوَا tu+xrj+uwA	أُخرُجَا Au+xruj+aA	أُخرُجوَا Au+xrj+uwA
2fd	خَرَجتُمَا xaraj+tumaA		تَخرُجَانِ ta+xruj+aAni		أُخرُجَا Au+xruj+aA	
2mp	خَرَجتُم xaraj+tum		تَخرُجُونَ ta+xruj+uwna		أُخرُجوَا Au+xruj+uwA	
2fp	خَرَجتُنَّ xaraj+tun~a		تَخرُجنَ ta+xruj+na		أُخرُجنَ Au+xruj+na	
3ms	خَرَجَ xaraja	خرَج xraj	يَخرُجُ ya+xruju	يُخرُج yu+xruj		
3fs	خَرَجَت xaraj+at	خَرِجت xarj+it	تَخرُجُ ta+xruj+u	تُخرُج tu+xruj		
3md	خَرَجَا xaraj+aA	خَرجُوَا xarj+uwA	يَخرُجَانِ ya+xruj+aAni	تُخرُجوَا tu+xrj+uwA		
3fd	خَرَجتَا xaraj+ataA		تَخرُجَانِ taxruj+aAni			
3mp	خَرَجُوَا xaraj+uwA		يَخرُجُونَ ya+xruj+uwna			
3fp	خَرَجنَ xaraj+na		يَخرُجنَ ya+xruj+na			

TABLE 2.5 – Conjugaison du verbe (خرج, xrj, « sortir ») en DT et ASM.

réduction des pronoms démonstratifs de l'ASM (هذَا, hðA, « *ce* ») et (هذه, hðh, , « *cette* »). En outre, le DT a le proclitique (+ع, ʕa+) et le proclitique (+م, m+) qui sont respectivement des formes réduites de la préposition (علَى, ʕlỳ, « *sur* ») et la préposition (من, min, « *de* ») ou la conjonction de coordination (مع, mʕ, « *avec* »). Par exemple, les syntagmes nominaux de l'ASM (هذَا الطفل, hðA AlTifl, « *cet enfant* »), (علَى الطَاولة, ʕlỳ AlTAwlħ, « *sur la table* »), (من الدَار, mn AldAr, « *de la maison* »), et (مع بعضنَا, mʕ bʕDnA, « *tous ensemble* ») deviennent, respectivement, en DT (هَالطفل, haAlTfil), (عَالطَاولة, ʕAlTAwlħ), (مَالدَار, mAldAr) et (مبعضنَا, mbʕDnA). Le tableau 2.6 présente l'ensemble de clitiques et d'affixes pour l'ASM et leur équivalent en DT.

Dérivation et Emprunt. En arabe, la plupart des mots se dérivent à partir d'une racine consonantique. On peut dériver selon des schèmes préétablis en impliquant par exemple

Clitiques / affixes	ASM	DT	Exemples	Signification en français
Enclitique pronominale	كَ ka	ك k	كتَابِك ktaAbik	*ton livre*
	كِ ki			
	ه h	ه h	ضَرْبُوه Darbuwh	*ils l'ont battu*
		و w	كتَابُو ktaAbuw	*son livre*
	هَا haA	هَا haA	كتَّابهَا ktaAbhaA	
	هُم hum	هُم hum	كتَّابهُم ktaAbhum	*leur livre*
	كُم kum	كُم kum	كتَّابكُم ktaAbkum	*votre livre*
	نَا naA	نَا naA	كتَّابنَا ktaAbnaA	*notre livre*
	ي y	ي y	كتَّابِي ktaAbiy	*mon livre*
Enclitique de négation	لن لم لَا lm lA ln	ش š	مَا قرِيتش maA qriytš	*je n'ai pas lu*
Enclitique d'interrogation	أ Â ou هل hal	-شِي -šiy	قرِيتشِي qriytšiy	*As-tu lu ?*
Proclitique	و w	-و w-	وَشْرِب wašarib	*et il a bu*
	ل li	-لِ li-	لِدَّار lidaAr	*pour la maison*
	ب bi	-بِ bi-	بِدَّار bidaAr	*dans la maison*
	علَى ʿalaý	-عَ ʿa-	عَالطَّاولَة ʿaAlTaAwlaħ	*sur la table*
	ك ki	-كِ ki-	كِالدَّار kiAld~aAr	*comme la maison*
	ال- Al-	ال- Al-	البَيت Albay.t	*la maison*
	مع mʿ	م- m-	مبعضنَا mbʿDnA	*ensemble*
	من الـ mn Al-	مَال- mAl-	مَالبَيت mAlbayt	*de la maison*

TABLE 2.6 – L'ensemble des clitiques et affixes pour le DT

une variation vocalique ou en ajoutant certains éléments consonantiques à partir de l'ASM [Mejri *et al.* 2009]. Ce phénomène est encore utilisé pour la constitution des mots du DT où l'ajout des affixes constitue l'opération la plus fréquente. On utilise la préfixation et à la suffixation pour former de nouvelles unités lexicales. On emprunte souvent de nouveaux suffixes à partir d'autres langues. Par exemple, on ajoute le suffixe d'origine turque (+جي, +jy) et le suffixe français (+يست, +ist, « -iste ») (*e.g.* (قهوَاجي, qahwaAjiy, « *un cafetier* »), (بنكَاجي, bankaAjiy, « *un banquier* »), (خبزيست, xubziyst, « *une personne dont la seule idéologie est de gagner son pain* »), etc.) ([Mejri *et al.* 2009] ; [Mejri & Baccouche 2003]). En plus, au niveau du DT, la dérivation peut se faire dans certains cas en utilisant des schèmes spécifiques combinés avec l'ajout des affixes. Par exemple, le mot (كوَارجي, kaw˜arjiy, « *footballeur* ») est dérivé à partir du mot (كورة, kuwraħ, « *ballon* ») en suivant le schème (فَعَّال, faʿ˜aAl) en ajoutant le suffixe (jiy, جي) [Mejri *et al.* 2009].

Le système verbal tunisien contient plusieurs verbes empruntés notamment au français [Ouerhani 2009]. Ces verbes sont intégrés dans le système dialectal à travers le moulage des schèmes [Ouerhani 2009]. La matière consonantique du verbe emprunté est versée dans un schème arabe lui permettant de se conjuguer [Ouerhani 2009]. Souvent les verbes empruntés sont transformés via le schème faʿlil. Par exemple, le verbe français « *jongler* » se transforme en jangil, de même pour le syntagme verbal « *avoir sa maîtrise* » qui sera transformé en ma-

triz [Ouerhani 2009]. On dérive à partir de ces verbes empruntés des paradigmes entiers contenant la conjugaison des verbes, des adjectifs et des noms. Dans certains cas, il suffit d'ajouter au verbe emprunté des affixes pour lui permettre de se conjuguer [Ouerhani 2009]. Par exemple, pour conjuguer le verbe français « *installer* » en DT, on ajoute l'affixe (ي, y), on obtient (يَنستَالِي, /yansta :li/, « *il installe* »). De même, à partir du mot emprunté /fuskupi :/ « *fausse-copie* », on peut dériver un verbe accompli (faska, « *il a fait une fausse-copie* »), un verbe inaccompli (yfaskyi, « *il fait une fausse-copie* »), un agent singulier (faskaAy, « celui qui fait une fausse-copie »), un agent pluriel (faskaAya, « *ceux qui font une fausse-copie* ») et un nom prédicatif (Masdar) (tfaskiyaħ, « *action de faire fausses-copies* »). Cet exemple est extrait à partir de [Mejri *et al.* 2009].

2.2.2.3 Lexique

Le lexique du DT est généralement constitué de mots issus de plusieurs phénomènes systémiques tels que la dérivation et l'emprunt d'autres langues [Mejri *et al.* 2009]. Une des principales raisons qui ont différencié le vocabulaire du DT de celui de l'ASM est l'utilisation massive de mots empruntés de l'italien, de l'espagnol, du français, du berbère et du turc. En effet, l'emprunt massif est le résultat de plusieurs événements historiques qui ont rendu la situation linguistique en Tunisie assez complexe. Le tableau 2.7 présente quelques exemples de mots empruntés en DT [Zribi *et al.* 2014].

Mots	Translittération	Signification en français	Origine
بَرَاكة	barAkaħ	cabine	Italien
بَانكة	baAnkaħ	banque	
دَاكوردو	daAkuwrduw	d'accord	
فيشطة	fiyšTaħ	Fête	
مَاكينة	maAkiynaħ	machine	
كروسة	karuwsaħ	carrosse	
بَابور	baAbuwr	bateau	Turc
سفنَارية	sfnaAriyaħ	carotte	
قهوَاجي	qahwaAjiy	serveur	
سبيطَار	sbiyTaAr	hôpital	
برنوس	barnuws	vêtement traditionnel	Berbère
كسكسي	kusksiy	couscous	
بطَانية	baTaAniyaħ	couverture	
صبَاط	SabaAT	chaussure	Espagnol
بوسطة	buwsTaħ	la poste	Français
بلَاصة	blaASaħ	Place	
بَاكو	baAkuw	paquet	

TABLE 2.7 – Exemples de mots en DT empruntés d'autres langues.

2.2.2.4 Syntaxe

Le DT est considéré comme une variante de l'ASM. Dans certains cas, les caractéristiques syntaxiques du système dialectal tunisien sont en rupture complète avec l'ASM.

D'une part, le système pronominal du DT est proche de celui de l'ASM, à quelques simplifications près. Il se caractérise par une confusion entre les genres de certains pronoms ainsi que la disparition complète de la forme duelle de l'ASM [Mejri *et al.* 2009]. Le système pronominal personnel est condensé à 7 pronoms personnels par opposition aux 12 pronoms pour l'ASM (*cf.* tableau 2.8).

ASM	DT	Traduction
أَنَا ÂanaA	آنَا ĀnaA	Je
أَنتَ Âan.ta	إِنتِي Ăintiy	Tu (masculin)
أَنتِ Âanti		Tu (féminin)
أَنتُمَا ÂantumaA	إِنتُومَا ou إِنتُوم ĂintuwmaA ou Ăintuwm	Vous (duel)
أَنتُم Âantum		Vous (masculin)
أَنتُنَّ Âantun~a		Vous (féminin)
نَحنُ naHnu	أَحنَا ÂaHnaA	Nous
هُوَ huwa	هُوَ huwa	Il
هِيَ hiya	هِيَ hiya	Elle
هُمَا humaA	هُومَا huwmaA	Ils (duel)
هُم hum		Ils
هُنَّ hun~a		Elles

TABLE 2.8 – Comparaison entre les pronoms personnels de l'ASM et ceux du DT.

Les pronoms démonstratifs du DT ont abandonné, aussi, leur forme duelle au féminin et au masculin ((هذَان, haðaAni) et (هَاتَان, haAtaAni)) [Mejri *et al.* 2009]. Le tableau 2.9 montre les pronoms démonstratifs en DT.

L'ASM possède plusieurs pronoms relatifs qui varient en fonction du nombre et du genre. Les pronoms relatifs de l'ASM ont été réduits à un seul pronom relatif (اللّي, Ailliy, « *qui* ») [Mejri *et al.* 2009].

ASM	DT	Genre	Nombre	Signification en français
هذَا hðA	هذَا hðA هذَايه hðAyh	masculin	singulier	ceci, celui-ci
ذلكَ ðalika	هذَاكه haðaAkah	masculin	singulier	cela, celui-là
هذه haðihi	هذي haðiy	féminin	singulier	celle-ci
تِلكَ tilka	هَاكي haAkiy	féminin	singulier	celle-là
هَؤُلَاء haŵulaA’	هذومَا haðuwmaA	masculin et féminin	pluriel	ceux-ci, celles-ci
أُولئِكَ Âuwlaŷika	هذوكم haðuwkum	masculin et féminin	pluriel	ceux-là, celles-là

TABLE 2.9 – Extrait des pronoms démonstratifs en DT et leurs équivalents en ASM.

D'autre part, l'ordre de mots dans les phrases n'est pas considéré comme un point de rup-

ture entre l'AD et l'ASM. En langue arabe, généralement, on distingue deux types de phrases : verbale et nominale.

La phrase verbale est une phrase contenant au moins deux éléments : un sujet (S) et un verbe (V) auxquels s'ajoute souvent un objet (O). Elle ne peut débuter que par un verbe. La structure d'une phrase de ce type peut être du type (VSO) ou (VOS). La phrase nominale est formée par le rapprochement de deux éléments : un sujet et un attribut. L'attribut peut être une phrase verbale. Elle doit commencer par un nom. La structure d'une phrase de ce type peut être soit (SVO) soit (SOV).

Les phrases en DT gardent les mêmes structures que celles des phrases en ASM, mais, la structure la plus dominante en DT est (SVO) alors qu'en ASM la structure la plus dominante est (VSO) [Saidi 2014].

2.2.2.5 L'alternance codique

Le DT est un mode de communication construit sur l'alternance codique entre le français et l'arabe dialectal tunisien [Ksouri 2013]. Bien que la langue officielle de la Tunisie soit l'ASM, la langue française continue à être utilisée depuis la colonisation française de la Tunisie. Ceci engendre une situation linguistique en Tunisie à la fois diglossique (ASM/DT) et bilingue : arabe/français [Saidi 2007]. Les Tunisiens utilisent spontanément trois langues : ASM, DT et le français dans une même conversation. Le changement de langue peut avoir plusieurs motivations, par exemple la facilité d'utiliser une langue dans un sujet donné et ce indépendamment de la pratique de la langue d'alternance. Les personnes qui pratiquent l'alternance codique ne sont pas forcément bilingues. Ils ont une connaissance bilingue à des degrés différents, et peuvent quelquefois ignorer totalement la langue française.

En effet, les Tunisiens utilisent dans leur quotidien des mots et même des expressions de la langue française sans aucune modification au niveau phonologique ou morphologique (*e.g.* « *ça va* », « *désolé* », « *rendez-vous* », « *mécanicien* », « *plombier* », « *technicien* », etc.). Les exemples (1) et (2) présentent deux phrases en DT contenant des mots et expressions en français. De même, l'alternance entre deux langues en Tunisie affecte le lexique du dialecte. Ceci permet d'introduire de nouveaux mots dialectaux dérivés des langues étrangères.

1. لَازم نعمل confirmation للّ réservation متَاع l'hôtel

   lAzm nçml confirmation *ll* réservation *mtAç* l'hôtel

   « *Je dois confirmer la réservation d'hôtel* ».

2. شني حوَالك يَا ولدي ? ça va ?

   šny HwAlk yA wldy ? ça va ?

   « *Comment vas-tu mon fils ? ça va ?* »

2.3 Motivations pour le traitement du dialecte tunisien

Le DT se veut une variété de langues, ainsi, quelques linguistes se trouvent du mal à accepter son traitement. C'est pourquoi ils estiment plus bénéfique de concentrer les efforts sur la langue mère : l'ASM. Il est donc nécessaire de préciser les raisons conduisant à l'analyse des dialectes arabes de façon générale et le DT plus précisément.

Dans cette section, nous présentons nos motivations pour traiter le DT.

— La langue maternelle des tunisiens.

Le DT est la langue maternelle de la majorité des tunisiens [Saidi 2007]. C'est la première langue qu'on apprend dès la naissance. Elle est acquise spontanément sans aucun apprentissage formel quelconque [Mejri *et al.* 2009]. Le DT est surtout parlé dans le cadre d'un dialogue quotidien entre Tunisiens et au sein de la famille. Les pratiques du DT se retrouvent également dans divers secteurs socioprofessionnels et dans les arts et spectacles (théâtre, musique, cinéma et littérature) [Mzoughi 2015]. Les tunisiens sont parfois incapables de comprendre et de s'exprimer spontanément en ASM. Face à cette situation sociolinguiste en Tunisie, le besoin d'analyser et de comprendre cette langue sera plus évident pour les tunisiens et même pour les étrangers qui veulent bien s'installer au pays. Ainsi, après les évènements de la révolution tunisienne, une version de la constitution a été élaborée en DT afin de la rendre plus compréhensible par les tunisiens.

— Développement des technologies de traitement de la parole.

De nos jours, les applications utilisant la voix et la parole et plus généralement les technologies de traitement de parole ne cessent de se développer [Graja 2015]. Diverses applications visent à permettre à un utilisateur humain d'accéder aux informations ou aux services disponibles sur un ordinateur ou sur Internet en utilisant la langue parlée comme moyen d'interaction. Plusieurs systèmes commercialement disponibles sont capables d'utiliser la parole comme une entrée ou une sortie. Parmi ces applications, citons les systèmes de contrôle et commande, les systèmes de reconnaissance de parole, les systèmes de dictée automatique, les applications de téléphonie mobile (*e.g.* Siri, lecteur d'écran, lecteur de SMS, etc.), etc. [Zribi *et al.* 2014]. Ils offrent une variété de services en utilisant diverses ressources et outils développés pour la langue parlée.

Le développement actuel de ce type de technologie justifie l'importance du traitement automatique de la forme parlée du DT qui est la langue natale des tunisiens à l'opposé de l'ASM. Il sera avantageux de trouver des applications comme Google Voice Search « OK Google »[2] ou Siri capable de comprendre les Tunisiens en parlant avec leur langue maternelle sans d'être obligé d'utiliser une autre langue comme le français ou l'anglais que la majorité des Tunisiens ne les maitrisent pas parfaitement.

— Volume des ressources en DT sur le Web.

En effet, le contenu en DT est en évolution constante sur le Web. Le nombre important de réseaux sociaux, de blogs, des sites de chat, etc. sont la principale cause de développement du

2. https ://www.google.com/search/about/

contenu dialectal sur le net [Mubarak & Darwish 2014]. Les internautes utilisent souvent leurs dialectes pour commenter un contenu du Web : les vidéos, leurs photos et publications, etc. Par ailleurs, la publication en AD peut également provenir d'une méconnaissance de l'ASM.

Le besoin de déchiffrer ce nombre important de productions écrites en dialecte et de les analyser est ainsi devenu progressivement une nécessité. De même, on peut également exploiter ce contenu dialectal pour enrichir les ressources de l'ASM à travers l'application de nombreuses méthodes de TAL (*e.g.* les outils de traduction automatique) permettant l'exploitation de ce genre de contenu [Abo Bakr *et al.* 2008].

2.4 Les travaux réalisés pour le dialecte tunisien

Depuis les années de 2010 et surtout après la révolution tunisienne, certains chercheurs se sont concentrés sur le traitement automatique du DT. Leurs travaux sont encore préliminaires. Quelques chercheurs se sont intéressés au développement des ressources (corpus, Wordnet, lexique, etc.). D'autres se sont focalisés sur le développement d'outils de TAL. Dans cette section, nous présentons un aperçu sur les principaux travaux réalisés au profit du DT.

2.4.1 Construction de corpus et de ressources lexicales pour le dialecte tunisien

La création des corpus en traitement automatique des langues est une étape nécessaire au traitement d'une langue particulière. À partir de cette ressource, on peut créer d'autres telles que des lexiques et/ou des applications en TAL. Le DT est une langue peu dotée avec une absence quasi totale de ressources capables d'être exploitées directement. La création de corpus pour le DT constitue donc une étape importante au traitement automatique de ce dialecte.

2.4.1.1 Corpus écrits

Acquisition des données à partir de diverses ressources. [McNeil & Faiza 2011] ont construit leur corpus en se basant sur plusieurs ressources. Ainsi, pour collecter un corpus de 859 814 mots en DT, [McNeil & Faiza 2011] ont utilisé trois types de ressources : les sources traditionnelles écrites (les folklores, les poèmes et chansons folkloriques, des collections de proverbes, les émissions de télévision, les pièces de théâtre et les scénarios des feuilletons en DT), les nouvelles sources écrites (les blogs sur le Web, les forums et Facebook) et les transcriptions de fichiers audio (transcriptions des émissions radiophoniques).

Le corpus de [McNeil & Faiza 2011] est organisé sous forme d'une application Web qui gère et organise les fichiers de corpus et les métadonnées, et accomplit les traitements linguistiques de base (comme des listes de fréquences, les collocations, et une concordance).

Bien qu'il soit librement consultable, il ne suit pas une convention pour la transcription des mots en dialecte. De plus, il est formé de plusieurs textes et récits en ASM et non plus en DT. Ainsi, nous ne pouvons pas le considérer comme une ressource textuelle spécifique

au DT. Notons que ce corpus a été utilisé pour construire un dictionnaire bilingue ANG/DT [McNeil & Faiza 2011].

Traduction des ressources de l'ASM vers le DT. Dans le cadre de la modélisation d'un système de reconnaissance automatique du DT, [Boujelbane *et al.* 2014a] ont proposé une méthode permettant le développement de ressources textuelles à partir d'un ensemble de ressources pour l'ASM. Il s'agit de convertir un corpus d'ASM en DT. [Boujelbane *et al.* 2014a] ont utilisé le corpus *Arabic TreeBank* (ATB) [Maamouri & Bies 2004] pour développer des dictionnaires (ASM/DT). À partir de ces dictionnaires, un corpus pour le DT est généré. En effet, [Boujelbane *et al.* 2014a] ont développé trois lexiques bilingues (ASM/DT) : un lexique pour les verbes, un autre pour les noms et finalement un lexique pour les mots-outils. Il s'agit principalement de collecter les données à partir du corpus ATB et de proposer une ou plusieurs traductions pour chaque terme traité. Ainsi, pour construire le corpus, un ensemble de règles syntaxiques permettant de convertir les structures syntaxiques vers des structures équivalentes en DT ont été proposées en suivant l'approche proposée par [Chiang *et al.* 2006].

Ce travail de recherche a abordé la tâche de création des ressources pour le DT. Il repose cependant sur le traitement d'une forme très particulière du dialecte qui se base souvent sur une alternance codique entre ASM et DT, ce qui n'est pas une caractéristique majeure du DT. La plupart des chercheurs en linguistique définissent en effet la situation linguistique en dialecte comme une situation dont l'alternance codique entre le français et le D est le principal trait du DT.

Collection de messages téléphoniques et des réseaux sociaux. La création d'un corpus pour le DT était, également, l'objet du travail de [Younes & Souissi 2014]. Ce travail s'est concentré sur les écritures avec les lettres latines pour le DT. [Younes & Souissi 2014] ont présenté plusieurs méthodes pour collecter un corpus pour le dialecte qui s'appuie sur les messages SMS. Ils ont tout d'abord collecté les messages SMS envoyés à travers les téléphones mobiles. Comme deuxième source, ils ont assemblé les messages en employant un formulaire créé via Google docs avec lequel les utilisateurs sont invités à écrire ou copier leurs messages en DT. De plus, [Younes & Souissi 2014] ont collecté les modèles de messages envoyés lors des cérémonies, les fêtes, etc. Les réseaux sociaux sont une source de données en dialecte utilisée par plusieurs travaux de collecte de corpus. [Younes & Souissi 2014] ont utilisé aussi cette ressource pour collecter les messages, les publications et les commentaires de Facebook. Ce corpus est enrichi par des messages proposés par un ensemble d'étudiants dans le cadre des mini-projets réalisés dans leur université. Tous ces efforts ont permis de collecter 43 222 messages.

De même, [Masmoudi *et al.* 2015] ont collecté un corpus pour le DT composé de messages et de commentaires des réseaux sociaux. Ils ont collecté des messages SMS, des commentaires de Facebook et des commentaires de YouTube. Ce corpus est composé de 870 904 mots en DT.

2.4.1.2 Corpus oraux

Enregistrement de parole. Deux corpus oraux ont été développés par [Masmoudi *et al.* 2014] et [Graja *et al.* 2013]. Ces corpus sont enregistrés dans les stations de la gare de la *Société Nationale des Chemins de Fers Tunisiens*[3]. Ils présentent des conversations entre un agent de la gare et des voyageurs qui demandent des informations sur les horaires des voyages, le prix des billets, etc. Le premier corpus TuDiCoI (*Tunisian Dialect Corpus Interlocutor*) est celui de [Graja *et al.* 2013], qui a été développé dans le cadre d'un projet de compréhension automatique du DT. Il s'agit de 1 825 dialogues composés de 12 182 énoncés de 1 831 utilisateurs. Ce corpus est annoté sémantiquement. Aucune information n'est mentionnée sur la convention de transcription utilisée ni sur l'outil d'aide à la transcription utilisée.

Le deuxième corpus TARIC (*Tunisian Arabic Railway Interaction Corpus*) [Masmoudi *et al.* 2014] est développé dans le cadre d'un projet de reconnaissance de parole. Il est composé de 20 heures de parole transcrites en utilisant l'outil d'aide à la transcription Transcriber[4]. Ce corpus est composé de 71 684 mots et transcrit en respectant la convention de transcription « CODA-TUN » [Zribi *et al.* 2014] que nous proposons dans notre thèse.

L'inconvénient majeur de ces deux corpus est leur vocabulaire qui est très restreint. Ils ne présentent qu'une partie réduite du lexique du DT, ce qui fait que l'utilisation de ces corpus est non pertinente pour notre sujet de thèse.

Transcription des émissions télévisées. [Boujelbane *et al.* 2014a] ont présenté leur corpus du DT collecté à partir de la transcription des émissions des débats politiques et les journaux télévisés. Ce type d'émissions regroupe des discussions dont la langue utilisée est très riche en ASM avec une alternance codique du DT. En effet, le pourcentage de contenu en DT ne dépasse pas les 37,2 % dans le meilleur des cas. Il s'agit principalement de traiter le dialecte intellectualisé qui représente un mélange entre le DT et l'ASM. [Boujelbane *et al.* 2014a] ont effectué la transcription (en utilisant Transcriber) pour 5 heures et 20 minutes d'enregistrements. Elle est composée de 37 964 formes. Le corpus de [Boujelbane *et al.* 2014a] comme celui de [Masmoudi *et al.* 2014] a aussi respecté notre convention orthographique du DT CODA-TUN [Zribi *et al.* 2014]. En fait, il a été utilisé pour mesurer l'impact de la couverture lexicale et de la perplexité d'un modèle de langage appris sur un corpus en DT issu d'une traduction d'un corpus en ASM dans le contexte de la reconnaissance automatique de la parole.

2.4.1.3 Wordnet

Les ressources pour le DT ne sont pas limitées à la création des corpus et des lexiques, mais, il existe deux travaux qui se sont engagés à créer un Wordnet pour le DT. Le premier est

3. http ://www.sncft.com.tn/
4. http ://trans.sourceforge.net/en/presentation.php

celui de [Bouchlaghem *et al.* 2014]. Une approche à base de corpus a été exploitée pour créer un Wordnet TunDiaWN pour le DT. Cette approche est composée de quatre étapes. La création d'un corpus est la première étape de cette approche. [Bouchlaghem *et al.* 2014] ont construit un corpus MultiTD pour le DT collecté à partir de plusieurs ressources : le Web (les commentaires et les statuts sur Facebook, Twitter et TripAdvisor), les transcriptions de quelques enregistrements audio et d'autres ressources en DT. Ce corpus est composé de 32 848 mots. La deuxième étape est une étape d'extraction de mots. Elle est suivie par une étape de groupement de mots en utilisant un algorithme de k-modes qui vise à suggérer des organisations pour les mots en DT suivant des groupes significatifs. La dernière étape est la validation et l'enrichissement de TunDiaWN.

Une autre approche est proposée par [Ben moussa & Alimi 2015] pour construire un Wordnet pour le DT. La construction de ce Wordnet est basée sur deux ressources qui sont le dictionnaire ANG/DT « *Peace corps dictionary* » et le corpus de [McNeil & Faiza 2011]. [Ben moussa & Alimi 2015] ont réussi à créer un Wordnet sous le standard ISO-LMF composé uniquement de 8 455 lemmes.

2.4.1.4 Lexique

[Boujelbane 2015] a développé un lexique bilingue ASM/DT. Pour créer ce lexique, elle a adopté une méthodologie de transformation fondée sur les catégories grammaticales des mots du corpus ATB [Maamouri & Bies 2004].

À partir de ce corpus, [Boujelbane 2015] a construit des lemmes, des schèmes et des racines verbaux pour le DT. La construction est basée sur une traduction manuelle de l'ASM vers le DT. À l'issue de ce processus, [Boujelbane 2015] a construit un lexique composé de 1 500 verbes ayant comme entrée le couple (schème, racine) en ASM et son correspondant en DT. De même, [Boujelbane 2015] a construit manuellement une base de noms composée de 1 050 lemmes traduits en DT à partir de l'ATB. Elle a identifié pour ces lemmes ses racines et ses schèmes de dérivation. Le lexique résultat a comme entrée le triplet (lemme, schème et racine) en ASM et son correspondant en DT. Pour créer un lexique ASM/DT pour les mots-outils, [Boujelbane 2015] a proposé un ensemble de transformations basées sur l'étude des différents contextes des mots-outils de l'ASM afin de les traduire en DT.

2.4.2 Les outils pour le dialecte tunisien

2.4.2.1 Segmenteur de mots en DT

[McNeil 2012] a abordé la tâche de segmentation des mots en DT en proposant une méthode hybride combinant une analyse à base de règles et de mesures statistiques. La méthode développée par [McNeil 2012] repose sur plusieurs étapes. La première est une étape de normalisation des mots en DT qui consiste à extraire les mots à partir du corpus [McNeil 2012], filtrer les mots étrangers et translittérer les mots suivant le système Buckwalter de translitération modifié selon la prononciation du DT. Dans une seconde étape, [McNeil 2012]

identifie les affixes du DT qu'elle a utilisés pour implémenter un analyseur reposant sur des grammaires de la morphologie du dialecte. [McNeil 2012] propose pour chaque mot un ensemble d'analyses. L'analyse correcte est identifiée par une approche statistique : les analyses avec la plus grande fréquence dans le corpus d'apprentissage sont retenues. Pour choisir les analyses erronées, [McNeil 2012] a testé deux mesures : « Word-Root Frequency Ratio » de [Dasgupta & Ng 2007] et la mesure de « Suffix Level Similarity ». La première mesure n'a pas pu résoudre le problème, alors que la seconde mesure a permis d'améliorer les mesures d'évaluation.

L'évaluation du processus de segmentation a été faite sur un ensemble de 2 000 mots segmentés manuellement. L'évaluation de la première étape a apporté 0,45 comme rappel, 0,98 comme précision et 0,61 comme F-mesure et une exactitude de 0,66. L'ajout de la mesure « Suffix Level Similarity » a amélioré les résultats de 0,4 pour le rappel, de 0,30 pour la F-mesure et 0,22 pour l'exactitude.

2.4.2.2 Analyseur morphologique

[Hamdi 2015] ont adapté l'analyseur morphologique « MAGEAD » [Habash *et al.* 2005] afin de traiter le DT. Les changements concernent principalement la représentation des connaissances linguistiques. Ils ont modifié la hiérarchie des classes de comportement morphologique afin de traiter les patrons et les voyelles du DT. L'ordre des morphèmes abstraits a été modifié et de nouveaux morphèmes ont été ajoutés. Les auteurs ont également modifié les règles morpho-phonémiques pour le DT.

2.4.2.3 Étiqueteur morphosyntaxique

[Boujelbane 2015] a proposé d'utiliser le corpus généré à partir d'une traduction automatique de l'ATB de l'ASM [Boujelbane *et al.* 2014b] afin de refaire l'apprentissage de l'étiqueteur morphosyntaxique Stanford [Toutanvoa & Manning 2000] de l'ASM. Ce système offre une exactitude de 78,5 % sur un corpus issu d'une transcription des débats politiques pour le DT.

[Hamdi 2015] a exploité un ensemble de ressources pour l'ASM afin d'étiqueter morphosyntaxiquement le DT. Il a proposé une méthode composée de trois étapes.

La première propose de convertir une phrase en DT en un treillis en ASM. Le processus de conversion se déroule en trois sous-étapes dont la première est l'analyse morphologique avec l'analyseur adapté vers le DT « MAGEAD-DT » pour générer pour chaque mot plusieurs analyses. La racine et le patron proposés pour un mot donné sont ensuite traduits en leurs équivalents en ASM. La dernière étape est la génération des mots en ASM grâce à l'analyseur morphologique « MAGEAD-DT ».

Ce treillis passe, ensuite, par une étape de désambiguïsation qui permet de transformer les mots en pseudo-phrases en ASM. Cette étape identifie le meilleur chemin dans le treillis afin d'être analysé par l'étiqueteur morphosyntaxique qui permet d'étiqueter le treillis avec

la séquence des étiquettes les plus probables. [Hamdi *et al.* 2015] ont choisi les modèles de Markov caché (HMM) pour réaliser cette tâche.

Le tableau 2.10 présente un inventaire sur les différentes ressources développées pour le DT en montrant les différentes licences de distribution.

Auteurs	**Ressource**	**Taille**	**Licence de distribution**
[McNeil & Faiza 2011]	Corpus écrit	859 814 mots	Droit d'auteurs
[Boujelbane *et al.* 2014a]		517 080 mots	Libre
[Younes & Souissi 2014]		420 897 mots	Non mentionné
[Masmoudi *et al.* 2015]		870 904 mots	Non mentionné
[Graja *et al.* 2013]	Corpus oral	21 551 mots	Non mentionné
[Masmoudi *et al.* 2014]		71 684 mots	Non mentionné
[Boujelbane *et al.* 2014a]		37 964 mots	Libre
[Bouchlaghem *et al.* 2014]	Wordnet	32 848 mots	Non mentionné
[Ben moussa & Alimi 2015]		8 455 lemmes	Non mentionné
[Boujelbane 2015]	Lexique	2 550 entrées	Libre
[McNeil 2012]	Segmenteur de mots	-	Droit d'auteurs
[Hamdi 2015]	Analyseur morphologique	-	Licence non libre
[Boujelbane *et al.* 2014b]	Étiqueteur morphosyntaxique	-	Libre
[Hamdi 2015]		-	Licence non libre

TABLE 2.10 – Tableau récapitulatif des travaux réalisés pour le DT.

2.5 Conclusion

Dans ce chapitre, nous nous sommes intéressés au thème principal de notre thèse, à savoir, le traitement automatique du DT. D'abord, nous avons présenté le DT en exposant ses différentes caractéristiques aux niveaux phonologiques, morphologiques, syntaxiques et lexicaux. Nous avons, ensuite, exposé nos motivations. Enfin, nous avons donné un bref survol des principaux travaux réalisés pour la création des outils et des ressources en DT.

La deuxième et troisième partie de notre rapport sont consacrées à l'exposé des différentes ressources et outils que nous avons créés au profit du DT. Dans le chapitre 3, nous proposons deux nouvelles conventions de transcription orthographique pour le DT.

Deuxième partie

Création des ressources textuelles pour le dialecte tunisien

CHAPITRE 3

Transcription du dialecte tunisien parlé

Sommaire

3.1 Introduction

La transcription, au sens linguistique, est la représentation systématique du langage sous forme écrite [Moukrim 2010]. La transcription orthographique est un type parmi d'autres qui présente plusieurs défis pour une langue peu dotée ne possédant pas des règles orthographiques. Les dialectes arabes sont considérés comme des formes bâclées de l'ASM, essentiellement parlés sans règles orthographiques pour les écrire. Depuis quelques années, les chercheurs en TAL ont commencé à étudier les dialectes arabes. La tâche de transcription et la proposition des règles orthographiques ont été traitées par quelques-uns, mais, elle reste encore préliminaire.

Le DT comme étant un dialecte de la langue arabe appartenant au groupe des dialectes maghrébins possède des caractéristiques qui rendent son traitement complexe ainsi

que sa transcription. Dans la littérature, nous distinguons quelques efforts pour la transcription manuelle du tunisien ([Mejri & Baccouche 2003] ; [Ksouri 2013] ; [Mzoughi 2015] ; [Saidi 2014] ; [McNeil & Faiza 2011] ; etc.) afin de créer des corpus. La transcription n'était pas un objectif pour ces travaux. Nous remarquons une absence totale de tout effort de normalisation de la transcription.

Dans ce chapitre, nous proposons une méthode pour la transcription orthographique du DT. D'abord, nous présentons les difficultés liées à la transcription manuelle du DT. Ensuite, nous présentons nos deux conventions de transcriptions du DT. Enfin, nous terminons ce chapitre par une comparaison avec les autres travaux réalisés dans ce cadre.

3.2 Les difficultés de la transcription manuelle

La transcription consiste à remplacer chaque son et phonème d'un signal audio par un ensemble de graphèmes d'un système d'écriture [Moukrim 2010]. La tâche de transcription est une tâche pénible. Elle nécessite une série de décisions à prendre avant d'entamer la tâche. Le choix du mode de transcription (à posteriori ou à la volée) [Bazillon 2011], les conventions de transcription, la segmentation du signal, l'alignement au signal, les outils d'aide ou de transcription, etc. constituent autant de décisions à prendre avant la transcription elle-même [Moukrim 2010]. La tâche se complique, aussi, lorsqu'il s'agit de la transcription d'une langue principalement parlée qui ne dispose pas de standards orthographiques.

Dans ce cas, le transcripteur doit résoudre plusieurs types de difficultés : celles qui sont liées au caractère spontanée de la langue parlée (et qui sont indépendants de la langue à transcrire) et celle liées à l'absence de conventions orthographiques ainsi que la particularité de la langue cible.

3.2.1 Difficultés liées à l'oral

Les problèmes liés à la perception de la parole, à l'écoute et à l'oral sont les principaux obstacles que nous avons rencontrés lors de la transcription de notre corpus pour le DT. Ces problèmes sont fréquents. Ils sont communs à tous les corpus oraux et indépendamment de la langue qu'on souhaite transcrire.

La perception de la parole produit quelques obstacles lors de la transcription [Moukrim 2010]. En effet, le transcripteur, dans certains cas, n'arrive pas à décoder le signal produit. Dans d'autres cas, l'auditeur ne perçoit pas certains phonèmes dans une séquence bien déterminée ou il perçoit des phonèmes au lieu d'autres. Les erreurs de l'écoute sont souvent dues à l'auditeur qui essaie de donner sens à une séquence de mots [Moukrim 2010]. Généralement, les problèmes liés à la perception proviennent de plusieurs phénomènes existant dans l'oral comme le bruit, la qualité de l'enregistrement, la présence de plusieurs locuteurs, le chevauchement, etc.

Par ailleurs, la nature du discours influence sa transcription. [González Ledesma *et al.* 2004] ont prouvé que la complexité de la tâche de transcrip-

tion s'augmente quand il s'agit d'un discours où le degré de spontanéité est très élevé. Par ailleurs, [Bazillon 2011] a remarqué que la transcription d'un segment spontané nécessite un temps de transcription et segmentation plus élevé.

Les phénomènes spécifiques à l'oral comme les pauses remplies, les mots incomplets, les chevauchements, etc. causent également des problèmes de transcription. Il est très fréquent que le transcripteur néglige involontairement certains de ces phénomènes et les corrige en complétant par exemple un mot incomplet ou en corrigeant un élément afin de lui attribuer un sens. Ces actes involontaires réduisent la qualité des transcriptions.

En outre, les problèmes de transcription augmentent lorsque ces phénomènes de l'oral sont couplés avec le dialogue spontané où les locuteurs se croisent, hésitent, etc. Il est dans ce cas difficile d'identifier le locuteur et quand il parle.

3.2.2 Difficultés liées au dialecte tunisien

Le DT ne possède pas de standard orthographique pour écrire de façon unique ses mots. Les caractéristiques du DT (les différences avec l'ASM, la richesse de son lexique et les emprunts massifs d'autres langues étrangères) engendrent de nombreuses difficultés lors de la transcription du DT.

En effet, le DT se distingue par de nombreuses différences par rapport à l'ASM. Avec l'absence de conventions orthographiques, une solution peut être proposée : transcrire le DT suivant les règles orthographiques de l'ASM ou en utilisant des mots apparentés.

Prenons l'exemple de la phrase (1). La transcription de cette phrase en utilisant la forme orthographique de l'ASM donne la phrase (2) qui peut être considérée comme une phrase de l'ASM notamment si on omet les voyelles courtes.

1. « On tire un grand profit de la boisson de l'eau. »
2. ثمة فَائِدة كبيرة من شرب المَاء
   *θ*m*ħ* fAŷd*ħ* kbyr*ħ* mn šrb AlmA’

La transcription du DT suivant les règles de l'ASM donne deux problèmes. D'une part, on perd les traits majeurs du DT en traduisant le dialecte vers l'ASM. D'autre part, le DT est riche par des mots dialectaux qui n'ont pas d'équivalent en ASM. Leur transcription avec les règles de l'ASM est difficile voire impossible. En outre, la transcription de ces mots pose de nombreuses ambiguïtés. Prenons l'exemple du mot /barša/ dont trois formes orthographiques sont possibles : (برشَا, bršA), (برشة, brš*ħ*) et (برشه, brš*h*). Donc, choisir une forme orthographique est un défi à surmonter.

Par ailleurs, les locuteurs du DT utilisent dans leurs discours plusieurs mots de la langue française : « *déjà* » , « *donc* », « *mécanicien* », « *plombier* ». La plupart de ces mots se prononcent avec des allophones différents de ceux de l'ASM. On ne peut pas transcrire ces emprunts avec des alphabets arabes. En outre, la transcription de certains mots empruntés avec les alphabets arabes engendre quelquefois une ambiguïté sémantique. Le mot (مدَام, mdAm) par exemple

peut avoir deux traductions différentes : « *madame* » et « *tandis que* ». Par conséquence, le choix de l'alphabet (arabe ou latin) présente un autre défi de la transcription orthographique du DT.

Le DT se caractérise par la présence de mots avec des variantes morphologiques. On ne sait pas quelle variante doit être transcrite. Plusieurs entre elles peuvent être correctes simultanément (*e.g.* (قَالُوا, qaAluwA) et (قَالُو, qaAluw)). L'élision de certains phonèmes est un phénomène très fréquent au niveau du DT. Parfois, plusieurs phonèmes sont réduits. Les formes de négations (موش, mwš) et (مش, mš) et les verbes (قلتلك, qltlk) et (قتلك, qtlk) sont des exemples de mots ayant des formes réduites.

En guise de conclusion, la transcription de la langue parlée est confrontée généralement à plusieurs difficultés qui diffèrent d'une langue à une autre, mais, elles sont plus importantes quand il s'agit d'une langue essentiellement parlée sans règles orthographiques à savoir le DT.

Dans la section suivante, nous proposons deux conventions de transcription qui tentent à proposer des solutions à certains problèmes rencontrés lors de la transcription du DT, notamment l'absence de standards orthographiques pour le dialecte, tout en fournissant un ensemble d'annotations pour enrichir les transcriptions adaptées aux phénomènes de l'oral.

3.3 Deux Conventions de Transcription pour le Dialecte Tunisien

« Comment écrire ce mot ? », « Comment représenter ce phénomène ? », etc. sont autant de questions que pose un transcripteur notamment confronté au DT. L'absence de standards orthographiques pour le DT est parmi les principaux obstacles que le transcripteur rencontre lors de la transcription de l'oral en même temps que la représentation des phénomènes de l'oral.

Dans la littérature, peu de travaux ont proposé des conventions de transcription pour l'AD (*cf.* section 1.4.1 du chapitre 1). La plupart de ces travaux se sont focalisés sur le traitement des dialectes égyptiens et levantins. Sachant qu'il existe des différences majeures entres les dialectes arabes, l'application directe de ces conventions n'est pas possible. De ce fait, nous développons nos conventions de transcription et d'annotation pour le DT.

Le développement d'une convention de transcription pour le DT porte sur deux axes. Le premier se focalise sur la transcription orthographique du DT. Il s'agit de définir un ensemble de règles définissant l'orthographe des mots en dialectal. Dans ce cadre, nous avons défini deux conventions. La première se base sur une transcription orthographique basée sur la phonologie du dialecte en respectant certaines règles de l'ASM (orthographe reflétant la phonologie). Cette convention a été proposée dans le cadre de la convention d'annotation et de transcription « OTTA : Orthographic transcription of Tunisian Arabic » (*transcription orthographique de l'arabe tunisien*). La deuxième convention se base sur une transcription orthographique reposant essentiellement sur les règles orthographiques de l'ASM. Elle présente une extension de l'orthographe conventionnelle de l'AD « CODA : Conventional Orthographic Dialectal Arabic » [Habash *et al.* 2012a]. CODA est un projet qui vise à développer des conventions de transcription pour les dialectes arabes. La convention CODA a d'abord été proposée pour le dialecte

égyptien. Elle a été adapté ultérieurement pour le dialecte algérien et palestinien.

Le deuxième axe de développement d'une transcription conventionnelle est la proposition d'un ensemble d'annotations pour enrichir les transcriptions du DT pour refléter les phénomènes de l'oral. L'ensemble de ces annotations fait l'objet de la deuxième partie de la convention OTTA.

3.3.1 La convention OTTA « Orthographic Transcription of Tunisian Arabic »

Le manque de règles d'annotation et de transcription pour le DT nous a conduit à proposer, dans un premier temps, la convention OTTA [Zribi *et al.* 2013a] qui a été définie, comme une étape préliminaire pour la création de notre corpus pour le DT.

Cette convention est composée essentiellement de deux parties. La première est consacrée à la définition d'un ensemble de règles orthographiques pour la transcription du DT. La deuxième est destinée à la présentation des enrichissements et des annotations appliquées sur les formes orthographiques pour marquer les phénomènes de l'oral.

L'objectif de cette convention est d'assurer une forme orthographique pour le DT proche de l'orthographe de l'ASM (utilisation des lettres arabes, la segmentation des mots en ASM, l'utilisation du ة _ħ « Ta Marbuta »), etc. tout en respectant les spécificités phonologiques du DT telles que l'utilisation de nouveaux phonèmes, la suppression des voyelles courtes, etc.

Une telle méthode de transcription qui reflète la prononciation d'une langue permet l'amélioration les travaux de traitement automatique de la parole tels que les systèmes de synthèse de parole, les systèmes de transcription automatique, etc. Par la suite, nous obtenons un corpus pour le DT qui pourra être utile à la création des outils de traitement automatique du DT tels que les analyseurs morphologiques, les étiqueteurs morphosyntaxiques, etc.

L'objectif de notre thèse est le traitement automatique du DT. Il serait intéressant de créer une ressource qui marque bien les spécificités orales de la langue traitée (les pauses, les allongements, les disfluences, etc.). Nous définissons ainsi un ensemble d'annotations à respecter lors de la transcription d'un corpus pour le DT.

3.3.1.1 Les annotations utilisées lors de la transcription

La transcription d'un discours oral est une description fidèle de toutes ses caractéristiques. Il s'agit de fournir une présentation orthographique pour les mots et de marquer (ou annoter) les phénomènes qui ont été prononcés. Dans la littérature, plusieurs chercheurs ont défini des conventions de transcription et d'annotation de la langue parlée. Vu le partage de certaines caractéristiques de l'oral tunisien avec la langue française, comme la présence des caractères omis, l'existence des homophones, etc., nous choisissons d'adapter la convention « TOE : Transcription Orthographique Enrichie » proposée par [Bertrand *et al.* 2008]. Cette convention a été développée pour la transcription du corpus conversationnel de la langue française. L'existence de différences entre la langue française et la langue arabe, nous conduit cependant à ajouter des nouvelles annotations et supprimer les annotations qui ne peuvent

pas être appliquées pour le DT. Nous appelons cette convention d'annotation « TOE-TUN : Transcription Orthographique Enrichie du dialecte TUNisien ».

Le principe général de cette convention d'annotation est de transcrire ce qui est entendu. TOE-TUN définit des annotations pour marquer les élisions, les disfluences, les liaisons, les prononciations atypiques, les rires, etc. Les règles proposées par cette convention sont groupées selon le phénomène à annoter : les règles typographiques, les notations de prononciation, les notations pour les séquences incompréhensibles, les notations pour les événements non linguistiques et les pauses et les notations pour le discours rapporté.

Les règles typographiques. Les règles typographiques définissent la présentation typographique des mots en DT en ajoutant dans certains cas des annotations et des symboles. Elles élucident des abréviations, les titres, les acronymes, les patronymes et les toponymes.

Règle n°1 : Pour respecter le principe général de la convention orthographique TOE-TUN, aucune abréviation ne doit être utilisée lors de la transcription. Tous les mots seront transcrits à la lettre en respectant les règles orthographiques de la convention OTTA et la prononciation des mots. Étant donné que le principe de la méthode est d'avoir une transcription qui reflète la prononciation des mots, nous n'utilisons donc pas les acronymes dans les transcriptions. Ceci dit, nous penchons sur la transcription des mots tels qu'ils sont prononcés. Nous n'utilisons ni les signes d'exclamation, ni les signes d'interrogation. La transcription des nombres aussi respecte le même principe. Elle doit être à la lettre.

Règle n°2 : Les titres de livres, de films, les journaux, les émissions radiophoniques et télévisées, etc. sont écrits entre deux guillemets. Lors de la transcription, les entités nommées sont, aussi, identifiées et encadrées par deux crochets, en identifiant son type. Nous utilisons le code « شخ-مك-شع ». Les symboles شخ, مك et شع sont utilisés pour désigner respectivement les patronymes, les toponymes et les acronymes. On utilise cette annotation : *[*شخ-مك-شع*, la forme orthographique]*.

Règle n°3 : Les lettres qui sont prononcées séparément ne se transcrivent pas par des lettres séparées. On transcrit la prononciation des lettres comme des mots.

Règle n°4 : Nous définissons une liste des onomatopées fréquemment utilisées par les locuteurs du DT qui sont transcrits en ajoutant le symbole « ÷ » à gauche. Même si la prononciation se diffère un peu de cette liste, on doit bien respecter l'orthographe utilisée dans cette dernière : ÷أَي, ÷بَاهِ, ÷أَوه, ÷إِم, ÷بون, ÷آك, ÷اه, ÷بَاه, ÷هأَه, ÷إِيوَاه, etc.

Règle n°5 : L'utilisation massive des mots empruntés des autres langues (notamment le français) est une caractéristique du DT. Ces mots ne doivent pas être négligés dans les transcriptions de notre corpus oral puisqu'ils sont porteurs d'informations. Ils sont écrits en alphabets latins et sont encadrés par deux crochets en précisant la langue et la forme orthographique du mot dans sa langue d'origine *[lan :langue, orthographe]*.

Règle n°6 : Comme on l'a déjà présenté dans la section 2.2.2 du chapitre 2, il existe des mots avec des variantes morphologiques qui sont transcrites entre accolades.

Les notations de prononciation.

Règle n°7 : Les locuteurs du DT se caractérisent par une rapidité lors de la prononciation des mots. Ce phénomène génère parfois la non-prononciation de certains phonèmes. Par exemple, le mot (موش, muwš) « *pas du tout* » est souvent prononcé en (مش, /muš/) en réduisant l'allongement vocalique issu de la voyelle longue (و, w). De ce fait, nous proposons de transcrire les mots selon leurs formes bien prononcées mais on encadre la ou les lettres non prononcées par deux parenthèses. On doit transcrire (م(و)ش, m(w)š) au lieu de (مش, muš), de même pour (عمتلك, ʕmtlk, « *j'ai fait pour toi* ») qui doit être transcrit (عم(ل)ت لك, ʕm(l)t lk).

Règle n°8 : Les cas d'accords atypiques sont transcrits comme ils sont prononcés. On ne corrige pas ces erreurs. Même, le transcripteur ne corrige pas celle de prononciation du locuteur. Par exemple, si le locuteur replace une lettre par une autre incorrecte, le transcripteur ne corrige pas cette erreur. Mais, il utilise cette annotation pour montrer le ou les caractères corrects : *xx{lettre correcte, lettre prononcée}xxx*.

Règle n°9 : Toute langue parlée comporte des phénomènes spécifiques, et notamment les disfluences. Il s'agit d'un phénomène apparaissant dans toute production orale spontanée ; ils consistent à l'interruption du cours normal du discours [Heeman & Allen 1994]. C'est pourquoi, nous proposons d'améliorer les annotations utilisées pour marquer ces faits. En ce qui concerne l'annotation des mots incomplets, nous proposons d'utiliser le petit tiret « - » pour marquer les mots non complets et non pas le signe d'allongement de l'arabe (*e.g.* عس- et non pas عسـ). Par exemple, le mot (عس-, ʕs-) est un mot tronqué du mot (عسلَامة, ʕslAmħ) « *Bonjour* ».

Règle n°10 : Le DT se distingue par la présence d'un phonème de liaison entre deux mots. Par exemple, le phonème /n/ est utilisé pour lier entre un numéral et le nom suivant. Nous utilisons deux symboles « = » entre les deux mots de liaison.

Règle n°11 : Le DT utilise trois phonèmes (/v/, /g/ et /p/) qui ne correspondent pas à des consonnes arabes. Ces dernières peuvent être présentés par ces alphabets (ڤ, pour le /v/, ڤ pour le /g/ et پ pour le /p/). Pour faciliter la transcription de ces consonnes en DT, on peut ajouter le caractère « ´ » après les caractères ف, ق et ب.

Les séquences incompréhensibles.

Règle n°12 : Nous utilisons « * » pour marquer une séquence incompréhensible. Au cas où, le discours serait peu accessible, mais pas quasiment incompréhensible, nous écrivons la meilleure séquence que nous comprenions entre doubles parenthèses *((séquence incompréhensive))*.

