AIX-MARSEILLE UNIVERSITÉ

ED 356 COGNITION, LANGAGE, ÉDUCATION

LABORATOIRE D'INFORMATIQUE FONDAMENTALE
CNRS-UMR-7279

Thèse présentée pour obtenir le grade universitaire de docteur

Discipline : Sciences du langage
Spécialité : Traitement automatique des langues

Ahmed HAMDI

Traitement automatique du dialecte tunisien à l'aide d'outils et de ressources de l'arabe standard : application à l'étiquetage morphosyntaxique

Soutenue le 04/12/2015 devant le jury :

Nabil HATHOUT	Université de Toulouse 2	Rapporteur
Imed ZITOUNI	Microsoft	Rapporteur
Alexis NASR	Aix-Marseille Université	Directeur de thèse
Núria GALA	Aix-Marseille Université	Co-directrice de thèse

Résumé

Le développement d'outils de traitement automatique pour les dialectes de l'arabe se heurte à l'absence de ressources pour ces derniers. Comme conséquence d'une situation de diglossie, il existe une variante de l'arabe, l'arabe moderne standard, pour laquelle de nombreuses ressources ont été développées et ont permis de construire des outils de traitement automatique de la langue. Étant donné la proximité des dialectes de l'arabe, avec l'arabe moderne standard, une voie consiste à réaliser une conversion surfacique du dialecte vers l'arabe moderne standard afin de pouvoir utiliser les outils existants pour l'arabe standard.

Dans ce travail, nous nous intéressons particulièrement au traitement du dialecte tunisien. Nous proposons un système de conversion du tunisien vers une forme approximative de l'arabe standard pour laquelle l'application des outils conçus pour ce dernier permet d'obtenir de bons résultats.

Afin de valider cette approche, nous avons eu recours à un étiqueteur morphosyntaxique conçu pour l'étiquetage de l'arabe standard. Ce dernier permet d'assigner des étiquettes morphosyntaxiques à la sortie de notre système de conversion. Ces étiquettes sont finalement projetées sur le tunisien.

Notre système atteint une précision de 89% suite à la conversion qui représente une augmentation absolue de $\sim$20% par rapport à l'étiquetage d'avant la conversion.

Mots clés : traitement automatique des langues, étiqueteur en parties de discours, outils, ressources, conversion, arabe moderne standard, dialecte tunisien.

Abstract

Developing natural language processing tools usually requires a large number of resources (lexica, annotated corpora, ...), which often do not exist for less-resourced languages. One way to overcome the problem of lack of resources is to devote substantial efforts to build new ones from scratch. Another approach is to exploit existing resources of closely related languages.

Taking advantage of the closeness of standard Arabic and its dialects, one way to solve the problem of limited resources, consists in performing a conversion of Arabic dialects into standard Arabic in order to use the tools developed to handle the latter.

In this work, we focus especially on processing Tunisian Arabic dialect. We propose a conversion system of Tunisian into a closely form of standard Arabic for which the application of natural language processing tools designed for the latter provides good results.

In order to validate our approach, we focused on part-of-speech tagging. Our system achieved an accuracy of 89% which presents $\sim$20% of absolute improvement over a standard Arabic tagger baseline.

Keywords : natural langage processing, part-of-speech tagger, resources, tools, conversion, modern standard Arabic, Tunisian dialect.

Table des matières

Liste des figures

Liste des tableaux

Introduction

Cette thèse s'inscrit dans le cadre général du traitement automatique du langage naturel (noté TAL dorénavant) et plus particulièrement dans celui du traitement automatique de l'arabe. Nous nous intéressons plus précisément à une variante de l'arabe, qui est l'arabe tunisien (noté par moments TUN dans le reste de ce document) et à la relation qu'il entretient avec l'arabe moderne standard, plus connu sous son acronyme anglais MSA (*Modern Standard Arabic*), que nous utiliserons ici.

Les applications phares du TAL, telles que la traduction automatique, la transcription automatique de la parole ou l'extraction d'informations à partir de textes, nécessitent toutes des ressources linguistiques importantes pour atteindre un niveau de qualité raisonnable. Ces ressources prennent des formes diverses, telles que des corpus bruts ou annotés, des lexiques ou des grammaires. Les connaissances qu'elles recèlent concernent des niveaux linguistiques variés, allant de la phonétique à la sémantique, en passant par la morphologie et la syntaxe. Le développement de telles ressources est un processus lent et coûteux et suppose des moyens humains, financiers et institutionnels importants. La conséquence de tout cela est que peu de langues possèdent de telles ressources. Les autres ont été rangées dans la catégorie des langues peu dotées, parmi lesquelles on trouve le tunisien. Le MSA, en revanche appartient à l'autre catégorie, celle des langues dotées.

L'idée générale sous-jacente à cette thèse est de mettre à profit la proximité du tunisien et du MSA pour développer des outils de TAL pour le tunisien. Plus précisément, nous proposons de convertir le tunisien en une approximation du MSA qui est assez proche de ce dernier pour pouvoir utiliser des outils existant pour le traitement automatique du MSA. Le processus de conversion que nous proposons repose principalement sur la morphologie et le lexique. Sa mise en œuvre nous a amené à nous intéresser de près à la morphologie de l'arabe, tant du MSA que du tunisien, ainsi qu'aux outils informatiques pour le traitement de la morphologie. Du point de vue du lexique, nous nous sommes intéressé au développement de lexiques "bilingues", tunisien, MSA en proposant notamment des moyens semi-automatiques pour les développer.

Une particularité de notre travail est d'avoir cherché à réaliser une analyse morphologique fine et profonde du tunisien, allant jusqu'à l'extraction de racines et de schèmes, dans le but de simplifier le processus de conversion du

tunisien vers le MSA. Le traitement automatique de la morphologie de l'arabe du fait de sa richesse et de sa complexité est un défi pour TAL. Les outils standard de traitement automatique de la morphologie, développés généralement pour le traitement automatique des langues indo-européennes ne sont souvent pas adéquats pour traiter les langues sémitiques. Nous avons collaboré, dans le cadre de cette thèse, avec des spécialistes du traitement automatique de la morphologie de l'arabe pour développer notamment un analyseur morphologique du tunisien, qui constitue une brique importante du système que nous proposons.

Dans le but de valider notre approche, nous avons choisi un outil de TAL standard : un étiqueteur morphosyntaxique. L'idée est d'utiliser un étiqueteur morphosyntaxique pour le MSA sur la sortie de notre outil de conversion et d'en étudier les performances. D'autres outils auraient pu être utilisés dans ce but, notamment un analyseur syntaxique. Nous avons préféré avoir recours à un étiqueteur morphosyntaxique car il s'agit à la fois d'un outil simple et très utilisé en TAL.

L'évaluation d'un étiqueteur morphosyntaxique du tunisien est confronté à l'absence de corpus étiqueté en parties de discours pour le tunisien. Nous nous sommes donc attelés au développement d'un tel corpus, dans un but d'évaluation. Ce développement est lui même confronté au problème de l'absence de conventions orthographiques pour les dialectes de l'arabe en général et le tunisien en particulier (Habash, 2010), contrairement au MSA pour lequel un système orthographique standard a été établi. Habash *et al.* (2012) fait partie des rares travaux qui se sont intéressés à l'établissement des conventions orthographiques pour les dialectes arabes. Il propose des conventions [a] communes qui peuvent être partagées pour tous les dialectes du monde arabe.

La méthodologie que nous proposons dans ce travail pour traiter le tunisien à l'aide d'outil développés pour le MSA peut être appliqué à tous les dialectes de l'arabe et même à d'autres dialectes qui se trouvent dans la même situation, à savoir l'existence d'une variante pour laquelle de nombreuses ressources ont été développées.

Ce document est composé de cinq chapitres regroupés en deux parties principales : une première partie composé de trois chapitres (incluant cette introduction) constitue l'état de l'art de notre travail. Cette partie se focalise sur la description des caractéristiques morphosyntaxiques de l'arabe ainsi que les différentes méthodes et techniques informatiques pour réaliser le traitement automatique de la morphologie complexe de l'arabe. Dans la deuxième partie, qui se compose de deux chapitres, nous décrivons en détail notre méthode et nous l'évaluons sur l'étiquetage en partie de discours du TUN.

Le chapitre 1 aborde quelques notions liées à la langue arabe. Dans ce chapitre, nous mettons en relief les phénomènes d'agglutination, de flexion et de dérivation en arabe et nous illustrons les problèmes qui y sont liés. Nous donnons,

a. Ce travail rentre dans le cadre du projet CODA (*Conventional Orthography for Dialectal Arabic* (Habash *et al.*, 2012)).

également, un aperçu sur les deux variantes de l'arabe qui nous intéressent dans ce travail à savoir le MSA et l'arabe dialectal et plus particulièrement le dialecte tunisien. Nous effectuons une étude profonde pour distinguer les différences et les similarités morphosyntaxiques du MSA et du TUN.

Le chapitre 2 porte sur le traitement automatique de la morphologie en général en insistant sur les outils spécifiques au traitement des langues gabaritiques. Ce chapitre présente également les analyseurs morphosyntaxiques les plus connus pour le traitement de l'arabe.

Le chapitre 3 est consacré à la description des outils et des ressources auxquels nous avons eu recours pour la réalisation de la méthode que nous proposons pour répondre à notre problématique.

Enfin, le chapitre 4 porte sur l'évaluation. Nous donnons dans ce chapitre tous les résultats obtenus à l'issue des différentes expériences que nous avons effectuées.

Le reste de ce chapitre est composé de deux sections. La première propose un rapide survol des variantes de la langue arabe et la seconde aborde la problématique du traitement automatique des langues peu dotées.

Survol des variantes de la langue arabe

La langue arabe était la langue de quelques tribus nomades (Holes, 2004). Actuellement, elle est la langue de plus que 300 million d'arabophones vivant dans 23 pays arabes (cf. figure 0.1). Cependant, la langue arabe présente d'une part des variantes stables partagées par tous les arabophones et couvre, d'autre part, une diversité de dialectes changeant d'un pays à un autre.

Figure 0.1.: Monde arabe : répartition des arabophones

À l'époque pré-islamique, l'arabe était la langue de communication de quelques peuplades vivant principalement dans la péninsule arabe. Depuis l'apparition de l'Islam au septième siècle, l'arabe a connu une expansion géographique considérable. Suite à la propagation de l'Islam, l'arabe s'est diffusé sur un grand empire qui couvre la péninsule arabique, le moyen-orient, l'Afrique du nord et le sud de l'Espagne. On distingue trois registres principaux de l'arabe, l'arabe classique, le MSA et l'arabe dialectal.

L'arabe classique constitue la variante la plus ancienne de la langue. C'est la langue employée dans le Coran, le livre sacré des musulmans. Actuellement, grâce aux textes religieux et l'étude de la poésie ancienne, l'arabe classique est encore présent dans les systèmes éducatifs des pays arabes quoiqu'il reste généralement très à l'écart de l'arabe utilisé dans nos jours.

Le MSA est la langue officielle de tous les pays arabes. Il constitue la version modernisée et standardisée de l'arabe classique. En effet, avec le temps, des termes anciens ont disparu ou bien ont été remplacés par d'autres et des termes nouveaux sont apparus pour répondre notamment aux évolutions de la société. Citons comme exemple de substitution le terme دجّة *dj∼ħ* "*obscurité*" qui a disparu avec l'arabe classique et dont le sens est exprimé en MSA par ظلمة *Ďlmħ*. On peut citer comme exemple de termes nouveaux مكننة *mknnħ* "*mécanique*" et تقنية *tqnyħ* "*technologie*" liés au progrès scientifique. En dépit de ces changements lexicaux entre l'arabe classique et le MSA, ce dernier a maintenu les systèmes

morphologique, grammatical et syntaxique de l'arabe classique.

Le MSA est employé dans les domaines administratifs et éducatifs ainsi que dans la communication formelle écrite et orale des pays arabes. Bien que le MSA représente la langue commune de toute la population arabe du golfe à l'océan, il ne constitue la langue maternelle d'aucun arabophone. En effet, les arabophones acquièrent dès leur petite enfance un dialecte arabe en fonction de leur lieu de naissance.

Les dialectes arabes représentent ainsi les langues vernaculaires. Ils sont utilisés dans les conversations quotidiennes des arabophones. Ces dialectes sont le résultat de l'interférence linguistique entre la langue arabe et les langues locales ou voisines, à l'issue d'une influence culturelle due principalement à la colonisation, aux mouvements migratoires, et récemment aux médias (Bassiouney, 2009). Ils sont en perpétuelle évolution, incluant constamment de nouveaux mots empruntés la plupart du temps à des langues occidentales géographiquement proches comme le français, l'espagnol, l'italien ou l'anglais. Ce sont les dialectes qui sont utilisés pour la communication de tous les jours dans les pays concernés. Récemment, depuis quelques années, l'emploi des dialectes n'est plus restreint à l'oral, ils commencent à être utilisés pour la communication écrite informelle dans le web (forums, blogs, réseaux sociaux...).

Les arabophones ne considèrent pas que le MSA et l'arabe dialectal sont deux langues séparées. Cette perception conduit à une situation particulière de coexistence de deux formes d'une même langue. Cette situation est appelée diglossie (Ferguson, 1959) : "*Diglossia is likely to come into being when the following three conditions hold in a given speech community : (1) There is a sizable body of litterature in a language closely related to the natural language of the community, and this litterature embodies, whether as source or reinforcement, some of the fondamental values of the community. (2) Literacy in the community is limited to a small elite. (3) A suitabe period of time, of the order of several centuries, passes from the establishment of (1) and (2)*".

Bien que ces deux variantes sont clairement prédominantes dans deux domaines différents, l'écrit formel (MSA) et l'oral informel (dialectes), il existe également une forme qui combine les deux variantes (Bassiouney, 2009) utilisée dans les réseaux sociaux, les forums et les débats télévisés. En effet, cette forme assemble, dans une même phrase, des termes de l'arabe standard et d'autres termes de l'arabe dialectal. Les termes eux-mêmes peuvent être dérivés d'un mot d'une variante et subissent l'agglutination ou la flexion de l'autre variante.

La classification des dialectes arabes dépend principalement de deux facteurs distincts : un facteur géographique et un autre sociologique.

Distinction sociologique Cette distinction est liée principalement à des facteurs religieux et sociaux. Au niveau religieux, plusieurs variations sont distinguées selon la religion, la secte et la doctrine. Du point de vue social, deux variantes majeures existent dans un dialecte : un dialecte citadin qui est parlé par

les habitants des cités et considéré comme étant plus prestigieux et un dialecte bédouin qui est moins bien considéré. La différence entre ces variantes concerne généralement les aspects phonologiques et lexicaux, les systèmes morphologique et syntaxique restant globalement invariables.

Distinction géographique Du point de vue géographique, chaque pays arabe utilise son propre dialecte. Des variations linguistiques existent également entre les régions d'un même pays. D'autres classifications regroupent les dialectes des pays voisins. Ce regroupement est fondé sur le partage de la majorité du lexique. Brustad (2000) et Bassiouney (2009) par exemple proposent la classification [b] suivante :

- dialecte égyptien (EGY) : couvre le dialecte de l'Égypte et du Soudan
- dialecte levantin (LEV) : parlé au Liban, en Syrie, en Jordanie et en Palestine
- dialectes du Golfe (GLF) : inclut les dialectes des pays du Golfe
- dialectes du Maghreb (MAG) : couvre les dialectes tunisien, libyen, algérien et marocain

Le dialecte arabe qui nous intéresse dans ce travail est le dialecte tunisien. Ce dernier est parlé par 12 millions de personnes habitant généralement en Tunisie. Cette variante de l'arabe est sous l'influence constante d'autres langues. Ceci est liée généralement à l'histoire du pays. En effet, toutes les nations qui ont transité par la Tunisie ont laissé des traces dans le dialecte des tunisiens à l'instar des phéniciens, des grecs, des romains, des vandales, des byzantins, des arabes, des turcs et des français, pour ne citer que les plus marquants. Cette diversité a fourni au dialecte tunisien de nouveaux termes et l'a rendu riche et variée.

Traitement automatique des langues peu dotées

Deux approches principales s'offrent à celle ou celui qui s'intéresse au traitement automatique d'une langue peu dotée. La première consiste à changer son statut de langue peu dotée en développant des ressources pour cette langue. C'est dans cette optique que Al-Sabbagh et Girju (2012) et Mohamed *et al.* (2012) ont annoté des corpus égyptiens extraits des réseaux sociaux pour entrainer des modèles de segmentation et d'étiquetage en parties de discours pour l'égyptien.

La deuxième approche, qui est moins coûteuse, se base sur l'utilisation de ressources et d'outils d'une langue proche. Plusieurs méthodes ont été proposées dans ce cadre pour traiter une langue (L_1) avec des outils développés pour une langue (L_2).

b. Cette classification est une parmi plusieurs et ne signifie pas que chaque dialecte est entièrement homogène d'un point de vue linguistique.

La première méthode consiste à adapter un outil existant pour L_2. Cette adaptation peut prendre plusieurs formes. Bernhard *et al.* (2013), par exemple, ont adopoté une telle approche pour adapter un étiqueteur de l'allemand à l'alsacien. Ils montrent que l'annotation manuelle d'une petite liste de mots provoque une amélioration significative de la précision de l'étiquetage. Dans le même esprit, Feldman *et al.* (2006) ont adapté un un étiqueteur morphosyntaxique de l'espagnol au portuguais et au catalan. Selon une méthode légèrement différente, Garrette et Baldridge (2013) ont montré l'efficacité de cette approche en adaptant des outils de traitement automatique de L_2 grâce à l'annotation de corpus de taille limitée pour L_1.

Une deuxième méthode consiste à traduire des corpus annotés de L_2 vers L_1 et de se servir de cette traduction comme corpus d'apprentissage pour des outils de L_1. Scherrer *et al.* (2009) par exemple, s'est intéressé à la traduction depuis l'allemand vers différents dialectes suisses allemands. Ce système repose sur une analyse syntaxique de l'allemand et c'est à l'issue de l'analyse syntaxique qu'un mécanisme de transfert permet de générer une traduction en dialecte. Plus proche de nou d'un point de vue linguistique, Shaalan *et al.* (2007) ont décrit un système de transfert de l'égyptien vers le MSA. Dans ce cas, le transfert est effectué au niveau des lemmes.

Une troisième méthode consiste à rapprocher L_1 de L_2 afin de pouvoir y appliquer des outils développés pour L_2. Le cas extrême consiste à traduire automatiquement L_1 en L_2, comme le proposent Yarowsky *et al.* (2001), Das et Petrov (2011) et Duong *et al.* (2013). Une telle approche n'est, bien entendu possible que s'il existe des corpus parallèles L_1, L_2. Certains travaux se sont servis de dictionnaires au lieu des corpus parallèles (Li *et al.*, 2012) et d'autres ont combiné les deux ressources (Täckström *et al.*, 2013). Vergez-Couret (2013) ont montré que de bons résultats peuvent être atteints en se limitant à la traduction des 300 mots les plus fréquents. Ce travail a été testé sur l'occitan avec le français d'une part, et l'occitan avec le castillan d'autre part.

Nos travaux se situent dans la troisième approche. L'idée que nous explorons consiste à *convertir* le TUN vers le MSA afin de pouvoir y appliquer des outils conçus pour le MSA. Nous avons utilisé à dessein le verbe *convertir* et non le verbe *traduire*. La raison est que nous ne cherchons pas une traduction de notre entrée en TUN vers une version en MSA qui soit intelligible pour un lecteur humain. Nous souhaitons nous approcher suffisamment du MSA afin que des outils développés pour ce dernier puissent donner de bons résultats sur cette approximation, que nous appellerons dorénavant pseudo-MSA. Nous verrons dans le chapitre 4 des sorties du système de conversion qui ne constituent pas des formes acceptables du MSA, mais sur lesquelles un étiqueteur morphosyntaxique permet de prédire la séquence d'étiquettes correcte.

De façon plus précise, la conversion que nous proposons repose largement sur la morphologie et le lexique. C'est en effet à ces deux niveaux que se manifestent la majorité des différences entre les variétés de l'arabe. Le système proposé relève

d'une architecture à transfert. Un mot en TUN est analysé sous la forme d'une racine, d'un schème et de traits morphologiques. Un lexique bilingue permet alors de convertir la racine et le schème source vers une racine et un schème cible (MSA). La racine et le schème cible, ainsi que les traits morphologiques vont alors permettre de générer un ou plusieurs mots cibles. Un étiqueteur en parties de discours entraîné sur des corpus MSA existant sera ainsi appliqué sur les mots cibles pour assigner les parties de discours adéquates aux mots MSA cibles. Ces étiquettes seront enfin projetées sur le texte tunisien.

Notre système réalise une analyse morphologique profonde, de manière à identifier la racine du mot cible plutôt, et non une analyse surfacique, qui aurait généré son lemme. La raison de ce choix est double : d'une part, la morphologie dérivationnelle de l'arabe est très régulière, l'identification de la racine peut être réalisée, de manière fiable et économique, à l'aide de règles. D'autre part, le fait de réaliser le transfert au niveau des racines permet de minimiser la taille du dictionnaire bilingue. On estime en effet à 7502 le nombre total de racines de l'arabe et à 2900 celui des racines fréquemment utilisées (Altabbaa *et al.*, 2010), ce qui permet de définir une borne supérieure de notre dictionnaire.

Le système que nous proposons est bi-directionnel : tous les modules qui le composent sont réversibles, ce qui permet de réaliser la traduction depuis le TUN vers le MSA et vice-versa. Notre système de conversion peut donc être utilisé dans le cadre de la deuxième approche que nous avons évoqué ci-dessus.
[c]

Peu de travaux se sont intéressés au traitement morphosyntaxique du tunisien : Zribi *et al.* (2013), par exemple, ont étendu la couverture d'un analyseur morphologique existant du MSA pour couvrir le TUN. Ils se sont servi d'un lexique MSA/TUN pour alimenter l'analyseur avec des racines spécifiques au TUN. Le même lexique a été ensuite exploité par Boujelbane *et al.* (2014) pour traduire des corpus d'apprentissage volumineux du MSA vers le TUN. Le corpus déduit a été utilisé pour entrainer un étiqueteur morphosyntaxique du dialecte tunisien.

Contrairement à Boujelbane *et al.* (2014) qui traduit des corpus annotés en MSA vers le TUN, notre système convertit le TUN vers le MSA. La raison principale pour laquelle nous avons choisi ce sens de conversion est que le processus de conversion est généralement ambigu et que les sorties du système de conversion doivent être désambiguïsées. Or la désambiguïsation à l'aide d'un modèle de langage, par exemple, peut être réalisé du côté MSA, pour lequel de grands corpus existent. En revanche, elle est beaucoup plus difficile à réaliser du côté tunisien.

Une autre différence importante entre notre travail et ceux de Boujelbane *et al.* (2014) et Zribi *et al.* (2013) est que notre système de conversion peut être limité à la génération de lemmes. Nous verrons dans le chapitre 4 que certains traitements, notamment l'étiquetage morphosyntaxique donne de meilleurs résultats

c. La traduction du MSA vers le TUN peut être intéressante dans une application de transcription automatique de la parole : on traduit en dialecte un corpus MSA afin de construire un modèle de langage pour le dialecte.

sur les lemmes.

1. Systèmes phonologique et morphosyntaxique de l'arabe

La langue arabe présente des phénomènes phonologiques et morphologiques spécifiques. Nous introduisons, dans ce chapitre, tous les termes linguistiques spécifiques à la langue arabe qui sont nécessaires pour bien appréhender le contexte linguistique de ce travail. Ces termes sont les suivants : diacritique, clitique, radical, racine, schème. La terminologie que nous adoptons suit les deux références (Al-Dahdah, 1996) [a] et (Al-Ghulayaini, 2006) [b]. La présentation que nous faisons de ces concepts est inspirée par le TAL. Nous nous focalisons notamment sur la description des ambiguïtés auxquelles le traitement morphosyntaxique est confronté. Pour bien décrire ces ambiguïtés, nous avons sélectionné des exemples qui illustrent les ambiguïtés dans les différentes étapes du traitement.

Dans la section 1.1, nous présentons le système phonologique du MSA. La section 1.2 donne une description détaillée des phénomènes morphosyntaxiques de l'arabe qui sont l'agglutination, la flexion et la dérivation. Un aperçu sur les catégories grammaticales utilisées dans la langue arabe est donné dans la section 1.3. Nous décrivons également la morphologie de l'arabe dialectal dans la section 1.4 et plus particulièrement le dialecte tunisien (TUN) dans la section 1.5. Afin de bien présenter les différences et les similarités entre le MSA et le TUN, nous avons eu recours à une analyse en profondeur de la morphologie.

1.1. Système d'écriture de l'arabe

L'écriture arabe est constituée d'un ensemble de symboles écrits de droite à gauche. Deux types de symboles existent : des consonnes qui constituent l'alphabet et des voyelles. Afin de faciliter la lecture aux non-arabophones, nous représentons les symboles arabes avec des caractères latins, cette opération est

a. (Al-Dahdah, 1996) expose les paradigmes verbaux et nominaux de l'arabe dans des tableaux. Il décrit d'une manière simple les systèmes de conjugaison verbale et de déclinaison nominale de cette langue riche par ses termes et complexe dans sa grammaire.

b. (Al-Ghulayaini, 2006) présente les différentes catégories grammaticales de l'arabe et donne les caractéristiques morphologiques et syntaxiques de chaque catégorie.

appelée translittération [c]. Dans tout ce document, nous suivons la translittération proposée par Habash *et al.* (2007).

1.1.1. Alphabet

L'alphabet arabe est composé de 28 consonnes (lettres). Ces dernières possèdent plusieurs formes qui dépendent principalement de leurs positions dans le mot. Le tableau 1.1.1 fournit la liste des consonnes, leurs noms, leurs formes et leurs translittérations.

Hormis les cinq consonnes (د *d*, ذ *ð*, ر *r*, ز *z*, و *w*) qui ne se lient pas avec les consonnes qui les suivent dans le sens de l'écriture (à gauche), toutes les consonnes arabes s'attachent avec les consonnes voisines. Les consonnes و *w* et ي *y* sont dites des semi-consonnes étant donné qu'elles peuvent être utilisées comme des voyelles (cf. section 1.1.2).

La première lettre de l'alphabet arabe (*hamza*) est particulière, elle s'écrit souvent à l'aide d'un support. Ce dernier peut être un *alif* (ا), un *wāw* (و) ou un *yā'* (ى), sa forme dépend des voyelles qui l'entourent. Bien que la *hamza* possède plusieurs formes أ, ؤ, ئ, elle se prononce toujours de la même façon /'/. Dans les écrits arabes la *hamza* أ écrite à l'aide du support *alif* est optionnelle. On retrouve généralement un *alif* simple ا à la place de أ. La translittération de la *hamza* évoque la forme de son support. Par exemple, un accent circonflexe est ajouté aux ا*A*, و *w*, ي *y* pour marquer leurs formes respectives avec la *hamza* أ*Â*, ؤ *ŵ*, ئ *ŷ*.

D'autres particularités existent dans le système d'écriture arabe telles que : le symbole **tā' marbūTa** ة *ħ* qui marque généralement le genre féminin des noms. Il apparait uniquement à la fin des noms et ne peut être suivi que d'une voyelle courte. Ce symbole est prononcé comme /t/ dans l'arabe moderne standard et reste souvent muet dans les dialectes arabes. Également, le symbole **alif maqSūra** ى *ý* apparait uniquement à la fin des mots et n'est précédé que de la voyelle courte /a/. Ce symbole marque des verbes défectifs (cf. section 1.3.2) et des noms féminins (cf. section 1.3.3).

Les lettres de l'alphabet sont classées en lettres lunaires et lettres solaires. L'identification du type de lettre est réalisée à l'aide de l'article défini ال*Al*, qui est invariant en genre et en nombre. En effet, la lettre initiale d'un mot déterminé est dite lunaire si la lettre ل *l* du déterminant est prononcée. Cette lettre est muette quand le mot commence par une lettre solaire. La lettre ق *q*, par exemple, est

c. contrairement à l'écriture arabe, la translittération est lue de gauche à droite.

lettre (forme isolée)	nom	forme			translittération
		initiale	médiale	finale	
ء	hamza	آ إ أ	ـئـ ـؤ ـأ	ء ئ ـؤ أ	’
ب	bā’	بـ	ـبـ	ـب	b
ت	tā’	تـ	ـتـ	ـت	t
ث	t̲ā’	ثـ	ـثـ	ـث	θ
ج	ǧīm	جـ	ـجـ	ـج	j
ح	ḥā’	حـ	ـحـ	ـح	H
خ	ḫā’	خـ	ـخـ	ـخ	x
د	dāl	د	ـد	ـد	d
ذ	d̲āl	ذ	ـذ	ـذ	ð
ر	rā’	ر	ـر	ـر	r
ز	zāy	ز	ـز	ـز	z
س	sīn	سـ	ـسـ	ـس	s
ش	šīn	شـ	ـشـ	ـش	š
ص	ṣād	صـ	ـصـ	ـص	S
ض	ḍād	ضـ	ـضـ	ـض	D
ط	ṭā’	طـ	ـطـ	ـط	T
ظ	ẓā’	ظـ	ـظـ	ـظ	Ď
ع	’ayn	عـ	ـعـ	ـع	ς
غ	ġayn	غـ	ـغـ	ـغ	γ
ف	fā’	فـ	ـفـ	ـف	f
ق	qāf	قـ	ـقـ	ـق	q
ك	kāf	كـ	ـكـ	ـك	k
ل	lām	لـ	ـلـ	ـل	l
م	līm	مـ	ـمـ	ـم	m
ن	nūn	نـ	ـنـ	ـن	n
ه	hā’	هـ	ـهـ	ـه	h
و	wāw	و	ـو	ـو	w
ي	yā’	يـ	ـيـ	ـي	y

Table 1.1.: Alphabet arabe

lunaire car la lettre ل de l'article défini du mot القمر *Alqamar* /alqamar/ "*la lune*" est prononcée. En revanche, la lettre ل *l* est muette dans le mot الشمس /aššams/ *Alšms* "*le soleil*". Par conséquent, la lettre ش *š* est dite solaire. Les mots cités comme exemples "*la lune*" et "*le soleil*" expliquent la classification des lettres en lunaires et solaires.

1.1.2. Voyelles

Toutes les consonnes présentées dans la section 1.1.1 s'accompagnent de voyelles. Deux types de voyelles existent en arabe : voyelles courtes et voyelles longues.

Voyelles courtes

Les voyelles courtes ou diacritiques sont des symboles situés au-dessus ou au-dessous des consonnes auxquelles ils sont affectés. Les diacritiques se regroupent en trois catégories :

(i) Diacritiques simples : ce sont des petits sons que l'on ajoute aux consonnes. On distingue quatre diacritiques simples : ـَ *a*, ـُ *u*, ـِ *i* et ـْ qui indique l'absence de tout son.

(ii) Diacritiques doubles : ce sont des diacritiques situés à la fin des noms arabes indéfinis[d]. Ces diacritiques se prononcent de la même manière que leurs homologues simples, sauf qu'on y ajoute le son /n/.

(iii) Chadda : elle se situe au dessus d'une consonne et a pour effet le doublement de cette dernière. Le symbole "*chadda*" ـّ ~ est toujours accompagné d'un diacritique simple.

(iv) Alif madda : c'est un diacritique qui permet de prononcer plus longuement la *hamza* أ *Â*. La *madda* آ *Ā* est utilisée toujours avec le support *alif* ا *A*.

Le tableau 1.2 donne la liste des diacritiques arabes, leurs translittérations et les sons qu'ils produisent.

d. Contrairement au français, le caractère défini ou indéfini des noms arabes est distingué par deux moyens différents : un article marque le défini et un diacritique marque l'indéfini.

type	diacritique	nom	translittération	transcription
Diacritique simple	ـَ	فتحة *fatHaħ*	a	/a/
	ـُ	ضمّة *ðam∼ħ*	u	/u/
	ـِ	كسرة *kasraħ*	i	/i/
	ـْ	سكون *sukuwn*	.	pas de son
Diacritique double	ـً	تنوين فتح *tanwiyn fatH*	ã	/an/
	ـٌ	تنوين ضمّ *tanwiyn ðam∼*	ũ	/un/
	ـٍ	تنوين كسر *tanwiyn kasr*	ĩ	/in/
Chadda	ـّ	شدّة *šad∼aħ*	∼	doublement de la consonne
madda	آ	مدّة *mad∼aħ*	Ā	long /'/

Table 1.2.: Diacritiques arabes

Voyelles longues

Les voyelles longues sont au nombre de trois. Elles permettent de prononcer plus longuement la vocalisation utilisée. La voyelle longue est composée d'une voyelle courte ـَ *a*, ـُ *u* ou ـِ *i* suivie respectivement d'un support ا*A*, و *w* ou ى *y*. Le tableau 1.3 présente la liste des voyelles longues de l'arabe, leurs translittérations et leurs transcriptions.

voyelle longue	translittération	transcription
ـَا	aA	/ā/ long a
ـُو	uw	/ū/ long u
ـِي	iy	/ī/ long i

Table 1.3.: Voyelles longues arabes

Tandis que les lettres présentées dans la section 1.1.1 sont obligatoires dans l'écriture, les diacritiques sont optionnels[e]. En effet, les textes arabes peuvent

e. Hormis les textes religieux qui sont entièrement diacrités, les textes arabes sont généralement non-diacrités.

être non-diacrités, partiellement diacrités ou entièrement diacrités. L'absence des diacritiques dans les écrits arabes pose des problèmes d'ambiguïtés pour le lecteur. La proportion des mots ambigus dans le lexique arabe (qui possèdent plus qu'une diacritisation potentielle) est estimée à 90.5% (Debili et Achour, 1998).

Les diacritiques jouent un rôle important dans la morphologie et la syntaxe de l'arabe (Hamdi, 2012). En morphologie, par exemple, la voix d'un verbe arabe est parfois rendue par des diacritiques. Les verbes كَتَب *katab* "*il a écrit*" et كُتِب *kutib* "*il a été écrit*" représentent respectivement la voix active et passive du verbe كتب *ktb* "*écrire*" et possèdent la même forme non-diacritée. En syntaxe, d'autre part, les diacritiques peuvent déterminer la fonction syntaxique d'un mot. Prenons comme exemple la phrase extraite du coran يخشى اللّه العلماء [f] *yxšý All∼h AlʕlmA'*. Au niveau syntaxique, cette phrase non-diacritée est ambigüe, elle possède deux interprétations différentes :

يخشى اللّهَ العلماءُ
*yxšý All∼h**a** Alʕlm A'**u***
Les savants craignent Dieu

يخشى اللّهُ العلماءَ
*yxšý All∼h**u** Alʕlm A'**a***
Dieu craint les savants

Bien que l'ordre des mots dans ces deux interprétations soit le même, la phrase peut être lue de deux façons différentes. Seules les voyelles (mises en gras dans l'exemple) situées à la fin des mots اللّه *All h* et العلماء *Alʕlm A'* permettent de distinguer le sujet de l'objet dans cette phrase [g]. Ces voyelles sont appelées voyelles casuelles. D'autres voyelles situées en début et en milieu de mot sont dites voyelles lexicales, leur apport se situe au niveau morphologique et lexical.

La diacritisation, dite aussi voyellation, est l'opération qui consiste à placer automatiquement des diacritiques dans un mot arabe n'en contenant pas. Le nombre de diacritiques d'un mot est égal au nombre de ses consonnes. Plusieurs travaux tels que (Vergyri et Kirchhoff, 2004), (Nelken et Shieber, 2005) et (Zitouni *et al.*, 2006) ont proposé des systèmes de diacritisation automatique des textes arabes. Cette opération est étroitement liée à la désambiguïsation morphosyntaxique. En effet, Hamdi (2012) a montré que, en restituant les diacritiques, les performances d'un analyseur morphosyntaxique passent de 84.91% à 95.59%. Nous revenons sur les différents processus de traitement morphosyntaxique de l'arabe dans le chapitre suivant.

f. cette phrase est extraite du verset coranique وَيَخشَى اللَّهَ مِن عِبَادِهِ العُلَمَاءُ *wayaxšaý All∼ha min ʕibAdihi AlʕulamA'u*

g. L'interprétation correcte de cette phrase présentée dans le coran est bien entendu la première.

Pour mieux illustrer l'ambiguïté liée à l'absence des diacritiques dans les textes arabes et leur impact sur les applications de traitement automatique, nous donnons l'exemple de traduction de la phrase بكين من الحزن *bkyn mn Al.hzn* "*elles ont pleuré de tristesse*". L'absence des diacritiques dans cette phrase a conduit à la fausse traduction proposée par Google "*Beijing de tristesse*". En effet, la forme non-diacritée du mot بكين *bkyn* est ambigüe, elle peut correspondre aux mots بِكِينْ *bikiyn* "*Beijing*" ou بَكَيْنَ *bakayna* "*elles ont pleuré*".

1.2. Phénomènes morphologiques de la langue arabe

La morphologie est l'étude de la structure interne des mots (Habash, 2010). Elle s'intéresse à la décomposition d'un mot en plusieurs unités. Ces unités sont appelées morphèmes (ou encore unités morphologiques) qui sont les plus petites unités porteuses du sens.

La langue arabe dispose de trois moyens pour déterminer les unités morphologiques d'un mot. Ces moyens sont l'agglutination, la flexion et la dérivation.

1.2.1. Morphologie agglutinante

L'agglutination concerne essentiellement le rattachement des clitiques aux mots dans un ordre bien précis. Les clitiques sont des morphèmes qui possèdent les mêmes propriétés que les affixes (cf. section 1.2.2) mais ils peuvent être réalisés comme des éléments autonomes puisqu'ils possèdent des fonctions syntaxiques indépendantes. Ils sont invariants, optionnels et ne changent pas la signification de base du mot auquel ils se rattachent. Des multiples clitiques peuvent apparaitre dans un même mot. Le mot qui contient des clitiques est nommé un mot agglutiné alors que le mot qui n'a aucun clitique est appelé mot simple. Le mot agglutiné وسيطعمونكم[h] *wasayuTˁimuwnakum* "*et ils vous nourriront*", par exemple est composé du mot simple يطعمون *yuTˁimuwna* "*ils nourrissent*" augmenté par les clitiques *wa* "*et*", *sa* "*particule de futur*" et *kum* "*vous*".

On distingue deux types de clitiques : des proclitiques qui se rattachent au début du mot et des enclitiques situés à la fin de ce dernier. Les proclitiques sont regroupés en plusieurs classes selon leurs fonctions grammaticales. Dans les tableaux 1.4, nous donnons la liste des proclitiques présentés selon leurs positions, du plus éloigné au plus proche du mot.

h. À l'instar de la majorité des écrits arabes, tous les mots arabes donnés dans ce manuscrit sont non-diacrités. Par contre, nous remettons les diacritiques dans la translittération.

classe	catégorie	proclitique	glossaire
QST	particule interrogative	+أ Â+	*est-ce que*
CNJ	conjonction	+و wa+	*et*
		+ف fa+	*puis, alors*
PRP	préposition	+بـ bi+	*par, avec*
		+كـ ka+	*comme*
		+لـ li+	*pour, à*
PRT	particule de futur	+سـ sa+	"*particule de futur*"
	particule de négation	+لا lA+	*ne ... pas*
		+ما mA+	
DET	déterminant	+الـ Al+	*le, la, les*

Table 1.4.: Liste de proclitiques

Le mot arabe connait un seul enclitique, le pronom, qui peut être soit un complément d'objet pour le cas des verbes et des particules, soit un pronom possessif pour les noms. L'enclitique varie en genre et en nombre. Nous présentons les différents enclitiques arabes selon la personne, le genre et le nombre dans le tableau 1.5.

Les clitiques ne sont pas toujours compatibles avec un mot donné, leur compatibilité dépend de la catégorie grammaticale du mot. Prenons comme exemple trois catégories grammaticales nom, verbe et particule, leurs structures peuvent être décrites respectivement par les expressions régulières suivantes :

QST ? CNJ ? PRP ? nom POSS ?

QST ? CNJ ? (PRP ?|PRT ?) verbe OBJ ?

QST ? CNJ ? particule OBJ ?

Le symbole '?' dans ces expressions indique que les clitiques sont optionnels contrairement à la forme simple qui constitue le cœur d'un mot arabe.

L'agglutination pose des problèmes d'ambiguïté pour le lecteur arabophone et par conséquent pour les processus de traitement automatique de la morphologie arabe. En effet, dans certains cas, plusieurs lectures sont possibles, comme dans le cas du verbe وعد *wςd* qui peut être reconnu comme forme simple *wςd* "*promettre*" ou bien composée en *w*+*ςd* "*et compter*". À l'instar de وعد *wςd*, les mots arabes qui commencent ou se terminent par des lettres qui peuvent représenter des clitiques sont potentiellement ambigus. À titre d'exemple, la forme المهم

personne	genre	nombre	enclitique
1	masculin	singulier	ـي+ +iy / ـني+ +niy
	féminin	pluriel	ـنا+ +nA
2	masculin	singulier	ـك+ +ka
		duel	ـكما+ +kumA
		pluriel	ـكم+ +kum
	féminin	singulier	ـك+ +ki
		duel	ـكما+ +kumA
		pluriel	ـكنّ+ +kunna
3	masculin	singulier	ـه+ +hu
		duel	ـهما+ +humA
		pluriel	ـهم+ +hum
	féminin	singulier	ـها+ +hA
		duel	ـهما+ +humA
		pluriel	ـهنّ+ +hunna

Table 1.5.: Liste d'enclitiques

Almhm commence par ال *Al* qui peut être l'article défini. Elle se termine par هم *hm* qui peut être également un clitique *leur*. Par conséquent, deux découpages sont possibles pour cette forme :

– ال+ مهم *Al+ mhm* "*l'important*"

– هم+ الم *Alm +hm* "*leur douleur*"

Le découpage هم+ م +ال *Al+ m +hm* n'est pas possible car le morphème م *m* n'existe pas dans le lexique de l'arabe.

Le mot والي *wAly* qui peut être reconnu comme un mot simple "*gouverneur*" commence et se termine par des lettres qui peuvent être des clitiques. La lettre و *w* représente la conjonction "*et*" alors que ي est le pronom possessif "*mon*". Ainsi, d'autres découpages possibles se rajoutent à la forme simple والي *wAly* "*gouverneur*" :

– و+ الي *w+ Aly* "*et automatique*"

– و+ ال +ي *w+ Al +y* "*et mon clan*" autres que la forme simple

L'ambiguïté est plus importante lorsque les diacritiques ne sont pas représentés. L'absence de diacritiques et l'agglutination présentent deux problèmes mutuels. En effet, la restitution des diacritiques permet de réduire le nombre de

découpages possibles d'un mot. De la même façon, le découpage d'un mot peut contribuer à lever l'ambiguïté vocalique du mot.

1.2.2. Morphologie flexionnelle

La flexion d'un mot repose sur la concaténation d'affixes à un radical pour construire une forme fléchie. Le radical, obligatoire dans le mot, porte le sens de base du mot. Les affixes possèdent trois types : les préfixes, qui se situent avant le radical, les suffixes, qui se situent après le radical et les circonfixes qui l'entourent. La détermination des affixes repose sur les valeurs des traits morphologiques. Au niveau flexionnel, seules les particules restent invariables et ne possèdent pas de formes fléchies. Elles possèdent, en revanche, des formes agglutinées comme nous l'avons évoqué dans la section 1.2.1.

Traits morphologiques du verbe

Les traits morphologiques associés à un verbe sont :

– l'aspect : l'arabe distingue trois aspects différents. **L'accompli** (الماضي *Al-mADy*) dit aussi le perfectif, utilisé quand l'action est accomplie. C'est l'aspect le plus simple d'un point de vue morphologique. Utilisé avec la troisième personne du singulier, il représente la forme canonique d'un verbe, à l'instar de l'infinitif en français. **L'inaccompli** (المضارع *AlmDAr*ς) appelé aussi l'imperfectif indique que l'action est en train de se réaliser, sans être achevée. Il exprime le présent, et permet d'exprimer le passé et le futur à l'aide de particules. **L'impératif** (الأمر *AlÂmr*) indique l'injonction. Il ne peut être conjugué qu'à la deuxième personne.

