Le traitement automatique de l'arabe dialectalisé : aspects méthodologiques et algorithmiques

Houda Saadane

▶ **To cite this version:**

Houda Saadane. Le traitement automatique de l'arabe dialectalisé : aspects méthodologiques et algorithmiques. Linguistique. Université Grenoble Alpes, 2015. Français. NNT : 2015GREAL022 . tel-01692998

HAL Id: tel-01692998
https://theses.hal.science/tel-01692998

Submitted on 25 Jan 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTE UNIVERSITÉ GRENOBLE ALPES

Spécialité : **Informatique et Science du Langage**

Arrêté ministériel : 7 août 2006

Présentée par

Houda SAADANE

Thèse dirigée par **Prof. Mathieu GUIDERE**

préparée au sein du **Laboratoire de Linguistique et Didactique des Langues Etrangères et Maternelles (LIDILEM – EA 609),** dans **l'École Doctorale Langues, Littératures et Sciences Humaines.**

Le traitement automatique de l'arabe dialectalisé : aspects méthodologiques et algorithmiques

Thèse soutenue publiquement le **14 décembre 2015**, devant le jury composé de :

Madame Lamia Hadrich Belguith
Professeur, Université de Sfax (Rapporteur)
Madame Lynne Franjié
Professeure, Université de Lille 3 (Rapporteur, Présidente)
Monsieur Olivier Kraif
MCF (HDR), Université de Grenoble 3 (Examinateur)
Monsieur Christian Fluhr
Directeur Scientifique, GEOLSemantics (Examinateur)
Monsieur Nasredine Semmar
Chercheur, CEA-LIST (Examinateur)
Monsieur Mathieu Guidère
Professeur, Université de Toulouse 2 (Directeur)

SOMMAIRE

INTRODUCTION GENERALE---------- 3
PARTIE I : DESCRIPTION DE L'ARABE STANDARD ET DIALECTAL---------- 13
CHAPITRE 1 LA LINGUISTIQUE DE LA LANGUE ARABE---------- 14
CHAPITRE 2 INTRODUCTION AUX DIALECTES ARABES ---------- 32
PARTIE II : ANALYSE LINGUISTIQUE DE LA LANGUE ARABE---------- 54
CHAPITRE 3 ANALYSE MORPHOSYNTAXIQUE ---------- 55
CHAPITRE 4 IDENTIFICATION ET TYPAGE DES ENTITES NOMMEES ---------- 95
PARTIE III : TRAITEMENT DES DIALECTES ARABES---------- 120
CHAPITRE 5 ANALYSE PHONOLOGIQUE---------- 121
CHAPITRE 6 ANALYSE MORPHOLOGIQUE VERBALE ---------- 142
CHAPITRE 7 ANALYSE MORPHOLOGIQUE NOMINALE ---------- 191
CHAPITRE 8 ANALYSE MORPHOLOGIQUE ADJECTIVALE---------- 224
PARTIE IV : CONTEXTE ET MATERIEL / GENERATION AUTOMATIQUES DES RESSOURCES 246
CHAPITRE 9 CREATION DES LEXIQUES---------- 247
CHAPITRE 10 TRANSLITTERATION DES NOMS PROPRES ARABES ---------- 266
CHAPITRE 11 CONSTITUTION DES CORPUS ---------- 285
PARTIE V : RESULTATS EXPERIMENTAUX ET EVALUATION---------- 322
CHAPITRE 12 SYSTEME D'EVALUATION ET D'EXTRACTION DE CONNAISSANCES DU MSA 323
CHAPITRE 13 SYSTEME D'EVALUATION DE LA TRANSLITTERATION DES NOMS PROPRES 341
CONCLUSION GENERALE---------- 352
BIBLIOGRAPHIE---------- 358

Introduction générale

Le Traitement automatique du langage naturel (TALN) regroupe à la fois la linguistique, l'informatique et l'intelligence artificielle. Cette discipline est devenue un axe de recherche essentiel pour analyser et traduire la grande masse d'informations disponible, qui évolue sans cesse. De plus, les enjeux cognitifs du traitement automatique des langues sont importants, et varient selon les applications. De nos jours, il existe plusieurs applications du traitement des langues telles que la reconnaissance de l'écriture manuscrite (la détection de la langue), le résumé automatique, le traitement de la parole, l'annotation sémantique, l'indexation et la recherche de documents, l'extraction d'informations, la traduction, etc.

Le traitement morphosyntaxique automatisé de la langue arabe n'est pas récent, il a fait l'objet depuis plusieurs décennies de travaux novateurs, en particulier en France par des équipes de recherche qui se sont progressivement spécialisées dans le traitement de l'information multilingue. En ce qui concerne la recherche d'informations, la problématique de la recherche interlingue a été une motivation importante qui a conduit au développement de projets tels que EMIR (European Multilingual Information Retrieval), et son extension à l'arabe le projet ALMA (Arabic Language Multilingual Applications). Ainsi, comme nous venons de le montrer, plusieurs projets européens ont porté sur le traitement de l'arabe. Plus récemment, un réseau d'excellence européen a permis de regrouper la plupart des acteurs européens pour échanger des informations et produire des ressources linguistiques (dictionnaires, corpus étiquetés, logiciels) dans le cadre des projets NEMLAR (Network for Euro-Mediterranean LAnguage Resources) puis MEDAR (Mediterranean Arabic Language and Speech Technology).

Dans tous ces projets, le traitement automatique de la langue arabe écrite s'est focalisé, de façon presque exclusive, sur l'arabe classique ou standard, laissant de côté les dialectes et les phénomènes liés à l'usage dialectal de la langue arabe. Mais la prolifération de rédacteurs de blogs sur Internet et les contributions diverses et variées sur les forums de discussion en ligne a fait apparaître des usages langagiers de l'arabe standard fortement teintés de dialecte local, ou mixés avec une langue étrangère comme le français ou l'anglais, ou encore directement transcrits en lettres latines, ce qui nous conduit à nous poser des questions par rapport à l'état de la recherche en la matière.

L'arabe moderne standard, qui est pratiqué dans les journaux écrits, radiodiffusés et télévisés, a fait l'objet de nombreux travaux tant pour la reconnaissance de la parole que pour l'analyse et la recherche d'informations. Toutefois, l'essentiel des échanges entre personnes du monde arabe se fait dans le dialecte parlé localement. Même dans les émissions qui sont censées n'utiliser que l'arabe moderne standard, un nombre non négligeable d'expressions dialectales trahissent l'origine de la personne qui s'exprime. La prise en compte de l'arabe dialectal concerne aussi bien les applications sécurité (terrorisme, drogue, trafic d'armes, blanchiment) que les applications purement civiles comme l'analyse d'opinion, la reconnaissance de la parole et d'une manière générale tout dialogue ou instruction donnée à un appareil (téléphone ou autre) au moyen de la voix.

L'intérêt de traiter les dialectes a été reconnu depuis déjà un certain temps, toutefois la difficulté réside dans le coût de constitution de corpus représentatifs, en particulier pour la reconnaissance de la parole. La constitution de tels corpus est coûteuse aussi bien par la difficulté de recueillir des sources représentatives, que par le travail nécessaire pour leur transcription. La reconnaissance de l'origine communautaire de commentaires rédigés en arabe dialectal apparait néanmoins l'objet d'une vague d'intérêt récente, et qui s'amplifie.

Problématique du sujet :

En prenant en compte ces travaux, nous avons choisi de faire des recherches sur les aspects peu étudiés jusqu'ici. La problématique de notre sujet de thèse concerne les traits morphosyntaxiques et rédactionnels de l'arabe standard dialectalisé. Cela revient à poser plusieurs questions : comment distinguer, dans les productions, les usages relevant de l'arabe moderne, des usages relevant de l'arabe dialectal ? Comment reconnaître les traits spécifiques à chaque dialecte arabe ?

Ces questions problématiques nous incitent à envisager plusieurs axes d'étude et d'analyse :

1. Analyser les usages nouveaux introduits par le recours à la médiation de l'ordinateur et aux téléphones portables dans l'écriture de messages de diverses natures.
2. Établir pour chaque corpus d'arabe dialectal étudié les « écarts » observables par rapport à la langue arabe standard, que ce soit du point de vue lexical ou morphosyntaxique.
3. Identifier et définir des traits discriminants du corpus propre à chaque situation de communication, notamment en référence à la région géographique des rédacteurs (Maghreb vs Machrek, etc.).

La problématique de notre sujet porte donc à la fois sur la collecte de données d'étude par des moyens automatisés, sur une analyse automatique des données collectées pour faire apparaître les écarts par rapport à l'arabe standard et pour la mise en évidence du caractère discriminant de certains de ces écarts pour une population localisée géographiquement. Elle a aussi des liens évidents avec le problème plus général de l'évolution des modes d'interaction introduits par les nouveaux outils de communication et par la pratique des réseaux sociaux et autres communautés virtuelles.

Nous avons, dans cette perspective, constitué des corpus « locaux » pour mieux comprendre le phénomène de l'influence des langues locales et des langues occidentales mais aussi, vu l'origine des textes, l'influence des habitudes acquises par l'utilisation du web et des nouvelles technologies sur la langue arabe moderne.

Pistes d'investigation empiriques :

Pour étudier cette problématique dans une perspective de traitement automatique, nous allons constituer divers types de corpus et répondre à plusieurs questions touchant le traitement informatique de ces corpus :

- **Q1 :** Quelles sont les sources qui pourraient être exploitées pour constituer des corpus représentatifs de la langue utilisée dans les blogs et les forums de langue arabe ?
- **Q2 :** Comment peut-on identifier la région dont relève le dialecte considéré ; pouvoir trier, classer et regrouper des productions langagières par origine géographique ?
- **Q3 :** Quelles sont les techniques scripturaires utilisées par les rédacteurs (écriture en arabe, écriture latine de mots arabes, écriture simplifiée de type SMS, écriture mixte relevant du *code switching*, etc.) ?
- **Q4 :** Comment traiter l'écart existant avec l'arabe standard moderne, en particulier lorsqu'il relève du lexique ?

Ces questions nécessitent une étude approfondie des traits morphosyntaxiques de l'arabe standard et de l'arabe dialectal. C'est pourquoi nous allons constituer un corpus étendu et comparé des productions écrites des rédacteurs de blogs, des interventions dans les forums ou des messages courts sur les réseaux sociaux disponibles en langue arabe.

Ce corpus sera utilisé pour notre travail de thèse mais pourra être largement partagé pour que la communauté des chercheurs sur les dialectes arabes, de même que les spécialistes de la didactique de l'arabe, puissent réaliser d'autres investigations. Notre corpus sera traité par un outil d'analyse automatique de l'arabe classique et standard qui fera dans un premier temps ressortir le vocabulaire inconnu. L'étude de ce vocabulaire inconnu va permettre de classer les mots suivant des critères permettant de les situer par rapport à l'arabe standard afin d'enrichir le dictionnaire afférent du système automatique, avec des mots issus de l'arabe local de chaque région, et des termes provenant d'autre langues (français, anglais, tamazight,…).

L'étude que nous menons vise à permettre la réalisation d'analyses automatiques complètes des textes intégrant ces diverses variétés d'arabe. En effet, une fois les divers écarts par rapport à l'arabe standard identifiés et normalisés, nous les intégrerons dans le système d'analyse générale, et nous mettrons en place des méthodologies statistiques pour faire ressortir les traits les plus discriminants. Nous avons proposé des méthodes linguistiques et statistiques sur nos corpus pour identifier les origines géographiques des textes, qui pourront ensuite être appliquées sur de nouveaux textes pour en déterminer l'origine.

Nous avons choisi comme point de départ de l'étude les pays suivants : pays du Maghreb (Maroc, Algérie, Tunisie) et du Machrek (Égypte), l'objectif étant de donner un aperçu suffisamment représentatif de la diversité des apports à l'arabe dialectalisé par d'autres langues.

Contexte de mes travaux de recherche

Le travail de recherche dans ce manuscrit vise à construire des systèmes automatiques d'analyse linguistique de l'arabe standard afin de réaliser une extraction de connaissances dans des textes arabes. Il vise aussi le développement de ressources linguistiques et d'outils pour les dialectes arabes. Le développement d'une approche de reconnaissance du dialecte s'appuyant sur des dictionnaires de termes propres à chaque dialecte prépare de futurs travaux sur l'analyse linguistique du contenu de ces dialectes.

Les travaux de recherche présentés dans cette thèse se sont déroulés dans le cadre d'une bourse CIFRE (Conventions Industrielles de Formation par la REcherche), menés au sein du Laboratoire de *Linguistique* et *Didactique* des *Langues Etrangères* et*Maternelles* (*Lidilem*), Axe 1 « Descriptions linguistiques : syntaxe, sémantique, pragmatique et traitement automatique de la langue (TAL) » et de la société GEOLSemantics. Ils s'inscrivent dans le cadre de la mise en place d'une analyse morphosyntaxique de l'arabe standard et dialectal.

Les travaux de la thèse font partie de deux projets :

Le projet SAIMSI (Suivi Adaptatif Interlingue et MultiSource des Informations)

Le projet SAIMSI[1]., financé par l'ANR, conduit à développer une plateforme d'intégration d'informations multi-sources ouvertes multilingues concernant des entités nommées pour la détection de signaux faibles dans le cadre des missions de protection des citoyens face aux menaces intérieures ou extérieures. La plateforme agrège des informations de toutes sources (base de données existante, rapports, publication ou flux public internet,

[1] http://www.agence-nationale-recherche.fr/Colloques/WISG2013/articles/Projet-SAIMSI.pdf

web 2.0...), de média (texte, parole) ou de langue et de système d'écriture (français, anglais, arabe, russe…). Elle doit permettre de discriminer les informations sur des entités homonymes. Elle doit aussi permettre d'attribuer un texte ou une parole à un auteur même si ce texte n'est pas signé ou le locuteur authentifié. Les technologies utilisées feront un large appel à des analyses linguistiques multilingues profondes, à une extraction et normalisation inter-lingue d'informations structurées en fonction des besoins métiers et à une normalisation des entités nommées (personnes, sociétés, lieux, dates, mesures).

Ce projet regroupe cinq partenaires : des industriels et des laboratoires de recherche.

- GEOLSemantics est le leader du projet, développe le text mining interlingue, des bases de données textuelles et du traitement de la parole.
- Cassidian est l'architecte du système basé sur la plate-forme Weblab. Il est aussi l'intégrateur des différents modules et en fournit certains.
- Mondeca est en charge de la construction de la base de connaissance, des raisonnements automatiques, de la gestion et de l'utilisation de la base de connaissance.
- Le LIP6 est en charge des technologies de reconnaissance de l'auteur ainsi que des méthodes d'évaluation.
- L'IREENAT est en charge des aspects juridiques et déontologiques du projet.
- SAIMSI Partenariat : biométrie vocale d'Agnitio, transcription de parole Vocapia research

Le projet ORELO (Origine des Rédacteurs et des Locuteurs)

Le projet ORELO a été financé dans le cadre du programme RAPID par la DGA et la DGE. Ce projet a pour but de mettre au point des techniques d'identification de l'origine dialectale arabe d'un texte écrit en caractères arabes ou en écriture latine ou d'une parole. Cette reconnaissance permet d'apprécier l'origine géographique et communautaire des internautes de langue maternelle arabe. Ce travail est une étape indispensable pour permettre ultérieurement de suivre leurs échanges afin de recueillir les traces de leur parcours sur Internet.

De plus, les ressources linguistiques ainsi constituées (corpus et dictionnaires) sont une première étape vers une analyse du contenu des textes écrits en dialecte. En effet, la constitution de corpus de textes arabes en écriture latine, donc phonétique, permet de constituer à faible coût des modèles de langage pour la reconnaissance de la parole. D'autre part, l'approche de reconnaissance du dialecte proposée par GEOLSemantics s'appuyant sur des dictionnaires de termes propres à chaque dialecte prépare de futurs travaux sur l'analyse linguistique du contenu de ces dialectes.

Le projet ORELO est l'œuvre d'une réflexion commune entre GEOLSemantics, société spécialisée dans le traitement sémantique multilingue pour la sécurité, et la société Vocapia Research et le laboratoire LIMSI qui sont tous deux reconnus dans le domaine de la transcription automatique de la parole multilingue.

Le cadre des projets SAIMSI et ORELO a permis de mettre en place un contexte précis pour mes recherches qui se sont focalisées essentiellement sur le traitement automatique de la langue arabe standard et dialectal, ainsi que sur la construction et l'amélioration des systèmes d'analyse automatique complète des textes intégrant ces diverses variétés d'arabe.

Organisation de la thèse

Ce manuscrit comprend cinq parties principales. La première partie théorique est constituée de deux chapitres. Le premier chapitre présente la linguistique de la langue arabe standard et sa morphologie, puis le deuxième chapitre présente la langue arabe dialectale.

Dans le chapitre 1, nous décrivons brièvement la linguistique de la langue arabe standard. Le système d'écriture de la langue arabe est présenté. Nous présentons de même le lexique et la grammaire ainsi que la morphologie flexionnelle. Par la suite, nous décrivons les problèmes d'analyse qui posent le traitement automatique de la langue arabe.

Le chapitre 2 est dédié à une présentation de la langue arabe dialectale et de ses spécificités. Nous avons commencé par présenter la langue arabe ainsi que ses variantes utilisées, à savoir : l'arabe classique, l'arabe moderne standard (MSA) et l'arabe dialectal. Ensuite, nous avons mis l'accent dans ce chapitre sur les variétés de l'arabe dialectal. Par la suite, nous décrivons (une section y a été consacrée) un état de l'art sur la situation linguistique de la langue dans le monde arabe. Cela nous a conduit à donner un aperçu historique de l'arabe algérien. Finalement, nous faisons une étude qui compare l'arabe algérien, tunisien, égyptien et l'arabe standard sur plusieurs niveaux : phonologique, morphologique, orthographique, lexical et syntaxique.

La deuxième partie présente notre système d'analyse linguistique profonde de la langue arabe, et est constituée de deux chapitres.

Dans le chapitre 3, nous décrivons notre système de l'analyse morphosyntaxique. Nous passons en revue les travaux effectués pour le traitement automatique de l'arabe standard. Ensuite, nous présentons le fonctionnement ainsi que les différentes étapes de notre analyseur linguistique : la tokenisation ; l'analyse morphologique qui permet la segmentation des formes agglutinées. La désambiguïsation ainsi que les transformations morphologiques sont présentés dans ce chapitre. Finalement, nous décrivons la phase d'analyse syntaxique qui permet d'identifier les relations syntaxiques dans les groupes nominaux et verbaux.

Le chapitre 4 est consacré au traitement des entités nommées (ENs) en arabe (problématique de repérage et de typage des entités nommées en arabe). La typologie des entités nommées ainsi que les principales applications qui utilisent les entités nommées sont présentées dans ce chapitre. Par la suite, nous exposons les particularités de la langue arabe liée à la détection des entités nommées. Nous décrivons ensuite un éventail des travaux ayant comme focus la proposition de systèmes de reconnaissance des entités nommées en arabe. Ces systèmes sont à base de règles, statistiques ou hybrides. Notre approche de détection et de typage des entités nommées est décrite dans ce chapitre. Finalement, nous détaillons la méthode de reconnaissance des noms propres de type personne, lieu et organisation ainsi que la méthode de reconnaissance des expressions numériques.

La troisième partie de cette thèse est consacrée à l'étude complète et approfondie de la morphologie dialectale de la langue arabe. Cette partie a été inspirée de deux références principales (Gadalla, 2000) et (Marçais, 1902). Elle est constituée de quatre chapitres.

Le chapitre 5 est dédié à une analyse phonologique de la langue arabe (standard et dialectale). Nous présentons les principaux préliminaires phonologiques, qui sont répartis dans les systèmes consonantiques et vocaliques. De ce fait, nous présentons et comparons les systèmes consonantiques de l'arabe standard (MSA) et de l'arabe dialectal. Ensuite nous décrivons leurs systèmes vocaliques. Finalement, nous passons en revue les alternances phonologiques, appelées aussi les variations ou dégradations phonologiques à savoir :

l'assimilation, la métathèse, l'emphase, l'épenthèse, l'élision, et le raccourcissement.

Dans le chapitre 6, nous présentons une étude détaillée de l'analyse morphologique verbale, en comparant le MSA et l'arabe dialectal égyptien, tunisien et quelques particularités de l'algérien. Nous décrivons les différentes classes de verbes : les verbes trilitères (les verbes sonores, géminés, glottalisés et les verbes faibles) et les verbes quadrilatères. Puis, nous exposons les différents traits de flexion utilisés en MSA et en arabe dialectal. Ces traits comportent : l'aspect, le mode ainsi que la voix.

Le chapitre 7 est consacré à une présentation de l'analyse morphologique nominale d'une part en arabe standard (MSA), et d'autre part en arabe dialectal (égyptien et algérien). Nous décrivons les principales classes des noms : les noms primaires qui sont directement dérivés de la racine, et les noms déverbaux qui sont, eux, dérivés des verbes. Ensuite, nous exposons les formes des racines des noms primaires ainsi que les modèles des noms déverbaux. Par la suite, nous présentons la différence entre les noms définis et indéfinis. Finalement, nous décrivons les différents traits de flexion des noms, à savoir : le cas, le genre et le nombre.

Dans le chapitre 8, nous présentons la morphologie des adjectifs en MSA et arabe égyptien (AE) et arabe algérien (AA). Nous décrivons les formes des racines adjectivales ainsi que la différence entre les adjectifs définis et indéfinis. Ensuite, nous exposons les différents traits de la flexion des adjectifs, à savoir : le cas, le genre et le nombre. Puis, nous présentons les différents degrés de la flexion. Finalement, nous présentons les adjectifs relationnels.

La quatrième partie présente essentiellement nos contributions à la constitution des ressources, et est constituée de trois chapitres. Chaque chapitre permet de répondre à l'une ds questions que nous nous sommes posées tout au long de cette thèse. Comme nous l'avons déjà mentionné, depuis plus d'une décennie, la constitution des lexiques et des corpus dialectaux constitue un champ d'investigation très animé, qui a attiré l'attention de nombreux chercheurs. Cette tâche a pour objectif de pallier la carence en ressources en arabe dialectal, nécessaires pour le développement d'outils de traitement automatique des langues. Cette constitution des ressources linguistiques est la principale tâche à laquelle nous nous sommes intéressés et que nous avons traitée dans ce manuscrit. La première question que nous nous sommes posées était donc :

Q1 : Quelles sont les sources qui pourraient être exploitées pour constituer des lexiques dialectaux ?

Le chapitre 9 répond à cette question en décrivant notre méthode de constitution des lexiques dialectaux à partir des lexiques MSA et à partir des mots translittérés en écriture latine. D'une part, l'approche de constitution de lexiques dialectaux a été décrite en détails. D'autre part, Nous présentons deux méthodes différentes pour la constitution de ces lexiques. La première méthode consiste à dériver à partir des ressources MSA des lemmes dialectaux, alors que la deuxième approche consiste à utiliser des dictionnaires et des corpus écrits en latin, et de proposer une approche de translittération afin d'exploiter cette source. Lors de cette tâche de transcription, nous avons été confronté à la problématique d'absence de convention de transcription admise par la communauté scientifique. À ce sujet, il convient de signaler qu'il n'existe pas de norme commune ni de stratégie unifiée pour la transcription automatique du dialecte. Pour résoudre cette carence, nous avons développé une convention d'écriture nommée CODA (Saâdane et Nizar, 2015). Ensuite, nous avons décrit les principales lignes directrices de CODA (la Convention Orthographique des Dialecte).

Lors de la constitution des lexiques dialectaux ainsi que lors de la tâche de la

reconnaissance et de typage des entités nommées, nous avons mis l'accent sur le phénomène de la transcription/ translittération des mots et surtout les mots empruntés ou encore les noms propres étrangers. Dans le même registre, nous notons qu'une forme transcrite peut donner une indication sur l'origine de l'auteur (francophone ou anglo-saxonne). Afin de réduire l'impact d'un tel problème, nous avons développé un système de transcription/translittération des noms propres (et qui a été étendu et utilisé pour la transcription des mots). La translittération connait un essor important en raison du caractère de plus en plus multilingue de l'Internet et des besoins exponentiels dans le domaine de la recherche d'information interlingue. Cela est d'autant plus vrai pour la recherche d'entités nommées (noms de personnes, de lieux, de sociétés, d'organisations, etc.), mais ces dernières présentent une pluralité de formes écrites, d'orthographes et de transcriptions selon les langues et les pays. Le cas des noms propres en arabe illustre cette situation complexe et multiforme. Le meilleur exemple pour montrer cette pluralité est le nom معمر القذافي (Mouammar Kadhafi) qui est transcrit en latin par plus de 60 formes, parmi lesquelles : Muammar Qaddafi, Mo'ammar Gadhafi, Muammer Kaddafi, Moammar El Kadhafi, etc. Ceci nous a mené à nous poser les questions suivantes :

Q2 : Quelle est la stratégie des pays arabe dans le domaine de la translittération ?
Q3 : Y a-t-il une stratégie arabe unifiée en ce domaine ?
Q4 : Existe-t-il une stratégie de translittération au niveau de chaque pays arabe?

Dans le chapitre 10, nous étudions la translittération des noms arabes en écriture latine et inversement. Nous présentons dans ce chapitre les différents aspects liés au sujet de la translittération, à savoir l'aspect linguistique, l'aspect cognitif et dialectologique. Nous dressons ensuite un état de l'art sur le domaine de la translittération (les principaux travaux connexes au domaine de la translittération) suivi d'une description des approches que nous avons utilisées pour développer notre système de translittération automatique des noms arabes voyellés et non voyellés vers les différentes transcriptions possibles en écriture latine. Puis, nous présentons notre méthode de transcription des noms arabes en écriture latine vers l'arabe. Nous validons notre technique dans en présentant des expérimentations utilisant des moteurs de recherche de référence.

D'autres questions nous intéressent :

Q5 : Quelles sont les sources qui pourraient être exploitées pour constituer des corpus représentatifs de la langue utilisée dans les blogs et les forums de langue arabe ?
Q6 : Comment peut-on identifier la région dont relève le dialecte considéré ? Comment trier, classer et regrouper des productions langagières par origine géographique ?

Le chapitre 11 répond à ces questions en décrivant notre système de constitution des corpus dialectaux rédigés à la fois en écriture arabe et latine. Nous présentons un éventail de travaux ayant comme focus la constitution des corpus pour les dialectes arabes. Ensuite, nous détaillons les différentes étapes et démarches suivies pour la constitution de ces corpus. Nous commençons par effectuer une étude sur les sites identifiés et exploités pour la constitution des corpus. Par la suite, nous décrivons les outils utilisés pour la récupération des données, ainsi que les étapes d'extraction des données. Nous présentons par la suite la démarche adoptée pour l'annotation des corpus et l'identification des dialectes, autrement dit l'identification de l'origine dialectale des internautes. Nous présentons dans ce chapitre un aperçu sur les difficultés de l'identification des dialectes, ainsi que les applications qui

l'utilisent. Nous rappelons que l'annotation est faite au niveau des mots et des textes écrits en arabe et en caractères latins (Arabizi). Nous présentons également notre interface d'annotation, permettant de visualiser les résultats, et qui, par conséquent, facilite la validation des résultats de notre analyse linguistique d'une part, et permet d'annoter manuellement les mots hors vocabulaire afin d'enrichir nos dictionnaires initiaux d'autre part. Finalement, nous exposons quelques traits extraits pour la reconnaissance automatique des dialectes arabes.