Les événements non linguistiques.

Règle n°13 : Les rires sont notés avec le symbole « && ». Quand les paroles et les rires se chevauchent, on encadre les paroles par un double symbole de « && ». Au niveau de l'exemple (3), l'intervenant rit puis, il dit son discours.

Règle n°14 : Les pauses courtes d'une durée inférieure à 200 ms sont marquées par le symbole « + ».

Règle n°15 : On utilise les étiquettes suivantes pour marquer le bruit (Nous distinguons 8 types de bruits).

— (تنفس) pour marquer la respiration du locuteur lors de la parole.
— (ضجيج) pour marquer le bruit extérieur.
— (نفح) pour marquer le halètement.
— (فم) pour marquer les bruits effectués par la bouche.
— (سعَال) pour marquer la toux.
— (عطَاس) pour marquer l'éternuement.
— (صفير) pour marquer le sifflement.
— (موسيقَى) pour marquer la musique.

Discours reporté. Le discours reporté est noté entre deux « \ ». Ce caractère est précédé et suivi par un blanc.

La phrase 3 est un exemple extrait de notre corpus qui est enrichi avec les différentes annotations proposées.

3. (تنفس) "برنَامج المجلة الصحية" في [منصف الشلي, شخ] دكتور - د مرحبَا بيك + عَالسلَامة &&
÷مة آ{ف ,ث} اللّي \ يقول خبر اللّي على تعلق لنَا بَاش ((اليوم)) * [lan :FR, donc]
الشكَلَاط\ [lan :MSA, جرَاء] من [قَابس, مك] في {تسمّموا, تسمّمو} طفل = ن=خمسطَاش
« *Bonjour et bienvenue d- docteur Moncef Chelly dans notre émission "la magasine de la santé" (respiration) donc * ((aujourd'hui)) tu nous parleras de l'information qui précise que \ quinze enfants de Gabès ont été empoisonnés par le chocolat* \. »

3.3.1.2 Les règles orthographiques pour transcrire le dialecte tunisien

Dans le cadre de la convention OTTA, nous proposons un ensemble de règles orthographiques pour schématiser le DT. Ces règles respectent d'une part les caractéristiques de la langue arabe et d'autre part schématisent les différences morphologiques, phonologiques et syntaxiques du DT.

L'étude lexicographique du DT nous a montré trois principales classes de mots : des mots de l'ASM avec des différences phonologiques, des mots dialectaux et des mots empruntés d'autres langues. En fait, la transcription orthographique des mots appartenant à la première ensemble de cette classification pourra être très proche de celle de l'ASM vu que les différences

résident uniquement au niveau du système vocalique. Dans ce cas, on garde les règles orthographiques de l'ASM pour les transcrire. Les mots dialectaux pourraient être, aussi, écrits en suivant les règles de transcription de l'ASM. En revanche, ces règles ne seront pas appliquées dans le cas de présence des différences phonologiques et morphologiques telles que l'utilisation des phonèmes non arabes tels que /v/, /g/ et /p/ et l'utilisation de nouveaux affixes tels que (ع, ى) et (م, m). Par conséquent, on a besoin de définir des règles de transcription pour les mots qui sont spécifiques au dialecte. L'écriture des mots empruntés présente une des problèmes de la tâche de transcription du DT. Nous avons proposé au niveau de la convention TOE-TUN de la section précédente une règle pour résoudre cette difficulté.

Nous présentons dans ce qui suit les principales règles de transcription de l'ASM que nous gardons lors de la transcription du dialecte et nous définissons un ensemble de règles de transcription basées sur les spécificités phonologiques et morphologiques du DT.

Les règles orthographiques basées sur l'ASM.

Alphabet arabe. Notre méthode de transcription garde les principes de transcription de l'ASM. Nous transcrivons le DT avec les caractères arabes en utilisant les voyelles courtes. (*e.g.* شرِب šrib « *il a bu* » et شَرَّب šar˜ab « *il a fait boire* »).

Segmentation des mots. Nous appliquons les règles de segmentation des mots de l'ASM dans certains cas. En effet, nous transcrivons certains affixes comme des mots. Prenons l'exemple de la conjonction d'interrogation (آش, Āš, « *quoi* ») qui est parfois réduite à une seule lettre (ش š) concaténée au mot suivant. Cette conjonction remplace la conjonction d'interrogation (مَاذَا mAðA « *quoi* ») en ASM. Afin de se rapprocher de la structure des phrases en ASM, on a choisi de transcrire cette conjonction comme un mot séparé avec sa forme étendue et non pas sa forme réduite. En outre, nous appliquons la segmentation des mots dans le cas de la négation et dans le cas de l'utilisation du groupe prépositionnel d'objet. Par exemple, le mot (/qultlwu :/, « *je lui ai dit* ») doit être transcrit en deux mots séparés par des espaces : قلت لو qult luw et non pas قلتلو qltlw. Cette segmentation est justifiée par le fait que le groupe prépositionnel d'objet (لو lw « *à lui* ») ne doit pas être agglutiné au verbe, pareillement, pour la conjonction de négation (مَا maA) (*e.g.* مَا قلتش mA qltš « *je n'ai pas dit* » et non pas مَاقلتش mAqltš ou مقلتش mqltš).

Agglutination. En DT, la préposition (علَى ىlỳ « *sur* ») est transformée en une seule lettre (ع ى). Puisqu'en ASM, la lettre doit être toujours agglutinée au mot qui la suit, cette préposition est transcrite comme un préfixe (*e.g.* عَالطَاولة ىAlTAwlħ « *sur la table* »). Ce principe s'applique pour la conjonction de coordination (و w « *et* ») et les clitiques (ف f « *dans* »), (م m « *de* ») et (ك k « *comme* »).

Les règles orthographiques basées sur la phonologie du DT.

Le pronom personnel ه /w/. L'étude linguistique menée sur le DT (cf. au chapitre 2) montre que le dialecte connaît plusieurs différences par rapport à l'ASM. Parmi lesquelles, citons l'introduction du (و, /w/) comme une nouvelle prononciation du suffixe pronominal de l'ASM (ه/ـه, /h/, « *son* ») lorsqu'il est agglutiné aux mots qui se terminent par une consonne. En revanche, le pronom (ه /ـه, /h/, « *son* ») est prononcé correctement lorsqu'il est collé aux mots qui se terminent par l'une des voyelles longues. En conséquence, nous proposons d'utiliser la forme (و, w) du pronom (ه/ ـه, /h/, « *son* ») dans nos transcriptions. Par exemple, le mot (/Ta :wiltu :/, « *sa table* ») qui est transcrit selon la convention de [Maamouri *et al.* 2004b] en (طَاولته, TAwlth) avec la transformation du (و, w, « *son* ») en (ه, hu), sera écrit avec nos règles en (طَاولتو, TaAwiltuw).

La lettre Hamza. La prononciation de la lettre Hamza présente, aussi, une autre différence par rapport à l'ASM. La lettre Hamza est, parfois, remplacée par l'une de ces voyelles (و, ي ou ا). Par exemple, on remplace le phonème /'i/ du mot (فَائِدة, fAŷdħ, « *le profit* ») par le phonème /y/ (فَايدة, fAydħ, « *le profit* »).

Si la lettre Hamza est située au début d'un nom défini, c'est-à-dire, après les lettres (ال, Al), alors elle est remplacée par le phonème /l/. On dit (لولَى, lwlỳ, « *la première* ») au lieu de (الأولَى, AlÂwlỳ, « *la première* »). Comme nous envisageons d'avoir des transcriptions qui reflètent la prononciation des mots et vu la difficulté de prédiction de la forme d'origine du mot en ASM, nous suggérons de transcrire la consonne Hamza uniquement lorsqu'elle est prononcée.

Nous gardons uniquement la transcription de la lettre Hamza lorsqu'elle est située à la fin des mots même si elle n'est pas prononcée. Nous justifions notre choix par l'ambiguïté qui peut apparaître à cause de la non-transcription de Hamza à la fin du mot. Par exemple, le mot (مَا, mA) peut désigner une conjonction de négation ou un nom « *l'eau* » qui est écrit en ASM avec (مَاء, mA').

Transcription des mots spécifiques au DT. Le lexique du DT connaît évidemment la présence de mots qui n'ont pas de racines dans la langue arabe : des mots spécifiques au DT. La transcription de ces mots connaît des différences d'un transcripteur à un autre. Afin d'avoir un corpus homogène avec des mots transcrits d'une manière unique, nous proposons définir des règles qui consistent à combiner les règles de l'ASM et les compositions phonémiques.

— Transcription de /v/, /p/ et /g/.

Le DT connaît l'ajout de nouveaux phonèmes tel que le /g/. L'utilisation des lettres arabes pour transcrire ces phonèmes engendre des ambiguïtés ainsi que le changement du sens du mot. Par conséquent, nous proposons d'utiliser les lettres persiennes (ڤ, /v/), (ڤ, /g/) et (پ, /p/) pour transcrire les mots connaissant l'utilisation de ces nouveaux phonèmes. On n'utilise

pas les lettres (ف, ق et ب). On transcrit (ڤرُون /gru :n/) pour « *les cornes* » et (قرُون /qru :n/) pour « *les siècles* »).

— Le phonème /a/ à la fin d'un mot dialectal.

En DT, plusieurs mots dialectaux se termine par le phonème /a/. On trouve souvent plusieurs formes orthographiques pour ces mots. Par exemple, le mot « *beaucoup* » en DT peut avoir au moins trois formes orthographiques possibles : (برشة, brš*ħ*), (برشه, brš**h**) et (برشَا, brš**A**). Pour uniformiser la transcription de ces mots, nous proposons cette règle : si le dernier phonème d'un mot est équivalent à /a/, alors le mot doit se terminer avec la lettre muette (ة, _*ħ*). On applique cette règle uniquement pour les noms. (*e.g.* كرهبة krhb*ħ* « *voiture* », برشة brš*ħ* « *beaucoup* »).

— Allongement vocalique à la fin d'un mot dialectal.

Si le dernier phonème d'un mot se caractérise par un allongement vocalique (/a :/, /u :/ ou /i :/) alors le mot doit s'écrire avec une de ces voyelles (ا, A), (ي, y) ou (و, w) (*e.g.* ياخي yAxy راهو rAhw). [1]

3.3.2 La convention orthographique du dialecte tunisien CODA « A Conventional Orthography for Tunisian Arabic »

En 2012, [Habash *et al.* 2012a] ont proposé une convention pour la transcription orthographique de l'AD. L'objectif de cette convention est d'alléger les difficultés que cause l'absence de standards orthographiques lors du développement d'applications de TAL. La conception de CODA est basée sur cinq objectifs [Habash *et al.* 2012a]. Tout d'abord, CODA est une convention cohérente et consistante pour l'écriture de l'AD. Elle est créée en faveur du développement d'applications de TAL en utilisant les alphabets arabes. CODA est conçue comme un cadre unifié pour l'écriture des dialectes arabes. Enfin, elle vise à trouver un certain équilibre en gardant les caractéristiques dialectales et en établissant des conventions sur la base des similitudes AD/ASM.

L'étude des principes de la convention CODA ainsi que la version développée en faveur du dialecte égyptien, nous ont motivés pour étendre cette convention pour notre dialecte. Les objectifs de la convention CODA convergent avec nos besoins pour la conception d'une orthographe conventionnelle pour le DT. En effet, notre objectif pour définir une orthographe pour le DT consiste, d'une part, à avoir une façon unique pour la transcription des mots en DT afin de créer un ensemble de ressources textuelles ainsi que de développer des outils de TAL pour le DT. D'autre part, étant donné que notre travail pour la proposition d'une convention orthographique est le premier travail développé pour le DT, nous envisageons de développer une convention orthographique standard et réutilisable par les chercheurs en TAL du DT.

En outre, notre orthographe conventionnelle développée dans le cadre de l'OTTA partage une grande partie des règles avec la convention CODA notamment nos règles orthographiques à base d'ASM. Par conséquent, le développement d'une convention orthographique du DT en

1. On n'utilise pas le simple Alif (ا, A) après le (و, w). On n'écrit pas (راهوا, rAhwA).

suivant les objectifs et les principes de travail de la CODA [Habash *et al.* 2012a] ne nécessitent qu'une simple amélioration des règles de transcription d'OTTA et l'ajout d'autres précisions et modifications.

Dans cette section, nous présentons le « CODA-TUN », qui est une extension de la CODA, développé pour couvrir le DT. Cette extension est le résultat d'une collaboration avec l'auteur de la CODA, Nizar Habash [2], professeur associé au département d'informatique à la faculté de New York à Abu-Dhabi et un ensemble de chercheurs de notre groupe de recherche ANLP group [3].

L'élaboration de la convention CODA s'est basée sur une stratégie qui détecte les différences phonologiques, morphologiques, phono-lexicales et lexicales avec l'ASM pour définir des règles correspondant à ces particularités. Nous choisissons, ainsi, à garder cet ordre pour présenter les conventions orthographiques de CODA-TUN. Rappelons que nous définissons des règles orthographiques pour le dialecte utilisé par les médias car c'est la forme dialectale la plus comprise par la majorité des tunisiens.

3.3.2.1 Les consonnes et les voyelles en DT

Notre convention CODA-TUN se base sur les consonnes et les voyelles de la langue arabe à l'exception des diphtongues /ay/ et /aw/. La voyelle longue /e :/ issue des emprunts d'autres langues est largement utilisée en DT. Nous représentons cette voyelle avec la voyelle courte (ِ, i) suivie par un Alif (ا, A).

Transcription du Šadda (˜). La langue arabe se caractérise par la présence du symbole de la šadda (˜) qui remplace la deuxième lettre répétée dans une séquence de lettres. Nous gardons, au niveau de la convention CODA-TUN, cette règle pour la transcription des lettres répétées.

La lettre Hamza. Le graphème de Hamza connaît plusieurs formes selon sa position dans le mot et la voyelle courte qui le précède. En CODA-TUN, nous gardons les mêmes règles orthographiques de l'ASM pour choisir la bonne forme. Cette règle ne s'applique qu'en cas de prononciation de Hamza.

La transcription des mots de l'ASM contenant une Hamza qui est omise en DT suit quatre règles :

(1) Hamza de l'ASM se transforme en (ي, y) si elle est précédée par la voyelle courte (i, ِ).

(2) Hamza de l'ASM se transforme en simple Alif si elle est précédée par la voyelle courte (َا, a).

(3) Hamza de l'ASM se transforme en (و, w) si elle est précédée par la voyelle courte (ُا, u).

(4) Hamza de l'ASM située à la position finale du mot est omise dans le DT.

2. www.nizarhabash.com
3. https ://sites.google.com/site/anlprg/

(5) Hamza de l'ASM située à la position finale du mot est transformée en une allongement vocalique /a :/.

Le tableau 3.1 présente la transcription de quelques mots avec la lettre Hamza et leurs équivalents en DT.

Règle	Mot en ASM	Mot équivalent en DT	Traduction en français
(1)	الدقَائِق AldaqaAŷiqu	دقَايق dqaAyiq	Les minutes
	ذئِب ðŷb	ذيب ðyb	Un loup
	قَائِد qAŷid	قَايد qaAyid	Un commandant
	فِئْرَان fiŷrAn	فيرَان fiyraAn	Des souris
(2)	كَأْس kÂs	كَاس kaAs	Un verre
	فَأْر fâr	فَار faAr	Une souris
	مرَأَة mrÂħ	مرَا mraA	Une femme
(3)	رُؤُوس ruŵuws	روّوس ruwws	Des têtes
	مُؤْمِنون muŵminwn	مومنين muwmniyn	Des croyants
(4)	هوَاء hawaA’	هوَا hwaA	L’air
	سَمَاء samaA’	سَمَا smaA	Le ciel
(5)	بَدَأَ badÂ	بدَا bdaA	il a commencé

TABLE 3.1 – Transcription de quelques exemples de mots contenant la lettre Hamza.

Notons qu'il existe un seul cas d'exception où la lettre Hamza est précédée par le sukūn. Dans ce cas, le lettre Hamza est prononcée en DT. Par exemple, les mots (سَأَل, saÂala, « *il a posé une question* ») et (أَسئِلة, Âsŷlħ, « *des questions* ») de l'ASM gardent leurs formes orthographiques en les transcrivant en DT.

En cas de transformation de la lettre Hamza en une allongement vocalique /a :/, CODATUN propose de transformer la lettre Hamza (أَ, Â) de l'ASM en une simple Alif (ا, A) et non pas une Alif Maqsura (ى, ŷ). Par exemple, le verbe /yibda/ « *il commence* » est transcrit comme (يِبدَا, yibdaA) et non pas (يِبدَى, yibdaY).

— Agglutination du Hamza.

En DT, nous remarquons qu'au niveau de certains mots qui débutent avec Hamza agglutinée aux clitiques (و, w) ou (ب, b) et l'article défini (ال, AL), le phonème correspondant à Hamza est omis. Prenons l'exemple des mots (أَحلَام, /ÂHle :m/) et (أُستَاذ, /Âusta :ð/) se prononcent respectivement comme /waHle :m/ et /wAsta :ð/ en cas d'agglutination au clitique (و, w). Dans ce cas, nous utilisons Hamza Wasl (une simple Alif avec une voyelle courte) pour représenter ces cas d'omission du phonème correspondant à Hamza. D'où, les mots /’aHle :m/ et /’usta :ð/ seront transcrits en (اَحلَام, AaHliAm) et (اُستَاذ, AustaAð) et non pas (أَحلَام, ÂaHlaAm) et (أَستَاذ, ÂustAað). Ainsi, les mots qui commencent par Hamza Wasl s'écrivent normalement en suivant les mêmes règles de transcription de l'ASM. Par exemple, les mots /bilHaq/ et /wilfikr/ s'écrivent respectivement (بَالحق, bAlHq) et (وَالفِكر, wiAlfikr) non pas (بِأَلحق, bil’alHaq) et (وِأَلفِكر, wi’ilfikr).

Transcription du simple Alif. En DT, le phonème /i/ qui correspond à la prononciation de la lettre Alif avec la voyelle courte (إِ, i), s'ajoute au début des mots, plus précisément aux verbes en forme passive (*e.g.*, /itqtil/, /itDlm/, /itkas˜ar/). En CODA-TUN, nous ne transcrivons pas le simple Alif. Donc, les verbes (/itqtil/, /tDlm/ et /itkas˜ar/) doivent se transcrire comme suit : (تقتل, tqtil), (تظلم, tDlm) et (تكسر, tkas˜ar).

Notons que cette règle ne s'applique pas aux mots qui suivent la même prononciation de l'ASM. Par exemple, le verbe (استَأْذن, AstÂðn, « *il demande la permission* ») ne se transcrit pas comme (ستَأْذن, stÂðn). En outre, les mots de l'ASM commençant par une Hamza Wasl qui est omise en DT dû aux changements de patrons, présentent aussi une exception à cette règle. Ils s'écrivent sans Hamza. Par exemple, le mot /bin/ « *le fils de* » se transcrit comme (بن, bin) et non pas (ابن, iAbin).

Transcription du Ta Marbouta ة. Le Ta Marbouta est prononcé /ap/, /ip/, /t/ ou /it/ selon sa position dans le mot. Mais, nous le transcrivons toujours dans sa forme (ة, *ħ*). Prenons l'exemple du mot « *voiture* » qui se prononce en DT comme /karhba/, /karhabit/, /karhabtu/ et /karhabitha :/ lorsqu'il est seul, suivi par un nom ou agglutiné à un enclitique : (كَرهبَة, karhba*ħ*), (كَرهَبِة أَحمد, karhabi*ħ* ÂHmd), (كَرهَبتُه, karhabth) et (كَرهَبتِهَا, karhabithA).

3.3.2.2 Les exceptions phonologiques

Le DT connaît plusieurs différences phonologiques en le comparant avec l'ASM. D'abord, la phonologie des voyelles du DT (courtes et longues) se caractérise par quelques particularités :

- Les voyelles longues situées à la position finale des mots sont réduites.
- Plusieurs voyelles longues sont permises (*e.g.*. مَاعون, « *ustensiles* »).
- Les voyelles non accentuées [4] ne peuvent pas être longues.
- L'ajout des affixes et des clitiques change les patrons accentués et interagit avec la longueur des voyelles.
- Les phonèmes des voyelles longues ont des allophones courtes mais, les phonèmes des voyelles courtes n'ont pas des longues allophones.
- Les voyelles omises sont considérées comme faisant partie de la forme phonémique du mot.

La règle générale de CODA-TUN est d'écrire les mots comme ils sont prononcés, sauf s'il y a des exceptions phonologiques, morphologiques ou lexicales.

Nous préservons les voyelles longues dans les mots en DT, même si elles se raccourcissent dans des contextes différents. Prenons l'exemple du mot (قلتلك, qltlk, « *je t'ai dit* »), il se prononce /qutlik/. Au niveau de cette prononciation, nous remarquons plusieurs phénomènes : l'omission du phonème correspondant à la lettre (ل, l) et la fusion des syllabes de deux com-

4. Les voyelles non accentuées sont des voyelles dans les mots avec plus d'une syllabe qui se trouve dans la partie non accentuée du mot.

posants : le verbe et le groupe prépositionnel d'objet (لك, lk). Selon CODA-TUN, ce syntagme doit être transcrit en deux mots séparés par un espace (قلت لك, qlt lk). Le même principe s'applique sur le syntagme مَا غير mA γyr /maγir/ « *sans* » avec un espace entre les deux mots.

En outre, nous faisons une distinction entre les deux voyelles longues /a :/ et /e :/. Pour se faire, nous ajoutons la voyelle courte (a) et la voyelle (i) respectivement après le simple Alif (ا, A). Par exemple, le mot حرام est transcrit comme (حرَام HraAm /Hra :m/) lorsqu'il se traduit par « *péché* » et il est transcrit comme (حرِام HriAm /Hre :m/) lorsqu'il désigne « *une couverture* ».

En ce qui concerne les mots étrangers et les noms des lieux, le CODA propose de les transcrire en suivant leur orthographe proposée en ASM. Cependant, s'il existe une orthographe régionale, nous suivons la forme proposée en dialecte. Par exemple, le mot « *garage* » est écrit en égyptien (جرَاج, jrAj), en levantin (غرَاج, γrAj) et en tunisien (قرَاج, qrAj).

Pour les phonèmes non arabes /g/, /p/ et /v/, nous utilisons respectivement les lettres (ق, q), (ب, b) et (ف, f). Par exemple, les mots /ga :tu :/ « *gâteau* », /pa :pa :/ « *papa* » et /mgariv/ « *en grève* » se transcrivent respectivement comme قَاتو, بَابَا et مقَرِف.

3.3.2.3 Les exceptions phono-lexicales

Comme déjà présenté dans la section 2.2.2.1 du chapitre 2, le DT connaît plusieurs consonnes avec une double prononciations qui sont parfois différentes de leurs équivalents en ASM. Par exemple, les lettres (س, s) et (ص, S) peuvent être prononcées /s/ ou /S/.

Le tableau 3.2 montre la liste des consonnes avec double prononciation en DT.

L'orthographie de la CODA-TUN	Les variantes de prononciation dialectale	Exemples de mots dialectaux écrits en suivant CODA-TUN
س	/s/ ou /S/	/ru :S/ رُوّوس « *des têtes* » /raSu :l/ رَسُول « *un prophète* »
ص	/s/ ou /S/	/saAyγi :/ صَايغِي « *un bijoutier* » /saba :t/ صَبَّاط « *une chaussure* »
ج	/j/ ou /z/	/zaza :r/ جَزَّار « *un boucher* » /zarzi :s/ جَرجِيس « *Zarzis* »
س	/s/ ou /z/	/fuzdaq/ فُسْتَق « *une pistache* »
ص	/S/ ou /z/	/zda :q/ صدَاق « contrat de mariage » /zafar/ صَفَر « *il a sifflé* »
ط	/T/ ou /t/	/tay~a :ra/ طَيَّارَة « *un avion* »
ث	/v/ ou /f/	/fam~a/ ثَمّة « *il y a* »
ع	/ʕ/ ou /H/	/mta :Hha/ متَّاعهَا « *c'est-à-dire* »
غ	/γ/ ou /x/	/xasa :la/ غَسَالَة « *un lave linge* » /xaslit/ غَسلِت « *elle a lavé* »
أ	/Â/ ou /h/	/shal/ سَأَل « *il a posé une question* »
ق	/q/ ou /G/	/baGra/ بَقرَة « *une vache* »

TABLE 3.2 – Les consonnes avec double prononciations.

Le phonème /g/ est largement utilisé dans les mots du DT. Ce phonème correspond à la consonne (ق, q) si la racine a une origine commune avec l'ASM. Par exemple, le mot (بقرة, bqrħ, « *vache* ») se prononce en DT /bagra/. De même, il existe plusieurs mots qui contiennent le phonème /G/. Par exemple, /bilgda :/ n'a pas d'équivalent en ASM. En CODA-TUN, nous utilisons la lettre (ق, q) pour présenter le phonème /G/. Donc, le mot /bilgda :/ est transcrit en (بَالقدَا, biAliqdA, « *très bien* »).

Par ailleurs, on utilise la forme apparentée de l'ASM pour transcrire la lettre (ض, D) lorsqu'elle est prononcée /D/ or /Ď/. Par conséquent, /Ďa :biT/ est transcrit en (ضَابِط, DaAbiT, « *officier de police* ») non pas (ĎaAbiT, ظَابِط). Cette règle s'applique aussi aux mots dérivés à partir du patron ($Air_1tar_2ar_3$). Par exemple, les deux mots /nafTariD/ et /iSTaETa/ se transcrivent respectivement en نَفتَرِض naftariD et استعطَى AstʕTỳ.

Le DT, comme d'autres dialectes arabes, se caractérise par l'ajout du phonème /n/ à la position finale du numéral, dans le cas d'un groupe nominal. Par contre, ce phonème est absent dans le cas où le numéral est prononcé seul. Nous proposons dans CODA-TUN de conserver le phonème /n/ dans le cas où il est prononcé (*e.g.*, /xmasta :šin/ (xmsTAšn Âlf) (خمسطَاشن أَلف) vs. (xmsTAš) (خمسطَاش)).

3.3.2.4 Les exceptions morphologiques

Le DT partage avec l'ASM plusieurs caractéristiques. Tout d'abord, tous les affixes et clitiques sont ajoutés au mot sans changer son orthographe en suivant le même principe de l'ASM (*cf.* exemple(4)). Ainsi, on n'ajoute pas le šadda en cas de doublement des lettres lors de l'ajout des clitiques à l'exception de l'ajout de (ي, ya) (*cf.* exemple(5)). De même, nous conservons la règle de segmentation des mots qui exige la séparation de la particule de négation مَا mA et les clitiques d'objet indirect (*e.g.* لِيش liyš) du verbe (*cf.* exemple (6)).

4. /bilmaka :tib/ => /bi+Al+mkAtib/ => (biAlmkAtib) (بَالمكَاتِب)
5. (saman+ny) (سمّنني), (šabh+hA) (شبهّها) et (fiy+˜a) (فِيَّ)
6. /wmaqalli :š/ => (wmA qAl liyš) (ومَا قَال لِيش) non pas (wmaqal˜iyš) (ومَقَلِّيش)

Au niveau de CODA-TUN, nous proposons de transcrire tous les affixes et clitiques en leurs formes phonétiques allomorphiques. Au niveau du DT, nous notons deux exceptions à cette règle. La première concerne la transcription des affixes de la troisième personne du pluriel. Nous ajoutons le simple Alif (ا) à la position finale du verbe. Par exemple, le suffixe /u :/ est transcrit comme suit (وا, uwA). La deuxième touche la transcription de l'article défini (ال, Al). Ainsi, il est écrit en suivant sa forme morphémique, c'est-à-dire, nous appliquons dans ce cas les règles de l'ASM en cas de présence des lettres solaires et les lettres lunaires (الشّمس Alš˜ams /iššams/ et non pas الشمس Ailšams ou اشّمس Aiš˜ams).

En CODA-TUN, l'orthographe des clitiques suit trois règles générales :

— Toutes les particules composées d'une seule lettre sont attachées au mot suivant. Par contre, les particules ou les clitiques composées de plusieurs lettres sont écrits séparément. La seule exception à ce cas est l'article défini (+ال, Al+).
— Les enclitiques pronominaux, la particule de négation (+ش ŝ+) et la particule d'interrogation (+شي ŝy+) sont toujours localisées à la position finale du mot. De même, nous exigeons dans CODA-TUN de ne pas attacher le groupe prépositionnel (l+pronom) au mot qui le précède.
— On ne change pas la base du mot lorsqu'on ajoute des clitiques au mot même si la phonologie du lemme ou du clitique change à l'exception de :
  - L'ajout du clitique (ال+ل+Al) qui se transforme en (للّ ll) (*e.g.* ل+البيت => للّبيت).
  - Le Ta Marbouta ة se transforme en (ت, t) (*e.g.* معلمة+هم => معلمتهم).
  - Le (وا, wA) du pluriel est supprimé en ajoutant des enclitiques (*e.g.* كتبوا+ها => كتبوها).
  - Le Alif Maqsura est transformée en simple Alif (ا, A) ou (ي, y) en dépendant du mot (*e.g.* حكى+هم => حكاهم et على+هم => عليهم).

Le DT connaît plusieurs clitiques dialectaux qui ne sont pas définis en ASM. Le tableau 3.3 présente quelques clitiques.

Clitiques dialectaux	Type	Exemples	Traduction
+شي	Clitique d'interrogation	عملت+شي ʕmlt+šy	as-tu fait ?
+ع	Préposition	عالطاولة ʕAlTAwlħ	sur la table
+م	Préposition	مالبيت mAlbyt	de la chambre
	Conjonction de coordination	مبعضهم mbʕDhm	ensemble
+ش	Clitique de négation	ما كتبّش mA ktb+š	il n'a pas écrit

TABLE 3.3 – Quelques clitiques dialectaux.

3.3.2.5 Les exceptions lexicales

Comme CODA-EGY, nous définissons une liste de mots qui ont une orthographe exceptionnelle ou qui peuvent être écrits selon plusieurs formes. Nous avons défini cette liste pour que la transcription de ces mots soit unique. Nous citons dans le tableau 3.4 quelques exemples de mots.

Type	Orthographe CODA	Autres formes orthographiques
Pronom personnel	انتي Anty	انت Ant, إنت Âint
	انتوما AintuwmA	نتوما ntwmA
Pronom démonstratif	هذاكه hðAkh	هذاك hðAk, هاذاكة hAðAkħ, هذاكة hðAkħ
Noms	عالسلامة ʕAlslAmħ	عسلامة ʕslAmħ

TABLE 3.4 – Quelques exemples de mots suivant la convention CODA-TUN.

De plus, plusieurs mots étrangers sont utilisés et intégrés dans le lexique tunisien. Dans ces mots, nous distinguons les phonèmes non arabes /g/, /v/ et /p/. Nous utilisons les consonnes arabes (ق, q), (ف, f) et (ب, b) pour représenter respectivement ces phonèmes.

3.3.3 Comparaison entre CODA-TUN et OTTA

Les deux conventions OTTA et CODA-TUN proposées dans cette thèse présentent plusieurs différences. La principale différence est au niveau phonologique. En OTTA, on essaye de réduire au minimum les différences entre la phonologie et la forme orthographique du DT. Par contre, CODA-TUN essaye de réduire les différences entre la forme orthographique du DT et de l'ASM.

Nos deux conventions sont adéquates pour la création des corpus oraux pour le DT. L'objectif visé de la création d'un corpus oral permet de choisir entre l'application de la première et la deuxième convention de transcription.

Une transcription suivant notre première convention qui reflète les spécificités phonologiques du DT oral est utile pour diverses applications de TAP. Le développement des applications de reconnaissance vocale, de synthèse de parole, etc. est plus réalisable avec notre convention OTTA. En revanche, cette dernière peut présenter des difficultés lors de l'analyse linguistique du DT. Dans ce cas, notre deuxième convention CODA-TUN avec une transcription plus proche de l'ASM permet de résoudre ces difficultés. CODA-TUN entre dans le cadre de l'approche de transcription à base de l'ASM. Cette approche facilite l'adaptation des outils de traitement linguistiques de l'ASM en faveur du DT. Ainsi, le nombre de modifications et de traitements apportés lors de l'adaptation est plus réduit vu le partage de plusieurs caractéristiques (les règles de segmentation des mots, la dérivation, etc.). On a besoin uniquement de cerner les différences entre la forme standard et la forme dialectale (*e.g.* identification de nouveaux affixes, la détermination des règles de segmentation, etc.).

Le tableau 3.5 montre un extrait de notre corpus transcrit avec nos deux conventions de transcription orthographique.

OTTA	وباش نحكيو زادة عالزيادات في أسوام القاز ونسألوا عالإنعكاسات متاعو عالمواطن. wbAš nHkyw zAdħ ϛAlzyAdAt fy ÂswAm Alq.Az wnsÂlwA ϛAlĂnϛkAsAt mtAϛw ϛAlmwATn
CODA-TUN	وباش نحكيوا زادة عالزيادات في اسوام القاز ونسألوا عالانعكاسات متاعه عالمواطن wbAš nHkywA zAdħ ϛAlzyAdAt fy AswAm AlqAz wnsÂlwA ϛAlAnϛkAsAt mtAϛh ϛAlmwATn.
Traduction	Nous allons aussi parler de l'augmentation des prix du gaz et nous posons des questions sur son impact sur le citoyen.

TABLE 3.5 – Un extrait de notre corpus transcrit avec les deux conventions de transcription orthographique OTTA et CODA-TUN

3.4 Différences avec d'autres conventions

Dans la littérature, peu de travaux ont abordé la tâche de transcription de la langue arabe parlée, plus précisément l'arabe dialectal. Les deux conventions proposées dans ce chapitre ne présentent pas beaucoup de différences par rapport à celles proposées par les autres chercheurs. En effet, nous proposons, en premier lieu, OTTA qui regroupe des conventions ortho-

graphiques pour le DT et qui a été ensuite améliorée par la convention CODA-TUN. Aussi, OTTA propose des annotations à ajouter pour améliorer les transcriptions de la langue parlée.

Comparons nos conventions avec les travaux existants. D'abord, à notre connaissance, nos conventions présentent le premier travail qui a établi une convention d'écriture pour le DT. Il ne traite pas seulement le dialecte de point de vue langue écrite comme déjà fait pour le dialecte égyptien et algérien dans le cadre de la CODA [Habash *et al.* 2012a] et [Saadane & Habash 2015], mais il prend en considération l'aspect oral.

Ainsi, nos conventions sont conçues aussi pour transcrire et annoter les dialectes comme étant une langue parlée en prenant en compte les caractéristiques de l'oral. Notons que [Maamouri *et al.* 2004a] ont proposé un ensemble de conventions pour créer un corpus pour le dialecte levantin et ont proposé aussi des annotations pour marquer la langue parlée, mais l'inconvénient majeur est que ce travail est basé sur l'outil AMADAT pour transcrire et annoter le dialecte. Or cet outil n'est pas libre et accessible par tous les chercheurs en TAL. Par contre, pour nos conventions, les annotations proposées ne se basent pas sur un outil spécifique. Elles peuvent être appliquées sur les outils d'aide à la transcription. De même, notre travail est basé sur le TOE [Bertrand *et al.* 2008] qui propose plusieurs annotations non traitées par [Maamouri *et al.* 2004a], comme par exemple les annotations des homophones.

3.5 Conclusion

La transcription orthographique pour une langue peu dotée sans tradition orthographique présente une tâche ardue. Dans ce chapitre, nous avons étudié la transcription orthographique du DT. Nous avons présenté tout d'abord les difficultés liées à la transcription manuelle du DT. Ensuite, nous avons proposé deux conventions orthographiques afin de transcrire et annoter le DT. La première se base sur une transcription orthographique basée sur la phonologie du dialecte (orthographe reflétant la phonologie). La deuxième se base sur une transcription qui se base essentiellement sur les règles orthographiques de l'arabe standard. Nous avons terminé ce chapitre par une comparaison avec les autres travaux réalisés dans ce cadre.

Au niveau du chapitre suivant, nous présenterons l'application de nos conventions orthographiques afin de créer un corpus pour le DT.

CHAPITRE 4

Création d'un corpus pour le dialecte tunisien parlé

Sommaire

4.1 Introduction

Les corpus sont fondamentaux pour le développement des ressources (*e.g.* les lexiques, etc.) et des applications de TAL. Pour les dialectes arabes, les corpus sont très limités, en particulier pour le DT. Au cours des dernières années et depuis les événements de la révolution, la présence du DT dans les interviews, les journaux et les programmes de débats s'est encore renforcée. L'utilisation croissante des technologies linguistiques pour de nombreuses langues parlées [Boujelbane *et al.* 2013] et le développement continu des travaux sur les technologies de la parole (*e.g.* Siri) nécessitent une énorme quantité de corpus oraux. Dans la littérature, il existe peu de corpus développés pour le DT. Généralement, la majorité de ces contributions visent à développer des corpus en considérant le dialecte comme étant une langue écrite. En outre, un ensemble de transcriptions ont été proposées, mais, pour un domaine très limité sans annotations linguistiques ([Masmoudi *et al.* 2014] ; [Graja *et al.* 2013]).

Après le développement des conventions de transcription orthographique, nous nous sommes fixés comme objectif de construire un corpus pour le DT oral ; un corpus général qui traite une variété de thèmes et est enrichi avec plusieurs types d'annotations (linguistiques et spécifiques à l'oral).

Ce chapitre détaille la construction de notre corpus STAC (Spoken Tunisian Arabic Corpus). La section 4.2 présente une description de ce dernier. Nous décrivons, aussi, notre méthode pour la collecte et la transcription de STAC. Ensuite, au niveau de la section 4.3, nous évaluons cette ressource en calculant l'accord inter-annotateurs et intra-annotateur. Enfin, nous exposons, dans la section 4.4, les différents types d'annotations que nous avons ajoutés pour enrichir notre corpus afin de le rendre plus utile pour nombreuses applications de TAL.

4.2 Description du corpus

Avant de présenter notre corpus « STAC : Spoken Tunisian Arabic Corpus » [Zribi *et al.* 2015], nous commençons par discuter les corpus existants pour le DT.

Dans la littérature, quelques travaux ([McNeil & Faiza 2011] ; [Graja *et al.* 2013] ; [Masmoudi *et al.* 2014] ; [Boujelbane *et al.* 2014a]) ont créé des corpus pour le tunisien. [McNeil & Faiza 2011] ont eu comme but de collecter une grande quantité des données textuelles pour le DT. Vu les rares ressources textuelles pour le DT, ils ont fait recours à plusieurs ressources pour collecter leur corpus. Le point faible de leur méthode est que la collecte a recueilli une importante quantité de textes en ASM. De ce fait, on ne peut pas utiliser le corpus de [McNeil & Faiza 2011] sans passer par une étape de filtrage préalable. Pareillement, le corpus de [Boujelbane *et al.* 2014a] est construit à la base de transcriptions abordant une forme très spécifique du DT avec l'alternance codique avec l'ASM. De même, les corpus oraux construits par ([Graja *et al.* 2013] ; [Masmoudi *et al.* 2014]) traitent un lexique spécifique relatif au transport ferroviaire. Malgré les tailles de ces corpus qui sont importantes, l'exploitation de ces ressources pour le développement reste limitée vu l'absence d'annotations linguistiques nécessaires pour le développement d'outils génériques pour le TAL.

En effet, notre objectif est le développement des outils en faveur du DT. La performance de ces outils (des analyseurs morphologiques, des étiqueteurs morphosyntaxiques, des analyseurs syntaxiques, etc.) se mesure par le taux réduit des mots hors vocabulaire. D'où, le premier avantage de notre corpus est son vocabulaire générique qui couvre le maximum du lexique du DT. D'autre part, nous traitons la langue parlée qui se caractérise par plusieurs phénomènes de l'oral (les pauses remplies, les pauses, les auto-corrections, etc.). D'où, le deuxième avantage de notre corpus est la présence des caractéristiques de la langue parlée. Ainsi, nous envisageons de créer le corpus STAC qui sera un échantillon significatif du DT reflétant les principales caractéristiques du DT ainsi que les particularités de l'oral [1]. .

Le processus de création d'un corpus oral se compose de deux principales étapes. La pre-

1. Les ressources et outils développés dans cette thèse sont disponibles auprès de l'auteur sous une licence libre.

mière consiste à collecter les données vocales. La seconde est la transcription de ces dernières en suivant des conventions orthographiques.

4.2.1 Collection et description des données

Le Web est une immense base de ressources disponibles en un clic de souris. Il contient des centaines de milliards de mots de textes, de vidéo et d'audio qui peuvent être utilisés pour toutes sortes de recherche linguistique [Kilgarriff & Grefenstette 2001]. Dans la littérature en TAL, plusieurs chercheurs ([Diab *et al.* 2010] ; [Zaidan & Callison-Burch 2011] ; [Cotterell & Callison-Burch 2014] ; etc.) ont exploité les ressources textuelles existant dans les commentaires des blogs, des journaux, etc. pour collecter leurs corpus en AD ainsi que les lexiques pour divers dialectes arabes (*e.g.* égyptien, levantin, irakien, etc.).

Notre objectif n'est pas très éloigné de ces travaux. Nous visons la création d'un corpus oral pour le DT. La première étape de notre méthode consiste à fournir les données vocales pour les transcrire. Pour collecter ces dernières, plusieurs chercheurs ont recours à la méthode d'enregistrement. Il s'agit d'enregistrer les paroles d'un ensemble de locuteurs parlant la langue cible discutant un ou plusieurs sujets dans des studios spécialisés, les rues, les stations de la gare, les aéroports, etc. selon la finalité du corpus. L'inconvénient de cette méthode est le coût élevé de la réalisation de ces enregistrements (en matière d'argent et de temps).

Pour faciliter notre tâche de collection des données vocales, nous avons suivi la méthode « télécharger et enregistrer » proposée par [Waibel *et al.* 2004] pour chercher et enregistrer les fichiers audio dont les locuteurs s'expriment en DT. Depuis les événements de la révolution, le volume de données en DT (oral et écrit) est en évolution continue. Le DT est devenu fréquemment utilisé dans les interviews, les journaux télévisés, les émissions de débats, etc. qui se rediffusent sur le Web [Boujelbane *et al.* 2013]. Pour constituer notre corpus, nous sommes essentiellement basés sur les audio disponibles dans le Web afin de les transcrire et annoter. La majorité de ces ressources sont disponibles gratuitement sur le Web.

Le choix du type et du contenu des enregistrements audio à télécharger est une étape liminaire afin de garder les objectifs visés pour notre corpus. Nous rappelons que notre corpus a deux principaux objectifs : (*i*) un corpus oral riche en phénomènes de l'oral (*ii*) un corpus pour le DT montrant toutes les caractéristiques linguistiques du DT. Ainsi, nous essayons de sélectionner les sources pour télécharger des enregistrements audio conformément aux objectifs visés.

Dans le Web, il existe une multitude de vidéos et d'enregistrements en DT traitant plusieurs thèmes, enregistrés dans des conditions différentes avec des qualités sonores variables. Par exemple, on trouve des séquences qui peuvent être une bonne ressource pour le DT où les locuteurs parlent avec un lexique riche en montrant les spécificités du DT, mais, la qualité médiocre de l'enregistrement audio (la présence de bruits, de la musique en arrière-plan, etc.) nous empêche de retenir ces enregistrements.

Le choix des meilleures séquences audio, avec une bonne qualité, nous oblige à passer

par une écoute parfois répétitive et minutieuse. Ainsi, pour réduire la difficulté de la tâche, nous choisissons de télécharger les séquences à partir des sites Web des chaînes télévisées et radiophoniques. Ces séquences sont la rediffusion des émissions et des programmes en télévision ou en radio. Le téléchargement à partir de ces sites est gratuit. De même, la qualité de ces enregistrements est considérée comme excellente.

Quatre chaînes de télévision tunisiennes sont utilisées (Télévision nationale tunisienne [2], El Hiwar Ettounsi [3] et Hannibal TV [4]) et deux stations de radio (Mosaïque FM [5] et Radio Sfax [6]). Ces médias proposent un ensemble riche d'émissions et de programmes. Nous cherchons celles qui contiennent le maximum de séquences en DT et relativement riches en phénomènes de l'oral. Ces données vocales sont généralement des débats télévisés, des programmes interactifs où le grand public est invité à participer à la discussion par téléphone.

Nous avons sélectionné trois types de programmes. Le premier programme présente des débats où la présentatrice propose un sujet sur lequel les intervenants parlent et discutent. Ces interventions sont réalisées via des appels téléphoniques. Chaque intervenant présente son idée de façon spontanée et la présentatrice intervient en posant diverses questions. La durée de chaque intervention peut varier de quelques secondes à quelques minutes. Ce programme traite des thématiques différentes : santé, sociale et politique.

Le principe du deuxième programme est de laisser les auditeurs de la radio intervenir et discuter de façon libre des sujets divers. Chaque intervenant a la liberté de choisir le sujet de l'intervention téléphonique. La conversation peut durer quelques secondes (en moyenne 30 secondes). Notons que les locuteurs (hommes et femmes) ont des âges variés et appartiennent à diverses classes sociales. Le tableau 4.1 présente quelques statistiques sur les locuteurs du corpus.

Cette caractéristique crée une variété dans le lexique utilisé. Souvent, on trouve des personnes qui alternent entre le français et le DT et d'autres utilisent dans leur discours l'ASM. Ceci a engendré une variété lexicale de notre corpus.

	Nombre de locuteurs	**Pourcentage**	**Durée totale des conversations (en %)**
Femme	118	38.94 %	40.73 %
Homme	185	61.06 %	59.27 %
Total	**303**	**100 %**	**100 %**

TABLE 4.1 – Des statistiques sur les locuteurs présents du corpus.

Au niveau du troisième programme choisi pour notre corpus, l'animateur parle d'un sujet particulier (*e.g.* religion) en racontant plusieurs histoires. La séquence audio de ce programme se caractérise par la quasi-absence des phénomènes de l'oral. Cependant, nous avons choisi d'inclure ce type de programme dans notre corpus pour le rendre plus générique et inclure toutes les variétés de discours en DT. Notons que le pourcentage de ce type de discours est

2. https ://www.watania1.tn/
3. https :// http ://www.elhiwarettounsi.com/
4. https ://www.hannibaltv.com.tn/
5. http ://www.mosaiquefm.net/
6. http ://www.radiosfax.tn/

assez petit en le comparant avec les autres types d'enregistrements. Le tableau 4.2 montre quelques statistiques concernant le corpus STAC.

	Parole préparée	Parole semi-spontanée	Parole spontanée
Durée	00 :12 :38	01 :52 :47	02 :45 :06
Pourcentage	4.35%	38.82%	56.83%

TABLE 4.2 – Des statistiques sur la nature de parole dans le corpus STAC

Ainsi, nous avons collecté plusieurs enregistrements pour les différents types d'émissions déjà présentées. Nous avons téléchargé plus de 5 heures d'audio contenant plusieurs éléments qui ne sont pas importants pour notre corpus. Ils contiennent des séquences bruitées, de la musique et de la parole incompréhensible, etc.

Avoir une bonne quantité d'enregistrements est fondamental dans la conception du corpus. En outre, un son de haute qualité est nécessaire et utile pour les utilisations ultérieures du corpus, (*e.g.* le développement d'un système de reconnaissance vocale, système de dictée automatique, etc.). Pour maintenir la bonne qualité de notre corpus, nous avons sauvegardé seulement les séquences dans lesquelles les locuteurs n'interviennent que sur un seul sujet à la fois. De même, nous avons choisi les séquences avec une bonne qualité sonore. Parfois, la qualité sonore de l'enregistrement peut varier considérablement au fil du temps. Par conséquent, les fichiers correspondant aux 5 heures que nous avons téléchargées nécessitent une étape de filtrage afin d'éliminer les bruits, les musiques, les séquences qui contiennent beaucoup de paroles en langue française ou en ASM qui peuvent affecter la qualité de notre corpus.

Donc, nous procédons au filtrage des séquences bruitées (musique ou autres bruits) qui durent plus d'une seconde. Pour ce faire, la seule méthode fiable pour le filtrage est l'écoute minutieuse. Il s'agit d'écouter les données audio et de supprimer les séquences non intéressantes pour notre corpus en utilisant l'outil de traitement de signal « Audacity [7] ».

Nous veillons à ce que tous les enregistrements contiennent des discours spontanés et le pourcentage du contenu dialectal est très supérieur à l'ASM ou du contenu en français. En outre, de nombreuses séquences d'émissions ont été supprimées particulièrement vu qu'elles contiennent des paroles non spontanées. Ainsi, nous avons divisé les données vocales en un ensemble d'enregistrements de petite taille qui varie de quelques secondes à une vingtaine de minutes. Cette décomposition a pour objectif de faciliter la transcription en utilisant l'outil d'aide à la transcription et aussi pour différencier entre les séquences qui sont spontanées, peu spontanées et préparées.

Le résultat de l'étape de filtrage est 4 heures et 28 minutes de discours en DT collecté à partir de plusieurs chaînes de télévision et stations de radio. Ces données vocales présentent la première partie de notre corpus STAC [Zribi *et al.* 2015].

Pour prendre en considération l'aspect oral, nous choisissons d'ajouter à notre corpus un ensemble de conversations enregistrées dans la gare. Il représente la deuxième partie de notre corpus qui dure environ 30 minutes extrait à partir du corpus « TuDiCoI : Tunisian Dialect

7. https ://audacity.fr

Corpus Interlocuteur » [Graja *et al.* 2013]. Le corpus de [Graja *et al.* 2013] est composé de dialogue en DT qui regroupe un ensemble de conversations enregistrées dans la station de chemins de fer entre le personnel et les clients en demandant des informations sur le temps du départ des trains, le prix des billets, faire des réservations, etc. [Graja *et al.* 2013]. Nous avons refait la transcription de cette partie suivant nos conventions de transcription et d'annotation.

La diversité des thèmes et des locuteurs en DT rend notre corpus plus générique. Il inclut de la parole spontanée, de la parole moins spontanée et parfois préparée. En outre, le nombre relativement important de locuteurs (environ 303 locuteurs) dans notre corpus parlant chacun avec son propre style fait de notre corpus un échantillon représentatif du DT. Ainsi, STAC contient de la parole conversationnelle et individuelle. Il permet d'identifier les différents aspects de la parole en DT.

Les enregistrements de radio et de télévision ont un contenu varié (une grande variété de locuteurs et de thèmes (social, santé, religieux, politique et autres)). Cette variété pourra être très utile pour des futures applications comme la classification de thèmes [Bischoff *et al.* 2009].

Le corpus contient des données vocales regroupant les dialectes de différentes régions tunisiennes mais, le dialecte de Tunis (la capitale de la Tunisie) est le plus dominant. C'est le dialecte utilisé dans les médias tunisiens. Il présente environ 90% de la totalité de notre corps. Le tableau 4.3 présente les différentes portions de notre corpus selon le thème. La figure 4.1 présente le pourcentage des langues dans le corpus STAC.

Thèmes	**Durées**
Social	01 :01 :35
Santé	01 :30 :46
Religieux	00 :12 :38
Politique	00 :50 :50
Autres	01 :14 :42
Total	**04 :50 :30**

TABLE 4.3 – Taille du corpus STAC selon le thème.

Il existe quelques travaux réalisés pour la création des corpus pour le DT ([Masmoudi *et al.* 2014] ; [Graja *et al.* 2013] ; [Boujelbane *et al.* 2014a]). À notre connaissance, notre corpus est la première ressource pour le DT qui contient différents types d'annotations et enrichissements. En outre, il est composé des enregistrements manuellement (partie 2) et des enregistrements téléchargés à partir du Web (partie 1) traitant des thèmes différents.

4.2.2 Schéma de transcription et d'annotation

La seconde étape du processus de création d'un corpus oral est la transcription orthographique. Il existe plusieurs logiciels d'aide à la transcription orthographique et l'annotation des fichiers audio (Transcriber, Praat, Anvil, Elan, etc.), dont chacun a une spécificité particulière (annotation de gros fichiers, annotation prosodique, etc.) [Bazillon 2011].

La transcription de notre corpus STAC est faite via l'outil d'aide à la transcription

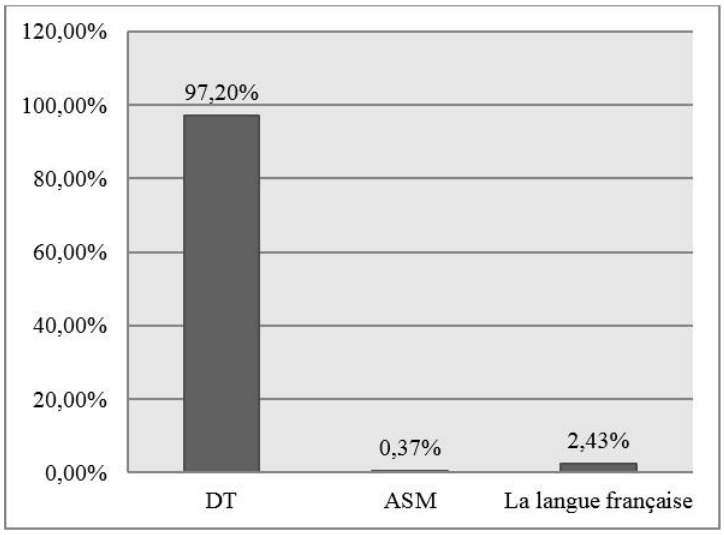

FIGURE 4.1 – Pourcentage des langues dans le corpus STAC.