– le mode : trois modes sont définis en arabe. **L'indicatif** (المرفوع *Almrfw*ς) employé dans une proposition principale. **Le subjonctif** (المنصوب *Almn*S*wb*) employé dans une proposition subordonnée. **L'apocopé** (المجزوم *Almjzwm*) dit aussi le jussif exprime la négation, l'interdiction ou le conditionnel. Le mode s'applique uniquement à l'aspect imperfectif.

– la personne, le genre et le nombre du sujet : comme le français, l'arabe distingue trois personnes et deux genres, **le masculin** (المذكّر *Alm*ð*k*∼*r*) et **le féminin** (المؤنّث *Almwn*∼θ). En revanche, l'arabe distingue trois valeurs pour le nombre **le singulier** (المفرد *Almfrd*), **le duel** (المثنى *Almθný*) et **le pluriel** (الجمع *Aljm*ς).

Le tableau 1.6 donne la liste des différents affixes des verbes dans l'aspect accompli selon les valeurs des traits morphologiques : personne, nombre et genre. Nous illustrons la flexion verbale sur le verbe *ÂaTçam* "*nourrir*".

personne	nombre	genre	affixe	Exemple [ÂaTςam]
1	singulier	-	+tu	ÂaTςam**tu**
	pluriel	-	+nA	ÂaTςam**nA**
2	singulier	masculin	+ta	ÂaTςam**ta**
		féminin	+ti	ÂaTςam**ti**
	duel	masculin	+tumA	ÂaTςam**tumA**
		féminin	+tumA	ÂaTςam**tumA**
	pluriel	masculin	+tum	ÂaTςam**tum**
		féminin	+tunna	ÂaTςam**tunna**
3	singulier	masculin	+a	ÂaTςam**a**
		féminin	+at	ÂaTςam**at**
	duel	masculin	+A	ÂaTςam**A**
		féminin	+tA	ÂaTςam**tA**
	pluriel	masculin	+uwA	ÂaTςam**uwA**
		féminin	+na	ÂaTςam**na**

Table 1.6.: Affixes des verbes pour l'aspect accompli

Le signe '+' utilisé dans le tableau 1.6 indique la position du radical par rapport l'affixe. Avec l'aspect accompli, uniquement des suffixes sont utilisés, contrairement à l'aspect inaccompli et impératif où des circonfixes sont possibles. La forme fléchie du verbe أطعم *ÂaTςam* dans le mode indicatif de l'aspect inaccompli avec la troisième personne du masculin pluriel est يطعمون *yTςmwn* "*ils nourrissent*".

Traits morphologiques du nom

À l'image de la flexion verbale, la flexion nominale se base sur la détermination des affixes liés au radical selon les valeurs des traits morphologiques suivants :

- l'état : un nom peut être **défini** (معرّف *mςr∼f*) à l'aide d'un article ou d'une construction possessive ou **indéfini** (نكرة *nkrħ*). Comme nous l'avons indiqué dans la section 1.1.2, l'état indéfini est marqué par un diacritique double.
- le cas : ce trait est fondamental pour distinguer la fonction syntaxique du nom. On dénombre trois valeurs différentes pour ce trait. **L'accusatif** (منصوب *mnSwb*), **le nominatif** (مرفوع *mrfwς*) et **le génitif** (مجرور *mjrwr*).
- le genre et le nombre : à l'image des verbes, les noms arabes possèdent deux genres et trois nombres.

Le tableau 1.7 présente les différents affixes des noms indéfinis selon les valeurs des traits morphologiques avec l'exemple du nom مطعم *muTςim* "*nourris-*

seur". Nous donnons entre parenthèses les affixes des noms définis.

genre	nombre	cas	affixe	exemple $muTʕim$
msaculin	singulier	nominatif	+un (+u)	muTʕim**un**
		accusatif	+an (+a)	muTʕim**an**
		génitif	+in (+i)	muTʕim**in**
	duel	nominatif	+Ani (+Ani)	muTʕim**Ani**
		accusatif	+Ani (+Ani)	muTʕim**Ani**
		génitif	+Ani (+Ani)	muTʕim**Ani**
	pluriel	nominatif	+uwna (+uwna)	muTʕim**uwna**
		accusatif	+iyna (+iyna)	muTʕim**iyna**
		génitif	+iyna (+iyna)	muTʕim**iyna**
féminin	singulier	nominatif	+ħun (+ħu)	muTʕimħ**un**
		accusatif	+ħan (+ħa)	muTʕimħ**an**
		génitif	+ħin (+ħi)	muTʕimħ**in**
	duel	nominatif	+atAni (+atAni)	muTʕim**atAni**
		accusatif	+atAni (+atAni)	muTʕim**atAni**
		génitif	+atAni (+atAni)	muTʕim**atAni**
	pluriel	nominatif	+Atun (+Atu)	muTʕim**Atun**
		accusatif	+Atan (+Atan)	muTʕim**Atan**
		génitif	+Atin (+Atin)	muTʕim**Atin**

Table 1.7.: Affixes nominaux de l'arabe

Contrairement à la flexion verbale qui est régulière, la flexion nominale ne l'est pas toujours. Le suffixe ـة +ħ par exemple, qui marque souvent le genre féminin peut apparaître dans des noms masculins tel que خليفة *xaliyfaħ* "*calife*". D'autre part, le passage du singulier au pluriel peut être effectué à l'aide de schèmes au lieu d'affixes. Nous revenons sur cette irrégularité dans la section 1.2.3.

À l'instar de l'agglutination, la flexion provoque de l'ambiguïté essentiellement quand le radical commence ou se termine par un affixe potentiel. Le mot تراجع *trAjʕ*, par exemple est ambigu car sa première lettre ـت *t* peut être réalisée comme un affixe ou faire partie du radical. Dans le premier cas, le verbe est conjugué à l'accompli actif avec la troisième personne du masculin singulier qui n'est autre que le radical تراجع *trAjʕ* "*il a diminué*". Dans la deuxième situation, le verbe تراجع *trAjʕ* "*elle/tu révise(s)*" représente la concaténation du préfixe ـت *t* au radical راجع *rAjʕ* "*réviser*". L'affixe ـت *t* est utilisé pour la troisième personne du féminin singulier et la deuxième personne du masculin singulier.

1.2.3. Morphologie dérivationnelle

Le processus de dérivation est basée sur la combinaison d'une racine (جذر *jðr*) et d'un schème (وزن *wzn*) pour former un radical. La racine est une séquence de trois, quatre ou cinq lettres qui définit une notion abstraite. La racine ك ت ب *k t b*, par exemple, est associée à la notion d'écriture alors que la racine د ر س *d r s* et liée à la notion d'étude. Le schème, appelé aussi gabarit ou patron, définit le format du radical. Un schème peut être représentée par une séquence composée de chiffres et de lettres tel que 1A2a3, ma12a3[i]. Le processus de dérivation (cf. figure 1.1) consiste à remplacer chaque chiffre du schème par les lettres de la racine dans l'ordre. Reprenons l'exemple de la racine ك ت ب *k t b*, en remplaçant les chiffres 1,2 et 3 des schèmes 1A2a3 et ma12a3 par les lettres correspondantes de la racine, donne naissance aux mots كاتب *kAtab* "*correspondre avec*" et مكتب *maktab* "*bureau*" respectivement.

racine	k		t		b
schème	1	A	2	a	3
radical	k	A	t	a	b

racine			k	t		b
schème	m	a	1	2	a	3
radical	m	a	k	t	a	b

Figure 1.1.: Processus de dérivation dans l'arabe

Un schème est porteur d'un sens général, tel que le factitif, le nom prototypique de la personne qui effectue l'action, le résultat de l'action... Le sens d'un mot dérivé d'une racine et un schème est généralement la combinaison de la notion définie par la racine et le sens véhiculé par le schème.

Les schèmes verbaux marquent l'aspect et la voix (on distingue l'actif et le passif sans agent). Le schème prend des formes différentes selon les valeurs de l'aspect et de la voix du verbe. L'arabe définit dix schèmes (I[j], II, X) pour les verbes trilitères[k] et deux schèmes (QI, QII) pour les verbes quadrilitères (Ha-

i. Il existe d'autres manières pour représenter les schèmes dans la littérature. La lettre C peut être utilisées à la place des chiffres pour indiquer la position des lettres de la racine. À l'aide de cette représentation, nos exemples deviennent CACaC et maCCaC respectivement.

j. Pour représenter les schèmes verbaux, les linguistes ont eu recours aux chiffres romains au lieu d'écrire explicitement la forme de schème. Le schème 1a2a3, par exemple, est représentée par I alors que 1a22a3 est représenté par II... Nous présentons cette correspondance dans le tableau 1.8

k. les verbes trilitères sont les verbes dont la racine est composés de trois lettres alors que les

bash, 2010).

Le tableau 1.8 présente les schèmes des verbes arabes pour l'aspect accompli et l'inaccompli ainsi que leurs significations. Nous indiquons entre parenthèse les schèmes de la voix passive. Nous donnons également les verbes trilitères résultant de la combinaison des schèmes avec la racine ب ت ك *k t b* et les verbes quadrilatères résultant de la combinaison des schèmes avec la racine ب ع ث ر *b ς θ r*.

	accompli	inaccompli	signification	verbe
I	1a2a3 (1u2i3)	a12a3 (u12a3)	sens de base	katab *écrire*
II	1a22a3 (1u22i3)	u1a22i3 (u1a22a3)	intensification	kattab *faire écrire*
III	1A2a3 (1uw2i3)	u1A2i3 (u1A2a3)	interaction	kAtab *correspondre avec*
IV	Âa12a3 (Âu12i3)	u12i3 (u12a3)	causalité	Âktab
V	ta1a22a3 (tu1u22i3)	ata1a22a3 (uta1a22a3)	forme réflexive de II	takattab
VI	ta1A2a3 (tu1uw2i3)	ata1A2a3 (uta1A2a3)	forme réflexive de III	takAtab
VII	Ain1a2a3 (Âun1u2i3)	an1a2i3 (un1a2a3)	forme passive de I	Ainkatab
VIII	Ai1ta2a3 (Âu1tu2i3)	a1ta2i3 (u1ta2a3)	exagération	Aiktatab
IX	Ai12a33 (Âu12u33)	a12a33 (u12a33)	transformation	Aiktabb
X	Aista12a3 (Âustu12i3)	asta12a3 (usta12a3)	exigence	Aistaktab
QI	1a23a4 (1u23i4)	u1a23i4 (u1a23a4)	sens de base	baς*θ*ar
QII	ta1a23a4 (tu1u23i4)	ata1a23a4 (uta1a23a4)	forme réflexive de Q	tabaς*θ*ar

Table 1.8.: schèmes verbaux arabes

Comme les schèmes verbaux, les schèmes nominaux véhiculent un sens général lié à l'action, tel que le nom de la personne qui effectue l'action (participe actif), la personne qui subit l'action (participe passif) ou le nom du lieu où l'action est réalisée ... La racine ب ت ك *k t b*, dans cet ordre, peut se croiser avec divers schèmes. Les différents mots générés suite à la combinaison d'une racine

verbes quadrilatères possèdent quatre lettres dans la racine. Dans certains ouvrages, les termes tri-consonantiques et quadri-consonantiques sont utilisés

avec des différents schèmes constituent une famille sémantique. Le tableau 1.9 présente quelques noms dérivés de la racine ك ت ب *k t b*. Comme nous l'avons évoqué précédemment, dans certains cas le passage du singulier au pluriel ne repose pas sur les affixes mais sur les schèmes. Le pluriel bâti sur un schème est appelé pluriel brisé (جمع التكسير *jmς Alt~ksyr*).

schème	signification	nom	glose
1A2i3	participe actif	kAtib	écrivain
ma12uw3	participe actif	ma12uw3	écrit
1i2A3aħ	forme infinitive	kitAbaħ	écriture
ma12a3	nom du lieu	maktab	bureau
ma12a3aħ	nom du lieu	maktabaħ	bibliothèque

Table 1.9.: schèmes nominaux arabes

Le processus de dérivation de l'arabe n'est pas systématique. C'est à dire qu'une racine ne s'applique pas à tous les schèmes. En effet, étant donné que l'ensemble de racines arabes est composé par n racines r_1, r_2, $r_3 \ldots$ r_n et l'ensemble de schèmes est composés de m schèmes sch_1, sch_2, $sch_3 \ldots$ sch_n, les diverses combinaisons entre les deux ensembles définissent la totalité du lexique potentiel de l'arabe dont le lexique réel ne constitue qu'une partie.

L'ambiguïté produite par la dérivation est parfois produite par l'ambiguïté engendrée apr la flexion. L'ambiguïté flexionnelle du mot تراجع *trAjς* conduit à une ambiguïté dérivationnelle. La forme qui correspond à la glose "*elle révise*" se dérive de la racine ع ج ر *r j ‘* et le schème 1A2a3. Alors que l'autre forme correspond au croisement de cette même racine avec le schème ta1A2a3. De plus, l'ambiguïté peut provenir des lettres qui peuvent appartenir potentiellement à la racine ou au schème. À titre d'exemple, le mot استرق *Astrq* possède deux dérivations différentes qui dépendent de la lettre س *s*. Cette lettre peut faire partie de la racine ou du schème :

– racine *s r q* et schème *Ai1ta2a3* "*espionner*"
– racine *r q q* et schème *Aista12a3* "*asservir*"

1.3. Catégories grammaticales

Dans la grammaire classique de l'arabe, on distingue trois catégories grammaticales qui sont la particule (حرف *Hrf*), le verbe (فعل *fςl*) et le nom (إسم *Ăsm*) (Al-Dahdah, 1996; Al-Ghulayaini, 2006). Ce jeu de catégories, bien que rudimentaire, regroupe entièrement toutes les catégories qui partagent les mêmes

propriétés. L'adjectif par exemple est considéré comme nom dans cette classification. En effet, l'adjectif possède les mêmes traits morphologiques que le nom tel que l'état. Dans l'expression الولد الصغير *Alwld AlSγyr* "*le petit enfant*", l'adjectif الصغير *AlSγyr* est défini "*le petit*" à l'aide de l'article ال *Al*. Nous nous focalisons dans cette section sur la description de ces trois catégories.

1.3.1. Particules

Les particules sont des mots (parfois des clitiques) qui n'ont pas de sens autonome (Al-Ghulayaini, 2006). Elles ne possèdent pas de formes fléchies et sont en nombre limité. La classification des particules arabes est une tâche complexe. En effet, il n'existe pas une classification commune à tous les grammairiens arabisants. Dans ce mémoire, nous citons les principales catégories proposées par (Al-Dahdah, 1996) :

- adverbes : certains adverbes (pas tous) sont considérés comme des particules tels que فقط *faqaT* "*seulement*", أيضا *ÂayðA* "*aussi*", أبدا *ÂabadA* "*jamais*"...
- conjonctions : و *wa* *et*, فـ *fa* "*alors*", ثمّ *θum∼* "*puis*"...
- prépositions : من *min* "*de*", إلى *Ăilaý* "*à*", عن ς*an* "*à propos*", على ς*alaý* "*sur*", في *fiy* "*dans*"...
- particules de conditions : لو *law* "*si*", إن *Ăin* "*si*"...
- particules d'interrogation : هل *hal* "*est-ce que*", مَا *mA* "*qu'est ce que*"...
- particules de négation : ما *mA*, لا *lA* "*ne .. pas*".
- particules de futur : سوف *sawfa*.
- proclitiques : بـ *bi* "*avec*", كـ *ka* "*comme*", سـ *sa*...

1.3.2. Verbes

Le système verbal de l'arabe est à la fois simple et complexe : il est simple dans le sens où sa flexion est régulière (Larcher, 2012). Il est complexe du fait que les variations entre ses groupes sont multiples, ce qui rend sa classification difficile. Les verbes arabes peuvent être regroupés selon la racine ou selon le schème.

Classification des verbes selon la racine

Les verbes peuvent être classés selon le nombre de lettres dans la racine, on distingue :

(i) les verbes trilitères (الأفعال الثلاثيّة *AlÂfςAl AlθlAθyħ*) qui possèdent trois lettres dans la racine tels que كتب *ktb* "*écrire*" et درس *drs* "*étudier*".

(ii) les verbes quadrilitères (الأفعال الرباعيّة *AlÂfçAl AlrbAçyħ*) qui contiennent quatre lettres dans leurs racines, à l'instar des verbes دحرج *dHrj* "*rouler*" et عرقل ç*rql* "*entraver*".

En outre, les verbes arabes peuvent être regroupés selon la nature des lettres de leurs racines. On distingue les verbes sains (الأفعال الصحيحة *AlÂfçAl AlSHyHħ*) et les verbes malsains (الأفعال المعتلّة *AlÂfçAl Almçtl∼ħ*). Ces derniers contiennent une lettre défective (و *w* ou ي *y*) dans leurs racines.

(i) verbes sains : ces verbes peuvent être "hamzés", redoublés ou parfaitement sains. **Le verbe parfaitement sain** n'a aucune particularité, il suit la dérivation et la flexion régulières utilisées dans l'arabe. **Le verbe "hamzé"** (الفعل المهموز *Alfçl Almhmwz*) est un verbe qui contient la lettre *hamza* dans sa racine. Cette lettre peut figurer dans la première, la deuxième ou la troisième position de la racine tels que أخذ *Âxð* "*prendre*", سأل *saÂal* "*questionner*", بدأ *badaÂ* "*commencer*" respectivement. Leur particularité réside dans la forme de la *hamza* qui dépend de la position et les voyelles voisines. **Le verbe redoublé** (الفعل المضاعف *Alfçl AlmDAçf*) se caractérise par une racine dont la deuxième et la troisième lettres sont identiques, à l'instar de ردّ *rad∼* "*rendre*" et ملّ *mal∼* "*s'ennuyer*". Sa spécificité provient de sa dérivation qui consiste à éliminer la deuxième voyelle du schème à partir duquel le verbe se dérive.

(ii) verbes malsains : ces verbes contiennent une lettre défective dans leur racine. On distingue trois groupes de verbes malsains selon la position de la lettre défective dans la racine. Cette lettre est située en première position de la racine pour *le verbe assimilé* (الفعل المثال *Alfçl Almθal*). La lettre défectueuse occupe respectivement la deuxième lettre de la racine du **verbe creux** (الفعل الأجوف *Alfçl AlÂjwf*) et la troisième lettre du verbe défectueux (الفعل الناقص *Alfçl AlnAqS*).

Morphologiquement, les verbes malsains se distinguent des verbes sains au niveau de la dérivation. En effet, leurs schèmes subissent des transformations radicales selon la valeur de l'aspect auquel le verbe est conjugué. Nous revenons en détail sur les transformations subies par les schèmes suivant le type du verbe dans l'annexe. Nous donnons quelques exemples de verbes malsains dans le tableau 1.10.

position de la lettre défectueuse	type	racine	verbe
1	مثال *miθAl* assimilé	و ص ل w S l	وصل *wSl* arriver
2	أجوف *Âjwaf* creux	ق و ل q w l	قال *qAl* dire
3	ناقص *nAqiS* défectueux	م ش ي m š y	مشى *mašaý* marcher

Table 1.10.: Exemples de verbes malsains

Deux autres classes se rajoutent à ces six catégories, les verbes possédant deux lettres défectueuses dans les lettres de la racine. Le verbe lié مقرون *maqruwn* comptant deux gildes successifs et le verbe séparé مفروق *mafruwq*. Un récapitulatif des catégories des verbes arabes qui dépendent de la racine est donné dans la figure 1.2.

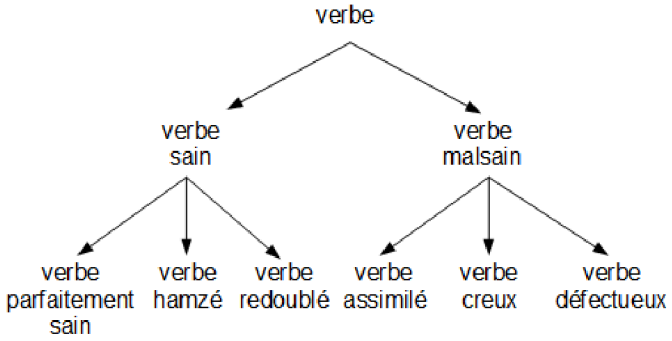

Figure 1.2.: classification des verbes selon la racine

Classification des verbes selon leur schème

Le système verbal distingue les verbes simples ou non-augmentés (الأفعال المجرّدة *AlÂfçAl Almjrdħ*) des verbes augmentés (الأفعال المزيدة *AlÂfçAl Almzydħ*) :

(i) verbes simples : ce sont les verbes qui suivent le schème trilitère I et le schème quadrilitère Q. Les trilitères possèdent trois formes différentes 1a2a3, 1a2u3 et 1a2i3. Les verbes 1a2a3 sont des verbes d'actions, les verbes

1a2u3 sont des verbes d'états et les verbes 1a2i3 sont l'un ou l'autre (Larcher, 2012). Quant aux quadrilitères, ils possèdent une seule forme 1a23a4.

(ii) verbes augmentés : ils sont au nombre de neuf pour les schèmes trilitères. Les schèmes possédant quatre radicaux ont une seule forme augmentée.

Le tableau 1.11 résume la classification des verbes arabes selon leurs schèmes :

non-augmenté	trilitère	I	1a2a3	كتب katab	écrire
		I	1a2u3	قرب qarub	se rapprocher
		I	1a2i3	ركب rakib	monter
	quadrilitère	Q	1a23a4	عرقل ʕrql	entraver
augmenté	trilitère	II	1a2 a3	نقّح nq∼H	réviser
		III	1A2a3	ساعد sAʕd	aider
		IV	Â12a3	أسرع Âsraʕ	se dépêcher
		V	ta1a2 a3	تعلّم taʕal∼am	apprendre
		VI	ta1A2a3	تراجع tarAjaʕ	décroître
		VII	Ain1a2a3	انتقل Aintaqal	se déplacer
		VIII	Ai1ta2a3	استمع Aistamaʕ	écouter
		IX	Ai12a3	احمرّ AiHmar	se rougir
		X	Aista12a3	استقبل Aistaqbal	acceuillir
	quadrilitère	IQ	ta1a23a4	تبعثر tabaʕθar	s'éparpiller

Table 1.11.: classification des verbes selon leur schème

1.3.3. Noms

Les noms arabes sont regroupés en deux catégories principales : les noms solides (الأسماء الجامدة *AlÂsmA' AljAmdħ*) qui échappent généralement à toute dérivation et les noms déverbaux (الأسماء المشتقّة *AlÂsmA' Almštq∼ħ*) qui dérivent d'une racine verbale (Al-Dahdah, 1996; Al-Ghulayaini, 2006).

Noms solides

La morphologie nominale arabe classe les noms solides en plusieurs sous-catégories, parmi lesquelles les pronoms, les nombres, les noms interrogatifs, les adverbes, les noms propres et les noms communs.

– noms interrogatifs : utilisés dans les phrases interrogatives tels que كيف *kyf* "*comment*", متى *mtý* "*quand*", أين *Âyn* "*où*"...

– noms numératifs et quantitatifs : les numératifs concernent les nombres simples tels que les unités de zéro (صفر *Sfr*) à neuf (تسعة *tsʕħ*), les dizaines de dix (عشرة ʕšrħ) à quatre-vingt-dix (تسعون *tsʕwn*), les centaines etc, et les nombres composés comme واحد وعشرون *wAHd wʕšrwn* "vingt-et-un" et واحد وثلاثون *wAHd wθlAθwn* "trente-et-un"... les quantitatifs comme كلّ *kl∼* "*tout*", كلّ *kl∼* "*tout*"...
– pronoms (ضمائر *DamAŷr*) : ce sont les **pronoms personnels** (ضمائر منفصلة *DamAŷr mnfSlħ*), à titre d'exemple pour la première personne il y a أنا *ÂnA* "*je*" et نحن *nHn* "*nous*", les **pronoms démonstratifs** (أسماء الإشارة *ÂsmA' AlĂšArħ*) tels que هذا *hðA* "*ce*", هؤلاء *hŵlA'* "*ces*"... Les pronoms relatifs comme الّذي *Al∼ðy* "*celui*" et الّتي *Al∼ty* "*celle*"... Cette sous-catégorie regroupe également les clitiques pronoms d'objets pour les verbes et les pronoms possessifs pour les noms.
– noms propres : c'est un nom qui réfère à une entité unique de personne comme محمّد *Mohammed*, de lieux tels que باريس *Paris*, etc.
– noms communs : ce sont des noms employés pour désigner tous les éléments d'un même ensemble, il dispose d'une définition et d'une signification (Goosse et Grevisse, 1993). Parmi ceux-ci il y a des noms bilitères فم *fm* "*bouche*", trilitères comme أنف *Ânf nez*, كلب *klb chien*, quadrilitères, etc.

Les trois premières sous-catégories sont non-dérivationnelles. En revanche, contrairement aux pronoms qui sont conjugables, les noms interrogatifs, numératifs et quantitatifs n'admettent pas de formes fléchies, morphologiquement ils sont similares aux particules. Les noms communs sont également non-dérivationnels mais conjugables. Leur flexion est généralement irrégulière et ne dépend pas des affixes. Des schèmes sont utilisés pour distinguer le singulier et le pluriel. Bien que les noms communs soient dérivés de schèmes arabes, ils sont considérés comme non-dérivationnels puisque les racines que l'on en déduit ne définissent pas une notion. Prenons comme exemple les noms أرنب *Ârnb*"*lapin*" et بلاد *blAd* "*pays*" qui possèdent respectivement les schèmes 'a12a3 et 1i2A3. Les racines déduites de l'équation racine×schèmes = nom_commun, sont ر ن ب *r n b* et ب ل د *b l d*, elles n'existent pas dans le lexique arabe. Concernant les noms propres, hormis les noms de personnes, ils sont à la fois non-conjugables et non-dérivationnels. En effet, les noms de personnes sont dérivés d'une racine et d'un schème et admettent généralement une signification liée à la notion définie par la racine comme كريم *krym* "*généreux*" et سالم *sAlm* "*sain*". Cette particularité rend

le nom propre arabe ambigu, le mot حسن *Hsn* par exemple peut être un nom propre "*Hasan*", un adjectif "*bon*" ou un verbe "*bonifier*".

Déverbaux

Contrairement aux noms solides, les déverbaux connaissent une flexion et une dérivation régulières. Ils sont dérivés de verbes ; en effet, chaque verbe fournit neuf catégories de déverbaux (Al-Ghulayaini, 2006). Chacun d'eux correspond à une relation sémantique entre le verbe et le déverbal (cf. table 1.12).

1	اسم الفاعل	participe actif
2	اسم المفعول	participe passif
3	مصدر	forme infinitive
4	اسم المكان	nom du lieu
5	اسم الزمان	nom du temps
6	اسم الآلة	nom de l'instrument
7	صفة مشبّهة	adjectif analogue
8	اسم التفضيل	adjectif comparatif
9	صيغة المبالغة	forme exagérée

Table 1.12.: Noms déverbaux arabes

Du verbe (1) نظر *naĎar* "*observer*" et (2) انتظر *AintaĎar* "*attendre*", plusieurs noms déverbaux peuvent être dérivés, tels que, ناظر *nAĎir* "*observateur*", منظور *manĎuwr* "*observé*" de (1) et منتظر *muntaĎir* "*la personne en attente*", منتظر *muntaĎar* "*attendu*" de (2). Ces déverbaux représentent respectivement les participes actifs et passifs de ces verbes. Ils sont dérivés de la même racine que le verbe et des différents schèmes nominaux qui dépendent des schèmes verbaux. La table 1.13 donne les schèmes de quelques déverbaux dérivés à partir des deux premiers schèmes verbaux.

schème verbal	déverbal	schème nominal MSA	exemple
1a2a3	1	1A2i3	kAtib "*écrivain*"
	2	ma12uw3	maktuwb "*écrit*"
	3	1i2A3aħ	kitAbaħ "*écriture*"
	4	ma12a3	maktab "*bureau*"
1a22a3	1	mu1a22i3	mudarris "*instituteur*"
	2	mu1a22a3	mudarras "*étudié*"
	3	1i2A3aħ	dirAsaħ "*étude*"
	4	ma12a3aħ	madrasaħ "*institut*"

Table 1.13.: Exemple de déverbaux dans l'arabe

La table 1.13 représente un échantillon d'une grande table des déverbaux qui définit pour chaque schème verbal tous les déverbaux qui dérivent d'eux. Généralement, un schème nominal unique est défini pour chaque déverbal. En revanche, dans quelques cas, tels que les noms des lieux et des instruments, on peut retrouver plusieurs schèmes nominaux.

1.4. Dialectes arabes : variations lexicales et morphologiques

Les dialectes arabes présentent plusieurs similarités linguistiques avec le MSA. En revanche, ils se diffèrent à un degré variable aux niveaux phonologique, lexical, morphologique et syntaxique.

1.4.1. Variations phonologiques

Le système phonologique varie entre le MSA et ses dialectes d'une part et entre les dialectes eux-mêmes d'autre part. Parmi ces variations on peut citer :

- les consonnes ث /θ/ et ذ /ð/ sont réalisées comme /t/ et /z/ respectivement en LEV et EGY. Par contre leurs prononciations originelles sont maintenus en MAG.
- la consonne ج /j/ est prononcée comme /g/ en EGY et /y/ en GLF.
- le consonne ق /q/ est réalisée comme /'/ en EGY et en LEV alors qu'en GLF, elle est prononcée /g/.
- la consonne de MSA ظ /Ď/ est réalisée comme /Z/ en EGY et en LEV.

1.4.2. Variations lexicales

Le nombre de variations lexicales entre l'arabe standard et ses dialectes est significatif. Le tableau 1.14 illustre ces variations avec des exemples en MSA,

MRC (dialecte marocain), TUN, EGY, LEV et IRQ (dialecte irakien) :

MSA	TUN	MRC	EGY	LEV	IRQ	glossaire
حسن *Hasan*	باهي *bAhy*	زوين *zwiyn*	كويّس *kway∼is*	منيح *mniyH*	خوش *xuwš*	*bien*
حذاء *HiðA'*	صبّاط *Sab∼AT*	صبّاط *Sab∼AT*	جزمة *jazmaħ*	صبّاط *Sab∼AT*	كندرة *kindraħ*	*chaussure*
مجنون *majnuwn*	مهبول *mahbuwl*	مهبول *mahbuwl*	عبية *ʕabiyT*	مجنون *majnuwn*	خبل *xibil*	*fou*
سيّارة *say∼Araħ*	كرهبة *karahbaħ*	طوموبيل *Tuwmuwbiyl*	عربيّة *ʕarabiy∼aħ*	سيّارة *say∼Araħ*	سيّارة *say∼Araħ*	*voiture*
طاولة *TAwilaħ*	ميدة *miydaħ*	ميدة *miydaħ*	طربيزة *Tarabiyzaħ*	طاولة *TAwlaħ*	ميز *miyz*	*table*

Table 1.14.: Exemples de variations lexicales entre le msa et ses dialectes

Deux traits principaux marquent les variations lexicales entre le MSA et les dialectes. Premièrement, les dialectes maintiennent parfois la même racine. Le changement, dans ce cas, est restreint sur le schème. Deuxièmement, les variations proviennent de l'emprunt de nouveaux termes à langues voisines.

1.4.3. Variations morphologiques

Les dialectes arabes s'inspirent beaucoup de l'arabe standard. En effet, on retrouve des phénomènes morphosyntaxiques partagés par toutes les variantes de l'arabe. Morphologiquement, les dialectes arabes sont moins riches que le MSA. Au niveau flexionnel, certains traits morphologiques utilisés dans l'arabe standard n'existent pas en arabe dialectal tels que le cas et le mode. D'autre part, les dialectes ne distinguent pas le duel et le pluriel ni le genre féminin et masculin au pluriel des formes verbales. Quant à la dérivation, ils utilisent des schèmes qui n'existent pas dans le MSA tel que 1i2i3. En effet, le verbe *katab* du MSA devient *kitib* en EGY et LEV. Au niveau de l'agglutination, une particule progressive qui n'a pas d'analogue en MSA est utilisée dans certains dialectes. Elle apparait comme +بـ *b*+ en EGY et en LEV, comme +د *d*+ en IRQ et +ك *k*+ en MRC. Quant à la particule de futur +سـ *s*+, elle devient +حـ *H*+ en EGY et LEV et apparait comme +غـ γ en MRC. Au niveau des noms, les dialectes arabes utilisent un clitique démonstratif +هـ *h*+ employé toujours strictement avant l'article défini +ال *Al*+. Les dialectes utilisent aussi les proclitiques +عـ ʕ+ et +مـ *m*+ qui représentent des formes réduites des prépositions على *ʕalaý* "*sur*" et من *min* "*de*" respectivement.

1.4.4. Variations syntaxiques

Les différences syntaxiques entre les dialectes arabes et le MSA sont généralement réparties : les mêmes phénomènes existent dans des conditions différentes à la fois en MSA et dans les dialectes. Au niveau de l'ordre des mots de la phrase, on retrouve les deux modèles SVO et VSO dans les deux variantes avec relativement plus de fréquence de l'ordre VSO en MSA. L'ordre de mots en MSA possède plus de flexibilité par rapport aux dialectes puisqu'il utilise des marqueurs de cas qui sont absents dans les dialectes.

1.5. Dialecte tunisien

Le dialecte tunisien est le fruit de différentes cultures développées sur le territoire tunisien. Plusieurs linguistes ont étudié la situation linguistique en Tunisie, par exemple, Mejri *et al.* (2009) ont décrit les systèmes phonologiques, morphologiques et syntaxiques du tunisien. Ouerhani (2009) ont étudié les phénomènes d'interférence entre la morphologie verbale du tunisien et celle de l'arabe standard d'une part, et la relation entre les verbes tunisiens et français (le cas de l'emprunt) d'autre part.

La diversité des peuples qui ont vécu en Tunisie particulièrement les berbères, les arabes, les turcs et les français a marqué le système phonologique et lexical du dialecte tunisien. Au niveau phonologique, le tunisien présente trois phonèmes supplémentaires par rapport à l'arabe standard à savoir /p/, /v/ et /g/. Au niveau lexical, il existe des multiples mots d'origines différentes qui ont enrichi le lexique tunisien. Nous trouvons ainsi des mots d'origine berbère كرموس *karmuws* "*figue*", français سبيطار *sbiyTAr* "*hôpital*", turc بابور *bAbuwr* "*bateau*", italien بوسطة *buwsTaħ* "*bureau de poste*", maltais قطّوس *qaT∼uws* "*chat*", anglais كرهبة *karhbaħ* "*voiture*", espagnol صبّاط *Sab∼AT* "*chaussure*", arabe classique أكحل *ÂakHil* "*noir*"...

Bien que ces mots proviennent de plusieurs origines, ils suivent les règles morphologiques de l'arabe standard. En effet, la morphologie du TUN se base sur celle du MSA, on retrouve ainsi les phénomènes d'agglutination de flexion et de dérivation décrits dans la section 1.2 mais avec quelques différences que nous décrivons ci-dessous.

1.5.1. Agglutination

Au niveau de l'agglutination, deux phénomènes distinguent le tunisien du MSA. D'une part, certains clitiques MSA sont réalisés sous la forme de particules indépendantes en tunisien et vice-versa. La préposition +لـ *li+ pour* et le proclitique de futur ne sont plus rattachés aux verbes. Tous deux se traduisent par

la particule indépendante باش *bAš* qui se situe avant le verbe : les formes لتكتب *litaktuba* "*pour que tu écrives*" et ستكتب *sataktubu* "*tu écriras*" sont exprimées en tunisien par باش تكتب *bAš tiktib*. Inversement, des particules indépendantes en MSA telles que على *ʕalaý* "*sur*" et من *min* "*de*" sont réalisées respectivement comme des clitiques +عـ *ʕ*+ et +مـ *m*+ quand elles sont suivies par des noms définis par l'article ال *Al*. D'autre part, la forme de certains clitiques change. Le proclitique d'interrogation MSA +أ *Â*+ "*est-ce que*", par exemple, devient en tunisien l'enclitique +ش *š*+. La forme verbale MSA أكتبت *Âkatabta* "*est-ce que tu as écrit*" se traduit en tunisien par كتبتش *ktibtiš*.

1.5.2. Flexion

De manière générale, la flexion en TUN est plus pauvre que celle du MSA. Au niveau des verbes, le mode n'est plus marqué alors que le cas n'est plus distingué pour les noms. Les valeurs du nombre qui étaient trois en MSA (singulier, duel et pluriel) sont réduites à deux (singulier et pluriel). Quant au genre, il n'est spécifié que lorsqu'il s'agit de la troisième personne du singulier. La liste des affixes sujet des verbes tunisiens dans l'aspect accompli est donnée dans le tableau 1.15. Ce dernier peut être mis en regard du tableau 1.6.

personne	nombre	genre	affixe	Exemple [ktib]
1	singulier	-	+t	ktib**t**
	pluriel	-	+nA	ktib**nA**
2	singulier	masculin	+t	ktib**t**
		féminin	+ti	ktib**ti**
	pluriel	masculin	+tuwA	ktib**tuwA**
3	singulier	masculin	-	ktib
		féminin	+it	kitb**it**
	pluriel	masculin	+uwA	kitb**uwA**

Table 1.15.: Affixes des verbes tunisiens dans l'aspect accompli

D'autre part, contrairement au MSA qui marque la voix dans le schème verbal, le tunisien marque la voix passive sous la forme du préfixe تـ *t*[1]. La forme MSA passive كتب *kutiba* "*il est écrit*" devient en tunisien تكتب *tiktib*.

1. le passif dans le dialecte tunisien peut être aussi exprimé avec les schèmes, en ajoutant un /t/ au début de chaque schème de la voix active.

1.5.3. Dérivation

Hormis les emprunts, les radicaux tunisiens dérivent d'une racine arabe et d'un schème, comme pour le MSA. Il y a en général correspondance bi-univoque entre un schème MSA et un schème TUN sauf dans certains cas où un schème MSA peut correspondre à deux schèmes TUN ou bien à aucun schème TUN. Les schèmes TUN se caractérisent généralement par la chute de la voyelle affectée à la première lettre de la racine. On retrouve, en effet, les schèmes TUN 12A3, 12iy3 et 12a3 qui correspondent respectivement aux schèmes MSA 1i2A3, 1a2iy3 et 1a2a3. Un échantillon de la correspondance entre les schèmes MSA (cf. section 1) et les schèmes TUN est donné dans le tableau 1.16.

accompli		inaccompli	
schème_MSA	schème_TUN	schème_MSA	schème_TUN
1a2a3	12a3	a12a3	a12a3
1a22a3	1a22a3	u1a22i3	1a22a3
1A2a3	1A2a3	u1A2i3	1A2a3
ta1A2a3	t1A2a3	ata1A2a3	it1A2a3
1a23a4	1a23i4	u1a23i4	1a23i4
ta1a23a4	ta1a23i4	ata1a23i4	ta1a23i4

Table 1.16.: Correspondance des schèmes msa et tun

Concernant les mots empruntés, grâce à leurs flexions régulières, le radical peut être distingué aisément (en enlevant les affixes de la forme fléchie). À partir du radical, une racine empruntée peut être déduite de l'équation :

`racine×schème = radical`.

Conclusion

La structure interne des mots arabes décrite dans ce chapitre, nous conduit à considérer la morphologie arabe comme étant à la fois concaténative et gabaritique. Elle est concaténative dans le sens où les clitiques et les affixes sont rattachés aux radicaux par une simple opération de concaténation et gabaritique puis que le radical est le résultat de la combinaison d'une racine et un schème. À l'issue de ce premier chapitre décrivant les phénomènes morphologiques de l'arabe, nous sommes en mesure de présenter les différents travaux et approches concernant le traitement automatique de la morphologie arabe.

2. Traitement automatique de la morphologie arabe

Après avoir décrit dans le chapitre précédent les caractéristiques morphologiques de la langue arabe, nous présentons, dans ce chapitre, différentes approches proposées dans la littérature pour en réaliser l'analyse et la synthèse automatiques.

Dans la section 2.1, nous décrivons les opérations élémentaires réalisées par les systèmes de traitement de la morphologie. Cette section décrit, également, les deux approches gabaritique et concaténative utilisées dans l'analyse morphologique. Nous présentons, ensuite, dans la section 2.2 les principaux modèles computationnels utilisés pour réaliser ces traitement à savoir le modèle à deux-niveaux et le modèle multi-bande. Enfin, un survol des principaux analyseurs morphologiques de l'arabe est proposé dans la section 2.3.

2.1. Traitement morphologique arabe : processus de base

Le traitement automatique de la morphologie arabe a fait l'objet de plusieurs travaux de recherche. Ces travaux ont commencé dans les années soixante avec le premier analyseur morphologique proposé par (Cohen, 1970). Suite à ce travail, de multiples systèmes d'analyse et de génération morphologiques de l'arabe ont été construits. Ces systèmes permettent d'identifier les différents morphèmes d'un mot et de leur associer des traits morphologiques.

Deux types de morphèmes existent en arabe :

- morphèmes "*concaténtifs*" : ils se déclinent en trois catégories : des clitiques, des affixes et des radicaux. Ils se combinent entre-eux à l'aide de l'opération de concaténation pour produire des formes agglutinées et des formes fléchies.
- morphèmes "*gabaritiques*" : ils se déclinent en deux catégories : les racines et les schèmes. Ils peuvent se combiner grâce à l'opération de croisement pour former des radicaux.

Un système de traitement automatique de la morphologie est composé principalement d'un lexique et d'un ensemble de règles. Le lexique permet de stocker les connaissances lexicales spécifiques, tel que les clitiques, les affixes, les radicaux, les racines et les schèmes. Les règles, elles, définissent l'ordre des morphèmes dans le mot et permettent de réaliser les ajustements phonologiques et orthographiques nécessaires suite à une opération de concaténation ou de croisement de morphèmes.

Plusieurs choix sont possibles pour répartir la connaissance entre le lexique et les règles. Un choix extrême, que l'on appellera choix de niveau 0, consiste à tout représenter dans le lexique. Dans ce cas, le lexique est composé de formes agglutinées et associe à chacune d'entre elles son analyse morphologique. Dans ce cas le processus d'analyse morphologique se limite à un accès au lexique. Dans la pratique, une telle approche n'est pas viable pour la famille des langues sémitiques, elle conduit à des lexiques de taille déraisonnable et dont la maintenance est quasiment impossible.

Trois autres choix sont possibles selon que l'on stocke dans le lexique des clitiques et des formes fléchies (niveau 1) ou bien des clitiques, des affixes et des radicaux (niveau 2) ou encore des clitiques, des affixes, des racines et des schèmes (niveau 3).

Bien entendu, plus le niveau augmente plus le système de règles associé au lexique est complexe. Aux niveaux 1 et 2, seules des règles de concaténation sont nécessaires. Dans le premier cas, elles vérifient la compatibilité entre formes fléchies et clitiques, dans le second, elles vérifient aussi la compatibilité entre radicaux et affixes. La concaténation de morphèmes nécessite, dans certains cas, des ajustements orthographiques qui sont aussi modélisés à l'aide de règles.

Au niveau 3, des règles de croisement assurent la compatibilité d'une racine avec un schème et permettent de les combiner et de réaliser les ajustements morphologiques, phonologiques et orthographiques nécessaires à la suite du croisement.