La cinquième partie de cette thèse est consacrée aux expérimentations et évaluations qui ont été réalisées. Elle est constituée de trois chapitres

Ce chapitre est consacré à la présentation du système d'extraction de GEOLSemantics. Nous décrivons par la suite la chaine de traitement qui est divisée en trois modules complémentaires. Les deux premiers modules reposent sur une expertise acquise depuis des années dans le domaine du traitement automatique des langues. A partir d'un texte en langage naturel donné en entrée (la langue arabe dans notre cas), nous procédons à une analyse syntaxique profonde afin d'identifier les relations syntaxiques entre les différents unités de la phrase. Vient par la suite, l'extraction de connaissances consistant à formaliser ces relations sous forme sémantique. A l'issue de ces deux modules, nous disposons d'une extraction des connaissances formalisée en RDF. L'étape de mise en cohérence complète le traitement. Elle aide à pallier quelques lacunes dans le résultat RDF dues au traitement intraphrase des deux analyses précédentes.

Pour estimer l'efficacité de notre système, nous avons mené deux types d'évaluations : une évaluation quantitative concernant la phase de segmentation et la phase d'extraction d'entités nommées, et une évaluation qualitative de l'extraction de connaissances. Une comparaison de notre outil à un autre outil de segmentation a été réalisée. Les résultats montrent que notre outil est aussi performant que l'autre outil au niveau de la segmentation. La particularité de notre outil est qu'il est beaucoup plus rapide et analyse toutes les entrées lexicales. Ensuite, nous avons effectué nos expériences sur notre système d'extraction d'entités nommées. Finalement, une évaluation qualitative a été effectuée pour estimer la performance de nos règles d'extraction de connaissances.

Avant de passer à la phase de reconnaissance des dialectes, nous avons d'abord procéder à la vérification des résultats établis lors de la construction de nos ressources linguistiques, ce que nous avons développé comme lexiques dialectaux. Une série d'expérimentations et de tests d'évaluation de la couverture des ressources linguistiques développées pour les quatre dialectes a été effectuée dans le deuxième chapitre.

Le deuxième aspect concerne l'identification du dialecte aussi bien sur de l'arabe dialectal écrit en écriture latine qu'en écriture arabe. Notre approche consiste à utiliser des dictionnaires, en particulier des dictionnaires des mots les plus discriminants. Elle permet plus facilement de donner une valeur de rejet si le texte n'appartient à aucune des langues ou dialectes considérés. Elle permet aussi de déterminer les changements de langue.

Dans la perspective d'évaluer l'impact de l'utilisation de la translittération de noms propres sur la qualité d'un lexique bilingue français-arabe produit par l'outil d'alignement de mots intégrant la translittération, nous présentons dans le troisième chapitre, d'une part un outil d'alignement de mots simples et composés à partir de corpus de textes parallèles français-arabe, et d'autre part, les résultats d'évaluation de ce lexique bilingue selon deux approches différentes :

– une évaluation manuelle comparant les résultats de notre aligneur de mots par rapport à un alignement de référence,

– une évaluation de l'impact de cet alignement sur la qualité de traduction du système de traduction automatique statistique Moses

Les résultats obtenus montrent que la translittération améliore aussi bien la qualité de l'alignement que celle de la traduction.

Nous concluons ce mémoire de recherche en rappelant l'ensemble des contributions réalisées, puis nous exposons les différentes perspectives ouvertes par nos travaux.

Partie I : description de l'arabe standard et dialectal

Chapitre 1 La Linguistique de la langue arabe

Introduction

Ce chapitre est consacré à la définition et à la présentation de la langue arabe moderne standard (MSA) et de ses spécificités. Dans la section 1.1, nous avons commencé par une présentation générale de la langue arabe. Nous présenterons également le système d'écriture de l'arabe dans la section 1.2. La section 1 .3 est dédiée à une présentation du lexique et de la grammaire de la langue arabe. Nous exposons ensuite la morphologie flexionnelle dans la section 1.4. Finalement, la section 1.5 est consacrée à exposer les problèmes d'analyse du traitement automatique de la langue arabe

1.1. Présentation de la langue arabe

La langue arabe est l'une des langues les plus parlées et utilisées dans le monde. Elle est la langue officielle de plus de 22 pays parlée par plus de 320 millions de personnes et elle est utilisée comme vecteur de transmission religieux pour tous les croyants musulmans au nombre de 1 milliard et demi à travers les cinq continents du globe. Elle constitue ainsi un élément principal dans la culture et la pensée d'une partie importante de l'humanité et du patrimoine mondial.

A l'origine, les peuples de la péninsule arabe tenait le monopole de cette langue qui est sémitique (comme l'hébreu ou l'araméen), mais du fait qu'elle est la langue du coran elle s'est étendue au-delà du golfe arabo-persique, atteignant l'Afrique du nord et l'Asie mineur. De plus, l'expansion territoriale de l'empire musulman a fait de l'arabe une langue d'administration, de culture et de sciences à travers son utilisation dans la définition et la rédaction des contrats et des lois, la rédaction de manuscrits et de livres, la transmission et la formation, etc. Par ailleurs, la diversité des populations arabes et de leurs cultures ont fait émerger différentes variantes de l'arabe allant de l'arabe classique utilisé dans le coran, à l'arabe standard moderne (ASM) – sur lequel nous avons focalisé notre étude dans ce chapitre - représentant l'arabe officiel employé actuellement dans la presse, les documents officiels, etc; en passant par l'arabe dialectal influencé par les spécificités historiques et culturelles locales des populations constituant le monde arabe.
Historiquement, l'arabe tient ses origines au 2ème siècle et malgré son utilisation les premières traces écrites comme on la connait actuellement remontent au 6ème siècle. Ce fait peut être expliqué par l'analphabétisme des populations de l'époque qui communiquaient plus oralement que par écrits. L'apparition de l'islam a fait sortir l'arabe de son territoire d'origine et lui a donné une dimension internationale, en raison de son utilisation comme langue seule et unique pour tous les devoirs et rituels religieux, et du fait que le coran, comme texte sacré, ne peut être lu ou écrit qu'en arabe. Cette nouvelle dimension a multiplié considérablement l'utilisation de l'arabe dans les communications et échanges oraux et surtout écrits.

Cette expansion à la fois géographique et fonctionnelle a rapidement généré des réflexions sur la structuration et l'organisation de cette langue, mais aussi des intégrations et des emprunts de mots depuis et vers d'autres langues comme le français, le perse, le turc, etc. Vers le 9ème siècle, deux écoles linguistiques sont apparues en Irak et ont mis en place les bases d'une science du langage basée sur l'arabe. La controverse entre ces deux écoles, en occurrence celle de Basra (drivé par al-Mazini et al-Mubarrid) et celle de Kufa (mené par al-Kisä'i et la'lab), a permis de développer la grammaire de l'arabe[2].

[2] Voir : http://www.universalis.fr/encyclopedie/grammaires-histoire-des-la-tradition-arabe/

1.2. Système d'écriture de l'arabe

Comme mentionné dans la section précédente, l'arabe est classé sous le groupe des langues sémitiques contemporaines qui s'écrit de droite à gauche. Son système graphique se compose d'un alphabet arabe de type abjad constitué de 28 lettres. Cet alphabet contient 25 consonnes et 3 voyelles longues « و» ,« ا» et « ي». L'écriture arabe comporte aussi des voyelles courtes qui sont généralement facultative mais essentielles dans les textes religieux (Coran, Hadith, etc.). Il existe de plus, une série d'autres diacritiques dont les plus courants comme l'indication de l'absence de voyelle (سكون- *sukun*) et la gémination des consonnes (شدة - *shadda*). En arabe les mots indéfinis, qui ne sont pas associé à des articles ou à des compléments du nom, prennent les désinences (nounatation ou tanwine) notées par des diacritiques spéciaux.

Voyelle courte	Transcription	Nom
ـَ	A	Fatha
ـُ	U	Damma
ـِ	I	Kasra
ـْ	E	Sukun
ـّ	Doublement	Shadda
ـً	Aa	Fathatan
ـٌ	Uu	Dammatan
ـٍ	Ii	Kasratan

Tableau 1. 1. Les voyelles courtes en arabe

Nous signalons également que les notions de lettres majuscules et de lettres minuscules n'existent pas dans la langue arabe (l'écriture est donc monocamérale). Aussi, l'arabe est semi cursive dans le sens où son alphabet est unique mais la forme des lettres change en fonction de la position qu'elles occupent dans le mot. Chaque lettre possède une forme spécifique en fonction de sa position dans un mot (au début, au milieu ou à la fin) ou si elles sont utilisées de façon isolée.

1.3. Lexique et grammaire

Dans cette section nous donnons une présentation sommaire du lexique et grammaire de la langue arabe, tout en mettant l'accent sur les éléments qui seront pris en charge en priorité dans notre étude. Nous trouvons différentes structuration du lexique de l'arabe, basées essentiellement sur les sous-ensembles : noms, verbes et particules, et augmentées avec d'autres sous-ensemble afin d'avoir suffisamment d'éléments pour un traitement automatique de la langue. Nous trouvons entre autres les classifications de (Kouloughli, 1991) et (khoja et al., 2001). Nous considérons dans étude une classification proche de celle de (khoja et al., 2001) ayant les éléments suivants :

1.3.1. Nom

Est une entité ou un élément qui exprime un sens indépendamment du temps pour désigner un objet ou un être. Nous pouvons répartir les noms en trois catégories selon le système morphologique comme suit :

a. **Les primitifs :** sont les noms qui constituent le glossaire fondamental de la langue arabe, et représentent les noms qui ne peuvent pas être rattachés à une racine verbale. Cette catégorie inclue aussi les noms propres, les noms communs et les racine bilitères. Par exemple, nous citons رأس *raa's* 'tête', محمد 'Mohammed' et فم *fam* 'bouche'.

b. **Les dérivés :** sont les noms formés à partir d'une racine verbale. Le statut de cette dernière détermine la nature et le nombre de ces formes. Nous trouvons dans cette catégorie les participes actifs (– ضَارِبٌcelui qui frappe), les participes passif (– مضروب frappé), les noms de lieux ou de temps (– مَضْرِبُlieu de frappe), le nom d'instrument (مِضْرَبٌ- raquette), le nom d'une fois (– ضربةune frappe), etc.

c. Les nombres : ce sont les numéros simples représentant les unités (de ‘– ’صفرsifr : zéro- à ‘– ’تسعةtis'at : neuf-), les dizaines (‘– عشرونishruwn :vingt-) et les centaines (‘ ’مئة- : cent-), etc ; et les numéros composés comme les cardinaux, par exemple ‘ ستة – ’عشرseize.

1.3.2. Verbe

Est une entité portant un sens dépendant du temps et qui exprime une action, ou un événement. Les verbes arabes sont formés sur des radicaux de trois consonnes comme le verbe "دَخَلَ" (dakhala - entrer) et encore sur quatre consonnes comme le verbe "لَمْلَمَ" (lamlama - …). Ces racines peuvent donner naissance à d'autres schèmes ou patrons à travers des transformations morphologiques, comme le redoublement d'une consonne ou allongement d'une voyelle, donnant lieux à ce que nous appelons les racines à schème augmentées. Selon ces racines nous avons la classification de verbe suivante :

- **Verbe à racine simple :** verbe à trois consonnes et associer au schème "فَعَلَ" *(fa'ala)*. Si le verbe ne contient pas une voyelle longue, on l'appelle verbe sain (صحيح). Dans le cas contraire, appelé verbe معتل(*mou3tale*), nous distinguons les cas suivants en fonction de la voyelle longue et de sa position :
  - ✓ *verbe mahmouz* (مهموز) : si l'une des consonnes radicales est le glide "أ" (hamza), quel que soit sa position dans le verbe ;
  - ✓ *verbe assimilé (مثال- mithal)* : si la 1ère consonne radicale est le glide "و" *(w - wâw)* ou "ي" *(y – yâ')*
  - ✓ *verbe creux (أجوف- ajwaf)* : si la 2ème consonne radicale est "و" *(w)* ou "ي" *(y)*
  - ✓ *verbe défectueux (ناقص- naâqis)* : si la 3ème consonne radicale est l'un des glides "و" *(w)* ou "ي" *(y)*

  Par ailleurs, une autre classe de verbe existe et s'appelle *verbe redoublé (مضاعف- mudaâ'if)*. Elle est caractérisée par la présence dans un verbe de deux consonnes identiques en deuxième et troisième position du radical

- **verbe à racine augmentée :** ce type de verbe est obtenu, comme indiqué ci-dessus, par des opérations morphologiques appliquées à des racines simples afin de donner un sens particulier. Il existe différentes opérations utilisées, mais au final ces opérations intègrent une ou plusieurs lettres de l'ensemble rassemblé dans le mot *(سَأَلْمُتُونِيهَا- saaltemouniha)*. Parmi les fonctions morphologiques utilisées, nous citons :
  - ✓ **le redoublement :** qui consiste généralement à redoubler la deuxième consonne radicale du verbe, les verbes obtenus suivent le schème « فَعَّلَ» *(fa''ala)*
  - ✓ **l'allongement :** cette opération est réalisée par l'ajout du glide "ا" (alif) à la première consonne radicale, ce qui donne le nouveau schème « فَاعَلَ» *(faâ'ala)*
  - ✓ **l'adjonction :** cette opération permet d'ajouter une ou plusieurs lettres à la racine radicale dans des positions différentes tel que :
    - adjonction d'un morphème des trois consonnes "اِسْتَ" *(ista)* au début de la racine radicale du verbe. Cette opération donne naissance à nouveau schème qui a la forme « اِسْتَفْعَلَ» *(istaf'ala)*
    - adjonction du glide "ا" *(alif)* au début de la racine radicale et l'ajout du morphème consonantique "ت" *(t)* après la première consonne, les verbes

obtenus suivent le nouveau schème « اِفْتَعَلَ» *(ifta'ala)*

- adjonction du morphème consonantique "ت" *(t)* pour les verbes à racine quadratique (racine de quatre lettres) donnant le schème « تَفَعْلَلَ» *(tafa'lala)*
- adjonction du glide "ا" *(alif)* au début de la racine quadratique et l'ajout du morphème consonantique "ن" *(n)* après la deuxième consonne. Cette opération morphologique produit le schème « اِفْعَنْلَلَ» *(if'alala)*

1.3.3. Pronoms

En arabe, les noms invariables sont appelés des pronoms, et ils possèdent une structure et une flexion uniques quelle que soit leur place dans la phrase. Ils contiennent un type particulier et jouent une fonction syntaxique précise dans la langue. Nous citons entre autres les types suivants :

- **Les pronoms personnels** : sont des noms utilisés pour remplacer un nom ou désigner une personne ou un objet qu'ils soient absents, auditeurs ou locuteurs. Dans notre travail, nous étudions les pronoms personnels isolés et collés.
  a. ***Pronoms personnels isolés (ضمائر منفصلة) :*** Il s'agit des pronoms qui ne collent pas ni aux noms ni aux verbes. Ils s'écrivent seuls et détachés du nom. Nous classons ces pronoms dans les trois catégories résumées dans le tableau suivant :

Type de la personne	**Genre**		**Pronom**
1ère personne (المُتَكَلِّم-El motakalim : locuteur)	*Singulier*		'أَنَا' (anaâ - je)
	Pluriel		'نَحْنُ' (nahnu - nous).
2ème personne (المُخَاطَب-Elmokhatab : auditeur)	*Singulier*	*Masculin*	'أَنْتَ' (âanta – tu)
		Féminin	'أَنْتِ' (âanti – tu)
	Duel		'أَنْتُمَا' (ântimaâ – vous)
	Pluriel	*Masculin*	'أَنْتُمْ' (âantum – vous)
		Féminin	'أَنْتُنَّ' (âantunna – vous)
3ème personne (الغَائِب- El gha'ib : absent)	*Singulier*	*Masculin*	'هُو' (huwa – il)
		Féminin	'هِيَ' (hiya – elle)
	Duel		'هُمَا' (humaâ – ils)
	Pluriel	*Masculin*	'هُمْ' (hum – ils)
		Féminin	'هُنَّ' (hunna – elles)

Tableau 1. 2. Les pronoms personnels isolés

  b. ***Pronoms personnels collés (الضَّمَائِر المُتَّصِلَة) :*** les pronoms de ce type se trouvent collés à la fin des noms ou des verbes. Parmi ces pronoms nous citons :
    - ✓ *Ha' El gha'yba (هاء الغائبة) :* en français peut désigner *son, sa, ses*.
    - ✓ *Kaf el khitab (كاف الخطاب) :* en français peut désigner *ton, ta, tes, votre , vos.*
    - ✓ *Ya' El motakalim (ياء المتكلم) :* en français peut désigner *mon, ma, mes.*
    - ✓ *Noun El motakallimine (نون المتكلمين)* : en français peut désigner *notre, nos, nous (quand c'est collé à un verbe).*

- **Les pronoms démonstratifs (– أسماء الإشارةasmaâ' ichaâra)** : ces pronoms sont utilisés pour indiquer une ou plusieurs entités. Cette indication permet de situer l'entité dans l'espace, le temps ou tout simplement dans les textes et cela en fonction du contexte de la phrase. Ces pronoms ne sont déclinables qu'au duel. Les pronoms démonstratifs sont classés en deux classes selon la distance qu'ils désignent, les démonstratifs de proximité ((hawulaâ - ceux-ci) هَؤُلَاء, et هَذَا(hadaâ – lui-ci) et les

démonstratifs d'éloignement par exemple’ ’أُولَئِك(ûuwlaâyika - ceux-là). Le tableau ci-après résume l’ensemble de ces pronoms démonstratifs.

Démonstratif de proximité		**Démonstratifs d’éloignement**		**Caractéristiques**	
هُنَاك/ هُنَالِك		هُنَا		**Propre au lieu**	
Subjonctif & génitif	**Nominatif**	**Subjonctif & génitif**	**Nominatif**	**Cas du nom**	
ذاك / ذلك	ذاك / ذلك	هذا	هذا	Singulier	Masculin
ذينك	ذانك	هذين	هذان	Duel	
أولئك/أولالك	أولئك/أولالك	هؤلاء	هؤلاء	Pluriel	
تلك	تلك	هذه	هذه	Singulier	Féminin
تينك	تانّك	هاتين	هاتان	Duel	
أولئك/أولالك	أولئك/أولالك	هؤلاء	هؤلاء	Pluriel	

Tableau 1. 3. Les pronoms démonstratifs

- **Les pronoms relatifs (– أسماء موصولةasmaâ' mawsuwla)** : Il s'agit d'un nom placé avant une phrase appelée lien de conjonction contenant une information qui complète le sens de la phrase principale. L’ensemble de ces pronoms sont résumés dans le tableau suivant :

	Nominatif	**Subjonctif & génitif**	**Cas du nom**	
Pronom	الّذي	الّذي	**Singulier**	**Masculin**
	اللّذَان	اللّذَين	**Duel**	
	اللّذِين	اللّذِين	**Pluriel**	
	الّتِي	الّتِي	**Singulier**	**Féminin**
	اللّتَان	اللّتَين	**Duel**	
	الّلَاتِي / الّلائي	الّلَاتِي / الّلائي	**Pluriel**	

Tableau 1. 4. Les pronoms relatifs

1.3.4. Les mots outils

Sont des entités ou des particules clés employées pour situer des objets et des faits par rapport au temps et à l’espace et assurer ainsi un enchainement cohérent du texte. De plus, ils constituent des éléments importants dans l’interprétation du sens d’une phrase. Nous distinguons plusieurs types de ces mots, comme introduction, explication, conséquence, en fonction de leur sémantique et rôle dans la phrase. Parmi ces mots nous citons à titre d’exemple les éléments suivants :

- ***Les prépositions :*** فِي, فَوْق
- ***Les particules :*** كَيْفَ, لَمْ, لَنْ
- ***Les conjonctions de coordination :*** ثُمّ, وَ, أَمْ
- ***Les conjonctions de subordination :*** حَيْثُمَا, بَيْنَمَا
- ***Les quantificateurs :*** بَعْض, كُلّ
- ***Les adverbes :*** أَبَدًا, أَخِيرًا

1.4. Morphologie flexionnelle:

La flexion en linguistique est une opération de dérivation, qui ne crée pas de nouveaux

mots, mais qui permettant d'appliquer des modifications sur un lemme afin de dénoter des traits grammaticaux souhaités. Elle possède deux catégories : la déclinaison pour le système nominal et la conjugaison pour les verbes. Toute langue utilisant cette opération est appelée langue flexionnelle, et l'arabe en est une. En arabe, la flexion se concrétise par l'ajout des suffixes et préfixes (Blachère et Gaudefroy, 1966) aux lemmes pour refléter des indices d'aspects, de mode, de temps, de personne, de genre, etc. Dans la suite de cette section nous détaillons ces opérations selon les deux axes : déclinaison et conjugaison.

1.4.1. Flexions des verbes (conjugaison)

Les verbes possèdent une forme particulière de variation, que l'on appelle la conjugaison. La plupart des mots en arabe, dérivent d'un verbe de trois lettres, par conséquent chaque verbe est la racine d'une famille de mots. Ces mots sont obtenus en en ajoutant des suffixes ou des préfixes à leur racine. Le paradigme de la conjugaison est déterminé par certaine valeurs liées au genre, mode et structure morphosyntaxique du verbe. Ces valeurs peuvent être résumées comme suit :

- Le temps (accompli, inaccompli)
- Le nombre du sujet (singulier, duel, pluriel)
- Le genre du sujet (masculin, féminin)
- La personne (première, deuxième et troisième)
- Le mode (actif, passif).

Nous présentons dans le tableau suivant une idée globale sur la répartition de ces valeurs de conjugaison.

<table>
<tr><td colspan="3">Genre du verbe</td></tr>
<tr><td rowspan="2">Intransitif</td><td colspan="2">Transitif</td></tr>
<tr><td>Voix Active</td><td>Voix Passive</td></tr>
<tr><td colspan="3">Paradigme de conjugaison</td></tr>
</table>

1.4.1.1. Genre du verbe :

En arabe, un verbe peut se contenter seulement d'un sujet pour accomplir le sens de la phrase, comme dans la phrase → نام الطفل l'enfant a dormi, dans ce cas-là nous l'appelons verbe *intransitif*, le cas échéant nous l'appelons verbe *transitif*. Toutefois, les verbes transitifs peuvent avoir besoin de un ou plusieurs compléments pour compléter le sens de la phrase, par exemple dans la phrase → كتب الشاعر قصيدة l'écrivain a écrit un poème, nous trouvons un seul complément, contrairement à la phrase → سأل الصحافي الوزير سؤالا le journaliste a posé une question au ministre, qui comporte deux compléments nécessaires pour la compréhension de la phrase. Quel que soit le verbe, il est conjugué dans la forme active, sauf pour les verbes transitifs, de par leur besoin d'un complément, qui peuvent être exprimés dans la forme passive où l'agent est éliminé et le complément du verbe actif devient pro-sujet. Par exemple : la phrase dans la forme active, → أَكَلَ الولدُ التفاحةَ l'enfant a mangé la pomme, devient dans la forme passive : → أُكِلَتْ التفاحةُ la pomme a été mangée.

1.4.1.2. Paradigme de conjugaison du verbe

Il existe trois modes en arabe pour la conjugaison des verbes : *accompli, l'inaccompli et l'impératif*. Ces modes sont caractérisés par l'ajout de suffixe ou de préfixes traduisant les marques de personnes, genre et le nombre. Cependant, nous signalons que le mode accompli est caractérisé par l'ajout seulement de suffixes ce qui n'est pas le cas des deux autres modes.

- ✓ ***L'accompli (الماضي) :*** indique un fait ou une action qui s'est accompli ou effectué au passé ou au moment où on parle. Les verbes dans ce paradigme sont conjugués en ajoutant à la racine des suffixes permettant d'exprimer le type de personne, le genre, le nombre et le mode du sujet. Par exemple le verbe 'شَرَحَ' (charaha – expliquer), se conjugue pour la 2ème personne au duel par l'ajout le suffixe 'تُمَا' pour obtenir la forme شَرَحْتُمَا(charahtoûma – vous avez expliqué (duel)). Le tableau suivant donne l'ensemble des suffixes utilisés de manière générale, en prenant le verbe 'شَرَحَ' comme exemple.

'أنَا' (je)	'نَحْنُ' (nous)	'أنْتَ' (tu)	'أنْتِ' (tu)	'أنْتُمَا' (vous-2)	'أنْتُمْ' (vous)	'أنْتُنَّ' (vous)	'هُو' (il)	'هِيَ' (elle)	'هُمَا' (ils-2)	'هُمْ' (ils)	'هُنَّ' (elles)
شَرَحْتُ	شَرَحْنَا	شَرَحْتَ	شَرَحْتِ	شَرَحْتُمَا	شَرَحْتُمْ	شَرَحْتُنَّ	شَرَحَ	شَرَحَتْ	شَرَحَا	شَرَحُوا	شَرَحْنَ

- ✓ ***L'inaccompli (المضارع) :*** Il sert à exprimer tout fait ou action qui n'est pas écoulé, c'est-à-dire le présent ou le futur. Il dispose de préfixes et de suffixes à ajouter à la racine du verbe. Ce paradigme se caractérise par le fait que les marques de personne, genre, nombre et mode sont constituées de préfixe ainsi qu'une ou plusieurs infixations (due à des transformation) à travers des transformations morphologiques, comme le redoublement d'une consonne ou substitution des voyelles, comme c'est le cas du verbe مَسَّ *massa* 'toucher' conjugué en نَمَسُّ *namassu* 'nous touchons', تَمْسَسْنَ *tamsasna* 'vous touchez'. Nous distinguons les variantes suivantes de ce paradigme :
  - ➢ *Inaccompli indicatif (مرفوع) :* ce paradigme est du mode réel où le locuteur énonce le caractère réel (réalisé, devant être réalisé, en cours de réalisation, etc.) du procès-verbal qui désigne le déroulement dans le temps de la situation décrite par le verbe et il correspond soit à un état, soit à un processus, soit à un évènement. Il est structuré sur la voyelle /u/ qui indique par défaut l'indicatif.
  - · *L'inaccompli futur :* correspond à une action qui se déroulera au futur et est marqué par l'ajout de la lettre ' سa' ou de la particule سوف *sawfa* au début du verbe conjugué à l'inaccompli indicatif. Par exemple, pour le verbe كتب *kataba* écrire) nous obtenons سيكتب*sayaktubu* pour 'il écrira' ou سوف يكتب*sawfa yaktubu* qui signifie 'il va écrire'.
  - ➢ *Inaccompli subjonctif (منصوب) et apocopé (مجزوم) :* ces deux paradigmes sont de mode potentiel (sauf pour les deux négations لَمْ et لَنْ) où le locuteur se contente d'en énoncer la nature possible ou virtuelle du procès-verbal (Blachère et Gaudefroy, 1966). Il est nécessaire de préciser que la voyelle finale /a/ caractérise le subjonctif et l'absence de voyelle finale ou soukoun pour l'apocopé.