Praat[8] qui est un logiciel libre pour l'analyse, la manipulation et l'annotation de sons [Boersma & Weenink 2016]. Il est considéré comme un outil complet pour l'étude de la parole. Ses interfaces graphiques et ses menus simplifiés le rendent pratique même pour les non-experts en traitement de la parole grâce à ses différentes fonctionnalités (enregistrement de fichiers audio, transcription, étiquetage, segmentation de fichiers, analyse phonétique et acoustique, analyse des paramètres prosodiques, manipulation et modification du signal de la parole, etc.) ([Delais-Roussarie & Durand 2003] ; [Goldman 2006]). Praat permet, aussi, l'alignement du son avec le texte ; ce qui offre une meilleure présentation pour les données vocales. (*cf.* figure 4.2).

Le choix de cet outil est justifié comme suit. D'abord, c'est le logiciel utilisé pour la transcription et l'annotation des paroles dans le projet OTIM[9] et nous visons la compatibilité de notre corpus avec le format utilisé dans ce projet. Ainsi, Praat permet l'analyse phonétique de la parole et soutient également la synthèse de la parole, y compris la synthèse articulatoire[Boersma & Weenink 2016]. Il peut fournir une transcription alignée entre la parole et le texte. Il facilite, aussi, l'étiquetage et la segmentation de la parole grâce aux tires fournis par son interface graphique. Ses fonctionnalités permettent l'usage de notre corpus pour d'autres travaux d'analyse prosodique et phonétique.

L'outil Praat offre plusieurs fonctionnalités aidant à la transcription et à l'annotation du signal acoustique. Un parmi les objets utilisés pour l'annotation du signal est TextGrid. Ce dernier permet d'annoter le signal sur un ou plusieurs niveaux d'annotation. Ces niveaux s'appellent « tire d'annotation » (en anglais Tier) [Boersma & Weenink 2016]. Les tires aident à faire la segmentation du signal en utilisant les frontières, qui marquent la fin et le début d'un intervalle de temps. Chaque intervalle de temps peut marquer la présence d'un phénomène particulier de l'oral ou bien d'un mot ou d'une séquence de mots.

8. http ://www.fon.hum.uva.nl/praat/
9. http ://www.lpl-aix.fr/˜otim/index.html

Pour notre corpus, nous choisissons d'utiliser les tires pour marquer les paroles d'un seul locuteur. Chaque tire porte le nom du locuteur qui parle dans cet intervalle de temps. Ainsi, l'utilisation des tires pour la transcription aide à marquer le phénomène de chevauchement des paroles ce qui n'est pas le cas pour d'autres outils d'aide à la transcription (*e.g.* Transcriber).

Avant d'entamer la transcription, nous faisons la segmentation du signal en utilisant une fonction de Praat, qui détecte les pauses ayant une durée minimale de 200 ms. La détection automatique des pauses exige une étape de rectification et de correction des frontières des unités inter pausales. Nous marquons ces zones par le symbole « # ». De même, nous détectons les phénomènes non linguistiques (*e.g.* les rires, les respirations, la musique, etc.) et nous les marquons selon l'annotation définie dans notre convention d'annotation (OTTA : Orthographic Transcription of Tunisian Arabic).

Après la segmentation du signal, nous procédons à la transcription de la parole de chaque locuteur dans la tire portant son nom. En cas de chevauchement, on transcrit la parole de chaque locuteur dans sa tire. Après la transcription, nous faisons l'alignement du signal aux mots transcrits. Nous définissons pour chaque mot les intervalles correspondants au début et à la fin du mot. La figure 4.2 montre un exemple d'objet TextGrid.

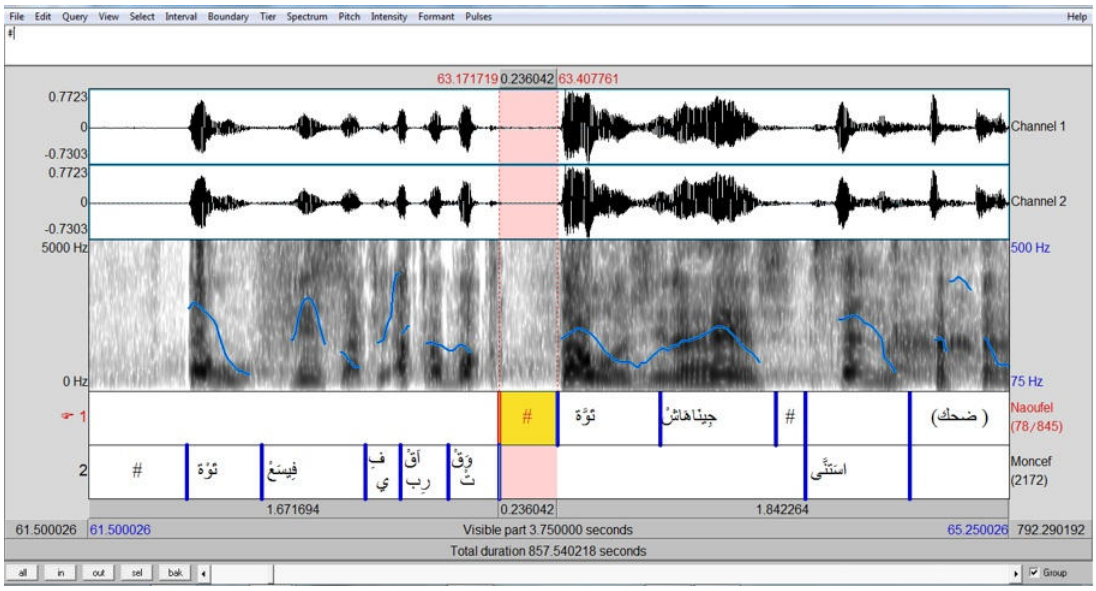

FIGURE 4.2 – Exemple d'objet TextGrid.

4.3 Évaluation de la transcription du corpus

La transcription d'un corpus oral est confrontée à plusieurs problèmes liés à l'oral et d'autres spécifiques à la nature du DT. Ces difficultés troublent les transcripteurs et affectent la qualité de la transcription du corpus. Les transcripteurs sont parfois perplexes et font des choix au hasard des formes orthographiques des mots ou même au niveau des annotations. La consistance de la transcription est importante pour les diverses applications prenantes comme

entrée l'oral et ses transcriptions. La question qui se pose est « Comment peut-on vérifier la bonne qualité d'une transcription ? ».

En effet, la tâche de transcription orthographique du signal audio, en utilisant un outil d'aide à la transcription, consiste en une annotation orthographique de tous les phénomènes observés dans le signal (les paroles, les bruits, les chevauchements, etc.). L'annotateur (ou le transcripteur pour notre cas) doit choisir, pour une séquence audio donnée, un mot du lexique du DT ou une annotation spécifique à l'oral.

La qualité des annotations dépend de l'annotateur (*i.e.* la motivation du transcripteur, sa responsabilité, etc.) et des influences externes (*i.e.* pression du temps, fatigue, etc.). L'annotateur, dans plusieurs cas, peut faire des erreurs d'inattention ou d'incompréhension de la tâche demandée qui peut s'avérer non claire (tâche ambigüe). Ces faits nous empêchent de faire une confiance totale à un annotateur [10].

Généralement, l'évaluation de la qualité des annotations nécessite une annotation de référence s'il s'agit d'une annotation automatique. On mesure le niveau de rapprochement entre les annotations automatiques et les annotations de référence. Les mesures de rappel, de précision et de f-mesure sont utilisées pour évaluer l'annotation automatique. En revanche, lorsqu'il s'agit d'une annotation manuelle, l'évaluation nécessite une comparaison avec une autre annotation humaine. Elle mesure le taux d'accord entre les différentes annotations. Dans ce cas, on calcule le taux d'accord inter et/ou intra-annotateur pour valider la qualité des annotations manuelles.

Pour notre cas, le calcul d'accord inter et intra-annotateur, nous aide à vérifier la bonne compréhension de nos directives de transcription orthographique et d'annotation par les annotateurs [Fort & Claveau 2012]. Les annotateurs peuvent se mettre d'accord sur des annotations et dans d'autres cas, les annotations peuvent se converger. Parmi les mesures les plus utilisées pour le calcul d'accord inter et intra-annotateur, nous citons le coefficient de *Kappa de Cohen* (k) (*voir équation* (4.1)) [Cohen 1960] qui est destiné à mesurer l'accord observé en fonction d'un accord dû au hasard [Fort & Claveau 2012]. Il calcule un rapport entre la probabilité d'accord Pa de deux annotateurs et la probabilité d'un accord aléatoire Pe.

$$k = \frac{Pa - Pe}{1 - Pe} \tag{4.1}$$

Le calcul d'accord inter-annotateurs se base sur un ensemble bien défini d'annotations et aussi pour les éléments à annoter. Dans notre cas, la définition des annotables (les éléments à annoter) et les annotations présente un problème puisque la tâche à évaluer est la transcription orthographique et la validité des règles de transcriptions.

Nous proposons donc de calculer l'accord sur les mots, c'est-à-dire, les annotables sont les unités audio et les annotations sont les mots. Nous segmentons le son en unités en insérant des frontières au niveau de chaque tire d'annotation de Praat d'un enregistrement audio. Chaque unité intonative correspond à un mot. Par conséquent, nous proposons aux transcripteurs (les

10. http ://www.modyco.fr/fr/documents/m1-plurital/725-aia/file.html

annotateurs) d'écouter et de transcrire le mot ou les phénomènes correspondants au son (*e.g.* rire, bruit, musique, parole, etc.).

Le premier accord inter-annotateurs calculé concerne l'écoute des annotateurs. Il s'agit de mesurer l'accord de l'écoute des annotateurs pour la même séquence. L'objectif est de savoir si les deux annotateurs ont écouté le même élément ou non. Nous supposons qu'il existe un accord entre les deux annotateurs si on détecte les mêmes lettres ordonnées.

Dans un second temps, nous calculons l'accord inter-annotateurs pour la transcription orthographique, c'est-à-dire, l'application de la convention orthographique CODA-TUN. Nous mesurons à quel point les transcripteurs ont compris la convention et dans quels cas ils ont appliqué les règles de CODA-TUN.

Enfin, nous calculons l'accord intra-annotateur pour mesurer la qualité du travail du transcripteur lui-même en allant de début de la tâche jusqu'à la fin de la tâche de transcription. Nous nous basons sur le nombre d'erreurs émis par rapport à un corpus bien transcrit selon la convention CODA-TUN.

4.3.1 Corpus d'évaluation

Une grande partie de notre corpus STAC a été transcrite par un seul transcripteur en respectant les règles de transcription et d'annotation d'OTTA et CODA-TUN. Afin de vérifier la qualité de notre corpus, une partie de STAC a été transcrite par deux autres transcripteurs. Cette deuxième transcription nous permet de calculer l'accord inter-annotateurs.

Les transcriptions ont été faites par trois transcripteurs qui sont des locuteurs natifs du DT et qui parlent l'ASM ainsi que la langue française. Les trois transcripteurs sont :

— Tr1 est une doctorante en informatique.

— Tr2 est une ingénieure en informatique.

— Tr3 est une étudiante à la faculté des lettres et des sciences humaines de Sfax.

La tâche des transcripteurs consiste à écouter le son et à transcrire et annoter les paroles de chaque enregistrement audio.

Avant de donner le corpus aux transcripteurs, le corpus a passé par une étape de prétraitement. L'identification des locuteurs, la création de tires pour chaque locuteur et la segmentation des tires sont des tâches qui ont été effectuées avant d'entamer la transcription orthographique. En outre, les effets non linguistiques comme la musique, les bruits, etc. sont marqués.

La tâche d'un annotateur (transcripteur) consiste à écouter et de transcrire ce qu'il écoute en se basant sur ses propres connaissances de la langue arabe et du DT. Ensuite, dans un deuxième passage, nous lui expliquons nos directives d'annotation et de transcription. Puis, il refait la transcription en appliquant nos règles. Les guides de transcription et d'annotation fournis aux transcripteurs sont sous forme de règles où on explique chacune par quelques exemples montrant les différentes formes orthographiques et d'annotation possible selon la prononciation.

Nous choisissons de calculer l'accord inter-annotateurs (*Kappa de Cohen*) [Cohen 1960] pour deux parties de corpus choisis au hasard. La première partie est composée de 2 heures et 22 minutes. Nous avons utilisé cette partie pour calculer l'accord entre deux transcripteurs (Tr1 et Tr2). La deuxième partie est composée de 26 minutes. Nous avons utilisé ce sous-ensemble pour calculer l'accord entre les trois transcripteurs (Tr1, Tr2 et Tr3).

Pour garantir la qualité des transcriptions, le corpus d'évaluation est divisé sur plusieurs fichiers. D'abord, nous affectons aux transcripteurs une petite tâche. Il s'agit de transcrire un petit fichier afin de tester la compréhension des transcripteurs de l'objectif de la tâche demandée. Ensuite, lorsqu'ils terminent, nous discutons avec eux les problèmes rencontrés et nous répondons à différentes questions. Enfin, nous attribuons aux transcripteurs les autres fichiers à transcrire.

4.3.2 Résultats d'évaluation

L'évaluation de la qualité des transcriptions des données orales est mesurée en calculant l'accord inter-annotateurs. Pour l'évaluation de nos données, nous avons recours à la mesure k de Cohen [Cohen 1960]. Le calcul de ce coefficient se base sur la création d'une matrice de contingence. À partir de cette matrice, on calcule la probabilité d'accord Pa (*voir équation* (4.2)) et l'accord aléatoire Pe (*voir équation* (4.3)).

$$Pa = \frac{1}{n} \sum_{i=1}^{p} n_{ii} \tag{4.2}$$

$$Pe = \frac{1}{n^2} \sum_{i=1}^{p} n_{i.} n_{.i} \tag{4.3}$$

Pour notre tâche, la transcription orthographique des données, nous considérons les séquences audio (les segments de son) comme des annotables et les annotations sont les formes orthographiques correspondantes pour le son. De ce fait, la matrice de contingence sera de la forme suivante (*voir* tableau 4.4).

Annotateur 1 / **Annotateur 2**	fo_1	fo_2	fo_2	∅	...	fo_i	**Total**
fo_1	4	0	0	0			
fo_2	0	12	0	2			
fo_2	0	0	15	3			
∅	2	0	0	40			
...							
fo_i	.	.	.	.	.	n_{ii}	
Total							N

TABLE 4.4 – Un exemple de la matrice de contingence.

Les fo_i représentent toutes les formes orthographiques des mots utilisées pour les deux transcripteurs. Le nombre n_{ij} est le nombre de fois où il existe un accord entre la forme orthographique i et la forme orthographique j. Le symbole « ∅ » signifie le vide, c'est-à-dire, le

transcripteur n'a pas attribué une forme orthographique pour le segment de son correspondant. L'absence d'annotation est due à une erreur d'écoute ou le transcripteur a oublié d'écrire le mot ou les mots correspondants à cette séquence audio.

4.3.2.1 Accord inter-annotateurs

Premièrement, nous mesurons l'accord entre les transcripteurs concernant l'écoute de la même séquence pour le même segment de son. Pour ce faire, nous supposons que les deux transcripteurs ont écouté la même séquence si nous détectons les mêmes consonnes dans le même ordre. Généralement, les différences entre les transcriptions résident généralement au niveau de l'ajout ou suppression des voyelles courtes, des voyelles longues et/ ou les espaces. Prenons l'exemple de l'expression « *je n'ai pas dit* ». Cette expression est transcrite en DT en utilisant plusieurs formes orthographiques : (ما قلتش, mA qltš), (ماقلتش, mAqltš), (مقلتش, mqltš), etc. De ce fait, nous supposons que les transcripteurs ont écouté le même segment si et seulement s'ils ont choisi des formes orthographiques qui respectent le même ordre des consonnes et la distance de Levenshtein [Levenshtein 1966] entre les deux formes orthographiques est inférieure à la longueur minimale de la forme.

Prenons l'exemple du mot (مشى, mšy) proposé par le transcripteur 1 et le mot (مشا, mšA) proposé par le deuxième transcripteur. D'où, nous considérons que les deux transcripteurs ont écouté le même son malgré les différences entre les deux formes orthographiques. La matrice de contingence sera de la forme suivante (*voir* tableau 4.5).

Ann$_2$ \ Ann$_1$		**fo$_1$**		**fo$_2$**				∅	…	**fo$_i$**		**Total**
		مشى	مشا	ما قالش	م قالش	ماقالش	مقالش		…	**fo$_{1i}$**	…	
fo$_1$	مشى	4		0				0	0	…		…
	مشا							0	2	…		…
fo$_2$	ما قالش	0		0				15	3	…		…
	م قالش											
	ماقالش											
	مقالش											
∅		2		0				0	40	…		…
…	…	…		…				…	…	…		…
fo$_i$	**fo$_{1i}$**	…		…				…	…	n$_{ii}$		…
	…											
Total		…		…				…	…	…		**N**

TABLE 4.5 – Un exemple de la matrice de contingence pour l'accord de l'écoute.

Le tableau 4.6 présente les valeurs de l'accord inter-annotateurs calculées pour l'accord de l'écoute et l'accord pour la transcription orthographique avant l'application de la convention CODA et après la convention CODA.

Parmi les problèmes de la transcription, nous avons cité les problèmes liés à l'écoute. Parfois le transcripteur n'écoute pas la même séquence de mots qu'un autre transcripteur ou il oublie de transcrire un mot ou une séquence de mots à cause de plusieurs raisons. Ainsi, nous calculons l'accord de l'écoute pour vérifier la qualité des transcriptions, c'est-à-dire, elles présentent tous les phénomènes de l'oral (les paroles, les bruits, les hésitations, etc.). Nous re-

			Partie1		Partie2	
			% d'accord brut	*kappa*	*% d'accord brut*	*kappa*
Accord d'écoute		Tr1/Tr2	71,57	0,71	70,31	0,69
		Tr2/Tr3	-	-	68,01	0,67
		Tr1/Tr3	-	-	81,04	0,80
Transcription orthographique	Avant CODA-TUN.	$Tr1_c$/Tr2	64,20	0,63	63,31	0,63
		Tr2/Tr3	-	-	60,95	0,60
		$Tr1_c$/Tr3	-	-	65,59	0,65
	Après CODA-TUN.	Tr1/Tr2	79,49	0,8	65,94	0,65
		Tr2/Tr3	-	-	65,41	0,64
		Tr1/Tr3	-	-	78,10	0,77

TABLE 4.6 – Les valeurs de l'accord inter-annotateurs.

portons une valeur d'accord kappa égale à 0,71 qui est un accord fort entre les transcripteurs Tr1 et Tr2 pour la première partie. Pour la deuxième partie de l'évaluation, nous avons reporté aussi un accord fort pour les trois transcripteurs Tr1, Tr2 et Tr3 (la valeur moyenne égale à 0,72). L'accord entre le transcripteur Tr1 et Tr3 est le meilleur accort d'écoute trouvé.

Pour se comparer, nous utilisons les transcriptions du transcripteur Tr1 comme une référence. Nous avons collecté les cas de différence d'écoute. Généralement, les cas de différence sont reliés aux phénomènes de l'oral : les pauses remplies, les mots incomplets et les répétitions. L'analyse des résultats montre que le transcripteur Tr2 néglige dans la plupart des cas les effets spéciaux de l'oral (les pauses remplies, les mots incomplets, les répétitions, etc.). Il a tendance à écrire que les mots complets. Par contre, le transcripteur Tr3 oublie dans 2 % des cas d'écrire des séquences de mots dans un intervalle de temps bien défini. Nous remarquons que les cas d'oubli sont souvent au milieu des transcriptions.

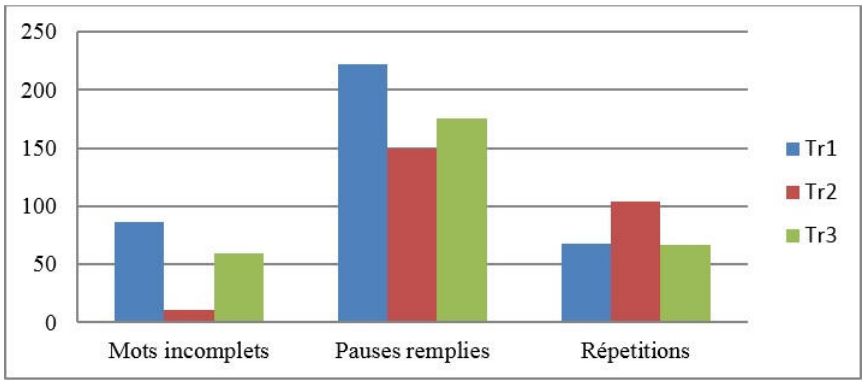

FIGURE 4.3 – Comparaison entre le nombre des mots incomplets, les pauses remplies et les répétitions détectés par les trois transcripteurs.

En outre, nous notons que la présence de plusieurs cas de désaccord en écoute sont dus au remplacement des mots par leurs homophones (*e.g.* (بالّله سامحني, bAllh sAmHny) et (بالّله صدقني, bAllh Sdqny)). Nous remarquons, aussi, que les deux transcripteurs (Tr2 et Tr3) ont échoué, parfois, d'écouter et de transcrire plusieurs mots comme les marqueurs de discours comme (باهي, bAhy), (فهمتني, fhmtny), « *voilà* », etc. De la même manière, dans plusieurs cas de désaccord d'écoute, nous remarquons que le transcripteur remplace le mot par un de ses syno-

nymes. Par exemple, nous détectons le remplacement de l'expression (عالسلامة أختي, ʕAlslAmħ Âxty) par l'expression (مرحبا أختي, mrHbA Âxty) qui sont deux expressions synonymes.

En effet, les valeurs de kappa et de l'accord brut calculées sans l'application des conventions orthographiques CODA-TUN sont encourageantes. Les valeurs reportées sont alentour de 0,62 (valeur moyenne) qui présente un accord fort. Le taux élevé d'accord montre que notre convention CODA-TUN est très proche de la transcription d'une personne normale. Plusieurs mots en DT en commun avec l'ASM gardent leurs formes orthographiques. En outre, notre convention orthographique propose souvent de transcrire les mots en gardant leur phonologie. Ces formes orthographiques sont, souvent, utilisées dans les médias, les enseignes publicitaires, etc. L'accord fort trouvé montre que nos conventions de transcription orthographiques sont simples à acquérir pour transcrire le DT et partagent les connaissances de base d'un transcripteur en DT.

Pour calculer l'accord lors de la transcription orthographique du DT, nous expliquons nos directives de transcription et d'annotation OTTA et CODA-TUN aux transcripteurs. D'abord, nous donnons à chaque transcripteur un fichier de petite taille (3 minutes) pour tester sa compréhension de la convention. Ensuite, nous discutons avec lui les erreurs effectuées. Enfin, chaque transcripteur termine la transcription de la partie restante du corpus d'évaluation.

Nous remarquons que le taux de désaccord a baissé de 15 % jusqu'à 5 % pour la petite partie d'apprentissage. Ainsi, nous constatons que la plupart des erreurs détectées sont au niveau de l'application des règles de segmentation des mots en dialecte, notamment la transcription de la conjonction de coordination (و, w) et du groupe prépositionnel d'objet (له lh, لك lk, etc.). La présence de ces erreurs de transcription est justifiée par le faux apprentissage des principes de transcription de base de la langue arabe dont les arabophones ont une tendance à ajouter ou supprimer des espaces respectivement après les conjonctions de coordination et le groupe prépositionnel d'objet. Ainsi, les scores de kappa reportés après l'application de la convention orthographique varient entre 0,64 et 0,8 qui sont des résultats encourageants. Elles prouvent la bonne qualité pour notre corpus STAC.

Nous remarquons que les valeurs de kappa sont améliorées de 0,17 pour la première partie. Par contre, nous signalons une faible amélioration pour la deuxième partie du corpus d'évaluation. L'amélioration varie entre 0,02 et 0,12. En effet, la richesse de ce la deuxième partie du corpus d'évaluation par les formes agglutinantes, les nombres et les mots étrangers a engendré cette faible amélioration des valeurs de kappa. Nous remarquons l'ajout des clitiques dialectaux aux mots et aux expressions de la langue française. Les transcripteurs ne savent pas comment transcrire ces clitiques. Dans certains cas, ils les écrivent avec des lettres latines, des lettres arabes et dans d'autres cas avec des lettres arabes attachées au mot ou à l'expression française (*e.g.* w ça va, و ça va, وça va). De même, les transcripteurs ont tendance à écrire les chiffres et non pas d'appliquer la règle qui exige la transcription des nombres à la lettre.

4.3.2.2 Accord intra-annotateur

Afin d'analyser le comportement de transcripteurs lors de la transcription, nous mesurons le taux d'accord intra-annotateur entre les transcriptions avant et après l'application des conventions de transcription. Le calcul de l'accord intra-annotateur pendant l'annotation, permet de vérifier que les annotateurs sont-ils cohérents avec eux-mêmes. Les taux d'accord lors de la transcription orthographique sont présentés dans le tableau 4.7.

	Partie1	**Partie2**
Tr1	95,45 %	100 %
Tr2	93,63 %	90,30 %
Tr3	-	93,93 %

TABLE 4.7 – Les valeurs de l'accord intra-annotateur.

Comme les résultats le montrent, les transcripteurs Tr1, Tr2 et Tr3 ont une tendance à donner aux mots les mêmes formes orthographiques dans 90 % des cas. Les cas d'incohérences sont généralement au niveau de la transcription des mots-outils qui ne sont pas définis dans la liste des mots-outils définis dans la convention CODA-TUN. De même, nous remarquons des erreurs au niveau de choix des formes orthographiques de certains mots étrangers. Les transcripteurs sont perplexes lors du choix des alphabets latins ou arabes. Nous prenons l'exemple du mot « madame » de la langue française qui est transcrit dans plusieurs cas comme (مدام, mdAm).

En revanche, nous avons choisi de calculer l'accord de l'application de quelques règles de transcription orthographique proposées dans la convention CODA-TUN. Nous avons choisi de mesurer l'applicabilité de deux règles : la première est de la segmentation de verbes dans leur forme de négation et la deuxième est de la transcription du pronom personnel singulier (ه, h). Pour calculer l'accord, nous avons collecté tous les noms qui doivent s'appliquer avec ces deux règles et nous avons calculé les pourcentages d'échec et le taux d'accord.

Le tableau 4.8 présente le taux d'application de quelques règles de transcriptions.

Règle de transcription de CODA	**Transcripteur**	**% d'échec**	**% de réussite**
Règle de segmentation des verbes en forme de négation « *mA+espace+verbe+š* »	Tr1	3,97	96,03
	Tr2	44,30	55,70
	Tr3	8,45	91,55
Règle de transcription du pronom personnel singulier (ه, h).	Tr1	9,21	90,79
	Tr2	27,86	72,14
	Tr3	12,12	87,88

TABLE 4.8 – Le taux d'application de quelques règles de transcription.

Selon les valeurs reportées, nous remarquons que la règle de segmentation des verbes en forme de négation est appliquée dans les 81,09 % des cas pour les trois transcripteurs. L'analyse des cas d'échec montre que les erreurs généralement se situent à la fin des fichiers de transcription. Ceci est dû, généralement, à la fatigue des transcripteurs. Par contre, la règle de transcription du pronom personnel singulier (ه, h) a été appliquée dans 83,60 % des cas. Cependant, nous avons remarqué que les transcripteurs font une confusion entre les verbes

conjugués en troisième personne du pluriel comme par exemple (قالوا, qAlwA) /qa :lu :/ « ils ont dit » et la forme (قاله, qAlh, « il l'a dit »)/qa :lu :/.

Nous remarquons, quelquefois, des erreurs dans la transcription de quelques mots-outils comme (راهو, rAhw, « il est ») qui est transcrit en (راهوا, rAhwA) ou en (راهّه, rAhh). Notons que la règle de transcription du pronom personnel singulier (ه, h) ne s'applique pas pour ce mot outil.

4.4 Annotation du corpus

Les transcriptions de notre corpus STAC sont enrichies par des annotations proposées au niveau de notre directive d'annotation OTTA qui spécifie les annotations liées à l'oral comme les phénomènes non linguistiques (les rires, le bruit, etc.), les amorces, les pauses remplies, etc. et des annotations pour marquer et identifier les entités nommées et la langue de quelques mots et expressions empruntés.

Pour mener à bien des travaux de traitements linguistiques sur des corpus, de nombreuses annotations ont été proposées qui permettent d'obtenir diverses informations morphologiques, morphosyntaxiques, syntaxiques et sémantiques.

La plupart des outils de traitement linguistiques sont, cependant, conçus à partir d'un ensemble de données annotées selon la tâche pour laquelle l'outil est conçu. En effet, pour traiter automatiquement un corpus oral, des annotations basiques doivent être présentes dans un corpus. Un de nos objectifs dans cette thèse est de développer les outils de traitement du DT et par conséquent, nous nous limitons à ajouter des annotations traitant la morphologie et la morphosyntaxe et d'autres spécifiques aux caractéristiques de l'oral : les disfluences.

4.4.1 Annotation morphosyntaxique

4.4.1.1 Le système d'annotation

Le principe de l'annotation morphosyntaxique consiste à associer à chaque mot de l'énoncé la ou les catégories correspondantes [Bertrand *et al.* 2008]. Afin de réaliser cet objectif pour notre corpus STAC, nous avons choisi d'utiliser nos propres outils pour l'analyse morphologique et morphosyntaxique.

Il existe plusieurs systèmes qui permettent d'analyser morphologiquement et morpho syntaxiquement l'ASM, Mais, pour le DT, ces outils sont absents. Pour ces raisons, nous nous sommes basés sur l'adaptation de l'analyseur morphologique de l'ASM « Al-Khalil » [Boudlal *et al.* 2010] au DT « Al-Khalil-TUN » [Zribi *et al.* 2013b]. Plus de détails sur l'adaptation de l'analyseur morphologique seront présentés dans le chapitre suivant.

En effet, la méthode d'annotation morphosyntaxique que nous proposons pour annoter notre corpus STAC est composée de plusieurs étapes.

La première est la segmentation du corpus en des phrases (ou *utterances* en anglais). Il s'agit d'identifier manuellement les frontières des phrases en oral afin de pouvoir analyser

morpho-syntaxiquement le corpus. L'identification des frontières des phrases en oral est une tâche ardue. Comme déjà présenté dans les chapitres précédents, l'oral se caractérise par l'absence des marques de ponctuation qui permettent d'identifier les frontières des phrases comme est le cas pour les corpus écrits. Vu l'absence des outils de segmentation des corpus oraux pour le DT, nous avons effectué cette tâche manuellement [11]. Notre corpus STAC intègre la transcription de nombreuses conversations entre plusieurs locuteurs. Les paroles de chaque locuteur sont composées de plusieurs tours de parole. Nous collectons les tours de parole pour chaque locuteur dans un texte unique. Ensuite, nous le segmentons manuellement en des phrases. Nous considérons un ensemble de mots comme étant une phrase toute unité qui a une sémantique significative. Les transcriptions comprennent de nombreuses annotations. Quelques annotations sont très utiles dans le processus d'annotation morphosyntaxique. D'autres annotations telles que le bruit et la musique sont supprimées.

La seconde étape de notre méthode d'annotation du corpus est l'analyse morphologique. Nous utilisons Al-Khalil-TUN [Zribi *et al.* 2013b] pour segmenter les mots en DT en identifiant les affixes et les clitiques attachés. De même, il attribue à chaque composant du mot (lemme, affixes et clitiques) les différentes caractéristiques morphologiques possibles. Nous présentons ci-dessous (figure 4.4) un extrait du corpus étiqueté morpho-syntaxiquement.

L'application de l'analyse morphologique sur la totalité du corpus nous donne un corpus segmenté et enrichi avec les caractéristiques morphologiques.

```
<?xml version="1.0" encoding="UTF-8"?>
<sentences>
  <sent num="1">
    <word num="1" value="توة" numSol="1">
      <sol num="1">
        <dialect value="TA" />
        <asp value="asp = na" />
        <gen value="gen = na" />
        <num value="num = na" />
        <stt value="stt = na" />
        <vox value="vox = na" />
        <per value="per = na" />
        <suffix value="enc0 = 0 + suff0 = 0" />
        <pos value="pos = temp_adv" />
        <wordroot value="#" />
        <Wordpattern value="#" />
        <stem value="توة" />
        <prefix value="prc0 =  0" />
        <voweledword value="تَوَّة" />
        <true_false value="true" />
      </sol>
    </word>
    <word num="2" value="عمال" numSol="5">
    <word num="3" value="الجماعة" numSol="2">
    <word num="4" value="يكثرو." numSol="2">
  </sent>
  <sent num="2">
  <sent num="3">
```

FIGURE 4.4 – Extrait du corpus STAC annoté avec les étiquettes morphosyntaxiques.

Comme tout système d'analyse morphologique, un ensemble d'analyses morphologiques est proposé pour un mot. Ainsi, une étape de désambiguïsation sera nécessaire pour choisir la

11. Dans le chapitre 6, nous présentons une méthode pour automatiser la segmentation des corpus oraux pour le DT

bonne étiquette morphologique suivant son contexte et sa définition. Le choix manuel est très difficile et coûteux en temps. Par conséquent, nous présentons une méthode itérative permettant de faciliter la désambiguïsation morphosyntaxique (ou l'annotation) semi-automatique de notre corpus STAC.

L'idée principale de notre méthode consiste à diviser le corpus en 10 dossiers de taille équivalente. Habituellement, la version Al-Khalil-TUN retourne une liste d'analyses pour un mot avec des informations différentes (le genre, le nombre, la personne, les affixes, les enclitiques, la voix et la catégorie grammaticale). Nous gardons toutes ces caractéristiques morphologiques lors de l'annotation de notre corpus.

Nous commençons par analyser la première partie du corpus avec Al-Khalil-TUN. Nous choisissons, ainsi, la bonne analyse selon la position du mot dans la phrase. Lorsque l'analyseur échoue à donner une analyse pour un mot, nous déterminons l'ensemble de caractéristiques morphologiques correspondantes à ce mot.

Ainsi, nous nous entraînons une première version d'un système de désambiguïsation avec la première partie du corpus complètement annoté à la main. Nous utilisons le modèle résultant pour annoter le second dossier du corpus. Nous corrigeons manuellement la sortie du système. Nous ajoutons la partie corrigée au corpus d'apprentissage. Ensuite, nous réitérons ce processus avec les différentes parties du corpus.

À la fin de ce processus, nous obtenons un corpus annoté morphosyntaxiquement. Toutes les analyses morphologiques sont conservées, et la bonne analyse dans un contexte donné est marquée.

4.4.1.2 Statistiques

L'étape de détection des frontières de phrases dans le corpus a permis d'identifier un nombre assez important de phrases. Ces phrases peuvent être classées en quatre classes. Le premier type regroupe les phrases qui ont une structure proche de celle de l'ASM (type 1), c'est-à-dire, elles ont une structure grammaticale de la forme SVO ou VSO. Le deuxième type de phrases regroupe des phrases qui sont spécifiques à la parole (type 2). Il regroupe des phrases de salutation et de remerciement. Le troisième type de phrases rassemble des phrases qui n'ont pas de structure grammaticale correcte (type 3). Ces phrases sont sémantiquement correctes et sont souvent très utilisées dans le DT. Ainsi, vu que le corpus est composé de la parole spontanée, nous remarquons l'existence de plusieurs phrases incomplètes (type 4). La figure ci-dessous (cf. figure 4.5) présente le pourcentage des types de phrases dans le corpus STAC.

STAC est composé de 42 388 mots. Le tableau 4.9 présente le nombre d'occurrence de chaque catégorie grammaticale dans le corpus STAC.

Catégorie grammaticale	Nombre
Verbe	7 354
Adjectif	1 231
Nom	11 246
Pronom	3 626
Préposition	2 481
Particule	3 016
Conjonction	460
Numéral	1 141
Adverbe	2 914
Nom propre	1 151
Mot étranger	1 963
Mot incomplet	998
Pause	2 152
Interjection	1 159
Onomatopée	1 496
Total	**42 388**

TABLE 4.9 – Les catégories grammaticales figurant dans le corpus STAC

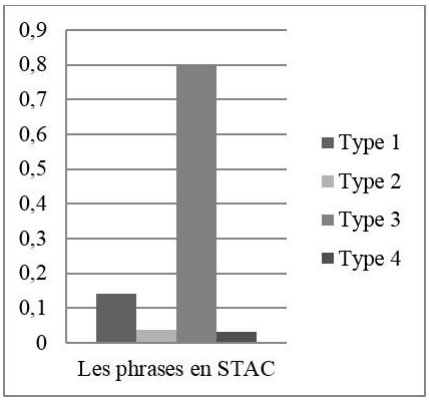

FIGURE 4.5 – Répartition des phrases du corpus STAC selon les types de classes.

4.4.2 Annotation des disfluences

L'annotation du corpus oral ne doit pas se limiter aux annotations habituelles de la langue écrite (l'étiquetage morphosyntaxique, l'étiquetage syntaxique, etc.). Un corpus oral est un ensemble de transcriptions de la parole qui présente une nouvelle forme de textes avec des spécificités qui constituent dans certains cas un défi pour l'analyse automatique de textes parlés [Dister *et al.* 2009]. Donc, l'annotation du corpus oral doit tenir en compte de ces spécificités, plus précisément, les « disfluences ».

En effet, les disfluences sont définies comme un phénomène se produisant fréquemment tout au long de la parole spontanée. Elles présentent une interruption de l'enchaînement normal du discours [Heeman & Allen 1994]. En fait, il existe différents types de disfluences : les pauses remplies, les répétitions, les mots incomplets, les expressions incomplètes, les autocorrections, etc. [Piu & Bove 2007]. Généralement, les disfluences sont combinées simultanément avec au moins deux phénomènes mentionnés ci-dessus. L'analyse de disfluences réalisée par [Shriberg 1994] a montré que le segment disfluent peut être divisé en trois régions

[Blache *et al.* 2010] :

— la reparandum : elle précède le point de rupture. Cette partie est obligatoire dans toutes les disfluences. Elle peut être sous forme d'un mot tronqué ou une phrase tronquée ;
— le point de rupture : c'est une partie facultative de la disfluence (pauses pleines, pauses silencieuses, etc.). Certains disfluences ne portent pas sur un événement spécifique.
— le reparans : c'est la partie qui suit le point de la rupture qui repère le reparandum. Dans certains cas de disfluences, cette partie est absente. Il s'agit d'une rupture syntagmatique ou syntaxique. Il n'y a ni complétude ni reprise.

4.4.2.1 Le système d'annotation

En se basant sur l'analyse de [Shriberg 1994], [Pallaud *et al.* 2008] ont défini un schéma d'annotation qui reflète cette structure de disfluences. Nous choisissons d'appliquer son guide d'annotation. Le guide d'annotation est développé afin d'annoter les disfluences à l'aide du logiciel Praat.

Le système d'annotation de [Pallaud *et al.* 2008] a défini les différents cas de disfluences qui peuvent exister. Les annotations sont inscrites sur une ou plusieurs tires (Tier en anglais) d'annotation de Praat dans le cas où les disfluences sont ou non enchâssées. Elles sont liées aux tokens en suivant la structure de la disfluence dans l'ordre chronologique : le Reparandum, le point de rupture et le Reparans. Chaque élément comporte plusieurs sortes d'informations qui vont être décrites et codées [Pallaud *et al.* 2008]. [Pallaud *et al.* 2008] ont distingué deux catégories de disfluences selon qu'elles seront suivies ou non d'une réparation (disfluences Réparées R et disfluences non réparées I). Le tableau 4.10 résume les différentes annotations utilisées pour schématiser les différentes parties des disfluences.

Compte tenu des spécificités du DT, en particulier l'alternance codique, nous avons remarqué la présence de certains cas de répétition disfluente avec alternance codique. La répétition avec alternance codique est un type de disfluence très spécifique pour le DT. La répétition disfluente est définie généralement par le fait qu'un locuteur hésite et répète le même mot ou expression, par exemple, l'expression (اعطيني آ ÷ وقت اعطيني, AʕTyny Ā ÷ AʕTyny wqt). Le locuteur répète le verbe (اعطيني, AʕTyny) de façon disfluente. Le même principe s'applique pour la répétition avec alternance codique. Le locuteur répète le même mot ou expression mais, cette fois-ci, il utilise une traduction du mot ou de l'expression. Par exemple, au niveau de l'expression « oui نعم » (oui, nʕm), le locuteur a traduit le mot (نعم, nʕm) en langue française. La répétition avec alternance codique peut être en traduisant le mot du DT vers la langue française ou vers l'ASM et vice versa.

Pour ce type de disfluences, nous avons ajouté une annotation qui décrit la répétition. Nous marquons ce type de disfluence par l'étiquette suivante : (rpcs) réparation en répétant le mot avec alternance codique. La figure 4.6 présente un exemple de disfluence avec répétition via une alternance codique.

La détection et l'annotation des disfluences sont faites de façon semi-automatique. Nous

Les disfluences suivies d'une réparation (R)	**Reparandum (R)**	***Élément affecté***	***Type de mot***	***Exemple***
		Syntagme (P)	Mot outil (tw)	اسمع (R,W,lw)
		Mot (W)	Mot lexical (lw)	
	Point de rupture (B)	Rien (B, no) Répétition du fragment (B, tr) Enoncé parenthétique (B, ps) Pause silencieuse (B, sp) Elément discursif (B, dc) Pause remplie (B, fp)		÷آ (B, fp)
	Reparans (RA)	***Position du Reparans***	***Fonctionnement du Reparans***	اسمع (RA,wr,co)
		Pas de reprise (nr)	Simple continuation (co)	
		Reprise au début du mot (wr)	Réparer le mot tronqué sans changement (wc)	
		Reprise au déterminant (dr)	Réparation en répétant (rp)	
		Reprise au début du syntagme (pr)	Réparer avec changement du mot tronqué (rc)	
		Autres types de reprise (or)	Réparation avec plusieurs changements (rm)	
Les disfluences non suivies d'une réparation (I)	**Reparandum (I)**	***Élément affecté***	***Type de mot***	***Exemple***
		Syntagme (P)	Mot outil (tw)	علاش (I, W, tw)
		Mot (W)	Mot lexical (lw)	
	Point de rupture (B)	Rien (B, no)		# (B, sp)
		Pause silencieuse (B, sp)		
		Pause remplie (B, fp)		
		Elément discursif (B, dc)		
		Répétition du fragment (B, tr)		
		Enoncé parenthétique (B, ps)		

TABLE 4.10 – Les annotations utilisées pour marquer les disfluences

détectons les hésitations et les répétitions couvrant les mots et les mots incomplets de façon automatique. La détection des autres types de disfluences et leur annotation sont réalisées manuellement. La figure 4.6 présente un exemple annoté extrait de notre corpus.

4.4.2.2 Statistiques

Nous présentons dans cette section quelques statistiques sur les disfluences dans notre corpus. Nous avons analysé les disfluences en se basant sur le schéma d'annotation utilisé. Ce schéma nous a permis d'étudier les disfluences selon leur type : avec réparation ou non. Aussi, il nous permet de cerner les différentes modalités de disfluences existantes dans le corpus STAC. Les proportions présentées dans le figure 4.7 sont calculées par rapport au nombre total de mots pour chaque locuteur dans le corpus.

4.5 Conclusion

La création d'une ressource textuelle fiable enrichie avec diverses annotations a fait l'objet de ce chapitre. D'abord, nous avons présenté la collection et la transcription de notre corpus STAC (Spoken Tunisian Arabic Corpus). Nous avons, ensuite, présenté une évaluation de notre ressource à travers le calcul d'accord des accords inter et intra annotateurs. Enfin, nous avons

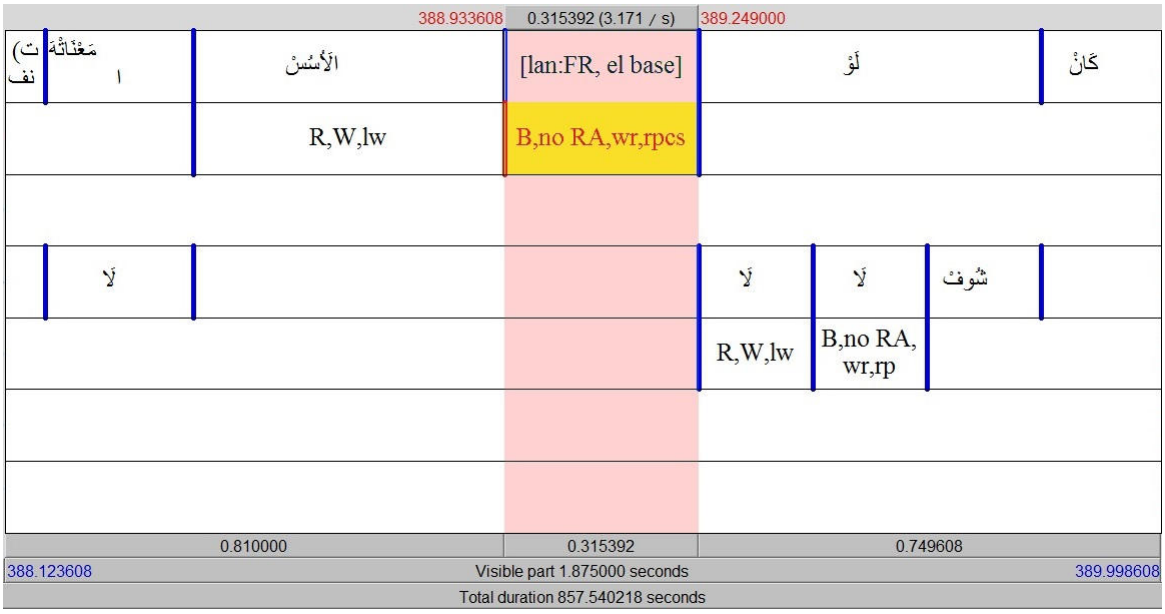

FIGURE 4.6 – Exemple d'une disfluence avec répétition via alternance codique.

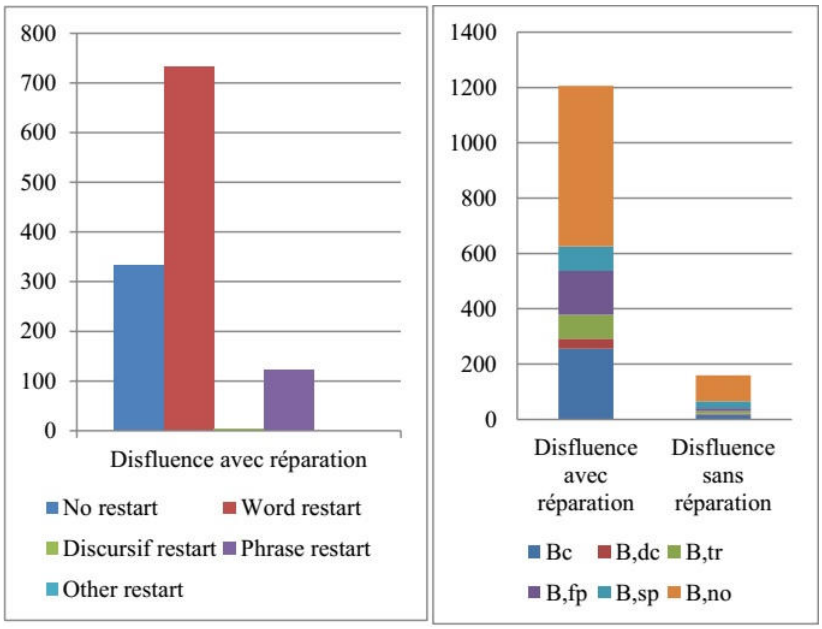

FIGURE 4.7 – Quelques statistiques concernant les disfluences.

discuté les différents types d'annotations que nous avons ajoutés pour enrichir notre corpus et le rendre plus utile pour de nombreuses applications du TAL.

La troisième partie de ce manuscrit de thèse sera consacrée pour présenter le développement des outils de traitement du DT. Nous présenterons dans le chapitre suivant la création d'un analyseur morphologique pour le DT.

Troisième partie

Création et adaptation d'outils pour le dialecte tunisien

CHAPITRE 5

Analyse morphologique du dialecte tunisien

Sommaire

5.1 Introduction

L'analyse morphologique est l'une des étapes fondamentales lors du traitement de la langue arabe. Elle détermine l'unité grammaticale minimale du mot en attribuant à chaque composant du mot ses caractéristiques morphologiques [Uchimoto *et al.* 2002]. Ces dernières années, les systèmes réalisés pour analyser les textes de la langue arabe sont devenus très performants. La bonne qualité de ces outils est due à l'énorme quantité de données étiquetées

traitant la forme écrite de la langue arabe. Cependant, leur qualité se dégrade lors de son application pour la langue parlée et plus précisément pour les dialectes arabes. Les cas d'échec sont dus aux différences entre les deux variétés de langues. Les textes de la langue parlée se caractérisent par plusieurs éléments qui se diffèrent de la langue écrite. Ces caractéristiques nécessitent des traitements particuliers.

Dans la littérature, deux approches ont été proposées pour analyser morphologiquement les dialectes arabes. Tout d'abord, l'adaptation des outils de l'ASM aux dialectes arabes a été proposée par plusieurs chercheurs en traitement automatique des dialectes arabes ([Almeman & Lee 2012], [Habash *et al.* 2013], etc.). D'autres ont préféré la création de nouveaux outils pour analyser morphologiquement les dialectes arabes [Habash *et al.* 2005], etc. En revanche, l'analyse morphologique de l'arabe parlé et plus précisément des dialectes arabes parlés n'a fait l'objet que de rares travaux en TAL. Dans ce chapitre, nous présenterons notre méthode proposée pour l'analyse morphologique du DT (la tokenisation et la détermination des différentes caractéristiques morphologiques).

Nous débutons ce chapitre par une étude linguistique du lexique du DT. Ensuite, nous proposons notre méthode pour l'adaptation d'un analyseur morphologique de l'ASM vers le DT. Au niveau de la section 5.4, nous présentons la création d'un lexique « racine - patron » pour le dialecte. La section 5.5 est consacrée pour présenter notre méthode pour la segmentation du DT. Puis, nous présentons l'intégration du lexique résultant dans un analyseur morphologique pour l'ASM. Enfin, nous clôturons le chapitre par la présentation de quelques expérimentations en discutant les résultats obtenus.

5.2 Étude du lexique du dialecte tunisien

Le DT comme la plupart des dialectes arabes est fondé sur le mariage entre plusieurs langues dont la langue de base est la langue arabe (ASM ou AC) [Baccouche 2011]. Les langues sur lesquelles se base le lexique dialectal varient selon le pays en question et ses circonstances historiques (*e.g.* les invasions islamiques) et géopolitiques (*e.g.* la colonisation, les frontières géographiques avec les autres pays, le commerce, etc.). Le français, l'espagnol, l'anglais, le turc, le berbère, l'italien etc. sont les principales langues composant le DT. Elles ont une nature morphologique et lexicale différente de celle de la langue arabe. Mais, elles sont bien intégrées dans le vocabulaire tunisien (et dans d'autres dialectes arabes) en suivant une stratégie de composition et/ou dérivation.

L'analyse lexicale du DT et la comparaison entre l'ASM et le DT, nous ont conduit à proposer une classification du vocabulaire tunisien suivant le processus de dérivation et composition. Dans cette section, nous présentons une classification du lexique du DT sur laquelle nous basons notre adaptation des outils pour le DT.

D'après les études menées sur le DT ([Baccouche 2009], [Mejri *et al.* 2009], [Mejri & Baccouche 2003], etc.), nous constatons que le lexique du DT est formé d'une partie partagée avec l'ASM, une partie composée des mots dialectaux (des mots qui sont

utilisés dans le dialecte et n'appartiennent ni au lexique de l'ASM ni au lexique de l'AC) et finalement un ensemble de mots issus des langues étrangères. Cette étude lexicale du DT nous a conduit à bien étudier chaque classe afin de proposer une classification plus fine.

Parmi les caractéristiques morphologiques du DT communes avec les dialectes arabes est la formation des mots via un processus de dérivation. Chaque unité lexicale est le résultat du croisement d'une « racine » et d'un « schème de dérivation ». Nous nous sommes basés sur l'étude de la forme dérivationnelle afin de classifier le lexique tunisien.

Une première étude porte sur la première classe du DT, c'est-à-dire, la partie commune avec l'ASM. L'étude de notre corpus STAC a montré que les mots appartenant à cette classe différaient de leurs équivalents de l'ASM au niveau phonologique. Les différences touchent principalement la prononciation des voyelles de l'ASM (la transformation des voyelles longues en leurs correspondantes courtes, etc.) et l'omission d'une ou plusieurs consonnes (la non-prononciation de Hamza, etc.).