La figure 2.1 décrit l'analyse de la forme agglutinée وسيطعموكم *wasayuTςimuwkum* "*et ils vous nourriront*" selon les différents niveaux.

niveau 0	forme agglutinée $wasayuT\varsigma imuwkum$				
niveau 1	proclitique + $wasa$		forme fléchie $yuT\varsigma imuwA$		+ enclitique kum
niveau 2	proclitique + $wasa$	préfixe + y	radical $uT\varsigma im$	+ suffixe uwA	+ enclitique kum
niveau 3	proclitique + $wasa$	préfixe + y	racine × schème $T\ \varsigma\ m \times u12i3$	+ suffixe uwA	+ enclitique kum

Table 2.1.: Niveaux de représentation d'un mot arabe

Les systèmes de traitement morphologiques de l'arabe peuvent être classés selon le niveau de représentation des morphèmes dans le lexique. La taille du lexique dépend fortement du choix du niveau de représentation. Afin de quantifier son influence sur la taille du lexique, nous avons identifié dans l'ATB le nombre de formes agglutinées, de formes fléchies, de radicaux et de racines différentes :

– niveau 0, formes agglutinées : $\sim 2M$
– niveau 1, formes fléchies : $\sim 300K$
– niveau 2, radicaux : $\sim 25K$
– niveau 3, racines : 2517

Comme on peut l'observer, l'influence du niveau de représentation sur la taille du lexique est très importante, le rapport est de l'ordre de 800 entre les niveaux extrêmes.

Dans le reste de cette section, nous décrivons les processus de base qui doivent être réalisés pour passer d'un niveau à un autre.

2.1.1. Segmentation

La segmentation est l'opération qui consiste à décomposer une forme agglutinée (délimité par des espaces dans le texte) en clitiques et forme fléchie (cf. section 1.2.1). Une forme agglutinée est composée généralement d'un nombre variable de proclitiques, d'une forme fléchie et éventuellement d'un enclitique. La séparation des clitiques de la forme fléchie est importante dans une perspective de TAL puisqu'elle permet, comme nous l'avons vu ci-dessus, de réduire considérablement la taille du lexique. En outre, la segmentation est nécessaire avant les opérations d'étiquetages grammatical et d'analyse syntaxique car les clitiques possèdent leurs propres parties de discours et fonctions syntaxiques.

Cette opération n'est pas toujours déterministe puisqu'un mot peut avoir plusieurs segmentations possibles comme nous l'avons illustré dans le chapitre 1. De plus, la concaténation des morphèmes peut conduire à des changements orthographiques. Une étape de normalisation orthographique peut ainsi s'imposer suite à la segmentation. Les principaux changements orthographiques en arabe dus à l'agglutination sont au nombre de quatre :

1. l'article défini ال *Al* subit des changements orthographiques qui consistent à omettre sa première lettre quand il suit la préposition ل *l* "*à*". Le mot المدرسة *Almdrsħ* "*l'école*" précédé par cette préposition devient للمدرسة *llmdrsħ* "*à l'école*"
2. la lettre ة *ħ* à la fin d'un mot devient ت *t* quand elle est suivie par un enclitique. Ainsi, la distinction entre les lettres ت et ة est perdue. Le mot مدرسة *mdrsħ* "*école*", après sa concaténation à l'enclitique هم *hm* "*leur*"devient

مدرستهم *mdrsthm* "*leur école*".

3. de la même manière, la lettre ى *ý* devient ا *A*. مستشفى *mstšfý* "*hôpital*" avec un enclitique se transforme en مستشفاهم *mstšfAhm* "*leur hôpital*".

4. La lettre ا *A* du suffixe verbal وا *wA* disparait quand un clitique est lié au verbe. De cette manière, كتبوا *ktbwA* "*ils ont écrit*" change à كتبوه *ktbwh* "*ils l'ont écrit*".

La normalisation pose également des problèmes d'ambiguïté. Le mot حكمتهم *Hkmthm*, par exemple, ne possède qu'une seule segmentation حكمة+هم *Hkmt+hm*. En dehors de tout contexte, ce mot peut être normalisé de deux façons différentes. En effet, la dernière lettre du premier segment حكمة peut être normalisée en ت *t* ou bien en ة *ħ*. Deux interprétations différentes sont donc produites : حكمت+هم *Hkmt+hm* "*elle les a gouvernés*" et حكمة+هم *Hkmħ+hm* "*leur sagesse*".

Un système de segmentation permet, étant donné une forme agglutinée, de générer tous ses découpages possibles et d'effectuer la normalisation de la forme simple après sa séparation avec les clitiques. Pour cela, un lexique de formes fléchies et une matrice de compatibilité des clitiques avec chaque forme fléchie sont nécessaires.

2.1.2. Analyse flexionnelle

L'analyse flexionnelle consiste à décomposer une forme fléchie en affixes et radical (cf. section 1.2.2). Les affixes permettent de déterminer les valeurs morphologiques du mot qui sont la personne, le genre, le nombre, le mode, le cas et l'état. Certains traits morphologiques ne sont pas véhiculés par des morphèmes, ils font partie du radical. C'est le cas de la voix et de l'aspect pour le verbe et parfois du nombre pour le nom.

L'identification du radical permet de déduire aisément le lemme de la forme fléchie. Cette opération, appelée lemmatisation, consiste à assigner à une forme fléchie le lemme qui lui correspond. Ce dernier représente la forme canonique d'un radical et correspond aux entrées lexicales dans un dictionnaire. Généralement, en arabe, les lemmes verbaux sont représentés par la forme fléchie à l'accompli de la troisième personne, masculin, singulier, et dans le cas des noms et des adjectifs, ils prennent la forme indéfinie du masculin singulier.

Les langues agglutinantes sont généralement caractérisées par une flexion concaténative (Beesley, 1998). Ainsi, un système morphologique, à ce stade de traitement, peut être fondé sur un processus de concaténations successives de morphèmes.

L'ordre de combinaison des clitiques, affixes et radical peut être représentée à l'aide d'une machines à états finis (cf. section 2.2) comme le montre la figure 2.1.

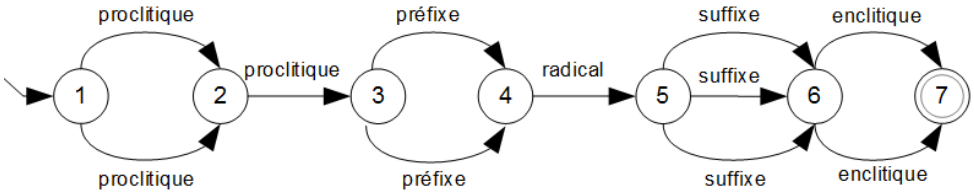

Figure 2.1.: Représentation de la morphologie concaténative à l'aide d'une machine à états finis

Un exemple de machine à états finis qui permet de générer tout le paradigme de flexion du verbe *ÂaTʕam* "*nourrir*" conjugué dans l'aspect inaccompli est donné dans la figure 2.2.

ÂuTʕimu	*je nourris*
nuTʕimu	*nous nourrissons*
tuTʕimu	*tu nourris*
tuTʕimiyna	
tuTʕimA	*vous nourrissez*
tuTʕimuwna	
tuTʕimna	
yuTʕimu	*il nourrit*
tuTʕimu	*elle nourrit*
yuTʕimAni	*ils nourrissent*
yuTʕimuwna	
yuTʕimna	*elles nourrissent*

Ce paradigme présente un échantillon de formes fléchies du verbe *ÂaTʕam*. Toutes ces formes simples peuvent être augmentées par des clitiques. La figure 2.2 présente un exemple de machine qui permet de générer cet échantillon.

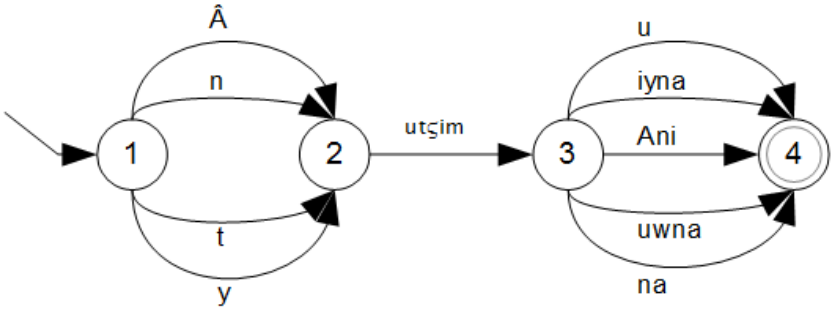

Figure 2.2.: Génération des formes fléchies du verbe *ÂaTʕam*

Toutes les formes fléchies du lemme verbal *ÂaTςam* acceptent des clitiques. Ces derniers peuvent être traduits simplement par des transitions dans la machine de la figure 2.2. La génération des formes agglutinées nécessite des transitions au début du mot qui correspondent aux proclitiques ordonnés alors que les enclitiques se rajoutent après les suffixes.

La machine présentée dans la figure 2.2 conduit à une surgénération de formes, elle produit des formes incorrectes telles que *nuTςimuwna* et *nuTςimAni*. Des règles morphologiques s'avèrent, ainsi, nécessaires pour bloquer la génération des formes incorrectes résultant de l'incompatibilité entre les affixes. Ces règles peuvent être elles-mêmes représentées par des machines à états finis.

2.1.3. Analyse dérivationnelle

L'analyse dérivationnelle consiste à décomposer un radical en une racine et un schème (cf. section 1.2.3). Cette opération permet de passer au niveau de représentation des mots le plus profond. Dans l'analyse morphologique de l'arabe, on retrouve des analyseurs qui ne réalisent pas ce niveau de traitement et se limitent à l'analyse de niveau 2.

L'extraction de la racine et du schème à partir d'un radical est elle aussi confrontée au problème de l'ambiguïté. À titre d'exemple, dans l'ATB (Maamouri *et al.*, 2004) 3250 radicaux possèdent plus d'une racine potentielle (Habash *et al.*, 2007).

L'ensemble des racines de l'arabe est ouvert : une racine peut être ajoutée pour des besoins de communication [a] ou encore empruntée pour combler une lacune dans le vocabulaire technique ou scientifique (Daniels, 2007). En revanche, l'ensemble de schèmes est clôt.

La compatibilité entre racines et schèmes peut être représentée par une matrice à deux dimensions à valeurs booléennes (cf. tableau 2.2). Une telle matrice indique les couples (racine, schèmes) valides.

La génération d'un radical à partir d'une racine et d'un schème nécessite, dans certains cas, des ajustements que l'on peut représenter à l'aide de règles phonologiques, morphologiques et orthographiques.

Voici, à titre d'exemple, deux règles d'ajustement :

1. La combinaison de la racine géminative م د د *m d d* et du schème verbal Ai1ta2a3 génère le radical امتدد *Aimtadad*. Une règle spécifique au traitement de racines géminatives assure la transformation de امتدد *Aimtadad* à امتدّ *Aimtad*∼ "*s'étendre*".

a. L'analyse morphologique au niveau des racines est avantageuse dans le traitement des dialectes où l'on voit fréquemment des mots empruntés. En effet, l'addition des racines qui proviennent d'origines différentes de l'arabe dans le lexique est plus simple que l'ajout des radicaux empruntés.

	R_1	R_2	R_3	R_4	R_5	...	R_n
S_1							
S_2							
S_3							
S_4							
S_5							
...							
S_k							

Table 2.2.: Matrice de compatibilité entre racines et schèmes

2. La combinaison de la racine ق و ل *q w l* et du schème Â12a3 génère le radical incorrect أقول *Âaqwal* "*virer*". Une règle morphologique spécifique aux racines creuses permet de remplacer la deuxième lettre de la racine par une voyelle longue. Appliquée à l'exemple précédent, elle permet de générer le radical أقال *ÂqAl*. L'application d'une telle règle peut être problématique. En effet, elle s'applique au radical استجوب *Ăistajwab* et le transforme en استجاب *ĂistajAb* "*accepter*". Or le radical استجوب *Ăistajwab* "*investiguer*" est valide mais il ne sera pas généré.

Le système de génération de radicaux pose aussi des problèmes de sur-génération. Les racines ف ت ح *f t H* et م ح ي *m H y*, par exemple, définissent respectivement les notions abstraites "*ouverture*" et "*nettoyage*". La combinaison de la racine avec le schème mi12A3 qui représente le nom d'outil d'un verbe produit, respectivement, les mots مفتاح *miftAH* "*clé*" et ممحاة *mimHAħ* "gomme". En revanche, la combinaison de ce schème avec la racine *k t b* qui définit la notion "*écriture*" génère le mot مكتاب *miktAb* qui devrait être, selon cette régularité, le synonyme de *stylo/crayon* cependant ce mot n'appartient pas au lexique.

En conclusion, un système de génération de radicaux par croisement de racines et de schèmes permet de réaliser des économies sur la taille du lexique. Mais il nécessite un système de règles complexe qui a tendance à sous-générer et à sur-générer.

2.1.4. Analyse et génération morphologique

L'analyse morphologique consiste à affecter à chaque mot d'un texte toutes les informations morphologiques qui lui sont associées. Elle repose sur les opérations élémentaires décrites ci-dessus. La table 2.3 décrit les différents traits morphologiques que chaque opération permet d'expliciter.

	trait morph.	valeurs possibles
Segmentation	conjonction	wa, fa, 0
	préposition	bi, ka, li, 0
	particule	sa, li, la, 0
	détermination	Al, 0
	enclitique	1S, 1P, 2MS, 2FS, 2D, 2MP, 2FP 3MS, 3FS, 3D, 3MP, 3FP, 0
Analyse flexionnelle	mode	indicatif, subjonctif, apocopé, 0
	personne	1, 2, 3, 0
	genre	masculin, féminin, 0
	cas	nominatif, accusatif, génitif, 0
	état	défini, indéfini, 0
	nombre	singulier, duel, pluriel, 0
Analyse dérivationnelle	nombre	singulier, duel, pluriel, 0
	aspect	perfectif, imperfectif, impératif, 0
	voix	active, passive, 0

Table 2.3.: Traits morphologiques d'un mot arabe

La table 2.3 est composée de trois parties (délimités par un doublement de lignes). La première partie concerne les clitiques, l'analyseur morphologique indique la présence du clitique dans le mot. La deuxième partie concerne la définition des informations morphologiques liées à la flexion. Ces informations sont traduites par des affixes. Enfin, les traits morphologiques de la troisième partie permettent de déterminer la racine et le schème du mot. La valeur 0 dans la première partie indique l'absence du clitique dans le mot alors que dans les deux dernières parties indiquent que le trait morphologique concerné ne s'applique pas au mot. Le nom et l'adjectif par exemple n'admettent pas de valeurs pour l'aspect, le mode et la voix alors que le verbe ne possède pas de valeurs pour l'état et le cas. Le nombre est présent dans l'analyse flexionnelle et l'analyse dérivationnelle puisque dans certains cas les affixes distinguent le nombre d'un nom et dans d'autres cas le schème détermine le nombre dans le cas du pluriel brisé (cf. sectionnoms).

L'analyse morphologique est indépendante généralement du contexte du mot. Une opération de désambiguïsation permet de choisir parmi toutes les analyses théoriques possibles d'un mot, la plus adéquate selon le contexte du mot. Cette opération est d'autant plus ambigüe que le niveau de représentation est profond. Le problème réside dans la façon d'effectuer le choix des analyses pertinentes parmi toutes les configurations combinatoirement possibles. En arabe, le nombre d'analyses possibles d'un mot en dehors de tout contexte est égal à $300,000$ [b], ce qui rend la tâche plus difficile en comparant ce nombre à celui de l'anglais 46 (Habash et Rambow, 2005).

b. Ce nombre est le résultat de la multiplication de nombre de valeurs possibles des traits morphologiques.

La génération morphologique est le processus inverse de l'analyse morphologique. Cette opération consiste à produire la forme surfacique d'un mot à partir de sa représentation morphologique, composée par des paires (traits morphologiques, valeurs). Cette opération n'est pas ambigüe, à une représentation morphologique entièrement spécifiée correspond à une forme surfacique au plus.

2.2. Morphologie à deux-niveaux

Depuis son introduction par Koskenniemi (1983) dans les années 80, la morphologie à deux niveaux est devenue le formalisme standard pour le traitement automatique de la morphologie.

Ce formalisme définit deux niveaux de représentation, le niveau surfacique et le niveau lexical. La correspondance entre les deux est réalisée à l'aide de machines à états finis, comme l'illustre la figure 2.3.

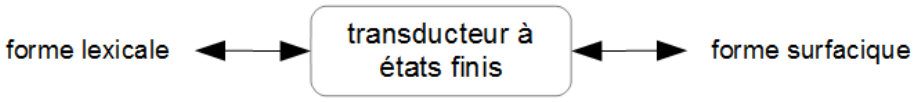

Figure 2.3.: Morphologie à deux-niveaux

Ces dernières se trouvent au cœur de nombreux systèmes de TAL, notamment pour le traitement automatique de la morphologie (Sproat, 1995; Roche et Schabes, 1997) et de la phonologie.

Plusieurs caractéristiques formelles ont contribué à leur succès dans le domaine du TAL. Parmi ces caractéristiques, nous pouvons citer :

La réversibilité Les machines à états finis sont réversibles. Une même machine peut servir à effectuer l'analyse et la génération.

La modularité Les machines à états finis peuvent être combinées entre elles par plusieurs opérations, telles que l'union, la concaténation, la fermeture de Kleene ou la composition. La combinaison de plusieurs machines réalisant des traitements différents, sous la forme d'une machine unique, permet de réaliser des traitements efficaces.

La rapidité Les machines à états finis, du moins les reconnaisseurs (voir la distinction ci-dessous) peuvent être rendus déterministes et peuvent effectuer la reconnaissance d'un mot de manière très efficace.

La compacité L'opération de minimisation permet de réduire autant que faire ce peut la taille des machines à états finis. Il est ainsi possible de représenter des objets importants, tel que des lexiques, de manière optimale.

Les machines à états finis se sont révélées particulièrement adaptées au traitement de la morphologie et de la phonologie. En effet, comme l'a montré (Karttunen, 1995) les règles morphologiques et phonologiques peuvent être représentées d'une manière simple et directe sous la forme de machines à états finis.

De plus, comme nous le verrons ci-dessous, les machines à états finis permettent d'implémenter des modèles de la morphologie concaténative ainsi que gabaritique.

Étant donné le rôle important que jouent les machines à états finis dans notre travail, nous commençons par donner une brève description de ces dernières et de leur utilisation pour la modélisation linguistique. Nous décrivons ensuite deux modèles standard utilisés dans le traitement de la morphologie arabe qui peuvent être implémentés à l'aide de machines à états finis : le modèle à deux niveaux et le modèle multibande. Ces deux modèles permettent d'implémenter respectivement les aspects concaténatifs et gabaritiques de la morphologie arabe.

Machines à nombre fini d'états

De nombreuses définitions des machines finies ont été proposées dans la littérature. Ces définitions sont généralement équivalentes mais varient entre elles notamment par la terminologie qu'elle adoptent. Les définitions et la terminologie que nous adoptons dans ce document est empruntée à (Roche et Schabes, 1997; Jurafsky et Martin, 2000). Nous nous intéresserons principalement à deux types de machines : les reconnaisseurs et les transducteurs. Avant de définir ces deux types de machines, commençons par donner quelques définitions des objets qu'elles manipulent.

L'élément le plus simple manipulé est le $symbole$ qui est une entité indivisible. Un ensemble fini de symbole est appelé $alphabet$, noté de manière conventionnelle par le symbole Σ. Des symboles d'un alphabet Σ combinés à l'aide de l'opération de la concaténation constituent un mot sur Σ. La longueur d'un mot m, notée $|m|$ est égale au nombre de symboles qui le constituent. Le mot de longueur zéro est noté conventionnellement ε.

L'ensemble de tous les mots que l'on peut former à l'aide de l'alphabet Σ est noté Σ^*. Un $langage\ L$ sur l'alphabet Σ est un sous-ensemble de Σ^*.

On appelle $reconnaisseur$ un algorithme ou une machine abstraite qui, étant donné un mot, décide, après un certain nombre d'étapes si ce mot est accepté ou pas. Un reconnaisseur définit ainsi un langage, qui est l'ensemble des mots qu'il reconnaît.

Automates On s'intéressera ici à un type de reconnaisseur simple : les automates finis, que l'on appellera simplement $automates$ dans la suite de ce document.

Un automate est un quintuplet $(Q, \Sigma, \delta, q_0, F)$, où :

– Q est un ensemble fini d'états

– Σ est un alphabet, appelé alphabet d'entrée
– δ est une *fonction de transition* définie comme suit :
  $\delta : Q \times \Sigma \cup \{\varepsilon\} \to \mathcal{P}(Q)$
  où $\mathcal{P}(Q)$ est l'ensemble des parties de Q.
– $q_0 \in Q$ est l'état initial
– $F \subseteq Q$ est l'ensemble des états d'acceptation

Voici un exemple d'automate possédant trois états et quatre transitions :
$A = (Q, \Sigma, \delta, q_0, F)$, avec :
– $Q = \{0, 1, 2\}$
– $\Sigma = \{a, b\}$
– $\delta(0, a) = \{0, 1\}$, $\delta(1, b) = \{1\}$, $\delta(2, b) = \{2\}$
– $q_0 = 0$
– $F = \{2\}$

Les automates sont souvent représentés sous la forme de graphes orientés dont les sommets et les arcs sont étiquetés. Les sommets du graphe correspondent aux états, et chaque transition est représentée par un arc. L'état initial est identifié à l'aide d'une flèche entrante et les états d'acceptation à l'aide d'un double cercle. On trouvera, figure 2.4, une représentation graphique de A.

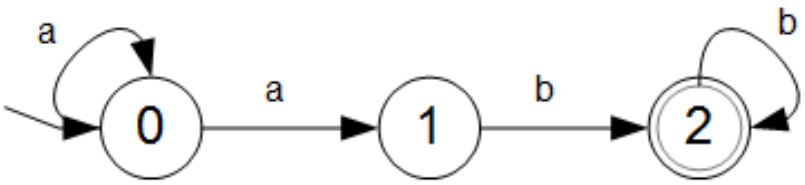

Figure 2.4.: Représentation graphique d'un automate

Le processus de reconnaissance d'un mot par un automate fait appel aux notions de *configuration* et de *mouvement*.

Une configuration décrit complètement l'état d'un automate lors du processus de reconnaissance d'un mot. Étant donnée un automate $A = (Q, \Sigma, \delta, q_0, F)$, une configuration est un couple $(q, m) \in Q \times \Sigma^*$.
– q représente l'état courant de A
– m est la partie du mot à reconnaître non encore lue. Le premier symbole de m (le plus à gauche) est le prochain symbole qui doit être reconnu par l'automate. Si $m = \varepsilon$ alors tout le mot a été lu.

Un mouvement, noté $\vdash$, permet de passer d'une configuration à une autre. Un mouvement entre une configuration (q, aw) et une configuration (q', w) est valide si la fonction de transition de l'automate permet de passer de (q, aw) à (q', w) :

$$(q, aw) \vdash (q', w) \text{ si } q' \in \delta(q, a)$$

lors de ce mouvement, l'automate passe de l'état q à l'état q' et le symbole a est consommé.

Etant donné un automate $A = (Q, \Sigma, \delta, q_0, F)$ et un mot $m \in \Sigma^*$ on définit de plus une *configuration initiale* : (q_0, m) qui est la configuration dans laquelle se trouve l'automate avant de commencer à lire le mot m et des *configurations d'acceptation* : (q, ε) avec $q \in F$, qui sont les configuration dans lesquelles se trouvent l'automate A si la reconnaissance a réussi.

Nous sommes maintenant en mesure de définir le processus de reconnaissance d'un mot m par un automate A ainsi que le langage reconnu par A, noté $L(A)$.

m est reconnu par A, s'il existe une séquence de mouvements valides menant de la configuration initiale (q_0, m) à une configuration d'acceptation (q, ε).

Le langage reconnu par A ($L(A)$) est l'ensemble des mots reconnus par A :

$$L(A) = \{m \in \Sigma^* | (q_0, m) \overset{*}{\vdash} (q, \varepsilon) \text{ avec } q \in F\}$$

Dans la représentation graphique de A, le processus de reconnaissance du mot m correspond à l'existence d'une séquence de transitions $t_1 \dots t_k$ menant de l'état initial de A à un de ses états d'acceptation, tel que la concaténation des symboles des transitions $t_1 \dots t_k$ forme le mot m.

L'ensemble des langages qui peuvent être reconnus à l'aide d'un automate fini est appelé l'ensemble des langage *reconnaissables*.

Les automates tel que nous les avons défini possèdent une propriété qui est le *non déterminisme*. Cette dernière provient de la définition de la fonction de transition qui peut associer à un couple composé d'un état et d'un symbole, plus d'un état. Cette propriété a une conséquence sur le processus de reconnaissance d'un mot. En effet, lors de la reconnaissance d'un mot, lorsque l'automate se trouve dans une configuration $c = (q, am)$ et que la fonction de transition associe plus d'une image au couple (q, a) alors plusieurs mouvements sont possibles à partir de c, menant à des configurations différentes. Le non déterminisme est illustré dans la représentation graphique des automates, par l'existence de plusieurs chemins étiquetés par le même mot permettant de rejoindre un état d'acceptation depuis l'état initial.

La conséquence de tout cela est que la reconnaissance d'un mot est un processus qui peut être coûteux. Il est, dans le pire des cas exponentiel en temps par rapport à la longueur du mot à reconnaître.

On peut définir une variante déterministe des automates de la manière suivante :

Un automate déterministe est un quintuplet $(Q, \Sigma, \delta, q_0, F)$, où :

– Q est un ensemble fini d'états
– Σ est un alphabet, appelé alphabet d'entrée
– δ est une $fonction\ de\ transition$ définie comme suit :
  $\delta : Q \times \Sigma \rightarrow Q$
– $q_0 \in Q$ est l'état initial
– $F \subseteq Q$ est l'ensemble des états d'acceptation

La différence entre les automates déterministes et les automates non déterministes réside dans la fonction de transition. Dans un automate déterministe, celle-ci associe à un couple composé d'un état et d'un symbole, au plus un état. Ainsi, à partir d'une configuration, il existe au plus un mouvement possible. Par conséquent, le processus de reconnaissance d'un mot sera composé au plus d'autant de mouvements que le mot possède de symboles. Le processus de reconnaissance est donc linéaire en fonction de la longueur du mot à reconnaître.

Les automates déterministes et non déterministes reconnaissent la même famille de langage, les langages reconnaissable. Pour tout langage reconnaissable, il existe un automate déterministe qui reconnaît le même langage. Le théorème de Rabin-Scott permet de construire, étant donné un automate non déterministe, un automate déterministe reconnaissant le même langage.

A titre d'exemple, la figure 2.5 représente un automate déterministe reconnaissant le même langage que l'automate déterministe de la figure 2.4

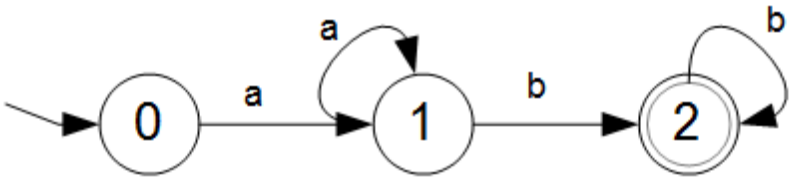

Figure 2.5.: Automate fini déterministe

Le déterminisme des automates est une propriété importante pour les aspects applicatifs en général et pour le traitement automatique de la langue en particulier car il garantit un temps de reconnaissance linéaire.

Application à la morphologie flexionnelle Les noms réguliers arabes se fléchissent en genre par la concaténation du caractère ة *ħ* lors du passage du masculin au féminin. Un nom masculin peut être représenté par un automate linéaire. À titre d'exemple, le nom سليم *slym* "*sain*" :

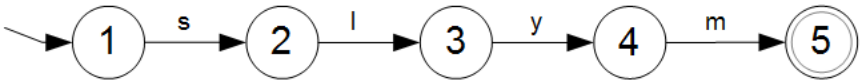

Figure 2.6.: Représentation du mot *slym* à l'aide d'un automate fini

Le lexique de tous les noms masculins peut être obtenu à l'aide d'un automate *noms-masculins*, obtenu par union des automates correspondant aux différents noms masculins.

L'obtention des noms féminins est réalisée à l'aide de la concaténation de $\hbar$ à la fin des noms masculins. Dans certains cas, le passage du masculin au féminin est irrégulier. Tous les noms féminins irréguliers peuvent être représentés par un automate *exceptions*. La liste des noms féminins peuvent ainsi être représentée par l'opération suivante :

(noms-masculins.ħ) $\cup$ *exceptions*

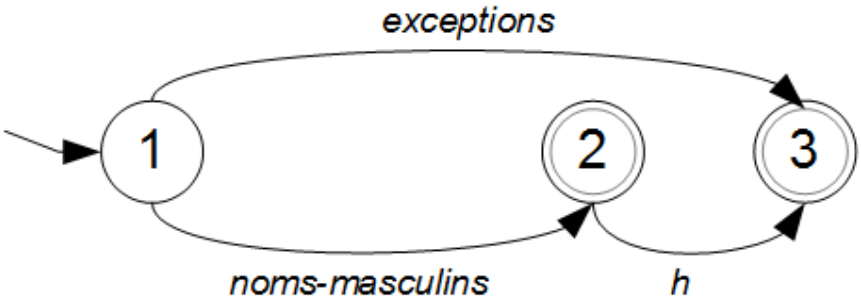

Figure 2.7.: Lexique des noms féminins représenté sous la forme d'un automate

Transducteurs Nous avons vu ci-dessus que les automates permettent de reconnaître des langages. Nous allons introduire maintenant un nouveau type de machine, les *transducteurs finis*, que nous appellerons simplement transducteurs dans la suite de ce document. Les transducteurs permettent de reconnaître des couples de mots $(u, v) \in \Sigma_1^* \times \Sigma_2^*$ ou, en d'autres termes, une relation sur $\Sigma_1^* \times \Sigma_2^*$.

Un transducteur peut être vu comme un reconnaisseur pour des couples de mots ou comme une machine produisant une sortie pour une entrée donnée.

Un transducteur est défini par un 6-uplet $(Q, \Sigma_1, \Sigma_2, \delta, q_0, F)$ où :

- Q est un ensemble fini d'états
- Σ_1 est un alphabet, appelé alphabet d'entrée
- Σ_2 est un alphabet, appelé alphabet de sortie
- δ est la fonction de transition définie comme suit :
  $\delta : Q \times \Sigma_1 \cup \{\varepsilon\} \times \Sigma_2 \cup \{\varepsilon\} \rightarrow \mathcal{P}(Q)$

– $q_0 \in Q$ est l'état initial
– $F \subseteq Q$ est l'ensemble des états d'acceptation

La représentation graphique des tranducteurs ressemble à celle des automates à la différence que les transitions sont étiquetés par des paires de symboles.

Un exemple de transducteur est représenté dans la figure 2.8. Ce transducteur permet de reconnaître la relation $\{(ab, df), (ac, fe)\}$.

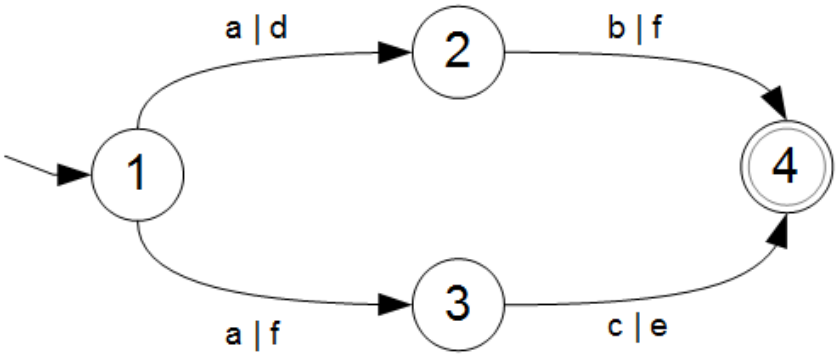

Figure 2.8.: Représentation graphique d'un transducteur

L'étude des propriété des transducteurs dépasse le cadre de la brève introduction que nous faisons ici. Il est néanmoins important d'aborder trois aspects des transducteurs important pour les applications en général et pour le traitement de la langue en particulier.

Le premier est l'opération de *composition*. Etant donné les transducteurs T_1 et T_2 tel que l'alphabet de sortie de T_1 et l'alphabet d'entrée de T_2 sont égaux. On note $T_1 \circ T_2$ la composition de T_1 et de T_2, qui consiste à fournir en entrée à T_2, la sortie de T_1. L'algorithme de composition permet de construire un transducteur T_3 qui réalise directement la composition $T_1 \circ T_2$. La possibilité de construire le transducteur qui réalise la composition est une propriété importante pour le traitement automatique des langues car il permet d'adopter une démarche modulaire dans la modélisation d'opérations complexes, tel que l'analyse morphologique. Il est en effet possible de décomposer un processus complexe en une séquence de processus élémentaires modélisé chacun par un transducteur puis de composer ces transducteurs entre eux pour obtenir, en fin de compte, un transducteur unique.

Le second aspect est le caractère inversible des transducteurs qui permet d'utiliser ces derniers de manière bidirectionnelle. Il s'agit là aussi d'une propriété importante car elle permet d'utiliser un même transducteur en analyse et en génération.

Le troisième aspect que nous aborderons ici est celui de l'ambiguïté. Un transducteur est dit ambigu si pour certaines entrées, il associe plus d'une sortie. Le transducteur qui reconnaît par exemple la relation $\{(ab, cd), (ab, de)\}$ est un transducteur ambigu car pour l'entrée ab il produira deux sorties. L'ambiguïté a un lien avec le déterminisme dans la mesure où un transducteur ambigu ne peut être rendu déterministe. Dans la pratique, les transducteurs utilisés pour modéliser la morphologie sont souvent ambigus car l'analyse morphologique est dans certains cas ambigüe.

Exemple de la morphologie arabe Un exemple de transducteur fini simple est donné dans la figure 2.9. Ce transducteur peut réaliser l'analyse en genre d'un nom arabe singulier. Il permet de reconnaitre un lemme nominal au singulier et de traduire son suffixe en valeur morphologique.

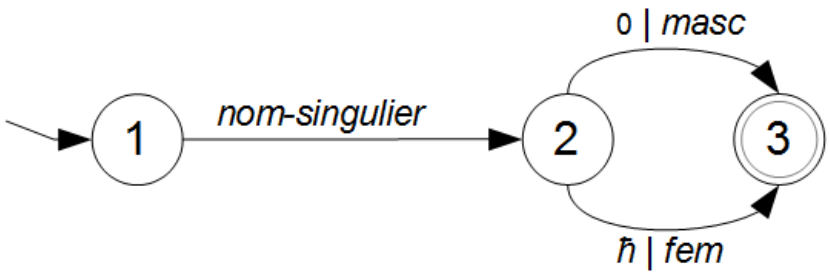

Figure 2.9.: Flexion nominale en genre à l'aide d'un transducteur

De la même manière, la flexion en nombre peut être réalisée à l'aide d'un transducteur. Dans ce cas, l'apparition du suffixe *wn* ou du suffixe *At* sont traduits par la valeur *pluriel* du nombre.

Les automates et les transducteurs sont parfaitement adaptés à la mise en œuvre de la segmentation et de la flexion. En revanche, ils ne sont pas adaptés à la modélisation d'opérations non concaténatives mises en œuvre dans la dérivation. Cette dernière nécessite un nouveau type de machine appelé automates multibande, décrits ci-dessous.

Automates multibande Les automates multibande ont été introduits par Rabin et Scott (1959) et Elgot et Mezei (1965). La définition que nous donnons ici nous est propre, elle est adaptée à l'usage qui est fait de ce type de machine pour la morphologie gabaritique.

Les automates multibande font appel à la notion de n-alphabet. Ces derniers sont composés de n-symboles qui sont des n-uplets de symboles. Un automate

à n bandes lit un $n-1$-symbole dont les $n-1$ symboles se trouvent sur $n-1$ bandes de lecture et écrit un mot sur la bande d'écriture.

Un automate à n bandes est un 6-uplet $(Q, \Sigma, \Sigma_n, \delta, q_0, F)$ où
- Q est l'ensemble fini d'états
- $\Sigma = \Sigma_1, \times \Sigma_2 \times \ldots \Sigma_{n-1}$ est le $n-1$ alphabet d'entrée
- Σ_n est l'alphabet de sortie
- δ est la fonction de transition définie comme suit :
  $\delta : Q \times \Sigma \times \Sigma_n^* \rightarrow Q$
- $q_0 \in$ Q est l'état initial
- F $\subseteq$ Q est l'ensemble des états d'acceptation

Les automates multibande sont utilisés en morphologie gabaritique. En arabe, par exemple, un automate 3-bande peut modéliser la dérivation d'un radical à partir d'une racine et d'un schème. Dans ce cas, deux bandes d'entrée et une bande de sortie sont utilisées. Le schème et la racine occupent les bandes d'entrée alors que le radical est généré sur la bande de sortie.

Un exemple d'automate 3-bande est donné dans la figure 2.10. Cet automate réalise la dérivation des radicaux *katab* et *daras*.

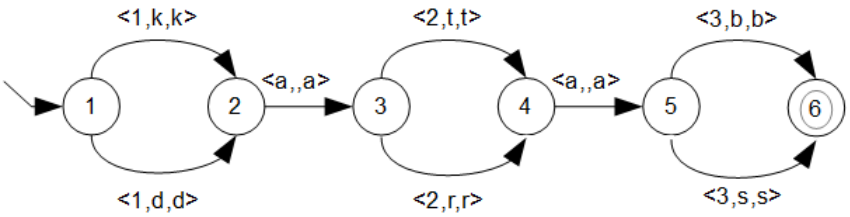

Figure 2.10.: Processus de dérivation à l'aide d'un automate multibande

Cet automate génère les radicaux *katab* et *daras* à partir du schème 1a2a3 et des racines *k t b* et *d r s* respectivement. La génération de ces radicaux est décrite dans la figure 2.11.

schème (entrée)	1	a	2	a	3
racine (entrée)	r_1		r_2		r_3
radical (sortie)	r_1	a	r_2	a	r_3

Figure 2.11.: Génération de radicaux à l'aide d'un automate multibande

Les bandes sont représentées verticalement (de haut en bas). La génération du radical à l'aide de cinq transitions représentées par des triplets. La première transition (1,r_1,r_1) lit le symbole 1 sur la première bande, le symbole $r_1 \in \{k, d\}$

sur la deuxième bande et écrit r_1 sur la troisième bande. La deuxième transition consiste à reproduire la voyelle de la première bande à la forme de surface. De cette manière, le radical est généré sur la bande d'écriture horizontalement de droite à gauche.

2.2.1. Modèle à deux niveaux

Comme nous l'avons évoqué, ce modèle définit deux niveaux de représentation. Un niveau lexical qui représente un suite de morphèmes et un niveau surfacique représentant la forme de surface. Les deux niveaux sont mis en correspondance à l'aide de règles, dites règles à deux niveaux.

Ces règles se présentent de la façon suivante

L:S OPERATEUR CG _ CD

où L est une forme lexicale et S une forme surfacique. CG et CD représentent respectivement le contexte gauche et le contexte droit dans lequel l'appariement L:S apparaît. Les contextes peuvent porter sur le niveau lexical, le niveau surfacique ou les deux.

Il existe quatre types de règles selon la valeur que prend OPERATEUR :

1. $\rightarrow$ l'appariement L:S n'est possible que dans le contexte CG _ CD
2. $\leftarrow$ le contexte CG _ CD force l'appariement L:S
3. $\leftrightarrow$ le contexte CG _ CD est nécessaire et suffisant pour observer l'appariement L:S
4. /$\leftarrow$ l'appariement L:S ne peut être observé dans le contexte CG _ CD

Un exemple formel de règle à deux niveaux est donné ci-dessous :

a:b $\leftrightarrow$ c: _ e:

Cette règle associe la forme lexicale a à la forme surfacique b si et seulement si a est précédé de c suivi de e.

Plus concrètement, en arabe, la flexion du mot féminin singulier مستشفى *mustašfaý* "*hôpital*" au pluriel مستشفيات *mustašfayAt* "*hôpitaux*" est réalisée à l'aide des trois règles à deux niveaux suivantes :

(1) X:X $\rightarrow$ _

(2) ý:y $\leftrightarrow$ _ +: A: t:

(3) +:0 $\rightarrow$ _

La première règle est indépendante du contexte, elle représente l'identité où tout caractère lexical est reproduit sur la forme de surface (X est ici une variable). La deuxième règle permet de remplacer le caractère ý par le caractère y s'il est suivi par le suffixe /+At/. La troisième règle, qui est indépendante du contexte, réalise la suppression du caractère +. L'obtention de la forme surfacique est réalisée après la suppression du symbole nul '0'.

Une règle à deux niveaux peut être compilée sous la forme d'un transducteur (Karttunen, 1995). Un ensemble de règles peut aussi être compilé sous la forme d'un transducteur unique qui réalise la correspondance entre forme lexicale et forme de surface comme l'illustre la figure 2.12.

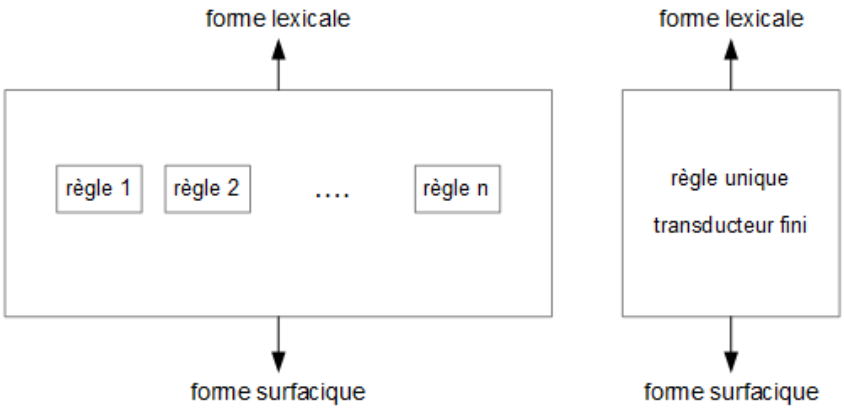

Figure 2.12.: Représentation d'une règle à deux-niveau par un transducteur

Les règles à deux niveaux proposées par Koskenniemi (1983) ne permettent d'apparier que des caractères uniques (L et S ne sont composés que d'un caractère). Black *et al.* (1987) propose une extension permettant d'apparier des séquences de caractères de longueur égale. Ruessink (1989) a introduit plus tard des règles qui relient des séquences de longueurs différentes. Pulman et Hepple (1993) a rajouté les paires (traits morphologiques/valeurs) dans la représentation lexicale. Par la suite, les règles sont représentées de la manière suivante :

CGS _ S _ CDS OPERATEUR CGL _ L _ CDS

CGS, CDS correspondent respectivement aux contextes gauche et droit de la forme surfacique et CGL, CDL représentent les contextes gauche et droit de la forme lexicale. La valeur de OPERATEUR est soit $\rightarrow$ ou $\leftrightarrow$. Le premier indique que si L apparait dans le contexte donné S apparait sur la forme de surface, alors que le deuxième ajoute la condition que si L apparait dans le contexte donné la forme de surface doit satisfaire S.

À l'aide de ces règles, la flexion du mot féminin singulier مستشفى *mustašfaý* "*hôpital*" au pluriel مستشفيات *mustašfayAt* "*hôpitaux*" est réalisée comme suit :

(1) * _ X _ * → * _ X _ *

(2) * _ ý _ +At ↔ * y _ +At

(3) * _ + _ * → * _ 0 _ *

Le symbole * indique un contexte vide qui représente une condition toujours satisfaite. La première et la troisième règles sont indépendantes du contexte. La deuxième règle permet de remplacer le caractère ý par le caractère y s'il est suivi par le suffixe /At/ marqué par le signe +. Le symbole ↔ indique que si la forme lexicale apparait alors la forme surfacique est générée et inversement. Le symbole → autorise un seul sens.