- ✓ ***L'impératif :*** il est utilisé pour exprimer un ordre, donner un conseil ou faire une suggestion ou une recommandation. Ce paradigme ne se conjugue qu'à la 2ème personne au singulier, duel et pluriel. La voyelle finale /i/ caractérise l'impératif (est structuré sur le soukoun) ou sur l'élimination du noun et de la lettre défectueuse du verbe non sain. Dans le tableau suivant nous donnons un exemple de conjugaison pour le verbe 'écrire' *kataba* 'كَتَبَ'.

أنْتَ' ،	أنْتِ' ،	أنْتُمَا' ،	أنْتُمْ' ،	أنْتُنَّ' ،

(tu)	(tu)	(vous-2)	(vous)	(vous)
أَكْتُبْ	أَكْتُبِي	أَكْتُبَا	أَكْتُبُوا	أَكْتُبْنَ

1.4.1.3. Trait grammatical

Par définition un trait grammatical, appelé aussi valeur, est une catégorie définie pour décrire les flexions morphologiques des mots variables. Ces traits concernent la nature (verbe, noms, adjectif), le genre (masculin, féminin), etc. Dans notre étude de la flexion des verbes en arabe, nous nous intéressons aux valeurs suivantes :

- ✓ **Aspectuelle :** cette valeur donne des informations sur la manière de déroulement du procès ou de l'état véhiculé par le verbe par rapport au moment où le procès a lieu, et non par rapport au moment où l'on parle. En arabe, cette valeur donne le caractère achevé ou inachevé du verbe indépendamment du moment où l'on parle. Exemple : ‘ أَكْتُبُ لَكَ هَذِهِ الرِّسَالَةَ لِأَسْأَلَكَJe t'écris cette lettre pour te demander …’ : processus en cours, inachevé. ‘ لَقَد كَتَبتُ لَكَ هَذِهِ الرِّسَالَةَJe t'ai bel et bien écrit cette lettre : processus accompli et achevé.
- ✓ **Modale :** cette valeur dénote la manière dont l'action exprimée par le verbe est conçue et présentée. elle exprime l'attitude du locuteur par rapport à ce qu'il dit (incertitude, souhait, etc.) ou à son destinataire (ordre). Sa combinaison avec la sémantique des verbes crée les aspects tel que : l'indicatif, le subjonctif, l'infinitif, etc.
- ✓ **Temporelle :** elle indique le moment où le procès a lieu. Elle permet d'exprimer la relation du déroulement du procès au temps passé, présent ou futur, par rapport au moment où l'on parle et à différents éléments du contexte ou de la situation. Exemples : ‘ سَأَرْجِعُje reviendrai’ (futur), ‘ رَجَعْتُje suis revenue’ (passé).

1.4.2. Flexions des noms

La déclinaison des noms en arabe est concrétisée en trois principaux cas : nominatif *(مَرْفُوع- marfu3)*, accusatif *(مَنْصُوب- mansub)*, génitif *(مَجْرُور- majrur)*. Ces déclinaisons sont faites en fonction du rôle du mot dans la phrase, à l'exception de certains cas particuliers. Les noms qui sont déclinable en arabe sont dits معربة(*mu'araba*). Ces déclinaisons sont traduites d'un point de vue graphique par un élément adjoint à la fin des formes nominales. La déclinaison est influencée par la forme du nom (simple ou diptote) et le nombre (singulier, duel ou pluriel) ainsi que le genre (féminin ou masculin). Dans la suite de cette partie, nous classons la flexion des noms en trois catégories selon le nombre de la forme comme suit :

1.4.2.1. Les déclinaisons au singulier

Selon la forme et la position dans la phrase du mot à décliner, nous distinguons les cas suivants :

- ➢ Le nom à décliner est défini par un article (‘’ال) ou par annexion (‘’إضافة), les désinences ou suffixes sont dhamma ' ُ' pour le cas nominatif; la fatha ' َ' à l'accusatif et la kasra ' ِ' pour le génitif.
- ➢ Le nom est indéfini, la déclinaison se concrétise par la nounatation (tanwin) par les trois signes ' ٌ' (- un), ' ً' (- an) et ' ٍ' (- in) pour les trois déclinaison respectivement. Notons que pour l'accusatif, la nounatation est associés avec un alif 'ا' sauf pour les noms se terminant par la lettre 'ة' (at) ou par 'اء'. A titre d'exemple, le nom 'درس' (dars – leçon), à l'accusatif nous obtenons 'درسًا' (darsäâ) contre le nom 'إمرأة' (imraât – femme) qui produit 'إمرأةً' (imraâtän – femme) à l'accusatif.
- ➢ Le diptote (الممنوع من الصرف: al-mamnu‘ min aṣ-ṣarf) est un nom qui ne respecte pas les règles de déclinaisons quand il est indéfini. Ces noms n'acceptent pas la

nounatation et prennent la même marque à l'accusatif et au génitif, à savoir la fatha ' ◌َ'. Ils existent des règles permettant de reconnaître ces noms, et dans le tableau suivant nous donnons certaines de ces règles accompagnées d'exemple :

Règle	**Exemple**
Noms propres féminins (العَلَم المُؤنّث)	هُدَى(Houda), زَيْنَبُ(Zaynab), جدَّةُ (Jeddah)
Un nom propre masculin, mais se terminant par le signe du féminin (عَلَمٌ مُؤنَّثٌ لِلْمُسَمَ المُذَكَّرِ)	حَمْزَةُ(Hamza), أسَامَةُ(Oussama), طَلْحَةُ (Talha)
Adjectifs et couleurs de schèmes af'alu (صِفَة ولَوْنٌ عَلَى الوَزْنِ أَفْعَلُ)	أَكْبَرُ(plus grand), أَسْوَدُ(noir), أَحْمَرُ (rouge)
Adjectifs de schèmes fa'lan (صِفَةٌ عَلَى الوَزْنِ فَعْلَانُ)	كَسْلَانُ(paresseux), عَطْشَانُ(assoiffé)
Les noms propres "Etrangers" (العَلَم الأعْجَمِي)	بَارِيسُ(Paris), إدْوَرْدُ(Edouard)

Tableau 1. 5. Certaines règles de déclinaison de diptote

- *Déclinaison des cinq noms :* c'est un ensemble de cinq exceptions bilitères qui se caractérisent par l'allongement de leur seconde syllabe lorsqu'ils sont définis par annexion. Autrement ils prennent les marques traditionnelles. Ces mots sont : أب (père), أخ(frère), حم(beau-père), فو(bouche), ذو(possesseur). Dans le tableau suivant nous donnons quelques exemples de flexion de ces mots :

Indéfini	Annexion	Indéfini	défini par l'article	annexion
أَبٌ	أَبُو بَكْرٍ	أَخٌ	الأَخُ	أَخُو مُحَمَّدٍ
أَبًا	أَبَا بَكْرٍ	أَخًا	الأَخَ	أَخَا مُحَمَّدٍ
أَبٍ	أَبِي بَكْرٍ	أَخٍ	الأَخِ	أَخِي مُحَمَّدٍ

1.4.2.2. Les déclinaisons au duel

Le duel est une sous-catégorie grammaticale pour représenter un ensemble de deux choses ou de deux personnes. C'est une catégorie qui se situe entre le singulier et le pluriel et qui possède les déclinaisons suivantes :

- Pour le mot indéfini ou défini par l'article, la suffixation est gérée par deux cas (ان) pour le nominatif, et (يْن) pour le génitif et l'accusatif. Par exemple, le duel du nom masculin رَجُل(rajül -homme) prend la forme رَجُلَان(*rajülAn* - deux hommes) au nominatif ou رَجُلَيْنِ(*rajül ayn* - deux hommes) à l'accusatif et au génitif
- Pour le mot féminin (se terminant par la lettre ة) une modification morphologique sera effectuée avant d'ajouter les suffixes. Cette modification consiste à transformer la lettre (ة) en (ت) afin d'ajouter les suffixes (ان) pour le cas nominatif ou (يْن) pour le génitif et l'accusatif.
- Pour le mot se terminant par l'un des glides suivants ((ا) (alif – â), 'و' (wâw- w) et (ي) (yâ' - y), des transformations seront appliquées sur ces glides pour obtenir le duel avec les suffixes décrit ci-dessus. Par exemple, le duel du mot مَلْهَى*(malha)* est obtenu en transformant la lettre (ى) « alif maksoura » en (ي- yaa) ensuite nous ajoutons les suffixations pour obtenir مَلْهَيَان(cas nominatif) et مَلْهَيَيْنpour les autres cas. Et pour le mot عصا(*'asaâ* – bâton) nous remplaçons d'abord le glide 'ا' par 'و' (wâw - w) ensuite nous ajoutons les suffixes ce qui donne pour le cas nominatif le mot عَصَوَان

(‘assawAn) et pour les autres cas le mot عَصَوَيْن *(‘assawayn)*.

1.4.2.3. Les déclinaisons au pluriel

Dans cette section nous présentons les deux types de pluriels suivants :

1.4.2.3.1. Le pluriel externe ou régulier :

Pour cette classe, le pluriel est obtenu par l’ajout d’un suffixe au singulier sans aucun changement au niveau de la structure du mot. Ces changements dépendent du genre du mot féminin ou masculin :

1.4.2.3.2. Le pluriel externe masculin : الجمع المذكر السالم

La flexion est réalisée par l’addition du suffixe 'ون' *(uwna)* dans le cas nominatif et du suffixe 'ين' *(iyna)* dans les cas accusatif et génitif. Par ailleurs, nous notons que si le mot est défini par annexion, nous supprimons la lettre 'ن' (noun) dans tous les cas. Dans le tableau suivant nous exhibons quelques exemples de ce type de flexion.

المعرفةDéfini	معرف بالإضافةDéfini par annexion	النكرةIndéfini	الجمع المذكر السالم
المُعَلّمُونَ	مُعَلّمُو الرّياضيات	مُعَلّمُونَ	Nominatif
المُعَلّمِينَ	مُعَلّمِي الرّياضيات	مُعَلّمِينَ	Accusatif & génitif

1.4.2.3.3. Le pluriel externe féminin : الجمع المؤنث السالم

A ce niveau la déclinaison au pluriel est réalisée par l’ajout du morphème ‘ ’اتau singulier féminin après la suppression de la lettre ‘ ’ة. Ensuite nous ajoutons la désinence dhamma ' ُ' pour le cas nominatif et la kasra ' ِ' à l'accusatif et au génitif. Le pluriel du mot ‘ سيارة’ (*sayArah* - voiture) est donné dans le tableau suivant :

المعرفةDéfini	النكرةIndéfini	الجمع المؤنث السالم
السَيَّاراتُ	سَيَّاراتٌ	Nominatif
السَّيَّاراتِ	سَيَّاراتِ	Accusatif & génitif

1.4.2.3.4. Le pluriel interne ou brisé (جمع التكسير)

Contrairement au pluriel interne où le singulier ne subit pas des transformations majeures, le pluriel externe brise la structure interne du mot au singulier suivant une diversité de règles complexes et dépendant du nom considéré. Par conséquent, les formes de ce pluriel sont nombreuses et généralement imprévisibles. De plus, pour certains mots nous trouvons deux types de pluriels : pluriel de multiplicité (جمع الكثرة) et pluriel pénurie (جمع القلّة). Des travaux ont été menés pour cadrer le pluriel interne, ce qui a donné des patrons pour les pluriels de pénurie et de multiplicité.

Pour le pluriel de pénurie, les quatre patrons suivant ont été établis : أَفْعُل، فِعْلَة، أَفْعَال، أَفْعِلَة (aaf-‘ilah, aaf-‘aal, fi‘-lah, aaf-‘ul), par exemple : (aaT-‘imah, Aliments) أَطْعِمَة, (aab-wab,portes) أَبْوَاب, (Sib-yah, Garçons) صِبْيَة, (aan-hur, riviéres) أَنْهُر. Pour le pluriel de multiplicité il existe 35 patrons du pluriel comme : (suhuwl, plaines) سُهُولselon le patron (fu‘uwl) فُعُوْلet (bul-daan, Pays) بُلْدَانsuivant le patron (fu‘-laan) فُعْلان.

1.5. Les problèmes d'analyse du traitement automatique de la langue arabe

L’arabe, comme toutes les langues naturelles, est caractérisée par un ensemble de phénomènes créant des difficultés et des problèmes qu’il faut prendre en considération lors

d'un traitement automatique. En plus des phénomènes classiques, comme l'ambiguïté, la coordination ou l'anaphore, nous trouvons aussi dans le cas de l'arabe d'autres phénomènes propres aux langues sémitiques tel que l'absence de voyelles, l'agglutination et l'ordre des mots dans une phrase. Dans la présente section, nous présentons les phénomènes que nous considérons les plus importants pour l'arabe.

1.5.1. L'absence de voyelle – voyellation –

Nous trouvons plusieurs définitions pour décrire le phénomène de la voyellation (Dichy, 1997 ; Debili et al., 2002), qui est concrétisée par l'absence des voyelles courtes, appelées aussi les diacritiques, dans les textes en arabe. Cette absence génère plusieurs cas d'ambiguïté compliquant ainsi le traitement automatique. Ces ambiguïtés lexicales sont dues essentiellement au fait que chaque consonne peut prendre l'une des sept voyelles de l'arabe, ce qui crée des combinaisons de mots dont le nombre diffère d'un mot non voyellé à un autre en fonction de l'existence de la combinaison obtenue dans le vocabulaire ou pas. Selon (Chalabi, 2000), l'absence de diacritiques en arabe entraîne une complexité de calcul d'un ordre de grandeur plus grand que la manipulation de ses homologues langues latines. Ce problème est d'autant plus complexe qu'un mot en arabe peut avoir différentes prononciations sans aucun effet orthographique en l'absence de diacritiques comme dans l'exemple suivant :

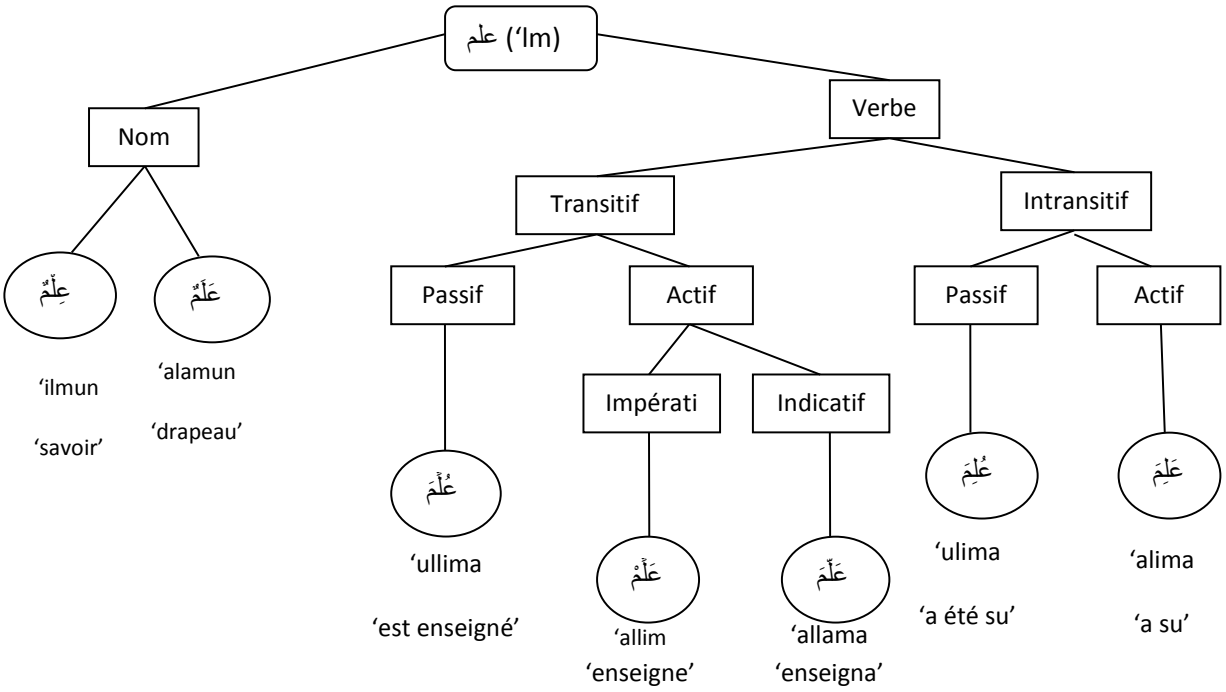


Figure 1. 1. Ambiguïté causée par le manque de diacritiques (Attia, 2008)

Dans cet exemple, nous voyons que le mot non voyellé ' 'علمpeut avoir sept voyellations différentes ayant pour chacune un sens particulier, réparties sur des catégories grammaticales différentes. Ceci engendre plusieurs cas d'ambiguïté lexicale comparable à celles posées par l'accentuation multiple des mots français non accentués. Pour illustrer cette comparaison, prenons le mot en français non accentué, *eleve*. Il peut être interprété comme *élève* (nom masculin ou Verbe, Présent de l'indicatif, Voix active, 1ère et 3ème personne, masculin/féminin, au singulier ou Verbe, Présent de l'impératif 2ème personne), ou *élevé* (adjectif masculin ou participe passé du verbe 'élever').

A travers ces différents exemples, nous voyons très bien les ambigüités que peut

engendrer ce type de phénomène, mais selon des études statistiques sur l'occurrence d'apparition de ce phénomène en français et en arabe, il a été démontré que ce phénomène est très fréquent en arabe par rapport à une faible fréquence en français (Debili et Achour, 1998) : 91.7% des mots du lexique français ne sont pas ambigües avec une moyenne de 1.1 accentuation possible par mot (El-Bèze, 1994), contre 19% des mots du corpus ne sont pas ambigües avec une moyenne de 6 voyellisations par mot (Debili, 2001; Ouersighni, 2002). Ces statistiques montrent qu'il est indispensable de prendre en compte cette problématique dans le cas d'un traitement automatique de l'arabe.

1.5.2. Agglutination

La langue arabe est une langue fortement agglutinée dans le sens où les mots peuvent être formés à partir d'une base à laquelle nous pouvons rajouter des affixes (préfixes et/ou suffixes) et des clitiques (enclitiques et/ou proclitiques). Dans le schéma suivant nous donnons une structuration globale d'un mot graphique en arabe, proposée par D. Cohen :

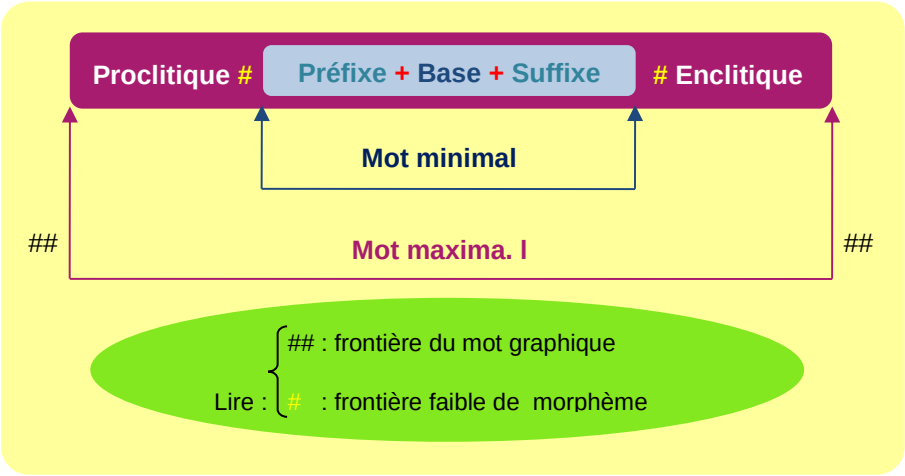


Figure 1. 2. Schéma général du mot graphique en arabe

Dans ce schéma nous voyons qu'un mot graphique contient essentiellement :

- *Une base* : qui représente la racine du mot à partir de laquelle l'agglutination est effectuée
- *Un mot minimal* : qui correspond à la forme fléchi de la base obtenue par la concaténation des préfixes et des suffixes à cette base
- *Un mot maximal* : unité décomposable en proclitiques, préfixes, base, suffixes et enclitiques. Elle peut être aussi analysé en proclitiques, mot minimal et enclitiques.

Ce mécanisme d'agglutination en arabe peut générer des mots qui peuvent être transcrits en une phrase complète en français. Par exemple la forme agglutinée 'أسنكفيكهمْ' (*asanakfikuhum*) qui peut être traduite en 'est ce que nous allons vous épargner de leur mal ?'. Les travaux de (Attia, 2008) supposent que le caractère riche et complexe des inflexions de l'arabe et l'agglutination des affixes et des clitiques permet de réduire l'ambiguïté plutôt que de l'augmenter en produisant une pyramide d'ambigüité reflétant les taux d'ambiguïté en fonction des structures du mot introduites ci-dessus, présentée dans la figure suivante :

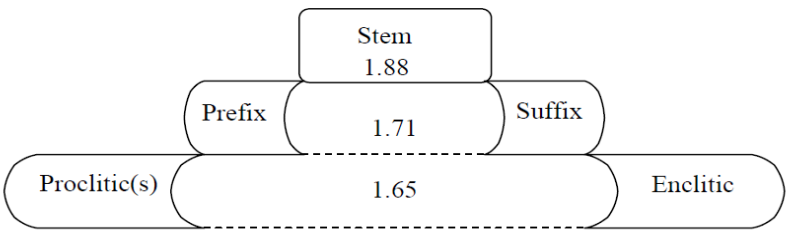


Figure 1. 3. Pyramide d'ambiguité (Attia, 2008)

Pour illustrer ce constat, nous prenons l'exemple de la racine ' 'كتب qui peut être interprétée en 'il a écrit' ou 'des livres' ou 'il a été écrit', et lorsque des clitiques lui sont ajoutées, l'ambiguïté est réduite : pour le mot ' 'يكتب (ajout d'affixe seulement) nous avons deux possibilités : 'il écrit' ou 'il s'écrit' et pour le mot ' 'يكتبه (ajout d'affixe et de clitique) nous avons que l'interprétation 'il l'écrit'.

Pour certains mots, l'agglutination peut entraîner une ambigüité morphologique au cours de l'analyse lorsqu'un clitique peut être assimilé à un caractère appartenant à la racine du mot. C'est le cas par exemple de la lettre ' 'ف (f) qui fait partie du mot ' 'فجر (aube, a fait exploser) et qui peut être aussi considérer comme un clitique collé au verbe ' 'جر (a tracté).

Dans le reste de cette section, nous allons décrire les clitiques (les proclitiques et les enclitiques) qui peuvent être collés à un mot minimal pour produire un mot maximal (ou la forme agglutinée).

1.5.2.1. Les proclitiques

Les proclitiques permettent de donner des traits syntaxiques (coordonnant, déterminant …) pouvant accompagner un mot arabe. Leur nombre est fini et peuvent se combiner entre eux pour être utilisés comme préfixes rattachés au 'mot minimal' ou détachés comme c'est le cas des conjonctions de coordination.

Lorsqu'un proclitique est rattaché à un verbe il dépend exclusivement de son aspect verbal, ainsi ils prennent tous les pronoms et par conséquent ils sont compatible avec tous les préfixes pris par l'aspect. Dans le cas des noms, les proclitiques dépendent du mode et du cas de déclinaison (Abbès, 2004). Nous pouvons répartir les proclitiques dans les catégories suivantes :

- *Les proclitiques réservés aux noms et adjectifs :*
  - L'article défini 'ال' (al- le)
  - L a préposition بِ (bi - avec), لِ (li - pour), ك (ka - comme)
- *Les proclitiques réservés aux verbes :*
  - La particule du subjonctif : nasb لِ (li - pour)
  - la particule du futur سَ (sa)
  - La particule de l'apocopé لِ (li - pour)
- *Les proclitiques généraux utilisés indépendamment de la catégorie des mots auxquels ils s'attachent :*
  - Les conjonctions de coordination ف (fa - et), et و (wa - et)
  - L'article d'interrogation أَ (a - est ce que)
  - Le marquer de corroboration لَ (la)

Cette classification n'omet pas le fait qu'il existe certaines exceptions de proclitiques qui peuvent jouer différents rôles, comme pour le proclitique ' 'و (wa) utilisé généralement comme particule de liaison (conjonction de subordination et de coordination), mais également peut être utilisé comme particule d'accompagnement (واو المعية) ou de serment (واو القسم).

Comme nous avons déjà mentionné les proclitiques peuvent se combiner entre eux et forment par conséquent des proclitique composé ('… ولك, ول, أفل, أفب, أف a-fa-li, 'a-fa-bi, 'a-fa, wa-li, wa-la-ka). Selon [Mesfar,] et [Habash, 2010], il existe quatre niveaux de clitisation selon la possibilité de leur apparition dans un proclitique composé, en respectant un ordre bien définit comme suit :

QST + [CNJ + [PRT + [DET + PRE + [BASE] + SUF + ENC]]]

1. *QST :* représente l'article d'interrogation ' أ'
2. CNJ : représente les conjonctions de coordination ' 'و(wa – et) et ' ' ف(fa, alors
3. PRT : représente l'ensemble de particules suivantes :
   - ✓ Les prépositions : بِ(bi - avec), لِ(li – avec) et ك(ka – comme)
   - ✓ La particule du subjonctif ' 'نصب(nasb) : لِ(li – pour)
   - ✓ La particule du futur : ' 'سَ(sa)
   - ✓ Le marqueur de corroboration ' 'تأكيد(taakiyd) : لَ(la)
   - ✓ La particule de l'apocopé ' 'جزم(gazm) : لِ(li)
4. DET: représente l'article défini ال(Al – el)

Pour illustrer ce propos, nous exposons la forme agglutinante suivante et qui est composée de plusieurs proclitiques en suivant leur position d'apparition : ' 'أَفَبِالبَيْتِdécomposable en ' بَيْتِ أَ' + فَ + بِ + ال +(*aa* + *fa* + *bi* + *l* + *bayti* – et + est ce qu+ à + la + maison ?).

Par ailleurs, nous signalons que la fusion des proclétiques n'est pas faite de façon aléatoire, elle suit deux types de contraintes exprimées par une relation d'ordre et un ensemble de règle de compatibilité comme suit :

- ***Une relation d'ordre :*** cette relation est établie en fonction d'un vecteur d'ordre selon (Dichy, 1984/89; 1994). Dans ce vecteur chaque proclitique est incompatible, à cause de la relation d'ordre strict, avec un proclitique de même position, c'est le cas par exemple des proclitiques *wa* et *fa* coordonnants (واو العطفet فاء العطف) qui occupent la position 2 dans le vecteur d'ordre. Nous notons aussi qu'un proclitique occupant une position d'antériorité par apport à un autre n'a aucune chance de se retrouver placé après ce dernier dans la construction d'un mot graphique. Par exemple, l'interrogatif 'a- (همزة الاستفهام) occupe toujours la première position dans un mot graphique maximal et il est impossible de le trouver précédé par un autre proclitique.
- ***Règles de compatibilité :*** pour des raisons syntaxiques et sémantiques, certains proclitique ne sont pas compatibles entre eux, c'est le cas par exemple des lettres بet ل(*bi-* et *li-) qui* ne peuvent pas se combiner, car elles sont des prépositions (حروف جر) ayant des sens différent (Dichy et Zmantar, 2009).