Prenons l'exemple des mots du DT 1, 2, 3 et 4 du tableau 5.1. Ils sont dérivés respectivement des racines de l'ASM : (ب-ع-ل, l-ع-b), (ل-ك-أ, ’-k-l), (ب-ت-ك, k-t-b) et (أ-ر-ق, q-r-’). De même, leurs équivalents en ASM sont dérivés de ces mêmes racines. Nous remarquons que les schèmes de dérivation pour ces mots en DT ne sont qu'une modification soit à la forme consonantique soit à la forme vocalique. Au niveau des mots 1 et 3 du tableau 5.1, les schèmes de dérivation sont le résultat d'une simple modification à la forme vocalique. Les schèmes (يِفعَل yir_1r_2ar_3) et (فعَال r_1r_2aAr_3) sont issus respectivement de la substitution de la voyelle courte (اَ, a) par la voyelle (اِ, i) et la suppression des voyelles courtes (اُ , u), (اِ, i) et (اُ, u) situées respectivement à la dernière position du schème yar_1r_2ar_3u et à la deuxième position et la dernière position du schème $r_1r_2aAr_3$.

Pour les exemples 2 et 4 du tableau 5.1, les schèmes sont les résultats de la suppression et la substitution de la consonne Hamza. Dans le premier schème, Hamza est remplacée par l'infixe (ا, A) situé à la position finale du schème. Alors que pour le deuxième schème, Hamza est remplacée par l'infixe (ي, y).

Le verbe (أَكَلَ, Âakala) est un verbe qui commence par Hamza. En DT, le verbe se transforme en (كلَا , klaA) en ignorant la première consonne (Hamza) et les deux voyelles courtes qui les suivent. Cette idée s'applique généralement à d'autres verbes qui commencent par une Hamza. Par exemple, le verbe (أَخَذَ, Âxð, « *prendre* ») se transforme, aussi, en (خذَا, *xðaA*). Nous comparons la conjugaison à l'inaccompli d'un verbe, en ASM, qui commence par Hamza avec celle de son correspondant en DT. Nous remarquons qu'elles partagent les mêmes caractéristiques. En effet, Hamza est transformée en une voyelle longue. La chute de Hamza est, donc, une caractéristique de la conjugaison de ce type de verbes en DT. Ce principe s'applique pour tous les mots appartenant à cette classe du lexique.

Donc, le mot en DT est le résultat de la dérivation d'une racine de l'ASM suivant un schème de l'ASM modifié soit à la forme consonantique ou à la forme diacritique. En effet, il s'agit de la suppression, l'ajout ou la modification d'une lettre ou une voyelle dans le schème. Dans la

suite du rapport, nous utilisons l'acronyme « *C1* » pour désigner cette classe du lexique du DT.

N°	Racine	ASM		DT		Traduction en français
		Mot	Schème de dérivation	Mot	Schème de dérivation	
1	ب-ع-ل-ç-b	يَلعَبُ yalçabu	$yar_1r_2ar_3u$	يلعَب yilçab	$yir_1r_2ar_3$	*il joue*
2	ل-ك-أ-k-l	أَكَلَ Âakala	$r_1ar_2ar_3a$	كلَا klaA	r_2r_3aA	il a mangé
3	ب-ت-ك ktb	كِتَابُ kitaAbu~	$r_1ir_2aAr_3u$~	كتَاب ktaAb	$r_1r_2aAr_3$	livre
4	ق-ر-أ qr'	قِرَاءَةٌ qiraA'aħ~u	$r_1ir_2aAr_3aħ$~	قرَايَة qraAyaħ	$r_1r_2aAyaħ$	lecture

TABLE 5.1 – Quelques exemples de mots en DT et leurs équivalents en ASM

La deuxième partie du lexique du DT regroupe les mots dialectaux, qui peuvent être classés en deux types. Premièrement, il existe des mots qui sont parfois d'origine non arabes et qui sont intégrés dans le vocabulaire tunisien suivant le processus de dérivation. Ces mots ont perdu leurs formes de la langue origine. Nous citons, par exemple, le verbe (ينَقِز, *ynaqiz*, « *il saute* ») qui est dérivé de la racine tunisienne (ز-ق-ن, n-q-z) et le schème de dérivation (يفَعِل, $yr_1ar_2ir_3$). Cette classe est le résultat de l'application des racines spécifiques au dialecte en appliquant des schèmes de dérivation de l'ASM avec des modifications soit au niveau diacritique ou consonantique. Nous désignons cette classe par l'acronyme « *C2* ». Deuxièmement, le deuxième groupe des mots dialectaux collectionne les mots dérivés des racines de l'ASM mais en appliquant de schèmes qui n'appartiennent pas à l'ASM. Généralement, ces schèmes sont améliorés par l'ajout des affixes non arabes tels que (جي, *-jy*) et (يست, *-yst*). Par exemple, le mot (قَهوَاجِي, qahwaAjiy, « *un serveur* ») est dérivé de la racine ASM (ق-ه-و, q-h-w) et le schème (فَعلَا, $r_1ar_2r_3aA$) auquel il s'attache le suffixe (جِي, jiy). Ce mot est traduit en ASM avec le mot (نَادل, naAdil). Ce groupe de mots est désigné dans le reste du rapport par l'acronyme (*C3*).

La dernière classe (*C4*) est pour les mots qui sont issus de langues étrangères, spécifiquement le français. Par exemple, le mot (يدَوِش, ydawiš) est dérivé de la phrase française « *il prend une douche* ».

5.3 Adaptation d'un analyseur en arabe standard pour le dialecte

5.3.1 Motivation

Rappelons que notre objectif dans ce chapitre est de mettre au point un analyseur morphologique pour le DT en tirant profit des ressources existantes de la langue arabe (plus précisément l'ASM). L'adaptation des ressources d'une langue bien dotée en faveur d'autres langues peu dotées est une approche qui a été adoptée par plusieurs chercheurs en TAL ([Zeman & Resnik 2008] ; [Walther & Sagot 2010] ; [Lindström & Müürisep 2009] ; [Das & Petrov 2011] ; etc.). Cette approche a été exploitée pour créer des ressources linguistiques (dictionnaire, TreeBank, etc.) et développer des applications de TAL (analyseurs syntaxiques, analyseurs morphologiques, systèmes de reconnaissance vocaux, etc.). Elle a été,

aussi, adoptée par la communauté des chercheurs travaillant sur les dialectes arabes. L'exploitation des outils et ressources de la langue arabe standard pour analyser un dialecte arabe et le recours aux ressources d'un dialecte liée avec un autre (deux dialectes se partagent plusieurs traits *e.g.* le dialecte égyptien et levantin) sont les principales méthodes d'adaptation qui ont été exploitées pour le traitement automatique de l'AD.

Nous adoptons la même idée pour développer un analyseur morphologique pour le DT. Nous proposons, dans une première étape, de tirer profit d'un lexique « *root-pattern* » (racine-patron [1]) de l'ASM en vue de produire un lexique pour notre dialecte d'étude. Dans une seconde étape, nous incorporons ce lexique dans un analyseur morphologique de l'ASM en faisant les modifications nécessaires pour analyser le tunisien.

Notre choix pour cette approche est motivé par plusieurs raisons. D'abord, l'adaptation d'une ressource ou un outil de l'ASM pour le DT est la solution la plus simple à réaliser et aussi demande moins de ressources et de temps. Le DT est une langue peu dotée fortement liée à l'ASM. Il partage avec l'ASM plusieurs caractéristiques en plusieurs niveaux phonologiques, morphologiques, lexicaux et morphosyntaxiques. L'ASM est une langue écrite. Un certain nombre de ressources et outils ont été développés depuis quelques années, justifiant l'intérêt d'exploiter ces ressources pour créer d'autres ressources pour le DT.

Par ailleurs, pour motiver notre choix d'utilisation d'un lexique « racine-patron » pour créer un lexique pour le dialecte, nous fournissons les constatations dégagées de l'étude linguistique menée sur le DT. Cette étude montre que la morphologie des mots en DT partage les mêmes phénomènes linguistiques avec l'ASM, telles que la dérivation des verbes et des noms à partir des racines composées de trois ou quatre lettres, l'agglutination, etc. De même, l'étude statistique du lexique DT montre qu'une grande partie du lexique provient des racines de l'ASM et/ou l'application de schèmes de dérivation de l'ASM (partiellement ou totalement modifiés).

Pour bien comprendre la morphologie de DT, nous avons réalisé une extraction manuelle des schèmes et racines pour toutes les différentes unités lexicales contenant dans STAC [2]. Cette étude lexicale menée sur notre corpus, nous a aidé à dégager les caractéristiques suivantes :

— 52,25 % des unités lexicales en DT (verbes et noms) partagent avec l'ASM les racines et/ ou les schèmes de dérivation. Dans ce groupe, les schèmes sont le résultat d'une simple modification soit à la forme consonantique (par exemple, par l'ajout d'un infixe), soit à la forme vocalique (par exemple, la suppression des voyelles courtes dans plusieurs cas). D'où, à partir d'un lexique « racine - patron » de l'ASM, la génération d'un lexique pour le DT n'est qu'une simple tâche de projection des racines sur les schèmes de dérivation. Le seul obstacle dans ce cas est la définition de la liste des schèmes de dérivation modifiés ainsi que les racines de l'ASM partagées avec le dialecte.

— Le DT partage dans 37,10 % des unités lexicales des racines de l'ASM en suivant les schèmes de dérivation spécifiques au DT. L'étude de la morphologie des mots apparte-

1. Un patron est un schème de dérivation enrichi avec un ensemble de caractéristiques morphologiques : catégorie grammaticale, genre, nombre, etc.

2. La liste complète des schèmes extraites est présentée dans l'annexe.

nant à ces 37,10 % nous a montré qu'avec une liste on peut connaître les schèmes de dérivation et par conséquent, nous déterminons les caractéristiques correspondant au schème.

— 10,65 % des unités lexicales du DT partagent avec l'ASM leurs schèmes de dérivation. Dans ce cas, il suffit de détecter les racines du DT pour générer les noms et les verbes en DT.

Avec ces constatations, nous finissons par conclure que la génération d'un lexique se base essentiellement sur un ensemble de racines et de schèmes de dérivation pour le DT et pour l'ASM. Dans la littérature, plusieurs lexiques ont été proposés pour l'ASM parmi lesquels les lexiques « racine-patron » qui ont été intégrés dans des analyseurs morphologiques de l'ASM. Par conséquent, nous proposons de commencer par générer un lexique pour le DT en partant d'un lexique « racine-patron » de l'ASM et un corpus pour le DT. Ensuite, nous proposons d'intégrer ce lexique dans un analyseur morphologique.

5.3.2 Méthode proposée

Un analyseur morphologique prend en charge un ou des modules employant un lexique contenant des informations sur les racines, les stems et les patrons. Il y a toujours la possibilité de fournir ces informations manuellement. Toutefois, cela est très coûteux. Nous décrivons, d'abord, comment acquérir un lexique à partir d'un corpus brut pour le DT et un lexique de l'ASM. La génération de ce lexique fait l'appel à plusieurs traitements qui englobent des traitements manuels, automatiques et d'autres semi-automatiques pour générer des patrons pour le DT. L'intégration de cette nouvelle ressource dans un analyseur morphologique est la deuxième principale étape de notre proposition.

Nous rappelons que l'adaptation des analyseurs morphologiques de l'ASM en profit du dialecte a déjà fait ses preuves dans plusieurs travaux de recherche ([Salloum & Habash 2014], [Salloum & Habash 2011], [Afify *et al.* 2006], [Abuata & Al-Omari 2015], etc.) qui proposent de créer un lexique compatible avec le format de l'analyseur. Cette idée a été adoptée par [Hamdi 2015] pour adapter l'analyseur morphologique MAGEAD afin de traiter le DT. Notre méthode proposée pour la création d'un analyseur morphologique pour le DT n'est pas très loin de celles proposées dans l'état de l'art. Elle suggère de créer un lexique compatible avec l'analyseur Al-Khalil-ASM. Généralement, la création des lexiques pour l'adaptation des outils d'analyse morphologique est manuelle. Dans notre travail, nous essayons de réduire au plus les traitements manuelles pour créer les ressources lexicales. Nous proposons des méthodes capables de générer notre lexique.

Notre méthode proposée repose sur cinq étapes. La première étape de la méthode consiste, tout d'abord, à créer un lexique pour le DT. Ensuite, nous l'intégrons dans un analyseur morphologique pour l'ASM. La troisième étape est la mise à jour des règles de segmentation des mots ainsi que la liste des affixes et clitiques. Nous terminons par effectuer un ensemble de modifications pour que l'analyseur morphologique résultat puisse analyser le DT en prenant en

compte les différentes spécificités de l'oral. La figure 5.1 présente les étapes de notre méthode proposée.

Une des originalités de la méthode proposée est la possibilité de l'appliquer à plusieurs dialectes arabes en connaissant les points de ressemblances et de différences avec l'ASM. Ainsi, elle prend profit des ressources créées manuellement, automatiquement et de façon semi-automatique.

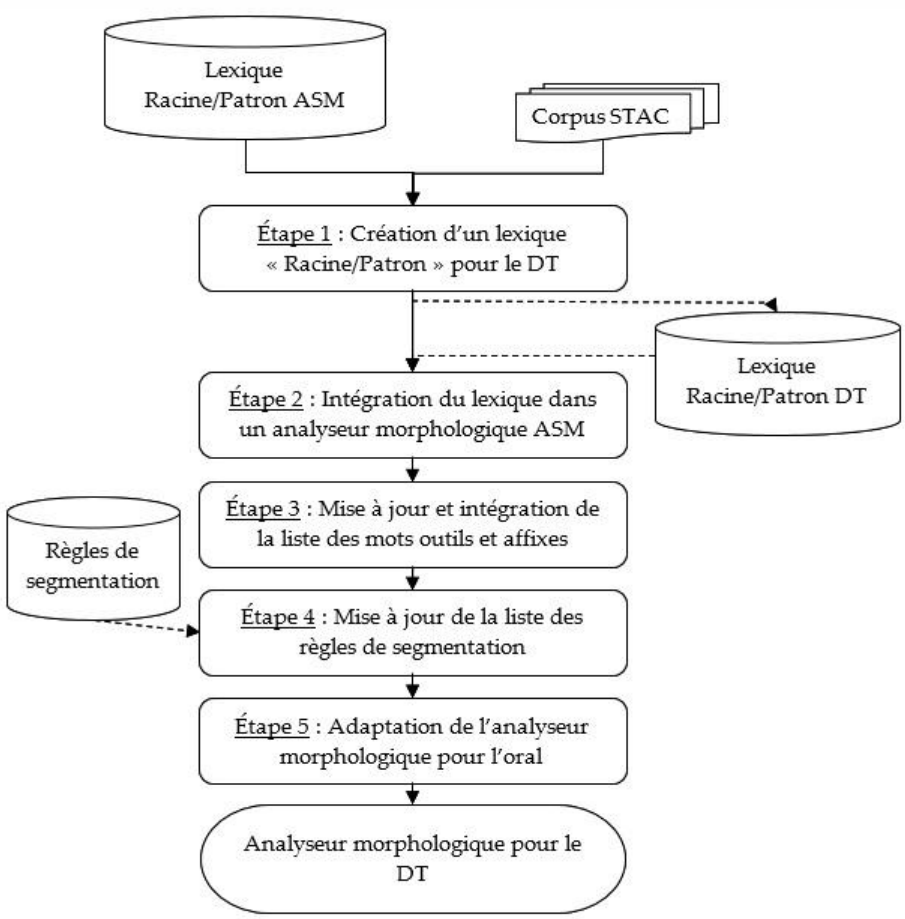

FIGURE 5.1 – Les étapes de l'adaptation d'un analyseur morphologique en faveur du DT.

5.4 Création d'un lexique « racine - patron » pour le dialecte tunisien

Un lexique contient souvent des informations sur les mots, les racines, les schèmes de dérivation, etc. Ces informations sont, généralement, procurées suite à des traitements manuels. Toutefois, ces actes sont très coûteux et ardus. Ainsi, nous proposons de créer un lexique pour le DT avec le minimum d'interventions manuelles. L'étude linguistique menée sur le DT nous a montré que les mots en DT étaient le résultat d'un moulage entre un ensemble de schèmes de dérivation et de racines. Par ailleurs, une grande partie de l'ensemble des racines de l'ASM est partagée avec le DT. De même, les schèmes de dérivation peuvent être déduits de leurs

équivalents en ASM. D'où, il est possible de les déduire via quelques traitements automatiques et manuels.

Notre méthode proposée pour la création d'un lexique est inspirée de celle proposée par [Hana 2008]. Hana a proposé une méthode hybride pour l'analyse morphologique de la langue tchèque : une langue peu dotée. Il a utilisé un ensemble de modules pour l'analyse morphologique ainsi que des ressources créées manuellement et des autres générées de façons non supervisées. Parmi les modules employés, [Hana 2008] a développé un devineur « *Guesser* ». Le rôle de ce devineur est de fournir les analyses possibles (les différents lemmes-racines-paradigmes) d'un mot inconnu en employant une petite liste de mots accompagnée d'informations sur les lemmes et les paradigmes correspondants. Il a proposé ainsi plusieurs analyses parmi lesquelles il existe des cas d'erreurs. Pour réduire le nombre d'analyses erronées, [Hana 2008] a utilisé un grand corpus. Il a supposé que l'analyse candidate a une probabilité d'être vraie si les lemmes et les paradigmes peuvent être extraits à partir d'autres parties de ce corpus.

Nous adoptons la même idée du devineur pour créer notre lexique pour le DT. Dans notre cas, le rôle du devineur est de dégager la liste des racines et des patrons pour les mots inconnus. En effet, nous créons, en premier lieu, une liste de racines et patrons pour le DT. Cette dernière englobe la partie en commun avec l'ASM. Ensuite, elle sera utilisée pour générer des racines et patrons du DT pour les mots inconnus.

Notre méthode repose sur trois principales étapes (*voir* figure 5.2) : la transformation des patrons de l'ASM en des patrons en DT, l'extraction des racines et des patrons du DT en se basant sur un corpus pour le DT, un lexique de l'ASM et le lexique résultat de la première étape. Enfin, nous finissons notre méthode par ajouter au lexique les mots étrangers et les mots dérivés des langues étrangères.

5.4.1 Transformation des patrons du dialecte tunisien à partir de l'arabe standard moderne

La première phase de notre méthode consiste à la détermination d'un ensemble de patrons du DT à partir des patrons apparentés de l'ASM en partant d'un lexique « racine-patron » de l'ASM. Pour se faire, nous essayons de mieux comprendre la morphologie des mots en DT et leurs équivalents en ASM. Nous commençons par prendre un exemple de deux verbes en ASM et nous essayons de chercher leurs équivalents en DT. Par exemple, les verbes (بدَأَ, bdÂ, « *il a commencé* ») et (ملَأَ, mlÂ, « *il a rempli* ») de l'ASM se transforment en (بدَا, bdA) et (ملَا, mlA) en DT. Ils sont dérivés respectivement des racines (ب-د-أ, b-d-') et (م-ل-أ, m-l-'), mais en suivant des schèmes de dérivation différentes. Ces verbes suivent le même modèle de dérivation en ASM (فَعَلَ, $r_1ar_2ar_3a$) ainsi qu'en DT (فعَا, r_1r_2aA), en gardant les mêmes caractéristiques morphologiques. La différence entre les deux schèmes est à la forme vocalique. De même, les deux noms (مَائِدَةٌ, maAŷidaħu, « *une table* ») et (فَائِدَةٌ, faAŷidaħu, « *intérêt* ») se dérivent suivant le schème de dérivation (فَاعِلَةٌ, $r_1aAr_2ir_3ap$) et les racines (م-أ-د, m-'-d) et (ف-أ-د, f-'-d).

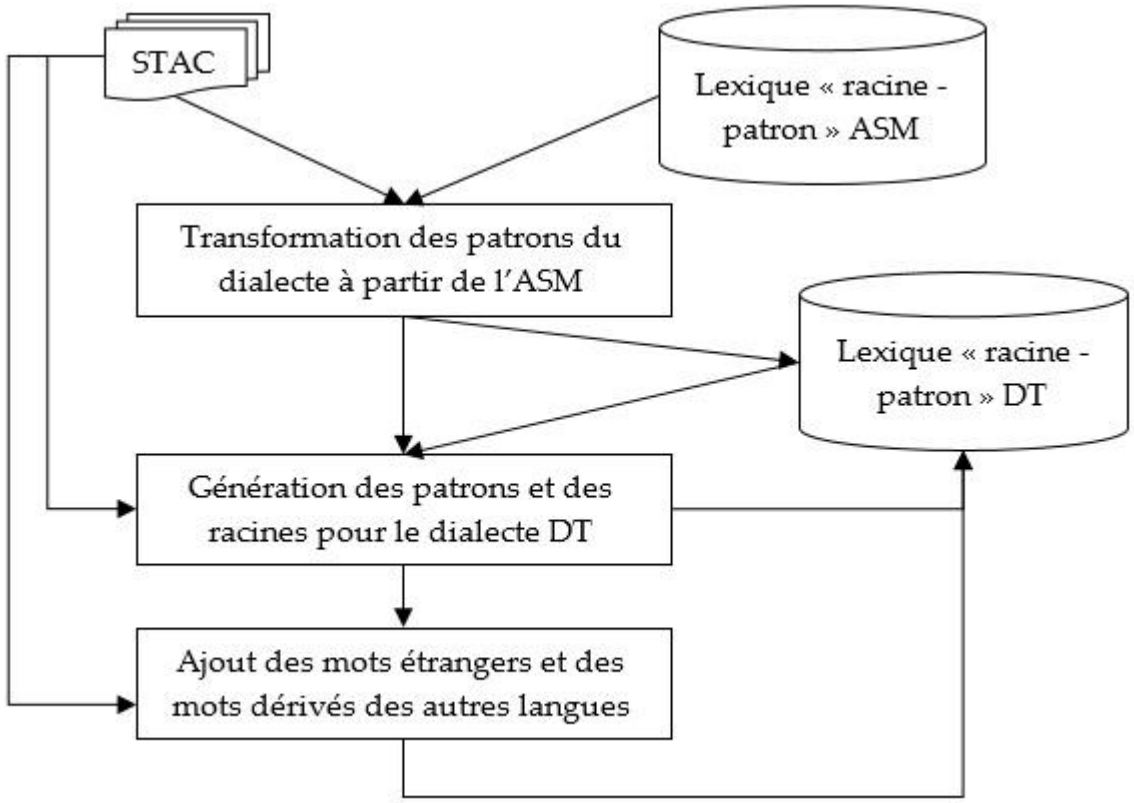

FIGURE 5.2 – Les étapes de la méthode de création d'un lexique pour le DT.

Leur équivalent en DT se dérive suivant le schème de dérivation (فَايلَة, r_1Ayr_3ap) et les mêmes racines de l'ASM (م-أ-د, m-'-d) et (ف-أ-د, f-'-d). Nous remarquons que les racines partagent des caractéristiques en commun. En fait, les deux racines contiennent Hamza (أ).

Les racines arabes peuvent être classées selon plusieurs critères : le nombre de lettres ajoutées à la racine pour former le mot, le nombre de lettres de la racine, la présence et la position de lettres défectueuses, etc. L'étude de quelques exemples des mots en DT et en ASM a montré que les mots qui se dérivaient de racines contenant des lettres défectueuses (ou des lettres saines) suivent les mêmes schèmes de dérivation en ASM et en DT. Notre étude est bien validée par l'étude de la morphologie du DT réalisée par le linguiste [Ouerhani 2009]. En effet, [Ouerhani 2009] a montré que les verbes appartenant à la classe *Mahmouz* (les racines contenant la lettre (أ, ')) partagent les mêmes schèmes de dérivation et les mêmes caractéristiques et suivent les mêmes règles, lorsqu'elles se transforment en DT. Ce raisonnement est également applicable à d'autres classes de racines. Nous prenons aussi l'exemple de la classe « défectueuse [3] ». Les deux verbes de l'ASM (مَشيَ, mašiya, « *il a marché* ») et (بَكِي, bakiya, « *il a pleuré* ») sont dérivés suivant le schème (فَعِلَ, $r_1ar_2ir_3a$). En DT, ces deux verbes se transforment respectivement en (مشَى, m.šaỳ) et (بكَى, bkaý). Nous remarquons qu'ils suivent le

3. Cette classe regroupe les racines qui se termine par une lettre défectueuse (ي, y) ou (و, w).

même schème de dérivation (فعَى, $r_1r_2$aý).

Pour dégager les schèmes de dérivation du DT, nous classons les racines selon la position et le nombre des lettres défectueuses dans la racine. Pour chaque classe, nous étudions les différentes dérivations possibles en appliquant les patrons correspondants à quelques exemples de racines. Ensuite, nous identifions, pour chaque dérivation, les schèmes qui coïncident avec ces unités lexicales en DT. Dans certains cas, nous notons que la sémantique des mots change lors du passage de l'ASM vers le DT. Enfin, nous appliquons ces schèmes pour générer des patrons en DT.

La figure 5.3 résume les étapes de transformation des patrons du dialecte à partir de l'ASM.

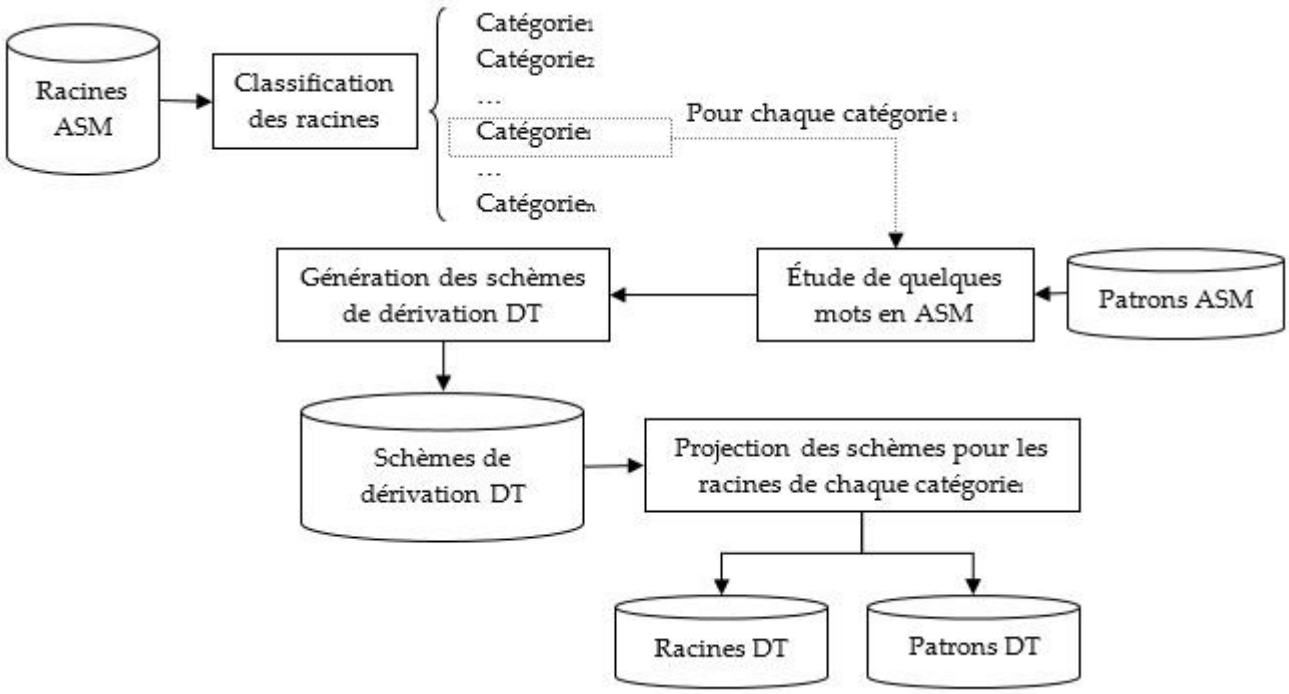

FIGURE 5.3 – Les étapes de transformation des patrons du DT à partir de l'ASM.

Nous prenons l'exemple de la racine (ك-ت-ب, k-t-b). Un ensemble de schèmes de dérivation est associé à cette racine. Le schème (فَعَلَة, $r_1ar_2ar_3a\hbar$) de l'ASM n'a pas d'équivalent en DT. Par contre, le schème (فَعَلَ, $r_1ar_2ar_3a$) de l'ASM se transforme en (فعِل, $r_1r_2ir_3$) en DT.

La figure 5.4 montre un extrait du fichier correspondant à la liste des schèmes de l'ASM et leurs équivalents en DT pour la classe Mahmouz.

À partir du lexique de l'ASM, nous avons pu dériver à partir de 13 271 schèmes de dérivation nominaux et 4 296 schèmes verbaux de l'ASM, respectivement 320 patrons nominaux et 331 patrons verbaux pour le DT. Au niveau de cette étape, nous avons retenu uniquement 1 213 racines en partant de 1 471 racines de l'ASM.

```
<?xml version="1.0" encoding="windows-1256" ?>
<verbRules>
    <rules ncg ="1" type="مم"><!--ماض مبني للمعلوم-->
        <rule diacSource="فَعَلْتُ" diacCible="غلِيتْ"/>
        <rule diacSource="فَعَّلْتُ" diacCible="وْعِلْتْ"/>
    </rules>
    <rules ncg ="3" type="مم"><!--ماض مبني للمعلوم-->
        <rule diacSource="فَعَلْتَ" diacCible="غلِيتْ"/>
        <rule diacSource="فَعَّلْتَ" diacCible="وْعِلْتْ"/>
    </rules>
    <rules ncg ="6" type="مم"><!--ماض مبني للمعلوم-->
        <rule diacSource="فَعَلْتُمْ" diacCible="غلِيتُو ا"/>
        <rule diacSource="فَعَّلْتُمْ" diacCible="وَعِلْتُو ا"/>
    </rules>
    <rules ncg ="2" type="مم"><!--ماض مبني للمعلوم-->
        <rule diacSource="فَعَلْنَا" diacCible="غلِينَا"/>
        <rule diacSource="فَعَلْنَا" diacCible="غلِينَا"/>
        <rule diacSource="فَعَّلْنَا" diacCible="وْعِلْنَا"/>
        <rule diacSource="فَعَّلْنَا" diacCible="وْعِلْنَا"/>
    </rules>
```

FIGURE 5.4 – Un extrait du fichier correspondant à la liste des schèmes de l'ASM et leurs équivalents en DT pour la classe Mahmouz. Les attributs *diacSource* et *diacCible* présentent respectivement les schèmes de dérivation de l'ASM et du DT. L'attribut *ncg* donne des informations sur le genre et le nombre du schème et l'attribut *type* montre la voix et l'aspect

5.4.2 Génération des patrons et extraction des racines pour le dialecte tunisien

La deuxième étape de la création d'un lexique pour le DT est la génération des patrons et l'extraction des racines pour le DT (*voir figure 5.5*). La première phase consiste à extraire des racines spécifiques au DT. Le but de cette phase est de couvrir la deuxième classe (*C2*) (voir section 5.2). Nous essayons d'extraire les racines à partir d'un corpus d'apprentissage qui contient des mots spécifiques au DT tels que le verbe (نقز, naqiz, « *il a sauté* ») et le nom (كرهبة, karhbaħ, « *une voiture* »).

Nous analysons morphologiquement tous les mots du corpus en utilisant le lexique généré de la première étape. Si l'analyse n'arrive pas à attribuer une racine et un patron pour les mots, alors ces mots passent par la phase d'extraction de racines. Nous essayons d'extraire les racines de ces mots qui coïncident avec les schèmes de dérivation issus de la première étape. Au niveau de cette étape, nous remarquons que le module d'extraction peut générer plusieurs racines pour un mot donné. Nous constatons que les patrons verbaux qui s'appliquent aux racines quadrilatérales engendrent, dans certains cas, des erreurs lors de la génération des racines pour ces verbes. L'ajout au lexique avec l'ensemble de racines générées affecte l'analyse morphologique en utilisant un lexique pareil. Les erreurs d'extraction des racines pour les mots en DT augmentent le nombre d'analyses morphologiques. De ce fait, le taux d'ambiguïté augmente.

Pour remédier à ce problème, nous proposons dans une première étape de classer les mots souhaitant extraire ses racines selon la possibilité que ces unités lexicales partagent la même racine. Pour se faire, nous divisons la liste de mots non reconnus en deux groupes : un pour

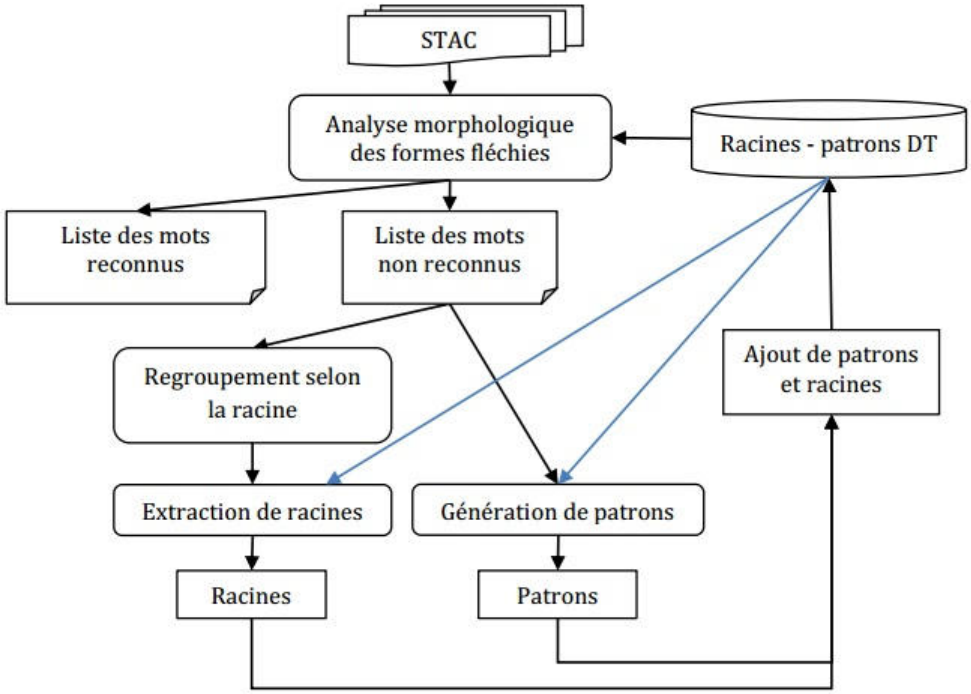

FIGURE 5.5 – Processus de génération de patrons et d'extraction de racines pour le DT.

les verbes et un pour les noms. Pour chacun, nous essayons de regrouper les unités lexicales selon la racine.

Le regroupement est basé sur le calcul de la similarité entre deux chaines de caractères. Nous cherchons pour chaque unité lexicale l'ensemble des caractères partagées avec toutes les autres unités. Nous utilisons le coefficient de recouvrement[4] (*Overlap coefficient*) : une mesure de similarité entre deux chaînes de caractères. C'est une mesure liée à l'index de Jaccard [Jaccard 1912] qui mesure le recouvrement de deux ensembles (*voir équation* (5.1)). Dans notre cas les deux ensembles sont les caractères formant les deux mots. Nous cherchons le coefficient de recouvrement de chaque paire d'unités lexicales. Pendant le calcul de ce coefficient, nous respectons l'ordre des lettres dans les deux unités lexicales. Les mots qui ont un coefficient compris entre 1 et 0,5, nous les regroupons dans un ensemble. Pour chaque ensemble, nous extrayons les racines possibles de chaque mot en utilisant la liste des schèmes pour les noms et les verbes. La racine la plus fréquente dans cet ensemble est retenue et ajoutée au lexique.

$$overlap\,(X, Y) = \frac{|X \cap Y|}{\min(|X|, |Y|)} \tag{5.1}$$

Nous prenons l'exemple des deux mots suivants : (يَنْقِز, ynaqiz, « *il saute* ») et (نَقْزُوا, naq-zuwA, « *ils ont sauté* »). Ces deux mots ne sont pas reconnus en utilisant le lexique résultant de la première étape de notre méthode.

En utilisant les deux patrons verbaux (يَفعِل, $yar_1r_2ir_3$) et (فَعلِل, $r_1ar_2r_3ir_4$), nous extrayons

4. https ://en.wikipedia.org/wiki/Overlap_coefficient

Ensemble	Les unités lexicales partagent la même racine
E1	ينَقِز ynaqiz , تنَقِز tnaqiz, ينقزوَا ynqzwA, نقز nqz
E2	كرهبة karhbaħ, كرَاهب kraAhib

TABLE 5.2 – Exemple de groupements de mots

respectivement les racines suivantes : (ن-ق-ز, *n-q-z*) et (ي-ن-ق-ز, *y-n-q-z*) pour le premier mot. De même, en utilisant le schème verbal (فَعلُوا, $r_1ar_2r_3uwA$), nous extrayons la racine (ن-ق-ز, *n-q-z*). Cette dernière est la plus fréquente dans cet ensemble. Donc, nous considérons que la racine de mots (ينَقِز, ynaqiz, « *il saute* ») et (نَقزوَا, naqzwA, « *ils ont sauté* ») est (ن-ق-ز, *n-q-z*). De ce fait, nous attribuons aux mots (ينَقِز, ynaqiz, « *il saute* ») et (نَقزوَا, naqzwA, « *ils ont sauté* ») respectivement les schèmes de dérivation (يفَعِل, $yr_1ar_2ir_3$) et (فَعلُوا, $r_1ar_2r_3uwA$) avec toutes leurs caractéristiques morphologiques.

Pour la liste des racines générées par notre module avec une fréquence d'apparition inférieure à 2, nous procédons à choisir manuellement les bonnes racines. Ainsi, une validation, par un expert, est effectuée pour garantir l'ajout des racines correctes à la base lexicale.

Dans la deuxième phase, nous suivons la même démarche de la première phase. Mais, cette fois-ci, nous cherchons un schème de dérivation pour les mots inconnus en utilisant la liste des racines de DT procurées dans la première étape de notre méthode. Ainsi, nous cherchons les schèmes de dérivation qui sont généralement spécifiques au DT comme (فَعلَاجِي, $r_1ar_2r_3aAjiy$).

Nous prenons l'exemple du mot (قَهوَاجِي, qahwaAjy). L'analyse morphologique en utilisant les lexiques résultant des étapes précédentes échoue à analyser ce mot vu l'absence d'un patron compatible avec la morphologie du mot. Le passage par la première phase de cette étape n'aboutit pas à générer une racine pour cette forme suite à l'absence d'un schème de dérivation qui coïncide avec cette forme orthographique. Le mot (قهوَاجي, qahwaAjy) passe par le module de génération de patrons. On essaie de chercher une racine de l'ASM qui coïncide avec ce mot. Puis, nous générons le ou les schèmes de dérivation qui peuvent être appliqués à ce mot. Dans cette phase, une validation par un expert est effectuée pour choisir le schème de dérivation correct et pour attribuer les différentes caractéristiques morphologiques tels que la catégorie grammaticale, le genre, le nombre, etc. Le résultat de cette phase est un ensemble de racines et patrons qui couvrent la troisième classe du lexique DT (*C3*).

En guise de conclusion, nous pouvons résumer les étapes de la génération des patrons et extraction des racines comme suit (*voir figure 5.5*) :

— Une étape d'analyse morphologique des différents mots d'un corpus.
— Une étape d'extraction des racines et/ou génération de partons est effectuée pour les mots non reconnus.
— Une étape d'ajout à la base lexicale qui est validée par des calculs statistiques (fréquence d'apparition) ainsi que des validations par un expert humain.

Notons que nous avons amélioré notre lexique par l'ajout d'environ 333 et 345 unités lexicales respectivement pour la première et la deuxième phase.

5.4.3 Enrichissement du lexique

La première et la deuxième étape de notre méthode pour la création d'un lexique pour le DT ne peuvent pas couvrir la totalité du DT. Selon notre étude lexicale du DT, ces deux étapes ne couvrent pas les mots issus des langues étrangères. Par exemple, les mots (سبيطَارَات, sbiyTaAraAt, « *les hôpitaux* ») et (بلَاصة, blaASaħ, « *une place* ») ne sont pas le résultats d'une dérivation en partant d'une racine et un schème de dérivation. Pour garantir une large couverture de notre lexique pour le vocabulaire du DT, nous avons amélioré notre lexique « racine-patron » par une liste regroupant les mots étrangers intégrés dans le DT. Cette liste couvre généralement les mots qui appartiennent à la classe (*C4*) de notre classification.

Au niveau de ce lexique, nous avons rassemblé toutes les formes existantes dans le corpus STAC. À chaque mot trouvé, nous cherchons sa forme en pluriel et/ou en singulier et aussi, sa forme en féminin et masculin. De même, pour les verbes, nous ajoutons à chaque verbe les différentes conjugaisons possibles.

Ce lexique est enrichi par les caractéristiques morphologiques de chaque entrée lexicale. Nous identifions la catégorie grammaticale, le genre, le nombre et la liste des clitiques qui peuvent être attachés, etc. Nous avons pu collecter 500 mots différents pour ce lexique.

À la liste des mots d'origine non arabe, nous avons aussi enrichi notre lexique par une liste des mots-outils du DT. Le DT garde dans certains cas les mots-outils de l'ASM et dans d'autres cas, il ajoute d'autres termes. La collecte de cette liste des mots-outils est basée principalement sur notre corpus STAC et aussi sur le lexique de l'ASM. Pour chaque mot outil, nous cherchons sa (ou ses) traductions possibles en DT. Les mots-outils de l'ASM se transforment en des mots-outils ou des clitiques en DT. Dans certains cas, on ne trouve pas d'équivalent pour le mot outil en dialecte. Par exemple, l'adverbe interrogatif (لمَاذَا, lmAðA) de l'ASM est traduit en DT en deux adverbes (علَاش, ʿlAš) et (علَاه , ʿlAh). Dans d'autres cas, le mot outil garde la même forme de l'ASM. Par exemple, la préposition (من, mn) est utilisée en ASM et aussi en DT, mais, elle se transforme en un clitique (م, m-). Par contre, la particule de restriction (حيثمَا, Hyθ mA) n'a pas de traduction en DT. Le tableau 5.3 présente quelques exemples de mots-outils pour l'ASM et leurs équivalents en DT[5].

Finalement, à partir de 89 mots-outils extraits de notre lexique de l'ASM, nous avons obtenu 56 mots-outils et 5 clitiques pour le DT. De même, à partir de notre corpus STAC, nous avons collecté 200 mots-outils qui n'ont pas d'équivalents en ASM.

5.5 Segmentation du dialecte tunisien

La tokenisation pour la langue arabe consiste à définir les frontières des mots et également les informations concernant les tokens qui les composent (le stem et les clitiques) [Attia 2009]. C'est une tâche non-triviale. Elle est étroitement liée à l'analyse morphologique.

5. La liste complète est présentée dans l'annexe 6.9

POS	ASM	DT	Traduction
Particule de futur	سوف swf	بَاش, bAš	*je ferai*
Adverbe interrogatif	لمَاذَا lmAðA	علَاش ىlAš	*pourquoi*
		علَاه ىlAh	
Adverbe	فقط fqT	كهو khw	*seulement*
Particule de restriction	حيثمَا HyθmA	-	*partout*
Préposition	من mn	من mn	*de*
		م m-	
Pronom démonstratif	-	أَوكة	*Là*
Pronom personnel	-	رَاهم	*Ils sont*

TABLE 5.3 – Exemples des mots-outils de l'ASM et leurs équivalents en DT

La première étape de la tokenisation consiste à définir la liste des clitiques et des affixes. Nous commençons par chercher l'équivalent de chaque clitique et proclitique en DT. Nous traduisons les clitiques en leurs équivalents en DT. Quelques clitiques de l'ASM sont traduits en DT en des mots-outils. Par exemple, le préfixe du futur de l'ASM (سـ, s-) « *je ferai* » est converti en mot outil (بَاش, bAš). Toutefois, le préfixe d'interrogation (أَ) est transformé en un suffixe (شي, *-šy*) « *quoi* ». Nous obtenons ainsi 56 suffixes et 5 préfixes pour le DT. Le tableau 5.4 présente quelques exemples d'affixes et de clitiques de l'ASM et leurs équivalents en DT

	Proclitiques /Préfixes		Suffixes/Enclitiques	
	ASM	DT	ASM	DT
Préfixe de futur	سـ	Mot outil : بَاش bAš	-	-
Préfixe d'interrogation	أَ Â	-	-	شي -šy
Suffixe de négation	-	-	-	ش -š

TABLE 5.4 – Exemples d'affixes et de clitiques de l'ASM et leurs équivalents en DT.

La deuxième étape de la tokenisation du DT est l'identification des composants de mots. Nous avons suivi une méthode à base des indicateurs pour segmenter les mots. Nous développons un ensemble de règles pour segmenter les mots en proclitique(s)+préfixe(s)+stem+suffixe(s)+enclitique(s). Nous définissons pour chaque clitique et affixe les catégories grammaticales auxquelles ils s'attachent. Nous identifions des clitiques qui s'attachent aux mots-outils (TW), aux verbes (V), aux noms (N) et aux noms propres (PN). La classe (C) est pour les affixes et clitiques communs. Ils s'attachent avec les noms, les verbes, les mots-outils et les noms propres. Le tableau 5.5 identifie les différents affixes et clitiques et les catégories grammaticales auxquelles ils s'attachent.

Le tableau 5.6 présente des exemples de combinaisons possibles de clitiques et d'affixes.

L'algorithme de tokenisation commence par chercher la chaîne de caractères correspondant à chaque combinaison de proclitiques et préfixes dans le mot à segmenter. Puis, il cherche la chaîne de caractères correspondante aux clitiques de fin. Le processus de segmentation s'arrête si le nombre de caractères du stem (ou du mot) est inférieur ou égal à 2.

Proclitique(s) + préfixe(s)	Classe	Suffixe(s) + enclitique(s)	Classe
ال Al, وَال wAl, ب b, م m, ف f, ل l	**TW**	همشي hmšy, هش hš, كش kš, كمش kmš, كشي kšy, هَاش hAš, همش hmš, نَاش nAš	**TW**
و w, وك wk, ك k	**C**	ك k, كم km, نَا nA, ني ny	**C**
		ه h, هم hm, هَا hA	
ب b, وك wk, و wb, ول wl, ك k, ل l	**PN**	ي y, يَا yA	**N**
وَال wAl, ال Al	**N**	شي šy, ني ny, نيش nyš	**V**
ب b, ل l, ك k		نَاهَا nAhA, نَاهم nAhm, نَاهش nAhš, نَاهَاش nAhAš, نَاهمش nAhmš	
ول wl, و wb, فب fb, لب lb, فل fl, وك wk		ش š, هش hš, كش kš, كمش kmš, كشي kšy, هَاش hAš, همش hmš, نَاش nAš	
ولّل wll, وعَال wʕAl, ومَال wmAl, وكَال wkAl, و بَال wbAl, بهَال bhAl, وهَال whAl, ولّل wll, هَال hAl, بَال bAl, مَال mAl, لّل ll, عَال ʕAl, كَال kAl		وه wh, وهش whš, ونِي wny, ونَا wnA, وهَا whA, وهم whm, ونَاش wnAš, ونيش wnyš, وهَاش wnAš, وهمش whmš	

TABLE 5.5 – Les affixes et clitiques du DT et les catégories grammaticales auxquelles ils s'attachent.

Verbe	Nom	Nom Propre	Mot outil
V+ك	N+فب	PN+و	TW+و
V+وك	N+و	PN+ل	TW+هم
V+وك+هَا	N+و+ي	PN+ول	TW+هم+و
V+و	N+ل+هم	PN+ك	TW+همشي+و
V+و+هَا	N+ومَال	PN+وك	
V+و+وهَا	N+عَال		
V+و+كمشي	N+نَا		

TABLE 5.6 – Quelques combinaisons de clitiques et d'affixes possibles.

La figure 5.6 présente les étapes de segmentation des mots en DT. L'algorithme de tokenisation propose pour chaque mot les segmentations suivantes :

— Stem,

— Proclitiques + Préfixes + Stem,

— Stem + Suffixes + Enclitiques,

— Proclitiques + Préfixes + Stem + Suffixes + Enclitiques.

La recherche effectuée par cet algorithme respecte bien la classe des clitiques de début et de fin. Par exemple, si on trouve des proclitiques et/ou des préfixes appartenant à la classe N (les noms), donc la recherche de suffixes et d'enclitiques sera limitée uniquement aux clitiques correspondants à la classe N.

Prenons l'exemple du mot (لعبنَاهَاش, lʕbnAhAš). Le processus de segmentation essaie en premier lieu de chercher la ou les clitiques de début. Il propose comme une première segmentation : (ل+عبنَاهَاش, l+ʕbnAhAš). Selon cette segmentation, le stem est traité comme un nom ou un nom propre. En deuxième lieu, le processus cherche la ou les clitiques de fin. Ainsi, il propose deux autres segmentations : (لعب+نَاهَاش, lʕb+nAhAš) et (لعبنَا+هَاش, lʕbnA+hAš).

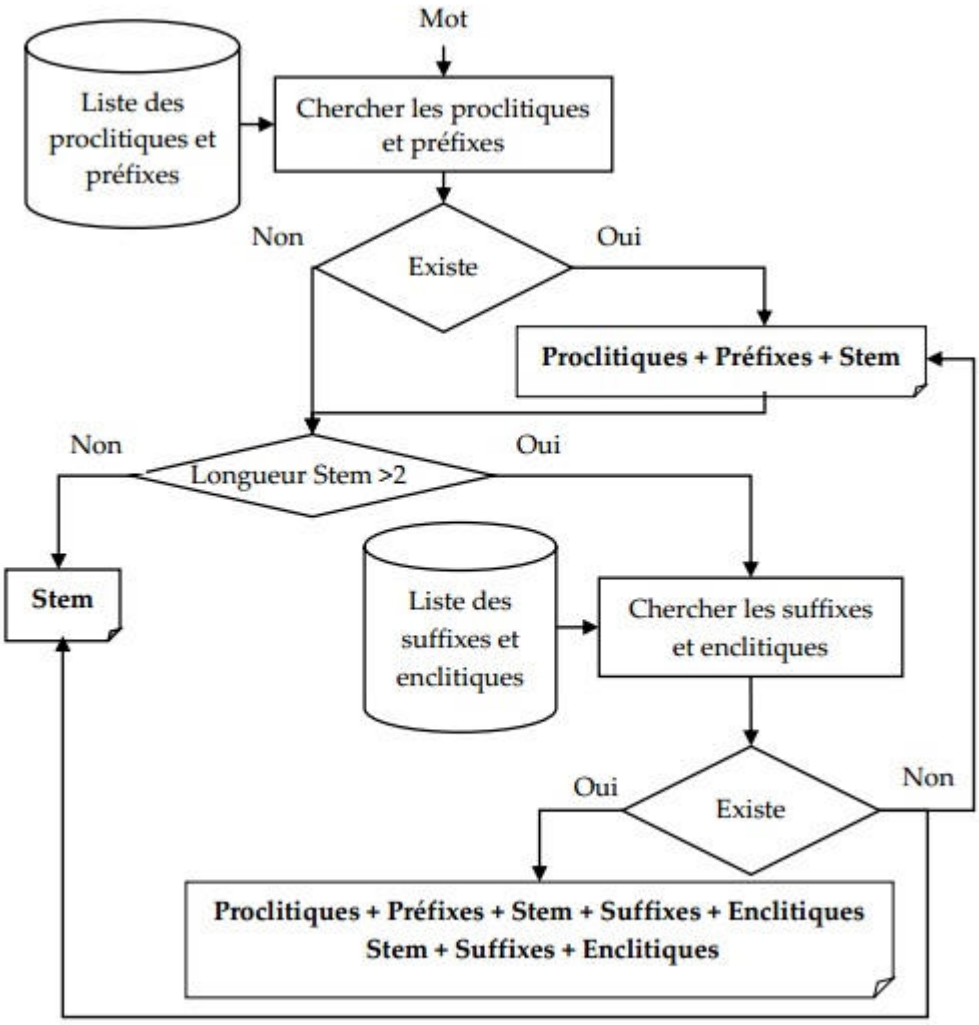

FIGURE 5.6 – Les étapes de segmentation des mots en DT.

Au niveau de ces deux segmentations, le stem est traité comme un verbe.

5.6 Intégration du lexique à l'analyseur morphologique Al-Khalil-ASM

5.6.1 Présentation de l'analyseur Al-Khalil-ASM

Le système Al-Khalil-ASM [Boudlal *et al.* 2010] est un analyseur morphologique de la langue arabe écrit en java mis en ligne en 2010. Il est considéré comme le meilleur système d'analyse morphologique disponible pour la langue arabe. En fait, Al-Khalil-ASM a remporté la première place parmi treize analyseurs morphologiques arabes, dans une compétition organisée par l'ALECSO[6] (Organisation arabe pour l'éducation, la culture et les sciences) en 2010 [Boudlal *et al.* 2010]. L'objectif de cet analyseur est de donner une description morpho-

6. The Arab League Educational, Cultural Scientific Organization.

logique détaillée de chaque mot à savoir : le mot sans voyelles courtes, le préfixe, le stem, le type, le schème, la racine, la catégorie grammaticale et le suffixe. Pour générer cette analyse, Al-Khalil-ASM utilise un ensemble de règles enregistrées dans des fichiers XML séparés, mais qui sont complémentaires d'une façon que l'un fait appel à l'autre au cours de l'exécution pour analyser morphologiquement un mot donné.