Dans le modèle à deux niveaux, la forme surfacique d'un mot est décrite comme une concaténation de morphèmes. Le mot *wasayutʕimuwkum* "*ils vous nourriront*" est le résultat de la concaténation de *utʕim* "*nourrir*" avec les proclitiques *wa+* "*et*", *sa+* "*futur*", le circonfixe *y+ +uwA* qui correspond à la troisième personne du masculin pluriel et l'enclitique *kum* "*vous*". La génération de la forme *wasayutʕimuwkum* est décrite comme suit :

w	a	+	s	a	+	y	+	u	T	ʕ	i	m	+	u	w	A	+	k	u	m	forme lexicale
1	1	3	1	1	3	1	3	1	1	1	1	1	3	1	1	2	3	1	1	1	
w	a	0	s	a	0	y	0	u	T	ʕ	i	m	0	u	w	0	0	k	u	m	forme surfacique

Les règles utilisées pour réaliser l'appariement de la forme lexicale et la forme surfacique sont :

1- identité * _ X _ * → * _ X _ *

2- suppression du A * _ A _ * → * _ 0 _ +

3- suppression du délimiteur '+' * _ + _ * → * _ 0 _ *

Les règles (1) et (3) sont indépendantes du contexte. Comme évoqué précédemment, (1) permet de reproduire les caractères de la forme lexicale sur la forme surfacique et (3) assure la suppression du signe + qui sépare les morphèmes de la forme lexicale. La règle (2), en revanche, constitue une règle d'ajustement standard de l'arabe. Elle supprime la lettre A du suffixe *+uwA* s'il est suivi d'un enclitique marqué par +

Bien que le modèle à deux niveaux soit bien adapté à l'agglutination et à la flexion de l'arabe, il ne permet pas de prendre en compte la dérivation.

Pour surmonter ce problème, Kay (1987) propose un modèle dans lequel les morphèmes peuvent être représentés sur plusieurs bandes et les symboles qui les composent peuvent être entrecroisés.

2.2.2. Modèle multi-bande

L'utilisation de multiples bandes pour représenter des morphèmes élémentaires apparait dans (Kay, 1987). Afin de prendre en compte les aspects concaténatifs et gabaritiques de la morphologie arabe, Kiraz (1994) a fusionné le modèle multibande et le modèle à deux niveaux. Le formalisme de Kiraz (1994) est conçu pour représenter tous les phénomènes morphologiques de l'arabe. Ce formalisme adopte exactement le modèle à deux niveaux décrit ci-dessus et l'étend à plusieurs bandes.

La forme lexicale est représentée par un $(n-1)$-uplet de symboles alors que la forme surfacique est représentée par une séquence unique de symboles. Pour mieux expliquer ce formalisme, prenons comme exemple une forme lexicale représentée sur trois bandes contenant respectivement des consonnes, des voyelles et des chiffres. Soient (a), (b) deux règles :

(a) `*` - `CV` - `*` $\leftrightarrow$ `*` - `(D,C,V)` - `*`

(b) `*` - `C` - `*` $\leftrightarrow$ `*` - `(D,C,)` - `*`

Les symboles `C`, `V` et `D` correspondent respectivement à une consonne, une voyelle et un chiffre. Le symbole `*` indique un contexte vide. La règle (a) permet, en lisant un chiffre `D` sur la première bande, une consonne `C` sur la deuxième bande et une voyelle `D` sur la troisième bande, de générer la forme surfacique `CV`. La règle (b) est identique à la règle (a) sauf que la troisième bande ne contient pas de voyelles. Ces règles permettent de générer le radical *daras* à partir de *123*, *drs* et *aa* :

1	2	3	1ère bande
d	r	s	2ème bande
a	a		3ème bande
(a)	(a)	(b)	
da	ra	s	4ème bande

L'exemple précédent relie la forme de surface *daras* à la forme lexicale composée de *123*, *drs* et *aa*. Ces règles permettent de générer tous les verbes de la forme (I) tels que *katab* "*écrire*" et *rakib* "*monter*". La bande de chiffres utilisées dans cet exemple indique l'ordre d'insertion de consonnes de la racine dans la forme de surface.

Nous présentons dans le reste de cette section deux systèmes morphologiques de l'arabe où on retrouve les deux modèles à deux-niveaux et multibande. Le premier système réalise l'analyse des verbes (McCarthy, 1981) alors que le deuxième a été implémenté pour l'analyse des noms (McCarthy, 1993).

2.2.3. Analyse de verbes

Pour décrire la morphologie verbale de l'arabe, McCarthy (1981) a proposé un modèle dans lequel un radical est représenté par trois types de morphèmes, un schème sous sa forme non diacritée représenté par des symboles `C V` qui indiquent, respectivement, une consonne et une voyelle quelconques, une racine composée de trois consonnes et un vocalisme composé de voyelles. Chaque morphème occupe une bande indépendante. Le radical *uTʕim*, par exemple, est généré à partir des trois morphèmes *VCCVC*, *Tʕm* et "*ui*".

V	C	C	V	C	bande de schème
	T	ʕ		m	bande de racine
u			i		bande de vocalisme

La génération de *uTʕim* à partir de ces trois morphèmes se réalise à l'aide de la projection des consonnes de la racine (deuxième bande) si la bande consacrée au schème contient `C` et des voyelles du vocalisme (troisième bande) quand le schème contient le symbole `V`.

Ces règles sont utilisées dans la génération de tous les radicaux des verbes sains (cf. section 1.3.2).

Lexique Une forme fléchie (`ff`) selon McCarthy (1981) est obtenue à l'aide de la formule :

```
ff = affixe + racine×schème×vocalisme + affixe
```

Le lexique utilisé est composé de 4 lexiques :

(a) lexique de schèmes : la forme du schème dépend des valeurs de l'aspect et de la voix.

```
1    [CVCVC, accompli actif]
     [CVCVC, accompli passif]
     [VCCVC, inaccompli actif]
     [VCCVC, inaccompli passif]
2    [CVCCVC, accompli actif]
     [CVCCVC, accompli passif]
     [CVCCVC, inaccompli actif]
     [CVCCVC, inaccompli passif]
   ..
   ..
10   [CCVCCVC, accompli actif]
     [CCVCCVC, accompli passif]
     [CCVCCVC, inaccompli actif]
     [CCVCCVC, inaccompli passif]
```

(b) lexique de racines : chaque racine est associée à la liste de schèmes.

```
1   ktb
2   drs
```

(c) lexique de vocalismes : le vocalisme détermine les voyelles du schème, il dépend également de l'aspect et de la voix.

```
1     [aa, accompli actif]
      [ai, accompli passif]
      [au, inaccompli actif]
      [ua, inaccompli passif]
2     [aa, accompli actif]
      [ui, accompli passif]
      [ui, inaccompli actif]
      [ua, inaccompli passif]
   ..
   ..
10    [a, accompli actif]
      [ui, accompli passif]
      [a, inaccompli actif]
      [ui, inaccompli passif]
```

(d) lexique d'affixes : les affixes sont déterminés à l'aide des traits morphologiques.

```
1   [tu : personne=1, nombre=s, aspect=accompli]
2   [ta : personne=2, genre=m, nombre=s, aspect=accompli]
3   [ti : personne=2, genre=f, nombre=s, aspect=accompli]
4   [a : personne=3, genre=m, nombre=s, aspect=accompli]
5   [at : personne=3, genre=f, nombre=s, aspect=accompli]
```

Règles Comme nous l'avons évoqué la forme lexicale selon McCarthy (1981) est regroupée sur trois bandes. Les affixes sont écrits sur la première bande avec le schème. McCarthy (1981) propose trois règles principales pour le traitement de verbes :

a- règles de base

R1 `* - X - * →* - (X„) - *` où $X \notin C, V$

R2 `* - X - * →* - (C,X,) - *`

R3 `* - X - * →* - (V, ,X) - *`

b- règle de borne

R4 `* - - * →* - + - *`

c- règle de propagation

R5 `* - X - * →(C,X,) ... - (C„) - *` où $X \neq +$

R5 `* - X - * →(V„X) ... - (V„) - *` où $X \neq +$

Les règles de base (a) sont indépendantes du contexte. R1 concerne les affixes, elle permet de projeter tous les caractères des affixes sur la forme de surface. Ces caractères sont présentés sur la première bande et sont obligatoirement différents de C ou V. R2 et R3 permet de produire les consonnes de la racine et les voyelles du vocalisme sur la forme de surface. La règle de borne R4, indépendante du contexte, permet de supprimer le délimiteur + qui sépare les affixes du schème sur la première bande.

Le nombre de symboles C dans le schème correspond généralement au nombre de consonnes de la racine. En revanche, dans certains cas, les schèmes trilitères géminatifs contiennent quatre symboles C. La règle R5 assure la propagation des consonnes sur la forme de surface. Pour mieux illustrer l'application de ces règles, prenons comme exemple la génération de la forme surfacique *darrasnA* "*nous avons enseigné*" :

C	V	C	C	V	C	+	n	A	bande de schème
d		r			s				bande de racine
	a			a					bande de vocalisme
R2	R3	R2	R5	R3	R2	R4	R1	R1	
d	a	r	r	a	s		n	A	forme surfacique

Dans cet exemple, tout symbole de schème qui n'appartient pas à l'ensemble {C, V, +} est projeté sur la forme de surface grâce à la règle R1. Les symboles C et V du schème sont traduits par la consonne et la voyelle de la deuxième et la troisième bande respectivement à l'aide de R2 et R3. La règle R5 met en correspondance la forme lexicale (C, ,) à la dernière consonne de la racine écrite sur la forme surfacique. Enfin, R5 remplace le signe +, délimitant le radical du suffixe, par un caractère vide.

2.2.4. Analyse des noms

Dans cette section, nous décrivons la méthodologie de McCarthy (1993) dans l'analyse des noms. Cette analyse consiste à représenter les noms arabes à l'aide

de syllabes. McCarthy (1993) a eu recours à trois syllabes génériques pour représenter les schèmes nominaux :

- s_1 : CV où C et V représentent respectivement une consonne et une voyelle simple.
- s_2 : CVV où VV est une voyelle longue [c].
- s_3 : CVC

McCarthy (1993) admet que tous les schèmes nominaux peuvent être obtenus à l'aide de la combinaison d'au plus deux syllabes à l'exception de s_2s_1 et s_3s_1. Il considère également que les noms arabes se terminent toujours par une consonne. Il a défini une syllabe additionnelle, notée s_0 pour représenter la dernière consonne.

À l'issu de cette terminologie, l'ensemble de schèmes nominaux est réduit à sept schèmes syllabiques :

(1) s_3s_0 : CVCC (*ʕilm*, "*savoir*")

(2) $s_1s_1s_0$: CVCVC (*ʕalam*, "*drapeau*")

(3) $s_1s_2s_0$: CVCVVC (*ʕuluwm*, "*savoirs*")

(4) $s_2s_1s_0$: CVVCVC (*šAmil*, "*exhaustif*")

(5) $s_2s_2s_0$: CVVCVVC (*jAmuws*, "*taureau*")

(6) $s_3s_1s_0$: CVCCVC (*maSnaʕ*, "*usine*")

(7) $s_3s_2s_0$: CVCCVVC (*jumhuwr*, "*public*")

Lexique Le lexique de noms est l'union de trois lexiques : racines, schèmes et vocalismes. Étant le schème nominal véhicule le nombre (singulier, pluriel), le lexique de schèmes précise pour chaque schème son nombre. Le lexique de racines associe pour chaque racine les listes de schèmes et de vocalismes compatibles à cette racine. Dans le lexique de vocalismes, chaque entrée distingue les voyelles du singulier et du pluriel.

(a) lexique de schèmes

1 s_3s_0 : CVCC
2 $s_1s_1s_0$: [CVCVC nombre=s]
3 $s_1s_2s_0$: [CVCVVC nombre=s]
4 $s_2s_1s_0$: [CVVCVC nombre=s]
5 $s_2s_2s_0$: [CVVCVVC nombre=s]
6 $s_3s_1s_0$: [CVCCVC nombre=s]
7 $s_3s_2s_0$: [CVCCVVC nombre=s]

c. McCarthy (1993) présente les voyelles longues aA, uw et iy (cf. section 1.1.2) comme une suite de deux voyelles courtes aa, uu et iy.

(b) lexique de racines

1 ʕlm : [$s_3 s_0$, sing_voyelle=i, plur_voyelle=u]
2 ʕlm : [$s_1 s_1 s_0$, sing_voyelle=a, plur_voyelle=a]
3 jmhr : [$s_3 s_2 s_0$, sing_voyelle=uu, plur_voyelle=ai]

(c) lexique de vocalismes

1 a : [sing_voyelle=ai]
2 ai : [sing_voyelle=ai]
3 au : [sing_voyelle=au]

Règles À l'image de verbes, chaque forme lexicale dans une règle est représentée sur trois bandes.

R1 * - X - * ↔ * - X - *
R2 * - C - * ↔ * - (s_0,C,) - (+„)
R3 * - CV - * ↔ * - (s_1,C,V) - *
R4 * - C_1VC_2 - * ↔ * - (s_3,C_1C_2,V) - (s_3,*,*)
R5 * - CVV - * ↔ * - (s_2,C,V) - (s_2,*,*)
R6 * - C_1VC_2 - * ↔ * - (s_3,C_1C_2,V) - (s_0,*,)
R7 * - CVV - * ↔ (S,*,*) - (s_2,C,V) - (s_0,*,) où S∈{s_1, s_2, s_3}
R8 * - - * ↔ * - (+„) - *
R9 * - CV - * ↔ (S,*,V) - (s_1,*,V) - * où S∈{s_1, s_2, s_3}
R10 * - CVV - * ↔ (S,*,V) - (s_2,*,V) - * où S∈{s_1, s_2, s_3}

La règle R1 représente l'identité, elle projette tout caractère lexical dans la première bande sur la forme de surface. R2 permet, étant donné la syllabe s_0 sur la première bande et une consonne C sur la deuxième bande de reproduire C sur la forme surfacique. Cette règle est obligatoire, elle concerne la syllabe finale du nom qui est toujours suivi d'un (+„) dans le contexte droit. R3 lit la syllabe s_1 sur la première bande, C sur la deuxième bande et V sur la troisième bande et écrit CV dans la forme de surface. D'une manière similaire, R4 et R5 permettent d'écrire CVC et CVV dans la forme de surface. Les contextes de droite (s_3,*,*) et (s_2,*,*) indique que ces règles sont appliquées aux premières syllabes. R6 et R7 sont analogues à R4 et R5 mais appliquées à la deuxième syllabe. Enfin, R8 permet de supprimer le signe '+'. R9 et R10 assurent la propagation de consonnes et de voyelles.

À titre d'illustration, nous donnons des exemples de génération des noms *ʕilm* "*savoir*", *ʕalam* "*drapeau*" et *jumhuur* "*public*" dans la figure 2.13.

s_3	s_3	+
ʕl	m	
i		

R4 R2 R8

ʕil	m	

s_1	s_1	s_0	+
ʕ	l	m	
a			

R3 R9 R2 R8

ʕa	la	m	

s_3	s_2	s_0	+
jm	h	r	
u	h	r	

R4 R10 R2 R8

jum	huu	r	

Figure 2.13.: Exemples de génération de noms basée sur des syllabes

2.3. Principaux analyseurs morphologiques de l'arabe

Dans cette section, nous exposons les principaux travaux portant sur l'analyse morphologique de la langue arabe. Le panorama d'analyseurs morphologiques de l'arabe dans cette section n'est pas exhaustif, néanmoins il permet d'établir les différentes méthodes et stratégies existantes. Afin de maintenir notre orientation basée sur une étude comparative entre l'approche concaténative et l'approche gabaritique, nous décrivons deux systèmes, parmi les plus répandus, de chaque approche. Chaque analyseur présente des innovations et des techniques plus sophistiquées par rapport à celui qui le précède.

2.3.1. Buckwalter Arabic Morphological Analyzer (BAMA)

L'analyseur morphologique BAMA (Buckwalter, 2002, 2004), est un système d'analyse morphologique qui adopte l'approche concaténative. Cet analyseur découpe un mot arabe en trois segments : un segment préfixal PREF (proclitiques et préfixe de flexion), un radical et un segment suffixal SUFF (suffixe de flexion et enclitique). BAMA définit un système de translittération et il réalise le traitement des mots translittérés.

Lexique BAMA contient trois lexiques : un lexique de radicaux, un lexique de PREF et un troisième lexique de SUFF. Un mot est obtenue suite à une concaténation d'un PREF, un radical et un SUFF. PREF et SUFF peuvent être nuls. Une catégorie contenant toutes les informations morphologiques et une glose en anglais sont assignées à chaque entrée lexicale. Un échantillon du lexique de BAMA est donnée dans la figure 2.14

PREF		
و/wa	Pref-Wa	*and*
ب/bi	NPref-Bi	*by/with*
و ب/wabi	NPref-Bi	*and + with/by*
ال/Al	Pref-Al	*the*
بال/biAl	Pref-BiAl	*with/by + Al*
و بال/wabiAl	Pref-WabiAl	*and + with/by + Al*
Radicaux		
كَتَب/katab	PV	*wrote*
كُتُب/kotub	IV	*write*
كُتِب/kutib	PV_Pass	*be written*
كتَب/ktab	IV_PASS	*be written*
PREF		
ة/ap	NSuff-ap	*[fem.sg]*
ات/At	NSuff-At	*[fem.pl]*
تان/atAni	NSuff-atAni	*two*

Figure 2.14.: Lexique de bama

Au total, le lexique de BAMA contient 82158 radicaux, 299 PREF et 618 SUFF.

Tables de compatibilité BAMA définit trois tables pour vérifier la compatibilité entre les morphèmes. Une table est implémentée pour chaque paire de morphèmes à savoir PREF/radical, radical/SUFF et PREF/SUFF. La table PREF/radical par exemple indique que l'article définie ال *Al*, n'est compatible qu'avec les radicaux de valeurs nominales.

Dans le cas où la décomposition d'un mot est valide selon le lexique et les tables de compatibilité, BAMA réalise leur concaténation et les ajustements nécessaires. Toutes les règles d'ajustement sont codées dans le lexique. Cet analyseur associe une étiquette contenant toutes les informations nécessaires appropriées au mot. À titre d'exemple, l'étiquette assignée au mot وسيةعمون *wasayutçimuwnkum* "*et ils vous nourriront*" à l'aide de BAMA est wa/CONJ+sa/PART+y/3MP+uTçim/IV+uwna/3MP.

2.3.2. Arabic Lexeme-based Morphological Generation and Analysis (ALMOR)

ALMOR est un analyseur morphologique fondé sur les bases lexicales de BAMA (Habash, 2004). En revanche, contrairement à BAMA qui se focalise uniquement sur l'analyse d'une forme surfacique, ALMOR permet également de générer la

représentation surfacique à partir de son analyse morphologique composée d'un lexème et des traits morphologiques. Par conséquent, ALMOR étend les étiquettes morphologiques de BAMA avec les traits morphologiques qui sont utilisés dans l'analyse et la génération.

L'analyse dans le système ALMOR est similaire à BAMA, le mot est décomposé à des triplets (PREF, lexème, SUFF). En revanche ALMOR rajoute une représentation intermédiaire en morphèmes abstraits qui relie la forme de surface à ses traits morphologiques.

ALMOR est utilisé dans l'analyseur morphosyntaxique MADA (Roth *et al.*, 2008; Habash *et al.*, 2009). Ce dernier réalise la segmentation, l'étiquetage grammatical, la diacritisation, la lemmatisation et l'analyse morphologique dans le même processus (Habash et Rambow, 2005). Étant donné un mot MADA se sert de ALMOR pour générer toutes les analyses potentielles d'un mot. Il associe ensuite des scores à ces analyses afin d'effectuer la désambiguïsation.

2.3.3. Xerox Finite State Machine (XFSM)

Contrairement à BAMA et ALMOR, XFSM (Beesley, 2001) est un analyseur morphologique basé sur des racines et des schèmes. Cet analyseur est construit à l'aide de machines à états finis. XFSM produit environ 85000 radicaux à partir d'un lexique de racines composé de 4930 racines et un ensemble de 400 schèmes. XFSM restitue les diacritiques de mots et leurs fournit des gloses en anglais. Ce système est bidirectionnel, il réalise l'analyse et la génération. Toutefois, XFSM à l'image des systèmes adoptant l'approche gabaritique permet de réduire la taille du lexique. En revanche, ce système connait un taux élevé d'ambiguïté, son système de règles complexe engendre beaucoup d'erreurs de sur-génération.

2.3.4. L'analyseur (ELIXIRFM)

ELIXIRFM (Smrž, 2007b,a) est basé sur des racines et des schèmes à l'image de XFSM. L'idée maîtresse qui a contribué au développement de ELIXIRFM, est la réduction du nombre de règles pour surmonter les problèmes de dérivation évoqués dans la section 2.1.3.

Smrž (2007a) a eu recours à la définition d'un ensemble de schèmes *artificiels*. Ces derniers ne font pas partie de l'ensemble de schèmes de l'arabe, ils permettent de simplifier le processus de dérivation irrégulière et servent à réduire considérablement le nombre de règles. À titre d'exemple, le verbe creux *AistaqAl* "*démissionner*" est dérivé de la racine *q w l* et le schème *Aista12a3*. Néanmoins, le croisement de telle racine et de tel schème génère la forme *Aistaqwal* (1) qui est différente à la forme de surface cible *AistaqAl* (2). Les systèmes morphologiques ont recours à une règle morpho-phonémique pour passer de (1) à (2). ELIXIRFM par contre introduit un schème *Aista1A3* qui définit parfaitement le format de *AistaqAl*. À l'aide de ce schème, le processus de dérivation devient régulier, il

est basé sur les substitutions des chiffres 1 et 3 du schème par la première et la troisième consonnes de la racine repectivement.

De la même manière, les déverbaux qui dérivent des verbes creux sont dérivés à l'aide du schème artificiel. Le déverbal *mustaqiyl*, par exemple est dérivé du schème *musta1iy3*.

Conclusion

Dans ce chapitre, nous avons présenté deux approches pour analyser la morphologie non-linéaire de l'arabe, l'approche concaténative qui se base sur la concaténation des clitiques et des affixes au radicaux. Un grand lexique de radicaux est ainsi utilisé avec moins de règles. L'approche gabritique se focalise sur le croisement des racines et des schèmes pour la génération des radicaux. Cette approche permet d'utiliser un lexique réduit basée sur les racines et une table de schèmes. L'inconvénient de cette approche réside dans le nombre énorme de règles qui devraient être définies. Ensuite, nous avons décrit les méthodes utilisées dans le traitement de la morphologie arabe plus particulièrement au niveau de la dérivation à savoir le formalisme de deux-niveau et le modèle basé sur les transducteurs multi-bande. Enfin, nous avons exposé quelques analyseurs morphologiques de l'arabe qui adoptent les différentes approches décrites.

3. Outils et ressources

Comme nous l'avons évoqué dans l'introduction, la stratégie que nous suivons pour réaliser le traitement automatique du TUN consiste à convertir ce dernier en une forme approximative du MSA (que l'on appelle pseudo-MSA). Le pseudo-MSA n'a pas pour vocation d'être compris par un être humain mais son traitement à l'aide d'un outil de TAL destiné au MSA fournit des résultats satisfaisants.

Notre système de conversion repose principalement sur les deux niveaux morphologique et lexical. En effet, nous avons eu recours à un analyseur morphologique du TUN, un lexique bilingue TUN-MSA et un générateur morphologique de MSA. Ces ressources sont décrites dans les sections 3.1 et 3.2.

Suite à la conversion, nous avons utilisé un étiqueteur en parties de discours entraîné sur le MSA pour l'étiquetage du pseudo-MSA. Cet outil est décrit dans la section 3.3. Enfin, le corpus TUN utilisé dans l'évaluation de notre méthode est décrit dans la section 3.4.

3.1. Système d'analyse et génération morphologique du MSA et de ses dialectes

Pour réaliser l'analyse morphologique du TUN et la génération morphologique du MSA, nous avons eu recours au système MAGEAD [a] (Habash *et al.*, 2005; Habash et Rambow, 2006).

Trois raisons principales nous ont poussé à utiliser ce système. Premièrement, MAGEAD est conçu pour le traitement morphologie de toutes les variantes de l'arabe, le MSA aussi bien que les dialectes. Deuxièmement, ce système réalise une analyse sous la forme de racines et de schèmes. Ce qui permet de décrire la conversion TUN→ MSA au niveau des racines et des schèmes. Troisièmement, MAGEAD est bi-directionnel, cela est important dans la perspective de conversion puisqu'il permet d'analyser une variante et de générer une autre variante. Dans notre cas, MAGEAD sert à analyser du TUN et à générer du MSA.

a. MAGEAD est l'acronyme de *Morphological Analyzer and Disambiguator for Arabic and its Dialects*.

3.1.1. Analyse et génération morphologique

MAGEAD utilise un modèle à deux niveaux (cf section 2.2.1) qui relie une forme lexicale composée d'une racine, d'un schème et d'un ensemble de traits morphologiques à une forme surfacique à travers un série de transformations. Ces transformations sont assurées par un transducteur (cf. figure 3.1). La forme surfacique يضطرّون *yaDTar∼uwna* "*ils s'obligent*", par exemple, est associée à la représentation profonde :

[ROOT:Drr][MBC:verb-VIII][ASP:i][MOD:i][VOX:a][PER:3][GEN:m][NUM:p]

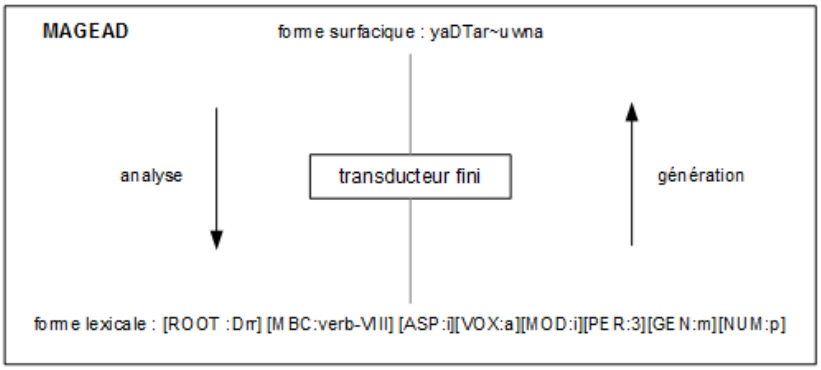

Figure 3.1.: Représentation simplifiée de magead

Dans le sens de la génération, les traits morphologiques sont d'abord traduits en morphèmes abstraits. Ces morphèmes sont ensuite ordonnés et traduits en morphèmes concrets. Ces derniers sont combinés entre eux pour générer une forme surfacique qui peut éventuellement subir des changements morpho-phonémiques et des ajustements orthographiques grâce à un ensemble de règles.

Pour relier la forme lexicale et la forme surfacique, MAGEAD définit quatre niveaux de représentation. Nous les décrivons ci-dessous dans le sens de la génération en nous appuyant sur l'exemple précédent : يضطرّون *yaDTar∼uwna*.

Représentation profonde : Ce niveau de représentation est identique pour toutes les variantes de l'arabe. Une forme lexicale est représentée sous la forme d'une racine (ROOT), d'une catégorie, appelée classe de comportement morphologique, notée MBC (pour *Morphologic Behavioural Class*) et de traits morphologiques non ordonnés. À ce niveau, notre exemple se présente sous la forme suivante :

[ROOT:Drr][MBC:verb-VIII][ASP:i][MOD:i][VOX:a][PER:3][GEN:m][NUM:p]

Les traits morphologiques utilisés dans cet exemple indiquent dans l'ordre les valeurs morphologiques de l'aspect (ASP), du mode (MOD), de la voix (VOX), de la personne (PER), du genre (GEN) et du nombre (NUM). MAGEAD utilise en outre quatre traits QST, CNJ, PRT et PRN pour la détermination des clitiques. Ils concernent respectivement les clitiques d'interrogation et de conjonction, les particules et les pronoms d'objet direct. Ils sont à valeurs booléennes et indiquent la présence ou l'absence d'un clitique dans le verbe.

Pour l'analyse des noms, MAGEAD utilise 8 traits morphologiques : GEN, NUM, STT, CAS, QST, CNJ, PRP et PRN. Les quatre premiers traits définissent respectivement les valeurs du genre, du nombre, d'état et du cas (accusatif, nominatif, génitif). Alors que les quatre derniers traits déterminent les clitiques rattachés à une forme nominale (interrogation, conjonction, préposition et pronom possessif).

MAGEAD définit 66 MBCs pour les verbes MSA parmi lesquelles 25 sont abstraites, utilisées uniquement pour des raisons d'organisation de la hiérarchie. Contrairement à l'analyse des verbes qui utilisent des systèmes de flexion et de dérivation réguliers, l'analyse des noms présentent de nombreuses irrégularités parmi lesquelles le pluriel brisé et les multiples pluriels. Le pluriel du mot مفتاح *miftAH* "*clef*", par exemple, repose sur le schème *ma1A2iy3* (مفاتيح *mafAtiyH*). Le mot كاتب *kAtib* "*écrivain*" possède trois formes différentes au pluriel : une forme régulière كاتبون *kAtibuwn* "*écrivains*" basée sur le rattachement du suffixe *uwn* à la forme au singulier, deux pluriels brisés كتبة *katabaħ* "*scribes*" et كتّاب *kut∼Ab* "*auteurs*". MAGEAD définit 962 MBCs pour les noms.

Représentation en morphèmes abstraits :

A ce niveau de représentation, une forme lexicale se présente sous la forme d'une séquence non ordonnée de morphèmes abstraits. Notre exemple se présente de la façon suivante :

```
[SUBJPREF_IV:3MP] [ROOT:Drr][PAT_IV:VIII][VOC_IV:VIII-act]
[SUBJSUF_IV:3MP_ind]
```

Le premier morphème ([SUBJPREF_IV:3MP]) correspond à un préfixe véhiculant la personne (3), le genre (M) et le nombre (P) du sujet. Les trois morphèmes qui suivent décrivent la racine ([ROOT:Drr]), le schème dépourvu de diacritiques ([PAT_IV:VIII]) et le vocalisme ([VOC_IV:VIII-act]). L'ensemble de ces trois morphèmes définissent un radical. Le dernier morphème ([SUBJSUF_IV:3MP_ind]) décrit un suffixe indiquant le mode (indicatif) ainsi que le genre, le nombre et la personne du sujet.

Le passage de la représentation profonde à la représentation en morphèmes abstraits est réalisée à l'aide de règles. Ces dernières permettent d'associer des

traits morphologiques à des morphèmes abstraits. On trouve, en partie gauche de telles règles un ou plusieurs traits et, en partie droite, un morphème abstrait.

À titre d'exemple, les traits morphologiques `[MOD:i]` `[ASP:i]` `[PER:3]` `[GEN:m]` `[NUM:p]` donnent naissance aux deux morphèmes abstraits `[SUBJPREF_IV:3MP]` et `[SUBJSUF_IV:3MP_ind]` (le circonfixe de la troisième personne du masculin pluriel) grâce aux règles suivantes :

`[ASP:i][PER:3][GEN:m][NUM:p]` ↔ `[SUBJPREF_IV:3MP]`
`[MOD:i][ASP:i][PER:3][GEN:m][NUM:p]` ↔ `[SUBJSUF_IV:3MP_ind]`

Les règles à appliquer à une représentation profonde dépendent du MBC. Un MBC peut donc être vu comme un ensemble de règles : les règles à appliquer pour la conversion en morphèmes abstraits.

Les MBCs sont organisées sous la forme d'une hiérarchie, elles héritent de leurs MBCs ancêtres un certain nombre de règles. C'est cette représentation hiérarchique qui permet de factoriser des règles communes à plusieurs MBC. La figure 3.2 donne une représentation simplifiée de la hiérarchie de classes morphologiques.

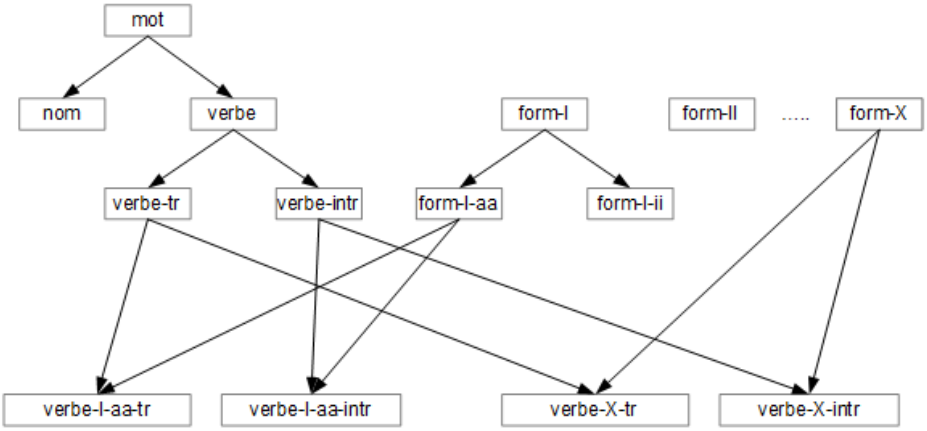

Figure 3.2.: Hiérarchie de classes de comportement morphologique

Le nœud de plus haut niveau de la hiérarchie est appelé mot. C'est à ce niveau que sont représentées les règles qui sont partagées par tous les mots arabes. On y trouve, par exemple, la règle associée au trait morphologique `[CONJ:w]`. Ce dernier correspond à la conjonction +و *w*+ "*et*". Ainsi, tous les mots des variantes de l'arabe acceptent ce clitique.

D'une façon analogue, tous les verbes transitifs qui correspondent à la MBC `Verbe-tr` et quels que soient leurs schèmes partagent les mêmes enclitiques pronominaux.

Dans notre cas, les 4 règles définies au niveau de la MBC `MBC:Verb-VIII` sont :

- [MBC:verb-VIII][ASP:i] ↔ [PAT_IV:VIII]
- [MBC:verb-VIII][ASP:p] ↔ [PAT_PV:VIII]
- [MBC:verb-VIII][ASP:i][VOX:a] ↔ [VOC_IV:VIII-act]
- [MBC:verb-VIII][ASP:p][VOX:p] ↔ [VOC_PV:VIII-pas]

Représentation en morphèmes concrets :

À ce niveau de représentation, une forme lexicale se présente comme une séquence ordonnée de morphèmes concrets. Notre exemple se représente de la façon suivante :

```
y + [Drr, VCtVCVC, aaa] + uwna
```

le circonfixe y+ +uwna indique la personne, le genre et le nombre du sujet. Le triplet [Drr, VCtVCVC, iaa] regroupe les trois composantes du radical : la racine, le schème et le vocalisme. Ils vont permettre de générer le radical proprement dit.

Le passage de la représentation morphématique abstraite à la représentation morphématique concrète est réalisée par une grammaire hors-contexte qui permet d'ordonner les morphèmes abstraits entre eux ainsi que par des règles de correspondance qui apparient morphèmes abstraits et morphèmes concrets.

La grammaire hors-contexte qui précise l'ordre des morphèmes abstraits est commune à toute les variantes de l'arabe à de rares exceptions près, comme on le verra dans le section 3.1.3.

La structure d'un mot selon MAGEAD est défini par la règle suivante :

```
[WORD] → [CONJ]? ( [NOUN] | [VERB] )
```

où le morphème [CONJ] correspond à une conjonction. Cette grammaire indique qu'une conjonction peut être rattachée à tous les mots et qu'elle se situe en début de mot. La structure du verbe [VERB] et du nom [NOUN] sont décrits dans les règles suivantes :

```
[NOUN] → [PREP]? [ART]? [INFLECTED_NOUN] [POSS]?
[VERB] → ( [PV_VERB] | [IV_VERB] | [CV_VERB] ) [OBJ]?
```

Les morphèmes [PREP], [ART] et [POSS] définissent respectivement une préposition, l'article défini et un pronom possessif qui peuvent être concaténé à une forme fléchie d'un nom ([INFLECTED_NOUN]). [PV_VERB], [IV_VERB] et [CV_VERB] indiquent des verbes fléchies dans l'aspect accompli, inaccompli et impératif. Ils peuvent tous prendre un pronom d'objet [OBJ].

Les règles de correspondance apparient morphèmes abstraits et morphèmes concrets. Elles ont pour partie gauche un morphème abstrait et pour partie droite un morphème concret. Ces règles sont généralement spécifiques à une variante de l'arabe. En MSA, par exemple, les morphèmes [SUBJPREF_IV:3MP] et [SUBJSUF_IV:3MP_IND] sont associés aux morphèmes concrets y+ +uwna à l'aide des règles suivantes :

```
[SUBJPREF_IV:3MP] ↔ y+
[SUBJSUF_IV:3MP_IND] ↔ +uwna
```

MAGEAD définit 92 règles de correspondance pour les verbes MSA. Elles concernent les clitiques et les affixes. Parmi ces règles, 3 sont utilisées pour les conjonctions, 6 pour les particules, 18 pour les suffixes sujet, 52 pour les préfixes sujet et 13 pour les pronoms objet. Concernant les noms, MAGEAD définit 359 règles.

Représentation de surface : Il s'agit de la représentation orthographique. Notre exemple se représente maintenant sous la forme

yaDTar∼uwna

qui est la translittération de la forme arabe يضطرّون. Le passage de la représentation en morphèmes concrets à la représentation de surface met en jeu trois types d'opérations. La combinaison d'une racine, d'un schème et d'un vocalisme, la concaténation des affixes et les règles d'ajustement orthographiques. Ces opérations sont réalisées par deux types de règles, les règles morpho-phonémiques et le règles orthographiques.

Ces règles sont compilées sous la forme d'un automate à 5 bandes. Les trois premières bandes sont des bandes de lecture, la quatrième bande est une bande de lecture écriture et la cinquième, une bande d'écriture. Sur la première bande est écrit le schème ainsi que les affixes et les clitiques. La deuxième bande est réservée à la racine et la troisième au vocalisme. La quatrième bande concerne la représentation morpho-phonémique et la cinquième bande la représentation orthographique.

La table 3.1 représente l'état des bandes avant application des règles dans le cas de notre exemple. Les bandes 4 et 5 sont à ce stade garnies de 0.

schème	y	+	V	C	t	V	C	V	C	+	u	w	n	a
racine				D			r		r					
vocalisme			a			a		a						
forme morpho-phonémique	0	0	0	0	0	0	0	0	0	0	0	0	0	0
forme orthographique	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 3.1.: État des bandes de l'automate avant application des règles

La première étape consiste à appliquer les règles morpho-phonémiques. Ces dernières lisent sur les bandes 1, 2 et 3 et écrivent sur la bande 4. À l'issue de cette étape, notre exemple `y+ [Drr, VCtVCVC, aaa] + uwna` est traduit en `y+aDTar0r+uwna` à l'aide des règles suivantes :

(1) `[X, , ,0]` → `X, X ∉ {C,V};`

(2) `[C,X, ,0]` → `X;`

(3) `[V, ,X,0]` → `X;`

(4) `[V, ,V,V]` → `0 / [2,%, ,X] _ [3+S,%, ,X+S] , S=[VOWEL];`

(5) `[t, , ,t]` → `T / [1,M, ,M] _ , M ∉ {STDZ};`

La première règle consiste à placer tous les symboles composant les affixes de la bande du schème sur la bande de la forme morpho-phonémique. La deuxième et la troisième règles permettent de remplacer les symboles C et V du schème par les consonnes de la racine et les voyelles du vocalisme. La quatrième règle est une règle géminative[b]. Elle permet de supprimer la voyelle située entre le deuxième et le troisième radical si le suffixe commence par une voyelle. Enfin, la cinquième règle provoque le voisement du son /t/ en /T/. Cette règle est appliquée uniquement lorsque la première lettre de la racine correspond à /S/, /D/, /Z/ ou /T/.

À l'issue de cette étape, les bandes de l'automate sont dans l'état représenté dans la table 3.2.

schème	y	+	V	C	t	V	C	V	C	+	u	w	n	a
racine				D			r		r					
vocalisme			a			a		a						
forme morpho-phonémique	y	+	a	D	T	a	r	0	r	+	u	w	n	a
forme orthographique	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 3.2.: Etat des bandes de l'automate après application des règles morpho-phonémiques

À l'issue de l'application des règles morpho-phonémiques, les règles orthographiques sont appliquées. Ces dernières lisent sur les 4 premières bandes et écrivent sur la cinquième. Dans notre exemple, cette étape génère `yaDTar∼uwna` à partir de `y+aDtar0r+uwna` à l'aide des règles suivantes :

(1) `[X, , ,X,0]` → `X, X ∉ {C,V,+};`

(2) `[C,X, ,X,0]` → `X;`

(3) `[V, ,X,X,0]` → `X;`

(4) `[+, , ,+,0]` → `X;`

(5) `[V, ,V,0,0]` → `∼;`

b. La gémination correspond au fait que la deuxième et la troisième lettres de la racine sont identiques.

À l'image des règles morpho-phonémiques, les trois premières règles permettent de projeter les caractères des quatre premières bandes sur la bande orthographiques. Les deux dernières permettent respectivement de supprimer le signe + et de remplacer le symbole 0 entre deux consonnes identiques par le caractère ∼. Ainsi la forme la forme `yaDTar∼uwna` est finalement générée. L'état des bandes de l'automate à ce stade apparaît dans la figure 3.3.

schème	y	+	V	C	t	V	C	V	C	+	u	w	n	a
racine				D			r		r					
vocalisme			a			a		a						
forme morpho-phonémique	y	+	a	D	T	a	r	0	r	+	u	w	n	a
forme orthographique	y	0	a	D	T	a	r	0	∼	0	u	w	n	a

Table 3.3.: Etat des bandes de l'automate après application des règles orthographiques

Les variétés de l'arabe se distinguent par certaines règles morpho-phonémiques et orthographiques mais en partagent d'autres. Pour le traitement des verbes MSA, par exemple, MAGEAD définit 69 règles morpho-phonémiques et 53 règles orthographiques. Au niveau des noms, 79 règles morpho-phonémiques et 77 règles orthographiques sont implémentées.

Dans la suite de cette section, nous décrivons l'architecture de MAGEAD (cf. section 3.1.2) et le processus d'adaptation de MAGEAD au TUN (cf. section 3.1.3).

3.1.2. Architecture de MAGEAD

Pour réaliser le traitement morphologique d'une variante de l'arabe, MAGEAD utilise des ressources (base lexicale et un ensemble de règles), un compilateur et un système morphologique. L'architecture générale de MAGEAD est donnée dans la figure 3.3. Cette figure est extraite de Altantawy *et al.* (2010).

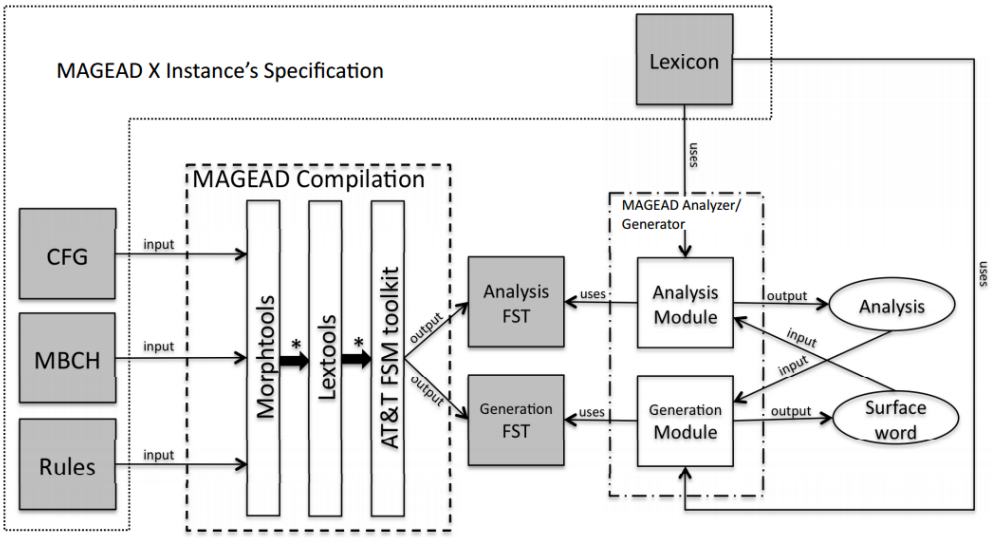

Figure 3.3.: Architecture de magead

La construction d'une instance de MAGEAD passe par deux phases principales. La première (encadrée par une forme de *L-renversé* dans la figure 3.3) consiste à créer manuellement les ressources linguistiques utilisées pour une instance spécifique de MAGEAD. Ces ressources sont :

- Un lexique de racines (`Lexicon`)
- Une hiérarchie de classes morphologiques (`MBCH`)
- Une grammaire hors-contexte qui ordonne les morphèmes abstraits (`CFG`)
- Des règles morpho-phonémiques et orthographiques ainsi que des règles associant les morphèmes abstraits aux morphèmes concrets correspondants (`Rules`)

La deuxième phase (représentée dans la figure 3.3 dans un rectangle avec des traits interrompus) est la compilation des ressources linguistiques d'une instance de MAGEAD pour produire deux transducteurs : un pour la génération et un autre (son inverse) pour l'analyse morphologique.