1.5.2.2. Les enclitiques

Les enclitiques présentent les pronoms suffixes qui s'attachent toujours à la fin du mot graphique, leur liste est constituée des 17 éléments suivants : نِي ي نَا كَ كِ كُمَا كُمْ كُنَّ هُ هَا هُمَا هُمْ هُنَّ هِ هِمَا هِمْ هِنَّ. Un mot graphique ne contient qu'un seul enclitique à la fois. Ils s'attachent aux verbes comme étant un complément-objet et aux noms et aux prépositions comme un complément du nom ou complément d'objet indirect. Leurs utilisations est régie par certaines restrictions.

Les enclitiques à la première personne tels que "نِي" (niy – moi / mon) ou "نا" (naa – nous/ notre) et ceux à la deuxième personne tels que "كَ" (ka – toi/ton) ou "كُمْ" (kum – vous/votre [masculin pluriel]) ont une forme invariable, mais ceux de la troisième personne sont variables et prennent différentes vocalisation suivants les règles suivantes :

- Dans le cas des verbes, l'enclitique peut varier en fonction de l'aspect du verbe et du pronom. La comptabilité entre les enclitiques et les verbes dépend de la propriété de transitivité du verbe. Ainsi, les verbes intransitifs et ceux conjugués à la forme passive ne prennent jamais des enclitiques. Par ailleurs, l'utilisation des enclitiques dans le cas

des verbes peut être répartie selon l'aspect du verbe comme suit (Mesfar, 2008) :

Aspect **Proclitique**	Inac-compli Actif	Inaccompli Subjonctif Actif	Inaccompli apocopé Actif	Impé-ratif	Accompli Actif	Accom-pli Passif	Futur	Pronoms
ـهُ(hu)	**X**							2éme personne, féminin, singulier
ـهُمَا(humaa)	**X**							
ـهُمْ(hum)	**X**							
ـهُنَّ(hunna)	**X**							
ـهِ(hi)	**X**							2ème ou 3ème personne, masculin ou féminin duel
ـهِمَا(himaa)	**X**							
ـهِمْ(him)	**X**							
ـهِنَّ(hinaa)	**X**							
ـهِ(hi)		**X**	**X**	**X**	**X**	**X**		2ème personne, féminin singulier
ـهِمَا(himaa)		**X**	**X**	**X**	**X**	**X**		
ـهِمْ(him)		**X**	**X**	**X**	**X**	**X**		
ـهِنَّ(hinaa)		**X**	**X**	**X**	**X**	**X**		
ـهُ(hu)		**X**	**X**	**X**	**X**	**X**	**X**	2ème ou 3ème personne, masculin ou féminin, duel
ـهُمَا(humaa)		**X**	**X**	**X**	**X**	**X**	**X**	
ـهُمْ(hum)		**X**	**X**	**X**	**X**	**X**	**X**	
ـهُنَّ(hunna)		**X**	**X**	**X**	**X**	**X**	**X**	

Tableau 1. 6. Utilisation des enclitiques dans le cas des verbes

- Dans le cas nominal, l'enclitique doit respecter une harmonie vocalique avec la voyelle casuelle de la forme à laquelle il se rattache, et dans le cas des noms se terminant par une voyelle double ou *tanwine*, ces derniers ne prennent jamais des enclitiques. Seul le mode déterminé par annexion est susceptible de prendre des enclitiques selon les règles suivantes :
  - ✓ Si le nom est fléchi au nominatif ou à l'accusatif, il nécessite l'utilisation des enclitiques suivants : ـهُ[PRON+3+m+s], ـهُمَا[PRON+3+m|f+d], ـهُمْ [PRON+3+m+p], ـهُنَّ[PRON+3+f+p]
  - ✓ Si le nom est fléchi au génitif, il nécessite l'utilisation des enclitiques suivants : ـهِ [PRON+3+m+s], ـهِمَا[PRON+3+m|f+d], ـهِمْ[PRON+3+m+p], ـهِنَّ[PRON+3+f+p].

Par ailleurs, certain mots nécessitent des transformations morphologiques avant de leur rattacher des enclitiques, c'est le cas des noms se terminant par une hamza, une "ى" ou une "ي" (y). Par exemple la forme مَلْهَى(*malha* - un manège), nécessite une transformation de celle-ci en "ا" avant sa suffixation pour produire la forme agglutinée مَلْهَاهُ(*malhAhu* - son manège).

1.5.3. Ambiguïté lexicale et syntaxique

L'un des problèmes centraux de l'analyse morphosyntaxique de l'arabe est l'ambiguïté lexicale et syntaxique, ce qui complique le travail des analyseurs lexico-syntaxique. Ces complications sont dues d'une part à la richesse des constructions et d'autre part à l'ambiguïté des segmentations en unités lexicales et à l'homographie polycatégorielle (Attia, 2006). Le traitement de ces ambiguïtés d'un point de vue informatique est alourdi par la combinatoire qu'elle engendre pour les analyseurs.

Par ailleurs, le problème ne réside pas dans l'analyse d'un langage ambigu en soi; mais c'est plutôt au niveau de son traitement de façon robuste et réaliste. En effet, après une première phase de segmentation du texte en unités lexicales, il est convenu de chercher dans

le lexique les interprétations correspondant à chacune d'entre elles. A chaque interprétation, nous associons une catégorie syntaxique reconnue par la grammaire.

L'un des aspects de la langue arabe qui cause cette ambiguïté, c'est le fait que beaucoup de mots en arabe sont *homographiques* : une même forme orthographique peut avoir des prononciations différentes. Cette homographie peut être accentuée lorsqu'elle est associée à d'autres phénomènes (absence de voyellation, morphologie flexionnelle et agglutinante, etc) ce qui donne des taux d'ambigüité assez élevés. Il a beaucoup de facteurs récurrents ayant contribués à ce problème, nous citons entre autres (Attia, 2006):

- Il existe dans l'arabe des mots homographes qui, sans flexion préalable, peuvent avoir différentes prononciations, des sémantiques différentes, voir généralement des catégories grammaticales différentes. C'est par exemple le cas du mot ذهب(dhb) qui a deux interprétations ذَهَبٌ(dahab) : or et ذَهَبَ(dahaba) : il est allé.
- La flexion des verbes contient des opérations morphologiques et orthographiques (suppression de caractères ou assimilation) qui produisent fréquemment des formes fléchies homographes. Ces formes peuvent appartenir à deux ou plusieurs lemmes. Dans l'exemple suivant nous montrons une forme verbale simple (– يعد y'd) qui peut être interprétée comme appartenant à cinq lemmes :

(أعاد يُعِد)	(عاد يَعُد)	(وعد يعِد)	(عد يَعُدّ)	(عد يُعِدّ)
yu'id, aa'âda	ya'ud, 'âda	ya'id, wa'ada	ya'udd, 'adda	yu'idd, aa'adda
il refait	il retourne	il promet	il compte	il prépare

- Le redoublement des lettres, au moyen de la lettre Shadda, crée des lemmes différents, sans que cela ne soit explicite à l'écrit. Le redoublement de la syllabe du milieu du mot درس.(drs) donne les deux lemmes suivants دَرَسَ(darasa) et دَرَّسَ(darrasa) ayant les interprétations 'il a étudié' et 'il a enseigné' respectivement.
- Plusieurs opérations de flexion induisent des changements légers dans la prononciation des mots sans que cela ait un effet orthographique explicit dû au manque de diacritique. Nous citons par exemple les ambigüités au niveau des formes fléchies du verbe كتبت(ktbt) :

كَتَبْتُ	كَتَبْتِ	كَتَبْتَ	كَتَبْتُ
katabtu – j'ai écrit	katabti – tu as écrit (féminin)	katabta – tu as écrit (masculin)	katabat – elle a écrit

- Les préfixes et les suffixes peuvent accidentellement produire une forme homographique avec un autre mot plein. Par exemple : le mot أسد('asd) qui peut signifier أَسُدّ أ+سد(*aasuddu* - je bloque) ou أَسَدٌ(*aasadun* - un lion).
- De même les proclitiques peuvent aussi accidentellement engendrer deux formes homographiques, comme c'est le cas de l'exemple suivant : علمي('lmy) qui donne suite à l'ajout des proclitiques علمي(*'ilmiyy* – scientifique) ou علم + ي)) علمي(*'ilm* + *y* - mes connaissances)

1.5.4. Irrégularité de l'ordre des mots dans la phrase

La construction des phrases en arabe est flexible, dans le sens où l'ordre des mots dans une phrase donnée est relativement libre. Généralement, un mot placé au début de la phrase est un terme sur lequel nous voulons attirer l'attention, s'en suit le terme le plus long ou le plus riche en sens ou en sonorité. Cette flexibilité provoque des ambiguïtés syntaxiques artificielles due à la prise en compte de toutes les règles de combinaison possibles des

composants d’une phrase. Pour illustrer cette propriété prenons les phrases suivantes :

- Verbe + sujet + complément :
  تأهلت الجزائر إلى كأس العالم(- L’Algérie s’est qualifiée pour la coupe du monde)
- Sujet + verbe + complément :
  الجزائر تأهلت إلى كأس العالم(- C’est l’Algérie qui s’est qualifiée en coupe du monde)
- Complément + verbe + sujet
  إلى كأس العالم تأهلت الجزائر(- C’est pour la coupe du monde que l’Algérie s’est qualifiée)

Chapitre 2 Introduction aux dialectes arabes

Introduction

La langue arabe est l'une des langues les plus parlées et utilisées dans le monde, elle occupe actuellement la cinquième place (Chung, 2008; Lewis, Simons et Fennig, 2013) avec plus de 330 millions d'arabophones, tout en devenant la langue officielle de plus de 22 pays, présentés dans la figure 1, répartis sur les régions suivantes :

- *Péninsule arabique (en arabe شبه الجزيرة العربية šibh al-jazīra al-'arabīya ou جزيرة العرب jazīrat al-'arab) :* est une vaste péninsule au sud-ouest de l'Asie, à la jonction entre ce continent et l'Afrique. Elle comprend les sept États suivants : l'Arabie saoudite, le Yémen, Oman, le Qatar, les Émirats Arabes Unis, le Koweït et le Bahreïn;
- *Moyen-Orient (en arabe الشرق الأوسط Ash-Shark al-awssat) :* cette région est comprise entre la rive orientale de la mer Méditerranée et la ligne tracée par la frontière entre l'Iran d'une part, le Pakistan et l'Afghanistan d'autre part. Cette région se trouve essentiellement en Asie mais est parfois étendue à l'Afrique du Nord. Elle comprend l'Irak, la Jordanie, le Liban, la Palestine et la Syrie. L'Égypte, avec sa péninsule du Sinaï en Asie, est généralement considérée comme faisant partie du Moyen-Orient
- *Afrique du Nord ou le Maghreb (en arabe : المغرب al-Maghreb) :* cette région inclut les états du Maghreb, à savoir l'Algérie, le Maroc, la Tunisie, la Mauritanie, la Libye, le Soudan, Djibouti et la Somalie.

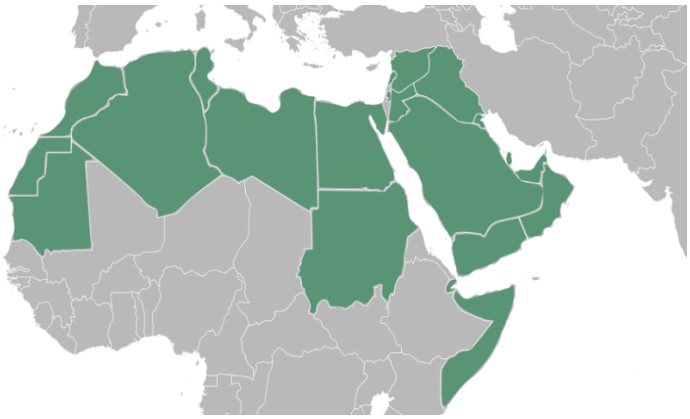


Figure 2. 1. Le monde arabe

Elle est par ailleurs la langue de la religion musulmane, ce qui étend son utilisation à tous les continents du globe constituant ainsi une communauté estimée à plus de 1 milliard et demi de croyants musulman. La langue arabe constitue ainsi un élément principal dans la culture et la pensée d'une partie importante de l'humanité et du patrimoine mondial.

D'un autre côté, l'arabe est une langue sémitique, comme l'hébreu et l'araméen, et en terme de nombre de parlers elle est actuellement la langue sémitique la plus parlée. De plus, l'arabe est une des langues naturelles les plus riches dans le monde en termes d'inflexion morphologique et de dérivation. Elle est caractérisée par le fait que l'arabe écrit diffère d'une manière non négligeable des différentes variétés parlées de la langue arabe ce qui a produit une situation diglossique où nous assistons à l'utilisation de deux variétés linguistiques d'une seule langue à savoir : l'arabe littéraire appelé *'variété élevée'* et l'arabe dialectal appelé *'variété basse'*.

Selon (Farghaly et Shaalan, 2009), l'arabe littéraire se divise en deux catégories : l'arabe classique et l'arabe moderne standard (MSA). L'arabe classique est utilisé pour les

textes et rituels religieux ainsi que les productions littéraires. Elle constitue la base de l'arabe moderne standard qui en constitue une forme moderne. L'arabe moderne (MSA) est utilisé dans les médias, les journaux et l'administration. Elle est aussi enseignée dans les écoles à partir du primaire.

Cependant, les locuteurs du monde arabe parlent en dialecte qui est une variante linguistique de l'arabe classique ayant des traits propres par pays ou par région, ces traits sont la conséquence d'une succession d'influences linguistiques, venues d'ailleurs comme le turc, le français, l'italien, et l'espagnol ou l'anglais, ou grâce à un mélange à des langues des peuples autochtones comme le berbère et le copte. Nous pouvons aussi considérer le dialecte comme un mélange homogène entre l'arabe moderne classique et l'arabe dialectal parlé par la population avec quelques différences d'une région à une autre et quelquefois au sein d'une même ville. Par conséquent, d'un point de vue scientifique, les dialectes peuvent être considérés comme des langues distinctes dans leur propre droit, un peu comme langues germaniques du Nord (Norvège, Suède, Danemark) et les langues slaves de l'Ouest (tchèque, slovaque, polonais) (Zaidan et Callison-Burch, 2014).

Par ailleurs, le MSA est la seule variété de l'arabe littéraire qui est normalisé, réglementé (standardisé). Elle est devenue indispensable pour la communication écrite et officielle. Quant aux dialectes, ils sont utilisés principalement pour la communication orale de tous les jours. Ils ne sont pas enseignés dans les écoles, et restent absent dans les communications écrites officielles. Cependant, il est possible de produire le dialectal en texte arabe, en utilisant les lettres utilisées dans le MSA et les mêmes règles d'orthographe du MSA, qui sont pour la plupart phonétique.

Ce chapitre est consacré à la définition et à la présentation de la langue arabe dialectale et de ses spécificités. La section 2.1 présente la langue arabe ainsi que ses variantes utilisées : l'arabe classique, l'arabe moderne standard (MSA) et l'arabe dialectal. Nous présenterons également les variétés de l'arabe dialectal dans la section 2.2. La section 2.3 est dédiée à une présentation de la situation linguiste de la langue dans le monde arabe. Nous donnerons ensuite un aperçu historique de l'arabe algérien dans la section 2.4. Enfin la section 2.5 est consacrée à une étude qui compare l'arabe algérien, égyptien et tunisien avec l'arabe moderne standard (MSA) sur plusieurs niveaux : phonologique, morphologique, orthographique, lexical et syntaxique.

2.1. Les variantes de langues arabes

La langue arabe est un terme vague qui fait référence aux nombreuses variétés existantes de la langue arabe. En effet, l'arabe possède plusieurs variantes depuis ses débuts. Il est à noter de ce fait que même à l'époque préislamique, l'arabe possédait déjà des dialectes distincts en un nombre considérable, comme c'était le cas entre des dialectes des tributs de Qahtane, Adnane et Himyar. Selon (Farghaly, 2010), il n'y a pas d'accord sur le nombre de variétés réellement utilisées aujourd'hui, et par conséquent il existe plusieurs classifications pour ces variétés. Par exemple (Ferguson, 1959a) définit deux variétés : la variété *élevée* ou l'arabe classique et la variété *basse* utilisée dans la communication quotidienne des arabophones (les dialectes). Nous citons aussi certaines classifications faites de manière locale comme celle du sociolinguiste (Badawi, 1973) réalisée pour l'arabe en Egypte qui met en avant les cinq variétés suivantes :

1. L'arabe classique patrimonial (fuSha al-turaaθ – فصحى التراث),
2. L'arabe classique contemporain (fuSha al-9aSr - فصحى القصر),
3. Le familier des éduqués (9aamiyyat al-muθaqqafiin - عامية المثقفين) ,

4. Le familier des éclairés (9aamiyyat al-mutanawwiriin - عامية المتنورين),
5. Le familier des analphabètes (9aamiyyat al-?ummiyyiin - عامية الأميين).

Cette classification a évolué entre temps, et son initiateur a proposé dans (Badawi, 1985), de nouvelles appellations aux variantes citées précédemment comme suit : 1. arabe classique, 2. arabe standard moderne, 3. arabe parle des instruits, 4. arabe parle des semi-instruits, et 5. arabe parle des analphabètes.

A l'époque moderne, l'arabe contient généralement au moins trois variétés qui coexistent côte à côte, à savoir l'arabe classique, l'arabe standard et l'arabe dialectal. La suite de cette section sera consacrée à la description de ces variétés.

2.1.1 L'arabe Classique

L'arabe classique est la variété la plus prestigieuse comme elle est la langue du Coran. C'est avec l'avènement de l'islam que la langue arabe a connu un véritable essor. Rappelons que pour les musulmans, la langue arabe classique est la langue sacrée de l'islam, de par le fait que le Coran a été révélé au prophète Mahomet par Dieu à travers l'archange Gabriel, en arabe classique, morceau par morceau, dans un arc de temps de 21 ou 22 ans et sous forme définitive. Selon (Djili, 2011), cette révélation du Coran en langue classique a marqué la naissance de cette dernière, et cette époque était appelée par certains linguistes et historiens, la première métamorphose de la langue arabe. La langue arabe est devenue une langue officielle du monde musulman en 685 quand le calife *Oumeya Abd Al Malik Ibn Marwan* arriva à Damas la capitale du monde musulman, avec pour objectif de centraliser son pouvoir politique : il a imposé donc l'arabe comme unique langue officielle. Le calife entreprend des réformes de l'écriture par la suite et prend de grandes décisions concernant les signes écrits. À partir du VIIIe siècle une codification au niveau de la grammaire fixa la langue dans sa forme classique définitive et facilita la diffusion de la langue par l'enseignement partout où la nouvelle religion 'l'islam' a pu pénétrer. C'est à cette époque que les premiers traités et dictionnaires sont apparus. Par conséquent, cette variété est bien définie, parce qu'elle a été codifiée par les premiers grammairiens arabes.

Elle s'est par ailleurs développée au fil du temps à travers son utilisation dans le développement des sciences et techniques, et dans la traduction des manuscrits grecs, de philosophie et de sciences, entre le VIIIe et le Xe siècle. Elle était aussi utilisée dans l'enseignement au sein des universités que ce soit à l'est de l'empire musulman, comme la maison de la sagesse à Baghdad, ou à l'ouest comme en Andalousie. Cette utilisation pour la science et la traduction a signé la seconde métamorphose de la langue arabe qui a fait d'elle une langue de civilisation qui a duré plus de quatorze siècles, et était arrivée jusqu'en occident.

Cet aspect a produit, au fil de l'histoire, un passage de la langue du Coran, comme expression de l'intelligence divine, et donc intouchable, inimitable et intraduisible, à la langue arabe comme expression de la perfection. A nos jours, il existe un consensus parmi les grammairiens arabes que la grammaire de l'arabe classique est complète comme elle décrit un corpus fermé contenant le patrimoine religieux et littéraire arabe.

2.1.2 L'arabe standard (MSA)

L'arabe moderne standard (MSA) est une forme de l'arabe, un peu différenciée de l'arabe classique, qui est utilisée chez les locuteurs arabes instruits dans les situations formelles. Le MSA est fondé syntaxiquement, morphologiquement et phonologiquement sur l'arabe classique avec un lexique plus récent. L'arabe moderne standard, appelé aussi *arabe formel*, est la forme de l'arabe utilisée dans la plupart des écrits administratifs, médiatiques, scientifiques, techniques, littéraires ainsi que dans la majorité des articles de presse et les journaux télévisés. Le MSA, constitue la langue écrite de tous les pays arabophones et de ce

fait elle est retenue comme langue officielle de ces pays, sans être la langue maternelle des populations de ces pays qui est généralement le dialecte. Cependant, le MSA n'est pas une variété bien définie car il n'a pas été complètement élucidé et décrit comme l'arabe classique. Le MSA se distingue de l'arabe dialectal par son système grammatical qu'il partage avec l'arabe classique, même s'il existe des constructions fréquentes dans l'un et qui sont considérées comme rares par l'autre.

Le MSA est donc la langue de communication non spontanée. Par conséquent, nous assistons, d'un point de vue sociologique selon (El Kassas, 2005) à deux mouvements en opposition. D'une part, l'apparition d'un langage des jeunes accentuant l'écart entre dialectes et normes de la langue, et d'autre part, un attachement à la langue classique et une envie de lui donner vie en tant que langue parlée. S'ajoute à cela, la globalisation qui donnera peut-être naissance à un futur MSA.

D'un autre côté, dire langue arabe, c'est donc parler d'un ensemble complexe dans lequel se déploient des variétés écrites et orales répondant à un spectre très diversifié d'usages sociaux, des plus savants aux plus populaires. Mais au-delà de cette diversité, les sociétés arabes ont une conscience aiguë d'appartenir à une communauté linguistique homogène. Elles sont farouchement attachées à l'intégrité de leur langue, d'où l'importance du MSA. Ce dernier constitue un terrain commun pour cette large population. Cet attachement est matérialisé de diverses manières : la multiplication des chaînes de télévision arabes par satellites et les sites arabes sur Internet ont contribué à augmenter la valeur et l'importance du MSA au sein de la société. Cette importance est augmentée d'avantage à travers la scolarisation, la constitution de grandes métropoles urbaines, les migrations interarabes, etc. Tous ces éléments constituent des facteurs qui ne font qu'accélérer le mouvement d'homogénéisation et d'harmonisation linguistique de l'arabe via la variante MSA. (El Kassas, 2005).

Le MSA possède par ailleurs des variations régionales. Par conséquent, nous pouvons détecter l'origine d'un texte marocaine, égyptienne ou en provenance des pays du Golfe. Cette variation est due à plusieurs facteurs parmi lesquels nous citons : i) les différences introduites par la création de nouveaux vocabulaires, ii) l'influence de l'histoire coloniale propres aux régions sur la syntaxe et la stylistique du MSA employé dans chaque région : les pays du Maghreb sont influencés par la littérature française alors que ceux du moyen orient sont influencés majoritairement par la littérature anglaise. Par exemple, الوزير الأول *alwaziir alawal* 'le premier ministre' traduit du français est le terme utilisé au Maghreb pour désigner le terme fréquent رئيس الوزراء *raʕiis alwuzaaraaʕ* 'le président des ministres' utilisé par ailleurs.

2.1.3 L'arabe dialectal

L'arabe dialectal est une autre forme de la langue arabe utilisé dans les communications quotidiennes, généralement appelée *ʻāmmiyya* "langue commune" ou *dārija* "langue courante". Cette variété possède également d'autres noms, parmi lesquels nous citons " l'arabe vernaculaire " -proposée par (Smith, 1917)- et "l'arabe parlé" (Salib, 1981). Elle est définie selon (Al-Toma, 1969) comme étant "la langue courante des activités quotidiennes, elle est généralement parlé, bien qu'elle soit parfois écrite. Elle varie non seulement d'un territoire arabe à un autre, mais aussi d'une région à une autre au sein du même territoire". Les dialectes populaires sont également bien définis; non pas parce qu'ils sont entièrement codifiée, mais parce qu'ils sont acquis naturellement par leurs locuteurs natifs.

Ainsi, presque tous les pays arabes ont leurs propres dialectes qui sont plus ou moins différents les uns des autres au sein du même pays, et plus naturellement, de ceux des autres pays. Ces différences dépendent considérablement de l'histoire de chaque pays et de son emplacement géographique. Prenons par exemple l'Algérie qui était une colonie française

après avoir été placée sous souveraineté de l'Empire ottoman. En dialecte algérien, le mot *table* emprunté du français et est dit طابلة *TaAblaħ* en dialecte algérien, de même pour le mot سكارجي *sukaArjiy* emprunté du turque qui signifie 'ivrogne'. Le dialecte algérien comprend également plusieurs termes qui dérivent du berbère comme par exemple قرجومة *Qarjuwmaħ* pour dire 'gorge'. Les systèmes grammaticaux des différents dialectes affichent de nettes divergences avec celui du MSA. Cependant, nous signalons que pour deux pays arabes frontaliers, les populations qui vivent des deux côtés de la frontière parlent des dialectes très proches partageant une bonne partie de leur syntaxe et lexique. Par exemple, dans la région qui se situe au Nord-Est de l'Algérie, regroupant les villes de Souk Ahras, Tébessa et Annaba; utilise un dialecte plus proche du dialecte tunisien que du dialecte algérien.

Par ailleurs, le dialecte, comme toute autre langue, se développe et s'adapte à chaque époque. Nous avons donc souvent de nouveaux termes qui apparaissent et qui peuvent dériver d'autres langues, sous la forme d'emprunt, comme mentionné dans les exemples ci-dessus. L'internet et les nouvelles technologies d'information et de communication ont aussi influencé les dialectes qui sont devenus de par leur utilisation de plus en plus comme langue d'écriture de ces supports. Les populations arabes utilisent le dialecte pour les échanges sur les forums, les SMS, le chat voir aussi les messages électroniques. Ces communications sont formulées soit en caractères arabes, ou aussi en caractères latins (arabe translittéré), selon les habitudes des utilisateurs avec les claviers arabes ou latins. Même s'il est écrit, le dialecte reste de l'arabe informel. En dialecte arabe, nous notons l'utilisation de plus des caractères arabes, des graphies qui n'appartiennent pas à la langue arabe, comme la lettre 'g' en dialecte tunisien ou algérien. Ces graphies sont utilisées pour écrire généralement des noms propres de villes ou de personnes. Les échanges sur les réseaux sociaux et les SMS ont aussi introduit l'utilisation des chiffres pour formuler certaines lettres arabe sans équivalent graphématique dans l'écriture latine, comme par exemple la lettre ح *H* qui est translittérée en chiffre '7', la lettre ص 'S' qui est translittérée en chiffre '9'.

D'un point de vue historique, selon (Farghali, 2010), il existe autant de théories sur l'origine des dialectes arabes modernes que des vues divergentes sur le nombre de premières langues arabe. Beaucoup de linguistes, comme (Versteegh, 1997), supposent que les dialectes arabes modernes se sont développés à partir d'un premier arabe dialectal parlé pendant les premiers jours des conquêtes arabes. La conquête islamique a étendu l'arabe à une vaste aire où diverses langues étaient parlées. Si les habitants des terres conquises ont parfois adopté la langue des conquérants, ils ont aussi été à l'origine d'un processus qui a conduit à l'émergence des dialectes.