La ressource lexicale utilisée par Al-Khalil-ASM est composée de plusieurs fichiers qui regroupent les patrons (les schèmes de dérivation avec leurs caractéristiques morphologiques) avec et sans voyelles courtes pour les noms ainsi que les verbes, les racines de mots, les affixes et les noms propres. Les étapes de l'analyse sont les suivantes : une étape de prétraitement, l'élimination de voyelles courtes, la segmentation (identification des éléments de chaque mot : les proclitiques + stem + les enclitiques) et l'analyse du stem [Boudlal *et al.* 2010]. L'architecture de l'analyseur morphologique arabe Al-Khalil-ASM est illustrée par la figure 5.9.

Nous présentons, dans ce qui suit, les différentes structures des fichiers XML manipulés par Al-Khalil-ASM pour générer une analyse morphologique complète d'un mot. Al-Khalil-ASM classifie la liste des racines suivant la première lettre de chaque racine.

La liste des patrons est stockée dans des fichiers XML suivant le nombre de caractères de la forme vocalique qui lui correspond. Les patrons sont stockés dans huit fichiers : le premier fichier contient les patrons dont leurs schèmes de dérivation sont composés de deux lettres et le dernier fichier contient les patrons dont leurs schèmes de dérivation sont composés de neuf lettres. Pour chaque patron verbal, on identifie les caractéristiques suivantes : sa forme canonique, sa forme vocalique, son temps de conjugaison, sa voix, son cas (nominatif, accusatif ou génitif), le pronom avec lequel le verbe est conjugué, sa transitivité du verbe et son augmentation (le verbe est augmenté par une ou plusieurs lettres ou non). De même, on identifie pour les caractéristiques suivantes pour les patrons nominaux : sa forme canonique, sa forme vocalique, son type (adverbe, adjectif, etc.), son cas (nom défini ou non), son nombre (singulier, duel ou pluriel) et son genre (féminin ou masculin). La figure 5.7 et figure 5.8 présentent respectivement deux exemples de patrons nominaux et verbaux.

```
<patterns>
    <pattern id="10" diac="آعَاءِيّ" canonic="فاعَاءِيّ" type="اسم" cas="نك" ncg="13"/>
    <pattern id="1000" diac="أفِعَاءَ" canonic="أفِعَاءَ" type="جا" cas="إض" ncg="18"/>
    <pattern id="10002" diac="فَعَّالاً" canonic="فَعَّالاً" type="مفا" cas="نك" ncg="7"/>
    <pattern id="10003" diac="فَعَّالَةٌ" canonic="فَعَّالَةٌ" type="جا" cas="نك" ncg="8"/>
    <pattern id="10004" diac="فَعَّالَةٌ" canonic="فَعَّالَةٌ" type="جا" cas="نك" ncg="2"/>
    <pattern id="10005" diac="فَعَّالَةٌ" canonic="فَعَّالَةٌ" type="جا" cas="نك" ncg="5"/>
```

FIGURE 5.7 – Exemple de patrons de dérivation nominaux.

```
<patterns>
    <pattern id="10" diac="آعَلَ" canonic="فاعَلَ" type="فعم" aug="جر" cas="ن" ncg="1" trans="ك"/>
    <pattern id="1014" diac="أفْلَ" canonic="أفْلَ" type="فعم" aug="جر" cas="ج" ncg="1" trans="ك"/>
    <pattern id="1015" diac="أفْلَ" canonic="أفْلَ" type="فعم" aug="جر" cas="ج" ncg="1" trans="ل"/>
    <pattern id="1021" diac="أفْلِ" canonic="أفْلِ" type="أ" aug="زي" cas="" ncg="3" trans="ك"/>
    <pattern id="1022" diac="أفْلِ" canonic="أفْلِ" type="أ" aug="زي" cas="" ncg="3" trans="ل"/>
    <pattern id="1023" diac="أفْلِ" canonic="أفْلِ" type="أ" aug="زي" cas="" ncg="3" trans="م"/>
```

FIGURE 5.8 – Exemple de patrons de dérivation verbaux.

Al-Khalil-ASM utilise aussi deux listes de clitiques : proclitiques et enclitiques lors de l'analyse morphologique d'un mot. On identifie pour les clitiques leurs formes sans et avec diacritiques et la classe de noms auxquels s'attachent les clitiques (nom, verbe ou les deux).

5.6.2 Adaptation de l'analyseur

Bien qu'Al-Khalil-ASM soit un très bon analyseur, il présente quelques bugs et erreurs dans sa base de données [Altabbaa *et al.* 2010]. Donc, avant de l'adapter au DT, nous corrigeons d'abord ces erreurs.

Les bases lexicales d'Al-Khalil-ASM contiennent des erreurs pour la définition de quelques traits morphologiques. Par exemple, les schèmes de dérivation contiennent des voyelles courtes erronées. Quelques traits pour certains schèmes de dérivation ont des valeurs manquantes. Dans d'autres cas, ils contiennent des valeurs erronées. Nous essayons de corriger ces erreurs et bugs avant de commencer l'adaptation d'Al-Khalil-ASM. Parmi les points faibles de l'analyseur morphologique Al-Khalil-ASM, nous citons l'affichage du résultat de l'analyse. En effet, Al-Khalil-ASM présente les résultats de l'analyse sous forme d'un fichier HTML qui regroupe les différentes analyses morphologiques possibles. La structure de ce fichier n'est pas très utile pour une utilisation future du résultat de l'analyse morphologique. Pour remédier à ce point faible, nous avons modifié Al-Khalil-ASM pour qu'il soit capable de stocker les résultats de l'analyse dans une structure XML. De même, Al-Khalil-ASM ne fait pas la différence entre les affixes et les clitiques. Ainsi, nous l'avons corrigé pour qu'il soit capable de traiter chaque partie du mot à part. Un exemple de la nouvelle forme du fichier sortie est présenté dans la figure 4.4 du chapitre 4.

Pareillement, nous avons amélioré le résultat de l'analyse avec d'autres détails. Nous avons bien spécifié les caractéristiques morphologiques de chaque terme traité telle que le genre, le nombre, l'aspect, le dialecte, etc.

Semblablement, nous avons mis à jour la liste des catégories grammaticales en suivant l'ensemble des étiquettes utilisées par l'analyseur morphologique MADA [Habash & Rambow 2005]. Les jeux d'étiquettes utilisées par Al-Khalil-TUN sont présentés dans le tableau 5.7.

L'adaptation de l'analyseur Al-Khalil-ASM consiste à intégrer dans une première étape le lexique du DT généré. Ensuite, nous améliorons cette ressource lexicale par l'ajout d'une liste de mots du DT qui sont généralement issus des langues étrangères et une liste des mots-outils. Puis, nous mettons à jour le module de tokenisation en spécifiant les nouvelles règles de tokenisation, en intégrant la liste des affixes et des clitiques du DT et en spécifiant bien la classe de chaque clitique.

De même, nous modifions Al-Khalil-ASM pour qu'il soit capable d'identifier les phénomènes spécifiques à l'oral. Ainsi, Al-Khalil-TUN est capable de détecter et annoter les pauses silencieuses, les pauses remplies, les mots incomplets, et les onomatopées.

Nous gardons les bases lexicales de l'ASM au niveau de la version tunisienne d'Al-Khalil

Symbole	Catégorie grammaticale	Genre			Nombre			Personne		
		f	m	na	s	p	na	1	2	3
adj	Adjectif	x	x		x	x				
adv_place	Adverbe de place	x	x	x			x			
adv_temp	Adverbe de temps	x	x	x			x			
interrog_adv	Adverbe interrogatif			x			x			
rel_adv	Adverbe relatif			x			x			
conj	Conjonction			x			x			
sub_conj	Conjonction de subordination			x			x			
interj	Interjection			x			x			
fw	Mot étranger	x	x		x	x				
TrunW	Mot tronqué			x			x			
noun_count	Nom de comptage			x			x			
number_noun	Nom de nombre	x	x		x	x				
prop_noun	Nom propre	x	x							
onom	Onomatopée			x			x			
part	Particule			x			x			
part_abst	Particule d'abstraction			x			x			
part_cond	Particule de condition			x			x			
part_fut	Particule de futur			x			x			
part_neg	Particule de négation			x			x			
part_restrict	Particule de restriction			x			x			
part_verb	Particule de verbe			x			x			
part_interrog	Particule interrogative			x			x			
part_voc	Particule voix			x			x			
break	Pause silencieuse			x			x			
FPause	Pause remplie			x			x			
prep	Préposition			x			x			
pron	Pronom	x	x	x	x	x	x	x	x	x
dem_pron	Pronom démonstratif	x	x	x	x	x	x			
ind_obj_pron	Pronom d'objet indirect	x	x		x	x		x	x	x
poss_pron	Pronom possessif				x	x		x	x	x
rel_pron	Pronom relatif			x			x			
verb	Verbe	x	x		x	x		x	x	x

TABLE 5.7 – Liste des étiquettes utilisées par l'analyseur morphologique du DT.

afin de proposer plus d'analyses pour les mots inconnus. Nous proposons d'analyser chaque mot en utilisant les bases lexicales du DT. Si aucun résultat n'est fourni, ce mot sera analysé par les bases lexicales de l'ASM.

Finalement, Al-Khalil-TUN est un analyseur morphologique pour le DT et aussi pour l'ASM qui prend en compte des spécificités de l'oral (les mots incomplets, les pauses remplies, etc.). La figure 5.9 illustre l'architecture de Al-Khalil-TUN.

5.7 Expérimentations et résultats

5.7.1 Ressources utilisées

Pour créer un analyseur morphologique pour le DT, nous nous sommes basés sur deux ressources : un lexique de l'ASM et le corpus STAC.

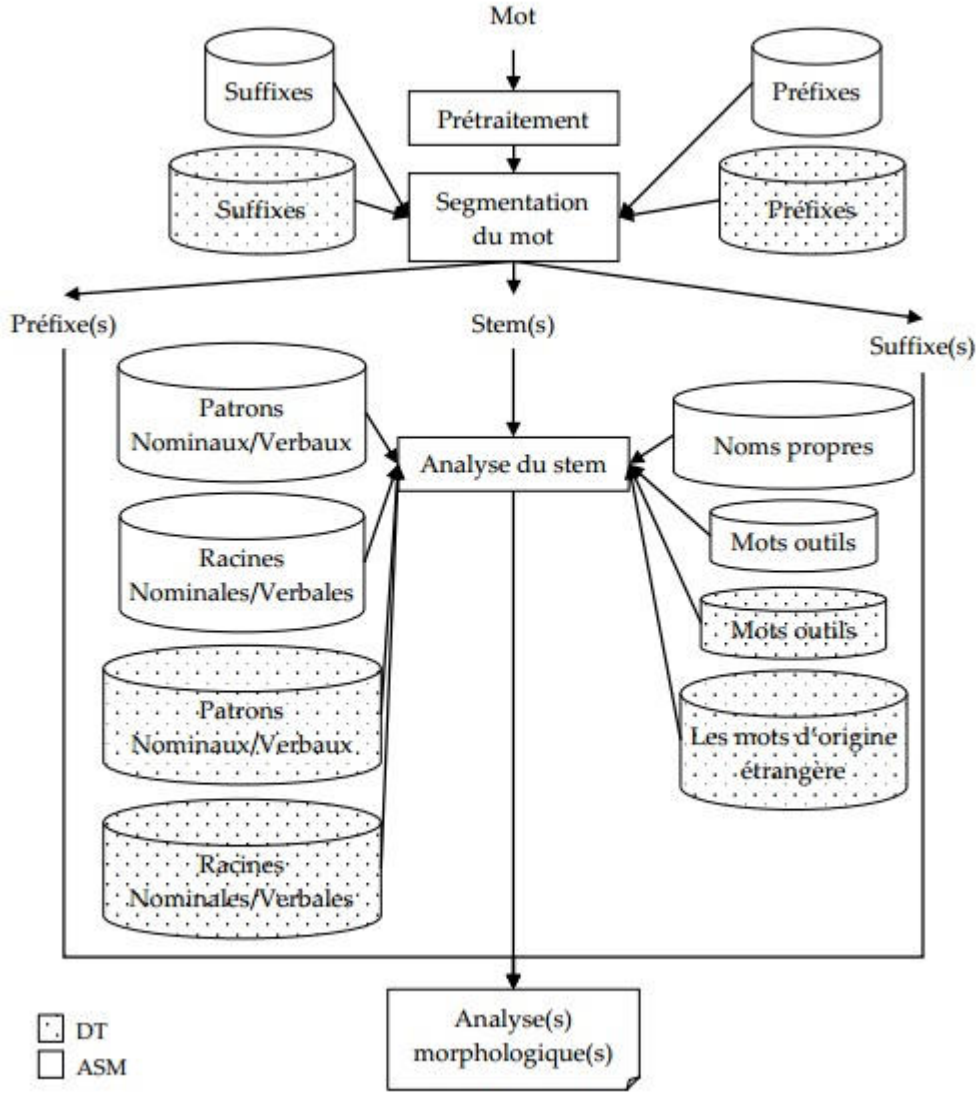

FIGURE 5.9 – Architecture du système Al-Khalil-TUN

5.7.1.1 Lexique de l'arabe standard

Nous exploitons le lexique de l'ASM utilisé par Al-Khalil-ASM. Le lexique suit une représentation sous la forme de « racine-patron ». Il est composé de 7 503 racines et 3 681 patrons (verbaux et nominaux) sans voyelles courtes. Les patrons et les racines définies dans ce lexique sont reliés. Pour chaque racine, un ensemble de partons verbaux et/ou nominaux est défini. De même, pour chaque patron, on identifie les différentes diacritisations en définissant les caractéristiques morphologiques correspondantes. Ce lexique couvre un nombre important de racines. Parfois, le nombre important de racines engendre l'augmentation des cas d'ambiguïtés lors de la création de notre lexique pour le DT. D'où, nous définissons une version allégée de ce lexique en réduisant le nombre de racines et patrons. Cette version est composée de 2 503 racines et 3 681 patrons (verbaux et nominaux) sans voyelles courtes.

5.7.1.2 Corpus de test et d'apprentissage

Pour la création et l'évaluation de notre analyseur morphologique Al-Khalil-TUN, nous avons divisé le corpus STAC en deux parties. La première partie a été utilisée pour la création du lexique TUN et le développement de l'analyseur morphologique Al-Khalil-TUN. Elle est composée de 24 003 mots et 1 035 mots types. Cette partie est composée de 234 verbes, 345 noms et 456 autres mots distincts (mots-outils et les phénomènes de l'oral). La partie d'apprentissage passe par un ensemble de prétraitements. Nous créons deux versions de ce corpus. La première version est utilisée pour la création du lexique. Elle regroupe les mots types sans voyelles courtes. Nous collectons à partir de cette partie d'apprentissage une liste de verbes et aussi une liste de noms. Nous filtrons aussi les mots des clitiques attachés pour ne pas falsifier l'étape d'extraction des racines et la génération des patrons. La deuxième version de la partie d'apprentissage est utilisée pour le développement et la validation de l'analyseur morphologique. Aucun traitement n'est effectué sur cette version du corpus. Nous supprimons uniquement les voyelles courtes de cette version du corpus.

Pour le test de l'analyseur Al-Khalil-TUN, nous avons utilisé la deuxième partie du corpus STAC composée de 12 054 mots et 589 mots types. Ce corpus de test est composé de 232 noms, 234 verbes et 123 autres mots types.

5.7.2 Mesures d'évaluation

Les performances de notre système d'analyse morphologique sont mesurées en fonction des analyses morphologiques proposées par Al-Khalil-TUN. Nous étudions le nombre d'analyses attribuées pour chaque mot, la justesse des analyses, etc. Nous nous intéressons, aussi, à mesurer le degré de sous génération ou surgénération des analyses morphologiques. Ainsi, nous mesurons la compétitivité de notre système en termes de précision, de rappel et F-mesure. Nous exploitons deux définitions pour ces mesures. Nous définissons ces mesures d'évaluation en fonction de nombre d'analyses attribuées par Al-Khalil-TUN. Premièrement, nous calculons le nombre de mots reconnus par l'analyseur morphologique par rapport au nombre total des mots. Nous mesurons le taux de couverture du lexique utilisé par rapport à l'ensemble de mots analysés. La définition du rappel et précision est définie comme suit :

$$Rappel_{eval1} = \frac{Nombre\ de\ mots\ correctement\ analysés}{Nombre\ total\ de\ mots\ analysés} \tag{5.2}$$

$$Précision_{eval1} = \frac{Nombre\ de\ mots\ correctement\ analysés}{Nombre\ de\ mots\ reconnus\ par\ l'analyseur} \tag{5.3}$$

Nous calculons les valeurs de rappel et précision en utilisant deux définitions pour le terme « un mot correctement analysé ». Dans une première évaluation, nous définissons « un mot correctement analysé » par un mot dont l'analyseur lui attribue au moins une analyse correcte. Dans une deuxième évaluation, nous considérons « un mot correctement analysé » un mot dont l'analyseur lui attribue toutes les analyses correctes possibles. Nous utilisons les acronymes

Rappel$_{eval1a}$, Précision$_{eval1a}$ et Rappel$_{eval1b}$, Précision$_{eval1b}$ respectivement pour les valeurs de rappel et précision de la première et la deuxième évaluation.

Deuxièmement, nous calculons le nombre d'analyses correctes attribuées par l'analyseur par rapport au nombre total d'analyses. Avec cette mesure, nous pouvons calculer le taux de génération des résultats.

$$Rappel_{eval2} = \frac{Nombre\ d'analyses\ correctes}{Nombre\ total\ d'analyses\ qui\ devraient\ \hat{e}tre\ donn\acute{e}es} \tag{5.4}$$

$$Pr\acute{e}cision_{eval2} = \frac{Nombre\ d'analyses\ correctes}{nombre\ total\ d'analyses\ donn\acute{e}es} \tag{5.5}$$

5.7.3 Expérimentations et évaluation

Pour valider notre méthode, nous avons choisi d'effectuer quelques expérimentations. Au niveau de la première expérimentation, nous testons deux versions d'Al-Khalil-TUN. Les ressources lexicales de la première version suivent la convention de transcription orthographique OTTA. La deuxième version est conçue pour traiter les mots transcrits en suivant la convention CODA-TUN. Ensuite, au niveau de la deuxième expérimentation, nous testons les différences entre l'utilisation d'une version réduite et une version longue du lexique d'Al-Khalil-ASM. Enfin, nous présentons une évaluation des différentes étapes de création de notre lexique. Ainsi, nous développons trois systèmes. Le premier utilise la liste des racines et des patrons issue de la première étape de la création du lexique. Le deuxième ajoute à cette liste les racines et patrons générés automatiquement et le troisième utilise une version complète de notre lexique.

Notons que le résultat de chaque expérimentation est considéré comme une entrée pour l'expérimentation qui suit.

5.7.3.1 OTTA vs. CODA-TUN

Nous avons proposé deux conventions de transcription orthographique pour le DT : OTTA et CODA-TUN. En effet, ces deux conventions présentent plusieurs points en commun mais évidemment des différences. Ces dernières sont principalement au niveau phonologique qui est dominant pour la convention OTTA en utilisant des suffixes et des proclitiques différents de CODA-TUN (*e.g.* le pronom personnel singulier (و, w) au lieu de (ه, h), etc.). Ainsi pour saisir les différences entre les deux conventions, nous développons deux versions d'Al-Khalil-TUN. La première version suit les règles de transcription orthographique de CODA-TUN alors que la deuxième version est pour les règles de transcription d'OTTA.

Généralement, la différence entre ces deux conventions peut affecter le niveau de segmentation des mots. De ce fait, nous calculons les pourcentages de segmentations réussies en transcrivant suivant les deux conventions de transcription CODA-TUN et OTTA. Le tableau 5.8 présente les résultats trouvés en utilisant uniquement le corpus de test. De même, nous calculons aussi les valeurs de rappel et précision pour l'évaluation eval$_{1a}$. L'objectif de cette

évaluation est de voir l'impact de l'utilisation d'une transcription orthographique à base phonétique sur la qualité de l'analyse morphologique. Nous voulons mesurer l'effet de l'utilisation d'une graphie proche de l'ASM sur le développement des outils de TAL. Cette évaluation nous permet, aussi, de justifier le développement d'une convention orthographique à base de l'ASM. Rappelons que notre corpus STAC est transcrit en deux versions. La première respecte les règles orthographiques d'OTTA et la deuxième respecte les règles de transcription de CODA-TUN. Le tableau 5.9 présente les résultats trouvés sur le corpus de développement et de test.

Catégorie grammaticale	**Segmentation correcte**		**Nombre de mots types non-distincts**
	OTTA	***CODA-TUN***	
Nom	93,94 %	95,03 %	4 424
Verbe	89,80 %	95,64 %	2 567
Mot outil	97,90 %	99,30 %	451
Total	92.70 %	95,50 %	7 442

TABLE 5.8 – Comparaison entre les segmentations des mots types (verbes, noms et mots-outils) en utilisant deux conventions de transcription orthographique OTTA et CODA-TUN.

	OTTA		***CODA-TUN***	
	Développement	**Test**	**Développement**	**Test**
Rappel_{eval1b}	0,764	0,7931	0,849	0,753
$\text{Précision}_{eval1b}$	0,826	0,7953	0,914	0,971
F-mesure_{eval1b}	0,794	0,7942	0,880	0,848

TABLE 5.9 – Comparaison entre les valeurs de Rappel_{eval1b},$\text{Précision}_{eval1b}$et F-mesure_{eval1b} reportées sur le corpus de développement et de test pour les deux versions de Al-Khalil-TUN en utilisant les conventions OTTA et CODA-TUN.

5.7.3.2 Deux lexiques de l'arabe standard

Le lexique intégré dans Al-Khalil-ASM est composé d'un grand nombre de racines et de patrons qui génèrent parfois des mots n'appartenant pas à l'ASM. D'où, nous proposons comme première étape d'utiliser ce lexique pour générer un pour le DT. Dans un deuxième temps, nous filtrons le lexique principalement des racines n'appartenant pas à l'ASM. Le filtrage est effectué manuellement à l'aide d'un expert qui essaye de dériver les racines de notre lexique avec les schèmes correspondants. L'ensemble des flexions est ensuite étudié. Notre expert essaye de chercher les flexions dans un dictionnaire de l'arabe classique pour décider si une racine doit être supprimée ou gardée.

Dans le tableau 5.10, nous présentons les valeurs de Rappel_{eval2}, Précision_{eval2} et F-mesure_{eval2}.

L'utilisation de deux versions de tailles différentes a pour objectif de voir l'impact de la taille du lexique de l'ASM sur notre première étape de la création d'un lexique pour le DT. Nous voulons mesurer l'effet de l'utilisation d'un lexique volumineux de l'ASM sur le nombre d'analyses morphologiques générées, c'est-à-dire, on va étudier la liaison entre la taille du lexique et le taux d'ambiguïté des résultats.

	Lexique entier		*Lexique réduit*	
	Développement	**Test**	**Développement**	**Test**
Rappel_{eval2}	0,968	0,964	0,994	0,702
Précision_{eval2}	0,693	0,682	0,846	0,953
F-mesure_{eval2}	0,807	0,799	0,915	0,809

TABLE 5.10 – Comparaison entre les valeurs de Rappel_{eval2}, Précision_{eval2} et F-mesure_{eval2} reportées sur le corpus de développement et de test en utilisant deux lexiques.

5.7.3.3 Évaluation d'Al-Khalil-TUN

La tâche de l'analyse morphologique du DT n'a pas pris l'attention de plusieurs chercheurs en TAL. Ainsi, nous choisissons d'utiliser Al-Khalil-ASM comme un système de référence pour comparer les performances de notre système puisque le seul analyseur morphologique pour le dialecte tunisien MAGEAD [Hamdi 2015] est sous licence non libre.

Ainsi, nous calculons les valeurs de précision et rappel pour les deux systèmes. Nous reportons le résultat de deux évaluations eval1 et eval2 (voir Tableau 5.11). Nous calculons aussi, les pourcentages des analyses correctes, erronées et le pourcentage des mots non reconnus lors de l'exécution de notre analyseur morphologique. Le tableau 5.12 présente les analyses reportées. Le taux d'analyses correctes est le pourcentage des analyses correctes prévues pour le mot. Le taux des analyses erronées est le pourcentage des analyses erronées qui ne peuvent pas être accordées à ce mot.

	Evaluation 1			**Evaluation 2**		
	Rappel_{eval1a}	$\text{Précision}_{eval1a}$	F-mesure_{eval1a}	Rappel_{eval2}	Précision_{eval2}	F-mesure_{eval2}
Baseline	0,602	0,780	0,679	0,450	0,076	0,130
Système 1	0,702	0,953	0,809	0,813	0,806	0,810
Système 2	0,722	0,984	0,833	0,972	0,888	0,928
Système 3	0,753	0,971	0,848	0,991	0,871	0,927

TABLE 5.11 – Les valeurs de Rappel_{eval2},Précision_{eval2}et F-mesure_{eval2} reportées sur le corpus de test.

Catégorie grammaticale	**Analyses correctes**	**Analyses erronées**
Nom	88,10%	11,90%
Verbe	85,59%	14,41%
Mot outil	98,41%	1,59%
Total	**90,70%**	**9,30%**

TABLE 5.12 – Les pourcentages d'analyses correctes, erronées et non reconnues de chaque type de catégorie grammaticale.

Al-Khalil-TUN utilise deux bases lexicales (de l'ASM et du DT). Donc, nous voulons mesurer le pourcentage des mots analysés par les deux bases lexicales. Nous calculons les résultats pour les trois catégories grammaticales (nom, verbe et mot outil). Le tableau 5.13 résume les résultats de l'analyse avec les deux bases lexicales.

Catégorie grammaticale	DT	ASM	Non reconnus	% des mots reconnus
Nom	55,67%	66,91%	71,52%	**94,39%**
Verbe	29,45%	26,84%	27,27%	
Mot outil	14,88%	6,25%	1,21%	
Total	**85,15%**	**9,24%**	**5,61%**	

TABLE 5.13 – Les pourcentages d'analyse des mots avec les deux bases lexicales de l'ASM et du DT.

5.7.4 Discussion des résultats obtenus

Nous discutons dans cette section les résultats de chaque expérimentation. L'objectif de la première expérimentation est de voir l'impact de l'utilisation de deux conventions orthographiques sur le résultat de la segmentation de mots lors de l'analyse morphologique. Les résultats reportés dans le tableau 5.9 montrent l'influence de ces deux conventions sur la segmentation. D'après les analyses trouvées, nous remarquons que les différences entre les deux conventions sont mineures (2,8 %). Ainsi, l'utilisation des affixes selon la convention OTTA augmente le nombre d'analyses erronées pour les mots reconnus par l'analyseur morphologique Al-Khalil-TUN. En effet, l'utilisation de l'enclitique (و, w) présente l'un des problèmes rencontrés lors de la segmentation et conduit à d'autres erronées. Prenons l'exemple du mot (باردو, bArdw, « *Bardo* ») : un nom propre non agglutiné. Le module de segmentation selon la convention OTTA a divisé ce mot en (بارد+و, bArd+w) « *froid+ son* » en considérant la lettre (و, w) à la position finale comme un enclitique. Or, cette segmentation est erronée. La même erreur se propage sur plusieurs mots qui se terminent par la lettre (و, w).

En effet, l'utilisation du proclitique (ه, h) a donné la possibilité à plusieurs mots d'être analysés en utilisant la base lexicale de l'ASM. Nous remarquons que le taux de mots non reconnus est diminué de 8,5 % pour le corpus du développement alors que pour le corpus du test la valeur est augmentée par 4 %. Nous justifions cette augmentation par le nombre important des mots d'origine ASM appartenant au corpus de développement à la différence de corpus du test. En revanche, les valeurs de précision pour le corpus de test et de développement ont augmenté lors de l'utilisation de la convention orthographique CODA-TUN.

Suite à cette évaluation, nous pouvons conclure que la convention orthographique OTTA n'est pas très appropriée pour les traitements linguistiques pour le DT. Ceci est dû notamment à la domination de la transcription à base de la phonétique qui perturbe les résultats de l'analyse morphologique.

Dans la deuxième expérimentation, nous avons mesuré l'impact d'utilisation de deux lexiques de la langue arabe afin de créer un lexique pour le DT. Nous remarquons que l'utilisation d'un lexique réduit a permis d'augmenter les valeurs de précision. Nous remarquons une amélioration de 20 % à 30 % pour les deux corpus (développement et test). En effet, une base lexicale riche peut engendrer des analyses erronées pour les mots du DT. Nous constatons que les mots-outils peuvent avoir un ensemble de catégories grammaticales erronées. Par exemple, le pronom démonstratif (هذَا) est analysé comme étant une dérivation de la racine

(ه-ذ-أ) suivant le schème de dérivation (فعَا). Cette analyse attribue à ce pronom démonstratif la catégorie verbe. Donc, le nombre important de racines de l'ASM engendre, dans certains cas, l'affectation des analyses morphologiques erronées pour un mot en DT. Cette constatation justifie bien la valeur élevée du rappel et la valeur modérée de la précision.

En comparant les valeurs de l'analyse morphologique à celles du Baseline, nous remarquons que le nombre d'analyses morphologiques données par Al-Khalil-ASM est très élevé lors de l'analyse du DT. Ceci engendre des valeurs de rappel et de précision très médiocres. Ce résultat déplorable s'explique par le nombre important des racines et de patrons dans sa base lexicale et aussi par les différences de diacritisations entre l'ASM et le DT. Par contre, nous remarquons que les valeurs de l'évaluation 1 sont encourageantes. La valeur de F-mesure est aux alentours de 60 %. Ainsi, Al-Khalil-ASM peut reconnaitre 60 % des mots en DT qui partagent avec l'ASM les racines et les schèmes de dérivation.

L'étude des deux étapes proposées à la création d'un lexique pour le DT a montré que la première étape a permis de connaître 70,2 % des mots analysés (les résultats du Système 1). Mais, la deuxième étape a amélioré uniquement la valeur de rappel de 2% (voir résultat du Système 2). Cette faible amélioration est due à la nature du corpus de test. Il contient peu de mots appartenant aux classes *C2* et *C3*. La Système 3 avec une version complète du lexique a amélioré la valeur du rappel de 5,1 % par rapport au système 1 et 3,1 % par rapport au système 2. Nous notons que la précision des analyses a augmenté, aussi, de 1,8 % (une faible amélioration).

Pour la deuxième évaluation, nous constatons qu'Al-Khalil-TUN attribue aux mots 99,1 % des analyses qui devraient être données avec un taux de précision égale à 87,1 %. Ces deux valeurs sont encourageantes. Elles montrent la qualité du lexique utilisé par Al-Khalil-TUN. Nous notons que la valeur de précision a diminué par rapport à celle reporté avec le système 2. Une diminution justifiée par le fait que l'ajout des listes des mots outils et des mots étrangers engendre dans certains cas de surgénération des analyses. Par exemple, le mot سلَاطة «*salade*» est reconnu par le lexique généré de la première étape via le moulage entre le schème $r_1r_2aAr_3aħ$ et la racine س-ل-ط et il est aussi reconnu par le lexique des mots étrangers.

Nous déduisons qu'Al-Khalil a pu donner des analyses pour 94,39 % des mots analysés. C'est un taux encourageant. De même, l'utilisation de la base lexicale de l'ASM a amélioré ce taux de reconnaissance des mots. Nous concluons, aussi, que 9,24 % des mots reconnus sont analysés avec la base lexicale de l'ASM et uniquement 5,61 % des mots ne sont pas reconnus. Ces mots présentent des mots issus des langues étrangères et des variétés dialectales différentes du DT traité.

5.8 Conclusion

La création d'un analyseur morphologique pour le DT a fait l'objet de ce chapitre. D'abord, nous avons présenté une étude du lexique du DT. Ensuite, nous avons proposé une méthode pour l'adaptation d'un analyseur morphologique pour le DT. Nous avons proposé aussi une

méthode pour la création d'un lexique « racine-patron » en utilisant deux ressources pour le DT et l'ASM. Enfin, nous avons présenté les expérimentations en discutant les résultats obtenus.

Évidemment, l'analyse morphologique basée sur un lexique génère toujours plusieurs analyses pour chaque mot. Les résultats ont montré que le taux d'ambiguïté est aux alentours de 70 %. L'utilisation directe du résultat d'analyse morphologique ne peut pas être utile pour plusieurs applications de TAL. Une étape de désambiguïsation s'avère importante pour détecter les catégories grammaticales appropriées. Cette étape sera l'objet du chapitre suivant.

CHAPITRE 6

Désambiguïsation morphosyntaxique du dialecte tunisien

Sommaire

6.1 Introduction

Le traitement automatique de la langue est souvent confronté aux problèmes liés à l'ambiguïté (morphologique, morphosyntaxique, sémantique, etc.). La langue arabe se caractérise par un niveau d'ambiguïté assez fort en la comparant aux langues indo-européennes. Cette ambiguïté complique son traitement automatique. Si nous nous concentrons particulièrement sur la morphosyntaxe, nous constatons la présence de plusieurs mots avec un nombre important de catégories grammaticales qui peuvent être attribuées. L'ambiguïté touche aussi la segmentation des mots en composants (suffixes, stem et préfixes). Le taux d'ambiguïté augmente lorsqu'on traite une langue peu dotée plus précisément le DT. L'analyse morphologique du DT nous a montré un taux d'ambiguïté d'environ de 70 %. En DT, un mot peut être à la fois un nom, une particule ou un verbe. Même s'il s'agit d'un verbe, ce dernier peut être conjugué de la même façon avec la troisième personne, la première personne et/ ou la deuxième personne du singulier. Ce niveau d'ambiguïté engendre plusieurs difficultés, lors des analyses linguistiques qui exploitent les caractéristiques morphologiques.

Nous nous intéressons dans ce chapitre au phénomène de l'ambiguïté morphosyntaxique dans le traitement automatique du DT. Nous présentons, dans la deuxième section, les difficultés de l'étiquetage de la langue parlée. Ensuite, nous justifions notre choix pour développer un nouvel outil pour la désambiguïsation morphosyntaxique du DT. Ensuite, nous décrivons au niveau de la section 6.4 la méthode proposée. Nous présentons, par la suite, notre méthode pour la segmentation des transcriptions en phrases. La section 6.7 est consacrée pour la description de l'analyse morphologique du DT parlé. L'adaptation de Al-Khalil-TUN est l'objet de la section 6.7. Nous détaillons dans cette section la méthode de désambiguïsation morphosyntaxique du DT. Enfin, nous clôturons le chapitre par une évaluation en discutant les résultats obtenus.

6.2 Les difficultés de l'étiquetage de la langue parlée

6.2.1 La segmentation des phrases

L'analyse morphosyntaxique consiste à attribuer, à chaque mot d'une phrase, sa catégorie grammaticale selon son contexte dans la phrase. Ainsi, la définition des frontières des phrases est une étape préliminaire et nécessaire à la tâche d'analyse morphosyntaxique. Cette tâche peut être négligeable pour les langues écrites qui se caractérisent par la présence des signes de ponctuation et la présence des lettres majuscules pour les langues indo-européennes [Belguith 2009].

Ces deux caractéristiques ne sont pas applicables à la langue arabe. D'une part, vu l'absence des lettres majuscules, et d'autre part, vu que les signes de ponctuation ne sont pas toujours utilisés et même lorsqu'ils y figurent, ils sont mis arbitrairement. Ainsi, on peut trouver tout un texte arabe avec un seul point orphelin à la fin [Belguith *et al.* 2014]. Notons que le DT partage ainsi avec l'ASM ces deux caractéristiques. En plus, la forme orale du dialecte connait

d'autres caractéristiques. Parmi lesquelles, nous pouvons citer l'absence totale des marques de ponctuations et l'abandon de la notion de phrase. En oral, les linguistes abandonnent la notion de phrase qui a une structure bien définie [Tellier *et al.* 2010]. Ils parlent de notion « *d'utterance* [1] » (énoncé) qui présente des petites unités de parole qui possèdent généralement un sens. Dans la suite du rapport, nous utilisons le terme phrase pour désigner un énoncé (*utterance*).

Pour le DT, cette tâche n'est pas triviale. Nous devons premièrement définir le terme phrase. En oral, nous détectons plusieurs types de phrases : des phrases bien formées et phrases incomplètes et des phrases avec des segments disfluents. D'où, il est nécessaire de définir les unités d'énoncé auxquelles nous proposons une méthode d'analyse morphosyntaxique. Toutes ces tâches alourdissent l'analyse morphosyntaxique du DT.

6.2.2 La présence des disfluences

La présence de disfluences dans le corpus oral présente un des problèmes qui alourdissent la tâche d'étiquetage morphosyntaxique de la langue parlée. En effet, les disfluences affectent la structure des phrases en impliquant plusieurs éléments de nature différente dans une phrase. Les mots incomplets, les pauses remplies, les pauses silencieuses, les répétitions, etc. affectent la structure syntaxique de la phrase. Donc, une étape d'identification de cet élément aide à accomplir la tâche d'étiquetage morphosyntaxique. De même, la présence des effets spéciaux de l'oral dans le discours augmente le nombre de mots non reconnus. Dans certains cas, elle engendre l'augmentation des taux d'ambiguïté. Par exemple, l'interjection (بَاهي, bAhy, « *d'accord* ») peut être annotée comme étant un adjectif dans un analyseur morphosyntaxique qui ne traite pas les phénomènes de l'oral. De même, les pauses remplies (اَ A, امّم Amm, etc.) qui sont assez fréquentes dans le discours oral, elles ne sont pas reconnues par l'application d'un simple analyseur morphosyntaxique pour la langue écrite. De même, la présence de reparandum et le reparans change la structure syntaxique de la phrase. On se trouve souvent en face des structures syntaxiques incorrectes. Par exemple, dans la phrase 1, le nom (بدني, bdny, « *mon corps* ») est normalement suivi par un adjectif. Mais, nous remarquons l'emploi d'un marqueur de discours, suivi par un adjectif. D'où, il est très difficile de détecter dans une transcription, des règles ou des patrons qui définissent la structure syntaxique des phrases qui contiennent les disfluences. Donc, ceci alourdit la tâche d'analyse morphosyntaxique.

1. نحس في بدني فهمتني مريض
   nHs fy bdny fhmtny mryD
   « *je suis, tu me comprends, malade* »

6.2.3 L'irrégularité de l'ordre des mots dans la phrase

Le DT est une variante de la langue arabe qui se caractérise par une irrégularité dans l'ordre des mots dans les phrases. Nous pouvons exprimer une seule phrase avec plusieurs structures

1. Utterance est l'unité la plus petite du discours.

syntaxiques : Sujet-Verbe-Objet (SVO), Verbe-Sujet-Objet (VSO) et Objet-Verbe-Sujet (OVS). Cette structure variable dans les phrases de la langue arabe cause des problèmes majeurs au niveau de l'étiquetage morphosyntaxique. Il est difficile ainsi de dégager des règles ou des patrons pouvant décrire d'une façon assez simple la structure des phrases de la langue arabe. Cette caractéristique est héritée par le DT. Mais, cette particularité en DT laisse la tâche d'analyse morphosyntaxique plus complexe. Nous commençons par étudier le DT comme étant une langue écrite. En effet, le dialecte est essentiellement une langue sans grammaire exacte ou des règles syntaxiques standards. Elle partage avec l'ASM quelques caractéristiques grammaticales et syntaxiques (les phrases verbales, les phrases nominales, les structures syntaxiques : SVO, VSO, etc.). Mais, l'utilisation massive des mots des langues étrangères et l'alternance codique engendrent dans certains cas une perte de la structure syntaxique des phrases en DT. Nous prenons l'exemple de la phrase suivante 2 en DT.

2. C'est vrai الّي آنَا مَا نفهمش في ال domaine امَا نجم نعطيك حل.

   C'est vrai Ally ĀnA mA nfhmš fy Al domaine AmA njm nʕTyk Hl

   « *C'est vrai que je ne m'y connais pas dans ce domaine mais je peux te donner une solution.* »

Cette phrase est composée des mots des expressions et des mots de la langue française. L'analyse morphosyntaxique d'une telle phrase est confrontée à plusieurs problèmes : comment traiter les mots étrangers, quelle étiquette morphosyntaxique doit-on attribuer à ces mots ?, etc.

Par ailleurs, nous passons maintenant à présenter les difficultés si on traite le DT comme étant une langue parlée. En effet, lorsqu'un locuteur parle, il s'exprime en utilisant des signes, des sons, etc. pour exprimer son discours. La transcription orthographique ne schématise pas tous ces effets de l'oral. De même, nous ne pouvons pas les considérer pour la tâche d'analyse morphosyntaxique. De même, le locuteur exprime son idée par un simple mot (un verbe, un nom ou même une pause remplie, etc.). Il peut, aussi, commencer une phrase sans terminer l'idée et un autre locuteur l'achève (*voir tableau 6.1*).

Locuteur A :	آنَا توة جيت بَاش نقول لك ...	
	ĀnA twħ jyt bAš nqwl lk ...	
Locuteur B :	آنَا خَارجة نعرف	
	ĀnA xArjħ nʕrf	

TABLE 6.1 – Exemple de conversation débuté par un locuteur et terminé par un autre.

On se trouve face à une grande variété de phrases avec des structures syntaxiques différentes et même des structures inachevées. Ceci pose plusieurs challenges pour l'analyse linguistique de façon générale et l'analyse morphosyntaxique particulièrement.

Pour conclure, les caractéristiques du DT associées avec les caractéristiques de l'oral présentent de nombreux défis à surmonter pour réussir la tâche d'analyse morphosyntaxique de l'oral tunisien spontané.

6.3 Importance de la création d'un nouvel outil pour l'analyse morphosyntaxique

L'approche que nous avons suivie lors du développement de nos outils pour l'analyse du DT est basée sur l'adaptation des ressources de l'ASM en faveur du DT. Cette approche a été suivie par plusieurs chercheurs en TALN dans le but de créer des outils et même des ressources pour plusieurs dialectes arabes.

Pour la tâche d'analyse morphosyntaxique, l'approche d'adaptation est rarement adoptée. Souvent, un étiqueteur est développé en faisant appel à une méthode stochastique ou statistique. Les méthodes proposées se basent sur la présence d'un large corpus annoté, à partir duquel un modèle est généré pour étiqueter morphosyntaxiquement un texte. Ces étiqueteurs proposent l'annotation d'un mot suivant un ensemble d'étiquettes qui utilisent les étiquettes de catégories grammaticales pour enrichir un texte. Une autre approche pour l'étiquetage morphosyntaxique a été suivie par les chercheurs en TALN. Elle consiste à désambiguïser le résultat de l'analyse morphologique. Cette approche exige évidemment la présence d'un analyseur morphologique pour la langue cible en utilisant souvent des règles de désambiguïsation ou en suivant des méthodes statistiques diverses pour réaliser cette tâche d'annotation.

Pour l'étiquetage morphosyntaxique du DT, nous adoptons une approche qui envisage de désambiguïser morphosyntaxiquement le résultat de l'analyse morphologique.

Nous justifions notre choix par deux principales causes. La première est que nous disposons d'un analyseur morphologique pour le DT. Il attribue à un mot toutes les analyses morphologiques possibles en spécifiant la catégorie grammaticale, la voix, le genre, le nombre, l'aspect, etc. Ainsi, il suffit de choisir la bonne combinaison entre ces caractéristiques pour un mot donné suivant son contexte dans une phrase. Cette approche de désambiguïsation a été proposée ([Habash & Rambow 2005], [Habash *et al.* 2013]) pour réaliser l'étiquetage morphosyntaxique de l'ASM et le dialecte égyptien. Cette approche a donné des résultats satisfaisants pour ces deux variétés de langues. De ce fait, nous proposons de tester cette approche pour étiqueter le DT.

De plus, le DT connait un niveau d'ambiguïté assez élevé par rapport à autres langues et même pour l'ASM. Connaitre la catégorie grammaticale pour un mot peut ne pas être suffisant pour dissiper l'ambiguïté. Par exemple, le verbe (رجعت, rjʕt) est un verbe ambigu qui peut être à la fois conjugué à la première personne du singulier, à la deuxième personne du singulier ou à la troisième personne du singulier au féminin. Par conséquent, un étiquetage morphosyntaxique en se basant uniquement à la catégorie grammaticale n'est pas suffisant pour dissiper l'ambiguïté des mots en DT.

6.4 Notre outil pour l'analyse morphosyntaxique du dialecte tunisien

L'analyse ou l'étiquetage morphosyntaxique passe généralement par trois étapes [Nguyen 2006] : la segmentation du texte en unités lexicales, l'étiquetage ou l'association à chaque occurrence de mot de toutes ses étiquettes possibles et finalement, la désambiguïsation ou la sélection parmi ces étiquettes possibles de la seule étiquette correcte.

Pour implémenter notre outil d'analyse morphosyntaxique, nous suivons principalement ces étapes. Nous leur ajoutons celle de segmentation des transcriptions en des phrases (énoncés). C'est une étape primordiale qui nous permet de bien cerner les segments que nous désirons désambiguïser ses composants lexicaux.

Donc, notre méthode d'analyse morphosyntaxique suit les étapes suivantes :

— **La segmentation du texte en phrases** : cette étape est réalisée par notre segmenteur de textes que nous avons implémenté dans le but de segmenter les transcriptions du DT.
— **L'analyse morphologique** : notre analyseur morphologique permet de réaliser les deux principales étapes de l'analyse morphosyntaxique à savoir : la segmentation du texte en unités lexicales et l'association à chaque occurrence de toutes les étiquettes possibles. La segmentation du texte en des unités minimales est l'étape de segmentation des mots en identifiant leurs composants : le lemme, les affixes et les clitiques. Étant donné que nous traitons l'oral transcrit, nous enrichissons l'analyseur par les étiquettes correspondantes aux composants spécifiques à l'oral.
— **La désambiguïsation morphosyntaxique** : cette étape peut choisir la bonne étiquette morphosyntaxique pour le mot à annoter en prenant en considération son contexte dans la phrase. La désambiguïsation se décompose en deux sous étapes que nous présenterons dans la section 6.7.

6.5 Segmentation des transcriptions en phrases

Les transcriptions d'une séquence sont une schématisation orthographique d'un discours et le signal audio. Cette représentation ne marque pas les frontières de phrases (les signes de ponctuations). De ce fait, pour segmenter nos transcriptions nous proposons en premier lieu d'adapter le segmenteur des textes arabes STAr [Belguith *et al.* 2005] en faveur du DT. Nous proposons aussi de tester une méthode statistique pour la segmentation et aussi l'hybridation de cette méthode avec l'outil adapté pour le DT.

Nous désirons segmenter les transcriptions en considérant chaque tour de parole comme étant un paragraphe. Ensuite, chaque paragraphe est segmenté en phrases.

6.5.1 Corpus

Les méthodes de segmentation présentées dans la section 6.6 se sont basées sur le corpus STAC segmenté manuellement avec deux experts différents. Nous proposons aux experts de segmenter chaque tour de parole de chaque transcription en des énoncés qui sont syntaxiquement corrects. La tâche de segmentation est réalisée comme étant un texte écrit. Les experts sont invités à se référencer aux audio en cas de perplexité. Nous obtenons, ainsi, une mesure de kappa de 0,86. Cette valeur d'accord présente un accord très satisfaisant. Pour choisir un corpus de référence, nous avons fait recours à un troisième expert qui joue le rôle d'un juge pour choisir la meilleure version segmentée du corpus. Ainsi, il a sélectionné la segmentation proposée par le premier expert.

Ce corpus a été divisée en trois parties :

— Corpus d'apprentissage : il est composé de 32 012 mots et 6 133 phrases.
— Corpus de validation : il est composé de 3 175 mots et 440 phrases
— Corpus de test : il est composé de 7 201 mots et 1 215 phrases.

6.5.2 Adaptation de STAr pour le dialecte tunisien

6.5.2.1 STAr : un outil pour la segmentation des phrases en arabe standard

STAr [Belguith *et al.* 2005] est un segmenteur des textes arabes fondé sur une approche de segmentation contextuelle qui utilise des signes de ponctuation et de certains mots-outils et particules. La segmentation de la langue arabe est confrontée à plusieurs difficultés comme l'ambiguïté vocalique des mots, l'ambiguïté dérivationnelle et l'ambiguïté structurelle, l'absence des signes de ponctuation et l'agglutination.

Afin de surmonter ces problèmes, [Belguith *et al.* 2005] ont proposé une approche de segmentation de textes arabes basée sur l'exploration contextuelle des signes de ponctuation, des mots connecteurs ainsi que certaines particules telles que les conjonctions de coordination. Ils ont employé l'exploration contextuelle pour étudier les contextes droit et gauche de chaque mot ou particule jouant le rôle de séparateur de phrases. STAr est construit à base de 183 règles de segmentation. [Belguith *et al.* 2005] ont classé les règles en trois classes relatives aux trois types de marqueurs déclencheurs à savoir les signes de ponctuation, les particules et les mots connecteurs [Baccour *et al.* 2003].

L'évaluation de STAr sur un corpus de l'ASM a prouvé la performance de la méthode proposée pour la segmentation des textes arabes. Ils ont obtenu respectivement 88,26 % et 80,65 % pour les mesures de rappel et de précision.

6.5.2.2 Méthode proposée

L'étude du corpus STAC nous permet de déduire que le DT partage plusieurs marqueurs pour la segmentation des phrases avec l'ASM. Par exemple, la conjonction de coordination (و, w, « *et* ») agglutinée à un verbe présente souvent une marque de début de phrase. De

même, le mot connecteur (امّا, AmA, « *mais* ») présente, dans un certain cas, une marque pour la segmentation des phrases en DT. De ce fait, nous proposons d'adapter le segmenteur STAr en proposant un ensemble de règles permettant la segmentation des transcriptions en DT [Zribi *et al.* 2016].

Nos règles de segmentation ont pour rôle de détecter le mot début d'une phrase ou les mots qui appartiennent aux contextes droits et gauche du premier mot d'une phrase. Ces règles suivent la même structure des règles proposées par [Belguith *et al.* 2005] lors du développement du segmenteur STAr. Elles ont la forme présente dans le tableau 6.2.

Contexte gauche	Marqueur	Contexte droit
G	X	D

TABLE 6.2 – Forme d'une règle contextuelle.

G, X et D présentent des unités lexicales qui peuvent être le début d'une phrase. Soit un marqueur déclencheur X si le contexte gauche de X est G et/ ou le contexte droit de X est D alors X ou D peuvent être le début d'une phrase.

Nous concevons nos règles en se basant sur un ensemble de marqueurs. Ces marqueurs sont dans certains cas spécifiques à la forme orale du DT. Dans d'autres cas, ils sont utilisés dans la forme écrite du dialecte. Ainsi, nous classons nos règles suivant ce critère.

Le premier ensemble regroupe des règles qui permettent souvent de segmenter les formes orales des énoncés. Ces règles utilisent des unités lexicales qui sont généralement utilisées dans la forme orale des transcriptions. Elles se basent aussi sur des marqueurs spécifiques à l'oral comme les pauses silencieuses et les pauses remplies. Les pauses silencieuses présentent dans 57,25 % des cas des marques de début de phrase. Elles jouent le rôle d'un point dans les textes écrits. Par contre, dans les autres 42,74 %, les pauses silencieuses sont situées dans le contexte du début d'une phrase. De même, les pauses silencieuses et les onomatopées peuvent être des marqueurs qui aident à localiser le point de départ d'une phrase. En se basant sur ces deux marqueurs spécifiques à l'oral, nous avons pu dégager 6 règles. Nous présentons dans ce qui suit (*voir tableau 6.3*) un exemple de règle qui utilise les pauses silencieuses.

Si le marqueur est égal à une pause silencieuse « # » et le contexte gauche appartient à cette liste de mots alors la pause est une marque de début de phrase.

Les expressions de remerciement (يعيشك yʕyšk « *merci* », يعطيكم الصحة yʕTykm AlSHħ « *j'espère que vous serez en bonne santé* », etc.), de salutation (صبَاح الخير SbAH Alxyr, صبَاح النور SbAH Alnwr, عَالسلَامة ʕAlslAmħ, مرحبَا mrHba, السلَام AlslAm, السلَام عليكم AlslAm ʕlykm), les expressions de supplication (ربي يهديك rby yhdyk, etc.), etc., un locuteur en DT les utilise souvent dans son discours. Ces expressions se sont situées dans 90 % des cas au début d'un énoncé. En utilisant ces expressions, nous pouvons dégager 10 règles. Par exemple, la règle suivante (*voir tableau 6.4*) est traduite comme suit : si le marqueur est égal à une expression de salutation, donc ce marqueur est le début d'une phrase.