Ces transducteurs permettent de produire la représentation morphologique profonde à partir d'une forme de surface ou l'inverse.

Plusieurs raisons ont poussé les développeurs de MAGEAD à utiliser cette architecture. D'abord, comme nous l'avons évoqué, MAGEAD est conçu pour réaliser le traitement des morphologies du MSA et ses dialectes dans un même système. Ainsi, cette architecture offre la possibilité d'adapter MAGEAD sur une variante arabe en se limitant uniquement à la création des ressources spécifiques à cette variante. En outre, l'architecture de MAGEAD permet d'exploiter certaines régularités partagées par les variantes de l'arabe.

Compilation de MAGEAD : La compilation d'une instance spécifique de MAGEAD se déroule en trois étapes successives. Premièrement, la grammaire, les MBCs et les règles sont représentées dans le format *morphtools* Deuxièmement, le format *morphtools* est compilé pour générer un format *lextools* qui représente une extension des outils *AT&T* (Mohri *et al.*, 2000) pour les machines à états finis (Sproat, 1995). Le format *lextools* est par la suite compilé pour produire les transducteurs désirés.

3.1.3. Adaptaion de MAGEAD au TUN

L'adaptation de MAGEAD à une variante de l'arabe consiste à créer manuellement les ressources linguistiques nécessaires à MAGEAD pour le traitement de cette variante. Pour cela, nous avons défini toutes les entités (traits, MBCs, morphèmes abstraits et concrets) spécifiques au TUN ainsi que les règles qui mettent en correspondance ces entités.

Hiérarchie de classes de comportement morphologique `MBCH` :
Bien que les MBCs sont sensées être valables pour toutes les variantes de l'arabe, nous avons dû néanmoins étendre la hiérarchie pour y inclure les morphèmes abstraits du TUN qui n'existent pas du côté MSA. Pour la hiérarchie de verbes TUN, seul le morphème abstrait `[NEG:š]` a été ajouté. Il correspond à l'enclitique de négation ش+ *+š*.

Au niveau de la hiérarchie des noms, les morphèmes abstraits qui correspondent aux prépositions +ع ʿ+, +م *m*+, +ه *h*+ et +ل *l*+ sont ajoutés. Ces prépositions sont réalisées d'une manière autonome en MSA.

Grammaire hors-contexte : Cette grammaire a été étendue suite à l'ajout du nouveau morphème abstrait. La position du morphème `QST` qui était en début de verbe MSA devient à la fin d'un verbe TUN. L'ordre des morphèmes abstraits d'un mot TUN, selon sa catégorie, est donné par les structures suivantes :

```
CNJ PRT PREF VERB SUFF OBJ NEG|QST
CNJ PREP PREF NOUN SUFF POS
```

Morphèmes abstraits ↔ morphèmes concrets : Nous avons construit les règles de correspondance des morphèmes abstraits et des morphèmes concrets TUN. L'ensemble des règles peut être considéré comme une table de deux colonnes où la première colonne contient les morphèmes abstraits et la deuxième colonne concernent les morphèmes concrets. Nous avons édité la table des morphèmes MSA. Mis à part le nouveau morphème abstrait TUNque nous avons ajouté dans la table, nous avons maintenu tous les morphèmes abstraits MSA

de la première colonne du côté TUN. Les morphèmes abstraits qui n'existent pas en TUN sont mis en correspondance avec des symboles nuls.

La majorité des changements ont été réalisés sur la deuxième colonne. En effet, la plupart des morphèmes concrets se réalisent différemment de ceux du MSA. Le préfixe sujet de la première personne du singulier, par exemple, se réalise comme +أ*Âa*+ en MSA et +ن *na*+ en TUN.

Ainsi, 16 morphèmes concrets ont été édités. Au total, une table de 28 règles est définie pour le TUN (cf. annexe A).

Règles morpho-phonémiques et orthographiques : La flexion en TUN est totalement différente de celle de MSA. Par conséquent, nous avons implémenté 94 règles morpho-phonémiques et 81 règles orthographiques spécifiques au traitement TUN. En TUN, par exemple, si la troisième lettre de la racine correspond à و *w* ou ي *y*, elle est remplacée par la voyelle longue أ *A* lorsque le suffixe sujet commence par la voyelle fermée /u/ ou /i/ (ce qui est le cas pour la troisième personne du singulier féminin et la troisième personne du pluriel). Le verbe مشى *mšayý* conjugué à la troisième personne du singulier féminin donne مشات *mšAt* alors qu'à la troisième personne du pluriel il donne مشاوا *mšAwA*.

D'autres règles définies en MSA sont éditées. La règle de gémination, par exemple, permet d'élaguer la voyelle entre la première et la deuxième lettres d'une racine verbale du côté MSA s'il est suivi par un suffixe qui commence par une voyelle. En revanche, du côté TUN, la gémination est appliquée toujours indépendamment du suffixe : مدّيت $mad\sim+iyt$ "*j'ai étendu*" et مدّت $mad\sim+it$ "*elle a étendu*".

Une particularité qui caractérise les verbes sains TUN de la forme CVCVC consiste à élaguer toujours une des deux voyelles du verbe. La voyelle élaguée dépend du suffixe sujet du verbe. Dans le cas où le suffixe commence par une consonne (première et deuxième personne), la forme du schème devient CCVC. Ce dernier prend la forme CVCC dans le cas où le suffixe commence par une voyelle (troisième personne). À titre d'illustration, la conjugaison du verbe TUN كتب *ktib* "*écrire*" est donné dans la table 3.4. Nous indiquons entre parenthèses la forme du schème.

	singulier		pluriel	
	masculin	féminin	masculin	féminin
1ère personne	ktibt CCVC+t		ktibnA CCVC+nA	
2ème personne	ktibt(iy) CCVC+t(iy)		ktibtuwA CCVC+tuwA	
3ème personne	ktib CCVC+0	kitbit CVCC+it	kitbuwA CVCC+uwA	

Table 3.4.: Conjugaison d'un verbe sain tun dans l'aspect accompli

Des diverses autres règles sont implémentées pour le TUN. Par exemple, le premier radical est remplacé par la voyelle longue ا *A* dans l'aspect inaccompli quand il correspond à ء *'* (*hamza*). Ainsi, la forme (يأكل $yÂkl$ devient ياكل $yAkl$ "*il mange*". D'une manière similaire, les verbes qui terminent par la lettre ء *'* se comporte de la même façon que les verbes pour lesquels le lettre finale de la racine est ي y dans l'aspect accompli. Les racines des verbes TUN بدينا $bdynA$ "*nous avons commencé*" et رمينا $rmynA$ *nous avons jeté* sont respectivement ب د ء $b\ d\ '$ et ر م ي $r\ m\ y$. Au total, nous avons défini 89 règles morpho-phonémiques et orthographiques pour l'analyse des verbes et des noms TUN.

Nous décrivons en détail dans l'annexe de ce manuscrit la hiérarchie de MBCs, la grammaire, les morphèmes (abstraits et concrets) et les règles (morpho-phonémiques et orthographiques) que nous avons définis pour le traitement du TUN.

3.2. Lexiques de transferts TUN→MSA

Le transfert lexical du TUN vers le MSA est assuré par trois ressources lexicales : un lexique de verbes, un lexique de noms et un lexique de particules. Le lexique des verbes ainsi que celui des particules sont construits de façon semi-automatique alors que le lexique des noms a été entièrement créé automatiquement. Il a été généré à partir du lexique des verbes et couvre uniquement des noms déverbaux.

3.2.1. Lexique des verbes

Le lexique des verbes est composé de paires (P_{MSA}, P_{TUN}) dont les éléments P_{MSA} et P_{TUN} sont des triplets (racine, MBC, lemme). Chaque entrée lexicale est enrichie de deux gloses, une en anglais et l'autre en français. Le lexique est composé de 1638 entrées et couvre 1478 lemmes verbaux distincts du côté MSA

et 920 lemmes différents TUN. Le tableau 3.5 donne deux exemples d'entrées extraites du lexique.

	TUN	MSA	ANG	FRA
racine	x l S	s d d	*to reimburse*	*rembourser*
MBC	II-ii	II		
lemme	xalliS	saddad		
racine	H l l	f t H	*to open*	*ouvrir*
MBC	I-ai	I-aa		
lemme	Hall	fataH		

Table 3.5.: Deux exemples d'entrées du lexique des verbes

Le lexique a été construit à partir du corpus ATB (Maamouri *et al.*, 2004) qui est composé de transcriptions d'émissions d'actualités en MSA diffusées par différentes chaînes de télévision arabes. Ce corpus comporte 29911 occurrences verbales qui correspondent à peu près à 1500 verbes différents. Ces occurrences sont des formes fléchies dont la racine et le schème ont été extraits afin de servir à la construction du lexique.

L'extraction des racines et des schèmes à partir des formes fléchies, a été réalisée grâce à l'analyseur morphologique ELIXIRFM (cf. section 2.3.4) qui permet, étant donné une forme fléchie en MSA, d'en extraire le lemme, la racine et le schème. Deux raisons principales nous ont conduit à choisir ELIXIRFM. D'une part, cet analyseur réalise des analyses profondes qui permettent de générer la racine et le schème d'un mot. D'autre part, contrairement à MAGEAD, ELIXIRFM produit explicitement des lemmes.

Suite aux analyses produites par ELIXIRFM, chaque occurrence d'un lemme MSA est traduite manuellement, en contexte, vers un lemme TUN. À ce stade, les entrées du lexique sont composées du côté MSA d'un lemme, d'une racine et d'un schème et uniquement d'un lemme du côté tunisien. Nous avons alors remplacé les schèmes d'ELIXIRFM par les MBC de MAGEAD correspondantes. Cette correspondance est généralement immédiate. Du côté TUN, nous avons assigné à chaque lemme un schème et une racine.

Lorsqu'un verbe tunisien ne correspond à aucune racine MSA, ce qui arrive dans $8,5\%$ des cas, une nouvelle racine TUN est créée. La création de la nouvelle racine TUN est réalisée grâce à une méthode déductive. En effet, étant donné l'équation `racine` $\times$ `schème = lemme`, lorsque nous disposons d'un lemme et d'un schème, il est possible d'en déduire une racine. Une centaine de nouvelles racines spécifiques au TUN ont ainsi été créées. La liste de ces racines est donnée dans l'annexe B de ce manuscrit.

À titre d'exemple, le lemme TUN نقّز *naq∼iz* "*sauter*" correspond au schème TUN *1a2∼i3*. L'équation dans ce cas est donc `racine` $\times$ *1a2∼i3* `=` *naq∼iz* . Cette équation admet une solution unique qui est `racine =` *n q z*.

Dans certains cas, l'équation admet plusieurs solutions. C'est notamment le cas du verbe TUN استّنى*Aistannaý* "*attendre*" auquel correspondent le quatre solutions suivantes :

1. la racine س ن ي *s n y* et le schème *Ai1ta2a3*
2. la racine ن ن ي *n n y* et le schème *Aista12a3*
3. la racine أ ن ي *' n y* et le schème *Aista12a3*
4. la racine أ ن ن *' n n* et le schème *Aista12a3*

Dans ces cas ambigus, la stratégie que nous avons suivie consiste à privilégier la paire contenant une racine de MSA. Ainsi, la troisième alternative, dans l'exemple précédent, est choisie. En effet, parmi les quatre racines *s n y*, *n n y*, *' n y* et *' n n*, seule la racine *' n y* existe en MSA, elle correspond à la notion de "*circonspection*".

Un échantillon du lexique apparaît dans la table 3.6. Les exemples donnés illustrent les variations lexicales (racines) et celles morphologiques (schèmes et MBC) entre le TUN et le MSA.

msa		**tun**		**glose française**
racine	**mbc/Schème**	**racine**	**mbc/Schème**	
Smt	I-aa / 1a2a3	skt	I-ii / 12i3	*se taire*
Hlq	I-aa / 1a2a3	Hjm	II-ii / 1a22i3	*coiffer*
rtb	II / 1a22a3	nZm	II-ii / 1a22i3	*organiser*
Hlq	II / 1a22a3	Tyr	I-a / 12a3	*voler/décoller*
xSm	III / 1A2a3	ϛrk	III-ii / 1A2i3	*disputer*
dhm	III / 1A2a3	hjm	I-ii / 12i3	*attaquer*
bhr	IV / Aa12a3	ϛjb	I-ii / 12i3	*être admiré*
xfy	IV / Aa12a3	xby	II-ai / 1a22a3	*cacher*
ršf	V / ta1a22a3	ršf	V-ii / t1a22i3	*déguster*
ϛjb	V / ta1a22a3	bht	I-ii / 12i3	*s'étonner*
šjr	VI / ta1A2a3	ϛrk	VI / t1A2i3	*se disputer*
ϛfy	VI / ta1A2a3	bry	I-aa / 12a3	*guérir*
xfD	VII / Ain1a2a3	nqS	I-uu / 12u3	*se réduire*
sHb	VII / Ain1a2a3	bTl	II-ii / 1a22i3	*démissionner/être éliminé*
nhy	VIII / Ai1ta2a3	kml	I-ii / 12i3	*finir*
Hdn	VIII / Ai1ta2a3	Hml	II-ii / 1a22i3	*porter*
dϛy	X / Aista12a3	ϛdy	X / Aista12a3	*inviter*
wfy	X / Aista12a3	kml	II-ii / 1a22i3	*compléter*

Table 3.6.: Échantillon du lexique des verbes TUN-MSA

Comme le montre la table 3.6, les verbes TUN et MSA du lexique peuvent être complètement différents comme le verbe MSA استوفى *Aistawfaý* "*compléter*" et le verbe TUN كمّل *kam∼il*. Ils peuvent partager une même racine, une même MBC ou le couple entier comme le verbe استدعى *Aistadʕaý* "*inviter*". La table montre, en outre, qu'une MBC du côté MSA peut correspondre à plusieurs MBC du côté TUN et vice-versa. Les entrées du lexique ont une racine identique du côté MSA et du côté TUN dans plus de 300 cas. Elles partagent, dans 193 cas le même LMM. Ce dernier représente la forme non diacritée du lemme. En d'autres termes, ces entrées possèdent la même racine et la forme non diacritée du schème. Enfin, 16 partagent le même lemme c'est à dire la racine et le schème sont identiques des deux côtés.

Dans sa forme actuelle, le lexique est composé de 1638 entrées. Du côté tunisien, le lexique couvre 920 lemmes verbaux TUN distincts et 1478 du côté MSA. Cette différence illustre l'ambiguïté lexicale que provoque le passage du TUN vers le MSA. En moyenne, un lemme TUN correspond à $1,8$ lemmes MSA et $1,1$ dans le sens inverse. L'ambiguïté est donc plus importante dans le sens TUN $\rightarrow$ MSA. Plus précisément, dans $63,8\%$ de cas, un lemme TUN correspond à un seul lemme MSA. Ce taux s'élève à $90,9\%$ dans le sens inverse. Le tableau 3.7 illustre la répartition de l'ambiguïté lexicale dans les deux sens. Nous donnons, pour chaque sens, la proportion des verbes sources correspondant à un nombre donné de verbes cibles.

Nombre de verbes cibles	1	2	3	4	≥ 5
TUN $\rightarrow$ MSA	48.2%	36.1%	9.2%	2.7%	3.4%
MSA $\rightarrow$ TUN	66.9%	27.6%	5.3%	0.2%	0%

Table 3.7.: Ambiguïté dans le lexique des verbes

Les taux d'ambiguïté donnés dans la table 3.7 confirment que le passage du TUN vers le MSA est plus ambigu que le passage inverse. Environ 702 (66.9%) verbes MSA correspondent à un seul verbe du côté TUN. Dans le sens inverse, 31 (3.4%) verbes TUN admettent plus de cinq verbes cibles du côté MSA. Ce nombre est égal à 0 dans le sens MSA $\rightarrow$ TUN.

L'ambiguïté maximale est égale à 16 dans le sens TUN $\rightarrow$ MSA et à 4 dans le sens opposé. Comme le montre la figure 3.7, le lemme TUN عمل *ʕmal*, par exemple, peut se traduire par 16 formes cibles du côté MSA. Le verbe MSA جمع *jamaʕ* de l'autre côté correspond à quatre lemmes TUN distincts.

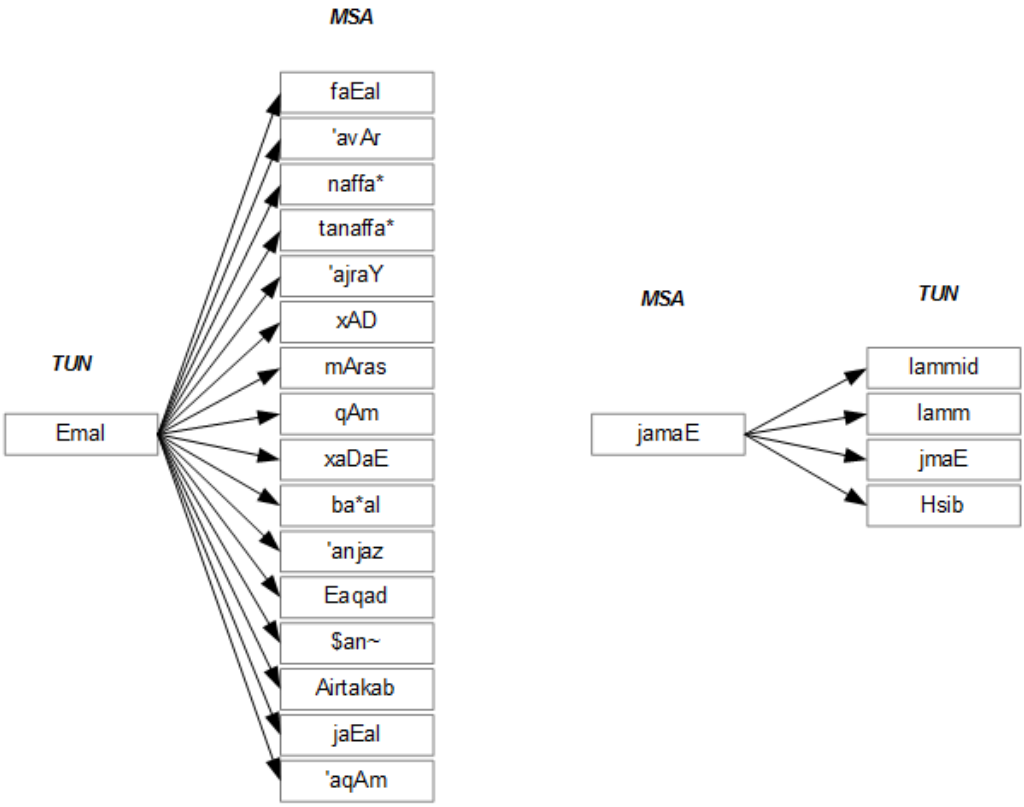

Figure 3.4.: Ambiguïté maximale entre verbes TUN et MSA

Lors du processus de conversion du TUN vers le MSA, la traduction d'un verbe TUN peut échouer à cause de l'absence du couple (racine, MBC) TUN, dans le lexique, bien que la racine existe soit présente. Afin d'augmenter la couverture du lexique des verbes, nous en avons extrait deux ressources différentes : un lexique de racines et une matrice de correspondance de MBCs.

Dans le cas où le couple TUN (racine, MBC) n'existe pas dans le lexique, la racine et la MBC seront traduites séparément. Nous montrerons, dans le chapitre suivant, que cette stratégie permet d'améliorer considérablement les performances de la conversion.

Lexique de racines

Ce lexique est constitué de couples de la forme (r_{MSA}, r_{TUN}), où r_{MSA} est une racine MSA et r_{TUN} une racine TUN. Cette ressource contient 1329 entrées correspondant à 1050 racines distinctes côté MSA et 646 côté TUN. Le lexique comporte 519 entrées composées d'une racine identique des deux côtés TUN et MSA.

Ce lexique montre deux points importants : d'une part, comme dans le lexique des verbes, l'ambiguïté est plus élevée dans le sens TUN → MSA. D'autre part, l'ambiguïté des racines est plus élevée que celle des lemmes verbaux. En moyenne, une racine TUN est reliée à $2,1$ racines MSA. Dans le sens opposé, le nombre est égal à $1,3$.

À titre d'exemple, les verbes TUN برد *brid* "*se refroidir*" et استبرد *Aistabrad* "*prendre froid*" partagent la même racine ب ر د *b r d*. Les verbes MSA qui correspondent à ces verbes sont respectivement برد *barad* et مرض *marið*. Ces verbes possèdent deux racines différentes ب ر د *b r d* et م ر ض *m r ð*. Par conséquent, les deux racines MSA sont associées à la seule racine TUN ب ر د *b r d*.

Matrice de correspondance de MBCs

La matrice de correspondance de MBC indique, pour chaque MBC MSA ou TUN, la liste des MBCs qui peuvent lui correspondre du côté opposé. La matrice indique également la fréquence d'apparition de la correspondance entre une MBC TUN et une MBC MSA dans le lexique des verbes. La matrice de correspondance de MBC est représentée dans la table 3.8. Chaque ligne de la matrice correspond à une MBC MSA et chaque colonne à une MBC TUN. La matrice indique, par exemple, que la MBC MSA I correspond à la MBC TUN I dans 434 cas et il correspond à la MBC TUN II dans 98 cas.

		TUN									
		I	II	III	IV	V	VI	VII	VIII	IX	X
MSA	I	$\mathbf{434}^{*}_{*}$	98	10	–	15	–	–	2	–	–
	II	39	$\mathbf{298}^{*}_{*}$	2	–	2	2	–	–	–	2
	III	24	19	$\mathbf{56}^{*}_{*}$	–	–	2	–	–	–	–
	IV	69	$\mathbf{118}^{*}$	4	–	6	–	–	–	–	–
	V	26	16	2	–	$\mathbf{88}^{*}_{*}$	–	–	–	–	3
	VI	18	14	2	–	7	$\mathbf{26}^{*}_{*}$	–	–	–	–
	VII	$\mathbf{13}^{*}$	7	2	–	–	–	–	–	–	–
	VIII	$\mathbf{41}^{*}$	24	5	–	16	4	–	$\mathbf{18}_{*}$	–	–
	IX	–	–	–	–	–	–	–	–	–	–
	X	17	24	2	–	3	–	–	–	–	$\mathbf{31}^{*}_{*}$

Table 3.8.: Matrice de correspondance de MBC s

Une astérisque en position d'indice ($\mathtt{X}_{*}$) indique la fréquence la plus élevée dans le sens TUN → MSA, en position d'exposant ($\mathtt{X}^{*}$) elle indique la fréquence la plus importante dans le sens opposé. La neuvième ligne de la matrice, par exemple, indique que la correspondance ($VIII_{\text{MSA}}$, I_{TUN}) est la plus fréquente

dans le sens MSA→TUN et la correspondance ($VIII_{\text{MSA}}$, $VIII_{\text{TUN}}$) est la plus probable dans le sens opposé.

La matrice présente plusieurs caractéristiques intéressantes. Premièrement, les MBCs de deux côtés MSA ou TUNne sont pas tous présentes dans notre lexique. Les MBCs IV et VII, par exemple, sont absentes du côté TUN. En effet, les verbes TUN qui suivent ces MBCs sont rares. Pour la même raison, la MBC IX est absente des deux côtés. Deuxièmement, le lexique révèle une tendance générale à maintenir la même MBC des deux côtés source et cible d'une entrée lexicale. Ceci est traduit par le fait que les cellules en gras sont souvent situées sur la diagonale. La cellules qui présentent la fréquence la plus élevée sur sa ligne et sa colonne à la fois est représentée par $\mathbf{X}^*_*$. Mis à part la MBC $VIII$, quand les MBCs sont présentes des deux côtés l'intersection des lignes et des colonnes contient la fréquence la plus élevée. Troisièmement, lorsqu'une MBC MSA ne correspond pas à une MBC TUN identique dans le lexique, elle est généralement associée à la MBC I.

Globalement, la matrice montre que la sélection de la racine cible et celle de la MBC cible ne sont pas deux processus indépendants. La décomposition du lexique de verbes en lexique de racines et en table de correspondance de MBCs provoque une perte d'informations. L'apport quantitatif de la division du lexique des verbes en deux ressources séparées sera étudié dans le chapitre 4.

L'extraction d'une table de correspondance de MBCs à partir de la matrice est simple : elle consiste à sélectionner pour chaque MBC de la langue source la MBC la plus fréquente dans la langue cible. Dans certains cas, la MBC la plus fréquente domine clairement les autres MBCs, comme le cas pour la MBC II MSA. Dans d'autres cas, la tendance n'est pas aussi claire, à l'instar de la MBC MSA IV.

3.2.2. Lexique des noms déverbaux

Dans la langue arabe, les noms sont généralement classés en noms déverbaux (qui dérivent d'une racine et d'un schème) et en noms solides (qui ne peuvent être analysés sous la forme d'une racine et d'un schème).

Les noms solides TUN sont généalement proches des noms solides MSA, à quelques variations phonologiques près. À titre d'exemple, l'unique différence entre les noms MSA حصان *HiSAn* "*cheval*", بلاد *bilAd* "*pays*" et leurs correspondants TUN *HSAn* et *blAd* réside dans l'élimination de la première voyelle.

En revanche, les différences entre noms déverbaux TUN et MSA sont plus nombreuses, mais elles ont tendance à être régulières. De plus, les noms déverbaux sont nombreux. Dans le lexique DIINAR (Dichy *et al.*, 2002), par exemple, parmi 109801 entrées nominales, 65% sont des déverbaux.

Ces raisons nous ont poussé à nous intéresser à ces noms et à les traiter dans notre système de manière analogue aux verbes : analyse morphologique profonde suivie d'une phase de transfert au niveau des racines.

Notre lexique de déverbaux a été construit d'une manière automatique à partir du lexique de verbes (Hamdi *et al.*, 2014). La méthode consiste à générer des

paires de déverbaux TUN et MSA d'une façon simultanée en nous servant du lexique des verbes décrit dans la section 3.2 et d'une table de correspondance de schèmes nominaux TUN, MSA, dans le but de générer des paires de déverbaux ($NOUN_{MSA}$, $NOUN_{TUN}$).

Cette méthode sur-génère et peut produire des erreurs du côté MSA ou du côté TUN. Une étape de filtrage s'avère ainsi nécessaire pour éliminer les paires candidates erronées. Nous utilisons pour cela une ressource du MSA existante.

Génération de paires de déverbaux

Comme nous l'avons vu dans le chapitre 1, neuf types de déverbaux peuvent être générés à partir d'un verbe arabe [c]. Ces déverbaux appartiennent au champ sémantique leur racine. Les schèmes qui leurs correspondent sont liés au schèmes verbaux associés à la racine.

Nous avons modélisé cette correspondance dans deux tables de schèmes de déverbaux pour le TUN et le MSA. Ces tables associent à un schème verbal des schèmes nominaux correspondant aux différents types de déverbaux. Ces tables sont présentées dans l'annexe C. Un échantillon des schèmes nominaux issus du schème IX du côté MSA et TUN est donné dans la table 3.9.

schème verbal	type de déverbal	schème nominal	
		MSA	TUN
IX	participe actif	mu12a33	mi12A3
	adjectif analogue	Âa12a3	Âa12a3
	forme infinitive	Ai12i3A3	12uw3iy∼aħ

Table 3.9.: Table de schèmes nominaux msa-tun

La table 3.9 indique que les verbes correspondant au schème verbal IX (qui correspond aux formes *Ai12a33* en MSA et *12A3* en TUN) construisent leur forme infinitive avec le shcème Ai12i3A3 du côté MSA et *12uw3iyy* du côté TUN. Nous avons défini ainsi, pour tous les schèmes verbaux, les schèmes nominaux TUN et MSA leur correspondant pour les neuf types de déverbaux. Au total, nous avons obtenu 54 schèmes nominaux pour MSA et 52 schèmes pour le TUN. À l'aide du lexique des verbes, nous avons combiné la racine, de chaque paire verbale, avec tous les schèmes nominaux correspondant au schème verbal du côté TUN et MSA. Ce processus produit des paires de la forme $((rac_{MSA}, scheme_{MSA}), (rac_{TUN}, scheme_{TUN}))$. Le principe de la génération des paires nominales est décrit dans la figure 3.5.

A ce niveau, environ 20 règles morphologiques et orthographiques développées manuellement sont appliquées sur les formes générées pour produire fi-

c. Ces déverbaux sont : participe actif, participe passif, forme infinitive, adjectif qualificatif, adjectif superlatif, nom d'outil, nom du lieu, nom du temps, forme exagérée (cf. section 1.3.3.)

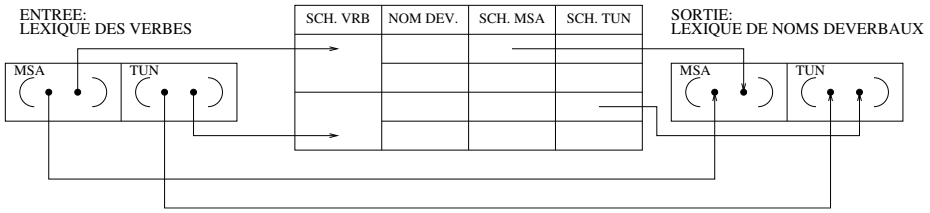

Figure 3.5.: Génération de paires de déverbaux nominaux tun-msa en utilisant les verbes

nalement des paires de lemmes. Citons comme exemple, la règle qui permet de changer le deuxième radical de la racine /y/ ou /w/ en /ŷ/ pour les participes actifs du côté MSA. De la même manière, une règle permet de changer la deuxième radicale /w/ en /y/ dans le côté TUN. Nous pouvons citer également une deuxième règle commune pour le MSA et le TUN qui nécessite de transformer les /t/ du schème verbal Ai1**t**a2a3 (VIII) et tous les schèmes nominaux qui en dérivent en /T/ si la première lettre de la racine correspond à /S/, /T/, /D/ ou /Z/ : e.g. la forme infinitive اضترَاب *AiDtirAb* devient اضطرَاب *AiDTirAb* "*trouble*".

À l'issue de l'étape de génération, un lexique de 137199 entrées nominales (Nom_{MSA}, Nom_{TUN}) est créé.

Le processus de dérivation, même s'il est généralement régulier, admet des exceptions. Le processus que nous avons suivi génère, par conséquent, des formes incorrectes. En effet, une racine ne peut pas se combiner à tous les schèmes nominaux ce qui engendre la création d'entrées contenant des paires nominales qui n'existe pas dans le lexique MSA ou dans le lexique TUN.

Filtrage du lexique

La méthode de génération décrite ci-dessus sur-génère : elle produire des paires correctes mais aussi des paires incorrectes. Quatre cas sont possibles :

1. les deux noms TUN et MSA sont corrects
2. le nom TUN est incorrect et le nom MSA est correct
3. le nom MSA est incorrect et le nom TUN est correct
4. les deux formes générées sont incorrectes

Ainsi, la paire verbale du lexique (فتح, حلّ) $(fataH_{MSA}, Hall_{TUN})$ "*ouvrir*", par exemple, peut générer ces quatre situations :

1. participe passif : (مفتوح, محلول) $(maftuwH_{MSA}, maHluwl_{TUN})$ "*ouvert*". Dans ce cas, les deux mots générés sont corrects.
2. forme exagérée : (فتّاح, حلّال) $(fattAH_{MSA}, HallAl_{TUN})$. Le nom TUN est incorrect dans cette figure alors que le nom MSA est correct "*conquérant*".

3. nom du lieux : (مفتح, محلّ) $(maftaH_{MSA}, mHall_{TUN})$, dans ce cas le nom TUN est correct "*boutique*" mais le déverbal MSA n'existe pas. Le mot tunisien est obtenu après l'application de la règle de gémination.
4. adjectif analogue : (فتيح, محلال) $(ftiyH_{MSA}, miHlAl_{TUN})$, les deux mots générés dans ce cas sont incorrects.

Dans notre cas, les situations (2) et (4) ne sont pas préjudiciables dans la mesure où nous n'analysons jamais une forme TUN incorrecte[d]. Les situations (3) et (4) peuvent être partiellement traitées en filtrant la partie MSA à l'aide d'une ressource existante. Pour cela, nous avons eu recours à trois ressources différentes :

– *un grand corpus* composé de dépêches de presse de l'AFP (*Agence française de presse*), qui contient environ $1,5$ million d'occurrences. À partir de ce corpus, nous avons extraits 10595 types de lemmes nominaux à l'aide de l'analyseur morphologique de l'arabe standard MADA. Seules les paires dont le nom MSA est attesté dans le corpus ont été maintenues. Suite à cette opération, un lexique de 20130 entrées a été obtenu. Ce lexique est composé de 8441 noms MSA associés 2636 noms TUN.
– *le lexique du* MSA à large couverture SAMA (Graff *et al.*, 2009) contenant 36935 lemmes nominaux. Le filtrage à l'aide de ce lexique a donné naissance à un lexique de 26486 entrées : 10647 déverbaux côté MSA et 4712 côté TUN.
– *l'union des deux ressources* composée de 40172 lemmes nominaux. En utilisant cette ressource, le filtrage a généré un lexique composé de 39793 paires a été obtenu. Ce lexique comporte 14804 lemmes MSA et 5017 lemmes TUN.

Afin d'évaluer ce lexique nous avons eu recours à un corpus d'évaluation TUN (voir section 3.4) contenant environ six mille occurrences parmi lesquels 1295 sont des déverbaux. Ce corpus a été divisé en deux parties égales, un corpus de test et un corpus de développement.

Deux métriques ont été utilisées pour l'évaluation du lexique des déverbaux généré. La première métrique est la couverture (qui correspond au rappel) qui mesure la portion des déverbaux du corpus présents dans le lexique généré. La deuxième métrique est l'ambiguïté qui constitue le nombre de déverbaux cibles en moyenne pour un déverbal source. Nous n'avons pas utilisé la précision comme mesure d'évaluation car le corpus d'évaluation est de petite taille. Un mot figurant dans le lexique mais pas dans le corpus ne peut être considéré comme incorrect.

Cette méthode présente deux sources d'ambiguïté :

– le lexique de verbes peut associer à un verbe source plusieurs verbes cibles à l'instar du verbe TUN مشَى *mšý* qui est associé aux deux verbes du MSA ذهب *ðhb* "*aller*" et مشَى *mšý* "*marcher*".

d. Nous supposons ici que notre entrée TUN ne comporte pas d'erreurs !

– la table de correspondance de schèmes nominaux TUN-MSA peut définir plusieurs schèmes cibles pour un schème source.

L'évaluation du lexique de déverbaux sur le corpus de test est donnée dans la table 3.10.

méthode de filtrage	nombre d'entrées	couverture	ambiguïté	
			MSA→TUN	TUN→MSA
sans filtrage	173407	67,2%	7,7	12,6
AFP	17896	60,1%	2,4	6,9
SAMA	33271	63,9%	3,5	7,2
AFP ∪ SAMA	35792	65,8%	2,6	7,4

Table 3.10.: Couverture du lexique de déverbaux sur l'ensemble de test

La table 3.10 montre que sans aucun filtrage la couverture du lexique atteint 67,2%. L'ambiguïté dans le sens TUN→MSA est égale à 12,6, c'est-à-dire, en moyenne, pour un déverbal TUN, 12,6 déverbaux MSA sont générés. Après le filtrage, la couverture du lexique baisse légèrement. Cela est dû au fait que la méthode du filtrage élimine dans certain cas des entrées TUN correctes correspondant à des noms MSA incorrects, ce qui correspond au cas 3 de la classification proposée ci-dessus. À l'aide du corpus AFP et du lexique SAMA, la couverture diminue à 60,1% et 62,7% respectivement. Enfin, la méthode qui consiste à filtrer le lexique par la ressource AFP ∪ SAMA, offre une meilleure couverture qui atteint 65,7% et une ambiguïté égale à 7,4.

La table 3.11 présente les valeurs de la couverture du lexique sur le corpus de développement. On observe une situation très proche de celle observée sur le corpus de test.

méthode de filtrage	nombre d'entrées	couverture	ambiguïté	
			MSA→TUN	TUN→MSA
sans filtrage	173407	66,1%	7,6	12,6
AFP	17896	59,2%	2,4	6,9
SAMA	33271	62,7%	3,5	7,2
AFP ∪ SAMA	3579	64,6%	2,6	7,4

Table 3.11.: Couverture du lexique de déverbaux sur l'ensemble de développement

Une analyse d'erreurs de la génération automatique des entrées lexicales sur l'ensemble de développement a montré que les erreurs (absence d'un nom TUN dans le lexique) proviennent principalement de trois origines différentes :

1. absence du verbe correspondant dans le lexique des verbes : les noms qui dérivent d'un verbe absent du lexique verbal ne sont pas générés dans le lexique des déverbaux.

2. absence de correspondances dans la table de correspondance de schèmes nominaux TUN-MSA
3. absence de règles morphologiques et orthographiques.

Dans le but d'estimer l'influence de la situation 1 sur la couverture, nous avons enrichi le lexique des verbes de sorte à couvrir tous les déverbaux du corpus de développement. À l'issue de ce processus, 92 entrées verbales ont été rajoutées.

La table 3.12 montre la couverture du lexique sur l'ensemble de développement. Bien qu'artificiel, ce résultat permet d'estimer la borne supérieure que la couverture peut atteindre dans le cas où le lexique des verbes est complet.

méthode de filtrage	nombre d'entrées	couverture	ambiguïté	
			MSA→TUN	TUN→MSA
sans filtrage	195917	87,3%	7,9	12,9
AFP	20130	81,5%	2,2	7,2
SAMA	36935	82,9%	3,7	8,1
AFP ∪ SAMA	39763	84,2%	2,9	8,2

Table 3.12.: Résultats sur l'ensemble de développement après l'enrichissement du lexique des verbes

Comme l'indique la table 3.12, après avoir garni le lexique des verbes, la couverture du lexique est passé de 66,2% à 87,3% avant l'étape de filtrage et de 64,6% à 84,2% après le filtrage par la ressource AFP ∪SAMA, pour une légère augmentation du taux d'ambiguïté.

La table 3.13 présente les résultats obtenus sur le corpus de test.

méthode de filtrage	nombre d'entrées	couverture	ambiguïté	
			MSA→TUN	TUN→MSA
sans filtrage	195917	72,9%	7,9	12,9
AFP	20130	65,9%	2,2	7,2
SAMA	36935	68,4%	3,7	8,1
AFP ∪ SAMA	39763	71,2%	2,9	8,2

Table 3.13.: Résultats sur l'ensemble de test après l'enrichissement du lexique des verbes

Le fait d'enrichir le lexique des verbes a contribué à améliorer considérablement la couverture du lexique de déverbaux sur le corpus de test. En effet, elle passe de 67% à 73% avant le filtrage et de 65% à 71% après le filtrage par la ressource AFP ∪SAMA. L'ambiguïté reste par contre stable.

L'expérience précédente a montré qu'une large portion des erreurs provient du manque de couverture du lexique des verbes. En ajoutant 92 entrées verbales, la couverture du lexique de déverbaux s'est élevée d'environ 6%. Parmi ces 92

entrées, la racine était absente dans 28 cas. Dans les 64 autres cas, la racine était présente, c'est le couple (racine, schème) qui était absent.

3.2.3. Lexique des particules

La catégorie des particules dans la langue arabe couvre de multiples catégories : conjonctions, prépositions, adverbes, particules, pronoms et clitiques. Notre lexique de particules est composée de 200 entrées environ. Ces entrées se présentent sous a forme de paires (PRT_{MSA}, PRT_{MSA}), elles couvrent toutes les catégories. De la même manière que pour le lexique des verbes (cf. section 3.2.1), les particules MSA ont été extraites de l'ATB puis elles ont été traduites manuellement en contexte vers le TUN (Boujelbane *et al.*, 2013).

Nous donnons dans la table 3.14 des exemples de particules MSA et TUN pour illustrer la différence lexicale entre les deux variantes. Notons que dans $37,5\%$ de cas, les entrées du lexique apparient les mêmes particules des deux côtés.

MSA	TUN		glossaire
فقط *faqaT*	كهو *kahaw*		c'est tout
أيضا *ÂyDA*	زادة *zAdħ*		aussi
مثل *miθla*	كيف *kiyf*	كمَا *kimA*	comme
لكن *lakin∼a*	أما *Âam∼A*		mais
لو *law*	كان *kAn*	إذا *ĂiðA*	si

Table 3.14.: Exemples de particules TUN et MSA

Dans sa version actuelle, le lexique relie 187 particules TUN à 143 particules MSA. Contrairement aux verbes et aux noms, l'ambiguïté est moins élevée dans le sens TUN → MSA. En effet, une particules TUN correspond en moyenne à $1,2$ particules du côté MSA. En revanche, une particule MSA correspond, en moyenne, à $1,7$ particules TUN. Nous donnons dans la table 3.15, l'ambiguïté moyenne pour chaque catégorie de particules dans les deux sens TUN → MSA et MSA → TUN.

catégorie	TUN → MSA	MSA → TUN
conjonctions	1,1	1,2
prépositions	1,1	1,3
adverbes	1,2	2,8
particules	1,2	1,6
pronoms	1,5	1,1
clitiques	1,1	1,1

Table 3.15.: Ambiguïté du lexique de particules TUN-MSA

3.3. Étiqueteur en parties de discours

Afin d'évaluer notre méthode sur l'étiquetage en parties de discours, nous avons eu recours à différents étiqueteurs entraînés sur le MSA. Ces étiqueteurs reposent sur le modèle des chaînes de Markov cachées que nous désignerons dans la suite de ce document par leur acronyme anglais HMM (pour *Hidden Markov Models*). Nous avons choisi ce modèle pour les raisons suivantes :

(1) lien avec les machines à états finis : tous les traitements mis en œuvre dans le cadre de ce travail reposent sur des machines finies. Or les HMM peuvent être implémentées sous la forme de telles machines. Choisir ce type de modèle permet d'aboutir à une chaîne de traitement d'une grande homogénéité.

(2) prise en compte naturelle d'entrées ambiguës : un autre avantage des HMM et de leur représentation sous la forme de machines finies est de pouvoir prendre en entrée de manière naturelle des hypothèses multiples. Cette caractéristique est particulièrement importante dans notre cas puisque notre processus de conversion est, comme nous l'avons évoqué, ambigu. Il propose pour une entrée TUN plusieurs sorties MSA.

(3) performances : bien que les HMM ne constituent pas le modèle le plus performant pour l'étiquetage morphosyntaxique, les différences de performances entre les HMM et d'autres modèles (tels que les CRF) pour l'étiquetage sont modestes.