D'autres grammairiens pensent que les dialectes modernes sont issus de l'arabe classique. A titre d'exemple, les gens qui ne savent pas comment parler correctement l'arabe classique, ont eu tendance à baisser les terminaisons de cas qui sont de ce fait prononcées avec un accent en introduisant de l'innovation lexicale. Un autre point de vue, qui est celui de (Ferguson, 1959b), réfute l'hypothèse précédente faisant un lien descendant/ascendant entre les dialectes et l'arabe classique. Il appuie son point de vue par l'énumération de quatorze caractéristiques linguistiques, essentiellement des traits phonologiques et morphologiques, que tous les dialectes partagent mais qui manquent en arabe classique. Il propose que tous les dialectes arabes proviennent d'une forme de l'arabe parlé, lors des contacts entre les populations des territoires conquis et les parlers des camps de bases militaires arabes positionnés dans ces territoires.

Enfin, nous signalons qu'il existe un grand nombre de différences linguistiques entre le MSA et l'arabe dialectal. Certaines de ces différences n'apparaissent pas sous une forme écrite mais ils sont au niveau voyelles courtes, qui sont omis dans le texte arabe de toute façon. D'autres différences se manifestent textuellement au niveau morphologique et

grammatical. La morphologie du MSA est plus riche que celle des dialectes en raison de la disparition des cas et des modes de flexion dans les dialectes. Par exemple, le MSA a une forme duale en plus des formes singulières et plurielles, alors que dans les dialectes manquent la plupart du temps la forme duale. Aussi, le MSA a deux formes plurielles, un masculin et un féminin, alors que de nombreux dialectes ne font souvent aucune distinction de genre au pluriel ou au singulier pour certains dialectes. D'autre part, les dialectes ont un système de cliticisation plus complexe que celui du MSA, ce qui permet la négation affixés (circonfixe), et l'attachement des pronoms aux objets qui agissent comme des objets indirects. Au niveau de la grammaire, le MSA dispose d'un système de cas complexe qui n'est pas présent dans les dialectes.

2.2. Les variétés dialectales de la langue arabe

La classification des dialectes arabes a intéressé les chercheurs et les observateurs depuis plusieurs années. Plusieurs classifications ont été proposées pour la répartition de ces dialectes au cours des années selon certains critères à savoir le critère géographique (horizontal) et le critère social (vertical). De ce fait, plusieurs grands groupes de dialectes, correspondant environ aux divers principes linguistiques, ont été proposés. Ces groupes répondent souvent à des divisions géographiques naturelles. Ce dernier constat est appuyé aussi par (Versteegh, 2011), qui avance que : '*les critères des classifications courantes ne sont pas toujours clairs. Dans une certaine mesure, ils semblent souvent ne refléter qu'une répartition géographique'*. Cette classification géographique, selon (Embarki, 2008), est relativement récente par rapport à d'autres classifications, comme la classification sociologique. La dialectologie arabe distingue généralement deux grandes zones ou familles principales de dialectes (Cohen, 1973; Barkat, 2000; Embarki, 2008; Saâdane et al., 2013, Baccouche, 1998) :

- *La zone occidentale* (l'Afrique du Nord, le Maghreb) : contient le groupe du Maghreb qui comporte l'Algérie, le Maroc, la Tunisie, la Libye et la Mauritanie,
- *La zone orientale* (le Machrek) : contient le groupe du Machrek comportant l'Égypte, la Syrie et les autres pays du Moyen-Orient (l'Irak, les Etats du Golfe, Yémen, Oman, Jordanie, etc.).

Selon (Baccouche, 1998) ces groupes sont séparés géographiquement et approximativement par l'Est libyen (du Sallûm au Tchad) et présentant plusieurs traits distinctifs morpho-phonologiques et lexico-sémantiques. Cependant ce découpage a été affiné, et la typologie qui en est issue, recueillant l'adhésion de plusieurs chercheurs, (Versteegh, 1997 et 2001; Habash, 2010), classe les parlers arabes modernes en cinq grandes aires dialectales (cf. Fig. 1), de l'Est à l'Ouest comme suit :

2.2.1 Les dialectes de la péninsule arabique (Golf)

Pour des raisons historiques le dialecte du golfe est le plus proche du MSA, étant donné que cette région constitue le berceau de la langue arabe d'une part, et d'autre part le MSA a évolué à partir d'une variété arabe originaire de la région du Golfe. Le dialecte du Golf conserve plus de traits du MSA par rapport aux autres dialectes, comme l'usage productif de la quatrième forme verbale ou le passif interne (Versteegh 2001). Cependant, le Golfe contient aussi des aspects le différenciant du MSA.

2.2.2 Les dialectes mésopotamiens (Irakien)

Ce dialecte est considéré parfois comme une variante du Golfe. Toutefois, il possède ses propres caractéristiques le distinguant du Golfe, notamment celles concernant les

prépositions, la conjugaison des verbes et la prononciation (Mitchell, 1990). D'un point de vue géographique, ce dialecte est utilisé par la population des bassins du Tigre et de l'Euphrate (Dajla et Alfwrat), en d'autres termes les parlers du nord de l'Irak et de l'Anatolie et ceux du sud de l'Irak. Nous signalons que plus nous nous rapprochons du sud de cette région plus les dialectes utilisés sont proches de ceux de la côte orientale d'Arabie.

2.2.3 Les dialectes levantins

Ce dialecte est utilisé par les parlers des pays suivants le Liban, la Syrie, la Jordanie et la Palestine. Cette région est connue aussi pour être un des bastions de la langue arabe depuis longtemps, elle fait partie des premières régions à être arabisée selon un processus rapide facilité par une forte présence arabe dans la région, et ce, dès avant l'islam. Les dialectes de cette région diffèrent quelque peu dans la prononciation et l'intonation, mais sont largement équivalent en écriture, et selon (Bassiouney, 2009) ils sont étroitement liés à l'araméen. Selon (Meillet et Cohen, 1981), les dialectes de cette catégorie peuvent être classés en trois groupes comme suit :

i. *Les dialectes libanais* qui concernent le dialecte de Beyrouth et celui de la Syrie (incluant celui de Damas),

ii. *Les dialectes du nord de la Syrie*, comme celui d'Alep par exemple,

iii. *Les dialectes palestino-jordaniens*, contenant certains dialectes de villageois et de citadins de la Jordanie et de la Palestine ainsi que ceux de certains parlers du sud de la Syrie.

2.2.4 Les dialectes égyptiens

Ces dialectes concernent l'Egypte essentiellement et constituent les dialectes les plus largement compris. Ce fait est dû essentiellement à l'influence politique de l'Egypte dans le monde arabe, surtout dans le 20ème siècle, ainsi que l'industrie cinématographique et télévisuelle de ce pays qui est très abondante, variée et massivement distribuée dans le monde arabe (Haeri, 2003). Les dialectes de cette catégorie sont classés par les dialectologues en quatre groupes :

i. Les dialectes du delta du Nil, qui se subdivisent eux-mêmes en dialecte de l'est et dialecte de l'ouest;

ii. Le dialecte du Caire considéré comme le dialecte le plus prestigieux comme c'est la langue de la capitale où se trouvent les bureaux de l'administration gouvernementale, c'est aussi la langue du cinéma, du théâtre et des divers médias. Ce dialecte est généralement parlé par un grand nombre de personnes instruites et cultivées.

iii. Les dialectes de la Moyenne-Égypte, s'étendant de Gizhz à Asyut,

iv. Les dialectes de la Haute -Égypte, qui s'étendent de Asyut jusqu'au sud du pays. Il convient d'ajouter également les parlers tchado-soudanais qui sont inclus dans l'aire égyptienne, particulièrement dans le sud de l'Égypte (Meillet et Cohen, 1981; Cohen, 2002).

2.2.5 Les dialectes maghrébins

Les dialectes de cette catégorie sont caractérisés par une forte influence des langues française et berbère. La plupart des dialectes considérés peuvent être inintelligible par l'orateur dans d'autres régions du Moyen-Orient, en particulier sous forme orale. La géographie du Maghreb lui procure une grande région, de ce fait elle présente une plus grande variation de dialecte, plus importante que celle perçue dans d'autres régions comme le Levant ou le Golfe. Elle peut être aussi divisée en d'autres sous-catégories.

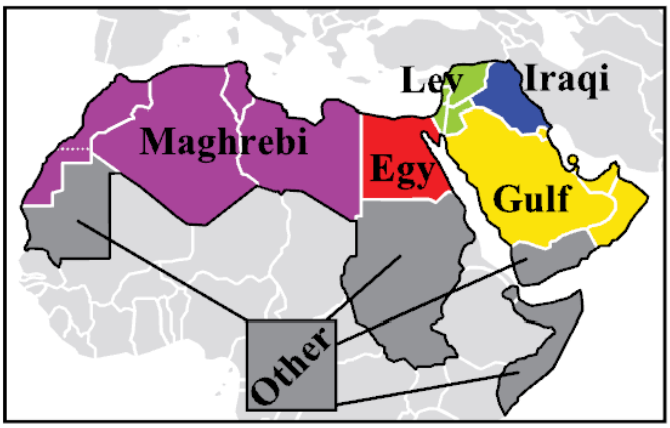


Figure 2. 2. Classification des parlers arabes modernes en aires dialectale

En plus de la géographie, le critère social est aussi proposé par certains chercheurs pour la stratification des dialectes, comme celle qui répartie les dialectes en deux groupes : groupes *citadins* et groupes *bédouins*. Cette classification est soutenue dans (Embarki, 2008) qui explique : '*les linguistes et autres observateurs de l'aire arabophone ont montré depuis longtemps que la plus petite localité comme la région la plus étendue sont traversées par une division entre ʿarab (nomades) vs ḥaḍar (sedentaires). Le terme ḥaḍar correspond à une population sédentaire, de type citadin ou villageois; quant à ʿarab, il englobe des populations nomades et semi-nomades*'. Ceci porte le nombre de classes dialectales à trois : 1) parlers bédouins nomades, 2) parlers bédouins sédentaires, et 3) parlers citadins.

2.3. La situation linguistique de la langue arabe

Linguistiquement parlant, la situation de la langue dans le monde arabe est caractérisée principalement par l'utilisation de deux variétés de l'arabe : l'arabe standard (MSA) et l'arabe dialectal. L'arabe standard est la langue de la littérature utilisée essentiellement dans la lecture et dans l'écriture des contenus des journaux, des revues, etc.; avec une utilisation quasi nulle dans les communications orales. Cependant, le dialecte est la langue de relations sociales quotidiennes, utilisée essentiellement pour les communications orales, elle est de ce fait la langue maternelle des populations du monde arabe.

Ces deux variétés partagent une grande partie de leur lexique, cependant il existe plusieurs indicateurs permettant, pour un lecteur/auditeur, de faire la distinction et d'identifier la variante utilisée. Parmi ces indicateurs, nous avons certains préfixes verbaux, la construction négative, la construction démonstrative et beaucoup de mots spécifiques à chaque variante, etc. Nous signalons à ce niveau que les mots spécifiques marquent une rupture nette entre le MSA et le dialecte. De plus les écarts entre les deux variantes peuvent aussi être illustrés, selon (Boukadida, 2008) par les éléments suivants :

- La disparition des désinences flexionnelles dans les dialectes;

- Le changement au niveau vocalique du système verbal. Pour le MSA, nous trouvons surtout une alternance vocalique, a/i et i/a dans l'opposition accompli/inaccompli, alors que pour les dialectes il y a une certaine similitude entre les deux : la voyelle de la deuxième consonne est l'élément le plus stable du schème et du mot;
- La variation syllabique du dialectal. Cette variation a entraîné une plus grande variation schématique ce qui introduit une souplesse structurelle plus étendue dans les mots et une possibilité d'intégration des emprunts et des néologismes.

Compte tenu des éléments introduits ci-dessus, nous pouvons dire que l'arabe dialectal possède un lexique très riche, surtout en vocables étrangers, en plus d'une morphologie et syntaxe simplifiées ce qui le distingue de l'arabe standard.

Comme mentionné, la société arabe utilise deux variantes de la même langue, ce qui constitue une 'diglossie' qui est un phénomène connu dans la littérature linguistique introduit pour la première fois par le linguiste (Marçais, 1930) dans le cadre des études faites pour caractériser la situation linguistique du monde arabe. Ce terme a été emprunté par la suite et défini par le linguiste (Ferguson, 1959a) dans un article intitulé « Diglossia » comme étant: « *Une situation de langagière relativement stable dans laquelle, en plus des dialectes primaires de la langue (qui peuvent inclure une ou plusieurs normes régionales), il existe une variété superposée, très divergente, hautement codifiée (souvent plus complexe du point de vue grammatical), elle véhicule d'une grande quantité de la littérature écrite vaste et respectée, soit à une époque antérieure soit dans une communauté linguistique. Cette variété est apprise essentiellement par l'enseignement et est utilisée pour la plupart des fonctions écrites et des fonctions orales à caractère formel, mais n'est pratiquée par aucun groupe de la communauté pour les conversations ordinaires.* »

Dans les situations diglossiques, les deux variétés linguistiques d'une seule langue sont baptisées : *variété H* (High, élevée) et *variété L* (Low, basse). La variété H est généralement utilisée dans le système éducatif, religieux et littéraire (les livres littéraires ou scolaires, les journaux, les publications gouvernementales, etc.) de par le fait qu'elle est standardisée, codifiée et normalisée. Elle jouit d'un statut social prestigieux et elle est très valorisée dans la société. Quant à la variété L, elle constitue le moyen de communication de vie quotidienne employée dans les conversations informelles, la littérature orale, les interviews, etc. Elle est généralement la langue maternelle acquise naturellement (sans apprentissage). Cependant, elle ne possède pas le même statut prestigieux et la même valorisation dont bénéficie la langue H. Le tableau suivant résume l'ensemble des domaines d'usage de ces variétés proposés par (Calvet, 1987) illustrant une situation diglossique :

Situations	**Variété haute**	**Variété basse**
Sermons, culte	+	
Ordre des ouvriers, serviteurs		+
Lettres personnelles	+	
Discours politiques, assemblées	+	
Cours universitaires	+	
Conversations privées		+
Informations sur les médias	+	
Feuilleton		+
Textes des dessins humoristiques		+
Poésie	+	

Littérature populaire		+

Tableau 2. 1. Cas d'usage des situations diglossiques

En conclusion, la situation sociolinguistique de la langue arabe s'inscrit amplement dans une conception diglossique, dans la mesure où il existe deux variantes de la langue arabe : d'une part, l'arabe standard moderne MSA (variété H), qui est une langue prestigieuse, valorisée, standardisée et reconnue comme langue officielle, et d'autre part, l'arabe dialectal (variété L), qui est réservée aux échanges informels de la vie quotidienne en plus du fait qu'elle est généralement la langue maternelle des arabophones.

2.4. Aperçu historique de l'arabe algérien (AA)

Les dialectes arabes ou langue vernaculaires constituent les langues maternelles des arabophones. Comme présenté ci-dessus, la dialectologie arabe distingue trois groupes différents de dialectes à l'intérieur du grand ensemble géographique que constitue le monde arabe. D'abord, les dialectes du Maghreb (le groupe de l'Ouest) où l'on trouve : l'Algérie, la Mauritanie, le Maroc, la Tunisie et la Libye. Ensuite, les dialectes du Machrek (le groupe de l'Est) où l'on trouve : l'Egypte, la Syrie, le Liban, la Jordanie et la Palestine. Enfin, les dialectes du Golfe où l'on trouve l'Arabie Saoudite, le Yémen, Oman, les Émirats arabes unis, le Qatar, le Bahreïn, le Koweït et l'Irak. Mais à l'intérieur de ces familles de géolectes, nous trouvons aussi bien des dialectes nationaux (natiolectes) que des dialectes régionaux (régiolectes) et même des dialectes locaux (topolectes), parlés sur un espace limité (village, localité) (Saâdane, 2011).

Le dialecte algérien, noté AA, est l'un des dialectes du Maghreb parlé en Algérie. Ce dialecte est aussi appelé دارجة *daArjaħ*[3], جزايري *jazaAyriy* ou دزيري *dziyriy* signifiant simplement 'algérien'. Ce dialecte est considéré comme un langage de basse variété (Faible variété). Ceci signifie que l'AA est faiblement normalisé et standardisé. Il est utilisé dans la presse, la télévision, la communication sociale, les échanges Internet, SMS, etc. Il est à mentionner que seules les communications officielles en lecture et en écriture n'utilisent pas le dialecte AA. Cependant, même si AA est parlé par la population de l'Algérie, estimée à 40 millions de personnes, il est caractérisé par une variation de ce même dialecte en fonction de l'emplacement géographique des locuteurs de l'AA. Ces variations ne créent généralement pas d'obstacles à comprendre le dialecte. En plus de AA, la population algérienne parle aussi le Berbère mais avec des rapports différents: AA est utilisé par 70 à 80% de la population, cependant la langue berbère est la langue maternelle d'une communauté importante de la population algérienne : 25% à 30% d'algériens sont des natifs berbérophones. La langue berbère est utilisée principalement dans le centre de l'Algérie (Alger et la Kabylie), l'Est de l'Algérie (Béjaia et Sétif), dans les Aurès (le chaoui), dans le Mzab (nord du Sahara) et il est utilisé par les Touaregs basés dans le sud du Sahara (Hoggar).

De plus, le dialecte AA est influencé principalement par trois langues : l'arabe, le berbère et le français. A ce titre, nous citons la définition du célèbre humoriste et comédien algérien, Mohamed Fellag, qui décrit le AA comme suit : « *L'algérien de la rue est une langue trilingue, un mélange de français, d'arabe et de berbère*. ». Cette diversité a contribué à avoir un paysage linguistique à la fois complexe et riche en Algérie comme l'avance (Taleb Ibrahimi, 2006) «*le paysage linguistique de l'Algérie, produit de son histoire et de sa*

[3]La translitération arabe est présentée dans (Habash et al., 2007). La transcription phonologique est présentée entre /…/ mais utilise les formes HSB (les schèmes Habash-Soudi-Buckwalter) des consonnes quand c'est possible afin de minimiser la confusion que peut engendrer les différents ensembles de symboles utilisés.

géographie, est caractérisé par la coexistence de plusieurs variétés langagières – du substrat berbère aux différentes langues étrangères qui l'ont plus ou moins marquée en passant par la langue arabe, vecteur de l'islamisation et de l'arabisation de l'Afrique du Nord.». De ce fait, le dialecte algérien ne peut pas être présenté comme un système linguistique homogène, mais il possède de multiples variétés linguistiques. Selon (Queffélec et al., 2002) nous distinguons quatre variétés linguistiques pour le dialecte algérien :

i. *L'Oranais :* cette variété est parlée dans l'ouest de l'Algérie, précisément depuis la frontière algéro-marocaine jusqu'aux limites de la ville de Ténès,

ii. *L'Algérois :* cette variété est largement répandue dans la zone centrale de l'Algérie jusqu'à Bejaia,

iii. *Le rural :* les locuteurs de cette variété sont situés dans l'est de l'Algérie comme Constantine, Annaba ou Sétif. Nous signalons aussi que les locuteurs situés plus à l'est, c'est-à-dire de Constantine à la frontière algéro-tunisienne, sont aussi considérés dans cette catégorie. Il est aussi à signaler qu'il existe des déclinaisons de cette variante propre à certaines villes, comme c'est le cas pour les villes d'Annaba et de Constantine.

iv. *Le Saharien :* est considéré comme le dialecte la population algérienne habitant la partie sud de l'Algérie, à partir de l'Atlas saharien.

Par ailleurs, nous signalons aussi que le dialecte AA est enrichi par les langues des groupes ayant colonisé ou géré la population algérienne au cours de l'histoire du pays. Parmi les langues de ces groupes, nous citons : le turc, l'espagnol, l'italien et plus récemment le français. Nous pouvons considérer de ce fait le dialecte AA comme une fertilisation croisée de nombreuses langues avec l'arabe du fait de l'histoire de l'Algérie, qui a fait de cette dernière un carrefour de multiples civilisations et une terre d'accueil. Le métissage linguistique qui a résulté de ce brassage des langues (Arabe, Berbère, Phénicien, Andalou, Mudéjar, Romain, Espagnol, Turc, Sicilien, Français, etc.), depuis des siècles, a donné lieu à une grande palette de variétés pour le dialecte Algérien. Cette palette prend des couleurs régionales, provinciales voir même locales. Ces variétés sont matérialisées par la présence de mots étrangers dans le dialecte et de systèmes de prononciation différents variant sensiblement d'une région à une autre. En plus des mots d'emprunt et l'intégration de certains d'entre eux dans la morphophonologie du dialecte algérien, l'influence des langues sur le AA a été matérialisée également par l'alternance codique *(le code switching)* souvent dans les conversations quotidiennes, en particulier du français, par exemple, 'lycée', 'salon', 'quartier', 'normal', etc. L'utilisation de ces mots est réalisée sans aucune adaptation de la phonologie.

Ceci crée une situation linguistique assez complexe. En effet, ce mélange de la langue a été étudié par de nombreux sociolinguistiques comme (Morsly, 1986; Ibrahimi, 1997; Benrabah, 1999; Arezki, 2008). Ils ont décrit le paysage linguistique de l'Algérie comme 'multilinguisme' ou 'poly-glossique' où plusieurs langues et variétés de langues coexistent. En d'autres termes, le dialecte AA présente le meilleur exemple d'une situation sociolinguistique complexe (Morsly, 1986).

Ce brassage de langues peut être expliqué d'un point de vue historique comme suit. D'abord, le berbère était la langue maternelle de la population du Maghreb en général et de l'Algérie en particulier avant la conquête islamique. La langue berbère est la langue maternelle d'une partie de la population algérienne. Le berbère intègre quelques mots arabes en raison des échanges commerciaux entre les populations locales d'Afrique du Nord et les arabes qui sont venus de l'Orient. L'arabisation des algériens a commencé avec les conquêtes islamiques qui ont introduit la langue arabe comme moyen de communication de base quelle que soit le domaine : la religion, l'économie, l'apprentissage, etc. Au XVIe siècle, les

Ottomans ont aidé l'Algérie contre l'invasion espagnole qui occupait les zones dans l'ouest de l'Algérie (Oran) (Guella, 2011). L'occupation espagnole, pendant trois siècles, a été la principale raison de l'existence de certains mots espagnols dans le dialecte algérien (ALG), et surtout dans l'ouest. Il était aussi le facteur de l'allégeance de l'Algérie à Ottoman Khalifa afin de déloger l'Espagne de l'ouest du pays. Par cette allégeance, Algérie est devenue une province ottomane où le turc est introduit dans différent domaine notamment dans l'administration, politique et des échanges économiques. L'arabe a continué à utiliser, mais progressivement, de nombreux mots turcs ont été introduits dans de nombreux domaines de la vie quotidienne, comme la nourriture, l'habillement, le commerce, etc. L'année 1830 marque le début de la colonisation française qui a tenté d'imposer le français comme l'unique moyen de communication pendant 132 années. Cette situation a provoqué une baisse significative de la langue arabe au détriment du dialecte, caractérisé par une grande influence du français et de l'introduction de certaines autres langues comme l'italien et l'espagnol en raison des flux migratoires en provenance de l'Europe, principalement d'Italie (installé dans la t côte Est) et d'Espagne (installé à l'ouest), en plus bien évidemment de la France.

2.5. Comparaison entre l'arabe algérien, égyptien, tunisien et le MSA

Dans cette section, nous présentons un ensemble de différences entre les dialectes suivants : l'arabe algérien (AA), l'arabe égyptien (EA), l'arabe tunisien (TN) et l'arabe standard moderne (MSA). Les différences entre ces dialectes sont nombreuses mais celles mises en exergue dans cette section concernent les niveaux phonologique, morphologique, orthographique et lexical. Cette présentation est basée sur les travaux effectués dans (Habash, 2010) et (Zribi et al., 2014) et (Saâdane et Habash, 2015). Nous renvoyons le lecteur à ces travaux pour d'avantage éléments sur la comparaison effectuée.

2.5.1 Variations phonologiques

Dans la liste ci-dessous, nous introduisons les principales différences phonologiques entre AA et les variétés EA, TA et MSA :

- La consonne (ق) /q/ en MSA est l'un des sons qui méritent une attention particulière. Ce son a de nombreuses variétés de prononciation dans le dialecte algérien. Ces variations peuvent être perçue entre les régions, les villes, et même entre les localités de l'Algérie. Ainsi, la prononciation du "q" de l'arabe standard peut être réalisée en tant que [q, g, ʔ , ou k] dans les dialectes arabes. Ces différentes prononciations sont décrites comme suit :
    - *uvulaire sourde* «ق » [q] : comme au Maroc et en Tunisie, cette prononciation est présente en AA dans différentes localités à l'instar de certaines villes urbaines comme Alger ou Constantine. Toutefois, nous signalons que cette prononciation et pratiquement inexistante en EA sauf pour certaines exceptions comme le mot القاهرة *qaAhra* 'le Caire'.
    - *palatale sonore* « ڤ » [g] : ce son est également utilisé à la fois au Maroc et dans les dialectes tunisiens tout comme pour le dialecte algérien. L'utilisation de ce son en Algérie est limitée dans certaines villes comme Annaba et Sétif, en plus des zones rurales (dialectes bédouins) où ce son est très répandu. Ce son est également présent dans le dialecte égyptien afin d'exprimer la consonne ج j du MSA, et dans certaines provinces comme la Haute-Egypte (صعيد مصر) pour prononcer la consonne q ق.
    - *glottale sourde* /ʔ/ : la présence de ce son est limitée à la seule ville de Tlemcen en Algérie contrairement en Egypte où il est très utilisé. Pour le reste des dialectes et la majorité de l'algérien, ce son est inexistant.

- *k post-palatal* : ce son est une particularité du dialecte AA que nous ne trouvons pas dans les autres dialectes d'Afrique du Nord. Ce son est utilisé dans les localités rurales et certaines villes comme la Kabylie, Jijel, Msirda et Trara.

En plus de ces types de sons, il existe quelques exceptions de prononciations ne pouvant pas être casées dans l'une des catégories citées ci-dessus. C'est le cas des mots où la prononciation est toujours la même quel que soit le dialecte n'utilisant pas la glottale sourde /?/, comme pour le mot بقْرة bagraћ 'vache' qui se prononce de la même manière en utilisant la consonne palatale sonore bagra. Nous avons aussi quelques cas où la prononciation crée des paires minimales surtout dans les dialectes urbains, par exemple : قرون qruwn /qru:n/ 'siècles' et /gru: n/ 'cornes'. Le phonème non standard /g/ est également utilisé dans de nombreux mots dialectaux qui ne disposent pas d'équivalent en MSA, à titre d'exemple بالقدا biAlqdA /bilgda:/ 'très bien'.