Le deuxième ensemble de règles est plus générique. Ces règles peuvent être appliquées à

Contexte gauche	Marqueur	Contexte droit
Adverbe interrogatif : (هل hl | علَاش ʕlAš | علَاه ʕlAh | قدَاش qdAš | قدَاه qdAh | وقتَاش wqtAš | وقتَاه wqtAh | كيفَاش kyfAš | كيفَاه kyfAh) **Expression qui marque le temps ou un adverbe de temps :** (و w |)(عَام كل kl ʕAm | شهر كل kl šhr | يوم كل kl ywm | نهَار كل kl nhAr | مرة كل kl mrħ | توة twħ | لحظة كل kl lHĎħ | دقيقة كل kl dqyqħ | ليلة كل kl lylħ | تو tw | اليوم Alywm | الآن AlĀn | غدوة γdwħ | بعد غدوة bʕd γdwħ | اللّيلة Allylħ | السنة Alsnħ | البَارح AlbArH | أَولت البَارح Âwlt AlbArH | بعد شوية bʕd šwyħ | آمس Āms | أَولت آمس Âwlt Āms) **Pronom personnel :** (و w |)(آنَا ĀnA | إِنتي Ănty | هي hy | هو hw | أَحنَا ÂHnA | نحنَا nHnA | إِنتم Ăntm | إِنتومَا ĂntwmA | هومَا hwmA) **Pronom démonstratif :** (لهذَا lhðA | لذَا lðA | هذَا hðA | هذية hðyħ | هذي hðy | هذَاية hðAyħ | هذَايَا hðAyA | هذَاكه hðAkh | هذم hðm | هذومَا hðwmA | هَاكه hAkh | هَاك hAk | هَاكم hAkm | هَاو hAw | آهو Āhw) **Etc.**	#	

TABLE 6.3 – Exemple d'une règle contextuelle basée sur la pause silencieuse

Contexte gauche	Marqueur	Contexte droit
	صبَاح الخير SbAH Alxyr| صبَاح النور SbAH Alnwr | عَالسلَامة ʕAlslAmħ| مرحبَا mrHbA | السلَام AlslAm| السلَام عليكم AlslAm ʕlykm	∅

TABLE 6.4 – Exemple d'une règle basée sur les expressions de salutation

des formes écrites du DT. Leur conception se base sur les mots connecteurs, les pronoms personnels, relatifs, les verbes, etc. Les structures syntaxiques du DT sont assez complexes. Ainsi, nous trouvons une difficulté de dégager des patrons pour détecter les frontières des phrases. Étant donné que notre corpus est un corpus oral avec un degré de spontanéité (spontané et semi-spontané) assez forte (95,65 %). Les phrases avec une structure syntaxique simple présente uniquement 15,86 % du nombre total de phrases. Par exemple, la règle suivante permet de détecter des phrases qui commencent par des verbes précédés par une des expressions définies dans le contexte droit.

Contexte gauche	Marqueur	Contexte droit
	Verbe	على الَاقل ʕlỳ AlAql | أَيَا ÂyA| مَا لَا mA lA | ومَا لَا wmA lA | لَا lA | وَإِلَا wĂlA | ولَا wlA| معنَاتهَا mʕnAthA | معنَاهَا mʕnAhA | يعني yʕny | بمعنَى bmʕnỳ | إِذَا ĂðA | لهنَا lhnA | يَا خويَا yA xwyA | بعد شوية bʕd šwyħ

TABLE 6.5 – Exemple d'une règle contextuelle basée sur les verbes.

Dans ce cadre, nous réussissons à dégager 7 règles. Ainsi, nous extrayons au total 23 règles contextuelles. Lors de la conception de nos règles, nous conservons uniquement les règles dont

leur précision est supérieure à 50 %. Ceci justifie le nombre relativement réduit des règles extraites en le comparant au nombre de règles proposées (183 règles) par [Belguith *et al.* 2005] lors de la conception du segmenteur STAr. L'implémentation de ces règles suit un ordre bien défini : le résultat de la règle n°i+1 rectifie le résultat de la règle n°i. Ainsi, nous dégageons un ensemble de marqueurs qui sont généralement utilisés pour désigner le début d'une phrase. Les règles qui se basent sur ces marqueurs rectifient le résultat de l'application des autres règles. Ces règles de correction se sont basées sur quelques marqueurs tels que par exemple les pauses silencieuses et les pauses remplies. Elles permettent de fusionner les phrases qui sont composées d'un seul mot (pause silencieuse, pause remplie, etc.) aux phrases suivantes. Nous fusionnons uniquement les mots qui ne peuvent pas être des phrases complètes. Nous prenons l'exemple du paragraphe suivant en DT :

آ ÷ لكن هو قَال إِنتي مشيت لّلدَار النشر # علَاش آ ÷ معنَاتهَا مَا قلت ليش

Ā÷ lkn hw qAl Ănty mšyt lldAr Alnšr # ʕlAš Ā÷ mʕnAthA mA qlt lyš

« *euh mais, il a dit que tu as visité la maison d'édition, pourquoi, euh, c'est-à-dire, tu n'as pas m'informé* »

La règle contextuelle n°1, qui se base sur les pronoms personnels, propose de chercher la combinaison (pronom + contexte droit). Si cette combinaison est trouvée alors les mots-clés correspondants aux mots-clés sont considérés comme le début de la phrase. L'application de cette règle à notre paragraphe permet de la segmenter en deux parties.

Contexte gauche	Marqueur	Contexte droit
	Pronom personnel	لكن lkn | إِن شَا الله Ăn šA Allh (بَالله bAllh| برب brby | وَالله wAllh) (معنَاتهَا mʕnAthA | معنَاهَا mʕnAhA | يعني yʕny |بمعنَى: bmʕnỳ) (هل hl | علَاش ʕlAš | علَاه ʕlAh | قدَاش qdAš | قدَاه qdAh | وقتَاش wqtAš | وقتَاه wqtAh | كيفَاش kyfAš | كيفَاه kyfAh | بَاش bAš) (لكن lkn) Etc.

TABLE 6.6 – Exemple d'une règle contextuelle basée sur les pronoms.

La règle contextuelle basée sur les pauses silencieuses segmente, aussi, le deuxième segment en deux autres parties. Nous remarquons que l'application de ces règles à segmenter ce paragraphe en trois phrases dont la première est erronée. Ainsi, nous proposons d'appliquer la règle de correction. Cette règle corrige les segmentations erronées en se basant sur les pauses silencieuses et les pauses remplies. Ainsi, elle fusionne les deux phrases en une seule. D'où, le processus de segmentation a segmenté le paragraphe en deux phrases (*voir* tableau 6.7).

L'application de ces règles sur notre corpus de développement a donné respectivement les valeurs suivantes : 62,56 pour le rappel, 85,26 pour la précision et 72,17 pour la F-mesure. Nous remarquons que ces valeurs sont acceptables mais, insuffisantes, car elles sont calculées sur la base du corpus de développement. Ce qui nous encourage à tester d'autres méthodes.

Phrase	آ ÷ لكن هو قَال إِنتي مشيت لّلدَار النشر # علَاش آ ÷ معنَاتهَا مَا قلت ليش Ā÷ lkn hw qAl Ănty mšyt lldAr Alnšr # ʕlAš Ā÷ mʕnAthA mA qlt lyš	
Application de la règle basée sur les pronoms personnels	لكن هو قَال أَنِتي مشيت لّلدَار النشر # علَاش آ ÷ معنَاتهَا مَا قلت ليش lkn hw qAl Ănty mšyt lldAr Alnšr # ʕlAš Ā÷ mʕnAthA mA qlt lyš	آ÷ Ā÷
Application de la règle basée sur les pauses silencieuses	لكن هو قَال إِنتي مشيت لّلدَار النشر lkn hw qAl Ănty mšyt lldAr Alnšr	# علَاش آ ÷ معنَاتهَا مَا قلت ليش # ʕlAš Ā÷ mʕnAthA mA qlt lyš	آ÷ Ā÷
Application de la règle de correction	آ ÷ لكن هو قَال إِنتي مشيت لّلدَار النشر Ā÷ lkn hw qAl Ănty mšyt lldAr Alnšr	# علَاش آ ÷ معنَاتهَا مَا قلت ليش # ʕlAš Ā÷ mʕnAthA mA qlt lyš

TABLE 6.7 – Exemple d'application des règles pour la segmentation d'un paragraphe en DT.

6.5.3 Une méthode statistique pour la segmentation des phrases

La méthode à base de règles pour la segmentation de notre corpus a donné des résultats acceptables mais insuffisants à notre tâche. De ce fait, nous choisissons d'expérimenter avec la méthode statistique PART [Mohamed *et al.* 2012]. Au niveau de cette méthode, nous convertissons la tâche de détection des frontières des phrases en une tâche de classification des mots en quatre classes différentes. La classe « B-S » est pour un mot situé au début de phrase. « I-S » et « E-S » sont deux étiquettes pour marquer les mots qui sont situés respectivement au milieu et à la fin d'une phrase. La classe « S » est pour marquer les phrases qui sont formées d'un seul mot [Zribi *et al.* 2016].

Nous avons expérimenté avec plusieurs méthodes de classification inclus dans l'outil WEKA. Cependant, PART a donné les meilleurs résultats pour notre tâche.

Le résultat d'un classificateur est fortement influencé par l'ensemble des attributs définis. Dans la littérature la tâche de détection des frontières de phrases pour la langue parlée est liée principalement aux deux types d'attributs : des attributs linguistiques et des attributs prosodiques.

L'analyse et les informations prosodiques sont totalement absentes dans le travail proposé dans cette thèse. Ainsi, nous nous limitons aux deux attributs qui sont utilisés au niveau de la prosodie. Ils sont les pauses silencieuses et les pauses remplies. Dans la conception de nos attributs, nous nous basons sur les attributs linguistiques. Nous définissons des attributs qui sont principalement liés aux mots-clés utilisés dans nos règles pour la conception de STAr version DT.

Ainsi, nous choisissons d'utiliser de même des attributs qui sont contextuels. Nous nous référons aux contextes avant et après les mots à classer. Ainsi pour fixer, la fenêtre utilisée, nous testons plusieurs contextes. Nous expérimentons avec une fenêtre nulle, une fenêtre de +/- 1 mot et une fenêtre de +/- 2 mots. Le choix d'une telle fenêtre est justifié par le fait que les phrases en DT ne sont pas assez longues. De même les mots-clés de la détection des mots de début de phrase sont situés dans une fenêtre qui ne dépasse pas deux mots en avant et deux mots en après de la frontière d'une phrase. De même, nous définissons des attributs dynamiques. Ces attributs prennent la classe des mots qui précèdent le mot à classer.

Le nombre de ces attributs suit la fenêtre que nous avons définie. Le tableau 6.8 présente la liste des attributs utilisés pour la tâche de segmentation des phrases en DT. Nous notons que ces attributs prennent deux valeurs possibles : vrai ou faux. Ils spécifient si un mot dans le contexte appartenant aux valeurs possibles définies.

Attributs	Exemple de valeurs possibles
Pause silencieuse	#
Pause remplie	آ÷ Ā÷, أَهّه Âhh, إِهّه Ăhh, أَيوَاه ÂywAh, etc.
Expression de début de phrase	لكن lkn, ولكن wlkn, سوَاء swA', سوَاءَا swA'A, etc.
Particule conditionnel	ولأَن wlÂn, ولأَني wlÂny, ولأَنه wlÂnh, ولأَنهَا wlÂnhA, , etc.
Marqueur discursif	معنَاتهَا mçnAthA, معنَاهَا mçnAhA, يعني yçny, فهمت fhmt, etc.
Expression pour marquer le lieu	ثمة θmħ, غَادي γAdy, etc.
Le verbe « *vouloir* »	حبيت Hbyt, نحبوَا nHbwA, تحبوَا nHbwA, نحب nHb, يحب yHb, etc.
Le verbe « *dire* »	تقول tqwl, يقول yqwl, نقولوَا nqwlwA, قولوَا qwlwA, قلت qlt
Verbes	Liste des verbes du DT
Pronom personnel	آنَا ĀnA, إِنتي Ănty, هي hy, هو hw, أَحنَا ÂHnA, نحنَا nHnA, etc.
Verbe « *être* »	كَان kAn, كَانت kAnt, تكون tkwn, كَانك kAnk, كنت knt, كَانك kAnk, etc.
Pronom relatif	وَاللّي wAlly, مَاللّي mAlly, اللّي Ally
Pronom démonstratif	هذية hðyħ, هذي hðy, هذَا hðA, هذَايَا hðAyA, هذَاكه hðAkh, etc.
Expression pour marquer le temps	كل عَام kl çAm, كل شهر kl šhr, كل نهَار kl nhAr, الحظةlHĎħ, etc.
Adverbe interrogatif	علَاه çlAh, قدَاش qdAš, وقتَاش wqtAš, كيفَاه kyfAh, بَاش bAš, etc.
Expression spéciale	بجَاه bjAh, بصرَاحة bSrAHħ, بكل صرَاحة bkl SrAHħ, بقدرة bqdrħ
Expression de salutation	عَالسلَامة çAlslAmħ, مرحبَا mrHbA, صبَاح الخير SbAH Alxyr, , etc.

TABLE 6.8 – Liste des attributs de la tâche de segmentation des phrases.

Nous remarquons que l'application d'un modèle statistique à la tâche de segmentation des phrases a donné des résultats nettement meilleurs que ceux donnés en appliquant STAr adapté pour le DT. Les valeurs de précision et de rappel calculées sur la base du corpus de développement ont été améliorées respectivement de 9,73 % et 7,83 %.

6.5.4 Une méthode hybride pour la segmentation des phrases

Après le test de la méthode à base de règles contextuelles et la méthode statistique pour la détection des frontières des phrases, nous remarquons que les valeurs de l'évaluation sur le corpus de développement sont insuffisantes. Ainsi, nous proposons de combiner ces deux méthodes [Zribi *et al.* 2016].

D'abord, nous proposons d'analyser, d'abord, les transcriptions avec STAr-TUN basé sur les règles contextuelles. Puis, nous analysons les phrases résultantes avec la méthode d'apprentissage. Nous considérons que les phrases sont bien segmentées si elles sont courtes, c'est-à-dire, leur nombre de mots est inférieur à un nombre n. Par contre, les autres phrases dont leur nombre de mots est supérieur à n peuvent être segmentées en d'autres phrases. De ce fait, nous choisissons ces phrases longues pour être segmentées de nouveau avec le segmenteur

basé sur le modèle d'apprentissage. Nous avons fixé n à 9 car c'est la valeur qui donne les meilleurs résultats pour les valeurs de rappel et précision calculées sur la base du corpus de développement. Nous avons reporté une valeur de précision de 86,06 % et de 69,16 % pour le rappel. Nous remarquons que ces valeurs sont nettement supérieures à ceux qui sont reportées en utilisant uniquement STAr adapté au DT. Par contre, cette méthode d'hybridation a diminué les valeurs de rappel et de précision en passant respectivement de 72,3 % à 69,16 % et 93,1 % à 86,06 % pour la méthode d'apprentissage.

La deuxième stratégie est l'inverse de la première stratégie. Elle consiste à analyser, d'abord, les transcriptions via le modèle généré par la méthode d'apprentissage. Puis, nous choisissons les phrases longues pour être analysées de nouveau avec le segmenteur basé sur les règles contextuelles. Le test de cette stratégie sur le corpus du développement a donnée des résultats moindres que ceux apportés avec la première stratégie d'hybridation. Les valeurs de rappel et de précisions sont diminuées respectivement de 5,13 % et 14,31 % pour le rappel et précision. Par contre, cette stratégie d'hybridation a amélioré la valeur de rappel en comparant avec STAr adapté pour le DT, mais la valeur de précision a diminué de 13,51 %. Ces valeurs sont calculées en appliquant la méthode sur le corpus de développement.

La troisième stratégie consiste à améliorer STAr-TUN basé sur les règles contextuelles en intégrant le modèle généré par la méthode d'apprentissage, c'est-à-dire, les deux méthodes de segmentation (à base de règles et à base d'apprentissage) fonctionnent de façon simultanée. Pour le faire, nous choisissons d'appliquer une des méthodes statistiques qui appartiennent à la famille des méthodes de classification à base de règles pour l'apprentissage de notre modèle statistique. Ainsi, les méthodes appartenant à cette famille fournissent un ensemble de règles dont le rôle de chacune est de proposer une classe pour l'instance à classer. Les règles résultantes ont, généralement, la forme suivante : « *Si condition Alors Conclusion* ». Nous proposons d'intégrer les règles générées par le modèle statistique dans le segmenteur STAr-TUN basé sur les règles contextuelles. Nous proposons de choisir uniquement les règles qui sont responsables d'attribuer la classe « B-S ». Ces règles sont capables de détecter les mots situés au début d'une phrase. Pour ne pas falsifier le résultat de STAr-TUN, nous pensons à intégrer uniquement les règles avec les meilleurs scores. Ainsi, nous attribuons à chaque règle un score. Ce dernier reflète la capacité de cette règle de bien classer les mots. Pour calculer les scores de chaque règle, nous appliquons cette règle de nouveau à la totalité du corpus de validation. Nous calculons le pourcentage de succès et d'échec de cette règle. Si le nombre de mots bien classés avec cette règle est supérieur à 75 % des mots classés, nous considérons que cette règle est pertinente et nous l'ajoutons à l'ensemble des règles contextuelles utilisées par la version STAr adaptée au DT. Pour choisir les règles pertinentes, nous avons utilisé un corpus de validation composé de 440 phrases. Nous avons calculé les scores pour chaque règle. Nous avons constaté que 40 règles classifient faussement les mots. Elles attribuent les étiquettes « B-S » aux mots qui sont situées au milieu des phrases. Les autres 5 règles sont des règles que nous ont extraits manuellement et elles sont déjà implémentées dans STAr.

Donc, cette stratégie d'hybridation a abouti à justifier notre choix pour une méthode de

classification à base de règles. En effet, cette méthode est capable de générer des règles qui sont proches de celle extraites manuellement.

6.6 Analyse morphologique du dialecte tunisien parlé

La définition de la liste des étiquettes pour l'étiquetage morphosyntaxique est une étape primaire afin de bien cerner le groupe de catégories que nous désirons attribuer aux mots de notre corpus pour le DT. Étant donné que notre tâche est la désambiguïsation du résultat de l'analyse morphologique, une étude de la liste des étiquettes utilisée par l'analyseur doit être réalisée. Ainsi, la liste des étiquettes proposée par Al-Khalil-ASM a été modifiée pour qu'elle soit appropriée pour l'analyse morphosyntaxique. Ainsi, nous rappelons que nous avons choisi d'utiliser la liste des étiquettes qui ont été exploitées par l'analyseur morphologique MADA [Habash & Rambow 2005] pour annoter les caractéristiques morphologiques de chaque mot analysé.

Al-Khalil-TUN est une version adaptée de la version développée pour l'ASM écrit. Dans le chapitre précédent, nous avons présenté une adaptation de cet analyseur en faveur du DT de façon générale sans prendre en considération la langue parlée. Ainsi, notre objectif tout au long de cette thèse est le traitement automatique du dialecte parlé. Nous adaptons l'analyseur pour analyser les mots qui se trouvent dans le discours oral. Nous avons amélioré la liste des étiquettes proposées par [Habash & Rambow 2005] par un ensemble des étiquettes de catégorie grammaticale pour marquer les onomatopées *onom*, les pauses remplies *FPause*, les pauses silencieuses *break* et les mots tronqués *TrunW*.

6.7 Désambiguïsation morphosyntaxique

6.7.1 Choix de la méthode

La désambiguïsation morphosyntaxique consiste à étudier le mot à annoter afin de choisir une étiquette parmi les étiquettes attribuées. La désambiguïsation peut être réalisée suivant trois différentes approches.

Un étiqueteur à base de règles intègre des règles de décisions qui sont souvent des règles génériques écrites à la main. Elles correspondront à un contexte approprié pour marquer la bonne étiquette pour un mot donné. La majorité des systèmes qui utilisent cette approche obtient de très bons résultats mais ils demandent un investissement humain conséquent. Cette approche est performante sur des textes bien écrits mais leur performance diminue considérablement sur des textes bruités (notamment ceux issus de la transcription orale).

La deuxième approche est l'approche numérique (ou stochastique). Ce type d'approche se base sur un algorithme d'apprentissage à partir de larges corpus de textes pré-étiquetés par un expert. L'étiqueteur apprend la probabilité de chaque mot pour avoir une étiquette. Le résultat de l'apprentissage sera sauvegardé pour être utilisé dans le module de désambiguïsation morphosyntaxique. Cette approche peut être très vite adaptée à tout type de textes, et plus

particulièrement, elle est robuste lorsque les textes sont bruités, c'est-à-dire, mal écrits. Les systèmes, qui utilisent cette approche, donnent des résultats moins précis que les systèmes à base de règles.

La troisième approche est une approche hybride qui regroupe les méthodes exploitées dans les deux autres approches. Souvent, cette approche donne des résultats satisfaisants en la comparant aux deux autres.

L'étude de l'état de l'art menée sur la langue arabe et ses variantes nous a prouvé que l'approche numérique était souvent utilisée lors du développement des étiqueteurs morphosyntaxiques suivant une étape de désambiguïsation. Les méthodes à base de chaine de Markov cachées (HMM), les champs conditionnels aléatoires (CRF), le séparateur à vaste marge (ou les machines à vecteur de support) (SVM), etc. sont les plus célèbres méthodes employées pour la tâche d'étiquetage. Ainsi, peu de travaux ont proposé des méthodes pour la désambiguïsation morphosyntaxique pour la langue arabe ainsi que pour les dialectes arabes. Le travail de [Habash *et al.* 2013] est à notre connaissance le seul travail qui a proposé une méthode de désambiguïsation du résultat de l'analyse morphologique pour accomplir la tâche d'étiquetage morphosyntaxique. En effet, [Habash *et al.* 2013] ont entrainé pour chaque caractéristique morphologique un classificateur SVM (séparateur à vaste marge). Ils ont combiné le résultat de ces classificateurs afin de choisir la bonne étiquette pour chaque mot. Leur méthode a été testée aussi pour étiqueter l'ASM.

Notre méthode pour la désambiguïsation est fortement inspirée de celle proposée par [Habash *et al.* 2013]. Nous suggérons de tester deux méthodes d'apprentissage statistique pour la désambiguïsation du résultat de l'analyse morphologique. Nous expérimentons deux classificateurs à base de règles et un séparateur à vaste marge (SVM). L'utilisation des machines à vecteur de support est approuvée par le fait que plusieurs chercheurs ont testé cette méthode pour la tâche de désambiguïsation que nous voulons bien la tester pour réaliser la tâche de désambiguïsation morphosyntaxique du DT.

Ainsi, les classificateurs à base de règles sont utilisés aussi pour la tâche de désambiguïsation de l'analyse morphosyntaxique pour l'ASM et le dialecte égyptien [Habash *et al.* 2013]. De ce fait, nous voulons tester la méthode utilisée par [Habash *et al.* 2013], *RIPPER*, et tester d'autres méthodes appartenant à l'approche statistique et voir l'impact de ces méthodes pour notre tâche. De même, notre choix est aussi justifié par le fait que ces algorithmes d'apprentissage génèrent un ensemble de règles qui sont compréhensibles que nous pouvons rectifier ou combiner avec d'autres développées manuellement afin d'améliorer les résultats de la tâche de désambiguïsation.

6.7.2 Présentation des méthodes statistiques

6.7.2.1 Classificateur à base de règles

Nous avons testé plusieurs méthodes de classification à base de règles. Nous présentons dans cette section deux méthodes qui ont reporté les meilleurs résultats. Nous rap-

pelons que nous avons choisi d'utiliser une méthode de classification à base de règles suite aux bons résultats reportés lors de la désambiguïsation morphosyntaxique de l'ASM [Habash & Rambow 2005]. *RIPPER* et *PART* sont deux classificateurs qui utilisent la technique d'induction des règles prédictives de la forme « *Si condition Alors Conclusion* ».

RIPPER (en anglais, Repeated Incremental Pruning to Produce Error Reduction) [Cohen 1995] qui signifie élagage incrémental répété pour réduire l'erreur, est parmi les plus fameuses techniques d'apprentissage de règles. Il permet de construire un ensemble de règles permettant de classifier avec précision les données d'apprentissage. Une règle extraite avec *RIPPER* est représentée sous la forme d'une conjonction de conditions : « *if T1 and T2 and ... Tn then class Cx* » où (T1 and T2 and ... Tn) est appelé le corps de la règle et *Cx* la classe à prédire. Une condition Ti teste une valeur particulière d'un attribut, et elle prend une des trois formes suivantes : « $An = v$ », « $Ac \geq \theta$ », « $Ac \leq \theta$ » ; avec *An* un attribut nominal et v une valeur parmi plusieurs valeurs que peut prendre *An* ; *Ac* est une valeur continue et θ une valeur quelconque pour *Ac*. En se basant sur un corpus d'apprentissage, l'algorithme *RIPPER* permet de produire des règles. Pour ce faire, il repose sur deux phases : une phase de création d'un ensemble de règles et une phase d'optimisation.

Au niveau de la première phase, RIPPER construit le premier ensemble de règles qui passent par une étape d'élagage. Elle consiste à supprimer toute séquence des conditions de la règle qui réduit la précision de celle-ci. L'objectif de l'élagage d'une règle est de diminuer le taux d'erreur sur les données non représentées dans la première partie du corpus d'apprentissage notamment lorsque l'ensemble contient des données bruitées. La deuxième phase de l'algorithme RIPPER est la phase d'optimisation. C'est l'optimisation de l'ensemble de règles obtenues de la première étape. L'algorithme RIPPER donne de meilleurs temps d'exécution (*Running Times*) comparé à d'autres algorithmes d'apprentissage de règles [Collins & Singer 1999]. Ainsi, il est plus efficace sur des données bruitées et il a une évolutivité quasi linéaire avec le nombre d'exemples dans un corpus d'apprentissage. Aussi, lors de la construction d'une règle, l'algorithme RIPPER essaie de trouver, de manière efficace, le test qui maximise le gain d'information pour un ensemble d'exemples d'apprentissage, en faisant un seul passage sur cet ensemble et ceci pour chaque attribut *A* et toutes les valeurs v qui apparaissent comme des éléments de cet attribut. Toutes les valeurs v sont considérées par l'algorithme *RIPPER* et aucune valeur ne sera négligée.

Le deuxième algorithme que nous avons utilisé est PART : un algorithme d'arbre de décision partiel. Il est développé à base des deux algorithmes C4.5 et RIPPER [Mohamed *et al.* 2012]. La principale différence entre PART et les deux algorithmes C4.5 et RIPPER est l'absence de la phase d'optimisation pour générer des règles précises. PART utilise la stratégie « *separate and-conquer* ». Cette stratégie [2] propose de construire une règle qui prédit au mieux sur une fraction d'instance (conquête), de retirer les instances couvertes par la règle de l'ensemble d'apprentissage (séparer) et réitérer jusqu'à ce qu'on ait épuisé les observations disponibles [Mohamed *et al.* 2012]. *PART* construit à chaque itération un arbre de

2. http ://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/fr_Tanagra_Rule_Induction.pdf

décision et il transforme la meilleure feuille en une règle.

6.7.2.2 Méthode statistique : SVM

Les machines à vecteurs de support ou séparateurs à vaste marge (en anglais *Support Vector Machine, SVM*) [Vapnik 1995] sont une technique largement utilisée pour résoudre les problèmes de classification et de régression. Les SVM ont été utilisés avec succès dans de nombreux domaines de recherche TAL, en général, et pour la tâche d'étiquetage morphosyntaxique en particulier.

Les SVM sont une généralisation des classificateurs linéaires les plus populaires. Elles sont robustes pour les données bruitées. Elles ont une capacité puissante de généralisation surtout en présence d'un grand nombre de caractéristiques et elles sont insensibles aux nombres d'exemples de données d'apprentissage (positives ou négatives).

Avec les SVM, les exemples de données d'apprentissage seront représentés en deux ensembles de vecteurs, un vecteur pour les exemples positifs et un pour les exemples négatifs, dans un espace multidimensionnel. À partir de ces ensembles de vecteurs, les SVM doivent déduire les meilleures combinaisons linéaires de caractéristiques à partir des exemples appropriés. Ces combinaisons sont appelées des vecteurs de support qui définissent un hyperplan dans l'espace des caractéristiques multidimensionnelles. Cet hyperplan sépare les exemples positifs des exemples négatifs en maximisant la marge entre les deux ensembles de données.

Weka[3] est un outil de fouille de données ou Data mining qui permet le classement, le tri, le rassemblement (clustering) des données [Guilleminot 2008]. Cet outil propose de nombreuses implémentations de différents algorithmes de data mining et de classification. Pour notre tâche, nous avons appliqué les algorithmes des SVM, RIPPER et PART pour générer nos modèles d'apprentissage. Pour l'algorithme des SVM, nous avons testé deux implémentations proposées par Weka : SMO [Platt 1998] et LibSVM [Chang & Lin 2011].

6.7.3 Les attributs utilisés

Un algorithme d'apprentissage utilise un ensemble de caractéristiques ou attributs qui présentent les éléments les plus influents sur le résultat de l'apprentissage. Nous avons choisi d'utiliser deux types d'attributs : des attributs morphologiques et des attributs contextuels.

Pour les *attributs contextuels*, nous avons choisi un contexte de +/- 2 mots ; c'est-à-dire, on étudie les deux mots qui sont situés avant et après le mot qu'on désire classer. Ceci aide à mieux désambiguïser les résultats de l'analyse morphologique. Nous justifions notre choix pour ce contexte par deux faits. D'une part, la longueur des phrases en DT est généralement courte. Nous avons comme une moyenne de longueur de phrases neuf mots. Donc une telle fenêtre peut décrire et désambiguïser les annotations morphologiques. D'autre part, les travaux effectués pour la langue arabe prouvent qu'un contexte de +/-2 mots est le meilleur pour la tâche de l'annotation morphosyntaxique.

3. http ://www.cs.waikato.ac.nz/ml/weka/

En outre, nous utilisons des *attributs dynamiques*. Ils permettent l'annotation d'un mot en prenant en compte les catégories grammaticales choisies comme étant correctes pour les mots précédents (les deux mots en avant). Ceci permet de bien choisir la bonne étiquette. En plus, nous définissons l'attribut position : un attribut contextuel. Cet attribut décrit la position du mot dans la phrase. Il peut avoir trois valeurs possibles : B, I, et E qui sont employées respectivement pour désigner le début, l'appartenance et la position finale dans la phrase.

En ce qui concerne les *attributs morphologiques*, nous avons choisi d'utiliser quelques caractéristiques, issues de l'analyse morphologique, comme des attributs pour l'algorithme de classification. Nous avons utilisé onze caractéristiques morphologiques du mot à savoir la catégorie du mot, le nombre, la personne, le genre, la voix, les clitiques et les affixes. Nous avons utilisé aussi une interprétation de quelques caractéristiques morphologiques : l'agglutination à un pronom, l'agglutination à une conjonction, l'agglutination à une particule de négation, l'agglutination à une particule interrogative, l'agglutination à une particule et la présence de marque d'un nom défini. Le tableau 6.9 résume les attributs morphologiques utilisés ainsi que les valeurs possibles de chaque attribut.

Attribut	Abréviation	Valeurs possibles
Catégorie grammaticale	POS	verb, noun, adv, etc. [4].
Personne	Per	1 (première personne), 2 (deuxième personne), 3 (troisième personne), na (n'est pas applicable).
Nombre	Num	s (singulier), d (duel), p (pluriel), u (indéfini), na.
Genre	Gen	f (féminin), m (masculin), na.
Voix	Vox	a (active), p (passive), na.
Agglutination à un pronom	Pron	Yes, no, na.
Agglutination à une conjonction	Conj	Yes, no, na.
Agglutination à une particule de négation	Neg	Yes, no, na.
Agglutination à une particule interrogative	Intero	Yes, no, na.
Nom défini	Def	Yes, no, na.
Agglutination à une particule	Part	Yes, no, na.

TABLE 6.9 – Liste des attributs morphologiques.

Notons que cette liste des attributs est inspirée de plusieurs travaux portant sur l'annotation morphosyntaxique de l'ASM et des dialectes arabes. Nous testons, ainsi, plusieurs combinaisons de ces attributs. La combinaison qui attribue le meilleur score est retenue.

6.7.4 Classification des résultats

Notre objectif est la désambiguïsation des résultats de l'analyse morphologique c'est-à-dire choisir la bonne étiquette morphologique pour un mot donné. Par conséquent, nous proposons de convertir la tâche de désambiguïsation en une tâche de classification. Nous proposons

4. adj, interrog_adv, adv_place, adv_temp, break, FPause, conj, sub_conj, fw, ind_obj_pron, interj, noun_count, prop_noun, number, number_noun, onom, part, part_abst, part_cond, part_fut, part_interrog, part_neg, part_restrict, part_verb, part_voc, prep, pron, dem_pron, poss_pron, rel_pron, rel_adv, sub_conj, TrunW, verb.

d'attribuer à chaque analyse soit l'étiquette « *true* » pour la bonne analyse selon le contexte du mot dans une phrase, soit l'étiquette « *false* » pour les autres analyses. Nous étudions les analyses du mot en cours et les analyses des autres mots situés au voisinage. Puis, nous attribuons l'étiquette « *true* » pour la combinaison qui est correcte. La figure 6.1 présente un exemple d'attribution des classes.

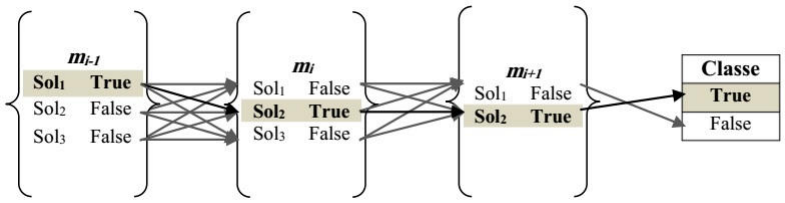

FIGURE 6.1 – Attribution des classes aux analyses du mot mi suivant son contexte.

6.7.5 Choix du résultat

Le rôle de chaque classificateur est d'attribuer une étiquette (true ou false) pour chaque analyse proposée par notre analyseur morphologique. Dans certains cas, le classificateur échoue à bien spécifier la bonne analyse pour un mot. Il peut attribuer plusieurs étiquettes « true » pour plusieurs analyses d'un mot ou il échoue à choisir une analyse comme étant correcte. Pour résoudre ce problème, nous proposons une stratégie pour choisir la bonne analyse suivant les caractéristiques morphologiques (catégorie grammaticale, genre, nombre, etc.).

Nous proposons de combiner le résultat du classificateur avec un autre classificateur (basé sur les n grams) pour réaliser cette tâche dans le cas d'échec. Ce dernier a pour rôle de chercher l'analyse correcte pour un mot donné. Il choisit une seule analyse pour le mot dans son contexte. Ce classificateur utilise des bi-grammes. Il annote les analyses suivant les fréquences d'apparition des bi-grammes d'étiquettes qui sont stockés dans un dictionnaire. L'étiqueteur choisit ainsi la solution qui a la plus grande fréquence d'apparition. Nous signalons que la méthode des modèles de langue a été utilisée par [Hamdi 2015] pour le choix des étiquettes morphosyntaxiques au niveau de l'étiquetage morphosyntaxique du DT.

6.8 Expérimentations et évaluation

6.8.1 Les mesures d'évaluation

Pour l'évaluation de notre segmenteur des transcriptions en des phrases, nous calculons les mesures de rappel, de précision et de F-mesure en nous basant sur le nombre de phrases correctement segmentées.

$$Rappel_{seg} = \frac{Nombre\ de\ phrases\ correctement\ segmentées}{Nombre\ total\ de\ phrases} \tag{6.1}$$

$$Précision_{seg} = \frac{Nombre\ de\ phrases\ correctement\ segmentées}{Nombre\ de\ phrases\ reconnues\ par\ le\ segmenteur} \tag{6.2}$$

Pour l'évaluation de notre analyseur morphosyntaxique, nous utilisons les mesures de rappel, de précision et F-mesure. Dans une première évaluation, nous nous concentrons sur l'évaluation de la tâche de classification proposée par nos classificateurs. Nous cherchons le nombre des analyses correctement classées. Dans ce sens, nous calculons deux mesures. La première consiste à évaluer nos données en attribuant aux attributs dynamiques uniquement les bonnes étiquettes. La deuxième prend en considération le dynamisme des attributs. Les mesures de rappel et précision pour chaque classe sont calculées suivant les deux formules ci-dessous.

$$Rappel = \frac{Nombre\ d'analyses\ correctement\ classés}{Nombre\ total\ d'analyses} \tag{6.3}$$

$$Précision = \frac{Nombre\ d'analyses\ correctement\ classés}{Nombre\ d'analyses\ pour\ une\ classe} \tag{6.4}$$

Les mesures de rappel, de précision et de F-mesure présentées, dans la section suivante, sont la moyenne pondérée des valeurs reportées pour toutes les classes (*true* et *false*).

Nous mesurons, aussi, la qualité de désambiguïsation des résultats de l'analyse morphologique. Nous cherchons le nombre de mots correctement désambiguïsés pour chaque classificateur. Nous reportons la valeur d'exactitude qui présente le rapport entre le nombre des mots correctement classés et le nombre total des mots annotés.

6.8.2 Évaluation de la segmentation des phrases

Pour évaluer les différentes méthodes proposées pour la segmentation des phrases, nous avons utilisé un corpus de test composé de 7 201 mots et de 1 215 phrases. Le tableau 6.10 résume les valeurs de mesures de rappel$_{seg}$, de précision$_{seg}$ et F-mesure$_{seg}$.

	Règles contextuelles RC	**Méthode statistique PART**	**Méthode d'hybridation 1**	**Méthode d'hybridation 2**
Rappel$_{seg}$	68,31	72,5	72,427	66,01
Précision$_{seg}$	90,841	94,8	89,16	73,917
F-mesure$_{seg}$	77,891	82,1	79,926	69,738

TABLE 6.10 – Les valeurs de l'évaluation du STAr-TUN, la méthode statistique PART et les deux méthodes d'hybridation.

En examinant les valeurs calculées sur le corpus de test, nous remarquons que les résultats sont très encourageants. En effet, la méthode statistique a montré son efficacité de bien annoter les mots du corpus. Elle a donné des meilleurs résultats par rapport à ceux trouvés par la méthode basée sur l'utilisation des règles contextuelles. Nous remarquons une amélioration des valeurs obtenues de 4 %. Elle a pu détecter un nombre de phrases plus grand que celui de la méthode à base de règles contextuelles (à peu près 4 % d'amélioration).

Ainsi, nous testons plusieurs méthodes d'hybridation. Nous remarquons que l'application de la première méthode d'hybridation permet d'améliorer la valeur de rappel en la comparant avec la valeur reportée en appliquant uniquement STAr-TUN basé sur les règles contextuelles. Nous remarquons une amélioration de 4,12 point. Par contre, la valeur de précision est diminuée de 5,64 points en la comparant avec celle de la méthode d'apprentissage.

Nous voyons aussi que l'application de la méthode statistique suivie par l'application des règles contextuelles a dégradé la valeur de rappel. La valeur a baissé en passant de 72,42 vers 66,01 (une diminution de 6,417 %). De même, la première méthode d'hybridation a donné des résultats supérieurs à ceux reportés avec la deuxième méthode. La valeur de précision est diminuée de 15,24 %. En effet, ces deux méthodes d'hybridation proposent de segmenter de nouveau les phrases longues qui ont un nombre de mots qui dépasse un (n= 9). Nous remarquons que lors du deuxième passage, les phrases longues se sont divisées en très petits segments. Cette segmentation augmente le nombre des phrases de 5 % pour la première méthode d'hybridation. Par contre, elle diminue la précision de la segmentation.

En guise de conclusion, nous remarquons que la méthode de segmentation à base des règles contextuelles ou la méthode de segmentation à base d'une méthode statistique donne des résultats nettement meilleurs que ceux des deux méthodes d'hybridation.

6.8.3 Expérimentations sur la désambiguïsation morphosyntaxique

La qualité de l'application d'une méthode d'apprentissage se base essentiellement sur des choix préalables à la phase d'apprentissage. Le choix des attributs, la définition des attributs, le choix d'une fenêtre, la définition des classes, etc. sont parmi les critères que le développeur doit prendre en considération pour réussir la génération de son modèle d'apprentissage.

De ce fait, nous proposons comme une première expérimentation de tester plusieurs paramètres (les attributs, la fenêtre et l'algorithme de classification) afin de choisir le meilleur pour construire nos classificateurs proposés. Dans une deuxième expérimentation, nous proposons de tester l'impact de l'ajout de l'étape de segmentation sur la qualité de la tâche de désambiguïsation morphosyntaxique. Ensuite, nous examinons l'apport de définition des étiquettes spécifiques à l'oral sur la qualité du résultat de l'analyse morphosyntaxique du DT.

Nous notons que le résultat d'une expérimentation sera l'entrée de l'expérimentation suivante.

6.8.3.1 Paramétrage des classificateurs

Nous avons présenté dans la section précédente un ensemble d'attributs qui ont été testés dans des applications d'étiquetage morphosyntaxique pour la langue arabe et aussi avec les SVM. Nous proposons, ainsi, à tester plusieurs attributs afin de choisir le meilleur ensemble pour notre tâche.

Premièrement, nous proposons de choisir l'ensemble d'attributs qui améliore les résultats de nos classificateurs. Nous proposons d'utiliser uniquement les attributs morphologiques en testant l'effet de l'utilisation de ces attributs. Nous expérimentons avec l'attribut catégorie grammaticale avec les attributs relatifs aux caractéristiques morphologiques (Per, Num, etc.). Ensuite, nous additionnons les attributs relatifs à la tâche de tokenisation (la conjonction, la négation, etc.).

Puis, nous expérimentons avec l'attribut dynamique. Enfin, nous combinons les attributs

morphologiques, dynamiques et contextuels (la position du mot dans la phrase, etc.). Le tableau 6.11 présente les valeurs de F-mesure reportées en appliquant un contexte de +/- 2 mots. Nous notons que les valeurs de rappel, de précision et de F-mesure sont calculées suivant une validation croisée « 10-fold cross-validation ».

Attributs		RIPPER	PART	SMO	LibSVM
Morphologique	POS + Gen, Num, Per, Vox, Asp	0,783	0,789	0,758	0,781
	+ Pron, Neg, Intero	0,786	0,791	0,758	0,781
	+Conj, Def, Part	0,78	0,799	0,783	0,788
Morphologique + dynamique		0,78	0,773	0,775	0,766
Morphologique + contextuel		0,879	0,881	0,863	0,865
Morphologique + dynamique + contextuel		**0,910**	**0,905**	**0,888**	**0,914**

TABLE 6.11 – Les valeurs de F-mesure reportées en utilisant différents attributs.

Nous avons testé deux implémentations proposées par Weka pour l'algorithme SVM : SMO [Platt 1998] et LibSVM [Chang & Lin 2011]. Nous constatons que ces deux implémentations donnent des résultats similaires. Dans certains cas, LibSVM donne des résultats supérieurs à ceux reportés par l'implémentation de SMO. Mais, le principal inconvénient de SMO est assez gourmand en termes de mémoire. De même, le temps nécessaire lors de son exécution est très supérieur à celui reporté en exécutant LibSVM. D'où, nous choisissons d'utiliser l'implémentation LibSVM pour l'algorithme les SVM. Nous remarquons que l'algorithme LibSVM a donné le meilleur résultat pour cette combinaison d'attributs. Par contre l'algorithme PART a montré sa performance pour la plupart des combinaisons des attributs.

D'après les résultats obtenus, nous remarquons que l'utilisation des attributs dynamiques, contextuels et morphologiques donne de meilleurs résultats. Nous constatons, aussi, que l'emploi de l'ensemble des attributs dynamiques couplé avec les attributs morphologiques n'a pas abouti à augmenter les valeurs de f-mesure. Par contre, l'ajout des attributs contextuels aux attributs morphologiques permet d'augmenter les valeurs de F-mesure (une amélioration de 7,7 % à 9,90 %).

Les attributs contextuels ont un rôle important dans la tâche de désambiguïsation. Il est important de chercher la bonne valeur de la fenêtre de mots. Dans la littérature, les travaux portant sur l'étiquetage morphosyntaxique ont employé une fenêtre de +/- 2 mots. Nous essayons d'expérimenter plusieurs contextes afin de choisir le meilleur pour notre tâche de désambiguïsation. Le tableau suivant montre les valeurs de rappel et précision et f-mesure pour une fenêtre de +/- 0, 1 et 2 mots.

D'après le tableau 6.12, nous remarquons que le contexte +/- 2 mots est la meilleure fenêtre pour notre tâche.

6.8.3.2 Apport de la segmentation sur la qualité de l'analyse morphosyntaxique

Nous avons ainsi testé trois corpus pour l'apprentissage et le test : un corpus non segmenté, un corpus segmenté manuellement et un corpus segmenté automatiquement avec les quatre méthodes de segmentation proposées dans la section 6.6. Pour tester l'impact de la segmenta-

		0	1	2
Ripper	**Rappel**	0,825	0,839	0,935
	Précision	0,831	0,828	0,923
	F-mesure	0,828	0,807	**0,910**
Part	**Rappel**	0,892	0,891	0,911
	Précision	0,896	0,886	0,903
	F-mesure	0,892	0,886	**0,905**
SVM	**Rappel**	0,876	0,861	0,937
	Précision	0,878	0,875	0,939
	F-mesure	0,875	0,829	**0,914**

TABLE 6.12 – Résultat de l'évaluation selon trois fenêtres différentes.

tion sur la qualité des classificateurs, nous avons ignoré l'étape d'utilisation de classificateur à base de bi-gramme lors du choix de la bonne étiquette pour les mots.

Le tableau 6.13 présente les valeurs d'exactitude de chaque méthode présentée. Nous remarquons que l'étiquetage morphosyntaxique avec des transcriptions non segmentées a donné la meilleure valeur d'exactitude lors de l'application de l'algorithme PART. Elle a apporté une exactitude égale à 71,88 %. Nous voyons que la segmentation a un impact sur la qualité de la désambiguïsation morphosyntaxique. En effet, la segmentation manuelle a augmenté la valeur d'exactitude pour les deux algorithmes (SVM et Ripper) : une faible amélioration aux alentours de 1 % par rapport à l'utilisation des transcriptions non segmentées. La segmentation automatique a permis d'améliorer les valeurs d'exactitude (une faible amélioration 2%) lors de l'application de la méthode hybride et la méthode à base des règles contextuelles. Mais, les valeurs reportées restent nettement inférieures à celles calculées sur des transcriptions non segmentées.

En étudiant, la segmentation proposée suivant les trois méthodes proposées, nous remarquons que ces dernières proposent de segmenter les transcriptions en des phrases courtes. En étudiant les étiquettes proposées pour les mots appartenant à ces phrases, nous remarquons que le taux d'erreur est très réduit pour les phrases composées d'un nombre petit de mots. Par contre, la méthode de segmentation à base de la méthode d'apprentissage qui donne des segmentations très proches de celles manuelles n'a pas abouti à améliorer la tâche de désambiguïsation morphosyntaxique.

Nous concluons qu'une segmentation plus fine peut être plus performante pour la tâche de désambiguïsation morphosyntaxique du dialecte tunisien parlé.

	Sans segmentation	**Segmentation manuelle**	**Segmentation automatique**			
			PART	**PART + RC**	**RC**	**RC + PART**
Ripper	62,53	63,92	61,69	63,92	**64,84**	64,20
Part	**71,88**	70,55	66,58	68,22	70,65	70,21
SVM	61,87	63,02	61,04	**63,66**	63,04	**63,39**

TABLE 6.13 – Les valeurs d'exactitude reportées sans segmentation, avec segmentation manuelle et avec segmentation automatique.

6.8.3.3 Apport de l'ajout des étiquettes de l'oral

Notre objectif, dans ce chapitre, est l'étiquetage morphosyntaxique du dialecte tunisien oral. De ce fait, nous avons ajouté des étiquettes spécifiques à l'oral à l'ensemble des étiquettes de catégories grammaticales. Pour voir l'impact de l'utilisation des étiquettes de l'oral, nous avons testé les deux ensembles d'étiquettes.

Nous avons ignoré l'étape d'utilisation de classificateur à base de bi gramme lors du choix de la bonne étiquette pour les mots et nous reportons le pourcentage de mots non reconnus pour les trois algorithmes : Ripper, PART et SVM. Le tableau 6.14 montre les valeurs trouvées.

	Avec étiquettes de l'oral	**Sans étiquettes de l'oral**
Ripper	34,83 %	48,49 %
Part	25,66 %	34,92 %
SVM	35,56 %	43,17 %

TABLE 6.14 – Les pourcentages des mots correctement classés avec et sans étiquettes de l'oral.

Nous remarquons une réduction considérable de taux mots inconnus pour les trois classificateurs PART, Ripper et SVM. Par exemple, lors de l'application du classificateur PART, nous constatons que pour certaines catégories grammaticales comme le nom et le verbe le taux de mot non reconnus a diminué respectivement de 0,95 % et de 1,57 %. Le tableau 6.15 montre le taux d'erreur pour quelques catégories grammaticales.

Catégorie grammaticale	**PART**		**RIPPER**		**SVM**	
	Sans	**Avec**	**Sans**	**Avec**	**Sans**	**Avec**
adj	18,52	16,67	25,93	25,93	25,93	25,93
adv	75	65,63	87,50	75	87,50	78,13
dem_pron	20	23,33	26,67	26,67	26,67	26,67
fut_part	40	50	100	90	90	90
ind_obj_pron	87,50	87,50	100	100	100	100
interj	71,43	57,14	100	100	100	100
interrog_adv	8,33	16,67	25,00	33,33	33,33	33,33
neg_part	55,17	68,97	82,76	79,31	89,66	82,76
noun	23,49	22,54	33,02	31,75	33,33	32,38
noun_count	63,64	72,73	81,82	81,82	81,82	81,82
number_noun	28,57	25	35,71	35,71	35,71	35,71
poss_pron	83,33	100	100	100	100	100
prep	31,18	32,26	39,78	37,63	40,86	39,78
prop_noun	10,53	7,89	15,79	13,16	15,79	13,16
restrict_part	90,91	63,64	100	100	100	100
sub_conj	60	40	100	80	100	80
verb	30,37	28,80	48,17	44,50	48,17	44,50

TABLE 6.15 – Les taux d'erreur pour quelques catégories grammaticales avec et sans l'emploi des étiquettes de l'oral.

6.8.4 Comparaison à d'autres systèmes

6.8.4.1 Système de référence « *Baseline* »

L'idée proposée pour le développement du « *Baseline* » est très simple. Elle consiste à attribuer à chaque mot l'annotation la plus fréquente. Pour ce faire, nous créons à partir de notre corpus d'apprentissage un lexique composé de mots avec les ensembles de catégories grammaticales possibles et les fréquences d'apparition de chaque combinaison dans le corpus d'apprentissage. Nous projetons ce lexique sur l'ensemble des mots à annoter et nous attribuons à chaque mot l'étiquette la plus fréquente.

6.8.4.2 L'étiqueteur Stanford appris pour le DT

Afin d'étiqueter morphosyntaxiquement le DT, [Boujelbane 2015] a adapté l'étiqueteur morphosyntaxique Stanford [Toutanvoa & Manning 2000] développé pour l'ASM. Elle a réappris cet étiqueteur en se basant sur un corpus du DT qui est le résultat de la traduction du corpus Arabic Treebank de l'ASM. Nous notons que le corpus utilisé pour l'apprentissage est issu d'une transcription orthographique des émissions d'actualité en ASM. Le corpus traduit est composé d'un grand pourcentage de mots en ASM.

6.8.4.3 Discussion des résultats

Pour comparer les trois systèmes, nous avons utilisé un corpus composé de 6 680 mots. Puisque le système Stanford appris pour le DT ne traite pas les phénomènes de l'oral (les mots incomplets, les mots répétés, les pause, etc.), nous avons filtré tous ces éléments du corpus de test. Nous remarquons aussi, la présence de quelques différences entre l'ensemble des étiquettes utilisées par les deux systèmes. D'où, nous avons essayé de réduire les différences entre ces deux et standardiser les sorties des deux systèmes. Le tableau 6.16 présente les valeurs d'exactitude reportées par ces trois systèmes.

	Exactitude
Notre système	85, 49
Baseline	68,51
Stanford DT	51,82

TABLE 6.16 – Comparaison entre les différents systèmes d'étiquetage morphosyntaxique.