Un étiqueteur HMM se présente sous la forme d'un transducteur P. Ce dernier est lui même la composition d'un transducteur pondéré E et d'un automate pondéré T ($P = E \circ T$). Le transducteur E associe à un mot m en entrée, ses différentes catégories possibles c_i. Les pondérations sont des probabilités d'émission $P(m|c_i)$. L'automate T associe une probabilité à une séquence de catégories [e]. Le calcul de la probabilité d'une séquence peut être réalisé par des modèles d'ordres variables. Nous avons eu recours à des modèles d'ordre 2 et 3 (2-gram et 3-gram), ce qui est la pratique courante pour la tâche d'étiquetage morphosyntaxique.

e. à l'image d'un modèle de langage dans un système de transcription automatique de la parole

L'entrée de l'étiqueteur se présente sous la forme d'un automate acyclique M qui représente la sortie de notre système de conversion. L'automate M peut être linéaire, il correspond alors à une séquence unique de mots MSA. Il peut aussi se présenter sous la forme d'une succession de faisceaux de transitions où les transitions de chaque faisceau correspondent à tous les mots MSA possibles pour un mot TUN.

L'étiquetage en parties de discours de M est réalisé par composition de ce dernier avec l'étiqueteur P. Cette opération est suivie d'une opération de recherche du chemin de moindre coût dans le transducteur issu de la composition. L'étiquetage correspond donc à la séquence d'opérations suivante :

$$BP(M \circ E \circ T)$$

où $BP(A)$ est l'opération de recherche du meilleur chemin dans l'automate acyclique A.

Lorsque l'automate M correspond à plusieurs hypothèses de conversions, l'étiqueteur P réalise simultanément l'étiquetage morphosyntaxique et la désambiguïsation. Nous étudierons plus en détails ce phénomène dans le chapitre 4.

L'étiqueteur est entraîné sur le corpus CATIB (Habash et Roth, 2009) qui correspond à la partie III de l'ATB. Le corpus est formé de $24K$ phrases composées de $330K$ occurrences et $30K$ types de mots en MSA. CATIB utilise un jeu d'étiquettes composé six catégories différentes : nom, nom propre, verbe, verbe passif, particule et ponctuation. Les étiquettes correspondantes à ces catégories sont respectivement : noun, prop, verb, verb-pass, part, pnx.

La table 3.16 montre les résultats de l'étiquetage en parties de discours du MSA par un étiqueteur 2-gram et 3-gram fondé sur le modèle décrit ci-dessus. Ces résultats sont légèrement inférieurs aux résultats obtenus par d'autres étiqueteurs du MSA existants. Pasha *et al.* (2014) arrive, par exemple, à des résultats légèrement supérieur (96%) en utilisant le système MADAMIRA sur les mêmes données. Nous n'avons pas utilisé cet étiqueteur car il ne permet pas de traiter des entrées ambigües contrairement à notre étiqueteur HMM.

2-gram	3-gram
94.52%	94.72%

Table 3.16.: Performances de l'étiquetage en parties de discours du MSA

3.4. Corpus d'évaluation tunisien

L'évaluation de notre méthode suppose d'avoir à notre disposition un corpus de référence en tunisien annoté en parties de discours. Une telle ressource n'existe pas, c'est la raison pour laquelle nous avons transcrit manuellement un

corpus TUN composé d'environ 800 phrases et de $11K$ mots. Il a été, par suite, segmenté et annoté manuellement. À chaque mot du corpus, nous avons assigné son lemme et sa partie de discours.

Le corpus est composé de phrases extraites de quatre sources différentes :

– des séries télévisées
– des débats politiques
– une pièce de théâtre transcrite (Dhouib, 2007)
– un corpus transcrit à partir des enregistrements de discussions entre des clients et un agent de la société nationale des chemins de fer tunisiens. Ce corpus a été enregistré pour entraîner un système TUN de reconnaissance de la parole (Masmoudi *et al.*, 2014).

Ces sous-corpus se distinguent sur différents points. Le premier est la variété lexicale. Ces sous-corpus correspondent en effet à des lexiques différents. Le second est le niveau de spontanéité.

La variété de ce corpus va permettre de tester notre modèle sur des données que nous pensons réalistes d'un point de vue linguistique. En revanche, elles ne sont pas réalistes par le fait que nous prenons en entrée du tunisien transcrit manuellement. Dans l'idéal il aurait été souhaitable d'utiliser les sorties d'un système de transcription automatique du tunisien.

La table 3.17 présente quelques statistiques du corpus d'évaluation.

statistiques	séries télévisées	débats politiques	pièce de théâtre	corpus SNCFT	total
phrases	203	199	205	195	802
formes occurrences	3032	2886	3163	2670	11551
formes types	689	622	712	433	2456
lemmes	474	442	502	331	1749
LMMs	431	407	466	360	1664

Table 3.17.: Statistiques sur le corpus d'évaluation tunisien

Outre le degré de spontanéité et le lexique, d'autres particularités caractérisent les différentes variantes qui composent le corpus d'évaluation. Les débats politiques, par exemple, constituent la variante la plus proche du MSA. En effet, les politiciens ont généralement une tendance à utiliser des phrases de l'arabe standard. Ainsi, un mélange de phrases et mots TUN et MSA est obtenu. De plus, nous retrouvons dans cette partie une combinaison de morphèmes MSA et TUN dans les mots. Le mot ننخرطوا *nnxrTwA*, par exemple est composé d'un circonfixe tunisien +ن +وا *n*+ +*wA* et le verbe MSA انخرط *AnxrT* qui correspond en TUN auxn+ verbes شارك *šArk* "*participer*" ou قيّد *qay∼id* "*s'inscrire*".

3.4.1. Conventions de transcription

La transcription du corpus repose sur les conventions CODA. Ces conventions visent à définir un cadre commun de transcription pour tous les dialectes de l'arabe. Contrairement à d'autres conventions et pour des raisons computationnelles, CODA définit une seule interprétation orthographique pour chaque mot.

Les conventions orthographiques sont fondées sur la similarité entre le MSA et ses dialectes. Ce choix a pour objectif de définir une seule convention qui réunit tous les dialectes arabes.

Au niveau phonétique, nous avons utilisé les lettres et les diacritiques arabes dans la transcription. Le TUN définit trois sons /g/, /p/ et /v/ qui ne sont pas représentés dans l'alphabet arabe. Dans ce cas, nous avons eu recours aux lettres du MSA qui produisent des sons proches phonétiquement. Les sons /g/, /p/ et /v/, sont transcrits ainsi par les lettres ق */q/*, ب */b/* et ف */f/* respectivement. Les mots /mung :ala/, /pArtiy/ et /viysta/ par exemple sont transcrits, respectivement, par منقالة *munqAlaħ* "*horloge*"/"*montre*", بارتي *bArtiy* "*match*" et فيستة *fiystaħ* "*veste*".

Contrairement à l'arabe standard où la lettre ة *ħ* est toujours prononcée, le dialecte tunisien, à l'instar des autres dialectes arabes, ne la prononce pas. Cette lettre marque les noms du genre féminin et elle est toujours précédée par le diacritique /a/ en MSA et en TUN. En tunisien, de nombreux mots se terminent par le son /a/ tels que /barša/ *beaucoup*, /famma/ "*il y a*". Parmi ces mots, nous avons rajouté la lettre ة *ħ* aux mots de valeur nominale et de genre féminin. Suivant cette convention, les transcriptions des mots /barša/ et /famma/ sont respectivement برشة *baršaħ* et فمّة *fam∼aħ*. Cette convention est établie pour maintenir la similarité entre le MSA et ses dialectes.

Un échantillon du corpus TUN est donné dans la table 3.18. Ce texte est accompagné par sa translittération, sa traduction en MSA et en français. Cet exemple illustre les variations et les similarités entre le TUN et le MSA.

3.4.2. Conventions de segmentation

Le corpus a été segmenté en tours de parole. Des signes de ponctuation ont été utilisés pour marquer les tours de parole. Les virgules désignent un temps de silence marqué par un locuteur. Nous nous sommes servis des points d'interrogation et d'exclamation pour dénoter respectivement les questions et les phrases qui indiquent un sens d'admiration ou de surprise. L'espace blanc a été utilisé, à l'image de l'écrit, pour séparer les mots.

Au niveau de l'agglutination, le TUN partage la majorité des clitiques du MSA. Dans ce cas, nous avons utilisé la même segmentation que le MSA. Le pronom d'objet direct de la troisième personne du singulier est réalisé comme ـه+ bien que non prononcé en TUN mais nous le maintenons pour être similaire à l'arabe

TUN	معقول معقول ياسر أمّا نكونوا هنا موضوعيّين أوّلا وقتلي بدات الحكومة الَانتقاليّة الأولى... لا أنا نذكّرك بالتّصريح زادة فمّة تصريح لسي الباجي قايد السبسي وقتها هو رئيس الحكومة وبعد أنا نخلّيك تحكي عالبداية قال وقتها يلزم عشرة سنين عَالأقلّ باش البلاد تنجّم تمشي. mṣqwl mṣqwl yAsr ÂmmA nkwnwA hnA mwDwṣyyiyn ÂwwlA wqtly bdAt AlHkwmħ AlĂntqAlyyħ AlÂwlý... lA ÂnA nðkrk bAltSryH zAdħ fmmħ tSryH lsy AlbAjy qAyd Alsbsy wqthA hw rŷys AlHkwmħ wbṣd ÂnA nxlyk tHky ṣAlbdAyħ qAl wqthA ylzm ṣšrħ snyn ṣAlÂqll bAš AlblAd tnjjm tmšy.
MSA	معقول معقول جدّا لكن لنكن هنا موضوعيّين أوّلَا لمّا بدأت الحكُومة الانتقاليّة الأولى... لا أنا أذكّرك بالتّصريح أيضا هناك تصريح لالسيّد الباجي قايد السبسي أنذَاك هو رئيس الحكومة وبعدهَا أنا أترلكّ تحكي عن البدَاية قال عندها يلزم عشر سنوات على الأقلّ لالبلد لتسطيع أن تّقدّم. mṣqwl mṣqwl jddA lkn lnkn hnA mwDwṣyyiyn ÂwwlA lmmA bdÂt AlHkwmħ AlĂntqAlyyħ AlÂwlý... lA ÂnA Âðkrk bAltSryH ÂyDħ hnAk tSryH lAlsyd AlbAjy qAyd Alsbsy ÂnðAk hw rŷys AlHkwmħ wbṣdhA ÂnA Âtrkk tHky ṣn AlbdAyħ qAl ṣndhA ylzm ṣšr snwAt ṣlý AlÂql lAlblAd ltstTyṣ Ân ttqdm.
français	*c'est logique c'est très logique mais restons ici objectifs* *premièrement lorsque le premier gouvernement transitoire a entamé...* *Non, Moi je te rappelle aussi de la déclaration, il y a une déclaration* *de M. El-bAji Qayid El-sibsi, à cette époque, il était chef du gouvernement* *ensuite je te laisse parler du début, il disait à l'époque* *il faut au moins dix ans pour que ce pays puisse progresser.*

Table 3.18.: Échantillon du corpus d'évaluation tunisien

standard. En effet le mot /ktibt :u/ en tunisien est exprimé en /ktibtuh/ "*je l'ai écrit*".

En revanche, le TUN se distingue du MSA, en terme d'agglutination, sur deux points. D'une part, certains clitiques MSA sont réalisés sous la forme de particules indépendantes en tunisien et vice-versa. En particulier, la préposition +لـ *li+ pour* et le proclitique du futur ne sont plus rattachés aux verbes. Tous les deux se traduisent par la particule indépendante باش *bAš* qui se situe avant le verbe : les formes لتكتب *litaktuba* "*pour que tu écrives*" et ستكتب *sataktubu* "*tu écriras*" sont exprimés en tunisien par باش تكتب *bAš tiktib*. Nous séparons cette particule du mot suivant étant donné que nous pouvons insérer un mot entre eux. Les prépositions من *min* "*de*" et على *ʻalaý* "*sur*" ainsi que le pronom démonstratif quels que soient son genre et son nombre sont réalisés sous la forme de clitiques en TUN. En effet, ils sont exprimés respectivement en +مـ *mi+*, +عـ ʕ*a+* et +هـ *ha+* en tunisien.

D'autre part, la forme de certains clitiques change. Le proclitique d'interrogation MSA +أ *Â+* "*est-ce que*", par exemple, devient en tunisien l'enclitique +شـ *+š*. La forme verbale MSA أكتبت *Âkatabta* "*est-ce que tu as écrit*" se traduit en tunisien par كتبتش *ktibtiš*.

3.4.3. Conventions d'annotation

Dans l'étiquetage en parties de discours du corpus, nous avons suivi les conventions de CATIB. Ce corpus définit un jeu d'étiquettes restreint composé de six catégories grammaticales. Les étiquettes sont inspirées de la grammaire traditionnelle de l'arabe qui classe les tous les mots en trois catégories اسم, فعل et حرف *nom, verbe et particule*. La simplicité de cet ensemble d'étiquettes a été suivi pour accélérer le processus d'annotation manuelle tout en gardant les distinctions importantes :

- **VRB** désigne toutes les formes verbales mises dans la voix active. Cette catégorie inclue l'ensemble des verbes incomplets (الأفعال الناقصة "AlÂfçAl AlnAqSℏ") à l'instar de كَان *kAn* "*être*", ليس *lys* "*ne pas être*", صار *SAr* "*devenir*" et مازال *mAzAl* "*demeurer*".
- **VRB-PASS** désigne toutes les formes verbales de la voix passive.
- **NOM** désigne les noms, les adjectifs, les adverbes, les déverbaux, les pronoms (personnels, possessifs, relatifs et démonstratifs), les numéraux et les interjections. Les pseudo-prépositions telles que قدّام *quddAm* "*devant*" et تحت *taHt* "*sous*" et les quantificateurs comme كلّ *kul* "*tout*"sont tous consi-

dérés comme des noms. Tous ces mots peuvent être déterminés avec l'article défini الـ *Al*.

- **PROP** désigne les noms propres. Contrairement au français qui marque les noms propres par une majuscule au début des mots, cette distinction n'existe pas en arabe. Le mot سليم *slym*, par exemple, est étiqueté nom s'il se traduit par "*sain*" et comme nom propre s'il se traduit par "*Salim*".
- **PRT** est utilisé pour toutes les particules. Cet ensemble inclut les prépositions, les conjonctions, les particules de futures, de négation, interrogatifs et les pseudo-verbes.
- **PNX** inclut tous les marqueurs de ponctuation.

Conclusion

Dans ce chapitre, nous avons présenté les différents outils et ressources développés pour la mise en œuvre de l'étiquetage morphosyntaxique du TUN. Dans le chapitre suivant, nous décrirons les expériences qui permettent l'application de ces outils dans le but de réaliser les pré-traitements nécessaires pour réaliser l'étiquetage.

4. Expérimentation et évaluation

Nous avons décrit dans le chapitre précédent les moyens nécessaires à la mise en œuvre du dispositif que nous proposons pour réaliser l'étiquetage en parties de discours du TUN. Nous présentons dans ce chapitre l'architecture générale de notre système ainsi que son fonctionnement en détaillant les trois processus impliqués : la conversion, la désambiguïsation et l'étiquetage.

Dans la section 4.1, nous revenons sur l'organisation du processus entier à travers un exemple illustratif. La section 4.2 concerne la description du processus de conversion. La désambiguïsation est décrite dans la section 4.3. Enfin, la section 4.4 décrit les expériences d'étiquetage.

4.1. Architecture générale

Comme nous l'avons évoqué au début de ce chapitre, la méthode que nous proposons pour réaliser l'étiquetage en parties de discours du dialecte tunisien est composée de trois étapes. Dans un premier temps, les différents mots d'un texte TUN sont convertis en un ou plusieurs mots en MSA. Plus précisément, étant donné la séquence de mots TUN t_1, t_2, ... t_n, chaque mot t_i est traduit un ou plusieurs mots cibles $m_{i,1}$, $m_{i,2}$, ... m_{i,k_i} en MSA. Cet ensemble peut être décrit à l'aide d'un automate fini acyclique qui permet de représenter toutes les séquences $m_1 \ldots m_n$ avec $m_1 \in M_1$ et $m_n \in M_n$ où M_i est un ensemble de forme $\{m_{i,1} \ldots m_{i,k_i}\}$. Un tel automate est représenté en figure 4.1.

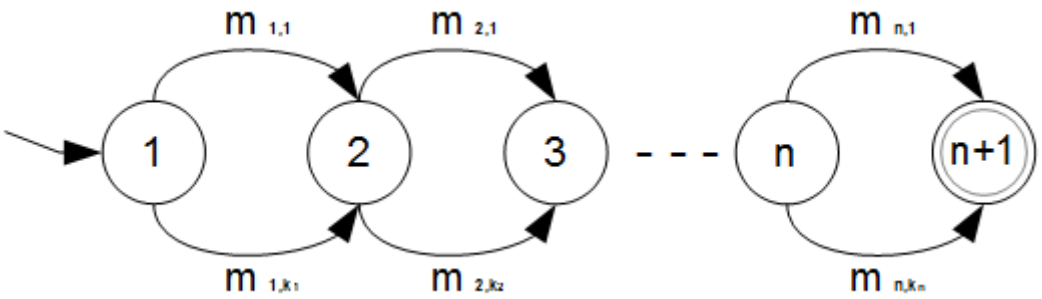

Figure 4.1.: Représentation de la sortie de la conversion à l'aide d'un automate acyclique

Dans un deuxième temps, une étape de désambiguïsation permet de sélectionner une séquence de mots $\hat{M} = \hat{m}_1, \hat{m}_2, \ldots \hat{m}_n$ parmi toutes les séquences reconnues par l'automate. La désambiguïsation réalisée à l'aide d'un modèle de langage de type n-gram. Cette séquence est celle à laquelle le modèle de langage attribue la probabilité la plus élevée.

Enfin, la séquence $\hat{M}$ est étiquetée en parties de discours par un étiqueteur morpho-syntaxique. Cet étiqueteur produit une séquences d'étiquettes : e_1, e_2, $\ldots e_n$. Ces étiquettes sont finalement projetées sur la phrase source t_1, t_2, $\ldots t_n$.

À titre d'illustration, prenons comme exemple la phrase TUN تجبر باش يقعد *tijbar bAš yuqʕud* "*il a été obligé de rester*". La séquence de parties de discours correspondant aux mots de cette phrase est '*verb-pass part verb*'.

Une traduction correcte en MSA de cette phrase est اضطرّ إلى البقاء *AiðTar∼a ĂilaÝ AlbaqA*'. Les étiquettes en parties de discours correspondant à cette phrase sont '*verb part nom*'. En effet, le verbe MSA *AiðTar∼a* est transitif indirect, son objet commence généralement par la préposition *ĂilaÝ*. Cette dernière est toujours suivie d'un nom ou d'un nom propre.

Comme nous l'avons mentionné précédemment, l'objectif général de ce travail n'est pas de produire un système de traduction du TUN vers le MSA mais de générer à partir d'une phrase TUN une version de cette dernière sous une forme approximative du MSA, de sorte que des outils de traitement automatique du MSA, tel qu'un étiqueteur en parties de discours puissent être utilisés sur cette nouvelle forme du texte avec des résultats satisfaisants. Suite à la conversion et à la désambiguïsation de notre exemple, le système produit la phrase cible اضطرّ سوف يجلس *AuðTur∼a sawfa yajlisu* "*il a été obligé il va s'assoir*". Bien que la traduction de la phrase ne soit pas correcte, elle reçoit la séquence d'étiquettes '*verbe-pass particule verbe*' qui correspond aux mots de la phrase source.

L'architecture générale de notre système est décrite dans la figure 4.2.

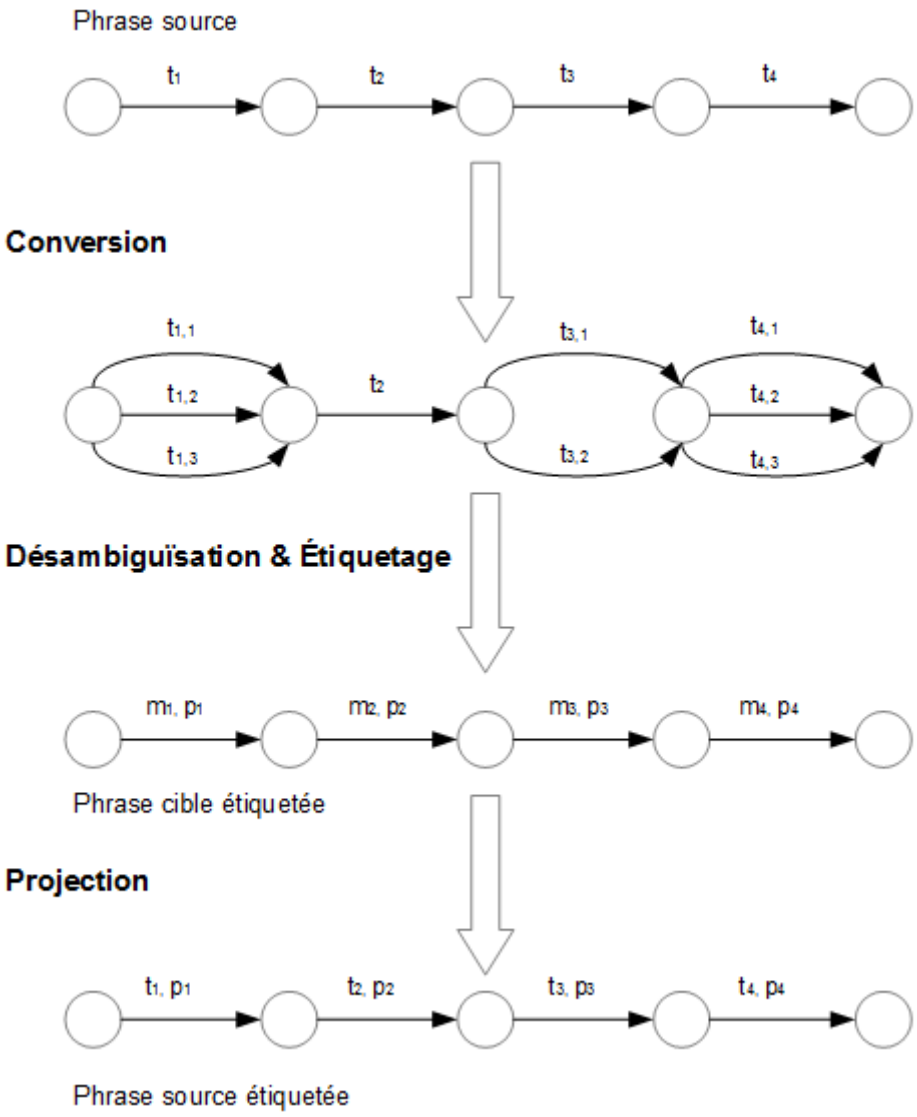

Figure 4.2.: Étiquetage en partie de discours d'une phrase en tun : architecture générale

Le dispositif d'étiquetage que nous proposons présentent deux particularités intéressantes dans le cadre de ce travail. D'une part, la conversion d'une séquence de formes TUN en pseudo-MSA peut produire des lemmes car la génération des formes cibles est réalisée à partir d'une analyse morphologique profonde. D'autre part, le processus désambiguïsation n'est pas indispensable car, comme nous l'avons vu au chapitre précédent, notre étiqueteur morpho-syntaxique peut prendre en entrée un ensemble d'hypothèses de conversion. Dans ce cas, c'est l'étiqueteur qui réalise la désambiguïsation et l'étiquetage.

Comme indiqué en tête du chapitre, chacune des étapes du processus fait l'ob-

jet d'une section.

4.2. Conversion du dialecte tunisien en arabe moderne standard

Comme nous l'avons vu au chapitre précédent, le processus de conversion d'une forme source en TUN en une ou plusieurs formes cibles en MSA se décompose en trois étapes :

1. L'analyse morphologique à l'aide de l'outil MAGEAD TUN décrit dans la section 3.1.3. MAGEAD TUN est composé de deux systèmes : MAGEAD_N pour le traitement des noms et MAGEAD_V pour le traitement des verbes. A l'issue de cette étape, des triplets (racine-source, MBC-source, traits morphologiques) sont produits.
2. Le transfert lexical au niveau des racines et des MBCs grâce à des lexiques MSA-TUN de verbes (cf. section 3.2.1) et de noms (cf. section 3.2.2).

   De manière plus précise, chaque paire (racine-source, MBC-source) permet de faire un accès aux lexiques pour extraire un ou plusieurs couples (racine-cible, MBC-cible). Les traits morphologiques sont quant à eux conservés tels quels. Les couples (racine-cible, MBC-cible) et les traits morphologiques constituent l'entrée du générateur morphologique MSA.

   Les formes sources TUN sont aussi utilisée sans analyse pour faire un accès au lexique des particules. Dans le cas d'un succès de l'accès, une ou plusieurs particules cibles sont générées.
3. la génération du mot cible en MSA grâce à l'outil MAGEAD MSA. Ce processus prend entrée des triplets (racine-cible, MBC-cible, traits morphologiques) et produit des formes MSA.

Rappelons que chacune de ces étapes est réversible et que l'on peut symétriquement traduire un mot en MSA en un mot en TUN.

Le processus complet est décrit dans la figure 4.3.

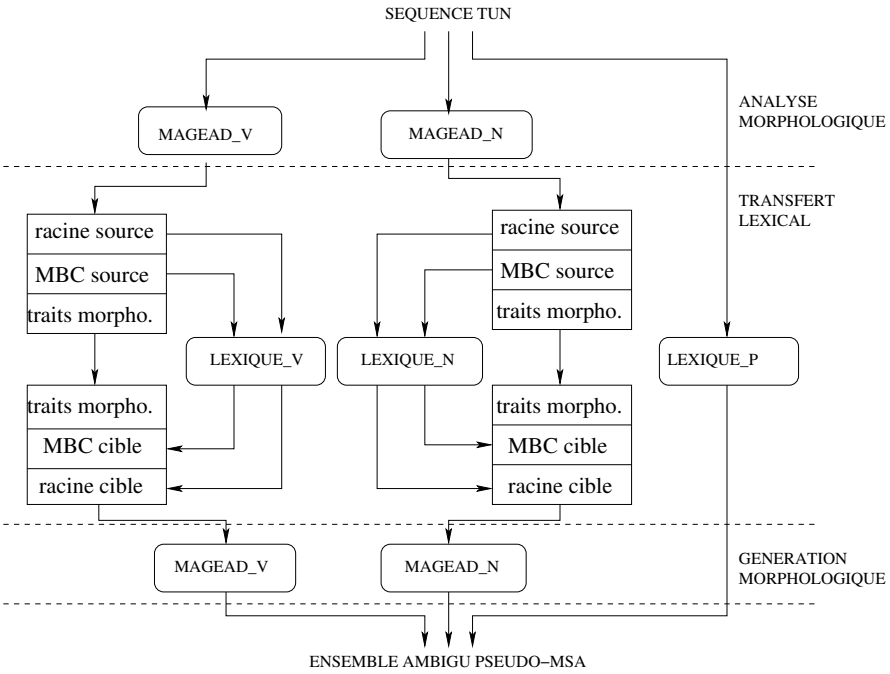

Figure 4.3.: Passage du dialecte tunisien à l'arabe standard

Ce modèle comporte deux sources d'ambiguïté :

– **morphologique** : l'analyse morphologique peut proposer plusieurs analyses pour un mot source. La génération morphologique est également ambigüe. En effet, la morphologie du MSA est plus riche que celle du TUN, qui ne réalise pas certains traits morphologiques, ou certaines valeurs de traits morphologiques. La génération peut ainsi produire plusieurs formes MSA à partir de l'analyse d'une forme TUN.

– **ambiguïté lexicale** : les lexiques de verbes et de noms peuvent mettre en correspondance plus d'un couple (racine-cible, MBC-cible) à un couple (racine-source, MBC-source). Le lexique de particules peut, lui aussi, proposer plusieurs cibles pour une particule TUN.

Le processus de conversion admet quatre variantes
qui se distinguent par la manière dont est réalisé le transfert lexical (Hamdi *et al.*, 2013a) :

1. transfert limité aux MBCs,
2. transfert de MBCs et de racines séparément,
3. transfert de couples (racine, MBC),
4. transfert de couples (racine, MBC) avec repli.

Afin de comparer ces méthodes, nous avons procédé à leur évaluations sur les verbes. Pour cela, nous avons extrait tous les verbes d'une pièce de théâtre

transcrite en TUN (Dhouib, 2007). Environ 1500 occurrences de formes verbales ont été identifiées et traduites en contexte vers le MSA. À l'issue de ce processus, un corpus contenant 1500 couples ($verbe_{\text{TUN}}$, $verbe_{\text{MSA}}$). Les verbes TUN et MSA du corpus sont présentés sous leurs formes fléchies. Cet ensemble d'évaluation a été divisé en deux parties égales. La première constituant un ensemble de développement et la seconde un ensemble de test.

Deux métriques ont été utilisées pour évaluer le processus : le rappel, qui indique la proportion de cas pour lesquels la forme cible correcte a été produite, l'ambiguïté, qui indique le nombre de formes cibles produites en moyenne.

Comme dans le chapitre précédent, nous n'avons pas utilisé le rappel et la précision pour l'évaluation car la référence ne contient qu'un seul verbe MSA. De plus, l'objectif de la conversion est plus de maximiser le rappel que de trouver un compromis raisonnable entre rappel et précision. En effet, la disponibilité de nombreuses ressources pour la désambiguïsation peut permettre de retrouver une forme correcte parmi plusieurs formes proposées. En revanche, l'absence de la forme TUN correcte dans les sorties du processus de conversion est irréparable.

Les expériences ont été réalisées dans le sens TUN→MSA et dans le sens MSA→TUN. Nous avons distingué les résultats sur les occurrences et sur les types. L'ensemble de développement a permis de combler quelques lacunes de l'analyseur et du générateur morphologique et d'enrichir le lexique de verbes.

L'évaluation a été réalisée sur les formes non diacritées bien que nous disposions des diacritiques des formes verbales aussi bien pour le TUN que le MSA. La raison pour cela est que les verbes dans la majortié des écrits arabes ne sont pas diacrités.

La première évaluation que nous avons faite consiste à ne pas réaliser de conversion. Le rappel est égal, dans ce cas, à $30,93\%$ sur les occurrences et à $29,44\%$ sur les types pour une ambiguïté de $1,0$. Cette référence indique le taux de formes fléchies verbales non diacritées TUN qui sont identiques aux formes du MSA dans l'ensemble de test.

Dans ce qui suit, nous présentons une série d'expériences avec des façons différentes de réalisation de transfert comme évoquée précédemment.

4.2.1. Transfert limité aux MBCs

Le processus de transfert le plus simple consiste à garder la racine source inchangée et à sélectionner la MBC cible par consultation de la matrice de correspondance de MBCs (cf. section 3.2.1). Cette expérience correspond à la situation pour laquelle nous ne disposons pas d'un lexique de transfert.

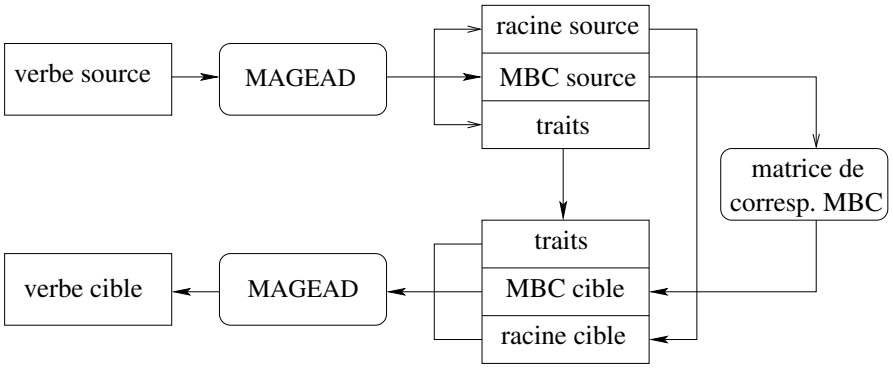

Figure 4.4.: processus de conversion d'une forme verbale source vers une forme verbale cible en utilisant une table de correspondance de MBCs

Le diagramme présenté dans la figure 4.4 présente ce processus de transfert. Les résultats de cette expérience sont donnés dans la table 4.1.

	rappel		ambiguïté	
	occurrences	types	occurrences	types
TUN→ MSA	$47,74\%$	$43,40\%$	$39,41$	$37,61$
MSA→ TUN	$52,55\%$	$48,05\%$	$5,89$	$7,12$

Table 4.1.: Rappel et ambiguïté dans l'ensemble de test en utilisant la matrice de correspondance de MBCs

Les résultats présentés dans la table 4.1 soulignent deux points importants. Premièrement, le rappel est assez bas, inférieur à 50%. En d'autre termes, en gardant la racine source pour produire la forme cible, nous obtenons uniquement une approximation grossière de cette dernière. Deuxièmement, l'ambiguïté dans le sens TUN→MSA est plus élevée que dans le sens MSA→TUN. Cela provient essentiellement du fait que le TUN ne distingue pas certains traits MSA comme le nombre duel ainsi que les genres au pluriel. À titre d'exemple, une forme verbale TUN fléchie au pluriel correspond à quatre formes verbales distinctes MSA fléchies duel masculin, duel féminin, pluriel masculin et pluriel féminin. L'absence de marquage du mode et du cas en TUN provoque une multiplication des formes cibles générées en MSA.

La même expérience a été réalisée en sélectionnant les deux MBCs cibles les plus probables étant donné la MBC source. La table 4.2 montre une légère augmentation du rappel. En effet, il s'élève à 51.65% sur les occurrences dans le sens TUN→MSA et $53,96\%$ dans le sens inverse. En revanche, l'ambiguïté augmente considérablement, le processus produit en moyenne environ 65 verbes MSA pour une occurrence en TUN.

	rappel		ambiguïté	
	occurrences	types	occurrences	types
TUN → MSA	51,65%	48,23%	66,98	64,69
MSA → TUN	53,96%	50,87%	9,81	10,68

Table 4.2.: Résultats sur l'ensemble de test en utilisant les deux MBCs cibles les plus fréquentes dans la matrice de correspondance de MBCs

4.2.2. transfert de MBCs et de racines d'une manière indépendante

Dans cette expérience, la MBC cible est sélectionnée, de la même manière que dans la dernière expérience, à l'aide de la matrice de correspondance de MBCs, néanmoins les racines cibles proviennent du lexique de racines (cf. section 3.2.1).

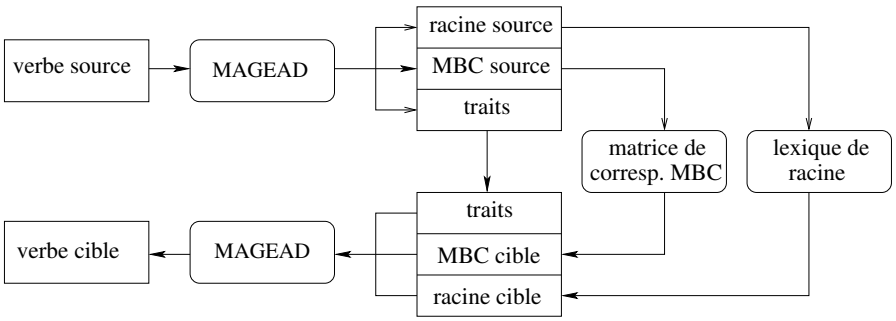

Figure 4.5.: Processus de conversion d'un verbe source vers un verbe cible à l'aide du lexique de racines et de la matrice de correspondance de MBCs

Le processus est décrit dans la figure 4.5. Les résultats sur l'ensemble de test sont donnés dans les tables 4.3 et 4.4.

	rappel		ambiguïté	
	occurrences	types	occurrences	types
TUN → MSA	68,98%	66,56%	74,37	72,89
MSA → TUN	72,37%	71,60%	13,70	14,52

Table 4.3.: Rappel et ambiguïté sur le corpus de test pour la conversion par le lexique de racines et la matrice de correspondance de mbcs

La table 4.3 montre une amélioration considérable du rappel. L'ambiguïté a également augmenté du fait de l'ambiguïté lexicale du lexique de racines. Rappelons qu'une racine TUN correspond, en moyenne, à 2,06 racines MSA. Ce nombre est égal à 1,26 dans le sens MSA→TUN.

En utilisant les deux MBCs cibles les plus fréquentes de la matrice de correspondance de MBCs, le processus de conversion provoque une augmentation du rappel et de l'ambiguïté, comme le montre la table 4.4.

	rappel		ambiguïté	
	occurrences	types	occurrences	types
TUN → MSA	81,77%	80,66%	126,44	122,45
MSA → TUN	86,12%	84,97%	21,92	22,56

Table 4.4.: Rappel et ambiguïté sur le corpus de test de la conversion en utilisant le lexique de racines et la table de correspondance de mbcs

Le rappel s'élève à $86,12\%$ pour les occurrences dans le sens MSA→TUN et atteint $81,77\%$ dans le sens inverse. En revanche, l'ambiguïté dépasse plus de 100 formes cibles dans le sens TUN→MSA.

4.2.3. Transfert de couples (racine, MBC)

Contrairement à l'expérience précédente où les racines et les MBCs cibles sont traduits d'une manière indépendante, le transfert, dans cette expérience consiste à utiliser les couples (racine, MBC) pour l'accès au lexique des verbes (Hamdi *et al.*, 2013b). Ce nouveau processus est décrit dans la figure 4.6 et les résultats apparaissent dans la table 4.5.

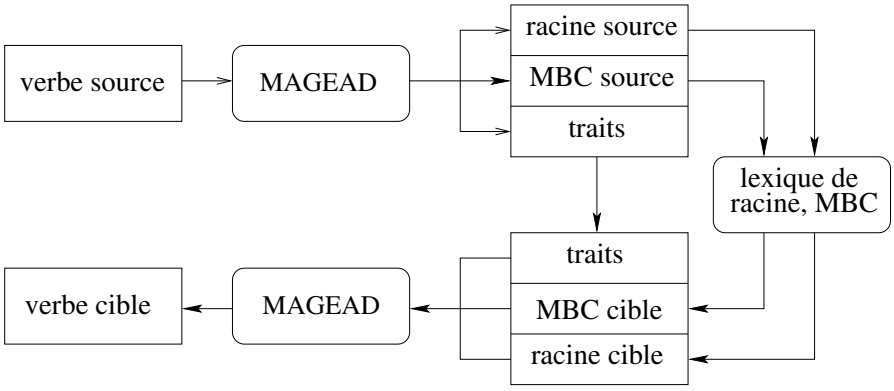

Figure 4.6.: Conversion d'un verbe source vers une forme cible par le lexique de racines et MBCs

La prise en compte simultanée d'une racine et d'une MBC lors d'une conversion a un effet positif sur la qualité du processus de conversion. La différence entre les résultats de cette expérience et de l'expérience précédente permet de quantifier ce que l'on perd en convertissant indépendamment la racine et la MBC.

	rappel		ambiguïté	
	occurrences	types	occurrences	types
TUN → MSA	76,43%	74,52%	26,82	25,57
MSA → TUN	79,24%	75,10%	1,47	3,10

Table 4.5.: Rappel et ambiguïté sur l'ensemble de test de la conversion par un lexique de racines et MBCs

Le principal inconvénient de cette méthode réside dans la couverture lexicale. En effet, la couverture du lexique n'étant pas parfaite, dans certains cas, l'accès au lexique échoue.

Afin de quantifier l'impact de la couverture lexicale, nous avons utilisé le corpus de développement. Celui-ci nous a servi à enrichir le lexique de sorte que l'accès lexical n'échoue jamais et produit toujours une cible correcte au couple (racine, MBC). Les résultats de cette expérience, quoique artificiels, permettent d'estimer une borne supérieure de notre processus. Dans le sens TUN→MSA, le rappel dans les occurrences s'élève à $87,65\%$ et $89,56\%$ dans le sens inverse.

La raison pour laquelle nous n'avons pas obtenu un rappel parfait (de 100%) dans cette expérience revient au fait que les deux systèmes morphologiques du TUN et du MSA ne produisent pas toujours les sorties correctes à l'issue de l'analyse et de la génération. Une analyse d'erreurs dans le sens TUN→MSA a montré que $21,8\%$ des erreurs proviennent du système de génération du MSA et $78,2\%$ du système d'analyse TUN. La plupart des erreurs sont dues aux phénomènes morphologiques non implémentés.

4.2.4. transfert de couples (racine, MBC) avec repli

Cette variante a pour objectif de pallier le problème de couverture. Dans le cas où le couple (racine, MBC) est absent du lexique, nous nous servons du lexique de racines et de la matrice de correspondance de MBCs pour sélectionner la racine et la MBC cible.

L'architecture du système est décrite dans la figure 4.7. Les traits interrompus représentent le chemin suivi par le système en cas de repli.

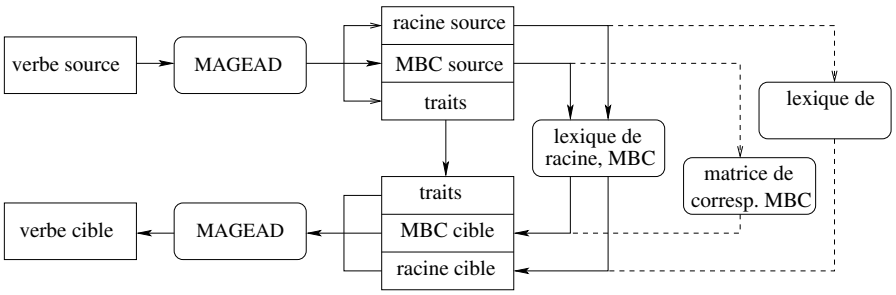

Figure 4.7.: Processus de conversion d'un verbe source vers une forme cible en utilisant un lexique de racines et MBCs avec repli

La table 4.6 montre que cette méthode augmente le rappel considérablement. Cette augmentation est elle-même le résultat d'une meilleure couverture. L'ambiguïté a également augmenté, cela est dû au fait que le processus de repli a tendance à augmenter l'ambiguïté, comme nous l'avons observé dans la section 4.2.2.

	rappel		ambiguïté	
	occurrences	types	occurrences	types
TUN → MSA	$79,71\%$	$78,94\%$	$29,16$	$28,44$
MSA → TUN	$84,83\%$	$84,03\%$	$3,47$	$4,95$

Table 4.6.: Rappel et ambiguïté sur l'ensemble de test en utilisant le lexique de racines et MBCs avec repli

C'est la sortie de ce processus de conversion que nous avons choisi pour la suite de nos expériences.

4.3. Désambiguïsation

Pour réaliser la désambiguïsation des solutions proposées à l'issue de la conversion, nous avons eu recours à deux moyens.

Le premier consiste à fournir en entrée à l'étiqueteur en parties de discours l'ensemble des solutions proposées par le processus de conversion. En effet, comme nous l'avons évoqué dans la section 3.3, l'étiqueteur peut prendre en entrée un ensemble ambigu de séquences d'observables représentées sous la forme d'un automate acyclique. Dans ce cas, la désambiguïsation est réalisée par l'étiqueteur.

Le deuxième moyen consiste à désambiguïser d'une manière indépendante de l'étiquetage. Elle repose sur des modèles de langage entraînés sur des corpus MSA. Ces derniers permettent d'associer des probabilités à chacune des solutions proposées par le processus de conversion.

Les modèles de langage peuvent être représentés sous la forme d'automates pondérés. Ainsi, le processus de désambiguïsation est réalisé par la composition de l'automate issu de la conversion suivi par une recherche de meilleur chemin.

Plusieurs modèles de langage ont été construits à partir de trois corpus : le premier (C_1) est formé de dépêches de presse de l'Agence France Presse (AFP). Le deuxième (C_2) est une collection de transcriptions de débats politiques d'*Aljazeera*. Enfin, le troisième (C_3) constitue l'union des deux corpus. Ce dernier couvre, par conséquent, les deux variantes orale et écrite du MSA. Chaque corpus est décomposé en deux parties : un ensemble d'entraînement et un ensemble d'évaluation.