- Il y aussi la prononciation de la consonne (ج) /j/ qui possède différentes formes spécifiques à une localité ou à un groupe de parlers, surtout en Afrique du nord. Cette consonne est prononcée [dj] à Alger et dans la plus grande partie du centre de l'Algérie comme dans le mot نجاح *ndjaH* 'succès', mais quand la consonne (ج) /j/ précède la consonne (د) /d/ elle est prononcée avec l'allophone [j] comme pour le mot جديد *jdid* 'nouveau'. En Égypte, cette consonne est prononcée comme /g/ palatale sonore. En Tunisie, Tlemcen et les habitants de l'est de l'Algérie, (ج) est réalisée en tant que /j/ ou /z/ lorsque le mot contient la consonne (س) /s/ ou (ز) /z/ comme dans les mots جبس *jibs* ou *djibs* 'plâtre' qui devient زبس *zebs*; et عجوز *ʕadjuwz* 'vieille femme' qui devient عزوز *ʕzuwz*.

- La consonne en MSA (غ) /ɣ/ est prononcée de manière différente selon certaines catégories de parlers. Dans l'est du Sahara algérien, comme M'sila et Bousaada, /ɣ/ est prononcée (ق) /q/, par exemple, les mots غالي *ɣaAliy* 'cher' et صغيرة *sɣayraћ* 'petite', sont prononcées respectivement */qaAliy/* et */sqayra/*. Parfois, elle est même prononcée (خ) /x/, comme pour les locuteurs tunisiens et ceux de l'est de l'Algérie qui prononcent, par exemple, le mot غسل 'lavé', */xssel/* ou */ɣssel/*.

- Il existe aussi d'autres prononciations qui consistent en l'assimilation de consonnes, comme pour la consonne (س) /s/ qui est assimilée à (ز) /Z/ et la consonne (ص) /S/, qui est assimilée à (س) /s/ dans certains dialectes, et prononcée (ز) /Z/ dans d'autres. C'est le cas du mot فازدة *faAzdaћ* 'corrompue' au lieu de فاسدة *faAsdaћ* et le mot سدر *sder* 'poitrine' au lieu de صدر *Sder*. Cette assimilation peut être expliquée par certaines causes phonétiques comme l'influence de la consonne voisine. D'un point de vue géographique, nous trouvons cette assimilation de consonnes dans certaines villes d'Algérie comme Tlemcen et à Annaba et la Tunisie.

- La consonne interdentale en MSA (ث) /θ/ peut être prononcée (ت) /t/, dans les trois dialectes AA, TA et EA comme pour le mot ثوم *θuwm* 'ail' qui est prononcé توم /tuwm/. Cette consonne est également prononcé /θ/ dans certains dialectes algériens et tunisiens urbains comme dans le mot ثوم *θuwm*. Elle est aussi prononcée (ف) /f/ comme dans les dialectes nomades de Mostaganem où par exemple le mot ثاني *θaAniy* 'également' est prononcé فاني *faAniy*; ou (س) /s/ dans certains cas dans le dialecte EA, par exemple, le mot ثابت *θaAbit* 'fixe' est prononcé سابت *saabit*.

- Une autre consonne interdentale en MSA a également des prononciations spéciales; il s'agit de la consonne (ذ) /ð/. Dans le dialecte EA, elle peut être prononcée (د) /d/, comme le mot ذهب *ðhab* 'or' qui est prononcé دهب *dhab*, ou (ز) /z/, par exemple le mot ذكي 'intelligent' est prononcé *zakiy*. Toutefois, dans le dialecte AA et TA, la consonne (ذ) /ð/ a l'une des prononciations suivantes: (ذ) /ð/ ou (د) /d/. Par exemple le mot ذراع 'bras' peut être prononcé *ðraAç* ou *draAç*. En outre, dans certaines régions en Algérie, comme Mostaganem, cette consonne est prononcée (ب) /v/, comme pour le mot ذهب *ðhab* 'or' est prononcé dans ces régions ڤهب *vhab*.

- Le phonème de la glottale sourde, qui apparaît dans de nombreux mots en MSA, possède dans le dialecte AA les différentes formes de prononciation suivantes :
    - *la glottale sourde devient longue :* cette prononciation est également présente dans les autres dialectes TA et EA. Nous pouvons donner comme exemple les mots : فأس *faÂs* /fa's/ → /fa:s/ فاس *faAs* 'pioche', ذئب *Diŷb* /Di'b/ → /Di:b/ ذيب *diyb* 'loup', et مؤمن *muŵmin* /mu'men/ → /mumin/ مومن *muwmin* 'croyant'.
    - *la disparition de la glottale sourde :* elle consiste à retirer simplement la glotte en prononçant le mot. Cette forme est également utilisée dans les dialectes TA et EA. Par exemple, prenons le mot suivant : زرقاء *zarqaA'* /zarqa:'/ → /zarqa:/ زرقا *zarqA* 'bleu'.
    - *la glottale sourde est remplacée par une semi-voyelle /w/ ou /y/ :* cette prononciation est présente dans les dialectes AA et TA et non pas dans le dialecte EA. Elle est utilisée par exemple dans le cas des mots أكّل */Âak~al/* 'faire manger' → وكّل *wuk~al*, أمس */Âams/* 'hier' → يامس *yaAmas*.
    - *la glottale sourde est remplacée par la lettre /l/ :* cette forme est également utilisée uniquement dans les dialectes AA et TA contrairement à leur homologue égyptien EA. C'est le cas des exemples suivants : أفعى */Âafça/* 'vipère' → لفعى */lafça/*, أرض */ÂaarD/* 'terre' → لرض */larD/*. Nous notons que les exemples donnés sont également des exceptions qui possèdent la même forme à la fois dans le cas défini et indéfini.
    - *la glottale sourde est remplacée par la lettre /h/ :* les dialectes AA et TA utilisent cette forme pour prononcer dans certaines cas la glottale sourde, par exemple dans le mot أجّالة *Âaj~aAlaħ* /Âajja:la/ 'veuve' → هجّالة *hajjaAlaħ* /hajja:la/, أمّالا *Âam~aAlaA* /Âamma:laA/ 'cependant' → همّالا *ham~aAlaA* /hamma:laA/. Cette forme de prononciation est inexistante dans le dialecte EA.

- Le dialecte AA comme la plupart des autres dialectes arabes, change et néglige les voyelles courtes, surtout quand elles sont placées à la fin d'une syllabe. Par exemple, le mot بابٌ *baAb-un* 'la porte' est transformé en بابْ *baAb* /ba:b/ en dialecte. Nous signalons, qu'en règle générale, la suppression de la première voyelle change la structure syllabique des unités lexicales, qui tendent à devenir pour certains mots monosyllabiques.

- Contrairement au dialecte égyptien, le dialecte algérien et tunisien élident de nombreuses voyelles courtes dans des contextes non stressées. Cette caractéristique est également présente dans les autres dialectes du Maghreb. C'est le cas des mots suivants : MSA جمل *jamal* 'Camel' (et EA */gamal/*) devient en AA */jmal/*. En outre, cette caractéristique introduit un élément intéressant pour distinguer les dialectes maghrébins du dialecte EA. Il s'agit de la présence d'une succession de deux consonnes au début du mot. Ceci se traduit par une particularité notable dans le schème verbal des dialectes AA et TA *'fçal'* à la place de *'façal'* dans le dialecte EA,

comme dans le verbe en MSA قتل */qatal/* 'il a tué' (et EA */'atal/*) devient en AA et TA /qtal/.

- Commencer des mots par des consonnes 'neutres', sans voyelles (avec un sukun) est l'une des caractéristiques marquantes de l'arabe dialectal maghrébin et qui le distingue à la fois du littéral et des dialectes orientaux. Par exemple en dialecte nous avons le mot كْتَب *ktab* (il a écrit) au lieu de كَتَبَ *kataba* en MSA. Cette particularité est particulièrement remarquable au niveau des prénoms comme بْرَاهِيم *brahim* au lieu de اِبْرَاهِيم *Ibrahim*; سْلِيمَان *slimân* au lieu de سُلَيمَان *Sulayman*.

- Les diphtongues *ay* et *aw* utilisées en MSA sont généralement réduites uniformément dans les dialectes à /i:/ et /u:/ respectivement. Par exemple, prenons les mots : حيط */hayT/* 'mur' qui devient en dialecte */hi:T/*, لون */lawn/* 'couleur' qui devient en dialecte */lu:n/*. Nous notons aussi que cette particularité se trouve chez la jeune génération des parlers; cependant, les locuteurs les plus âgés et les parlers ruraux conservent encore les diphtongues *ay* et *aw* dans certains mots et contextes, par exemple le mot عود est encore prononcé */ɕawd/* 'cheval' par certains vieux parlers.

- Les dialectes AA et TA sont aussi caractérisés par la prononciation, dans certains mots, de la voyelle longue /a:/ du MSA comme /e:/ et dans d'autres mots comme /a:/. Par exemple, le mot جَمَالْ */jam:al/* 'beauté' avec cette signification est prononcé avec /a:/ mais il est réalisé avec /e:/ dans le mot */jme:l/* signifiant 'chameaux'.

- Les dialectes AA et TA utilisent la particule 'n' pour la première personne du singulier comme les autres dialectes du Maghreb. Cette particule est généralement absente dans les dialectes du Machrek comme le dialecte EA. Dans ces dialectes la particule 'n' est remplacé par le 'a' comme le montre l'exemple suivant : نكتب */naktab/* 'J'écris' dans le dialecte AA est réalisé en EA comme اكتب */Aaktib/*.

2.5.2 Variations Morphologiques

Sur le plan morphologique, il existe plusieurs différences entre les dialectes, et principalement le dialecte AA, et le MSA au niveau de plusieurs aspects. Il est à noter aussi que les dialectes maghrébins possèdent en général des aspects morphologiques assez proches et qui consistent essentiellement en une simplification de certaines inflexions et l'inclusion de nouveaux clitiques comme suit :

- En termes d'inflexion dans le dialecte AA, comme les autres dialectes arabes, les cas des terminaisons dans les noms et les modes des verbes sont perdus. Nous notons que l'indicatif est utilisé par défaut, contrairement aux autres modes qui ne sont pas utilisés. En outre, le duel (masculin et féminin) et le pluriel féminin sont disparus; ils sont assimilés au masculin pluriel. Par exemple, le mot شَكَرْتُنَّ *šakartun~a* 'vous (féminin au pluriel) avez remercié' est normalisé dans le dialecte AA en شْكَرْتُوا *škar-tuwaA* 'vous avez remercié'. En outre, la première et la deuxième personne du singulier sont conjuguées de la même manière dans le dialecte, par exemple, dans le MSA nous disons شَكَرْتُ *šakartu* 'J'ai remercié' et شَكَرْتَ *šakarta* 'tu as remercié', ces deux formes sont normalisées en dialecte AA et TA dans la forme unique suivante : شْكَرْتْ *škart* 'j'ai/ tu as remercié' et en dialecte EA شَكَرْتْ *šakart*. Cette simplification peut conduire, de ce fait, à des ambiguïtés dans les dialectes.

- Le dialecte AA modifie la forme interne des verbes quand il fait sa flexion sous la forme imparfaite et impérative. Il introduit la gémination dans la première lettre et le déplacement de la voyelle de la seconde consonne du radical vers la première consonne du même radical. Cette modification est appliquée seulement pour former le pluriel et la 2ème personne du singulier au féminin. Pour illustrer cet aspect, la flexion en AA du verbe 'remercier' à la 3ème personne du singulier au masculin est يُشْكُرْ *yu-škur* 'il remercie' et pour la 3ème personne du pluriel au masculin nous avons يُشُّكْرُوا *yuš~ukr-uwA* 'ils remercient', cependant, en dialecte EA et TA le même cas est formulé en يُشْكُرُوا *yuškur-uwA*. Ce dernier exemple montre bien l'absence de la gémination dans les autres dialectes, ce qui fait d'elle un aspect propre à l'algérien.

- Le dialecte AA utilise seulement, comme les autres dialectes arabes, le suffixe ين /yn/ pour former le pluriel régulier. Cependant, les dialectes AA et TA élident les voyelles courtes dans des formes plurielles, comme dans les exemples suivants : مُلْحَدْ *mulHad* 'incroyant', au pluriel مُلْحْدِين *mulHdiyn*, مُهَنْدِسْ *muhandis* 'ingénieur', pl. مُهَنْدْسِينْ *muhandsiyn*. Mais il existe une exception pour le participe actif [Faa3iL] → [Faa3L-iyn] où l'élision au niveau de cette exception est maintenu quel que soit le dialecte comme pour le mot صَايِمْ *SaAyim* 'fasting' → صَايْمِينْ *SaAymiyn*.

- Le suffixe emphatique تيك /-tiyk/, décrit par (Cohen, 1912), en tant que caractéristique du dialecte d'Alger qui est utilisé pour exprimer les adverbes se terminant par /-a/, comme pour les mots قانا *gana* 'également' qui devient *ganaAtiyk*, زعما *zaçma* 'soi-disant' qui devient *zaçmaAtiyk*.

- Pour la forme استفعل "Aistaf3al", qui existe dans les différents dialectes, le dialecte AA introduit une nouvelle variante de cette forme. Cette variante est سفعل '*ssa-f3al*' et elle est employée essentiellement dans les parlers de l'ouest algérien. A ce sujet (Marçais, 1902) indique la réduction de la séquence [st] classique à [ss] que nous entendons fréquemment en un seul /s/. Par exemple, prenons le verbe اِسْتَكْلَفْ *Aistaklaf* 's'occuper de' peut également être utilisé comme سَّكْلَفْ *ssaklaf* ou سَكْلَفْ *saklaf*.

- Une autre caractéristique du dialecte AA, inexistante dans le système morphologique du MSA et généralement présente dans les dialectes, consiste en l'insertion de la voyelle /i:/ entre la racine et les suffixes consonantiques de la forme perfective du verbe géminé primaire. Cette caractéristique traduit l'écart entre l'arabe dialectal et l'arabe standard au niveau de la suffixation consonantique à la forme perfective (accomplie) du verbe géminé. Par exemple, dans le MSA, le verbe شدّ/شددت *šad~a/šadadtu* 'il/j'ai tiré' devient dans le dialecte AA شدّ/شديت *šad~/šad~iyt*. Cette caractéristique est également présente dans les autres dialectes arabes, comme indiqué précédemment, mais avec quelques modifications, comme c'est le cas avec le dialecte EA où l'insertion de la voyelle /ee/ est effectuée à la place de la voyelle /i:/.

- La voix passive existe aussi dans la variété dialectale mais avec quelques différences significatives par rapport à cette même voix dans le MSA. En MSA, la voix passive est le résultat d'un changement interne des voyelles du verbe, tandis qu'en dialecte, cette voix est ainsi formée par l'introduction de nouveaux morphèmes, généralement le [t-] et parfois, dans les dialectes AA et EA, le morphème [n-]. Ces morphèmes ajoutés sont préfixés à la forme perfective et infixés à la forme imperfective. Par exemple, le dialecte tunisien marque la voix passive du verbe exprimé en MSA par كُتِب

kutiba ‘il a été écrit’, par تِكتِب *tiktib*. Plus en détails, la forme passive dans le dialecte algérien est obtenue en faisant précéder le verbe avec l’un des éléments suivants:

- t- / tt-, par exemple : تبنى *tabnaý* ‘il a été construit’, ترفد *ttarfad* ‘il a été relevé’
- n-, par exemple : نفتح *nftah* ‘il a été ouvert’
- /tn- / ou /nt/, e.g., نتكل *ntkal* ‘il a été mangé’, تنقتل *tnaqtal* ‘il a été tué’. Nous notons que ce dernier élément est spécifique pour le dialecte AA.

- Plusieurs dialectes introduisent de nouveaux clitiques qui n’existent pas dans le MSA, comme la négation circonfixe ما + *mA*+ +ش +*š* qui est exprimée en MSA avec diverses particules comme : ما *mA*, لم *lam*, لن *lan* ‘ne … pas’. Par exemple ما قريتش *mA qriyteš* ‘je n’ai pas lu’. Un autre exemple spécifique au dialecte TA est le clitique d’interrogation verbale qui est exprimé en MSA par أ *Âa* et la particule هل *hal*. Ces clitiques sont substitués en TA par le clitique شي *šiy*.

- A l’instar de plusieurs dialectes (EA et TA), le dialecte AA comprend un ensemble de clitiques qui sont des formes réduites des mots MSA. A ce titre, le proclitique démonstrative +ه *ha*+ qui précède strictement l’article défini +ال *Al*+ utilisé en dialecte correspondent aux pronoms démonstratifs du MSA هذا *haðaA* et هذه *haðihi*, par exemple la phrase هذه الدنيا *haðihi AldunyaA* est exprimée en dialecte par *haAldinyaA* ‘cette vie’. Il y aussi le proclitique +ع *ça*+ utilisé dans les dialectes qui est une forme réduite de la préposition على */çalaý/* ‘sur’, comme dans l’exemple suivant : la phrase en MSA على الطاولة */çalaý AlTaAwilaħ/* est formulée en dialecte AA عالمايدة *çaAlmaAydaħ çaAlmaAydaħ* ‘sur la table’. La même remarque est valable pour les proclitiques + ف +*fa* et + م *m*+; qui sont la forme réduite des prépositions في *fiy* ‘dans’ et من *min* ‘de’ ou de la conjonction de la coordination مع *maça* ‘avec’ respectivement. Par exemple (MSA → AA) في الدار *fiy AldaAr* → فالدار *fiAldaAr* ‘dans la maison’, et la phrase en MSA من المدرسة *mina Almadrasaħ* donne dans le dialecte AA مالمسيد *miAlmsiyd* ‘de l’école’.

- Le dialecte AA a perdu en général les formes duelles nominales, qui sont remplacées par le mot *zudwj* /zu:dj/ ‘deux’ suivi du nom au pluriel. Par exemple, la forme duelle كتابين *kitaAbayn* en MSA est exprimée par la forme زوج كتب *zuwdj ktub* ‘deux livres’ en dialecte AA. Les dialectes tunisien et marocain utilisent le même procédé pour exprimer le duel avec l’utilisation du mot زوز *zuwz* /zu:z/ et جوج *juwj* /ju:j/ respectivement.

2.5.3 Variations Orthographiques

La variation orthographique dans l’écriture des mots en dialectes arabes est dû principalement à deux raisons: i) la non-existence d’une norme orthographique pour les dialectes arabes ces derniers ne sont pas codifiées et normalisées, et ii) les différences phonologiques entre le MSA et les dialectes arabes en générale, voir même au sein d’un même dialecte. Pour ces dialectes les mots peuvent être écrits phonétiquement ou étymologiquement en utilisant leurs formes correspondantes en MSA. Ce fait crée une certaine incohérence entre les écrivains des dialectes. Par exemple, le mot correspondant à ‘or’ peut être écrit دهب *dhab* ou ذهب *Dhab*. En outre, dans certains cas, la phonologie ou la morphologie sous-jacente se traduit par une écriture d’assimilation phonologique régulière, par exemple, طوموبيل *Tuwmuwbiyl* ‘voiture’ est aussi écrite طونوبيل *Tuwnuwbiyl*, إسماعيل *IsmaA’iyl* ‘Ismaël’ est aussi écrit إسماعين *IsmaA’iyn*, من بعد *men ba3d* ‘après’ est également écrit مم بعد *mem ba3d*.

De plus, ces différentes orthographes peuvent conduire à une certaine confusion sémantique, comme pour le mot شربو *šrbw* qui peut être شربوا *šarbuwA* ‘ils buvaient’ ou شربه

šarbuh ‘il l’a bu’. Enfin, les voyelles longues raccourcies peuvent être prononcées longues ou courtes. A titre d’exemple, شافوها/شفوها *šAfw+hA/ šfw+hA* ‘ils l’ont vu’, et مجابش *majaAbaš* ‘il n’a pas apporté’ qui peut être prononcé aussi ماجابش *mAjaAbaš*. Le dernier exemple est particulier où la particule ما *mA* en MSA, qui est la source du proclitique *ma-*, possède une autre orthographe en dialecte comme suit : ما جابش *mA jaAbaš* (en d’autres termes deux mots distincts). (Zribi et al, 2014) précise que pour le dialecte tunisien un certain nombre d’adverbes possèdent de multiples formes, par exemple, l’adverbe interrogatif آش *Āš* ‘quoi’ apparaît parfois comme un proclitique +ش *+š* et dans certains cas il est transcrit comme un mot séparé reflétant différentes prononciations, par exemple شقال *šqaAl* et آش قال *ĀšqaAl*.

2.5.4 Variations lexicales

Au niveau de la variation lexicale, il existe plusieurs aspects caractérisant les dialectes. Nous avons choisi de focaliser la présente section sur la présentation de deux aspects très répandus dans les dialectes, à savoir la dérivation et l’emprunt.

2.5.4.1. *La dérivation*

La dérivation dans la grammaire arabe est un phénomène régulier et utilisé pour construire à partir d’une racine consonantique plusieurs éléments et paradigmes exprimant l’agent, le patient, le locatif, les noms prédicatifs (masdar), le superlatif, etc. Cette construction ou dérivation est faite en suivant des schèmes préétablis avec l’implication d’une variation vocalique et l’ajout de certains éléments consonantiques. Pour les dialectes, la régularité de la dérivation constitue la colonne vertébrale du système morphologique dialectal. Selon (Mejri et al., 2009), la dérivation est néanmoins enrichie dans les dialectes par une présence relativement importante du système affixal qui concerne également la forme littérale moderne. Cet enrichissement est continu et est matérialisé, à titre indicatif, par l’incorporation dérivationnelle (Sfar, 2005 & 2006) ou l’ajout d’un certain nombre d’affixes spécifiques comme جي *jiy* qui indique la profession (Baccouche, 1994) : قهواجي *qahwaAjiy* ‘celui qui tient un café’, بنكاجي *bankaAjiy* ‘banquier’. Par conséquent la dérivation au niveau des dialectes diffère de celle de l’arabe standard au niveau quantitatif. De plus, nous notons que dans les dialectes, un autre type de dérivation est utilisé, non basé sur des schèmes spécifiques mais plutôt en combinant les schèmes aux affixes. C’est le cas par exemple du mot كوارجي *kawwarjiy* ‘footballeur’ qui ajoute au schème [Fa33aL], qui donne à partir de كورة *kuwra* ‘ballon’ le mot كوار *kawwaAr* le suffixe جي *jiy* utilisé pour exprimer une profession, ou du mot حيطيست *HiTist* qui incarne l’ajout du suffixe يست *ist*, emprunté du français pour exprimer une profession, afin de qualifier une personne dont la profession est d’adosser les murs (une manière ironique pour dire chômeur).

2.5.4.2. *L’emprunt*

L’emprunt est aussi une autre caractéristique lexicale fortement présente dans les dialectes arabes. D’un point de vue qualitatif et quantitatif, l’emprunt présente un dynamisme assez intéressant. Par ailleurs, l’emprunt est le reflet de l’influence des autres langues sur les dialectes, pour toutes les raisons citées auparavant, où dans les dialectes nous trouvons beaucoup de mots issus des différentes langues comme l’anglais, le français, le turc, l’espagnol, etc. Sur le plan qualitatif, (Mejri et al., 2009) avance qu’il existe trois points à retenir : l’introduction de nouveaux suffixes empruntés à d’autres langues, l’intégration systématique des unités empruntées dans les paradigmes construits par schèmes et l’impact phonologique qui agit par le bais de l’emprunt sur le système phonologique du dialecte.

En ce qui concerne l’introduction de nouveau suffixe, ces derniers sont issus des autres langues, comme le turc ou le français, afin d’exprimer certains paradigmes comme une profession. C’est le cas du suffixe turque جي *jiy ou* français يست *ist* décrits dans la section

précédente. Quant à l'intégration des emprunts par le biais des schèmes, nous signalons qu'à partir d'un mot emprunté, nous pouvons obtenir toutes les unités répondant à tous les schèmes disponibles en dialectal. Par conséquent, cette particularité reflète une grande capacité à la fois d'intégration et de création lexicale. Par exemple, à partir du mot emprunté 'business' en dialecte tunisien, et maghrébins en général, nous obtenons les mots suivants :

- Le verbe بزنس *baznas* 'il a fait un biseness'
- L'agent بزناس *baznaAs* 'celui qui fait du biseness' avec un pluriel بزناسة *baznaAsa*
- Le Masdar تبزنيس *tbazniys* 'action de faire des biseness'

Pour ce qui est de l'impact phonologique des emprunts sur les dialectes, nous citons par exemple l'introduction de voyelles nasales dans le dialecte maghrébin. Cet impact est matérialisé par la coexistence d'une nasalisation de la voyelle doublée et d'une présence assez timide de la consonne [n], comme c'est le cas pour le mot *ɛlɛktisyɛ* 'électricien'.

Nous terminons cette section par donner, dans le tableau (2.2), quelques exemples d'emprunts de mots, de différentes origines (berbère, turc, italien, espagnol et français), dans le dialecte algérien AA.

Mots	**Traduction**	**Translittération**	**Origine**
فكرون	tortue	*Fakruwn*	Berbère
شلاغم	Moustache	*šliAyam*	
قرجومة	gorge	*Qarjuwmaħ*	
تقاشير	Chaussettes	*tqaAšiyr*	Turc
سكارجي	Ivrogne	*sukaArjiy*	
زردة	Festin	*Zardaħ*	
فيشطة	Fête	*fiyšTaħ*	Italien
زبلة	Faute	*Zablaħ*	
صوردي	Money	*Suwrdiy*	
سيمانة	Semaine	*siymaAnaħ*	Espagnol
سبردينة	Espadrille	*Spardiynaħ*	
سُكويلة	Ecole	*Sukwiylaħ*	
طابلة	Table	*TaAblaħ*	Français
تيليفون	Téléphone	*Tiyliyfuwn*	
فرملي	infirmier	*Farmliy*	

Tableau 2. 2. Origine et sens de quelques mots empruntés utilisés dans le dialecte algérien

2.5.5 La variation syntaxique

Dans cette section, nous essayons de présenter l'écart entre l'arabe MSA et les dialectes au niveau syntaxique où à ce niveau la rupture avec le littéral est plus marquante et grande.