Nous remarquons que notre système a donné la meilleure exactitude. Nous remarquons, aussi, que le système Stanford pour le DT a donné une exactitude nettement inférieure à la valeur apportée par notre système. La différence entre les deux corpus d'apprentissage utilisé pour les deux systèmes est la principale cause de cet écart. En effet, le corpus d'apprentissage du système Stanford est composé de transcriptions d'émissions d'actualité qui traitent des thèmes généralement politique. Le dialecte tunisien utilisé par ce corpus est le dialecte intellectualisé. Par contre, notre corpus est composé de transcriptions qui traitent des thèmes variés. En plus, le pourcentage des mots en DT est très élevé par rapport à celui du corpus de

Stanford DT. En outre, la nature syntaxique du corpus utilisé pour l'apprentissage du Stanford [Boujelbane 2015] est différente de la nature de notre corpus.

6.9 Conclusion

Dans ce chapitre, nous avons proposé la création d'un étiqueteur morphosyntaxique pour le DT. D'abord, nous avons présenté les challenges d'analyse morphosyntaxique du DT. Puis, nous avons justifié notre choix pour développer un nouvel outil pour le DT. Ensuite, nous avons présenté notre méthode pour l'adaptation d'un analyseur morphologique pour le DT. Nous avons présenté notre méthode pour la segmentation des transcriptions en DT en des phrases. Puis, nous avons exposé notre méthode de désambiguïsation morphosyntaxique pour le DT. Enfin, nous avons présenté quelques expérimentations, en discutant les résultats obtenus.

Conclusion générale

La problématique du développement de ressources et d'outils pour le traitement automatique du Dialecte Tunisien (DT) a été abordée dans cette thèse. Pour la mettre en pratique, nous avons proposé et testé deux approches. La première consiste à l'adaptation des ressources de l'Arabe Standard Moderne (ASM) pour traiter le DT. La deuxième concerne le développement des nouveaux outils en faveur de ce dialecte.

Pour défendre nos propositions, nous avons présenté, dans le chapitre 1 et le chapitre 2 de la première partie, un état de l'art du traitement de dialectes arabes en général et du DT en particulier. Nous avons distingué deux approches pour développer les ressources de l'arabe dialectal (AD) : *(i)* une approche à base d'adaptation des ressources existantes de l'ASM pour traiter les dialectes et *(ii)* une approche proposant le développement de nouvelles ressources. La présentation de l'état de l'art et l'analyse des approches proposées nous ont permis de dégager les constations suivantes :

— La problématique de l'absence des règles orthographiques pour les dialectes arabes est peu abordée au niveau de l'état de l'art.

— La création de corpus pour la forme orale des dialectes arabes est rarement abordée dans la littérature.

— En focalisant sur les méthodes proposées pour la création des analyseurs morphologiques et les analyseurs morphosyntaxiques, nous avons constaté que l'approche d'adaptation est souvent exploitée pour aboutir à la création d'analyseurs morphologiques pour l'AD. Par contre, les étiqueteurs morphosyntaxiques sont le résultat de l'application de méthodes statistiques sans passer par l'adaptation des ressources de l'ASM.

Afin de réussir la transcription orthographique du DT qui souffre de l'absence des règles orthographiques, nous avons présenté, au niveau du chapitre 3, deux conventions pour la transcription orthographique et l'annotation des transcriptions, en prenant en compte les spécificités de l'oral. Pour la première convention « OTTA : Orthographic Transcription of Tunisian Arabic », nous avons proposé une transcription orthographique fidèle à la prononciation des mots dialectaux, en respectant les principales règles de transcription orthographique de l'ASM. Nous avons, aussi, proposé au niveau de cette convention un ensemble d'annotations pour enrichir les transcriptions de la forme orale du DT. Afin de réussir l'adaptation des ressources de l'ASM vers le DT, nous avons proposé une deuxième convention pour la transcription orthographique « CODA-TUN : Conventional Orthographic for Dialectal Arabic-TUN ». Cette convention vise à transcrire le DT en se rapprochant autant que possible de l'ASM. Il s'agit d'une extension de la convention orthographique proposée par [Habash *et al.* 2012a] pour transcrire orthographiquement les dialectes arabes, et qui a été proposée pour le dialecte égyptien et d'autres dialecte arabes.

La création du corpus « STAC : Spoken Tunisian Arabic Corpus » a fait l'objet du chapitre 4.

La méthode « télécharger et enregistrer » proposée par [Waibel *et al.* 2004] a été adoptée pour collecter les supports audio afin de les transcrire. Plusieurs transcriptions ont été réalisées pour mesurer la qualité de notre corpus et de nos directives d'annotation. Le coefficient kappa de [Cohen 1995] de 0,67 (en moyenne) a été reporté pour les transcriptions. Cette valeur présente un accord acceptable montrant la qualité de la transcription de notre corpus. Pour améliorer nos transcriptions, un ensemble d'annotations ont été proposées. Nous avons enrichi notre corpus avec les étiquettes morphosyntaxiques. De même, nous avons annoté les traits spécifiques à l'oral, en l'occurrence les disfluences.

La création d'un analyseur morphologique pour le DT a fait l'objet du chapitre 5. L'étude du lexique tunisien nous a mené à proposer une méthode permettant de créer, à partir d'un lexique « root-pattern » de l'ASM, un lexique pour le DT. La création de ce lexique s'est basée sur une méthode à base de règles et aussi une méthode peu supervisée pour son enrichissement. Nous avons exploité ce lexique pour adapter l'analyseur morphologique « Al-Khalil-ASM » [Boudlal *et al.* 2010] de l'ASM vers le DT. L'évaluation de l'analyseur adapté pour le DT a donné les valeurs de 0,991, 0,871 et 0,927 respectivement pour les mesures de rappel, précision et F-mesure.

Pour surmonter le problème de l'ambiguïté au niveau de l'analyse morphologique, nous avons proposé une méthode pour la désambiguïsation morphosyntaxique des résultats de l'analyse morphologique. La description détaillée de cette méthode a été l'objet du chapitre 6. Tout d'abord, nous avons essayé de surmonter les problèmes liés à l'oral lors de la désambiguïsation morphosyntaxique notamment la segmentation des transcriptions en des phrases. Ainsi, nous avons proposé l'adaptation du segmenteur de textes arabes STAr [Belguith *et al.* 2005] pour traiter le DT. Nous avons proposé, d'abord, une méthode à base de règles pour la détection des frontières des phrases en DT. L'évaluation de cette méthode n'a pas abouti à des résultats assez satisfaisants (F-mesure = 78,71). Ensuite, nous avons testé une méthode statistique (PART) pour l'extraction des règles permettant de classer les mots en quatre classes (B-S, I-S, E-S et S) selon leur position dans la phrase. Cette méthode statistique a permis d'améliorer les résultats de segmentation (F-mesure = 82,9). Les valeurs de l'évaluation ont montré que l'utilisation du mécanisme d'apprentissage a permis de remédier au problème de manque d'information linguistique.

Pour réaliser la désambiguïsation morphosyntaxique, nous avons testé plusieurs méthodes statistiques (les SVM, Ripper et PART) qui ont montré leur efficacité. Ces méthodes ont été testées, au début pour désambiguïser les transcriptions sans prendre en compte les spécificités de l'oral. Ainsi, les résultats reportés en termes d'exactitude sont de 85,49 % avec un taux de mots inconnus ou mal annotés de 14,51 %. Ainsi, pour permettre à l'analyseur morphosyntaxique d'annoter les transcriptions orales spontanées, nous avons enrichi notre ensemble des étiquettes avec celles spécifiques à l'oral telles que les mots incomplets, les pauses remplies, etc. Cette adaptation a permis d'améliorer les résultats de l'analyse en réduisant le nombre de mots hors vocabulaire de 34,92% à 25,66% pour l'algorithme PART.

Dans un futur proche, nous comptons tester les outils développés sur d'autres types de

corpus pour le dialecte tunisien tels que la constitution tunisienne transcrit en dialecte et un corpus collecté à partir des réseaux sociaux. Nous envisageons, aussi, de réaliser une évaluation extrinsèque des outils développés dans le cadre des applications du TALN réalisées au niveau de notre groupe ANLP-group telles qu'un système de reconnaissance de parole ou un serveur vocal (homme-machine) afin de mesurer l'apport des outils dans ces applications.

Un autre axe que nous souhaitons suivre est de traiter d'autres spécificités de l'oral au niveau du DT. Nous envisageons, ainsi, d'entamer le problème de détection et de correction des disfluences. Une brève étude de l'état de l'art dans ce cadre a montré que cette tâche est souvent résolue en utilisant des méthodes statistiques (*e.g.* les CRF, les SVM, etc.) et avec la présence d'un corpus. Nous rappelons que notre corpus est enrichi avec des étiquettes spécifiques au phénomène de disfluence. Ainsi, la détection et la correction des disfluences nous amèneront à améliorer la tâche d'analyse morphosyntaxique du DT. De plus, elle nous permettra de proposer des méthodes pour l'analyse syntaxique du DT oral.

Le développement d'autres outils linguistiques pour le DT est un autre axe que nous envisageons de le suivre. Nous souhaitons développer un analyseur syntaxique pour le DT en suivant l'approche d'adaptation des outils de l'ASM vu que les travaux récents sur l'analyse syntaxique au sein de notre laboratoire, sont en faveur de l'application concrète de cette perspective.

Bibliographie

[Abdul-Mageed & Diab 2011] Muhammad Abdul-Mageed et Mona Diab. *Subjectivity and sentiment annotation of modern standard arabic newswire.* In Association for Computational Linguistics, editeur, Proceedings of the 5th Linguistic Annotation Workshop, pages 110–118, Portland, Oregon, USA, June 2011. (Cité en page 23.)

[Abdul-Mageed & Diab 2014] Muhammad Abdul-Mageed et Mona Diab. *SANA : A Large Scale Multi-Genre, Multi-Dialect Lexicon for Arabic Subjectivity and Sentiment Analysis*. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA). (Cité en pages iv, 23, 24 et 26.)

[Abo Bakr *et al.* 2008] Hitham Mohamed Abo Bakr, Khaled Shaalan et Ibrahim Ziedan. *A hybrid approach for converting written Egyptian colloquial dialect into diacritized Arabic*. In The 6th International Conference on Informatics and Systems, INFOS2008, pages 27–33, Cairo, Egypt, 2008. (Cité en page 45.)

[Abuata & Al-Omari 2015] Belal Abuata et Asma Al-Omari. *A rule-based stemmer for Arabic Gulf dialect*. Journal of King Saud University - Computer and Information Sciences, vol. 27, no. 2, pages 104 – 11, 2015. (Cité en pages 32 et 101.)

[Afify *et al.* 2006] Mohamed Afify, Ruhi Sarikaya, Hong-Kwang Jeff Kuo, Laurent Besacier et Yuqing Gao. *On the use of morphological analysis for dialectal Arabic speech recognition*. In INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, September 17-21, 2006. ISCA, 2006. (Cité en pages 27 et 101.)

[Al-Kabi *et al.* 2016] Mohammed Al-Kabi, Mahmoud Al-Ayyoub, Izzat Alsmadi et Heider Wahsheh. *A Prototype for a Standard Arabic Sentiment Analysis Corpus*. The International Arab Journal of Information Technology, vol. 13, no. 1A, pages 163–170, 2016. (Cité en page 8.)

[Al-Sabbagh & Girju 2010] Rania Al-Sabbagh et Roxana Girju. *Mining the Web for the Induction of a Dialectical Arabic Lexicon*. In Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta. European Language Resources Association, 2010. (Cité en pages 23, 24 et 26.)

[Al-Sabbagh & Girju 2012] Rania Al-Sabbagh et Roxana Girju. *YADAC : Yet another Dialectal Arabic Corpus*. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, May 23-25, 2012, pages 2882–2889. European Language Resources Association (ELRA), 2012. (Cité en pages 17, 20 et 21.)

[Al-Saidat & Al-Momani 2010] Emad Al-Saidat et Islam Al-Momani. *Future Markers in Modern Standard Arabic and Jordanian Arabic : A Contrastive Study*. European Journal of Social Sciences, vol. 12, no. 3, pages 397–408, 2010. (Cité en pages 8, 9 et 10.)

[Alghamdi *et al.* 2008] Mansour M. Alghamdi, Fayez A. Alhargan, Mohamed I. Alkanhal, Ashraf Alkhairy, Munir Eldesouki et Ammar Alenazi. *Saudi accented Arabic voice bank*. In ISCA Tutorial and Research Workshop on Experimental Linguistics, ExLing 2008, Athens, Greece, August 25-27, 2008, pages 9–12, 2008. (Cité en pages 18 et 21.)

[Almeman & Lee 2012] Khalid Almeman et Mark Lee. *Towards Developing a Multi-Dialect Morphological Analyser for Arabic*. In CITALA, editeur, 4th International Conference on Arabic Language Processing (CITALA 2012), pages 19–25, Rabat, Morocco, May 2012. (Cité en pages 27 et 97.)

[Almeman *et al.* 2013] Khaled Almeman, Mark Lee et Ali Abdulrahman Almiman. *Multi dialect Arabic speech parallel corpora*. In 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA), 2013, pages 1–6, 2013. (Cité en pages 19 et 21.)

[Alorifi 2008] Fawzi Suliman Alorifi. *Automatic Identification of Arabic Dialects USING Hidden Markov Models*. PhD thesis, University of Pittsburgh, September 2008. (Cité en pages 10, 12 et 13.)

[Aloulou *et al.* 2003] Chafik Aloulou, Lamia Hadrich Belguith et Abdelmajid Ben Hamadou. *MASPAR : Un système multi-agent pour l'analyse syntaxique de textes arabes*. In troisièmes journées scientifiques des jeunes chercheurs en Génie Electrique et Informatique (GEI'2003), Mahdia, Tunisie, mars 2003. (Cité en page 3.)

[Altabbaa *et al.* 2010] Mohammad Altabbaa, Ammar Al-zaraee et Mohammad Arif Shukairy. An arabic morphological analyzer and part-of-speech tagger. Master's thesis, the Faculty of Informatics Engineering, Arab International University, Damascus, Syria, 2010. (Cité en page 114.)

[Appen 2006] Pty. Ltd. Appen. *Iraqi Arabic Conversational Telephone Speech and Transcripts*. In Linguistic Data Consortium, editeur, LDC Catalog Nos. : LDC2006T16, Philadelphia. Sydney, Australia, 2006. (Cité en page 25.)

[Attia 2009] Mohammed A. Attia. *Arabic Tokenization System*. PhD thesis, School of Informatics/The University of Manchester, Manchester, 2009. (Cité en page 110.)

[Baccouche 2009] Taieb Baccouche. *Dynamique de La Langue Arabe*. Synergies Tunisie 1, pages 17–24, 2009. (Cité en pages 8, 13 et 97.)

[Baccouche 2011] Karim Aissa Baccouche. L'alternance codique arabe-français dans les forums virtuels tunisiens. Master's thesis, Université de Jyväskylä, 2011. (Cité en page 97.)

[Baccour *et al.* 2003] Leila Baccour, Ghassen Mourad et Lamia Belguith Hadrich. *Segmentation de textes arabes en phrases basée sur les signes de ponctuation et les mots connecteurs*.

In Troisième journées scientifiques des jeunes chercheurs en génie électrique et informatique, Mahdia, Tunisie, mars 2003. (Cité en pages 3 et 130.)

[Bahloul 2009] Noureddine Bahloul. *L'arabe dialectal, un outil pour une intercompréhension en classe de langue*. Synergies Algérie, no. 4, pages 255–263, 2009. (Cité en page 10.)

[Bahou *et al.* 2008] Younès Bahou, Lamia Hadrich Belguith et Abdelmajid Ben Hamadou. *Compréhension Automatique de la Parole Arabe Spontanée : Intégration dans un Serveur Vocal Interactif*. In 9th International Business Information Management Conference (IBIMA'08), Session Spéciale sur le Traitement de l'Information en Arabe, pages 1250–1259, Marrakech, Morocco, January 2008. (Cité en page 2.)

[Bayoudhi *et al.* 2014] Amine Bayoudhi, Houssem Koubaa, Hatem Ghorbel et Lamia Hadrich Belguith. *Vers un lexique arabe pour l'analyse des opinions et des sentiments*. In the 5th International Conference on Arabic Language Processing CITALA'14, novembre 2014. (Cité en page 3.)

[Bazillon 2011] Thierry Bazillon. *Transcription et traitement manuel de la parole spontanée pour sa reconnaissance automatique*. PhD thesis, Université du Maine, 2011. (Cité en pages 54 et 78.)

[Belgacem 2009] Mohamed Belgacem. *Construction d'un corpus robuste de différents dialectes arabes*. In Actes des VIIIèmes RJC Parole, volume 33, Avignon, novembre 2009. (Cité en pages 18, 19 et 21.)

[Belguith *et al.* 2005] Lamia Hadrich Belguith, Leila Baccour et Ghassan Mourad. *Segmentation de textes arabes basée sur l'analyse contextuelle des signes de ponctuations et de certaines particules*. In TALN 2005, 2005. (Cité en pages 129, 130, 132 et 150.)

[Belguith *et al.* 2014] Lamia Hadrich Belguith, Mariem Ellouze, Mohamed Hédi Maâloul, Maher Jaoua, Fatma Kallel Jaoua et Philippe Blache. Natural language processing for semitic languages, chapitre Automatic summarization, pages 371–403. Springer, 2014. (Cité en pages 11, 15 et 125.)

[Belguith 2009] Lamia Hadrich Belguith. *Analyse et résumé automatiques de documents : Problèmes, conception et réalisation*. Habilitation universitaire, Faculté des Sciences Économiques et de Gestion de Sfax, Sfax, Tunisie, Mai 2009. (Cité en pages 8 et 125.)

[Ben moussa & Alimi 2015] Nadia Karmani Ben moussa et Mohamed Adel Alimi. *Construction d'un Wordnet standard pour l'Arabe tunisien*. In CEC-TAL'2015, 2015. (Cité en pages 48 et 50.)

[Bertrand *et al.* 2008] Roxane Bertrand, Philippe Blache, Robert Espesser, Gaelle Ferré, Christine Meunier et Stéphane Rauzy. *Le CID - Corpus of Interactional Data - Annotation et Exploitation Multimodale de Parole Conversationnelle*. Traitement Automatique des Langues, vol. X, 2008. (Cité en pages 57, 71 et 87.)

[Biadsy *et al.* 2009] Fadi Biadsy, Julia Hirschberg et Nizar Habash. *Spoken arabic dialect identification using phonotactic modeling*. In In Proceedings of the EACL Workshop on Computational Approaches to Semitic Languages, pages 53–61, 2009. (Cité en page 12.)

[Bikel 2002] Daniel M Bikel. *Design of a Multi-lingual, Parallel-processing Statistical Parsing Engine*. In Proceedings of the second international conference on Human Language Technology Research, pages 178–182, 2002. (Cité en page 27.)

[Bischoff *et al.* 2009] Kerstin Bischoff, Sava Claudiu, Raluca Paiu, Wolfgang Nejdl, Cyril Laurier et Mohamed Sordo. *Music Mood and Theme Classification - a Hybrid Approach.* In Proceeding of The International Society for Music Information Retrieval, pages 57–62, 2009. (Cité en page 77.)

[Blache *et al.* 2010] Philippe Blache, Roxane Bertrand, Brigitte Bigi, Emmanuel Bruno, Edlira Cela, Robert Espesser, Gaëlle Ferré, Mathilde Guardiola, Daniel Hirst, Elgar-Paul Magro, Jean Claude Martin, Christine Meunier, Marry Annick Morel, Elisabeth Murisasco, Irina Nesterenko, Pascal Nocera, Berthille Pallaud, Laurent Prévot, Béatrice Priego-Valverde, Julien Seinturier, Ning Tan, Marrion Tellier et Stéphane Rauzy. *Multimodal Annotation of Conversational Data*. In Association for Computational Linguistics, editeur, Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010,, pages 186-191, Uppsala, Sweden, July 15-16 2010. (Cité en page 90.)

[Boersma & Weenink 2016] Paul Boersma et David Weenink. *Praat : doing phonetics by computer [Computer program]*. Version 6.0.21, retrieved 25 September 2016 from http ://www.praat.org/, 2016. (Cité en page 78.)

[Bouamor *et al.* 2014] Houda Bouamor, Nizar Habash et Kemal Oflazer. *A Multidialectal Parallel Corpus of Arabic*. In In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 1240-1245, Reykjavik, Iceland, 2014. (Cité en pages 14, 20 et 21.)

[Bouchlaghem *et al.* 2014] Rihab Bouchlaghem, Aymen Elkhlifi et Rim Faiz. *Tunisian dialect Wordnet creation and enrichment*. In Arabic Natural Language Processing Workshop co-located with EMNLP 2014, Doha, Qatar, 2014. (Cité en pages 48 et 50.)

[Boudlal *et al.* 2010] Abderrahim Boudlal, Abdelhak Lakhouja, Azzedine Mazroui, Abdelouafi Meziane, Mohamed Ould Abdallahi Ould Bebah et Mohamed Shoul. *Alkhalil Morpho Sys : A Morphosyntactic analysis system for Arabic texts*. In Proceedings of ACIT2010, Riyadh, Saudi Arabia, 2010. (Cité en pages 27, 87, 112, 113 et 150.)

[Boujelbane *et al.* 2013] Rahma Boujelbane, Mariem Ellouze Khemekhem, Siwar BenAyed et Lamia Hadrich Belguith. *Building bilingual lexicon to create Dialect Tunisian corpora and adapt language model*. In Proceedings of the Second Workshop on Hybrid Approaches to Translation., 2013. (Cité en pages 34, 72 et 74.)

[Boujelbane *et al.* 2014a] Rahma Boujelbane, Mariem Ellouze, Frédéric Béchet et 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*. TAL.

2. Traitement automatique du langage parlé, vol. 55, pages 73–96, 2014. (Cité en pages 27, 46, 47, 48, 50, 73 et 77.)

[Boujelbane *et al.* 2014b] Rahma Boujelbane, Mariem Mallek, Mariem Ellouze et Lamia Hadrich Belguith. *Fine-grained POS tagging of Spoken Tunisian Dialect Corpora*. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pages 59–62, 2014. (Cité en pages 49 et 50.)

[Boujelbane 2015] Rahma Boujelbane. *Traitements linguistiques pour la reconnaissance automatique de la parole appliquée à la langue arabe : de l'arabe standard vers l'arabe dialectal*. PhD thesis, Université de Sfax et Aix-Marseille université, 2015. (Cité en pages 48, 49, 50, 147 et 148.)

[Bouraoui 2008] Jean-Léon Mehdi Bouraoui. *Analyse, modélisation, et détection automatique des disfluences dans le dialogue oral spontané contraint : le cas du Contrôle Aérien*. PhD thesis, Université Toulouse III - Paul Sabatier, Toulouse, 2008. (Cité en page 2.)

[Buckwalter 2004] Tim Buckwalter. *Buckwalter Arabic morphological analyzer version 2.0*. LDC catalog number LDC2004L02, ISBN 1-58563-324-0., 2004. (Cité en pages 27, 28 et 30.)

[Canavan *et al.* 1997] Alexandra Canavan, George Zipperlen et David Graff. *CALLHOME Egyptian Arabic Speech LDC97S45*. https ://catalog.ldc.upenn.edu/LDC97S45, 1997. Philadelphia : Linguistic Data Consortium. (Cité en pages 18 et 21.)

[Chaâben & Belguith 2004] Nouha Chaâben et Lamia Hadrich Belguith. *Implémentation du système MORPH2 d'analyse morphologique pour l'arabe non voyellé*. In quatrièmes journées scientifiques des jeunes chercheurs en Génie Electrique et Informatique (GEI'2004), 2004. (Cité en page 3.)

[Chang & Lin 2011] Chih-Chung Chang et Chih-Jen Lin. *LIBSVM : A Library for Support Vector Machines*. ACM Trans. Intell. Syst. Technol., vol. 2, no. 3, pages 27 :1–27 :27, may 2011. (Cité en pages 140 et 144.)

[Chiang & Rambow 2006] David Chiang et Owen Rambow. *The Hidden TAG Model : Synchronous Grammars for Parsing Resource-Poor Languages*. In In Proceedings of the Eighth International Workshop on Tree Adjoining Grammars and Related Formalisms (TAG+8), pages 1–8, Sydney, July 2006. (Cité en pages 26 et 27.)

[Chiang *et al.* 2006] David Chiang, Mona Diab, Nizar Habash, Owen Rambow et Safiullah Shareef. *Parsing Arabic Dialects*. In Proceedings of the European Chapter of ACL (EACL), 2006. (Cité en page 46.)

[Cohen 1960] Jacob Cohen. *A coefficient of agreement for nominal scales*. Educational and Psychological Measurement, vol. 20, pages 37–46, 1960. (Cité en pages 80, 81 et 82.)

[Cohen 1995] William W. Cohen. *Fast effective rule induction*. In In Proceedings of the Twelfth International Conference on Machine Learning, pages 115–123, 1995. (Cité en pages 138 et 150.)

[Collins & Singer 1999] William W. Collins et Yoram Singer. *A simple, fast and effective rule learner*. Proceedings of the 16th National Conference on Artificial Intelligence (AAAI-99), pages 335–342, 1999. (Cité en page 139.)

[Cotterell & Callison-Burch 2014] Ryan Cotterell et Chris Callison-Burch. *A Multi-Dialect, Multi-Genre Corpus of Informal Written Arabic*. In The 9th edition of the Language Resources and Evaluation Conference, 2014. (Cité en pages 20 et 74.)

[Das & Petrov 2011] Dipanjan Das et Slav Petrov. *Unsupervised Part-of-Speech Tagging with Bilingual Graph-Based Projections*. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 600–609, Portland, Oregon, June 2011. (Cité en page 99.)

[Dasgupta & Ng 2007] Sajib Dasgupta et Vincent Ng. *High-Performance, Language-Independent Morphological Segmentation*. In Proceedings of Human Language Technology (NAACL), 2007. (Cité en page 49.)

[Delais-Roussarie & Durand 2003] E. Delais-Roussarie et J. Durand, editeurs. Annoter et segmenter des données de parole sous PRAAT, chapitre Corpus et Variation en Phonologie. Presses Universitaires du Mirail, presses universitaires du mirail. édition, 2003. (Cité en page 78.)

[Diab & Habash 2007] Mona Diab et Nizar Habash. *Arabic Dialect Processing Tutorial*. In Association for Computational Linguistics., editeur, In Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 5–6, Rochester, April 2007. (Cité en page 11.)

[Diab & Habash 2008] Mona Diab et Nizar Habash. *Arabic Dialect Processing*, 2008. (Cité en page 13.)

[Diab *et al.* 2010] Mona Diab, Nizar Habash, Owen Rambow, Mohamed Altantawy et Yassine Benajiba. *COLABA : Arabic Dialect Annotation and Processing*. In In Proceedings of the LREC Workshop on Semitic Language Processing, pages 66–74, 2010. (Cité en pages 16, 17, 19, 21, 34 et 74.)

[Diab *et al.* 2014] Mona Diab, Mohamed Al-Badrashiny, Maryam Aminian, Mohammed Attia, Pradeep Dasigi, Heba Elfardy, Ramy Eskander, Nizar Habash, Abdelati Hawwari et Wael Salloum. *Tharwa : A Large Scale Dialectal Arabic - Standard Arabic - English Lexicon*. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 3782–3789, 2014. (Cité en pages iv, 23, 25 et 26.)

[Dister *et al.* 2009] Anne Dister, Matthieu Constant et Gérald Purnelle. *Normalizing speech transcriptions for Natural Language Processing*. In Proceedings of the 3rd International Conference on Spoken Communication (GSCP'09), 2009. (Cité en page 90.)

[Duh & Kirchhoff 2005] Kevin Duh et Katrin Kirchhoff. *POS Tagging of Dialectal Arabic : A Minimally Supervised Approach*. In Ann Arbor, editeur, Proceedings of the ACL Workshop

on Computational Approaches to Semitic Languages,, pages 55–62, June 2005. (Cité en pages 3, 30 et 31.)

[Duh & Kirchhoff 2006] Kevin Duh et Katrin Kirchhoff. *Lexicon Acquisation for Resource-Poor Languages Using Transductive Learning*. In UWEE Technical Report Number UWEETR-2006-0012 April 2006, 2006. (Cité en pages 22 et 26.)

[Fort & Claveau 2012] Karin Fort et Vincent Claveau. *Annotation manuelle de matchs de foot : Oh la la la ! l'accord inter-annotateurs ! et c'est le but*. In TALN 2012, editeur, Actes de la conférence conjointe JEP-TALN-RECITAL, volume 2, pages pages 383–390, Grenoble, 4 au 8 juin 2012. (Cité en page 80.)

[Gibson 1998] Michael Luke Gibson. *Dialect Contact in Tunisian Arabic : Sociolinguistic and Structural Aspects*. PhD thesis, The university of Reading, 1998. (Cité en page 34.)

[Gibson 2009] Michael Luke Gibson. Encyclopedia of arabic language and linguistics, volume 4, chapitre Tunisian Arabic. Brill, 2009. (Cité en page 34.)

[Goldman 2006] Jean-Philippe Goldman. *Tutoriel Praat*. Rapport technique, Université de Genève, 2006. (Cité en page 78.)

[González Ledesma *et al.* 2004] Ana González Ledesma, Guillermo De la Madrid Heitzmann, Manuel Alcántara Plá, Raúl De la Torre Cuesta et Antonio Moreno Sandoval. *Orality and Difficulties in the Transcription of Spoken Corpora*. In Proceedings of the Workshop on Compiling and Processing Spoken Language Corpora, LREC, Lisbon, 2004. (Cité en page 54.)

[Graff & Maamouri 2012] David Graff et Mohamed Maamouri. *Developing LMF-XML Bilingual Dictionaries for Colloquial Arabic Dialects*. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, May 23-25, 2012, pages 269–274. European Language Resources Association (ELRA), 2012. (Cité en pages 24, 25 et 26.)

[Graff *et al.* 2006] David Graff, Tim Buckwalter, Hubert Jin et Mohamed Maamouri. *Lexicon Development for Varieties of Spoken Colloquial Arabic*. In th International Conference on Language Resources and Evaluation, Genoa, May 22-28, 2006. (Cité en pages iv, 22, 23, 25 et 26.)

[Graja *et al.* 2013] Marwa Graja, Maher Jaoua et Lamia Hadrich Belguith. *Discriminative Framework for Spoken Tunisian Dialect Understanding*. In Statistical Language and Speech Processing - First International Conference, SLSP 2013, Tarragona, Spain, July 29-31, 2013. Proceedings, volume 7978 of *Lecture Notes in Computer Science*, pages 102–110. Springer, 2013. (Cité en pages 47, 50, 72, 73, 76 et 77.)

[Graja 2015] Marwa Graja. *COMPREHENSION AUTOMATIQUE DE LA PAROLE EN DIALECTE TUNISIEN DANS LE CADRE DES SYSTEMES DE DIALOGUE*. Thèse de doctorat, Faculté des Sciences Économiques et de Gestion de Sfax, 2015. (Cité en page 45.)

[Guilleminot 2008] Christian Guilleminot. *DÉCOMPOSITION ADAPTATIVE DU SIGNAL DE PAROLE APPLIQUÉE AU CAS DE L'ARABE STANDARD ET DIALECTAL*. PhD thesis, Université de Ranche-comté, 2008. (Cité en page 140.)

[Habash & Rambow 2005] Nizar Habash et Owen Rambow. *Arabic Tokenization, Part-of-speech Tagging and Morphological Disambiguation in One Fell Swoop*. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL '05, pages 573–580, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. (Cité en pages 29, 114, 128, 137 et 138.)

[Habash *et al.* 2005] Nizar Habash, Owen Rambow et George Kiraz. *Morphological Analysis and Generation for Arabic Dialects*. In Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, Semitic '05, pages 17–24, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. (Cité en pages 31, 32, 49 et 97.)

[Habash *et al.* 2007] Nizar Habash, Abdelhadi Soudi et Timothy Buckwalter. Arabic computational morphology : Knowledge-based and empirical methods. Springer Netherlands, Dordrecht, 2007. (Cité en pages 8 et 35.)

[Habash *et al.* 2012a] Nizar Habash, Mona T. Diab et Owen Rambow. *Conventional Orthography for Dialectal Arabic*. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, May 23-25, 2012, pages 711–718. European Language Resources Association (ELRA), 2012. (Cité en pages 12, 16, 17, 25, 34, 56, 63, 70 et 149.)

[Habash *et al.* 2012b] Nizar Habash, Ramy Eskander et Abdelati Hawwari. *A Morphological Analyzer for Egyptian Arabic*. In Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology, SIGMORPHON '12, pages 1–9, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. (Cité en pages 28 et 29.)

[Habash *et al.* 2013] Nizar Habash, Ryan Roth, Owen Rambow, Ramy Eskander et Nadi Tomeh. *Morphological Analysis and Disambiguation for Dialectal Arabic*. In Human Language Technologies : Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA, pages 426–432. The Association for Computational Linguistics, 2013. (Cité en pages 26, 29, 97, 128 et 138.)

[Habash 2010] Nizar Habash. *Introduction to Arabic Natural Language Processing*. In Introduction to Arabic Natural Language Processing, Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2010. (Cité en page 14.)

[Hamdi *et al.* 2015] Ahmed Hamdi, Alexis Nasr, Nizar Habash et Núria Gala. *POS-tagging of Tunisian Dialect Using Standard Arabic Resources and Tools*. In Proceedings of the Second Workshop on Arabic Natural Language Processing, pages 59–68, Beijing, China, 2015. (Cité en page 50.)

[Hamdi 2007] Rym Hamdi. *La variation rythmique dans les dialectes arabes*. PhD thesis, Université Lumière Lyon2 et de l'Université du 7 Novembre à Carthage, 2007. (Cité en pages 8, 9, 10 et 11.)

[Hamdi 2015] Ahmed Hamdi. *Traitement automatique du dialecte tunisien à l'aide d'outils et de ressources de l'arabe standard : application à l'étiquetage morphosyntaxique*. PhD thesis, Aix-Marseille Université, 2015. (Cité en pages 32, 49, 50, 101, 120 et 142.)

[Hammami *et al.* 2009] Souha Mezghani Hammami, Lamia Hadrich Belguith et Abdelmajid Ben Hamadou. *Arabic Anaphora Resolution : Corpora Annotation with Coreferential links*. The international Arab Journal of Information Technology, 2009. (Cité en page 3.)

[Hana 2008] Jirka Hana. *KNOWLEDGE- AND LABOR-LIGHT MORPHOLOGICAL ANALYSIS*. In OSUWPL, volume 58, pages 52–84, 2008. (Cité en pages 102 et 103.)

[Harrat *et al.* 2014] Salima Harrat, Karima Meftouh, Mourad Abbas et Kamel Smaïli. *Building resources for Algerian Arabic dialects*. In INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014, pages 2123–2127, 2014. (Cité en pages 19, 21, 27 et 28.)

[Heeman & Allen 1994] Peter A. Heeman et James F. Allen. *Detecting and Correcting Speech Repairs*. CoRR, vol. abs/cmp-lg/9406006, 1994. (Cité en pages 59 et 90.)

[Jaccard 1912] Paul Jaccard. *Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines*. Bulletin de la Société Vaudoise des Sciences Naturelles, vol. 37, pages 241–272, 1912. (Cité en page 106.)

[Jalloh 2006] Muhammad AL Amin Jalloh. Introduction to arabic. ATLANTIC INTERNATIONAL UNIVERSITY, HONOLULU, HAWAII, 2006. (Cité en pages 8, 9 et 10.)

[Jarrar *et al.* 2014] Mustafa Jarrar, Nizar Habash, Diyam Akra et Nasser Zalmout. *Building a Corpus for Palestinian Arabic : a Preliminary Study*. In In proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing. Association for Computational Linguistics (ACL), pages 18–27, Doha , Qatar, October 25 2014. (Cité en pages 16, 17 et 21.)

[Khalfaoui 2009] Amel Khalfaoui. *A COGNITIVE APPROACH TO ANALYZING DEMONSTRATIVES IN TUNISIAN ARABIC*. PhD thesis, THE FACULTY OF THE GRADUATE SCHOOL, 2009. (Cité en page 34.)

[Kilany *et al.* 1997] Hanaa Kilany, Hassan Gadalla, Howaida Arram, Ashraf Yacoub, Alaa El-Habashi et C. McLemore. *Egyptian Colloquial Arabic Lexicon LDC99L22*. https ://catalog.ldc.upenn.edu/LDC99L22, 1997. Philadelphia : Linguistic Data Consortium. (Cité en pages 24 et 26.)

[Kilgarriff & Grefenstette 2001] Adam Kilgarriff et Gregory Grefenstette. *Web as corpus*. In Lancaster University, pages 342–344, 2001. (Cité en page 74.)

[Ksouri 2013] Myriam Ksouri. *Les particularités morphologiques et sémantiques du langage en interférence. Cas de l'alternance français/arabe dans le dialecte tunisien*. PhD thesis, Institut supérieur des langues de Tunis, 2013. (Cité en pages 43 et 53.)

[Levenshtein 1966] VI Levenshtein. *Binary Codes Capable of Correcting Deletions, Insertions and Reversals*. Soviet Physics Doklady, vol. 10, page 707, 1966. (Cité en page 83.)

[Lindström & Müürisep 2009] Liina Lindström et Kaili Müürisep. *Parsing Corpus of Estonian Dialects*. In NEALT proceedings series vol. 8 . Proceedings of the NODALIDA 2009 workshop Constraint Grammar and robust parsing, pages 22–29, Odense, Denmark, 2009. (Cité en page 99.)

[Maalej 1999] Zouhair Maalej. *PASSIVES IN MODERN STANDARD AND TUNISIAN ARABIC*. MAS-GELLAS, 1999. (Cité en page 39.)

[Maamouri & Bies 2004] Mohamed Maamouri et Ann Bies. *Developing an Arabic Treebank : Methods, Guidelines, Procedures, and Tools*. In Proceedings Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004), 2004. (Cité en pages 46 et 48.)

[Maamouri *et al.* 2004a] Mohamed Maamouri, Ann Bies et Tim Buckwalter. *The Penn Arabic Treebank : Building a large scale annotated Arabic corpus*. In In NEMLAR Conference on Arabic Language Resources and Tools, Cairo, Egypt, 2004. (Cité en pages 16, 17, 18, 22, 23, 26 et 71.)

[Maamouri *et al.* 2004b] Mohamed Maamouri, Tim Buckwalter et Christopher Cieri. *Dialectal Arabic Telephone Speech Corpus : Principles, Tool design, and Transcription Conventions*. In NEMLAR International Conference on Arabic Language Resources and Tools, 2004. (Cité en pages 21 et 62.)

[Maamouri *et al.* 2006] Mohamed Maamouri, Ann Bies, Tim Buckwalter, Mona Diab, Nizar Habash, Owen Rambow et David Tabessi. *Developing and Using a Pilot Dialectal Arabic Treebank*. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-2006), Genoa, Italy, May 2006. European Language Resources Association (ELRA). ACL Anthology Identifier : L06-1329. (Cité en pages 24, 26 et 32.)

[Maamouri *et al.* 2009] Mohamed Maamouri, David Graff, Basma Bouziri, Sondos Krouna, Ann Bies et Seth Kulick. *Standard Arabic Morphological Analyzer (SAMA) Version 3.1*. In Linguistic Data Consortium LDC2010L01, 2009. (Cité en page 28.)

[Maamouri *et al.* 2012] Mohamed Maamouri, Ann Bies, Seth Kulick, Dalila Tabessi et Sondos Krouna. *Egyptian Arabic Treebank Pilot*, 2012. (Cité en page 29.)

[Masmoudi *et al.* 2014] Abir Masmoudi, Yannick Estève, Mariem Ellouze Khmekhem, Fethi Bougares et Lamia Hadrich Belguith. *Phonetic tool for the Tunisian Arabic*. In SLTU'2014, The 4th International Workshop on spoken Language Technologies for Under-resourced Languages, Saint-Petersburg (Russia), 2014. (Cité en pages 47, 48, 50, 72, 73 et 77.)

[Masmoudi *et al.* 2015] Abir Masmoudi, Nizar Habash, Mariem Ellouze, Yannick Estève et Lamia Hadrich Belguith. *Arabic Transliteration of Romanized Tunisian Dialect Text : A Preliminary Investigation*. In Proceedings of Computational Linguistics and Intelligent Text Processing - 16th International Conference, CICLing 2015, Cairo, Egypt, April 14-20, 2015 Part I, pages 608–619, 2015. (Cité en pages 47 et 50.)

[McNeil & Faiza 2011] Karine McNeil et Miled Faiza. *Tunisian Arabic Corpus : Creating a Written Corpus of an Unwritten Language*. In Workshop on Arabic Corpus Linguistics (WACL), Lancaster University, April 2011. (Cité en pages 46, 48, 50, 53 et 73.)

[McNeil 2012] Karen McNeil. *Tunisian Arabic Morphological Parser*. Rapport technique, Brown University, Providence, 2012. (Cité en pages 49 et 50.)

[Meftouh *et al.* 2012] Karima Meftouh, Najette Bouchemal et Kamel Smaïli. *A STUDY OF A NON-RESOURCED LANGUAGE : AN ALGERIAN DIALECT*. In Proc. 3td International Workshop on Spoken Languages Technologies for Under-resourced Languages (SLTU'12), Cape Town, South Africa, May 2012. (Cité en pages 18 et 21.)

[Mejri & Baccouche 2003] Salah Mejri et Taieb Baccouche. *L'Atlas linguistique de Tunisie : repères méthodologiques pour la description du système dialectal*. In Actes du colloque Langues et métissage dans le Maghreb, 2003. (Cité en pages 34, 40, 53 et 97.)

[Mejri *et al.* 2009] Salah Mejri, Mosbah Said et Inès Sfar. *Pluringuisme et diglossie en Tunisie*. Synergies Tunisie, vol. 1, pages 53–74, 2009. (Cité en pages 34, 35, 37, 40, 41, 42, 44 et 97.)

[Mohamed *et al.* 2012] W. N. H. W. Mohamed, M. N. M. Salleh et A. H. Omar. *A comparative study of Reduced Error Pruning method in decision tree algorithms*. In IEEE International Conference on Control System, Computing and Engineering (ICCSCE), pages 392–397, Nov 2012. (Cité en pages 134 et 139.)

[Moukrim 2010] Samira Moukrim. *Morphosyntaxe et sémantique du « présent » Une étude contrastive à partir de corpus oraux Arabe marocain, berbère tamazight et français (ESLO/LCO)*. PhD thesis, Université d'Orléans, 2010. (Cité en pages 53 et 54.)

[Mubarak & Darwish 2014] Hamdy Mubarak et Kareem Darwish. *Using Twitter to collect a multi-dialectal corpus of Arabic*. In Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), page 17, Doha, Qatar, 2014. (Cité en pages 17, 20, 21 et 45.)

[Mzoughi 2015] Inès Mzoughi. *Intégration des emprunts lexicaux au français en arabe dialectal tunisien*. Thèse de doctorat, Université de Cergy-Pontoise, juin 2015. (Cité en pages 37, 44 et 53.)

[Nguyen 2006] Thi Minh Huyen Nguyen. *Outils et ressources linguistiques pour l'alignement de textes multilingues français-vietnamiens*. PhD thesis, Université Henri Poincaré, Nancy 1 en Informatique, 2006. (Cité en page 128.)

[Ouerhani 2009] Béchir Ouerhani. *Interference entre le dialectal et le littéral en Tunisie : Le cas de la morphologie verbale*. In Synergies Tunisie n1, pages 75–84, 2009. (Cité en pages 39, 41 et 104.)

[Pallaud *et al.* 2008] Berthille Pallaud, Philippe Blache et Roxane Bertrand. *Codage des annotations de disfluences dans les corpus du CID.*, 2008. (Cité en pages 90 et 91.)

[Piu & Bove 2007] Marie Piu et Rémi Bove. *Annotation des disfluences dans les corpus oraux*. In RECITAL 2007, juin 2007. (Cité en page 90.)

[Platt 1998] John Platt. *Fast Training of Support Vector Machines Using Sequential Minimal Optimization*. In Advances in Kernel Methods - Support Vector Learning. MIT Press, January 1998. (Cité en pages 140 et 144.)

[Precoda *et al.* 2007] Kristin Precoda, Jing Zheng, Dimitra Vergyri, Horacio Franco, Colleen Richey, Andreas Kathol et Sachin S. Kajarekar. *Iraqcomm : a next generation translation system*. In INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, August 27-31, 2007, pages 2841–2844, 2007. (Cité en page 19.)

[Rambow *et al.* 2006] Owen Rambow, David Chiang, Mona Diab, Nizar Habash, Rebecca Hwa, Khalil Sima'an, Vincent Lacey, Roger Levy, Carol Nichols et Safiullah Shareef. *Parsing Arabic dialects*. Rapport technique, Columbia University, 2006. (Cité en pages 23 et 26.)

[Rapp 1999] Reinhard Rapp. *Automatic Identification of Word Translations from Unrelated English and German Corpora*. In 27th Annual Meeting of the Association for Computational Linguistics, Univeristy of Maryland, College Park, Maryland, USA, 20-26 June 1999. ACL, 1999. (Cité en page 23.)

[Riesa & Yarowsky 2006] Jason Riesa et David Yarowsky. *Minimally supervised morphological segmentation with applications to machine translation*. In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, Visions for the Future of Machine Translation, pages 185–192, Cambridge, Massachusetts, USA, August 8-12 2006. (Cité en pages 29 et 31.)

[Riesa *et al.* 2006] Jason Riesa, Behrang Mohit, Kevin Knight et Daniel Marcu. *Building an English-iraqi Arabic machine translation system for spoken utterances with limited resources*. In INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, September 17-21, 2006. ISCA, 2006. (Cité en page 29.)

[Saadane & Habash 2015] Houda Saadane et Nizar Habash. *A Conventional Orthography for Algerian Arabic*. In Proceedings of the Second Workshop on Arabic Natural Language Processing, pages 69–79, Beijing, China, July 2015. Association for Computational Linguistics. (Cité en pages 16, 17 et 70.)

[Saidi 2007] Darine Saidi. *Typology of Motion Event in Tunisian Arabic*. In LingO, pages 196–203, 2007. (Cité en pages 34 et 44.)

[Saidi 2014] Darine Saidi. *Développement de la compétence narrative en arabe tunisien : rapport entre formes linguistiques et fonctions discursives*. PhD thesis, Université Lyon 2, 2014. (Cité en pages 43 et 53.)

[Sajjad *et al.* 2013] Hassan Sajjad, Kareem Darwish et Yonatan Belinkov. *Translating Dialectal Arabic to English*. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 2 : Short Papers, pages 1–6, 2013. (Cité en page 26.)

[Salama *et al.* 2014] Ahmed Salama, Houda Bouamor, Behrang Mohit et Kemal Oflazer. *YouDACC : the Youtube Dialectal Arabic Comment Corpus*. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA). (Cité en pages 17, 20 et 21.)

[Salloum & Habash 2011] Wael Salloum et Nizar Habash. *Dialectal to Standard Arabic Paraphrasing to Improve Arabic-English Statistical Machine Translation*. In Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties, pages 10–21, Edinburgh, Scotland, July 2011. Association for Computational Linguistics. (Cité en page 101.)

[Salloum & Habash 2013] Wael Salloum et Nizar Habash. *Dialectal Arabic to English Machine Translation : Pivoting through Modern Standard Arabic*. In Human Language Technologies : Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA, pages 348–358, 2013. (Cité en page 26.)

[Salloum & Habash 2014] Wael Salloum et Nizar Habash. *ADAM : Analyzer for Dialectal Arabic Morphology*. Journal of King Saud University - Computer and Information Sciences, vol. 26, no. 4, pages 372 – 378, 2014. Special Issue on Arabic {NLP}. (Cité en pages 27, 28 et 101.)

[Shriberg 1994] Elizabeth Ellen Shriberg. *Preliminaries to a Theory of Speech Disfluencies*. PhD thesis, UNIVERSITY OF CALIFORNIA at BERKELEY, 1994. (Cité en page 90.)

[Tachicart *et al.* 2014] Ridouane Tachicart, Karim Bouzoubaa et Hamid Jaafar. *Building a Moroccan dialect electronic Dictionary (MDED)*. In The fifth edition of the International Conference on Arabic Language Processing (CITALA'14), pages 216–221, 2014. (Cité en pages 25 et 26.)

[Talmoudi 1983] Fathi Talmoudi. *Texts in the Arabic Dialect of Susa (Tunisia) : Transcription, Translation, Notes and Glossary*. Language, vol. 59, no. 3, page 700, sep 1983. (Cité en page 34.)

[Tellier *et al.* 2010] Isabelle Tellier, Iris Eshkol, Samer Taalab et Jean-Philippe Prost. *POS-tagging for Oral Text with CRF and Category Decomposition*. Research in Computing Science, vol. 46, pages 79–90, 2010. (Cité en page 126.)

[Toutanvoa & Manning 2000] Kristina Toutanvoa et Christopher D. Manning. *Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger*. In 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, pages 63–70, 2000. (Cité en pages 49 et 147.)

[Trigui *et al.* 2010] Omar Trigui, Lamia Hadrich Belguith et Paolo Rosso. *DefArabicQA, "Arabic Definition Question Answering System"*. In Workshop on Language Resources and Human Language Technologies for Semitic Languages, 7th LREC, Valletta, Malta., May 17th 2010. (Cité en page 3.)

[Uchimoto *et al.* 2002] Kiyotaka Uchimoto, Chikashi Nobata, Atsushi Yamada, Satoshi Sekine et Hitoshi Isahara. *Morphological Analysis of the Spontaneous Speech Corpus*. In COLING 2002 : The 17th International Conference on Computational Linguistics : Project Notes, 2002. (Cité en page 96.)

[Vapnik 1995] Vladimir N. Vapnik. The nature of statistical learning theory. Springer-Verlag New York, Inc., New York, NY, USA, 1995. (Cité en page 139.)

[Waibel *et al.* 2004] Alex Waibel, Tanja Schultz, Stephan Vogel, Christian Fügen, Matthias Honal, Muntsin Kolss, Jürgen Reichert et Sebastian Stüker. *Towards Language Portability in Statistical Machine Translation*. In Invited paper, Special Session on Multilinguality in Speech Processing, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. (Cité en pages 74 et 149.)

[Walther & Sagot 2010] Géraldine Walther et Benoît Sagot. *Developing a Large-Scale Lexicon for a Less-Resourced Language : General Methodology and Preliminary Experiments on Sorani Kurdish*. In Proceedings of the 7th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages (LREC 2010 Workshop), Valetta, Malta, 2010. (Cité en page 99.)

[Yang *et al.* 2007] Mei Yang, Jing Zheng et Andreas Kathol. *A semi-supervised learning approach for morpheme segmentation for an Arabic dialect*. In INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, August 27-31, 2007, pages 1501–1504, 2007. (Cité en pages 29 et 30.)

[Younes & Souissi 2014] Jihene Younes et Emna Souissi. *A quantitative view of Tunisian dialect electronic writing*. In 5th International Conference on Arabic Language Processing, Oujda, Morocco, 2014. (Cité en pages 17, 47 et 50.)

[Zaidan & Callison-Burch 2011] Omar Zaidan et Chris Callison-Burch. *The Arabic Online Commentary Dataset : an Annotated Dataset of Informal Arabic with High Dialectal Content*. In The 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011,

Portland, Oregon, USA - Short Papers, pages 37–41, 2011. (Cité en pages 20, 21 et 74.)

[Zawaydeh *et al.* 2003] Bushra Zawaydeh, Dave Stallard et John Makhoul. *Babylon Transcription Guidelines*. In Babylon Transcription Guidelines. http ://ldc.upenn.edu/Catalog/docs/LDC2005S08/BBN- Babylontranscription-guidelines.pdf, 2003. (Cité en pages 16 et 17.)

[Zbib *et al.* 2012] Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard M. Schwartz, John Makhoul, Omar Zaidan et Chris Callison-Burch. *Machine Translation of Arabic Dialects*. In Human Language Technologies : Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 3-8, 2012, Montréal, Canada, pages 49–59. The Association for Computational Linguistics, 2012. (Cité en pages 20 et 26.)

[Zeman & Resnik 2008] Daniel Zeman et Philip Resnik. *Cross-Language Parser Adaptation between Related Languages*. In Third International Joint Conference on Natural Language Processing, IJCNLP 2008, Hyderabad, India, January 7-12, 2008, pages 35–42. The Association for Computer Linguistics, 2008. (Cité en page 99.)

[Zouaghi *et al.* 2008] Anis Zouaghi, Mounir Zrigui et Georges Antoniadis. *Compréhension automatique de la parole arabe spontanée*. TAL, vol. Volume 49, no. 1, pages 141–166, 2008. (Cité en page 2.)

[Zribi *et al.* 2013a] Inès Zribi, Marwa Graja, Mariem Ellouze Khemakhem, Maher Jaoua et Lamia Hadrich Belguith. *Orthographic Transcription for Spoken Tunisian Arabic*. In A. Gelbukh (Ed.) : CICLing 2013, Part I, LNCS 7816, pp. 153-163, 2013., pages 153–163, 2013. (Cité en pages 34 et 57.)

[Zribi *et al.* 2013b] Inès Zribi, Mariem Ellouze Khemakhem et Lamia Hadrich Belguith. *Morphological Analysis of Tunisian Dialect*. In Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, Nagoya, Japan, October 14-18, 2013, pages 992–996, 2013. (Cité en pages 87 et 88.)