La table 4.7 donne les tailles des corpus.

		entraînement	évaluation
C_1	occurrences	1 550 713	9 077
	types	39 877	2 636
C_2	occurrences	900 109	21 658
	types	32 959	2 073
C_3	occurrences	2 450 822	30 735
	types	54 721	4 146

Table 4.7.: Comptes sur les corpus des modèles de langage

Les probabilités des modèles de langage ont été calculées à l'aide de l'outil SRILM (Stolcke *et al.*, 2002). Dans le cas de N-grams inconnus, nous avons utilisé la technique standard du repli qui consiste à diminuer l'historique d'un N-gram tant qu'il n'a pas été trouvé dans le modèle de langage.

Neuf modèles de langage ont été construits, qui se distinguent par leur ordre (1-*gram*, 2-*gram* et 3-*gram*) et par le corpus sur lequel leurs probabilités ont été estimées. Le modèle 1-*gram* n'a pas été utilisé pour réaliser la désambiguïsation, il a servi à pondérer les mots de l'automate issu de la conversion lorsque la désambiguïsation est réalisée par l'étiqueteur, ceci dans le but de pénaliser les mots rares ou incorrects.

Afin d'évaluer le processus de désambiguïsation, nous avons eu recours au calcul de la couverture des modèles de langage sur les corpus d'évaluation d'une part et à la mesure de la perplexité d'autre part. Cette dernière consiste à mesurer la capacité de prédiction des modèles de langage. En d'autres termes, étant donné un mot à prédire, la valeur de la perplexité représente le nombre moyen d'hypothèses associées à un mot dans un contexte donné. Ainsi, l'efficacité d'un modèle de langage est inversement proportionnelle à la valeur de la perplexité.

La table 4.8 donne les différentes valeurs de perplexité des neuf modèles de langage implémentés ainsi que le nombre de mots hors-vocabulaire (OOVs[a]) de

a. Cet ensemble représente les mots qui n'appartiennent pas au lexique d'entraînement des modèles de langage.

l'ensemble de test (nous donnons entre parenthèses le pourcentage de OOVs dans le corpus d'évaluation).

		perplexité	OOVs
C_1	1-gram	1290, 47	128 (1,4%)
	2-gram	282, 26	
	3-gram	284, 50	
C_2	1-gram	1054, 81	349 (1,6%)
	2-gram	177, 00	
	3-gram	169, 29	
C_3	1-gram	1262, 49	295 (0,5%)
	2-gram	245, 64	
	3-gram	241, 25	

Table 4.8.: Évaluation des modèles de langage

Les résultats présentés dans la table 4.8 montrent que quel que soit le corpus d'apprentissage, la perplexité est légèrement variable entre les modèles. Nous nous sommes donc basés sur la couverture pour effectuer le choix des modèles. Nous réalisons ainsi la désambiguïsation à l'aide des modèles entraînés sur le corpus C_3.

Suite à la désambiguïsation à l'aide des modèles de langage entraînés sur c_3, trois sorties peuvent être fournies à l'étiquetage : la meilleure séquence prédite par chacun des modèles 2-gram et 3-gram et l'automate pondéré à l'aide du modèle 1-gram.

4.4. Étiquetage en parties de discours

Dans cette section, nous décrivons les différentes expériences d'étiquetage du dialecte TUN. Pour cela, nous avons fait varier :

– la nature des entrées de l'étiqueteur qui peuvent être des formes ou des lemmes,
– les pré-traitements réalisés avant l'étiquetage

Concernant les pré-traitements, nous avons d'abord essayé d'étiqueter des données TUN sans aucun pré-traitement (1), puis des données converties et désambiguïsées à l'aide de modèles de langage (2) et enfin des données converties mais non désambiguïsées (3). Chacune des expériences (1), (2) et (3) font l'objet d'une sous-section.

4.4.1. Étiquetage sans conversion

La première expérience que nous avons réalisé consiste à appliquer l'étiqueteur en parties de discours décrit dans la section 3.3 sur des données TUN sans au-

cun traitement préalable. Cette expérience constitue la borne inférieure de notre évaluation et permettra par la suite de calculer la contribution de la conversion sur la qualité d'étiquetage.

Les performances d'étiquetage sur le TUN et le nombre de (OOVs) sont donnés dans la table 4.9. Nous donnons pour rappel dans cette table les résultats d'étiquetage du MSA à l'aide du même outil.

	MSA			TUN		
	formes	lemmes	LMMs	formes	lemmes	LMMs
performance (%)	94,72	97,63	96,94	69,04	67,41	71,41
OOVs	158	47	42	2891	4766	2705
(%)	0,57	0,16	0,15	26,9	44,35	25,17

Table 4.9.: Résultats d'étiquetage avant la conversion

Nous donnons également les résultats d'étiquetage des lemmes et des LMMs. Ces derniers sont les formes non diacritées des lemmes. Comme nous l'avons indiqué dans la section 4.2, l'un des avantages de notre système de conversion est qu'il peut générer à partir d'une forme source, les lemmes et les LMMs cibles.

Dans cette expérience, nous donnons les résultats d'étiquetage des lemmes et des LMMs de référence puisque nous ne disposons pas d'un système de lemmatisation du TUN. Les résultats de l'étiquetage en parties de discours des lemmes et des LMMs sont donnés pour des raisons comparatives. Nous les comparons aux résultats finaux de notre expérience sur les lemmes et les LMMs prédits par le système de conversion.

Les résultats montrent que l'étiquetage le plus performant est donné sur des LMMs du côté TUN et sur des lemmes du côté MSA. Ces résultats sont artificiels étant donné que nous avons utilisé les lemmes et les LMMs de référence. La tâche réelle est l'étiquetage des formes. Les résultats d'étiquetage des formes atteignent 69% du côté TUN et 94% du côté MSA. Ces résultats constituent les bornes de notre expérience.

La table 4.9 montre également l'intersection importante des lexiques TUN et MSA. Environ 75% de formes et des LMMs TUN appartiennent au lexique MSA. Ce taux n'est que de 55.65% pour les lemmes, ce qui était prévisible dans la mesure où, contrairement aux formes et aux LMMs, les lemmes sont entièrement diacritées.

La deuxième expérience que nous avons menée consiste à diviser notre corpus d'évaluation (cf. section 3.4) en deux ensembles : une ensemble d'entraînement composé de 600 phrases et un ensemble de test contenant 200 phrases. Comme nous l'avons indiqué dans le chapitre précédent, le corpus d'évaluation représente une collection de phrases extraites de quatre sources différentes : des séries télévisées, des débats politiques, une pièce de théâtre et un corpus transcrit à partir des enregistrements de discussions entre des clients et un agent de

la société nationale tunisienne des chemins ferrés. 150 phrases de chaque domaine ont été sélectionnées pour la construction du corpus d'entraînement et 50 phrases de chaque domaine ont construit l'ensemble de test.

Bien que la taille du corpus d'entraînement n'est pas suffisante pour l'apprentissage d'un étiqueteur robuste, nous avons réalisé cette expérience afin d'estimer les performances d'étiquetage du TUN à l'aide d'un étiqueteur entraîné sur le TUN.

	formes	lemmes	LMMs
précision (%)	71, 53	76, 74	84, 43
OOVs	1954	1633	980
(%)	58, 49	48, 88	29, 33

Table 4.10.: Résultats d'étiquetage du TUN

Les résultats de la table 4.10 montrent que l'entraînement de l'étiqueteur sur du TUN pour le traitement du TUN ne permet pas d'obtenir des résultats équivalents à ceux obtenus avec du MSA pour l'étiquetage du MSA. Cela est certainement du à la faible quantité de données utilisées dans l'entraînement du TUN comparativement au MSA (cf. table 4.9). De plus, nous remarquons que presque 50% des mots du corpus de test n'ont pas été observés lors de la phase d'entraînement.

4.4.2. Étiquetage après désambiguïsation à l'aide de modèles de langage

Notre expérience principale consiste à effectuer la conversion du TUN en pseudo-MSA avant le processus d'étiquetage. La conversion génère trois automates acycliques sur les formes, sur les lemmes et sur les LMMs.

L'automate composé de formes a été désambiguïsé à l'aide de trois modèles de langage qui varient selon l'ordre (1, 2, 3). Ainsi, trois entrées différentes composées de formes ont été fournies à l'étiqueteur.

- un automate pondéré avec des scores attribués aux formes grâce au modèle 1-gram
- la meilleure séquence de formes calculée à l'aide du modèle 2-gram
- la meilleure séquence donnée par le modèle 3-gram.

Pour désambiguïser les automates de lemmes et de LMMs, nous avons utilisé les résultats des modèles de langage entraînés sur les formes, en remplaçant chaque forme par le lemme (ou LMM) correspondant.

La figure 4.8 décrit la chaîne de traitement pour réaliser l'étiquetage en parties de discours du TUN.

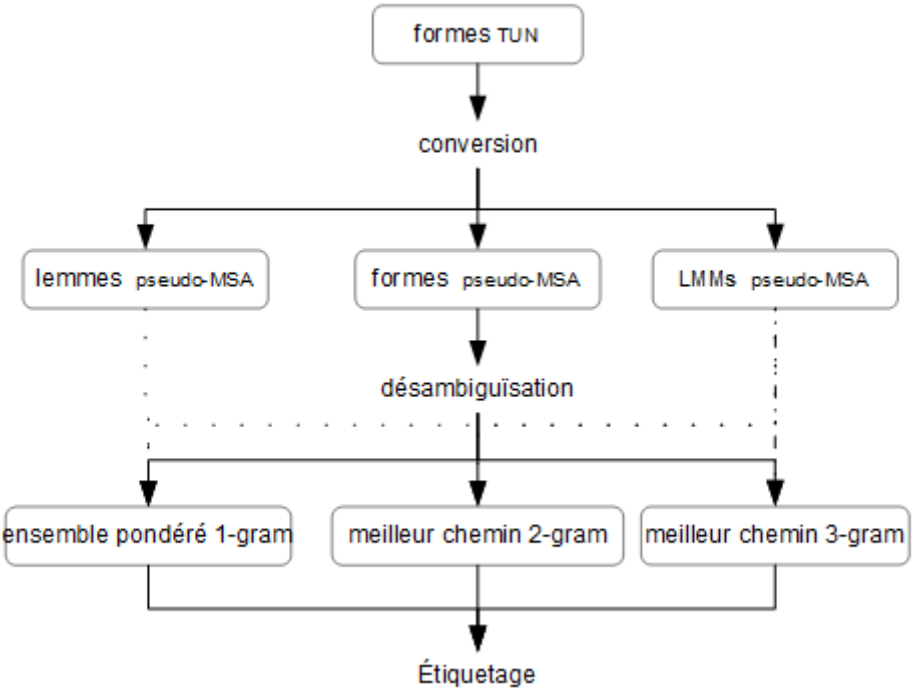

Figure 4.8.: Étiquetage en parties de discours du tun avant la conversion

La sortie finale de notre système présente la séquence des parties de discours pour la phrase source. Les résultats sont donnés dans la table 4.11.

	formes	lemmes	LMMs
1-gram	$77,21\%$	$78,87\%$	$79,35\%$
2-gram	$80,33\%$	$82,67\%$	$83,17\%$
3-gram	$80,51\%$	$82,32\%$	$83,25\%$

Table 4.11.: Résultats de l'étiquetage du pseudo-msa après désambiguïsation

Les résultats montrent que la conversion permet d'améliorer considérablement les résultats de l'étiquetage en parties de discours. En effet, ils passent de 69% à 80% sur les formes. Deux points importants méritent d'être notés. D'une part, la désambiguïsation à l'aide des modèles 3-gram est la plus performante. D'autre part, l'application de l'étiqueteur sur les LMMs est meilleure que les formes et les lemmes. Les résultats d'étiquetage de LMMs fournit une précision de ($\sim 83.5\%$).

4.4.3. Étiquetage en parties de discours sans désambiguïsation

La dernière expérience réalisée consiste à fournir à l'étiqueteur la sortie du processus de conversion. L'étiqueteur effectue ainsi la désambiguïsation et l'étiquetage en parties de discours simultanément.

Comme nous l'avons indiqué, la conversion génère trois automates qui varient selon la nature des sorties : des formes, des lemmes et des LMMs. La figure 4.9 décrit le nouveau processus d'étiquetage.

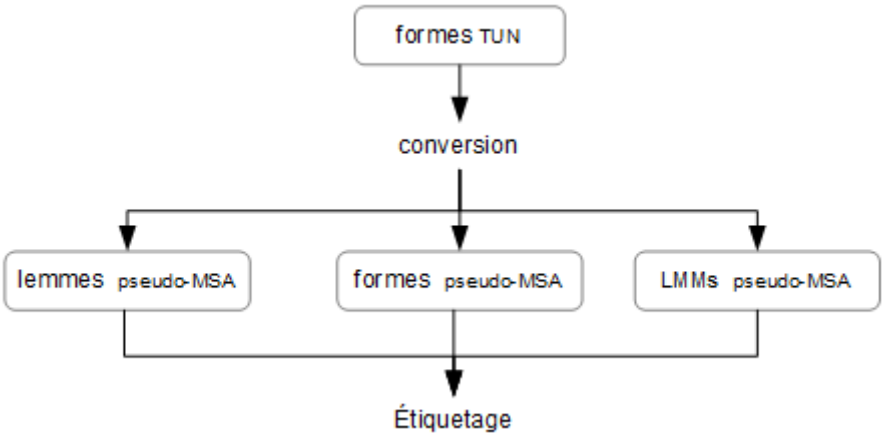

Figure 4.9.: Étiquetage en parties de discours des lemmes et des LMMs en pseudo-MSA

Dans le table 4.12, nous présentons les résultats de l'étiquetage en parties de discours de chaque sortie générée par le processus de conversion.

	formes	lemmes	LMM*s*
précision (%)	82.5%	86.9%	89.1%
OOVs	1456	669	538%
(%)	13.5%	6.2%	4.9%

Table 4.12.: Résultats de l'étiquetage du pseudo-MSA sans désambiguïsation

Comme le montre la table 4.12, l'approche qui consiste à ne pas désambiguïser préalablement à l'étiquetage obtient de meilleures performances. Cela est du à la différence d'origine des corpus d'entraînement des modèles de langage (MSA) et le corpus que nous souhaitons analyser (pseudo-MSA issu du TUN oral). On observe également que coupler les étapes de désambiguïsation et d'étiquetage améliore significativement les résultats. Ceci provient probablement du fait que chaque tâche compense certaines erreurs de l'autre.

Elle montre aussi qu'il est préférable de réaliser l'étiquetage sur les LMMs plutôt que sur les lemmes ou les formes. L'étiqueteur des LMMs permet en effet d'obtenir une précision de 89.1%. Ce résultat présente une augmentation absolue de 20% par rapport aux résultats de la première expérience (étiquetage du TUN à l'aide d'un étiqueteur MSA).

Afin de déterminer les sources d'erreurs, nous avons réalisé une analyse d'erreurs sur l'étiquetage des formes, des lemmes et des LMMs. La table 4.13 décrit le taux d'erreurs provenant de chaque processus de traitement à savoir la conversion et l'étiquetage. Naturellement, nous n'attribuons l'erreur à l'étiquetage que lorsque la conversion est correcte. Nous avons choisi de séparer les erreurs provenant de la phase d'étiquetage en deux types afin d'identifier si les erreurs sont des erreurs de désambiguïsation ou d'étiquetage (si la désambiguïsation est correcte).

	formes	lemmes	LMM*s*
conversion	62%	46%	34%
désambiguïsation	27%	39%	49%
étiquetage	11%	15%	17%

Table 4.13.: Analyse d'erreurs d'étiquetage du pseudo-msa

La table 4.13 montre qu'une conversion incorrecte est nocive pour la qualité de l'étiquetage en parties de discours. Une meilleure conversion fournit un meilleur étiquetage. Au niveau des LMMs 34% des erreurs proviennent de la conversion. Dans 49% cas, les erreurs proviennent de la désambiguïsation. Enfin, dans 17% cas, les erreurs sont issues de l'étiquetage en parties de discours, c'est-à-dire le système de conversion a généré un ensemble ambigu contenant un LMM correct. L'étiqueteur a sélectionné le bon LMM suite à la désambiguïsation mais lui a assigné une étiquette incorrecte.

Conclusion

Dans ce chapitre, nous avons comparé de multiples méthodes d'étiquetage en parties de discours du TUN à l'aide d'un étiqueteur entraîné sur le MSA. Avant l'étiquetage, le texte TUN est traduit en pseudo-MSA. Le processus de conversion est composé de trois étapes : une analyse morphologique d'un mot TUN source, suivi d'un transfert lexical et une génération morphologique des formes cibles MSA. Le système atteint une précision de 89% ($\sim$20% d'amélioration absolue par rapport à la précision donnée par l'étiquetage du TUN sans pré-traitement). Les expériences ont montré que les meilleurs résultats sont obtenus sur l'étiquetage des lemmes et plus précisément ceux non diacrités.

Conclusion générale et perspectives

Bilan de la thèse

Dans ce travail, nous avons proposé une méthode générique pour le traitement automatique des dialectes arabes. La méthode consiste à de mettre à profit les nombreux outils et ressources du MSA pour réaliser le traitement de dialectes. Nous nous sommes intéressés en particulier au dialecte tunisien.

Notre point de départ est de montrer que le traitement automatique d'une langue peu dotée L_1 peut être réalisé à l'aide de ressources et d'outils d'une deuxième langue mieux dotée, étymologiquement proche, L_2. Afin de traiter L_1, un processus de conversion convertit L_1 en une approximation de L_2. L'hypothèse sous-jacente est que le coût de développement d'un convertisseur de L_1 en L_2 est inférieur au coût de développement des ressources nécessaires pour produire directement des outils pour L_1.

Le processus de conversion que nous proposons repose sur l'analyse morphologique profonde et le transfert lexical. Pour cela, nous avons développé un analyseur morphologique du TUN ainsi qu'un lexique TUN-MSA. Une des originalité de notre approche est le recours à une analyse morphologique profonde qui analyse une forme agglutinée en une racine, un schème et des traits morphologiques. On peut alors réaliser le transfert au niveau des racines et des schèmes. Cela permet d'une part de se restreindre, si nécessaire à un lexique de racines, dont la taille est réduite et, d'autre part, cela permet, lors du processus de génération en L_2 de ne pas aller jusqu'à la génération des formes agglutinées et de s'arrêter avant. C'est à ce niveau qu'est alors appliqué l'outil pour $L2$.

Nous avons validé notre méthode sur la tâche d'étiquetage morphosyntaxique. Nous avons pour cela utilisé un étiqueteur MSA fondé sur les chaînes de Markov cachées dont nous avons estimé les paramètres sur un corpus étiqueté MSA. À l'issue de l'étiquetage, les étiquettes sont projetées sur les mots TUN. Afin d'évaluer la qualité de l'étiquetage ainsi produit, nous avons collecté, transcrit, segmenté

et annoté en partie de discours un corpus tunisien. Ce corpus est composé de 800 phrases et d'environ $10,000$ occurrences. Il couvre plusieurs variantes du TUN.

La première expérience que nous avons menée consiste à réaliser l'étiquetage en partie de discours du TUN avec un outil entraîné sur le MSA sans aucun traitement préalable du TUN. Cette expérience donne des performances d'étiquetage qui atteignent 69%. Ce résultat constitue la borne inférieure des résultats de nos expérimentations.

Nous avons ensuite réalisé la conversion du TUN en pseudo-MSA. La conversion permet de générer un corpus MSA ambigu. En effet, chaque mot source TUN est traduit par un ensemble de mots cibles en MSA. Les performances du processus de conversion ont été évaluées sur environ $1,500$ verbes TUN en utilisant deux métriques, le rappel, qui atteint 80% et l'ambiguïté, qui produit en moyenne 28 formes cibles pour une forme source.

L'ensemble de mots cibles constitue un ensemble ambigu qui nécessite un traitement de désambiguïsation. Cette opération peut être réalisée à l'aide de l'étiqueteur en parties de discours qui autorise l'étiquetage d'une entrée ambiguë. La désambiguïsation peut aussi être effectuée indépendamment de l'étiquetage, à l'aide d'un modèle de langage entraîné sur du corpus MSA dont on peut disposer en grande quantité. Le modèle de langage permet d'attribuer des poids aux différentes formes qui compose la sortie de la conversion ou de réaliser la désambiguïsation proprement dit.

La troisième étape de notre méthode repose sur l'étiquetage en parties de discours. Nous avons réalisé l'étiquetage sur trois entrées différentes :
- l'ensemble ambigu non pondéré généré par la conversion
- l'ensemble ambigu pondéré par les scores d'un modèle de langage
- le meilleur chemin proposé par le modèle de langage

L'étiquetage de ces données a donné environ 82%, 80% et 77% respectivement. Comme on peut l'observer, la désambiguïsation à l'aide d'un modèle de langage dégrade les performances. Ces résultats s'expliquent probablement par la différence du genre du corpus utilisé pour l'apprentissage des modèles de langage d'une part et, d'autre part, du corpus d'évaluation TUN.

Comme nous l'avons mentionné ci-dessus, un avantage de notre méthode est de pouvoir arrêter le processus de génération morphologique avant son terme et de produire ainsi des lemmes ou des LMMs qui sont des formes non diacritées des lemmes. L'étiquetage peut alors être réalisé sur ces derniers, à la condition, bien entendu, de ré-entraîner l'étiqueteur sur des données de cette nature. L'étiquetage des lemmes a permis d'atteindre une précision de 86% et celui des LMMs

une précision de 89%. Ceci constitue notre meilleur résultat, il est supérieur de 20 points au résultat de l'étiquetage sans conversion.

La mise en œuvre de notre méthode a nécessité le développement d'un analyseur morphologique du tunisien ainsi qu'un lexique TUN-MSA. Il est malheureusement difficile de quantifier le coût d'un tel développement car il a été réalisé en même temps que la mise point de notre modèle. Nous pensons que le temps de développement de telles ressources pour un autre dialecte de l'arabe devrait être bien inférieur à celui du développement de corpus annotés.

Perspectives

Plusieurs perspectives s'ouvrent à nous à l'issue de ce travail. Nous en développons trois dans les paragraphes suivants.

Le dialecte tunisien étant avant tout oral. Traiter des transcriptions manuelles constitue un objet artificiel. Notre travail trouvera toute sa justification lorsqu'il sera possible de prendre en entrée les sorties d'un système de transcription automatique du tunisien. Nous n'avons malheureusement pas pour l'instant à notre disposition un tel système.

Nous avons validé notre méthode sur la tâche d'étiquetage morphosyntaxique, qui a l'avantage d'être à la fois simple et utile. Dans le but de mieux valider notre méthode, nous envisageons de recourir à d'autres outils standard de traitement automatique des langues, tel que l'analyse syntaxique. Si les résultats sur cette tâche (ou d'autres) montre aussi de bons résultats, notre méthode prendra alors toute sa justification car l'utilisation de nouveaux outils ne nécessitera aucune modification à notre système, alors que le développement de ressources en TUN des ressources nécessaires pour entraîner de tels outils, un corpus annoté en syntaxe dans le cas de l'analyse syntaxique est une entreprise coûteuse.

Finalement, d'autres dialectes arabes peuvent être traités selon le même principe. Nous disposons en particulier d'une implémentation de l'analyseur/générateur morphologique MAGEAD pour le levantin et l'égyptien. Il ne reste donc plus qu'à développer des lexiques pour ces dialectes de l'arabe pour reproduire sur ces derniers les expériences que nous avons réalisées sur le tunisien.

Bibliographie

AL-DAHDAH, A. (1996). *muʕjam qawAʕid Alluγaħ Alʕarabyyaħ fiy jadAwil wa lawHAt* معجم قوَاعد الّغة العربيّة في جدَول ولوحَات. maktabaħ lubnAn nAširuwn, Beyrouth, Liban مكتبة لبنَان نَاشرون، بيروت، لبنَان.

AL-GHULAYAINI, M. (2006). *jAmʕ Aldrws Alʕrbyħ, Part II* جَامع الدروس العربية. dAr Alktb Alʕlmyyħ, Beyrouth, Liban دَار الكتب العلمية، بيروت، لبنَان.

AL-SABBAGH, R. et GIRJU, R. (2012). A supervised pos tagger for written arabic social networking corpora. *In Proceedings of KONVENS*, pages 39–52.

ALTABBAA, M., AL-ZARAEE, A. et SHUKAIRY, M. (2010). An arabic morphological analyser and part-of-speech tagger. *Actes de JADT*, page 50.

ALTANTAWY, M., HABASH, N., RAMBOW, O. et SALEH, I. (2010). Morphological analysis and generation of arabic nouns : A morphemic functional approach. *In LREC*.

BASSIOUNEY, R. (2009). *Arabic sociolinguistics*. Edinburgh University Press.

BEESLEY, K. R. (1998). Arabic morphological analysis on the internet. *In Proceedings of the 6th International Conference and Exhibition on Multi-lingual Computing*. Citeseer.

BEESLEY, K. R. (2001). Finite-state morphological analysis and generation of arabic at xerox research : Status and plans in 2001. *In ACL Workshop on Arabic Language Processing : Status and Perspective*, volume 1, pages 1–8.

BERNHARD, D., LIGOZAT, A.-L. *et al.* (2013). Hassle-free pos-tagging for the alsatian dialects. *Non-Standard Data Sources in Corpus Based-Research*.

BLACK, A., RITCHIE, G., PULMAN, S. et RUSSELL, G. (1987). Formalisms for morphographemic description. *In Proceedings of the third conference on European chapter of the Association for Computational Linguistics*, pages 11–18. Association for Computational Linguistics.

BOUJELBANE, R., BENAYED, S. et BELGUITH, L. H. (2013). Building bilingual lexicon to create dialect tunisian corpora and adapt language model. *ACL 2013*, page 88.

BOUJELBANE, R., MALLEK, M., ELLOUZE, M. et BELGUITH, L. H. (2014). Fine-grained pos tagging of spoken tunisian dialect corpora. *In Natural Language Processing and Information Systems*, pages 59–62. Springer.

BRUSTAD, K. (2000). *The syntax of spoken Arabic : A comparative study of Moroccan, Egyptian, Syrian, and Kuwaiti dialects*. Georgetown University Press.

BUCKWALTER, T. (2002). Buckwalter {Arabic} morphological analyzer version 1.0.

BUCKWALTER, T. (2004). Buckwalter arabic morphological analyzer version 2.0. ldc catalog number ldc2004l02. Rapport technique, ISBN 1-58563-324-0.

COHEN, D. (1970). Essai d'une analyse automatique de l'arabe. *Etudes de linguistique sémitique et arabe*, pages 49–78.

DANIELS, P. T. (2007). Mélanges david cohen : Études sur le langage, les langues, les dialectes, les littératures, offertes par ses élèves, ses collègues, ses amis ; présentés à l'occasion de son quatre-vingtième anniversaire (review). *Language*, 83(1):221–222.

DAS, D. et PETROV, S. (2011). Unsupervised part-of-speech tagging with bilingual graph-based projections. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies-Volume 1*, pages 600–609. Association for Computational Linguistics.

DEBILI, F. et ACHOUR, H. (1998). Voyellation automatique de l'arabe. *In Proceedings of the Workshop on Computational Approaches to Semitic Languages*, pages 42–49. Association for Computational Linguistics.

DHOUIB, E. (2007). El makki w zakiyya. Maison d'édition manshuwrat manara, Tunis.

DICHY, J., BRAHAM, A., GHAZALI, S. et HASSOUN, M. (2002). La base de connaissances linguistiques diinar. 1 (dictionnaire informatisé de l'arabe, version 1). *In Proceedings of the International Symposium on The Processing of Arabic, Tunis (La Manouba University)*, pages 18–20.

DUONG, L., COOK, P., BIRD, S. et PECINA, P. (2013). Simpler unsupervised pos tagging with bilingual projections. *In ACL (2)*, pages 634–639.

ELGOT, C. C. et MEZEI, J. E. (1965). On relations defined by generalized finite automata. *IBM Journal of Research and Development*, 9(1):47–68.

FELDMAN, A., HANA, J. et BREW, C. (2006). A cross-language approach to rapid creation of new morpho-syntactically annotated resources. *In Proceedings of LREC*, pages 549–554.

FERGUSON, C. A. (1959). Diglossia. *WORD-JOURNAL OF THE INTERNATIONAL LINGUISTIC ASSOCIATION*, 15(2):325–340.

GARRETTE, D. et BALDRIDGE, J. (2013). Learning a part-of-speech tagger from two hours of annotation. *In HLT-NAACL*, pages 138–147. Citeseer.

GOOSSE, A. et GREVISSE, M. (1993). *Le bon usage : grammaire française*. Duculot.

GRAFF, D., MAAMOURI, M., BOUZIRI, B., KROUNA, S., KULICK, S. et BUCKWALTER, T. (2009). Standard Arabic Morphological Analyzer (SAMA) Version 3.1. Linguistic Data Consortium LDC2009E73.

HABASH, N. (2004). Large scale lexeme based arabic morphological generation. *In Proceedings of Traitement Automatique du Langage Naturel (TALN-04)*.

HABASH, N. (2010). *Introduction to Arabic Natural Language Processing*. Morgan & Claypool Publishers.

HABASH, N., DIAB, M. T. et RAMBOW, O. (2012). Conventional orthography for dialectal arabic. *In LREC*, pages 711–718.

HABASH, N. et RAMBOW, O. (2005). 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*, pages 573–580. Association for Computational Linguistics.

HABASH, N. et RAMBOW, O. (2006). Magead : a morphological analyzer and generator for the arabic dialects. *In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 681–688. Association for Computational Linguistics.

HABASH, N., RAMBOW, O. et KIRAZ, G. (2005). Morphological analysis and generation for arabic dialects. *In Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, pages 17–24. Association for Computational Linguistics.

HABASH, N., RAMBOW, O. et ROTH, R. (2009). Mada+ tokan : A toolkit for arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization. *In Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt*, pages 102–109.

HABASH, N. et ROTH, R. M. (2009). Catib : The columbia arabic treebank. *In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 221–224. Association for Computational Linguistics.

HABASH, N., SOUDI, A. et BUCKWALTER, T. (2007). On Arabic Transliteration. *In* van den BOSCH, A. et SOUDI, A., éditeurs : *Arabic Computational Morphology : Knowledge-based and Empirical Methods*. Springer.

HAMDI, A. (2012). Apport de la diacritisation dans l'analyse morphosyntaxique de l'arabe. *JEP-TALN-RECITAL 2012*, page 247.

HAMDI, A., BOUJELBANE, R., HABASH, N. et NASR, A. (2013a). The effects of factorizing root and pattern mapping in bidirectional tunisian - standard arabic machine translation. *In MT Summit, Nice*.

HAMDI, A., BOUJELBANE, R., HABASH, N. et NASR, A. (2013b). Un système de traduction de verbes entre arabe standard et arabe dialectal par analyse morphologique profonde. *In Traitement Automatique des Langues Naturelles, Les Sables d'Olonnes*, page 395–406.

HAMDI, A., GALA, N. et NASR, A. (2014). Automatically building a tunisian lexicon for deverbal nouns. *COLING 2014*, page 95.

HOLES, C. (2004). *Modern Arabic : Structures, functions, and varieties*. Georgetown University Press.

JURAFSKY, D. et MARTIN, J. H. (2000). *Speech & language processing*. Pearson Education India.

KARTTUNEN, L. (1995). The replace operator. *In Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 16–23. Association for Computational Linguistics.

KAY, M. (1987). Nonconcatenative finite-state morphology. *In Proceedings of the third conference on European chapter of the Association for Computational Linguistics*, pages 2–10. Association for Computational Linguistics.

KIRAZ, G. A. (1994). Computational analyses of arabic morphology. *arXiv preprint cmp-lg/9408002*.

KOSKENNIEMI, K. (1983). Two-level model for morphological analysis. *In IJCAI*, volume 83, pages 683–685.

LARCHER, P. (2012). *Le système verbal de l'arabe classique, 2ème édition, revue et augmentée*. Presses universitaires de Provence.

LI, S., GRAÇA, J. V. et TASKAR, B. (2012). Wiki-ly supervised part-of-speech tagging. *In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1389–1398. Association for Computational Linguistics.

MAAMOURI, M., BIES, A., BUCKWALTER, T. et MEKKI, W. (2004). The penn arabic treebank : Building a large-scale annotated arabic corpus. *In NEMLAR conference on Arabic language resources and tools*, pages 102–109.

MASMOUDI, A., ELLOUZE KHMEKHEM, M., ESTÈVE, Y., HADRICH BELGUITH, L. et HABASH, N. (2014). A corpus and a phonetic dictionary for tunisian arabic speech recognition. *In of the Language Resources and Evaluation Conference, Iceland*.

MCCARTHY, J. J. (1981). A prosodic theory of nonconcatenative morphology. *Linguistic inquiry*, pages 373–418.

MCCARTHY, J. J. (1993). Template form in prosodic morphology. pages 187–218.

MEJRI, S., MOSBAH, S. et SFAR, I. (2009). Pluringuisme et diglossie en tunisie. *Synergies Tunisie n 1*, pages 53–74.

MOHAMED, E., MOHIT, B. et OFLAZER, K. (2012). Annotating and learning morphological segmentation of egyptian colloquial arabic. *In LREC*, pages 873–877.

MOHRI, M., PEREIRA, F. et RILEY, M. (2000). The design principles of a weighted finite-state transducer library. *Theoretical Computer Science*, 231(1):17–32.

NELKEN, R. et SHIEBER, S. M. (2005). Arabic diacritization using weighted finite-state transducers. *In Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, pages 79–86. Association for Computational Linguistics.

OUERHANI, B. (2009). Interférence entre le dialectal et le littéral en tunisie : Le cas de la morphologie verbale. *Synergies Tunisie n 1*, pages 75–84.

PASHA, A., AL-BADRASHINY, M., KHOLY, A. E., ESKANDER, R., DIAB, M., HABASH, N., POOLEERY, M., RAMBOW, O. et ROTH, R. (2014). Madamira : A fast, comprehensive tool for morphological analysis and disambiguation of arabic. *In In Proceedings of the 9th International Conference on Language Resources and Evaluation, Reykjavik, Iceland*.

PULMAN, S. G. et HEPPLE, M. R. (1993). A feature-based formalism for two-level phonology : a description and implementation. *Computer Speech & Language*, 7(4):333–358.

RABIN, M. O. et SCOTT, D. (1959). Finite automata and their decision problems. *IBM journal of research and development*, 3(2):114–125.

ROCHE, E. et SCHABES, Y. (1997). *Finite-state language processing*. MIT press.

ROTH, R., RAMBOW, O., HABASH, N., DIAB, M. et RUDIN, C. (2008). Arabic morphological tagging, diacritization, and lemmatization using lexeme models and feature ranking. *In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies : Short Papers*, pages 117–120. Association for Computational Linguistics.

RUESSINK, H. (1989). *Two-level formalisms*. Katholieke Universiteit.

SCHERRER, Y. *et al.* (2009). Un système de traduction automatique paramétré par des atlas dialectologiques. *Actes de TALN*.

SHAALAN, K., BAKR, H. et ZIEDAN, I. (2007). Transferring egyptian colloquial dialect into modern standard arabic. *In International Conference on Recent Advances in Natural Language Processing (RANLP–2007), Borovets, Bulgaria*, pages 525–529.

SMRŽ, O. (2007a). Elixirfm : implementation of functional arabic morphology. *In Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages : Common Issues and Resources*, pages 1–8. Association for Computational Linguistics.

SMRŽ, O. (2007b). *Functional Arabic Morphology. Formal System and Implementation*. Thèse de doctorat, Ph. D. thesis, Charles University in Prague, Prague, Czech Republic.

SPROAT, R. (1995). Lextools : Tools for finite-state linguistic analysis. Rapport technique, Technical Report 11522-951108-10TM, Bell Laboratories.

STOLCKE, A. *et al.* (2002). Srilm-an extensible language modeling toolkit. *In INTERSPEECH*.

TÄCKSTRÖM, O., DAS, D., PETROV, S., MCDONALD, R. et NIVRE, J. (2013). Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 1:1–12.

VERGEZ-COURET, M. (2013). Tagging occitan using french and castillan tree tagger. *In Proceedings of 6th Language & Technology Conference*.

VERGYRI, D. et KIRCHHOFF, K. (2004). Automatic diacritization of arabic for acoustic modeling in speech recognition. *In Proceedings of the workshop on computational approaches to Arabic script-based languages*, pages 66–73. Association for Computational Linguistics.

YAROWSKY, D., NGAI, G. et WICENTOWSKI, R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. *In Proceedings of the first international conference on Human language technology research*, pages 1–8. Association for Computational Linguistics.

ZITOUNI, I., SORENSEN, J. S. et SARIKAYA, R. (2006). Maximum entropy based restoration of arabic diacritics. *In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 577–584. Association for Computational Linguistics.

ZRIBI, I., KHEMAKHEM, M. E. et BELGUITH, L. H. (2013). Morphological analysis of tunisian dialect. *In Proceeding of International Joint Conference on Natural Language Processing (IJCNLP 2013), Nagoya, Japan*.

ANNEXES

A. Règles morphologiques du tunisien

Dans cette annexe, nous décrivons en détail toutes les ressources que nous avons créées dans MAGEAD pour le traitement du TUN.

Grammaire hors-contexte

```
[ENTREE] → [WORD] [RACINE]
[RACINE] → [RAD1] [RAD2] [RAD3]
[MOT] → [CONJ]([VRB]|[NOM])
[NOM] → [PREP]([INDEF_NOM]|[DEF_NOM])
[INDEF_NOM] → [NOM_STEM] [INDEF_CAS]
[DEF_NOM] → [DET] [NOM_STEM] ([DEF_CAS]|[NSUFF_NOPOSS])
[DEF_NOM] → [NOM_STEM] ([DEF_CAS]|[NSUFF_POSS])[POSS]
[DEF_NOM] → [DEF_NOM_STEM] [DEF_CAS]
[VRB] → ([PV_VRB]|[IV_VRB]|[CV_VRB])[OBJ][POST_VRB]
[POST_VRB] → ([POST:NEG]|[POST:nil])
[PV_VRB] → [PV_PRT] [PV_VRB_STEM] [SUBJSUF_PV]
[PV_PRT] → ([PRT:EMPHATIC]|[PRT:NEG]|[PRT:nil])
[IV_VRB] → [IV_PRT][IV_VRB_CONJUG]
[IV_PRT] → ([PRT:NEG]|[PRT:nil])
[IV_VRB_CONJUG] → [SUBJPRE_IV:1S][IV_VRB_STEM][SUBJSUF_IV:1S]
[IV_VRB_CONJUG] → [SUBJPRE_IV:1P][IV_VRB_STEM][SUBJSUF_IV:1P]
[IV_VRB_CONJUG] → [SUBJPRE_IV:2MS][IV_VRB_STEM][SUBJSUF_IV:2MS]
[IV_VRB_CONJUG] → [SUBJPRE_IV:2FS][IV_VRB_STEM][SUBJSUF_IV:2FS]
[IV_VRB_CONJUG] → [SUBJPRE_IV:2FP][IV_VRB_STEM][SUBJSUF_IV:2FP]
[IV_VRB_CONJUG] → [SUBJPRE_IV:2MP][IV_VRB_STEM][SUBJSUF_IV:2MP]
[IV_VRB_CONJUG] → [SUBJPRE_IV:3MS][IV_VRB_STEM][SUBJSUF_IV:3MS]
[IV_VRB_CONJUG] → [SUBJPRE_IV:3FS][IV_VRB_STEM][SUBJSUF_IV:3FS]
[IV_VRB_CONJUG] → [SUBJPRE_IV:3FP][IV_VRB_STEM][SUBJSUF_IV:3FP]
[IV_VRB_CONJUG] → [SUBJPRE_IV:3MP][IV_VRB_STEM][SUBJSUF_IV:3MP]
[CV_VRB] → [CV_VRB_STEM] [SUBJSUF_CV]
```