Nous rappelons que les dialectes arabes se caractérisent par la disparition des marqueurs flexionnels : les cas nominatif, accusatif et génitif pour les noms, ainsi que la perte de la distinction entre l'indicatif, le subjonctif et le jussif (impératif) pour les verbes. Cette perte pose un problème pour définir les fonctions syntaxiques des unités lexicales dans une phrase donnée. Ce fait que nous constatons est renforcé par les propos de (Merji, 2009) où il avance que « *Si le littéral relève des langues casuelles, qui, grâce aux flexions, marque les fonctions syntaxiques des unités lexicales dans le cadre de la phrase, il n'en est pas de même du dialectal qui substitue au marquage casuel une rigidité très contrainte dans l'ordre des mots et qui compense dans certains cas la disparition des formes fléchies par un recours plus important aux éléments prépositionnels* ». Afin d'illustrer ces propos, prenons l'exemple de la

phrase suivante :

- ضرب الرجل الطّفل (l'homme frappe l'enfant)
  Daraba ?al-rajul ?at-tifl
  frapper-[accompli]-l'homme-l'enfant

Dans la grammaire arabe, cette phrase donne lieu à deux interprétations différentes selon les marques casuelles, comme suit :

1. ضربَ الرجلُ الطّفلَ (l'homme a frappé l'enfant)
   daraba ?al-rajul-u ?at-tifl-a
   frapper-[accompli]-l'homme-[nominatif]-l'enfant-[accusatif])

2. ضرب الرجلَ الطّفلُ (l'enfant a frappé l'homme)
   daraba ?al-rajul-a ?at-tifl-u
   frapper-[accompli]-l'homme-[accusatif]-l'enfant -[nominatif]

Au niveau des contraintes liées à l'ordre des mots, le dialecte partage avec le français les mêmes contraintes. Cet ordre est réalisé de deux manières différentes selon le type de la phrase, comme suit :

- ***Verbales :***
  - ضربْ الراجلْ الطّفلْ → l'homme a frappé l'enfant
  - ضربْ الطّفلْ الراجلْ → l'enfant a frappé l'homme

- ***Nominales :***
  - الطّفل ضرب الجار → l'enfant a frappé le voisin
  - الجار ضرب الطّفل → le voisin a frappé l'enfant

Dans certaines phrases dans le dialecte, nous faisons recours à la préposition [fi] «dans» afin de marquer l'accusatif qui ne peut pas être marqué seulement par la position du mot dans la phrase. Pour illustrer ce cas, prenons cette phrase en MSA :

- MSA
  يأكل الطّفلُ الطماطمَ (l'enfant mange la tomate)
  ya?kulu ?at-tifl-u ?aT-TamaATim-a
  manger-inaccompli-3ème personne singulier- le-enfant-la-tomate

- Dialecte AA
  ياكل الطّفلْ في الطماطمَ (l'enfant mange la tomate)
  yaAkul t-tful T-TmaATam
  manger-inaccompli-3ème personne singulier- le-enfant-dans-la-tomate

L'accord entre le verbe et le sujet en fonction de la position du verbe dans la phrase constitue un autre écart entre le MSA et l'arabe dialectal. En MSA, nous avons deux types d'accords entre le verbe et le sujet : total et partiel. Cependant, en arabe dialectal, il existe seulement un accord complet quel que soit la position du verbe. Par exemple, pour la phrase : كتب الاولاد الدروس 'les enfants écrivent les leçons' nous avons les présentations suivantes :

- MSA

1) Verbe Sujet Objet (***accord Partiel***)
   كتب الاولاد الدروس
   kataba ?al-AwlaAd-u ?ad-duruws-a
   écrit$_{\textbf{mascSing}}$ les enfants les leçons
2) Sujet Verbe Objet (***accord Complet***)

الاولاد كتبوا الدروس
?al-AwlaAd-u katabuwA ?ad-duruws-a
Les enfants **écrivent$_{\textbf{mascPlural}}$** les leçons

- EGY

1) Verbe Sujet Objet
كتبو الاولاد الدروس
katabuw ?il-AwlaAd ?id-duruws
écrivent$_{\textbf{mascPluriel}}$ les enfants les leçons
2) Sujet Verbe Objet
الاولاد كتبو الدروس
?il-AwlaAd katabuw ?id-duruws
Les enfants **écrivent$_{\textbf{mascPluriel}}$** les leçons

Dans le même registre, la construction possessive إضافة 'Idafa' est une autre différence notable entre le dialecte et le MSA à signaler. Cette construction est réalisée dans le dialecte avec l'utilisation d'une particule entre le premier et le deuxième mot. Cette particule diffère largement entre les dialectes. Quant au MSA, la construction possessive est faite grâce à l'article défini attaché au deuxième mot. Prenons l'exemple suivant :

- ***MSA :*** Nom1 **de** Nom2
ملك المغرب (le roi du Maroc)
Malik ?al-maghrib
roi le-Maroc
- ***Dialecte :*** Nom1 **<particule>** Nom2
AA : الملك ديال المغرب
?al-malik **dyaAl** ?al-maghrib
Le-roi appartenant le-Maroc
- ***LEV :*** الملك تبع المغرب
?al-malik **taba'** ?al-maghrib

Enfin, nous signalons une dernière différence concernant la modification de la position de l'article démonstratif. En MSA, le pronom démonstratif est placé en première position avant le nom, contrairement aux dialectes où il est placé en deuxième position après le nom, comme l'illustre l'exemple suivant :

- ***MSA :*** هذا الرجل *haDaA ?ar-rajul* 'cet homme'
- ***AE :*** الراجل ده *?ir-raAguil dah* 'homme cet'
- ***AA :*** الراجل هادا *?ir-raAgil haAdaA* 'homme cet'

Partie II : Analyse Linguistique de la langue arabe

Chapitre 3 Analyse morphosyntaxique

Introduction

Ce chapitre est consacré à présenter les démarches suivies pour le développement de notre analyseur morpho-syntaxique dédié à l'arabe standard. Nous avons commencé par présenter un aperçu des travaux réalisés sur le traitement automatique de l'arabe dans la section 3.1. La section 3.2 est dédiée à présenter les démarches et les étapes effectués pour le développement de notre système d'analyse linguistique (proposé). Enfin, la section 3.3 est consacrée à présenter l'analyse syntaxique effectuée lors de cette analyse, tout en exposant les relations syntaxiques dans des phrases verbales et nominales.

3.1. Etat de l'art sur le traitement automatique de l'arabe

Les premières recherches sur le traitement automatique de l'arabe ont commencé vers les années 1970 (Cohen, 1970) et concernaient notamment le lexique et la morphologie.

Le traitement morphosyntaxique par ordinateur de la langue arabe n'est pas récent, il a fait l'objet depuis plusieurs décennies de travaux novateurs, en particulier en France par des équipes de recherche qui se sont progressivement spécialisées dans le traitement de l'information multilingue. En effet, dès le milieu des années 1970, les travaux de chercheurs tels que Yahya Hlal, puis ceux de Fathi Debili dans les années 1980, ont montré la possibilité d'un traitement automatique de la langue arabe. Dans les années 1990, on peut également citer aussi, toujours en France, les travaux de Joseph Dichy notamment dans le cadre du projet européen DIINAR-MBC (Dictionnaire informatisé de l'arabe, multilingue et basé sur corpus).

Plusieurs projets européens ont porté sur le traitement de l'arabe. Plus récemment un réseau d'excellence européen a permis de regrouper la plupart des acteurs européens pour échanger des informations et produire des ressources linguistiques (dictionnaires, corpus étiquetés, logiciels, dans le cadre des projets NEMLAR (Network for Euro-Mediterranean LAnguage Resources) puis MEDAR (MEDAR (Mediterranean Arabic Language and Speech Technology). La recherche sur le TAL arabe a été confrontée à de nombreuses difficultés qui relèvent de niveaux différents directement en lien avec notre sujet de thèse.

Tout d'abord, le niveau morphologique a posé des problèmes spécifiques en raison du système particulier de création lexicale et de dérivation de l'arabe standard. (Roman, 1999) démontre que le système syllabique de la langue arabe, constitué de sous-ensembles de consonnes et de sous-ensembles de voyelles permet l'attribution de fonctions différentes aux consonnes et aux voyelles dans la production langagière. Il démontre aussi que la langue arabe a construit son « système de nomination » sur des racines de consonnes et qu'elle a fondé son « système de communication » sur ses voyelles brèves, qui sont utilisées en fait comme des désinences casuelles. Dans cette étude de la morphologie de l'arabe, André ROMAN présente les oppositions entre les formes nominales et verbales de l'arabe dans la langue attestée par des textes de différentes époques mais n'aborde pas du tout les réalisations orales ou oralisées de l'arabe moderne.

Concernant la question de la dérivation qui constitue le cœur de son analyse, (Roman, 1999) oppose le couple {res - modus} dans la nomination et le couple {première voix – seconde voix} dans la communication, alors que les productions contemporaines observées notamment sur les forums et autres blogs montrent un mélange de ces deux systèmes couplés dans les réalisations langagières des locuteurs natifs de l'arabe.

Plusieurs chercheurs ont essayé d'apporter des solutions à cette spécificité morphologique pour le traitement automatique de l'arabe, mais c'est la question de l'écrit -et en particulier de la voyellation de l'écrit- qui a concentré l'essentiel des travaux. Ainsi, (Grainger, 2003) consacre une étude approfondie à « la reconnaissance du mot écrit en arabe », mais son « approche expérimentale » ne tient pas compte des productions effectives des locuteurs et part du système de la langue pour proposer une méthode de reconnaissance

théorique. De son côté, (Ghenima, 1998) consacre sa thèse de doctorat au problème de la voyellation, mais son analyse morphosyntaxique est loin de permettre une reconnaissance du mot écrit en arabe. La proposition de (Zaafrani, 1997) est plus convaincante parce qu'elle ne vise pas le « mot » mais les traits morphologiques de l'arabe,

Ensuite, concernant l'aspect sémantique de notre sujet, les études consacrées au traitement automatique de l'arabe ont été marquées au cours des deux dernières décennies par une concentration des travaux sur l'étude statistique du vocabulaire. Seul (Abbas-Mekki, 1998) a proposé une description des unités linguistiques en vue de l'indexation automatique, mais ses travaux ont porté exclusivement sur les textes écrits en arabe classique. Plus récemment, les synthèses proposées par (Abbès, 2002**)** et par (Abbès et Dichy, 2008) constituent une référence en matière de traitement statistique du vocabulaire arabe classique. Le premier a développé un fréquenceur (AraFreq) permettant le calcul de fréquences sur des formes dérivées ou non de l'arabe (lemmes) ; le second a utilisé le logiciel « AraConc » pour réaliser l'extraction automatique des fréquences à partir d'un corpus journalistique.

Il est clair cependant que la principale préoccupation des chercheurs durant cette période a été le développement d'outils permettant de constituer des bases de données lexicales, très recherchées pour l'arabe. (Ezzahid, 1996) avait proposé des pistes très intéressantes en se basant sur la théorie Sens-Texte d'Igor Mel'cuk. Suivant ces pistes, (Dichy, 1997) a fait l'inventaire des spécificateurs du mot en arabe et développé une base de données (DIINAR 1.0) enrichie de spécifications morphosyntaxiques, même si elle reste exclusivement axée sur l'arabe classique. À partir de cette base, il a été possible de mener des études locales concernant notamment les verbes en arabe classique en vue de l'enseignement (Abu Al-Chay, 1988) ou encore les verbes en arabe moderne en vue de la traduction (Franjié, 2003). Mais malgré cette diversification des objectifs, des problèmes de fond sont restés sans solution.

Enfin sur le plan méthodologique, le tournant intervient progressivement au cours des années 1990 grâce à un changement de perspective. En effet, on assiste à un passage des travaux théoriques sur la langue arabe en tant que système linguistique, aux travaux empiriques basés sur des corpus d'usages attestés et de productions effectives dans des situations réelles. Ce changement de perspective a été impulsé par l'intérêt suscité par l'Internet pour la recherche d'information, d'abord en mode monolingue arabe, puis en mode interlingue avec l'arabe. Les travaux de (Fluhr, 1997 & 1998) sont un signal fort de ce tournant, en particulier pour les études spécifiques consacrées au traitement « crosslingue » de l'arabe.

Dans la même optique, (Guidère, 2003 & 2005) propose un système de recherche d'information multilingue intégrant l'arabe et donne des recommandations précises pour la constitution de corpus arabes, l'alignement d'unités du discours, et l'élaboration d'ontologies en vue de la détection automatique d'entités nommées.

Les travaux de (Attia, 2000) sont parmi les premiers à adopter cette dimension empirique dans l'analyse morphologique à travers la proposition d'une approche hybride combinant à la fois règles de conjonction et statistiques, et propose de ce fait l'utilisation d'une liste de préfixes, suffixes et modèles pour la transformation d'une forme dérivée (stem) à une racine (root). Les combinaisons possibles entre préfixe-suffixe-modèle sont construites pour chaque mot afin d'en dériver les possibles racines. Ce système a été implémenté par le RDI[4] dans le développement du logiciel MORPHO3. Dans la ligné de l'analyseur MORPHO3, (Darwish, 2002) propose une approche de traitement plus automatique, implémentée dans «Sebawai », qui remplace le traitement manuelle, qui construit les règles et

[4] RDI : Research and Development International (Egypt)

les suffixes, par un traitement qui produit la racine de chaque mot par des règles dérivées automatiquement et statistiquement. Ce système comporte deux modules principaux : le premier utilise une liste de pairs en arabe (mot-racine) afin d'obtenir une liste des préfixes et suffixes, de construire des modèles de dérivation et de calculer l'apparition d'une vraisemblance à un préfixe, un suffixe, ou un modèle. Le second module prend en entrée les mots arabes, les tentatives de constructions possibles des combinaisons préfixe-suffixe-modèle, et renvoie en sortie une liste de classement des racines possibles.

D'autres méthodes ont été aussi proposées pour effectuer ces analyses comme l'*alignement des étiquettes morphologiques et syntaxiques.* (Lee, 2004) propose d'utiliser l'alignement des étiquettes morphologiques et syntaxiques du texte en arabe segmenté avec des étiquettes morphologiques et syntaxiques des textes en anglais, pour statuer sur la prise en compte des segmentations valables.

L'outil AMIRA développée par (Diab, 2009) implémente une approche différente basée sur la réalisation de la séparation des clitiques indépendamment de l'étiquetage morphosyntaxique et adopte l'apprentissage supervisé utilisant les Séparateurs à Vaste Marge (SVM).

Nous citons aussi MADA (Morphological Analysis and Disambiguation for Arabic) développé par (Habash, Rambow et Roth, 2009), qui est un outil d'analyse morphologique et de désambiguïsation pour la langue arabe. Cet outil effectue en premier lieux une translittération du texte arabe en entrée en utilisant l'encodage proposé par (Buckwalter, 2002). Il effectue un ensemble de traitements pour produire une liste d'analyses morphologiques potentielles de chaque mot du texte en entrée, indépendamment du contexte. Les segmentations possibles du mot sous la forme préfixe-racine-suffixe sont engendrées et les règles définies par la base de données BAMA (Buckwalter, 2004) sont employées pour vérifier la compatibilité bilatérale. Après la segmentation, MADA détermine l'analyse la plus probable d'un mot étant donné son contexte. Pour y parvenir, MADA s'appuie sur des scores calculés pour les analyses proposées, et ce calcul utilise 19 paramètres : 14 prédits par des modèles SVM (Support Vector Machines), 2 paramètres prédits avec l'outil SRILM[5] (Stolcke, 2002), 1 paramètre prédit à partir du modèle unigramme, et 2 heuristiques supplémentaires.

Par ailleurs, l'adaptation des outils de segmentation des autres langues à l'arabe est aussi un axe envisageable. Des travaux dans cette direction ont donné lieux à plusieurs résultats, comme l'outil MorphTagger (Mansour, 2010) qui était dédié initialement pour l'étiquetage morphosyntaxique de l'hébreu (Mansour, Sima'an et Winter, 2007) et qui s'appuie également sur l'analyseur morphologique BAMA. MorphTagger segmente l'arabe en se basant sur les modèles de Markov cachés (HMM). En termes de performance, il est plus rapide que MADA. L'étape de segmentation ainsi que quelques règles de normalisation ont été ajoutées à l'outil. L'architecture de MorphTagger est similaire à celle de MADA étant donné qu'il utilise la base de données BAMA ainsi que l'outil SRILM pour la désambiguïsation.

D'un point de vue opérationnel, MorphTagger prend en entrée un texte en arabe et il le fait passer à travers l'analyseur morphologique BAMA. Cette première étape produit pour chaque mot, toutes les analyses possibles ainsi que leurs étiquettes morphosyntaxiques puis la séquence d'étiquettes la plus probable en fonction du modèle. La sélection de l'analyse correcte est réalisée en choisissant le morphème le plus probable tout en tenant compte de l'étiquette morphosyntaxique. Afin de résoudre certains problèmes d'ambiguïtés au niveau des sorties, MorphTagger utilise l'outil SRILM. Enfin, nous signalons que ce segmenteur peut

[5] SRILM : The SRI Language Modeling Toolkit (http://www.speech.sri.com/projects/srilm)

effectuer éventuellement quelques étapes de normalisation de textes afin d'obtenir les formes correctes des mots.

Dans le même registre, (Gahbiche-Braham et al., 2012) ont proposé un analyseur morphosyntaxique permettant de segmenter le texte en arabe et de séparer les proclitiques. Cet outil est basé sur les champs markoviens conditionnels CRF. Leur approche procède de la manière suivante : les textes en arabe sont tout d'abord translittérés en utilisant l'encodage de (Buckwater, 2002). Ensuite, la prédiction des étiquettes morphosyntaxiques et de la segmentation est effectuée avec des modèles de prédiction construits à l'aide de l'outil Wapiti (Lavergne et al., 2010) permettant de construire des modèles intégrant un très grand nombre de descripteurs. L'étape de prédiction est suivie d'une étape de normalisation. Finalement des règles de segmentation ont été développées afin de segmenter le texte en arabe et séparer les proclitiques de la forme de base.

Les approches à base de règle ont été aussi investies pour effectuer l'analyse morpho-grammaticale comme c'est le cas de l'arabe G-LexAr proposé par (Debili et al., 2002). Ce système prend en entrée des textes voyellés ou non voyellés et procède de la manière suivante : i) il segment le texte d'entrée en unités morphologiques, ii) il filtre les chaînes de caractères qui ne relèvent pas de l'analyse morphologique de l'arabe, iii) il analyse les unités morphologiques indépendamment de leur contexte et iv) il produit en sortie pour chaque unité lexicale ses segmentations, voyellations, lemmatisations et étiquettes grammaticales possibles sous la forme d'un arbre.

AraParse est un analyseur morphosyntaxique des textes arabes (voyellé, semi ou non voyellé). Il est basé sur des ressources linguistiques à large couverture et utilise un lexique de lemmes généré à partir du dictionnaire DIINAR.1 (Ouersighni, 2002). Pour remédier au problème des mots inconnus, le système utilise une technique d'appariement approximatif implémentée avec le formalisme 'AGFL' et emploie l'opérateur de priorité entre les alternatives d'une règle et les expressions régulières.

De leur côté, (El Isbihani et al., 2006) proposent trois méthodes de segmentation de la langue arabe : 1) à base d'apprentissage supervisé, 2) à base des fréquences, et 3) à base des automates à états finis. Ils démontrent que l'utilisation de la troisième approche donne les meilleurs résultats et qu'elle est adaptable à différentes tâches. C'est la raison pour laquelle nous avons développé aussi un analyseur morphosyntaxique à base de règle et fondé sur les automates à états finis.

On ce qui concerne l'analyse syntaxique de la langue arabe, nous citons principalement les travaux de (Bahou et al., 2005) qui ont proposé un analyseur syntaxique de textes arabes non voyellés. Pour réaliser ce système, ils ont eu recours à l'adaptation et l'implémentation des grammaires HPSG pour la réalisation du système baptisé « SYNTAXE ». Ce système se compose de trois modules à savoir, le module de prétraitement qui construit les matrices attribut/valeur HPSG qui seront stockées dans l'Agenda (une structure de pile), le module d'unification qui sert à tester l'accord entre les constituants et le module d'analyse qui interagit et le module d'unification pour produire comme résultat les arborescences syntaxiques du texte. Ces arborescences seront stockées dans un fichier XML.

(El Kassas et Kahane, 2004) utilisent un arbre de dépendance afin de présenter la structure syntaxique des phrases en arabe. Les travaux de thèse de (El Kassas, 2005) visent le développement des systèmes de production d'énoncés cohérents, valides, compréhensibles et grammaticalement corrects. Les travaux ont porté sur l'analyse syntaxique de l'arabe moderne et sa correspondance avec la sémantique dans une interface syntaxique-sémantique bilingue (arabe – français). Elle a choisi la théorie Sens-Texte (TST) créée par I. Mel'čuk et A. Žolkovskij pour l'élaboration des données langagières.

3.2. Système d'analyse linguistique proposé

L'analyse linguistique profonde est nécessaire pour assurer une extraction d'informations sûre, pertinente et complète. Par exemple lier des éléments qui peuvent être éloignés dans une phrase. Nous pouvons avoir différentes définitions pour l'analyse linguistique, par exemple : selon (Laporte, 2000) : "l'analyse morphosyntaxique est l'ensemble des techniques qui concourent à passer d'un texte brut, exempt d'informations linguistique, à une séquence des mots étiquetés par des informations linguistiques". L'analyse que nous avons mise au point se divise en plusieurs étapes allant du découpage en lexèmes jusqu'aux relations que ceux-ci entretiennent au sein d'une phrase. Les principales étapes de cette analyse sont décrites par le schéma suivant :

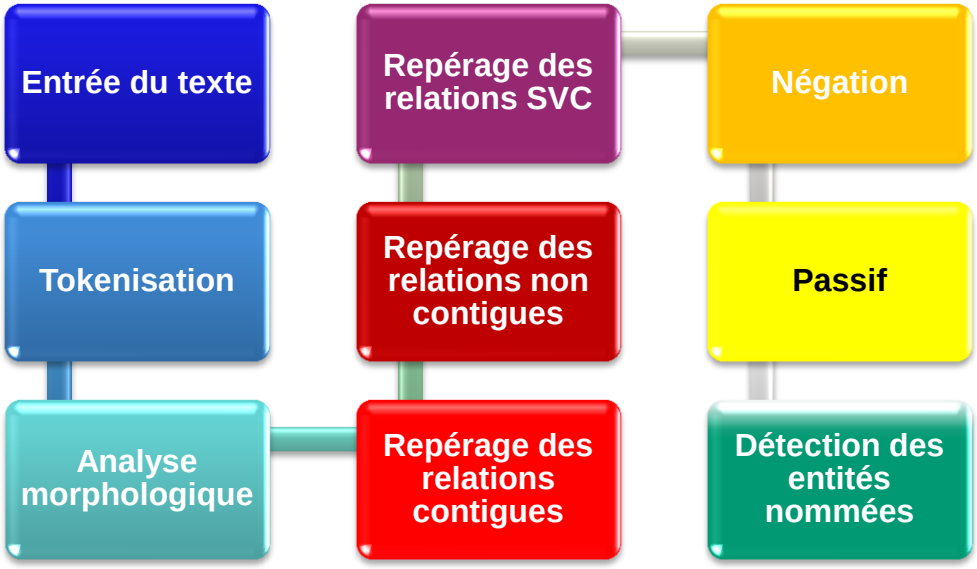


Figure 3. 1. Les étapes de l'analyse linguistique.

3.2.1. Segmentation locale (Tokenisation)

La tokenisation fait partie d'un processus global appelé segmentation. La segmentation est une étape nécessaire et non négligeable dans le traitement de la langue naturelle car elle est "étroitement liée à l'analyse morphologique" (Chanod et Tapanainen, 1996). C'est encore plus le cas avec les langues à morphologie riche et complexe comme l'arabe. Elle est considérée comme une étape primordiale dans un processus de traitement des corpus, des documents ou des textes permettant le découpage en unités lexicales ayant plusieurs niveaux de granularité : texte, phrase et mot. Ces unités sont aussi baptisées « les tokens ou les segments ». La segmentation a besoin de connaître la liste de toutes les limites des mots, tels que des espaces blancs et des signes de ponctuation, etc.

En Traitement Automatique de Langues, nous classons les langues, par rapport à leur système d'écriture, en deux groupes : les langues avec séparateurs et les langues sans séparateurs. Les langues avec séparateurs sont celles qui disposent d'un système d'écriture segmentée : des écritures délimitées par des espaces et où les mots sont nettement séparés par des délimiteurs (espace, signes de ponctuation, caractères spéciaux, ...). C'est le cas pour le français ou l'anglais. Quant aux langues dites sans séparateurs, elles s'appuient sur des systèmes d'écritures non segmentées où les mots ne sont pas séparés par des espaces avec des mots ayant des frontières qui ne sont pas explicites (elles ne sont pas nettes). C'est le cas du japonais, le chinois et le thaï.

Pour ce qui est de l'arabe, elle présente un système d'écriture combinant à la fois les propriétés des deux groupes présentés dessus (voir figure 3.2). C'est un système d'écriture composé d'une écriture segmentée, et d'une autre non segmentée dans laquelle des mots graphiques arabes correspondent à des mots minimaux séparés par des délimiteurs. Cependant, une partie des mots graphiques arabes sont composés d'une suite d'unités lexicales agglutinées pouvant être décomposée en termes de mots minimaux et de clitiques. Ces mots et clitiques doivent apparaître dans le résultat de la segmentation de ces mots composés.

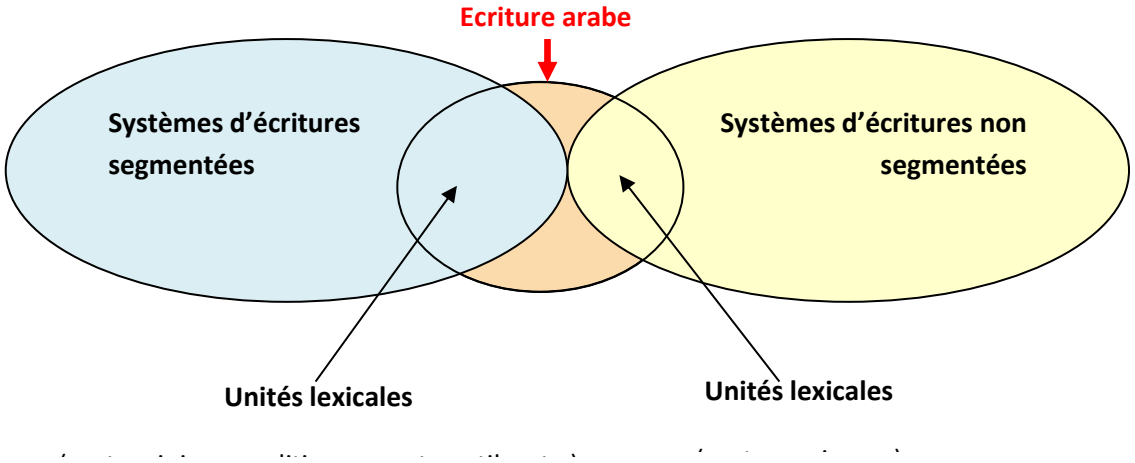


Figure 3. 2. Les groupes de langues par segmentation.

Nous distinguons plusieurs niveaux de segmentation selon le degré de granularité souhaité. Pour cela il existe trois types suivants :

- ***La segmentation lexicale (tokenization)*** : qui représente le découpage d'un texte en segments lexicaux (*tokens*). Ce type de segmentation est aussi appelé *itémisation*.
- ***La segmentation morphologique*** **:** ce type a pour but d'isoler les différents constituants des items lexicaux en unités distinctes, plus petites, qui sont les *morphèmes*.
- ***La segmentation syntaxique :*** ce type de segmentation permet d'identifier les différents constituants du texte en unités indépendantes, plus important que les mots, comme les propositions, les syntagmes, etc. Ce type de segmentation est aussi appelé *chunking*.

Dans le reste de la présente section, nous focalisons notre présentation sur ce que nous avons réalisé au niveau de la segmentation lexicale ou tokenisation qui, encore une fois, consiste à structurer le texte en passant d'un ensemble continu de caractères à une suite discrète d'items lexicaux. Ces items ou *tokens* peuvent être un mot, une expression de plusieurs mots, un chiffre ou un signe de ponctuation. Ces segments sont appelés '*les segments principaux'* et ils sont séparés soit par des signes de ponctuation ou par des espaces dans un texte analysé.