[Zribi *et al.* 2014] Inès Zribi, Rahma Boujelbane, Abir Masmoudi, Mariem Ellouze, Lamia Hadrich Belguith et Nizar Habash. *A Conventional Orthography for Tunisian Arabic*. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014., pages 2355–2361. European Language Resources Association (ELRA), 2014. (Cité en pages 12, 14, 34, 35, 37, 40, 42, 45, 47 et 48.)

[Zribi *et al.* 2015] Inès Zribi, Mariem Ellouze, Lamia Hadrich Belguith et Philippe Blache. *Spoken Tunisian Arabic Corpus « STAC » : Transcription and Annotation*. Research in computing science, vol. 90, 2015. (Cité en pages 73 et 76.)

[Zribi *et al.* 2016] Inès Zribi, Inès Kammoun, Mariem Ellouze, Lamia Hadrich Belguith et Philippe Blache. *Sentence boundary detection for transcribed Tunisian Arabic*. In Procee-

dings of the 12th Edition of the Konvens Conference, Bochum, Germany, september 2016. (Cité en pages 130, 134 et 135.)

Liste des publications

1. **Zribi, Inès**, Mariem Ellouze Khemakhem, et Lamia Hadrich Belguith. « Morphological Analysis of Tunisian Dialect. » International Joint Conference on Natural Language Processing, 14-18 October 2013 : 992-996.
2. **Zribi, Inès**, Marwa Graja, Mariem Ellouze Khemakhem, Maher Jaoua, et Lamia Hadrich Belguith. « Orthographic Transcription for Spoken Tunisian Arabic. » A. Gelbukh (Ed.) : CICLing 2013, Part I, LNCS 7816, pp. 153-163, 2013., 2013 : 153-163.
3. **Zribi, Inès**, Rahma Boujelbane, Abir Masmoudi, Mariem Ellouze, Lamia Belguith, et Nizar Habash. « A Conventional Orthography for Tunisian Arabic. » The Ninth International Conference on Language Resources and Evaluation (LREC'14), 2014 : 2355-2361.
4. **Zribi, Inès**, Mariem Ellouze, Lamia Hadrich Belguith, et Philippe Blache. « Spoken Tunisian Arabic Corpus "STAC" : Transcription and Annotation. » Research in Computing Science journal,volume 90, 2015.
5. **Zribi, Inès**, Inès Kammoun, Mariem Ellouze, Lamia Hadrich Belguith, et Philippe Blache. « Sentence boundary detection for transcribed Tunisian Arabic » Proceedings of the 12th Edition of the Konvens Conference, Bochum, Germany, September 19-21, 2016. Ruhr-University Bochum, 2016.
6. Boujelbane, Rahma, **Inès Zribi**, Syrine Kharroubi, et Mariem Ellouze. « An Automatic Process for Tunisian Arabic Orthography Normalization » HrTAL, 2016.

Annexe A : Extrait du corpus STAC

Nous présentons dans cet annexe un extrait de notre corpus STAC sous forme de fichier « .textgrid ».

```
File type = "ooTextFile"
Object class = "TextGrid"

xmin = 0
xmax = 216.236
tiers? <exists>
size = 5
item []:
    item [1]:
        class = "IntervalTier"
        name = "Presenter"
        xmin = 0
        xmax = 215.975
        intervals: size = 238
        intervals [1]:
            xmin = 0
            xmax = 0.3019375000000006
            text = "مُكَالَمَة"
        intervals [2]:
            xmin = 0.3019375000000006
            xmax = 0.6038750000000006
            text = "الهَاتفِيَّة"
        intervals [3]:
            xmin = 0.6038750000000006
            xmax = 1.2077500000000005
            text = "ال-"
        intervals [4]:
            xmin = 1.2077500000000005
            xmax = 1.5096875000000005
            text = "آ÷"
        intervals [5]:
            xmin = 1.5096875000000005
            xmax = 1.8116250000000005
            text = "آ÷"
        intervals [6]:
            xmin = 1.8116250000000005
            xmax = 2.1135625000000005
            text = "لِلْسَبِيْ"
        intervals [7]:
            xmin = 2.1135625000000005
            xmax = 2.717437500000001
            text = "[فَـنْجِي, شخ]"
        intervals [8]:
            xmin = 2.717437500000001
            xmax = 3.019375000000001
            text = "آ÷"
        intervals [9]:
            xmin = 3.019375000000001
            xmax = 3.3213125000000012
            text = "آ÷"
        intervals [10]:
            xmin = 3.3213125000000012
            xmax = 3.6232500000000014
            text = "[غَاشُوْ, شخ]"
```

FIGURE 2 – Extrait du corpus STAC sous forme d'un fichier « .textgrid ».

Annexe B : Liste des schèmes de dérivation présents dans le corpus STAC

Schème	Schème	Schème	Schème
اتَّعَلتُ At açaltu	مُفَعَّلَاتِ mufaç alaAti	فِعلِيَّ fiçliy a	فَاعِلُ fAçilu
اِستَفَال AistafaAl	مُفَعَّلَاتِ mufaç ilaAti	فِعلِيُّ fiçliy u	فَاعَل faAçal
استِفَالَاتُ AstifaAlaAtu	مَفعلَات mafiçlaAt	فِعَلِيُّ fiçaliy u	فَاعِل fAçil
استِفَالَة AstifaAlaћ	مُفَعَّلَات mufaç ilaAt	فِعلِيُّ fiçliy u	فَاعِل faAçil
استِفَالَةٍ AstifaAlaћĩ	مَفعِلَّة mafçil aћ	فَعَلِيّ façaliy i	فَاعِلًا faAçilãA
استِفعَاء AstifçaA’	مَفعِلَة mafiçlaћ	فَعلِيّ façliy i	فَاعِلَاتُ faAçilaAtu
استِفعَاءُ AstifçaA’u	مُفعِلَة mufiçlaћ	فَعَلَي façalay	فَاعِلَة faAçilaћ
استِفعَال AstifçaAl	مُفَعَّلَة mufaç ilaћ	فَعلَي façlay	فَاعِلَةَ fAçilaћa
استِفعَالَي AstifçaAlay	مفَعَلَة mfaç laћ	فَعَلِيَّاتٌ façaliy aAtũ	فَاعِلَةُ faAçilaћu
استَفعَل Astafçal	مَفعَلَة mafçalaћ	فَعلِيَّات façliy aAt	فَاعَلَت faAçalat
استَفعَلت Astafçalt	مِفعَلَة mifçalaћ	فُعَّلِيَّة fuç aliy aћ	فَاعَلِت faAçalit
استَفعَلنَا AstafçalnaA	مُفعِلَة mufçilaћ	فَعَلِيَّة façaliy aћ	فَاعلِت faAçlit
استَفعَى AstafçaY	مُفَعِّلَتَينِ mufaç ilatayni	فَعلِيَّة façliy aћ	فَاعَلت faAçalt
استَفعِيت Astafçiyt	مُفَعلَل mufaçlal	فُعلِيَّة fuçliy aћ	فَاعِلت faAçilt
استَفعَينَا AstafçaynaA	مُفَعلِلَةُ mufaçlilaћu	فِعلِيَّة fiçliy aћ	فَاعَلتُوا faAçaltuwA
اَعِل Ãçil	مَفعِلِيَّةُ mafçiliy aћu	فِعلِيلٌ fiçliylũ	فَاعَلنَا fAçalnA
اَفَاعِل ÂafaAçil	مَفعِلِينَ mafiçliyna	فِعلِيلَ fiçliyla	فَاعَلنَا faAçalnaA
اِفَالَات ĂifaAlaAt	مُفَعَّلِينَ mufaç aliyna	فَعَّلَينِ faç ilayni	فَاعلُوا faAçluwA
اِفَالَة ĂifaAlaћ	مُفَعِّلِينَ mufaç iliyna	فَعَلَينِ façalayni	فَاعِلُونَ faAçiluwna
افتَال AftaAl	مُفَّعَلِينَ muf açaliyna	فُعُلَينِ fuçulayni	فَاعِلِي fAçiliy
افتَالُوا AftaAluwA	مُفعَلِينَ mufçaliyna	فَعلَينِ façlayni	فَاعِلِيَّ faAçiliy a
افتِعَاءَا AftiçaA’aA	مُفَّعِلِينَ muf açiliyna	فَعلِيِّينَ façliy iyna	فَاعِلَي faAçilay
افتِعَالٌ AftiçaAlũ	مُفعِلِينَ mufçiliyna	فَعَلِيَّينِ façaliy ayni	فَاعَلِيَّةَ faAçaliy aћa
افتِعَالِ AftiçaAlĩ	مُفَعَّلَينِ mufaç alayni	فَعلِيَّينِ façliy ayni	فَاعِلِيَّةَ faAçiliy aћa
افتِعَالَ AftiçaAla	مُفَعَّلَينِ mufaç ilayni	فُعَّ fuç aåaAt	فَاعَلِيَّةُ faAçaliy aћu
افتِعَالُ AftiçaAlu	مُفعَلَينِ mufçalayni	فَعُّوا faç uwA	فَاعِلِيَّةُ faAçiliy aћu
افتِعَال AftiçaAl	مَفعُول mafçuwl	فُعُّوا fuç uwA	فَاعِلَينِ fAçilayni
افتِعَال AftiçaAl	مَفعُولَات mafçuwlaAt	فِعُّوا fiç uwA	فَاعِلِين faAçiliyn

Suite page suivante . . .

Schème	Schème	Schème	Schème
افتِعَالَاتٌ AftiçaAlaAtũ	مَفعُولَة mafçuwlaħ	فَعُول façuwl	فَاعُّوا faAç uwA
افتِعَالَات AftiçaAlaAt	مَفعُولِيَّةَ mafçuwliy aħa	فُعُول fuçuwl	فَاعُوا faAçuwA
افتِعَالِي AftiçaAlay	مَفعُولِيَّةُ mafçuwliy aħu	فعُول fçuwl	فَاعُول faAçuwl
افتَعَل Aftaçal	مَفعُولِينَ mafçuwliyna	فُعُولًا fuçuwlãA	فَاعَى fAçaY
افتَعِل Aftaçil	مَفعُولَينِ mafçuwlayni	فَعُولَة façuwlaħ	فَاعِي faAçiy
افتَعَلَت Aftaçalat	مَفعُولِين mafçuwliyn	فُعُولَة fuçuwlaħ	فَاعَّي faAç ay
افتَعَلُوا AftaçaluwA	مَفعَى mafçaY	فُعُولِيَّ fuçuwliy a	فَاعِّيَّة faAç iy aħ
افتَعُّوا Aftaç uwA	مُفعَى mufçaY	فُعُولِيَّاتٌ fuçuwliy aAtũ	فَاعِيَةٌ faAçiyaħũ
افتَعَى AftaçaY	مَفعِيَّة mafçiy aħ	فُعُولِيَّاتُ fuçuwliy aAtu	فَاعِيَةَ faAçiyaħa
افتِيَالٌ AftiyaAlũ	مُفَعَّينِ mufaç ayni	فُعُولِيّة fuçuwliy ħ	فَاعِيَةُ faAçiyaħu
افعَا AfçaA	مُفِعَّينِ mufiç ayni	فُعُولِيَّة fuçuwliy aħ	فَاعِيَةِ faAçiyaħi
أَفِعَّاء Âafiç aA’	مُفَوَّلِين mufaw liyn	فَعُولِيَّةَ façuwliy aħa	فَاعِّينَ faAç iyna
أَفعَالٌ ÂafçaAlũ	مُفِيل mufiyl	فَعُولِيَّةُ façuwliy aħu	فَاعِينَ faAçiyna
إِفعَالٌ ĂifçaAlũ	مُفِيلَةُ mufiylaħu	فَعُولَينِ façuwlayni	فَاعَّينِ faAç ayni
أَفعَالَ ÂafçaAla	مُفِيلَينِ mufiylayni	فَعَّى faç aY	فَاعِيَينِ faAçiyayni
إِفعَالَ ĂifçaAla	مُنفَعِلِين munfaçiliyn	فِعَى fiçaY	فَال fAl
أَفعَالُ ÂafçaAlu	مُوعلِين muwçliyn	فعَى fçaY	فَال faAl
أَفعَالِ ÂafçaAli	مِيعَال miyçaAl	فَعِّي faç iy	فَالَة faAlaħ
إِفعَال ĂfçaAl	نَاعِل naAçil	فِعِّي fiç iy	فَالَت fAlat
أَفِّعَال Af içaAl	نَاعلُوا naAçluwA	فِعِّيَّ fiç iy a	فَالِت fAlit
أَفعَال ÂafçaAl	نِتفَاعلُوا nitfaAçluwA	فِعِّيُّ fiç iy u	فَالُوا faAluwA
إِفعَال ĂfçaAl	نَتَفَاعَى natafAçaY	فَعِيٌّ façyũ	فَالِيَّات faAliy aAt
إِفعَال ĂifçaAl	نِتفَعَّاوَا nitfaç aAwA	فَعِيَ façya	فَالَينِ faAlayni
إِفعَالَاتِ ĂifçaAlaAti	نَتَفَعَّلُ natafaç alu	فِعِّيًّا fiç iy aA	فَايِ faAy
إِفعَالَات ĂifçaAlaAt	نَتَفَعَّل natafaç al	فِعِيَّاتِ fçiy aAti	فَائِلُ faA’ilu
أَفعَالِيَّاتُ ÂafçaAliy aAtu	نِتفَعَّل nitfaç al	فُعِّيَّات fuç iy aAt	فَايِل faAyl
إِفعَالِيَّات ĂfçaAliy aAt	نِتفَعِّل nitfaç il	فُعِّيَّة fuç iy aħ	فَايِيل faAyil
إِفعَالِيَّةُ ĂfçaAliy aħu	نَتَفَعلَلُ natafaçlalu	فِعِّيَّة fiç iy aħ	فَائِلًا faA’ilãA
افعَاوَا AfçaAwA	نِتفَعَّلُوا nitfaç luwA	فَعِيَةً façyaħã	فَايلَات faAylaAt
افعَل Afçal	نِتفَعَّى nitfaç aY	فَعِيَّةٍ façiy aħĩ	فَايلَة faAylaħ
أَفعَل Âafçal	نِستَفعِل nistafçil	فِعِيَةَ fiçyaħa	فَايِلَة faAyilaħ
افعُل Afçul	نِستَفعِلُوا nistafçiluwA	فُعِّيَّةُ fuç iy aħu	فَايِلِين faAyliyn
افعِل Afçil	نُفَاعِ nufAçi	فَعِيَّةُ façiy aħu	فَايِين faAyiyn
أَفعِلَاءُ ÂafçilaA’ũ	نُفَاعِل nufAçil	فَعِيَةُ façyaħu	فُعِّ fç u
أَفعِلَة Âafçilaħ	نفَاعِل nfaAçil	فُعِيَةُ fuçyaħu	فَعِّ faç u

Suite page suivante . . .

Schème	Schème	Schème	Schème
افعُلُوا AfçuluwA	نفَاعِلُوا nfaAçiluwA	فَعِيَّة façiy aħi	فَعّ faç
افعِلُوا AfçiluwA	نفَاعلُوا nfaAçluwA	فَعّيت faç iyt	فُعّ fuç
افعلُوا AfçluwA	نفَاعِيوَا nfaAçiywA	فعِيت fçiyt	فَع faç
أَفعَلِي Âafçaliy	نفَال nfaAl	فعِيتُوا fçiytuwA	فُع fuç
أَفعَلِيَّة Âafçaliy aħ	نِفتَال niftaAl	فعيل fçyl	فِع fiç
أُفعُول Âufçuwl	نِفتَعِل niftaçil	فِعّيل fiç iyl	فَعّء faç ’
أُفعُولَةُ Âufçuwlaħu	نِفتَعِلُوا niftaçiluwA	فعِيل fçiyl	فعَا fçA
افعِي Afçiy	نَفتَعِي naftaçiy	فَعِيل façiyl	فَعَّا faç ãA
افعِيوَا AfçiywA	نفَعّ nfaç i	فَعِيلًا façiylãA	فَعَّا faç˜aA
أَلَة Ãlaħ	نِفعِ nifçi	فَعِيلَات façiylaAt	فُعَّا fuç˜aA
انفِعَالٌ AnfiçaAlũ	نفِع nfiç	فَعيلَة façylaħ	فَعَا façaA
إيعَالَ Ăiyçaala	نِفعَا nifçaA	فِعّيلَة fiç˜iylaħ	فعَا fçaA
تَاعِل taAçil	نِفعَاوَا nifçaAwA	فَعِيلَة façiylaħ	فَعَاءٌ façaA’ũ
تَاعلُوا taAçluwA	نفَعّل nfaç˜il	فَعِيلِيّ façiyliy˜i	فَعَاءَ façaA’a
تِتفَاعِل titfaAçil	نِفعِل nifçil	فَعِيلِيَّة façiyliy˜aħ	فِعَاءَ fiçaA’a
تِتفَاعَى titfaAçaY	نفعِل nfçil	فِعّيِن fiç˜ayni	فِعَاءُ fiçaA’u
تِتفَال titfaAl	نَفعَلُ nafçalu	فَعِين façiyn	فَعَاءِ façaA’i
تِتفَع titfaç	نُفعل nfuçl	فَعّينَا faç˜iynA	فَعَاء façaA’
تَتَفَعَّلُ tatafaç˜alu	نفِعل nfiçl	فَعّينَا faç˜iynaA	فُعَاء fuçaA’
تَتَفَعّل tatafaç˜al	نفَعّل nfaç˜il	فعِينَا fçiynaA	فِعَاءَا fiçaA’aA
تِتَفَعّل titafaç˜al	نَفعَل nafçal	فُعّيُّوا fuç˜iy˜uwA	فِعَاءَاتٌ fiçaA’aAtũ
تِتفَعَّل titfaç˜al	نِفعَل nifçal	فِعّيوَا fiç˜iywA	فِعَاءَاتُ fiçaA’aAtu
تِتَّفعَل tit˜afçal	نُفَعلِلُ nufaçlilu	فَلَّت fal˜at	فَعَائِيٌّ façaA’iy˜u
تَتَفَعلَلَ tatafaçlala	نفَعلِل nfaçlil	فُلت fult	فَعَائِيًّا façaA’iy˜ãA
تَتَفَعلَلُ tatafaçlalu	نِفعلُوا nfiçluwA	فِلت filt	فَعَائِيَّةَ façaA’iy˜aħa
تِتفَعَّلُوا titfaç˜aluwA	نفَعّلُوا nfaç˜iluwA	فُلتُوا fultuwA	فِعَائِيَّةَ fiçaA’iy˜aħa
تِتَفَعّلُوا titafaç˜luwA	نَفعَلُوا nafçaluwA	فِلتُوا filtuwA	فِعَائِيَّةُ fiçaA’iy˜aħu
تِتفَعَّى titfaç˜aY	نِفعَلُوا nifçaluwA	فُلنَا fulnA	فَعَاة façaAħ
تَستَفعِلُ tastafçilu	نفَعلُوا nfaçluwA	فِلنَا filnaA	فَعَّات faç˜aAt
تِستَفعِل tistafçil	نفِعُّوا nfiç˜uwA	فَوَاعِل fawaAçil	فِعَّات fiç˜aAt
تَعِل taçil	نِفعَى nifçaY	فُوعَة fuwçaħ	فَعَات façaAt
تِعلَا tiçlaA	نفَعّي nfaç˜iy	فُول fuwl	فعَات fçaAt
تَفَاعِل tafaAçil	نِفعِي nifçiy	فيَال fyaAl	فَعَّالٌ faç˜aAlũ
تفَاعَل tfaAçal	نِفعِيوَا nifçiywA	فِيَال fiyaAl	فَعَالٌ façaA˜alũ
تَفَاعُل tafaAçul	نَفُولُ nafuwlu	فِيَالَة fiyaAlaħ	فَعَّالَ faç˜aAla

Suite page suivante . . .

Schème	Schème	Schème	Schème
تفَاعِل tfaAçil	نفُول nfuwl	فِيَالِيَّ fiyaAliy˜a	فَعَالَ façaA˜ala
تَفَاعُلَةُ tafaAçulaħu	نفُولُوا nfuwluwA	فِيَالِيَّةَ fiyaAliy˜aħa	فَعَالِ façaA˜ali
تفَاعَلت tfaAçalt	نِفِيل nifiyl	فِيَالِيَّةُ fiyaAliy˜aħu	فَعَّال faç˜aAl
تَفَاعَلتُوا tafaAçaltuwA	نفِيلُوا nfiyluwA	فِيَة fiyaħ	فُعَّال fuç˜aAl
تفَاعلُوا tfaAçluwA	نَنفَعِلُوا nanfaçiluwA	فِيلَ fiyla	فعَال fçaAl
تَفَاعَى tafAçaY	وعِّل wç˜il	فِيل fiyl	فَعَال façaAl
تفَاعِيت tfaAçiyt	وعِّلت wç˜lit	فَيِّل fay˜il	فُعَال fuçaAl
تفَال tfaAl	وَعلُوا waçluwA	فِيلَان fiylaAn	فِعَال fiçaAl
تفَالِت tfaAlit	يَاعِل yaAçil	فِيلَة fiylaħ	فعَال fçaAl
تِفتَّال tiftaAl	يَاعلُوا yaAçluwA	فَيَّلَة fay˜laħ	فَعَالًا façaAlãA
تِفتَعِل tiftaçil	يِتفَاعَاوَا yitfaAçaAwA	فِيلَة fyilaħ	فَعَّالَا faç˜aAlaA
تَفتَعلُوا taftaçluwA	يَتَفَاعَلُ yatafAçalu	فَيلَة faylaħ	فَعَالَاَ façaAlaAa
تُفتَعَى tuftaçaY	يُتَفَاعَلُ yutafAçalu	فَيَّلت fay˜alt	فِعَالَاتٌ fiçaAlaAtũ
تَفتَعِي taftaçiy	يَتَفَاعَل yatafAçal	فِيلُوا fiyluwA	فَعَّالَاتُ faç˜aAlaAtu
تفَعِّ tfaç˜i	يتَفَاعِل ytafaAçil	فَيَّلُوا fay˜aluwA	فَعَالَاتُ façaA˜alaAtu
تفَع tfaç	يِتفَاعِل yitfaAçil	مَاعلَة mAçlaħ	فَعَّالَاتِ faç˜aAlaAti
تفِع tfiç	يتَفَاعَلُوا ytafaAçaluwA	مَاعلِين maAçliyn	فَعَّالَات faç˜aAlaAt
تِفعَا tifçaA	يِتفَاعلُوا yitfaAçluwA	مُتَّعَل mut˜açal	فَعَالَات façaAlaAt
تفَعَّات tfaç˜aAt	يِتفَاعَى yitfAçaY	مُتَّعِلِينَ mut˜açiliyna	فِعَالَات fiçaAlaAt
تِفعَال tifçaAl	يِتفَال yitfaAl	مُتَفَاعِل mutafaAçil	فعَالَات fçaAlaAt
تفَعَّاوَا tfaç˜aAwA	يِتفَاولُوا yitfaAwluwA	مُتَفَاعِلَة mutafaAçilaħ	فَعَّالَة faç˜aAlaħ
تِفعَاوَا tifçaAwA	يِتفَع yitfaç	مُتَفَاعِلَةُ mutafaAçilaħu	فُعَّالَة fuç˜aAlaħ
تفَعِت tfaçi˜t	يِتفَعَّاوَا yitfaç˜aAwA	مُتَفَاعِلِين mutafaAçiliyn	فَعَالَة façaAlaħ
تفَعِتُوا tfaçituwA	يِتفعَاوَا yitfçaAwA	مُتَفَعَّل mutafaç˜alu	فِعَالَة fiçaAlaħ
تَفَعِّل tafaç˜il	يَتَفَعَّلُ yatafaç˜alu	مُتَفَعَّل mutafaç˜al	فَعَالَةُ façaA˜alaħu
تفَعِّل tfaç˜il	يَتَفَعَّل yatafaç˜al	مُتَفَعِّل mutafaç˜il	فَعَالِلَ façaAlila
تفَعِل tfaçil	يِتفَعِّل yitfaç˜il	متَفَعِّل mtafaç˜l	فَعَالِل façaAlil
تِفعِل tifçil	يتَّفعَل yt˜afçal	مُتفَعل mutfaçl	فعَالِلةٌ fçaAlilħũ
تَفَعُّلَ tafaç˜ula	يَتَفَعلَلُ yatafaçlalu	مُتَفَعِّلَة mutafaç˜ilaħ	فعَالِلةَ fçaAlilħa
تَفعَلَ tafçala	يِتفَعلِل yitfaçlil	مُتفَعِّلَة mutfaç˜laħ	فعَالِلةُ fçaAlilħu
تَفُعلُ tafuçlu	يتَفَعَّلُوا ytafaç˜aluwA	مُتَفَعِّلَةُ mutafaç˜ilaħu	فَعَّالِي faç˜aAliy
تفُعل tfuçl	يِتفَعلُوا yitfaçluwA	مُتَفَعلَلُ mutafaçlalu	فِعَالِي fiçaAliy
تفَعَّل tfaç˜al	يِتفَعُّوا yitfaç˜uwA	مُتَفَعلِلَةُ mutafaçlilaħu	فعَالِي fçaAliy
تَفَعُّل tafaç˜ul	يِتفَعَّى yitfaç˜aY	مُتَفَعِّلِينَ mutafaç˜aliyna	فَعَالِيَّ façaAliy˜a
تفَعِّل tfaç˜il	يُستَفَعُّ yustafaç˜u	مُتَفَعِّلَينِ mutafaç˜alayni	فَعَّالَي faç˜aAlay

Suite page suivante . . .

Schème	Schème	Schème	Schème
تَفَّعَل tf˜açal	يَستَفِعُّ yastafiç˜u	مُتَفَعّي mutafaç˜iy	فِعَالَي fiçaAlay
تَفعَل tafçal	يِستَفعِل yistafçil	مُتَفَعِّيَة mutafaç˜iyaħa	فَعَالَي façaA˜alay
تِفعَل tifçal	يُستَفعَى yustafçaY	مُستَاعِل mustaAçil	فَعَالِيَّة façaAliy˜aħ
تَفعُل tafçul	يِستَفِيل yistafiyl	مُستَفَع mustafaç	فُعَالِيَّة fuçaAliy˜aħa
تُفعِل tufçil	يَعِل yaçil	مُستَفِعَّاتٌ mustafiç˜aAtũ	فَعَالِيَّةُ façaAliy˜aħu
تفَعِل tfaçil	يُفَاعَ yufAça	مُستَفِعَّة mustafiç˜aħ	فَعَّالَينِ faç˜aAlayni
تَفعِل tafçil	يُفَاعِ yufAçi	مُستَفعَل mustafçal	فَعَالِين façaAliyn
تَفَعُّلَاتُ tafaç˜ulaAtu	يفَاعِل yfaAçil	مُستَفعِل mustafçil	فُعَان fuçĀn
تَفعُلَة tafçulaħ	يُفَاعِلُ yufAçilu	مُستَفعَلَاتُ mustafçalaAtu	فَعَّاوَا faç˜aAwA
تَفَعُّلَة tafaç˜ulaħa	يفَاعِل yfaAçil	مُستَفعِلَاتِ mustafçilaAti	فَعَاوَا façaAwA
تفَعلِت tfaçlit	يفَاعِلُوا yfaAçiluwA	مُستَفعِلَات mustafçilaAt	فعَاوَا fçaAwA
تَفَعَّلت tafaç˜alt	يفَاعلُوا yfaAçluwA	مُستَفعِلَة mustafçilaħ	فَعَّايَات faç˜aAyaAt
تفَعَّلت tfaç˜alt	يُفَالُ yufAlu	مُستَفعَى mustafçaY	فِعَايَات fiçaAyaAt
تِفعَلت tifçalt	يفَال yfaAl	مُستَفعِي mustafçiy	فَعَّايَة faç˜aAyaħ
تفَعِلت tfaçilt	يَفَالُوا yafAluwA	مُستَفِعِّينَ mustafiç˜iyna	فعَايَة fçaAyaħ
تَفَعلَل tafaçlala	يفَالُوا yfaAluwA	مُستَفِيل mustafiyl	فَعَائِلُ façaA'ilu
تفَعلِلِ tfaçlili	يَفتَالَ yaftAla	مَعَايلَة maçaAylaħ	فَعَائِل façaA'il
تفَعلِل tfaçlil	يَفتَالُ yaftAlu	مَعَائِلَة maçaA'ilħa	فَعَايِل façaAyil
تفَعَّلنَا tfaç˜alnA	يَفتَالُوا yaftAluwA	مُفَاعِ mufaAç	فَعَّة faç˜aħ
تفَعَّلنَا tfaç˜alnaA	يفتَعِل yiftaçil	مُفَاعَلُ mufaAçalu	فُعَّة fuç˜aħ
تفُعلُوا tfuçluwA	يفتَعلُوا yiftaçluwA	مُفَاعِلُ mufaAçilu	فِعَّة fiç˜aħ
تفَعَّلُوا tfaç˜aluwA	يُفتَعَى yuftaçaY	مُفَاعَل mufaAçal	فُعَةَ fuçaħa
تفَعِّلُوا tfaç˜iluwA	يُفَعَّ yufaç˜a	مَفَاعِل mafaAçil	فُعَةِ fuçaħi
تفَعّلُوا tfaç˜luwA	يُفَعِّ yufaç˜i	مُفَاعِل mufaAçil	فَعّت faç˜it
تَفعَلُوا tafçaluwA	يفعِ yifçi	مُفَاعَلَاتٌ mufaAçalaAtũ	فِعَّتَينِ fiç˜atayni
تَفعُلُوا tafçuluwA	يفعِ yfiç	مُفَاعَلَاتُ mufaAçalaAtu	فَعلٌ façlũ
تفَعلُوا tfaçluwA	يَفعَا yafçaA	مُفَاعَلَات mufaAçalaAt	فَعَلَ façala
تفَعُّوا tfaç˜uwA	يفعَاوَا yifçaAwA	مُفَاعَلَة mufaAçalaħ	فعُلُ fçulu
تِفعُّوا tfiç˜uwA	يُفَعِّلُ yufaç˜ilu	مُفَاعِلَة mufaAçilaħ	فَعِلِ façli
تفَعَّى tfaç˜aY	يَفعَلُ yafçalu	مُفَاعلَة mufaAçlaħ	فَعل façl
تِفعَى tifçaY	يُفعَلُ yufçalu	مُفَاعَلِينَ mufaAçaliyna	فُعل fuçl
تفَعِي tfaçiy	يَفعُلُ yafçulu	مُفَاعَلَينِ mufaAçalayni	فِعل fiçl
تِفعِي tifçiy	يفَعَّل yfaç˜al	مُفَاعِلَينِ mufaAçilayni	فَعَّل faç˜al
تفَعّي tfaç˜y	يُفَعّل yufaç˜il	مُفَاعَلِين mufaAçaliyn	فَعّل faç˜il
تَفَعّيَاتُ tafaç˜iyaAtu	يِفَعّل yifaç˜il	مُفَاعِلِين mufaAçiliyn	فَعَل façal

Suite page suivante . . .

Schème	Schème	Schème	Schème
تَفَعِّيَةُ tafaʕ˜iyaħu	يفَعِّل yfaʕ˜il	مَفَاعِيل mafaAʕiyl	فُعَل fuʕal
تَفعِيَةُ tafʕiyaħu	يَفعَل yafʕal	مَفَال mafaAl	فِعَل fiʕal
تفَعِّيت tfaʕ˜iyt	يِفعَل yifʕal	مِفتَال miftaAl	فعَل fʕal
تَفعِيل tafʕiyl	يُفعِل yufʕil	مِفتَالَة miftaAlaħ	فَعُل faʕul
تَفعِيلًا tafʕiylãA	يَفعِل yafʕil	مُفتَالِين muftaAliyn	فُعُل fuʕul
تَفعِيلَاتُ tafʕiylaAtu	يُفَعلِلُ yufaʕlilu	مُفتَع muftaʕ	فَعِل faʕil
تَفعِيلَة tafʕiylaħ	يفَعلِل yfaʕlil	مُفتَعَّةً muftaʕ˜aħã	فعِل fʕil
تَفعِيلَةَ tafʕiylaħa	يفَعلِلُوا yfaʕliluwA	مُفتَعَّةَ muftaʕ˜aħa	فَعل faʕl
تَفعِيلَةُ tafʕiylaħu	يفُعلُوا yfuʕluwA	مُفتَعَل muftaʕal	فُعل fuʕl
تَفعِيلَي tafʕiylay	يفَعَّلُوا yfaʕ˜iluwA	مُفتَعِل muftaʕil	فِعل fiʕl
تِفعِيوَا tifʕiywA	يفَّعِلُوا yf˜aʕiluwA	مُفتَعِلَةُ muftaʕilaħu	فَعَلًا faʕalãA
تفُول tfuwl	يُفعلُوا yufʕluwA	مُفتَعَلَي muftaʕalay	فَعلًا faʕlãA
تفِيل tfiyl	يفَعلُوا yfaʕluwA	مُفتَعِلِين muftaʕiliyn	فُعلًا fuʕlãA
تَفَيَّل tafay˜al	يَفعلُوا yafʕluwA	مُفتَعَى muftaʕaY	فِعلًا fiʕlãA
تفِيلُوا tfiyluwA	يَفعُوا yafʕuwA	مُفتَعَينَ muftaʕayna	فعِلَا fʕilaA
تَنفَعِل tanfaʕil	يفِعُّوا yfiʕ˜uwA	مُفتَعَّينِ muftaʕ˜ayni	فَعِلَا faʕilaA
ʕِil عِل	يُفَعَّى yufaʕ˜aY	مَفَع mafaʕ	فَعلَا faʕlaA
ʕlaA علَا	يُفعَى yufʕaY	مِفَع mifaʕ	فَعَلَاً faʕalAã
ʕlaAt علَات	يَفعَى yafʕaY	مُفِع mufiʕ	فَعلَاً faʕlAã
ʕlaAwA علَاوَا	يِفعَى yifʕaY	مِفعَال mifʕaAl	فُعَلَاءٌ fuʕalaA’ũ
ʕliyt علِيت	يفَعِّي yfaʕ˜iy	مَفَعَّة mafaʕ˜aħ	فُعَلَاءَ fuʕalaA’a
ʕliynaA علِينَا	يِفعِي yifʕiy	مِفَعَّة mifaʕ˜aħ	فُعَلَاء fuʕalaA’
ʕuwA عُوا	يِفعِيوَا yifʕiywA	مُفَّعِل muf˜aʕil	فِعلَات fiʕlaAt
faA فَا	يَفُولُ yafuwlu	مَفعِلَّ mafʕil˜a	فَعَلَات faʕalaAt
faAt فَات	يفُول yfuwl	مُفِعل mufiʕl	فَعلَات faʕlaAt
faAʕ˜ũ فَاعٌّ	يفَوِّل yfaw˜il	مُفَعَّل mufaʕ˜al	فُعلَات fuʕlaAt
faAʕ˜a فَاعَّ	يِفَوِل yifawil	مُفَعِّل mufaʕ˜il	فَعلَال faʕlaAl
faAʕaA˜atũ فَاعَاتٌ	يفُولُوا yfuwluwA	مفَعّل mfaʕ˜l	فِعلَال fiʕlaAl
faAʕaA˜atu فَاعَاتُ	يفِيل yfiyl	مَفعَل mafʕal	فَعَلَانِ faʕalaAni
faAʕ˜aAt فَاعَّات	يفَيِّل yfay˜il	مُفعَل mufʕal	فَعلَانِ faʕlaAni
faAʕ˜aħ فَاعَّة	يفِيلُوا yfiyluwA	مَفعِل mafʕil	فَعَلَان faʕalaAn
faAʕil فَاعِل	يوَعِّل ywaʕi˜l	مُفَعِّلَاتٌ mufaʕ˜ilaAtũ	فعلَان fʕlaAn
faʕlaAn فَعلَان	فَعلَة faʕlaħ	فَعَّلت faʕ˜ilt	فَعلَلَةَ faʕlalaħa
fuʕlaAn فُعلَان	فُعلَة fuʕlaħ	فعَلت fʕalt	فَعلَلَةُ faʕlalaħu
faʕlaAnaħ فَعلَانَة	فِعلَة fiʕlaħ	فَعِلت faʕilt	فَعلَلَت faʕlalat

Suite page suivante . . .

Schème	Schème	Schème	Schème
فَعَلَانِيَّة faʕalaAniy˜aħ	فَعلَة faʕlaħa	فعِلت fʕilt	فَعِلنَ faʕilna
فَعلَانَين faʕlaAnayni	فعِلت fʕilt	فعَلتُوا fʕaltuwA	فَعَّلنَا faʕ˜alnA
فعِلَّة fʕil˜aħ	فعَّلَت fʕ˜alat	فعِلتُوا fʕiltuwA	فَعَّلنَا faʕ˜alnaA
فُعلَة fuʕlaħ	فَعلِت faʕlit	فَعلَتَين faʕlatayni	فعَّلنَا fʕ˜alnaA
فَعَلَة faʕalaħ	فِعلِت fiʕlit	فُعلَتَين fuʕlatayni	فعَلنَا fʕalnaA
فُعَلَة fuʕalaħ	فَعَّلت faʕ˜alt	فَعلِل faʕlil	فعِلنَا fʕilnaA
فَعِلَة faʕilaħ	فعَّلت fʕ˜alt	فِعلِلَة fiʕlilaħ	فَعِلُوا faʕiluwA
فَعّلُوا faʕ˜luwA	فُعلُولِيَّة fuʕluwliy˜aħi	فَعَلِيَّ faʕaliy˜a	فُعلُوا fuʕluwA
فَعلُوا faʕluwA	فَعلَى faʕlaY	فَعلِيَّ faʕliy˜a	فَعَّلُوا faʕ˜aluwA
فُعلُولِيَّة fuʕluwliy˜aħa		فَعَلِيٌّ faʕaliy˜ũ	

TABLE 17 – La Liste des schèmes.

Annexe C : Liste des mots outils

Nous présentons dans cette annexe la transcription des nombres en DT (*voir* 18) et la transcription des mots-outils présents dans le corpus STAC (*voir* 19).

Nombre	Les nombres en DT	Nombre	Les nombres en DT
0	صفر Sfr	10	عشرة ʕšrħ
1	وَاحد wAHd	11	حدَاش HdAš
2	ثنين θnyn	12	ثنَاش θnAš
3	ثلَاثة θlAθħ	13	ثلتطَاش θltTAš
4	اربعة ArbʕAħ	14	اربعطَاش ArbʕTAš
5	خمسة xmsħ	15	خمسطَاش xmsTAš
6	ستة stħ	16	ستطَاش stTAš
7	سبعة sbʕħ	17	سبعطَاش sbʕTAš
8	ثمنية θmnyħ	18	ثمنطَاش θmnTAš
9	تسعة tsʕħ	19	تسعطَاش tsʕTAš
1000	الف Alf	100	مية myħ
2000	الفين Alfyn	200	ميتين mytyn

TABLE 18 – La transcription des nombres en dialecte tunisien.

Mot outil	Catégorie grammaticale	Mot outil	Catégorie grammaticale
متَاع mtAʕ	Pronom possessif	اهلَا AhlA	Particule de voix
متَاعه mtAʕh		اي Ay	
متَاعهَا mtAʕhA		ايه Ayh	
متَاعهم mtAʕhm		حَاصيلو HASylw	
متَاعك mtAʕk		ايَا AyA	
متَاعكم mtAʕkm		يَا yA	
متَاعنَا mtAʕnA		لكن lkn	Particule de restriction
متَاعي mtAʕy		مهمَا mhmA	
علَى ʕlỳ	Préposition	سوَا swA	
علي ʕly		غير γyr	
إلَى Alỳ		الَا AlA	
في fy		زوز zwz	Nom de nombre
من mn		لو lw	Conjonction de subordination
احنَا AHnA	Pronom	مع mʕ	
نحنَا nHnA		معَا mʕA	
آهم ’Ahm		اقل Aql	

Suite page suivante . . .

Mot outil	Catégorie grammaticale	Mot outil	Catégorie grammaticale
آهو ’Ahw		خَاطر xATr	
آهي ’Ahy		اوه Awh	
آنَا ’AnA		ابعد Abçd	
انتم Antm		اكثر Akθr	
هم hm		اقل Aql	
انتومَا AntwmA		إذَا ’ðA	
ويَانَا wyAnA		مخر mxr	Adverbe de temps
هي hy		تو tw	
ويَاكم wyAkm		توة twħ	
انتوم Antwm		بكري bkry	
ويَاك wyAk		تويكة twykħ	
نتي nty		الآن Al’An	
ويَاهم wyAhm		آمس Ams	
ويَاهَا wyAhA		غدويكة γdwyk	
ويَاه wyAh		قبيليكه qbylykh	
هو hw		قبيلي qbyly	
نحن nHn		قبيله qbylh	
هومَا hwmA		قريب qryb	
همَا hmA		غدوة γdw	
انتي Anty		مه mh	Interjection
حيث Hyθ	Adverbe relatif	مرحبَا mrHbA	
كَاين kAyn		نعم nçm	
كيف kyf		تَاتَا tAtA	
كيفمَا kyfmA		تي ty	
كيمَا kymA		ايّ Ayy	
اللّي Ally	Pronom relatif	بَاه bAh	
للّي lly		بَاهي bAhy	
امَا AmA	Particule de restriction	بف bf	
كَان kAn		دّه dh	
هكة hkħ	Adverbe	هَا hA	
كهو khw		ايوَاه AywAh	
لي ly		ايه Aayh	
ليك lyk		ايه Aiyh	
ليلو lylw		اي Ay	
ليلي lyly		آه ’Ah	

Suite page suivante . . .

Mot outil	Catégorie grammaticale	Mot outil	Catégorie grammaticale
لية lyħ		الو Alw	
لنَا lnA		آمين ’Amyn	
رمَا rbmA		معنَاهَا mʕnAhA	
حتَى Htỳ		معنَاتهَا mʕnAthA	
عوض ʕ		غَادي γAdy	Pronom démonstratif
بيك byk		غَاديكة γAdykħ γAdyk	
بيكم bykm		ذَا ðA	
قبل qbl		ذلك ðlk	
دمَا dymA		اوكة Awkħ	
تحت tHt		ايضَا AyDA	
شيرة šyr	Adverbe de place	هَاك hAk	
بعد bʕd		هَاكه hAkh	
بعديكش bʕdykš		هَاكي hAky	
بعديكه bʕdykh		هَاكية hAkyħ	
بين byn		هنَاك hnAk	
قدَام qdAm		هذيكة hðykħ	
فوق fwq		هذَاك hðAk	
لهنَا lhnA		هكومة htwmħ	
ثمَا θmA		هَاكم hAkm	
هونيكة hwnykħ		هَاذومَا hAðwmA	
ثمة θmT		هَاذم hAðm	
بحذَاية bHðħ		هذية hðyħ	
هوني hwny		هذي hðy	
وسط wsT		هذومَا hðwmA	
دَاخل dAxl		بش bš	Particule de futur
ورَا		بَاش bAš	
كذَا kðA	Particule d’abstraction	لهَا lhA	Pronom d’objet indirect
عن ʕn	Particule	لهم lhm	
اي Ay		لك lk	
عن ʕn		لكم lkm	
ام Am		لنَا lnA	
قد qd		لي ly	
رَاهَا rAhA		لو lw	
رَاهم rAhm		فيسع fysʕ	Particule de verbe
رَاهو rAhw		مَاهم mAhm	

Suite page suivante . . .

Mot outil	Catégorie grammaticale	Mot outil	Catégorie grammaticale
رَاهي rAhy		مَاهو mAhw	
رَاك rAk		مَاهوَاش mAhwAš	
رَاكم rAkm		مَاهومَا mAhwmA	
رَانَا rAnA		مَاهيَاش mAhyAš	
رَاني rAny		مَاكم mAkm	
ليت lyt		مَانَا mAnA	
ڠ جميع		مَاني mAny	
اكهو Akhw		مَاهي mAhy	
بَالقدَا bAlqdA		ايوَاه AywAh	
بيهم byhm		مذبي mðby	
فقط fqT		اذن Aðn	
هكَاكة hkAkℏ		شكون škwn	Adverbe interrogatif
لمَا lmA		شنو šnw	
لي ly		شنومَا šnwmA	
ليكم lykm		شنوة šnw	
ليلهَا lylhA		شني šny	
ليلهم lylhm		شنية šnyℏ	
ليلك lylk		ڠلَاش ɣlAš	
ليلكم lylkm		ڠلَاه ɣlAh	
ليلنَا lylnA		آش Aš	
لينَا lynA		امَا AmA	
منَا mnA		فَاش fAš	
لين lyn		فين fyn	
ثم θm		هل hl	
عند ɣnd		لوَاش lwAš	
هكَاية hkAyℏ		لوَاه lwAh	
هنَا hnA		لوين lwyn	
حسب Hsb		منَاش mnAš	
شوية šwy		منين mnyn	
بحذَا bHðA		وقتَاش wqtAš	
بعض bɣD		وقتَاه	
بلَاش blAš		وين wyn	
برك brk		يَالندرَا	
بي by		كيفَاش kyfAš	
بيه byh		كيفَاه kyfAh	

Suite page suivante . . .

Mot outil	Catégorie grammaticale	Mot outil	Catégorie grammaticale
byhA بيهَا		qdAš قدَاش	
bynA بينَا		qdAh قدَاه	
bynAt بينَات		AnA انَا	Particule d'interrogation
’an أَن	Particule de condition	انَاهي	
hAn هَان		Anỳ انَى	
HAl حَال		yAxy يَاخي	
HAlhm حَالهم		yndrA يندرَا	
HAlkm حَالكم		lA لَا	Particule de négation
HAlny حَالني		mš مش	
kArhA كَارهَا		mA مَا	
kArk كَارك		mwš موش	
kArnA كَارنَا		kl كل	Nom de comptage
kArw كَارو		klkyA كلكيَا	
’An إِن		zAdħ زَادة	Pronom démonstratif
HAlhA حَالهَا		hðA هذَا	
HAlk حَالك		hðAkh هذَاكه	
HAlnA حَالنَا		hðAyh هذَايه	
HAlw حَالو		hðdm هذم	
kArhm كَارهم		ðdwkh هذوكه	
kArkm كَاركم		hðwkm هذوكم	
kArny كَارني		Aw او	Conjonction

TABLE 19 – La Liste des mots-outils.

Annexe D : Sorties des outils développés

D.1. Sortie du système « Al-Khalil-TUN »

Nous présentons dans cette annexe la sortie de notre analyseur morphologique pour le dialecte tunisien.

Analysis Results

الدخل INPUT	الخرج OUTPUT												
	الكلمة المشكولة Voweled Word	السابق Prefix + Proclitic	الجذع Stem	الوزن Pattern	الجذر Root	الحالة الإعرابية POS Tags	اللاحق Suffix + Enclitic	الإسناد Person	المبني للمعلوم و المبني للمجهول Voice	التعريف State	العدد Number	النوع Gender	الزمن Aspect	اللهجة Dialect
عالسلامة	عَالسَّلَامَة	ع : prc0 = aa_prep + ال : prc1 = Al	سلامة	فعالَة	سلم	pos = noun	تْ : suff = fem_tah	Not applicable	Not applicable	Definite	Singular	Feminine	Not applicable	TA
مرحبا	مَرْحَبَا	prc0 = 0	مرحبا	#	#	pos = interj	enc0 = 0 + suff0 = 0	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	TA
بيكم	بِيكُم	prc0 = 0	بيكم	#	#	pos = adv	enc0 = 0 + suff0 = 0	2nd	Not applicable	Not applicable	Plural	Not applicable	Not applicable	TA

FIGURE 3 – Un exemple de sortie de l'analyseur morphologique « Al-Khalil-TUN ».

D.2. Sortie du système « STAr-Tun »

Nous présentons dans cette annexe la sortie de notre segmenteur de phrases pour le dialecte tunisien.

```
<?xml version="1.0" encoding="UTF-8"?>
<ن>
  <ق>
    <ج>السادة المشاهدين الكرام # صحة شريبتكم</ج>
    <ج>وربي يفوركم</ج>
    <ج>ويفضلكم بجاه القرآن العظيم اليوم # نشوفوا</ج>
    <ج>السنة الخامسة # من الهجرة ثمة غزوة كبيرة ياسر #</ج>
    <ج>ربي_سبحانو وتعالى # نجى فيها المومنين</ج>
    <ج>وعلمهم حاجة ما ينساوهاش المومنين الكل ليوم القيامة #</ج>
    <ج>شنية هو كون الإنسان</ج>
    <ج># ولو ضعيف</ج>
    <ج># ولو فقير</ج>
    <ج># ولو يحس بروحو # تاعب</ج>
    <ج>إذا شد في ربي # واخلص نيتو لله راهو</ج>
    <ج># ربي_سبحانو وتعالى</ج>
    <ج>ينصرو ويفوزو</ج>
  </ق>
</ن>
```

FIGURE 4 – Exemple de sortie du segmenteur STAr-TUN.

الخلاصة :

هذه الأطروحة تندرج في إطار المعالجة الآلية للغة المنطوقة وتهتم بموضوع خلق موارد لغوية للهجة التونسية. أولا، قمنا باقتراح طريقة لبناء المدونة STAC (مدونة اللهجة التونسية المحكية). وتبدأ هذه الطريقة باقتراح اتفاقيتين لكتابة كلمات اللهجة التونسية وتبيان الظواهر الناتجة عن النطق العفوي للهجة. ثم استخدمنا المدونة STAC ومعجم" جذر نموذج" للفصحى لإنشاء معجم للهجة التونسية. وقد استخدم هذا الأخير لتحليل الصرفي للهجة التونسية. لحل مشكلة الغموض الناجم عن التحليل الصرفي، اقترحنا طريقة إحصائية لاختيار تحليل صرفي واحد صحيح لكل كلمة. وأخيرا، اقترحنا طريقة هجينة تعتمد على مجموعة من القواعد السياقية وطريقة إحصائية للكشف عن حدود الجمل في اللهجة التونسية. تظهر نتائج التقييم أن الطرق المختلفة المقترحة لخلق الموارد اللغوية للهجة التونسية هي طرق واعدة ويمكن استغلالها لاقتراح طرق قادرة على كشف وتصحيح الآلي للأخطاء الناتجة عن النطق العفوي للهجة.

الكلمات الجوهرية : اللهجة التونسية، نسخ كتابي، المدونة STAC، إنشاء معجم، التحليل الصرفي، كشف عن حدود الجمل، توضيح أقسام الكلام.

Résumé :

Cette thèse s'intègre dans le cadre du traitement automatique de la langue parlée et s'intéresse à la création des ressources linguistiques pour le dialecte tunisien. D'abord, nous avons décrit une méthode pour la création du corpus STAC (Spoken Tunisian Arabic Corpus). Cette méthode commence par l'élaboration de deux conventions de transcription orthographique pour écrire les mots dialectaux et annoter les phénomènes dus au caractère spontané des productions orales. Ensuite, nous avons utilisé le corpus STAC et un lexique « racine-patron » de l'arabe standard afin de créer un lexique pour le dialecte tunisien. Ce dernier a été exploité pour analyser morphologiquement le dialecte tunisien. Pour résoudre le problème d'ambiguïté causé par l'analyse morphologique, nous avons proposé une méthode statistique permettant de choisir une seule analyse correcte pour un mot dans une phrase. Enfin, nous avons proposé une méthode hybride qui se fonde sur un ensemble de règles contextuelles et une méthode statistique afin de détecter les frontières des phrases en dialecte tunisien. Les résultats d'évaluation montrent que les différentes méthodes proposées pour le développement des ressources pour le dialecte tunisien sont prometteuses et elles peuvent être exploitées pour proposer des méthodes permettant la détection et la correction automatique des disfluences.

Mots clés : dialecte tunisien, transcription orthographique, corpus STAC, création de lexique, analyse morphologique, segmentation des transcriptions en des phrases, désambiguïsation morphosytaxique.

Abstract:

This thesis deals with the linguistic resources creation of spoken Tunisian Arabic. First, we described a method for creating the STAC corpus (Spoken Tunisian Arabic Corpus). Our method started with the definition of two orthographic transcription conventions for writing dialectal words and annotating spontaneous oral phenomena. Then, we proposed a method for creating a Tunisian Arabic lexicon based on the STAC corpus and a modern standard Arabic lexicon. This lexicon was exploited to morphological analyze the Tunisian Arabic. To solve the ambiguity caused by the morphological analysis, we proposed a statistical method that is able to choose one correct analysis for a word in a given sentence. We proposed a hybrid method based on a set of contextual rules and a statistical method in order to detect sentence boundaries. The obtained results show that the different methods proposed for resource development for the Tunisian dialect are promising and can be exploited to provide methods for the automatic detection and correction of disfluencies.

Keywords: Tunisian Arabic, orthographic transcription, STAC corpus, lexicon creation, morphological analysis, sentence boundary detection, morphosytactic disambiguation.