Hiérarchie de classes morphologiques

```
MBC-word
[cnj:f] : [CONJ:f]
[cnj:wa] : [CONJ:wa]
[cnj:wi] : [CONJ:wi]
[cnj:0] : [CONJ:nil]
[prt:0] : [PART:nil]
```

```
MBC-verb
[prt:l] : [PART:RESULT]
[prt:l] : [PART:SUBJUNC]
[prt:l] : [PART:EMPHATIC]
[prt:s] : [PART:FUT]
[prt:neg] : [PART:NEG]
[asp:P][per:1][num:s] : [SUBJ_SUF_PV:1S]
[asp:P][per:1][num:p] : [SUBJ_SUF_PV:1P]
[asp:P][per:2][gen:m][num:s] : [SUBJ_SUF_PV:2MS]
[asp:P][per:2][gen:f][num:s] : [SUBJ_SUF_PV:2FS]
[asp:P][per:2][num:d] : [SUBJ_SUF_PV:2D]
[asp:P][per:2][gen:m][num:p] : [SUBJ_SUF_PV:2MP]
[asp:P][per:2][gen:f][num:p] : [SUBJ_SUF_PV:2FP]
[asp:P][per:3][gen:m][num:s] : [SUBJ_SUF_PV:3MS]
[asp:P][per:3][gen:f][num:s] : [SUBJ_SUF_PV:3FS]
[asp:P][per:3][gen:m][num:d] : [SUBJ_SUF_PV:3MD]
[asp:P][per:3][gen:f][num:d] : [SUBJ_SUF_PV:3FD]
[asp:P][per:3][gen:m][num:p] : [SUBJ_SUF_PV:3MP]
[asp:P][per:3][gen:f][num:p] : [SUBJ_SUF_PV:3FP]
[asp:I][per:1][num:s] : [SUBJ_PRE_IV:1S]
[asp:I][per:1][num:p] : [SUBJ_PRE_IV:1P]
[asp:I][per:2][gen:m][num:s] : [SUBJ_PRE_IV:2MS]
[asp:I][per:2][gen:f][num:s] : [SUBJ_PRE_IV:2FS]
[asp:I][per:2][num:d] : [SUBJ_PRE_IV:2D]
[asp:I][per:2][gen:m][num:p] : [SUBJ_PRE_IV:2MP]
[asp:I][per:2][gen:f][num:p] : [SUBJ_PRE_IV:2FP]
[asp:I][per:3][gen:m][num:s] : [SUBJ_PRE_IV:3MS]
[asp:I][per:3][gen:f][num:s] : [SUBJ_PRE_IV:3FS]
[asp:I][per:3][gen:m][num:d] : [SUBJ_PRE_IV:3MD]
[asp:I][per:3][gen:f][num:d] : [SUBJ_PRE_IV:3FD]
[asp:I][per:3][gen:m][num:p] : [SUBJ_PRE_IV:3MP]
[asp:I][per:3][gen:f][num:p] : [SUBJ_PRE_IV:3FP]
[asp:I][per:1][num:s] : [SUBJ_SUF_IV:1S]
[asp:I][per:1][num:p] : [SUBJ_SUF_IV:1P]
[asp:I][per:2][gen:m][num:s] : [SUBJ_SUF_IV:2MS]
[asp:I][per:2][gen:f][num:s] : [SUBJ_SUF_IV:2FS]
[asp:I][per:2][num:d] : [SUBJ_SUF_IV:2D_Ind]
[asp:I][per:2][gen:m][num:p] : [SUBJ_SUF_IV:2MP]
[asp:I][per:2][gen:f][num:p] : [SUBJ_SUF_IV:2FP]
[asp:I][per:3][gen:m][num:s] : [SUBJ_SUF_IV:3MS]
[asp:I][per:3][gen:f][num:s] : [SUBJ_SUF_IV:3FS]
[asp:I][per:3][gen:m][num:d] : [SUBJ_SUF_IV:3MD]
```

```
[asp:I][per:3][gen:f][num:d] : [SUBJ_SUF_IV:3FD]
[asp:I][per:3][gen:m][num:p] : [SUBJ_SUF_IV:3MP]
[asp:I][per:3][gen:f][num:p] : [SUBJ_SUF_IV:3FP]
[asp:C][gen:m][num:s] : [SUBJ_SUF_CV:MS]
[asp:C][gen:f][num:s] : [SUBJ_SUF_CV:FS]
[asp:C][gen:m][num:p] : [SUBJ_SUF_CV:MP]
[asp:C][gen:f][num:p] : [SUBJ_SUF_CV:FP]

MBC-verb-Intr
[pro:0] : [OBJ:nil]

MBC-verb-Tr
[pro:1S] : [OBJ:1S]
[pro:1P] : [OBJ:1P]
[pro:2MS] : [OBJ:2MS]
[pro:2FS] : [OBJ:2FS]
[pro:2D] : [OBJ:2D]
[pro:2FP] : [OBJ:2FP]
[pro:2MP] : [OBJ:2MP]
[pro:3MS] : [OBJ:3MS]
[pro:3FS] : [OBJ:3FS]
[pro:3D] : [OBJ:3D]
[pro:3MP] : [OBJ:3MP]
[pro:3FP] : [OBJ:3FP]
[pro:0] : [OBJ:nil]

MBC-verb-I
[asp:P][pos:V]:[PAT_PV:I]
[asp:I][pos:V]:[PAT_IV:I]
[asp:C][pos:V]:[PAT_CV:I]
[asp:P][pos:V][vox:pas]:[VOC_PV:I-pas]
[asp:I][pos:V][vox:pas]:[VOC_IV:I-pas]

MBC-verb-I-aa
[asp:P][pos:V][vox:act] : [VOC_PV:I-aa-act]
[asp:I][pos:V][vox:act] : [VOC_IV:I-aa-act]
[asp:C][pos:V][vox:act] : [VOC_CV:I-aa-act]

MBC-verb-I-au
[asp:P][pos:V][vox:act] : [VOC_PV:I-au-act]
[asp:I][pos:V][vox:act] : [VOC_IV:I-au-act]
[asp:C][pos:V][vox:act] : [VOC_CV:I-au-act]

MBC-verb-I-ai
[asp:P][pos:V][vox:act] : [VOC_PV:I-ai-act]
```

```
[asp:I][pos:V][vox:act] : [VOC_IV:I-ai-act]
[asp:C][pos:V][vox:act] : [VOC_CV:I-ai-act]

MBC-verb-I-uu
[asp:P][pos:V][vox:act] : [VOC_PV:I-uu-act]
[asp:I][pos:V][vox:act] : [VOC_IV:I-uu-act]
[asp:C][pos:V][vox:act] : [VOC_CV:I-uu-act]

MBC-verb-I-ia
[asp:P][pos:V][vox:act] : [VOC_PV:I-ia-act]
[asp:I][pos:V][vox:act] : [VOC_IV:I-ia-act]
[asp:C][pos:V][vox:act] : [VOC_CV:I-ia-act]

MBC-verb-I-ii
[asp:P][pos:V][vox:act] : [VOC_PV:I-ii-act]
[asp:I][pos:V][vox:act] : [VOC_IV:I-ii-act]
[asp:C][pos:V][vox:act] : [VOC_CV:I-ii-act]

MBC-verb-II
[asp:P][pos:V] : [PAT_PV:II]
[asp:P][pos:V][vox:act] : [VOC_PV:II-act]
[asp:P][pos:V][vox:pas] : [VOC_PV:II-pas]
[asp:I][pos:V] : [PAT_IV:II]
[asp:I][pos:V][vox:act] : [VOC_IV:II-act]
[asp:I][pos:V][vox:pas] : [VOC_IV:II-pas]
[asp:C][pos:V] : [PAT_CV:II]
[asp:C][pos:V][vox:act] : [VOC_CV:II-act]

MBC-verb-III
[asp:P][pos:V] : [PAT_PV:III]
[asp:P][pos:V][vox:act] : [VOC_PV:III-act]
[asp:P][pos:V][vox:pas] : [VOC_PV:III-pas]
[asp:I][pos:V] : [PAT_IV:III]
[asp:I][pos:V][vox:act] : [VOC_IV:III-act]
[asp:I][pos:V][vox:pas] : [VOC_IV:III-pas]
[asp:C][pos:V] : [PAT_CV:III]
[asp:C][pos:V][vox:act] : [VOC_CV:III-act]

MBC-verb-IV
[asp:P][pos:V] : [PAT_PV:IV]
[asp:P][pos:V][vox:act] : [VOC_PV:IV-act]
[asp:P][pos:V][vox:pas] : [VOC_PV:IV-pas]
[asp:I][pos:V] : [PAT_IV:IV]
[asp:I][pos:V][vox:act] : [VOC_IV:IV-act]
[asp:I][pos:V][vox:pas] : [VOC_IV:IV-pas]
[asp:C][pos:V] : [PAT_CV:IV]
```

[asp:C][pos:V][vox:act] : [VOC_CV:IV-act]

MBC-verb-V
[asp:P][pos:V] : [PAT_PV:V]
[asp:P][pos:V][vox:act] : [VOC_PV:V-act]
[asp:P][pos:V][vox:pas] : [VOC_PV:V-pas]
[asp:I][pos:V] : [PAT_IV:V]
[asp:I][pos:V][vox:act] : [VOC_IV:V-act]
[asp:I][pos:V][vox:pas] : [VOC_IV:V-pas]
[asp:C][pos:V] : [PAT_CV:V]
[asp:C][pos:V][vox:act] : [VOC_CV:V-act]

MBC-verb-VI
[asp:P][pos:V] : [PAT_PV:VI]
[asp:P][pos:V][vox:act] : [VOC_PV:VI-act]
[asp:P][pos:V][vox:pas] : [VOC_PV:VI-pas]
[asp:I][pos:V] : [PAT_IV:VI]
[asp:I][pos:V][vox:act] : [VOC_IV:VI-act]
[asp:I][pos:V][vox:pas] : [VOC_IV:VI-pas]
[asp:C][pos:V] : [PAT_CV:VI]
[asp:C][pos:V][vox:act] : [VOC_CV:VI-act]

MBC-verb-VII
[asp:P][pos:V]:[PAT_PV:VII]
[asp:P][pos:V][vox:act]:[VOC_PV:VII-act]
[asp:P][pos:V][vox:pas]:[VOC_PV:VII-pas]
[asp:I][pos:V]:[PAT_IV:VII]
[asp:I][pos:V][vox:act]:[VOC_IV:VII-act]
[asp:I][pos:V][vox:pas]:[VOC_IV:VII-pas]
[asp:C][pos:V]:[PAT_CV:VII]
[asp:C][pos:V][vox:act]:[VOC_CV:VII-act]

MBC-verb-VIII
[asp:P][pos:V] : [PAT_PV:VIII]
[asp:P][pos:V][vox:act] : [VOC_PV:VIII-act]
[asp:P][pos:V][vox:pas] : [VOC_PV:VIII-pas]
[asp:I][pos:V] : [PAT_IV:VIII]
[asp:I][pos:V][vox:act] : [VOC_IV:VIII-act]
[asp:I][pos:V][vox:pas] : [VOC_IV:VIII-pas]
[asp:C][pos:V] : [PAT_CV:VIII]
[asp:C][pos:V][vox:act] : [VOC_CV:VIII-act]

MBC-verb-IX
[asp:P][pos:V] : [PAT_PV:IX]
[asp:P][pos:V][vox:act] : [VOC_PV:IX-act]

[asp:P][pos:V][vox:pas] : [VOC_PV:IX-pas]
[asp:I][pos:V] : [PAT_IV:IX]
[asp:I][pos:V][vox:act] : [VOC_IV:IX-act]
[asp:I][pos:V][vox:pas] : [VOC_IV:IX-pas]
[asp:C][pos:V]:[PAT_CV:IX]
[asp:C][pos:V][vox:act] : [VOC_CV:IX-act]

MBC-verb-X
[asp:P][pos:V] : [PAT_PV:X]
[asp:P][pos:V][vox:act] : [VOC_PV:X-act]
[asp:P][pos:V][vox:pas] : [VOC_PV:X-pas]
[asp:I][pos:V] : [PAT_IV:X]
[asp:I][pos:V][vox:act] : [VOC_IV:X-act]
[asp:I][pos:V][vox:pas] : [VOC_IV:X-pas]
[asp:C][pos:V] : [PAT_CV:X]
[asp:C][pos:V][vox:act] : [VOC_CV:X-act]

MBC-verb-XI
[asp:P][pos:V] : [PAT_PV:XI]
[asp:P][pos:V][vox:act] : [VOC_PV:XI-act]
[asp:P][pos:V][vox:pas] : [VOC_PV:XI-pas]
[asp:I][pos:V] : [PAT_IV:XI]
[asp:I][pos:V][vox:act] : [VOC_IV:XI-act]
[asp:I][pos:V][vox:pas] : [VOC_IV:XI-pas]
[asp:C][pos:V] : [PAT_CV:XI]
[asp:C][pos:V][vox:act] : [VOC_CV:XI-act]

MBC-verb-QI
[asp:P][pos:V] : [PAT_PV:QI]
[asp:P][pos:V][vox:act] : [VOC_PV:QI-act]
[asp:P][pos:V][vox:pas] : [VOC_PV:QI-pas]
[asp:I][pos:V] : [PAT_IV:QI]
[asp:I][pos:V][vox:act] : [VOC_IV:QI-act]
[asp:I][pos:V][vox:pas] : [VOC_IV:QI-pas]
[asp:C][pos:V] : [PAT_CV:QI]
[asp:C][pos:V][vox:act] : [VOC_CV:QI-act]

MBC-verb-QII
[asp:P][pos:V] : [PAT_PV:QII]
[asp:P][pos:V][vox:act] : [VOC_PV:QII-act]
[asp:P][pos:V][vox:pas] : [VOC_PV:QII-pas]
[asp:I][pos:V] : [PAT_IV:QII]
[asp:I][pos:V][vox:act] : [VOC_IV:QII-act]
[asp:I][pos:V][vox:pas] : [VOC_IV:QII-pas]
[asp:C][pos:V] : [PAT_CV:QII]

```
[asp:C][pos:V][vox:act] : [VOC_CV:QII-act]

MBC-verb-QIII
[asp:P][pos:V] : [PAT_PV:QIII]
[asp:P][pos:V][vox:act] : [VOC_PV:QIII-act]
[asp:P][pos:V][vox:pas] : [VOC_PV:QIII-pas]
[asp:I][pos:V] : [PAT_IV:QIII]
[asp:I][pos:V][vox:act] : [VOC_IV:QIII-act]
[asp:I][pos:V][vox:pas] : [VOC_IV:QIII-pas]
[asp:C][pos:V] : [PAT_CV:QIII]
[asp:C][pos:V][vox:act] : [VOC_CV:QIII-act]

MBC-NOM
[prt:l] : [PREP:l]
[prt:b] : [PREP:b]
[prt:k] : [PREP:k]
[prt:0] : [PREP:nil]
[det:Al] : [DET:Al]
[det:0] : [DET:nil]
[pro:1S] : [POSS:1S]
[pro:1P] : [POSS:1P]
[pro:2MS] : [POSS:2MS]
[pro:2FS] : [POSS:2FS]
[pro:2D] : [POSS:2D]
[pro:2FP] : [POSS:2FP]
[pro:2MP] : [POSS:2MP]
[pro:3MS] : [POSS:3MS]
[pro:3FS] : [POSS:3FS]
[pro:3D] : [POSS:3D]
[pro:3MP] : [POSS:3MP]
[pro:3FP] : [POSS:3FP]
[pro:0] : [POSS:nil]
```

Table de correspondance de morphèmes abstraits et mophèmes concrets

```
[CONJ:wa] : wa+
[CONJ:wi] : wi+
[CONJ:f] : fa+
[CONJ:nil] : ϵ
[PREP:b] : bi+
[PREP:k] : ki+
[PREP:l] : li+
```

[PREP:nil] : ϵ

[PART:RESULT] : la+
[PART:SUBJUNC] : bAc+
[PART:EMPHATIC] : la+
[PART:NEG] : lA+
[PART:NEG] : mA+
[PART:FUT] : bAc+
[PART:FUT] : mAc+
[PART:nil] : ϵ

[POST:NEG] : +c
[POST:nil] : ϵ

[DET:Al] : Al+
[DET:nil] : ϵ

[POSS:1S] : +I
[POSS:1P] : +nA
[POSS:2MS] : +ik
[POSS:2FS] : +ik
[POSS:2MP] : +kum
[POSS:2FP] : +kum
[POSS:3MS] : +h
[POSS:3FS] : +hA
[POSS:3MP] : +hum
[POSS:3FP] : +hum
[POSS:nil] : ϵ

[OBJ:1S] : +nI
[OBJ:1P] : +nA
[OBJ:2MS] : +ik
[OBJ:2FS] : +ik
[OBJ:2FP] : +kum
[OBJ:2MP] : +kum
[OBJ:3MS] : +h
[OBJ:3MS] : +U
[OBJ:3FS] : +hA
[OBJ:3MP] : +hum
[OBJ:3FP] : +hum
[OBJ:nil] : ϵ

[SUBJ_SUF_PV:1S] : +t
[SUBJ_SUF_PV:1P] : +nA
[SUBJ_SUF_PV:2MS] : +t
[SUBJ_SUF_PV:2FS] : +t
[SUBJ_SUF_PV:2MP] : +tuwA
[SUBJ_SUF_PV:2FP] : +tuwA

```
[SUBJ_SUF_PV:3MS] : +0
[SUBJ_SUF_PV:3FS] : +it
[SUBJ_SUF_PV:3FP] : +uwA
[SUBJ_SUF_PV:3MP] : +uwA

[SUBJ_PRE_IV:1S] : n+
[SUBJ_PRE_IV:1P] : n+
[SUBJ_PRE_IV:2MS] : t+
[SUBJ_PRE_IV:2FS] : t+
[SUBJ_PRE_IV:2MP] : t+
[SUBJ_PRE_IV:2FP] : t+
[SUBJ_PRE_IV:3MS] : y+
[SUBJ_PRE_IV:3FS] : t+
[SUBJ_PRE_IV:3MP] : y+
[SUBJ_PRE_IV:3FP] : y+

[SUBJ_SUF_IV:1S] : +0
[SUBJ_SUF_IV:1P] : +uwA
[SUBJ_SUF_IV:2MS] : +0
[SUBJ_SUF_IV:2FS] : +0
[SUBJ_SUF_IV:2MP] : +uwA
[SUBJ_SUF_IV:2FP] : +uwA
[SUBJ_SUF_IV:3MS] : +0
[SUBJ_SUF_IV:3FS] : +0
[SUBJ_SUF_IV:3MP] : +uwA
[SUBJ_SUF_IV:3FP] : +uwA

[SUBJ_SUF_CV:MS] : +0
[SUBJ_SUF_CV:FS] : +0
[SUBJ_SUF_CV:MP] : +uwA
[SUBJ_SUF_CV:FP] : +uwA

[PAT_IV:I][VOC_IV:I-aa-a-act] : [V12V3,XXX,aa]
[PAT_PV:I][VOC_PV:I-aa-a-act] : [1V2V3,XXX,aa]
[PAT_CV:I][VOC_CV:I-aa-a-act] : [V12V3,XXX,aa]
[PAT_IV:I][VOC_IV:I-aa-a-pas] : [VtV12V3,XXX,iia]
[PAT_PV:I][VOC_PV:I-aa-a-pas] : [tV1V2V3,XXX,iaa]

[PAT_IV:I][VOC_IV:I-aa-i-act] : [V12V3,XXX,ia]
[PAT_PV:I][VOC_PV:I-aa-i-act] : [1V2V3,XXX,ia]
[PAT_CV:I][VOC_CV:I-aa-i-act] : [V12V3,XXX,ia]
[PAT_IV:I][VOC_IV:I-aa-i-pas] : [VtV12V3,XXX,iia]
[PAT_PV:I][VOC_PV:I-aa-i-pas] : [tV1V2V3,XXX,iia]

[PAT_IV:I][VOC_IV:I-au-act] : [V12V3,XXX,uu]
[PAT_PV:I][VOC_PV:I-au-act] : [1V2V3,XXX,aa]
[PAT_CV:I][VOC_CV:I-au-act] : [V12V3,XXX,uu]
[PAT_IV:I][VOC_IV:I-au-pas] : [VtV12V3,XXX,iia]
```

[PAT_PV:I][VOC_PV:I-au-pas] : [tV1V2V3,XXX,iaa]

[PAT_IV:I][VOC_IV:I-ai-act] : [V12V3,XXX,ai]
[PAT_PV:I][VOC_PV:I-ai-act] : [1V2V3,XXX,aa]
[PAT_CV:I][VOC_CV:I-ai-act] : [V12V3,XXX,ai]
[PAT_IV:I][VOC_IV:I-ai-pas] : [VtV12V3,XXX,iaa]
[PAT_PV:I][VOC_PV:I-ai-pas] : [tV1V2V3,XXX,iaa]

[PAT_IV:I][VOC_IV:I-ii-act] : [V12V3,XXX,ii]
[PAT_PV:I][VOC_PV:I-ii-act] : [1V2V3,XXX,ii]
[PAT_CV:I][VOC_CV:I-ii-act] : [V12V3,XXX,ii]
[PAT_IV:I][VOC_IV:I-ii-pas] : [VtV12V3,XXX,iii]
[PAT_PV:I][VOC_PV:I-ii-pas] : [tV1V2V3,XXX,iii]

[PAT_IV:I][VOC_IV:I-uu-act] : [V12V3,XXX,uu]
[PAT_PV:I][VOC_PV:I-uu-act] : [1V2V3,XXX,uu]
[PAT_CV:I][VOC_CV:I-uu-act] : [V12V3,XXX,uu]
[PAT_IV:I][VOC_IV:I-uu-pas] : [VtV12V3,XXX,iia]
[PAT_PV:I][VOC_PV:I-uu-pas] : [tV1V2V3,XXX,iaa]

[PAT_IV:II][VOC_IV:II-aa-act] : [1V22V3,XXX,aa]
[PAT_PV:II][VOC_PV:II-aa-act] : [1V22V3,XXX,aa]
[PAT_CV:II][VOC_CV:II-aa-act] : [1V22V3,XXX,aa]
[PAT_IV:II][VOC_IV:II-aa-pas] : [Vt1V22V3,XXX,iaa]
[PAT_PV:II][VOC_PV:II-aa-pas] : [t1V22V3,XXX,aa]

[PAT_IV:II][VOC_IV:II-ii-act] : [1V22V3,XXX,ai]
[PAT_PV:II][VOC_PV:II-ii-act] : [1V22V3,XXX,ai]
[PAT_CV:II][VOC_CV:II-ii-act] : [1V22V3,XXX,ai]
[PAT_IV:II][VOC_IV:II-ii-pas] : [Vt1V22V3,XXX,iai]
[PAT_PV:II][VOC_PV:II-ii-pas] : [t1V22V3,XXX,ai]

[PAT_IV:III][VOC_IV:III-aa-act] : [1A2V3,XXX,a]
[PAT_PV:III][VOC_PV:III-aa-act] : [1A2V3,XXX,a]
[PAT_CV:III][VOC_CV:III-aa-act] : [1A2V3,XXX,a]
[PAT_IV:III][VOC_IV:III-aa-pas] : [Vt1A2V3,XXX,ia]
[PAT_PV:III][VOC_PV:III-aa-pas] : [t1A2V3,XXX,a]

[PAT_IV:III][VOC_IV:III-ii-act] : [1A2V3,XXX,i]
[PAT_PV:III][VOC_PV:III-ii-act] : [1A2V3,XXX,i]
[PAT_CV:III][VOC_CV:III-ii-act] : [1A2V3,XXX,i]
[PAT_IV:III][VOC_IV:III-ii-pas] : [Vt1A2V3,XXX,ii]
[PAT_PV:III][VOC_PV:III-ii-pas] : [t1A2V3,XXX,i]

[PAT_IV:V][VOC_IV:V-aa-act] : [Vt1V22V3,XXX,iaa]
[PAT_PV:V][VOC_PV:V-aa-act] : [t1V22V3,XXX,aa]
[PAT_CV:V][VOC_CV:V-aa-act] : [t1V22V3,XXX,aa]

[PAT_IV:V][VOC_IV:V-ii-act] : [Vt1V22V3,XXX,iai]
[PAT_PV:V][VOC_PV:V-ii-act] : [t1V22V3,XXX,ai]

```
[PAT_CV:V][VOC_CV:V-ii-act] : [t1V22V3,XXX,ai]
[PAT_IV:VI][VOC_IV:VI-act] : [Vt1A2V3,XXX,ii]
[PAT_PV:VI][VOC_PV:VI-act] : [t1A2V3,XXX,i]
[PAT_CV:VI][VOC_CV:VI-act] : [t1A2V3,XXX,i]
[PAT_IV:VIII][VOC_IV:VIII-aa-act] : [V1tV2V3,XXX,iaa]
[PAT_PV:VIII][VOC_PV:VIII-aa-act] : [AV1tV2V3,XXX,iaa]
[PAT_CV:VIII][VOC_CV:VIII-aa-act] : [V1tV2V3,XXX,iaa]
[PAT_IV:VIII][VOC_IV:VIII-ai-act] : [V1tV2V3,XXX,iai]
[PAT_PV:VIII][VOC_PV:VIII-ai-act] : [AV1tV2V3,XXX,iaa]
[PAT_CV:VIII][VOC_CV:VIII-ai-act] : [V1tV2V3,XXX,iai]
[PAT_IV:IX][VOC_IV:IX-act] : [V12A3,XXX,i]
[PAT_PV:IX][VOC_PV:IX-act] : [12AV3,XXX,u]
[PAT_CV:IX][VOC_CV:IX-act] : [V12A3,XXX,i]
[PAT_IV:X][VOC_IV:X-act] : [VstV12V3,XXX,iai]
[PAT_PV:X][VOC_PV:X-act] : [AVstV12V3,XXX,iai]
[PAT_CV:X][VOC_CV:X-act] : [VstV12V3,XXX,iai]
[NSUFF_POSS_PL_MASC] : [+iy]
[NSUFF_POSS_PL_FEM] : [+At]
[NSUFF_NOPOSS_PL_MASC] : [+iyn]
[NSUFF_NOPOSS_PL_FEM] : [+At]
```

Règles morpho-phonémiques et orthographiques

Dans cette section, nous donnons les règles nécessaires pour réaliser la flexion d'une classe de verbes TUN. Les règles sont suivi d'un exemple illustratif.

```
Règles morpho-phonémiques de base
[X„,0] → X, X=[PATTERNLETTER]
[C,X„0] → X
[V„X,0] → X
Règles orthographiques de base
[X„,X,0] → X, X=[PATTERNLETTER]
[C,X„X,0] → X
[V„X,X,0] → X
```

Verbes parfaitement sains (*ktib* "*écrire*")

```
Règles morpho-phonémiques des verbes sains
[V„X,X] → 0 / [2,%„%] _ [3+Y,%„%+Y], Y=[{Uui}], X=[{aui}]
[V„X,X] → 0 / _ [3+0+Y,%„%+0+Y], Y=[{Uui}], X=[{aui}]
[V„X,X] → 0 / [1,%„%] _ [2V3,%%,V,%V%], X=[{aui}], V=[{aui}]
[V„X,X] → 0 / [1,%„%] _ [2V3,%%,%,%Y0], X=[{aui}], Y=[{aui}]
```

```
[V„V,V] → 0 / _ [1V,%,Y,%Y], V=[{aui}], Y=[{Aaui}]
```

	accompli	inaccompli
1S	ktib**t**	**n**iktib
1P	ktib**nA**	**n**iktib**uwA**
2MS	ktib**t**	**t**iktib
2FS	ktib**tiy**	**t**iktib**iy**
2P	ktib**tuwA**	**t**iktib**uwA**
3MS	ktib	**y**iktib
3FS	kitb**it**	**t**iktib
3P	kitb**uwA**	**y**iktib**uwA**

Table .14.: Flexion des verbes parfaitement sains

Verbes défectifs (*mad*∼ "*étendre*")

```
Règles morpho-phonémiques des verbes défectifs
[V3,X,V,VX] → Y0 / [V2,%,%,%%] _ [+0„,+0], X=[{wy}], V=[{aui}]
[V3,X,V,VX] → A0 / [V2,%,%,%%] _ [+S„,+S], S=[{ui}], X=[{wy}],
V=[{aui}]
[V3,X,V,VX] → A0 / [tV12,%%,i,ti%%] _ [+S„,+S], S=[{ui}],
X=[{wy}], V=[{aui}]
[V3,X,V,VX] → Y0 / [tV12,%%,i,ti%%] _ [+0„,+0], X=[{wy}],
V=[{aui}]
[V„V,V] → 0 / _ [2V3,%X,%,%A0], X=[{wy}], V=[{aui}]
[V„V,V] → 0 / _ [2V3,%X,%,%Y0], X=[{wy}], V=[{aui}]
[2,w„w] → W / _ [V3,X,%,%y], X=[{wy}], V=[{aui}]
[V„V,V] → X / [S+1,%„S+%] _ S=[{nty}] , V=[{aui}]
[V3,X,V,VX] → Y0 / [V22,%,Z,Z%%] _ [+0„,+0], X=[{wy}], V=[{aui}],
Z=[{ai}]
[V3,X,V,VX] → A0 / [V22,%,Z,Z%%] _ [+S„,+S], S=[{ui}], X=[{wy}],
V=[{aui}], Z=[{ai}]
[V„V,X] → V
[V„V,V] → X / [S+1A,%„S+%A] _ , S=[{nty}] , V=[{aui}]
[V3,X,V,VX] → Y0 / [A2,%„A%] _ [+0„,+0], X=[{wy}], V=[{aui}]
[V3,X,V,VX] → A0 / [A2,%„A%] _ [+S„,+S], S=[{ui}], X=[{wy}] ,
V=[{aui}]
[S„,S] → 0 / [V3+,X,V,%%+] _ , S=[{ui}] , X=[{wy}], V=[{aui}]
[V„V,X] → V

Règles orthographiques des verbes défectifs
[V1,Y,X,ZY,0] → ZY, X=[{aui}], Z=[{ui}], Y=[{wy}]
[2,X„Y,0] → Y, X=[{wy}], Y=[{wy0}]
```

```
[3,X„Y,0] → Y, X=[{wy}], Y=[{wy0}]
[V„X,Y,0] → Y, X=[VOWEL], Y=[LONGVOWEL]
[V„Y,X,0] → X, X=[VOWEL], Y=[LONGVOWEL]
[V„X,Y,0] → Y, X=[VOWEL], Y=[VOWEL]
[X„,Y,0] → Y, X=[VOWEL], Y=[LONGVOWEL]
[X„,Y,0] → Y, X=[VOWEL], Y=[VOWEL]
```

	accompli	inaccompli
1S	rmiy**t**	**n**armiy
1P	rmiy**nA**	**n**armiy**wA**
2MS	rmiy**t**	**t**armiy
2FS	rmiy**tiy**	**t**armiy
2P	rmiy**tuwA**	**t**armiy**wA**
3MS	rmaý	**y**armiy
3FS	rmA**t**	**t**armiy
3P	rmA**wA**	**y**armiy**wA**

Table .15.: Flexion des verbes défectifs

Verbes creux (*bAς* "*vendre*")

```
Règles morpho-phonémiques des verbes creux
[V„V,V] → 0 / _ [1V2V,%X,%%,%%X%], V=[{aui}], X=[{wy}]
[V„V,V] → 0 / _ [12V,%X,%,%X%], V=[{aui}], X=[{wy}]
[V2V,X,VZ,VXZ] → A00 / [1,%„%] _ [3+S,%„%+S], S=[{0iuU}], X=[{wy}]
, V=[{aui}] , Z=[{aui}]
[2V,X,Z,XZ] → 0A / [tV1,%,%,t0%] _ [3+S,%„%+S], S=[{0iuU}],
X=[{wy}] , Z=[{aui}]
[2V,w,u,wu] → 0U / [+V1,%,%,+%%] _ [3,%„%]
[2V,y,i,yi] → 0I / [+V1,%,%,+%%] _ [3,%„%]
[V2V,w,VX,VwX] → 00u / [1,%„%] _ [3+S,%„%+S], S=[{tn}], V=[{aui}],
X=[{aui}]
[V2V,y,VX,VyX] → 00i / [1,%„%] _ [3+S,%„%+S], S=[{tn}], V=[{aui}],
X=[{aui}]
[V„V,V] → 0 / _ [2V3,%X,%,%A0], X=[{wy}], V=[{aui}]
[V„V,V] → 0 / _ [2V3,%X,%,%Y0], X=[{wy}], V=[{aui}]
[2,w„W] → w
```

	accompli	inaccompli
1S	qul**t**	**n**quwl
1P	qul**nA**	**n**quwl**uwA**
2MS	qul**t**	**t**quwl
2FS	qul**tiy**	**t**quwl**iy**
2P	qul**tuwA**	**t**quwl**uwA**
3MS	qAl	**y**quwl
3FS	qAl**it**	**t**quwl
3P	qAl**uwA**	**y**quwl**uwA**

Table .16.: Flexion des verbes creux

Verbes assimilés (*wSul* "*arriver*")

```
Règles morpho-phonémiques de verbes assimilés
[V1,w,X,Xw] → uw / [+„,+] _ , X=[VOWEL]
[V1,y,X,Xy] → iy / [+„,+] _ , X=[VOWEL]
```

	accompli	inaccompli
1S	wSil**t**	**n**uwSil
1P	wSil**nA**	**n**uwSl**uwA**
2MS	wSil**t**	**t**uwSil
2FS	wSil**tiy**	**t**uwSl**iy**
2P	mad∼iy**tuwA**	**t**mid∼**uwA**
3MS	wSil	**y**uwSil
3FS	wiSl**it**	**t**uwSil
3P	wiSl**uwA**	**y**uwSl**uwA**

Table .17.: Flexion des verbes assimilés

Verbes hamzés

```
Règles morpho-phonémiques de verbes hamzés
[V1,',X,X'] → A0 / [+„,+] _, X=[aui]
[V1,',X,X'] → A0 / [+tV„%,+t%] _, X=[{aui}]
[V3,',X,0'] → A0 / _ [+S„,+S], X=[{aui}], S=[{0iuU}]
[V3,',X,X'] → A0 / _ [+S„,+S], X=[{aui}], S=[{0iuU}]
[V„V,V] → 0 / _ [2V3,%X,%,%A0], X=[{'}], V=[{aui}]
[V3,',X,X'] → I0 / _ [+S„,+S], X=[{aui}], S=[{tn}]
[V„V,V] → 0 / _ [2V3,%X,%,%I0], X=[{'}], V=[aui]
[3,'„'] → 0 / [V„%,I] _
[S„,S] → 0 / [V3+,X,V,%%+] _, S=[{ui}], X=[{'}], V=[{aui}]
```

– hamza dans la première lettre de la racine (*klaý* "*manger*")

	accompli	inaccompli
1S	kliy**t**	**n**Akil
1P	kliy**nA**	**n**Akl**uwA**
2MS	kliy**t**	**t**Akil
2FS	kliy**tiy**	**t**Akl**iy**
2P	kliy**tuwA**	**t**Akl**uwA**
3MS	klý	**y**Akil
3FS	klA**t**	**t**Akil
3P	klA**wA**	**y**Akl**uwA**

Table .18.: Flexion des verbes contenant une hamza dans la première radicale

– hamza dans la troisième radicale (*bdA* "*commencer*")

	accompli	inaccompli
1S	bdiy**t**	**n**abdA
1P	bdiy**nA**	**n**abdA**wA**
2MS	bdiy**t**	**t**abdA
2FS	bdiy**tiy**	**t**abd**iy**
2P	bdiy**tuwA**	**t**abdA**wA**
3MS	bdA	**y**abdA
3FS	bdA**t**	**t**abdA
3P	bdA**wA**	**y**abdA**wA**

Table .19.: Flexion des verbes contenant une hamza dans la troisième radicale

Verbes redoublés (*mad*∼ "*étendre*")

```
Règles morpho-phonémiques des verbes redoublés
[3,X„X] → X / [V2V,X,%%,%%%] _, X=[CONSONANT]
[3,X„X] → X / [12V,%X,%,%%%] _, X=[CONSONANT]
[V„V,V] → 0 / _ [12V3,%%%,%,%%%X], V=[{aui}]
[V„V,V] → 0 / _ [3,%„X] , V=[SHORTVOWEL]
[2V,X,V,X0] → VX / [1,%„%] _ [3,X„X], V=[SHORTVOWEL],
X=[CONSONANT]
[+„,+] → I / [3,X„X] _ [S„,S], S=[{tn}], X=[CONSONANT]
[3,X„X] → X

Règles orthographiques des verbes redoublés
[2V,X,V,VX,00] → VX, V=[VOWEL], X=[CONSONANT]
[+„,I,0] → I
```

	accompli	inaccompli
1S	mad∼iy**t**	**n**mid∼
1P	mad∼iy**nA**	**n**mid∼**uwA**
2MS	mad∼iy**t**	**t**mid∼
2FS	mad∼iy**tiy**	**t**mid∼**iy**
2P	mad∼iy**tuwA**	**t**mid∼**uwA**
3MS	mad∼	**y**mid∼
3FS	mad∼**it**	**t**mid∼
3P	mad∼**uwA**	**y**mid∼**uwA**

Table .20.: Flexion des verbes redoublés

Verbes de la forme IX (*HmAr* "*rougir*")

```
Règles morpho-phonémiques de verbes de la forme IX
[AV„u,Au] → 0u / _ [3+S,%„%+S], S=[{tn}]
[AV„u,Au] → A0 / _ [3+Z,%„%+Z], Z=[{0iuU}]
```

	accompli	inaccompli
1S	Hmur**t**	**n**iHmAr
1P	Hmur**nA**	**n**iHmAr**uwA**
2MS	Hmur**t**	**t**iHmAr
2FS	Hmur**tiy**	**t**iHmAr**iy**
2P	Hmur**tuwA**	**t**iHmAr**uwA**
3MS	HmAr	**y**iHmAr
3FS	HmAr**it**	**t**iHmAr
3P	HmAr**uwA**	**y**iHmAr**uwA**

Table .21.: Flexion des verbes de la forme IX

B. Liste des verbes issus de racines TUN

1. gzr × 12a3 = gzar "*regarder*"
2. nqb × 12a3 = nqab "*percer*"
3. rkH × 12a3 = rkaH "*se calmer*"
4. rAD × 12a3 = rAD "*se calmer*"
5. Sdm × 12u3 = Sdum "*attaquer*"
6. lqf × 12i3 = lqif "*attraper*"
7. cwf × 12u3 = cAf "*voir*"
8. jbd × 12i3 = jbid "*tirer*"
9. qdm × 12i3 = qdim "*mordre*"
10. nqz × 1a22i3 = naqqiz "*sauter*"
11. gTs × 12u3 = gTus "*plonger*"
12. Srf × 12a3 = Sruf "*dépenser*"
13. Hšm × 12i3 = Hšim "*intimider*"
14. rʕš × 12u3 = rʕuš "*trembler*"
15. sxf × 12i3 = sxif "*avoir pitié*"
16. skr × 1a22i3 = sakkir "*fermer*"
17. nzl × 1a2i3 = nzil "*appuyer*"
18. ’nn × Aista1a2a3 = Aistannaý "*attendre*"
19. bkš × 12i3 = bkiš "*devenir muet*"
20. Hbs × 12a3 = Hbas "*isoler*"
21. rqd × 12a3 = rqad "*dormir*"
22. Sbb × 12a3 = Sabb "*verser*"
23. lHlH × 1a23a4 = laHlaH "*insister*"
24. blbz × 1a23i4 = balbiz "*défaire*"
25. lnsy × 1a23a4 = lansaý "*lancer*"
26. frhd × 1a23i4 = farhid "*amuser*"
27. fdd × 12a3 = fadd "*s’ennuyer*"
28. ʕss × 12a3 = ʕass "*surveiller*"
29. lzz × 12a3 = lazz "*obliger*"
30. mss × 12a3 = mass "*toucher*"
31. sdd × 12a3 = sadd "*bloquer*"
32. lmd × 1a22i3 = lammid "*rassembler*"

33. bTl × 1a22i3 = baTTil "*suspendre*"
34. lwj × 1a22i3 = lawwij "*chercher*"
35. frks × 1a23i4 = farkis "*chercher*"
36. njm × 1a22i3 = najjim "*pouvoir*"
37. hzz × 12a3 = hazz "*emporter*"
38. dzz × 12a3 = dazz "*pousser*"
39. jyb × 12a3 = jAb "*ramener*"
40. TyH × 12a3 = TAH "*tomber*"
41. šyx × 12a3 = šAx "*jouir*"
42. wly × 1a22a3 = wallaý "*devenir*"
43. xly × 1a22a3 = xallaý "*laisser*"
44. fyD × 12a3 = fAD "*déborder*"
45. msmr × t1a23i4 = tmasmir "*se fixer*"
46. xlS × 1a22i3 = xalliS "*rembourser*"
47. ςwm × 12a3 = ςAm "*se baigner*"
48. fwH × 1a22a3 = fawwaH "*épicer*"
49. kHH × 12a3 = kaHH "*tousser*"
50. HSl × 1a22i3 = HaSSil "*tromper*"
51. ςfs × 12a3 = ςfas "*fouler*"
52. ςTš × 12u3 = ςTuš "*avoir soif*"
53. Hqr × 12a3 = Hqar "*mépriser*"
54. wlm × 1A2i3 = wAlim "*adapter*"
55. bzq × 12a3 = bzaq "*cracher*"
56. srH × 12a3 = sraH "*rêver*"
57. ðbH × 12a3 = ðbaH "*égorger*"
58. fDH × 12a3 = fDaH "*diffuser*"
59. ssy × 1A2a3 = sAsaý "*mendier*"
60. br’ × 12a3 = brA "*guérir*"
61. Tyb × 1a22i3 = Tayyib "*préparer*"
62. Hws × 1a22i3 = Hawwis "*se promener*"
63. γšš × 1a22i3 = γaššiš "*énerver*"
64. tlf × 1a22i3 = tallif "*négliger*"
65. syb × 1a22i3 = sayyib "*laisser*"

66. sys × 1A2i3 = sAyis "*aider*"
67. Hmš × 1a22i3 = Hammiš "*provoquer*"
68. šyς × 1a22a3 = šayyaς "*emmener*"
69. Tyš × 1a22i3 = Tayyiš "*jeter*"
70. bnj × 1a22i3 = bannij "*anesthésier*"
71. xbš × 1a22i3 = xabbiš "*griffer*"
72. nHy × 1a22a3 = naHHaý "*enlever*"
73. kbš × 1a22i3 = kabbiš "*s'attacher*"
74. ςyT × 1a22i3 = ςayyiT "*crier*"
75. qrr × 1a22i3 = qarrir "*insister*"
76. wxr × 1a22i3 = waxxir "*reculer*"
77. xlT × 12a3 = xlaT "*rattraper*"
78. γlT × 12a3 = gluT "*se tromper*"
79. SHH × 1a22a3 = SaHHaH "*signer*"
80. zrq × 1a22a3 = zarraq "*vacciner*"
81. slf × 1a22i3 = sallif "*prêter*"
82. bws × 12a3 = bAs "*embrasser*"
83. qbH × t1a22a3 = tqabbaH "*affecter*"
84. msx × t1a22a3 = tmassax "*se salir*"
85. ςrD × 12u3 = ςruD "*rencontrer*"
86. rtH × 1a22a3 = rattaH "*reposer*"
87. xmj × 12i3 = xmij "*périmer*"
88. štH × 12a3 = štaH "*danser*"
89. dls × 1a22i3 = dallis "*falsifier*"
90. Dbς × 1a22a3 = Dabbaς "*devenir fou*"
91. rwm × 1a22i3 = rawwim "*dompter*"
92. zrb × 12i3 = zrib "*dépêcher*"
93. ςtb × 1a22i3 = ςattib "*franchir*"
94. fjς × 12a3 = fjaς "*effrayer*"

C. Table de déverbaux TUN-MSA

schème verbal	type de déverbal	schème nominal	
		MSA	TUN
I	participe actif	1A2i3	1A2i3 / 1A2a3
	participe passif	ma12uw3	ma12uw3
	forme infinitive	1a23 / 1u23/ 1i23	1a23 / 1u23/ 1i23
		1a2A3aħ	12iy3aħ
	nom du lieu	ma12a3	ma12i3 / 1u23aħ
		ma1A2i3	m1A2i3
	nom du temps	ma12i3	mu12u3
	nom d'outil	mi12A3 / ma1A2iy3	mi12A3 / m1A2a3
	adjectif analogue	1a2iy3	12iy3
	adjectif comparatif	Âa12a3	Âa12i3 / Âa12a3
	forme exagérée	1a22A3	1a22A3
II	participe actif	mu1a22i3	m1a22i3
	participe passif	mu1a22a3	m1a22i3
	forme infinitive	ta12iy3	ta12iy3
	adjectif analogue	1a22A3	1a22A3
III	participe actif	mu1A2i3	m1A2i3
	participe passif	mu1A2a3	m1A2i3
	forme infinitive	mu1A2a3aħ	12A3 / mu1A2a3aħ
IV	participe actif	mu12i3	mu12i3
	participe passif	mu12a3	mu12i3 / mu12a3
	forme infinitive	Ăi12A3	Ăi12A3
V	participe actif	muta1a22i3	mit1a22i3
	participe passif	muta1a22a3	mit1a22i3
	forme infinitive	ta1a22u3	1a2A3 / ta1a22u3
VI	participe actif	muta1A2i3	mit1A2i3
	participe passif	muta1A2a3	mit1A2i3
	forme infinitive	ta1A2u3	ta1A2i3
VII	participe actif	mun1a2i3	mun1a2i3
	participe passif	mun1a2a3	mun1a2i3
	forme infinitive	Ain1i2A3	Ain1i2A3
VIII	participe actif	mu1ta2i3	mu1ta2a3
	participe passif	mu1ta2a3	mu1ta2a3
	forme infinitive	Ai1ti2A3	Ai1ti2A3

schème verbal	type de déverbal	schème nominal	
		MSA	TUN
IX	participe actif	mu12a33	mi12A3
	participe passif	mu12a33	mi12A3
	forme infinitive	Ai12i3A3	12uw3iyya$\hbar$
X	participe actif	musta12i3	musta12i3
	participe passif	musta12a3	musta12i3
	forme infinitive	Aisti12A3	Aisti12A3
Q	participe actif	mu1a23i4	1a23A4
	participe passif	mu1a23a4	m1a23i4
	forme infinitive	1a23a4a$\hbar$	1a2i34a$\hbar$
QI	participe actif	muta1a23i4	mit1a23i4
	participe passif	muta1a23a4	mit1a23i4
	forme infinitive	ta1a23u4	t1a23iy4

Mes publications

2015 **Ahmed Hamdi**, Alexis Nasr, Nizar Habash, Núria Gala
POS-tagging of Tunisian Dialect Using Standard Arabic Resources and Tools
Workshop on Arabic Natural Language Processing
Annual Meeting of the Association for Computational Linguistics (ACL), Pékin, Chine

2014 **Ahmed Hamdi**, Núria Gala, Alexis Nasr
Building a Tunisian Lexicon for Deverbal Nouns
Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial) Workshop
International Conference on Computational Linguistics (COLING), Dublin, Irlande

2013 **Ahmed Hamdi**, Rahma Boujelbane, Nizar Habash, Alexis Nasr
The Effects of Factorizing Root and Pattern Mapping in Bidirectional
Tunisian - Standard Arabic Machine Translation
MT Summit, Nice, France

Ahmed Hamdi, Rahma Boujelbane, Nizar Habash, Alexis Nasr
Un système de traduction de verbes entre arabe standard et arabe
dialectal par analyse morphologique profonde
Traitement Automatique des Langues Naturelles (TALN), Les Sables d'Olonnes, France

2012 **Ahmed Hamdi**
Apport de la diacritisation dans l'analyse morphosyntaxique de l'arabe
Rencontre des Étudiants Chercheurs en Informatique
pour le Traitement Automatique des Langues (RECITAL), Grenoble, France