L'étude des corpus nous a permis d'identifier toutes les unités lexicales permettant de segmenter les textes. Parmi ces unités nous citons : l'espace, le point, les deux points, le point-virgule, le point d'interrogation, le point d'exclamation, parenthèse ouvrante, parenthèse fermante, crochet ouvrant, crochet fermant, le tiret, les guillemets, retour à la ligne, début de

ligne, tabulation, les chiffres arabes et les chiffres romains. En plus des chiffres arabes et romains, une bonne partie des pays arabes utilise les chiffres indiens que nous devons considérer aussi dans notre analyse. Des signes de ponctuation supplémentaires propre à la langue arabe tel que la virgule '،', le point d'interrogation '؟' et le point-virgule '؛'. La tokenisation ne permet pas d'avoir des tokens ayant pour l'instant qu'une position de début et de fin. Elle prend aussi en compte les balises, les dates abrégées et les abréviations, etc. Pour illustrer cette segmentation, montrons dans le tableau suivant l'ensemble des tokens que nous obtenons de la phrase en entrée :

Entrée :	كل وعاء يضيق بما جعل فيه إلا وعاء العلم؛ فإنه يتسع به .
Sortie :	كل | وعاء | يضيق | بما | جعل | فيه | إلا | وعاء | العلم | فإنه | يتسع | به | .

Tableau 3. 1. Un exemple sur les segments principaux.

3.2.2. Analyse morphologique

Comme décrit précédemment, la tokenisation permet d'obtenir, à partir d'un texte en entrée, des unités ou segments principaux. Ces résultats doivent être ensuite traités et analysés afin de détecter le rôle de chacun et leur structuration dans le texte ainsi que les règles régissant cette structuration. Cette étape du traitement du texte est du ressort du domaine de la morphologie qui étudie des mots considérés isolement (hors contexte), appelés morphèmes, sous le double aspect de la nature et les variations qu'ils peuvent subir ainsi que la façon dont ces derniers se combinent pour former des lemmes (flexion et dérivation).

La fonction principale de l'analyseur morphologique consiste à retrouver la forme de surface d'un mot stocké dans le lexique à partir de la forme canonique (*lemmatisation*) de ce dernier (infinitif du verbe, masculin singulier d'un adjectif, etc...) et d'attribuer à ces unités lexicales simples ou complexes divers types d'informations à partir de deux types d'étiquettes, d'une part l' étiquette syntaxique concernant les catégories grammaticales (nom, verbes, etc.) et d'autre part, l'étiquette morphologique concernant les traits morphologiques (genre, nombre, la voix, le mode, ...etc.). C'est à ce niveau, que l'ambiguïté morphologique se manifeste le plus souvent, lorsque l'analyse assigne à une unité lexicale plusieurs informations. Cette étape est primordiale lors de l'analyse linguistique. Elle se divise à son tour en plusieurs étapes : la consultation du dictionnaire des formes fléchies d'une part pour récupérer la normalisation du mot et d'autre part, pour permettre de récupérer les informations linguistiques (genre, nombre, catégorie grammaticale, etc.) concernant les mots à reconnaître. Cette analyse morphologique s'intègre comme étape essentielle dans un très grand nombre d'applications en traitement automatique des langues comme le résumé automatique, l'alignement des phrases dans des systèmes de TAO.

Cette étape d'analyse morphologique est d'autant plus importante et plus complexe à appréhender dans le cas de la langue arabe, car rappelons-le que l'une des particularités de cette langue est la présence des formes agglutinées (formes avec des proclitiques et des enclitiques). Ces formes ne sont pas présentes dans le dictionnaire des formes fléchies. Pour identifier ces formes et les traiter correctement, nous avons ajouté un segmenteur secondaire qui consiste à découper et séparer les formes agglutinées (segmentation morphologique), implémenté sous forme de transducteurs à état finis (grammaires morphologiques *HTFST*). Ce système a pour objectif de reconnaître toutes les segmentations possibles du mot en identifiant la forme canonique du mot et les différents affixes et clitiques qui lui sont collés. Cette analyse est encore complexifiée par l'absence ou la présence des voyelles dans les textes

analysés. Pour ceux qui sont semi voyellés ou non voyellés, une consultation du lexique permet de récupérer les formes voyellées correspondantes, c'est à dire leurs alternatives orthographiques lorsqu'elles existent. Dans le cas par exemple du mot non voyellé 'مدرسة' la recherche dans le dictionnaire donne les deux alternatives orthographiques suivantes: "Ecole" (Nom commun féminin singulier) et "Institutrice" (Nom commun féminin singulier). Notons aussi que cette analyse des expressions idiomatiques afin de grouper certains mots pour les considérer comme une seule unité (سكة الحديد : Chemin de fer). Cette reconnaissance se fait à l'aide de règles et de dictionnaires. Notons que les expressions idiomatiques et les mots composés sont inclut dans le dictionnaire général et analysés automatiquement au cours de la consultation du dictionnaire.

3.2.2.1. Segmentation des formes agglutinées

Rappelons d'une forme agglutinée en arabe est constituée d'une racine (lemme) à qui nous rajoutons des clitiques. Ces clitiques peuvent être enchaînées l'un après l'autre, ce qui les rend plus difficiles à manipuler et analyser. Un verbe, par exemple, peut avoir jusqu'à quatre segments secondaires : une conjonction, un complément, un lemme de verbe et un pronom d'objet. De même un nom peut comporter jusqu'à cinq segments secondaires : conjonction, préposition, l'article défini, lemme et pronom.

Nous définissons quatre degrés de cliticisation qui sont applicables dans un ordre strict à base de texte: QST + **[**CNJ+ **[**PRT+ **[**DET+ **[**BASE**]** +SUF = ENC**]** **]** **]** (Habash, 2010), où :

- **DET+ BASE +SUF + ENC :** la base peut avoir soit un article défini (+ *Al* + ال) ou un membre de la classe des enclitiques pronominaux, par exemple : هُمَّ *hm* 'leur / eux.
- **PRT :** classe de proclitiques de particules comme +ل l+ 'à / pour'.
- CNJ: le proclitique de conjonction comme + و w + 'et'.
- QST : la particule de question

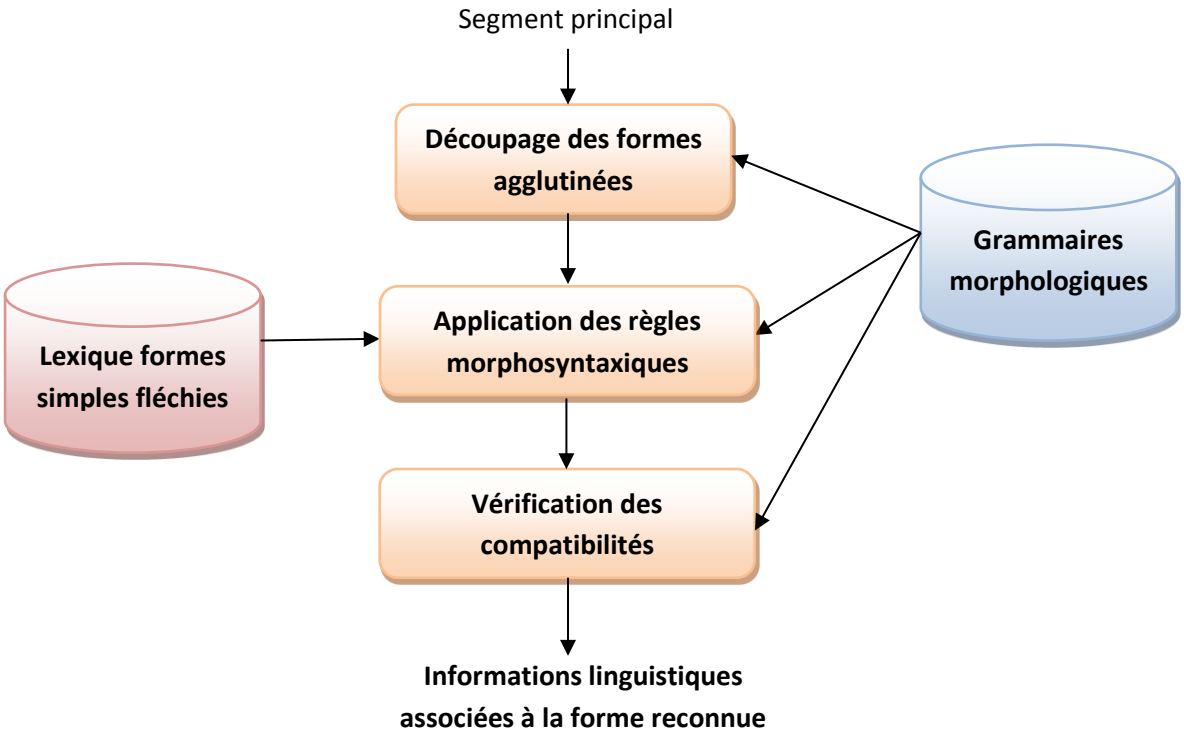


Figure 3. 3. Le schéma du processus d'analyse des formes agglutinées.

L'attachement des clitiques à des formes de mots n'est pas un processus de concaténation simple. Il y a plusieurs règles d'ajustement orthographiques et morphologiques qui sont appliqués sur les mots.

Le processus de segmentation des formes agglutinées, schématisé dans la figure (3.3), se déroule de la manière suivante :

1. Recherche de toutes les compositions possibles entre les clitiques (proclitique, enclitique) et le radical en utilisant les dictionnaires des proclitiques, enclitiques et formes fléchies.
2. Chaque radical est ensuite recherché dans le dictionnaire des formes fléchies. Si ce radical n'existe pas dans le dictionnaire, des transformations morphologiques sont appliquées avant leur suffixation en se basant sur des règles morphosyntaxiques (règles de réécriture qui seront détaillées dans les sections suivantes), enfin le radical résultat est de nouveau recherché dans le dictionnaire des formes fléchies. Par exemple, considérons la forme agglutinée «بسيارته» (avec sa voiture) et les clitiques inclus dans cette forme (ب, ه). Le radical récupéré «سيارت» n'existe pas dans le dictionnaire des formes fléchies. Mais après l'application de la règle de réécriture transformant la lettre «ت» en «ة» en fin de mot, le radical modifié «سيارة» (voiture) est trouvé dans le dictionnaire des formes fléchies et la forme agglutinée «بسيارته» est découpée en proclitique + radical + enclitique comme suit : بسيارته = ب + سيارة + ه (avec sa voiture).
3. Une étape supplémentaire permet de vérifier la relation d'ordre au sein d'une représentation des formants du mot sur un vecteur ordonné (Zmantar et Dichy, 2009). La principale propriété de celui-ci est que chaque proclitique est incompatible avec un proclitique de même position, en raison de la relation d'ordre strict qui régit les formants du mot graphique. Exemples : *wa* et *fa* coordonnants (واو العطف et فاء), qui occupent tous les deux la même position sur le vecteur d'ordre, sont incompatibles entre eux (ils ne peuvent pas apparaître dans un même mot). Cette étape doit aussi vérifier les règles, syntaxiques mais aussi sémantiques, de comptabilité et d'incompatibilité entre les proclitiques et les enclitiques.
4. Vérification de la compatibilité entre les étiquettes morphosyntaxiques des trois composants de la forme agglutinée après découpage (proclitique, radical, enclitique). Seules les segmentations valides sont gardées.

3.2.2.2. La désambiguïsation

Parfois, certains mots restent inidentifiables ou inconnus après les étapes d'analyse morphologique. Par conséquent, le système lui attribue une (des) catégorie(s) par défaut, en s'appuyant sur des informations révélées par sa forme de surface. Par exemple, s'il s'agit d'un mot en caractères latins majuscules, comme ONU, il sera étiqueté comme nom propre.

Dans le cas du traitement de la langue arabe, la majorité des mots restent ambigus à cause de l'absence des voyelles courtes arabes dans les textes (Debili et Suissi, 1998), ce qui est moins prononcé pour les autres langues. Cette ambiguïté, à la fois lexicale et grammaticale, constitue un problème majeur rencontré dans cette phase d'analyse. Il découle du fait que lorsqu'un mot est reconnu, l'analyseur morphologique peut fournir **plusieurs** interprétations qui renvoient à plusieurs catégories syntaxiques ou à plusieurs sens. Le rôle du désambiguïseur morphosyntaxique qui intervient par la suite, est de réduire le nombre des ambiguïtés grammaticales en utilisant des *matrices de désambiguïsation*.

Pour réaliser cette analyse nous nous appuyons sur un dictionnaire utilisé pour la segmentation. Il contient 167423 couples ayant la forme *(mot, catégorie)* et peut être associé à un poids. Les couples ayant au plus une occurrence dans le corpus sont dépourvus de

pondération. Le tableau (3.2) présente un extrait de ce corpus où la première colonne est un mot, la deuxième colonne représente la catégorie grammaticale du mot et la dernière colonne indique le poids associé au couple (mot, catégorie). Ce poids se base sur le nombre d'occurrences du couple (mot, catégorie) dans le corpus d'apprentissage. Il est calculé par la formule suivante :

$$poids(w_i, cat_j) = -log(\Sigma(w_i, cat_j)) \dots (eq01)$$

Où Σ(wi, catj) désigne le nombre d'occurrence du mot wi avec la catégorie catj dans le corpus d'apprentissage.

Mot	*Catégorie*	*Poids*
إداري	adj+	2.56494935746154-
موظفين	nom+	0-
نعلم	verbe+	0.693147180559945-
اقتراف	nom+	0-
نجيب	pre+nom	0.693147180559945-
جولدشتاين	np+	0-
مثلما	prep+	0-
أين	part+	1.6094379124341-

Tableau 3. 2. Un exemple de couples (mot, catégorie) pondérés.

Le modèle de langue s'applique sur des textes étiquetés et utilise des matrices de bi-grammes et trigrammes de catégories morphosyntaxiques obtenues à partir d'un corpus d'apprentissage LDC (Arabic Treebank, 6.0, 2007). Ce corpus est étiqueté et désambiguïsé manuellement. Ces n-grammes sont établis à partir du corpus, et permettent d'attribuer une pondération aux séquences de catégories afin de calculer la catégorie la plus probable d'un mot en contexte. Afin d'optimiser ce processus de désambiguïsation, nous avons modifié le corpus LDC avec un jeu de catégories morphosyntaxiques défini par notre équipe.

Uni-gramme 1	**Uni-gramme 2**	**Uni-gramme 3**	**Poids**
`+conjsubV'	`+verbe'	`+artd'	-0
`+verbe'	`+pron'	`+annp'	-1.09861228866811
`+nom'	`+prondem'	`+pointint'	-0
`+nom'	`+pronrel'	`+guill'	-0
`+prenom'	`+np'	`+2point'	-2.19722457733622
`+verbe'	`+prenom'	`+np'	-3.43398720448515
`+point'	`+verbe'	`+prep'	-1.6094379124341
`+prepN'	`+annp'	`+np'	-4.66343909411207

Tableau 3. 3. Un exemple de trigrammes de catégories.

Nous notons que les probabilités des différents chemins possibles sont calculées afin de résoudre les ambiguïtés de segmentation et de catégorisation. Le résultat de l'application des n-grammes nous permet d'obtenir la suite de couples mot-catégories la plus probable : à l'issu de ce traitement, seul le meilleur chemin est renvoyé par l'automate. L'ambiguïté lexicale est conservée à ce niveau afin d'être traitée plus tard.

3.2.2.3. Transformation morphologique (Règles réécriture)

Lors de la description du traitement des formes agglutinées, nous avons mentionné que si le radical n'existe pas dans le dictionnaire, des transformations morphosyntaxiques sont appliquées. Ces transformations sont formalisées dans des règles morphosyntaxiques appliquées aux différentes segmentations. Ces règles ont pour objectif la réalisation de la correspondance entre un radical traité non reconnu, et un mot du dictionnaire. Cette correspondance est effectuée par un ensemble de règles de réécriture à appliquer au radical ou à la segmentation afin d'arriver à une forme fléchie dans le dictionnaire. Par conséquent, la consultation du lexique des formes du dictionnaire est nécessaire tout au long du processus de la transformation.

Les règles de réécriture que nous proposons prennent en considération les contraintes morphologiques et orthographiques caractérisant la grammaire arabe. Parmi ces contraintes nous citons : l'ajout de lettres, la suppression ou la substitution. Pour chaque contrainte nous lui avons associé une règle de réécriture comme suit :

a) **Ajout de lettre :** cette règle permet d'ajouter une lettre au radical identifié. Nous appliquons cette règle dans le cas des verbes se terminant avec le 'Waw de pluriel'. Cette règle consiste à effectuer une opération de concaténation entre le verbe et la lettre 'Alif â'. La validation de cette segmentation passe par la prise en compte de certaines propriétés morphosyntaxiques comme :
   - le verbe doit être conjugué à la forme active et non pas à la forme passive
   - le verbe doit être transitif
   - le verbe doit être conjugué à la 3ème personne, masculin au pluriel

   Cette règle d'analyse morphologique d'une forme, comme le mot 'ضَرَبُوهُ' (Darabuwhu – ils l'ont frappé), nécessite la restitution de la voyelle longue finale avant la consultation du dictionnaire. L'application de la règle de l'ajout se déroule comme suit :
   - ***1ère étape :*** segmentation de la forme : en verbe + suffixe : 'ضَرَبُو + هُ' (Darabuw + hu)
   - ***2ème étape :*** ajout de la voyelle longue finale 'ا' au radical : 'ضَرَبُو' (Darabuw) → on obtient ضَرَبُوا (Darabuwâ)
   - ***3ème étape :*** la consultation de la forme obtenue dans le dictionnaire : ضَرَبُوا et هُ ' (Darabuwâ + hu) où 'ضَرَبُوا' est la forme fléchie à la troisième personne, masculin, pluriel, à l'accompli, voix active et 'هُ' et un pronom personnel.

b) **Suppression de lettre :** comme son nom l'indique, cette règle consiste à effectuer une opération de suppression de lettres. Là aussi, la prise en compte de certaines propriétés morphosyntaxiques est nécessaire pour la validation de cette segmentation. Les contraintes que nous considérons sont :
   - le verbe doit être conjugué à la forme active et non pas à la forme passive
   - le verbe doit être est transitif
   - le verbe doit être conjugué à la 2ème personne, masculin au pluriel

   Le processus de suppression de lettre d'une forme, comme celle du mot 'ضَرَبْتُمُوهن' (Darabtumuwhun – ils l'ont frappé), nécessite la restitution de la voyelle longue finale avant la consultation du dictionnaire. La règle de la suppression sur ce mot s'applique en trois étapes :
   - ***1ère étape :*** segmentation de la forme : en verbe + suffixe : 'ضَرَبْتُمُو +هُنّ'

(Darabtumuw + hun)

- ***2ème étape :*** suppression de deux voyelles ' وُ' (uw) : 'ضَرَبْتُمُو' (Darabtumuw) → on obtient ضَرَبْتُم (Darabtum)
- ***3ème étape :*** consultation de la forme obtenue dans le dictionnaire : ضَرَبْتُم et 'هُنّ' (Darabuwâ + hu) où 'ضَرَبْتُم' est la forme fléchie à la deuxième personne, masculin, pluriel, à l'accompli, voix active et ' هُ' et un pronom personnel.

c) Substitution de lettres : cette règle consiste à effectuer une opération de substitution de lettres. Elle est appliquée dans le cas des verbes et des noms se terminant par la lettre '*Alif maksoura*'. Pour le cas des verbes, des propriétés morphosyntaxiques doivent être prises en considération pour valider la segmentation obtenue :

- le verbe doit être conjugué à la forme active et non pas à la forme passive
- le verbe doit être transitif
- le verbe doit être conjugué à la 3ème personne, masculin au singulier

L'analyse morphologique pour la substitution de lettre dans une forme, par exemple dans le mot 'كَسَاهُم' (kasaAhum – ils l'ont frappé), nécessite la substitution de la voyelle longue finale avant la consultation du dictionnaire en suivant les étapes suivantes :

- ***1ère étape :*** segmentation de la forme en (verbe + suffixe) ou (nom + suffixe) : 'كَسَا +هُم' (kasaA + hum)
- ***2ème étape :*** substitution de la voyelle longue 'ا' (A - alif) en 'ى' (Y – yaa maksura) : 'كَسَا' (kasaA) → on aura كَسَى (Kasay)
- ***3ème étape :*** consultation de la forme obtenue dans le dictionnaire : كَسَى et 'هُم' (kasay + hmu) où 'كَسَى' correspond à la forme fléchie à la deuxième personne, masculin, singulier, à l'accompli, voix active et ' هُم' et un pronom personnel.

Cependant, dans le cas des particules et prépositions se terminant par Alif maksoura, la règle de substitution consiste à restituer la dernière voyelle longue de " (*Y* – yâ') en ى' (*Y – yaa maksura*) en suivant le même processus. Pour illustrer ce cas, prenons l'analyse morphologique de la forme 'عَلَيهِ' (*'alayhi* – sur lui). Cette analyse se déroule comme suit :

- ***1ère étape :*** segmentation de la forme : en préposition + suffixe : 'إِلَي +هِ' (Ilay + hi)
- ***2ème étape :*** substitution de la voyelle longue 'ي' (y - yaa) en 'ى' (Y – yaa maksura) : 'إِلَي' (Ilay) → on aura إِلَى (IlaY).
- ***3ème étape :*** consultation de la forme obtenue dans le dictionnaire : إِلَى et هِ' (kasay + hmu) où 'إِلَى' est préposition, et 'هِ' et un pronom personnel.

d) Restitution de l'article de définition : parmi ces contraintes, citons les phénomènes de transformation morphologique qui affectent les mots en fonction de la nature de leur lettre initiale. Ainsi, si le mot contient l'article AL (ال), il faut faire la distinction entre les lettres «solaires» et les lettres «lunaires». Les lettres solaires sont caractérisées par une absence de la prononciation du «L» tout en doublant la lettre qui le suit dans la prononciation et dans l'écriture (par le signe de gémination). Quant aux lettres lunaires, le «L» de l'article se prononce et la lettre qui le suit n'est pas dédoublée ni dans la prononciation ni dans l'écriture[6].

i. 1er cas : Lettre lunaire : l'analyse morphologique de la forme 'لِلْمَكْتَبَةِ' (*li-l-*

[6]Pour le détail des lettres «solaires» et «lunaires», voir la page suivante: http://fr.wikilingue.com/es/Lettres_solaires_et_lettres_lunaires

maktabati – ...) nécessite les étapes suivantes :

- ***1 ère étape :*** segmentation de la forme : مَكْتَبَةِ + لْ + لِ' (li+l+maktabati)
- ***2 ème étape :*** restitution de l'article défini 'لْ' en ' ال
- ***3ème étape :*** consultation de la forme obtenue dans le dictionnaire.

ii. **2ème cas : Lettre solaire :** l'analyse morphologique de la forme 'لِلّعِبِ' (*li-ll'ibi* – ...) nécessite les étapes suivantes :

- ***1 ère étape :*** reconnaissance de la préposition لِ (li) et la segmentation de la forme : + لّعِبِ + لِ' (li+ll'ibi)
- ***2 ème étape :*** suppression de la gémination qui occulte (implicite) une autre transformation liée à la restitution de l'article défini 'ال ' comme un proclitique
- ***3ème étape :*** la consultation de la forme obtenue dans le dictionnaire

e) **Ta-Marbouta :** cette règle a pour objectif la transformation orthographique de la forme agglutinée en substituant de la lettre 'ت' en 'ة'. L'analyse morphologique effectuée dans ce cas sur une forme comme le mot 'مَدْرَسَتِهِ' (*madrasatihi* – son école) se déroule de la manière suivante :

- ***1 ère étape :*** segmentation de la forme : en nom + enclitique : 'مَدْرَسَتِ +هِ' (madrasati + hi)
- ***2 ème étape :*** substitution de la lettre ' ت (t - Taa) en 'ة' (t –Taa marbuta) : 'مَدْرَسَتِ' (madrasati) → on obtient مَدْرَسَةِ (madrasati)
- ***3ème étape :*** consultation de la forme obtenue dans le dictionnaire : مَدْرَسَةِ et هِ' (madrasati + hi) où 'مَدْرَسَةِ' correspond à un nom, féminin au singulier mis au génitif et 'هِ' et un pronom personnel.

f) **Hamza :** ce cas concerne les formes nominales qui se terminent par la lettre 'ء' (*ʼ - hamza*). La règle de réécriture dans ce cas consiste à substituer la lettre supportant la hamza, waw ou yaa, par la lettre 'ء' (*ʼ - hamza*). L'identification des cas à substituer passe par la détection de la lettre casuelle qui détermine la lettre supportant la hamza. Cette règle tient compte aussi de la fonction grammaticale du mot. Par exemple, l'application de cette règle sur les deux formes دَوَائِه (dawa'ihu) et دَوَاؤُهُ dawa'uhu, donne la forme دَوَاءٌ (dawa'un - médicament). D'une manière générale :

- Si la hamza est accompagnée par une 'ضَمَّة - ُ ' (u – damma), elle prend la forme 'ؤ' (w – hamza 'alaa al-wâw); c'est le cas du nominatif.
- Si la hamza est accompagnée par une ' فَتحَة - َ ' (a – fatha), elle prend la forme 'أ' (a – hamza) ou 'ء' (hamza 'alaa es-satir); c'est l'accusatif.
- Si la hamza est accompagnée par une ' كَسْرَة - ِ ' (i – kasra), elle prend la forme 'ئ' (y – hamza 'la-ya'); c'est le génitif;
- Si la hamza est accompagnée par une ' سُكُون - ْ ' (sukun – signe de quiescence), elle prend la forme 'ء' (hamza 'alaa es-satir).

g) **Y-Shadda :** cette règle concerne le remplacement d'une double consonne par la chadda (dédoublement de la consonne). Elle est appliquée dans le cas des prépositions, par exemple 'فِي' (fi) + 'ي' (y - ya') → 'فِيَّ'. Ce remplacement est motivé par le fait que l'enclitique 'ي' (y – ya') ne se combine qu'avec les mots outils ayant au moins trois consonnes, le cas échéant, sa concaténation nécessite l'ajout de la chadda. Les propositions suivantes représentent des cas d'application de cette règle :

- 'عَنْ' (*'an - selon*) + 'ي' (*y - ya'*) → عَنِّي ('*anniy – selon moi*)
- 'دُونَ' (*duwna* - sans) + 'ي' (*y - ya'*) → 'دُونِي' (*duwniy - sans moi*)

Toutefois, nous signalons qu'il existe une exception dans cette règle. Elle concerne les

lettres assimilées à des verbes, *الأحرف المشبّهة بالفعل*, peuvent engendrer deux écritures différentes en se combinant avec un même enclitique, comme c'est le cas de la lettre لَعَلَّ (*la'alla – peut-être*) + 'نِي' (*niy* -) qui donne les deux formes : 'لَعَلَّنِي' (*la'allaniy* – je pourrai) et 'لَعَلِّي' (*la'alliy* – je pourrai).

Nous pouvons résumer les règles présentées dans le tableau suivant :

Nom de la règle	Condition	Résultat	Décomposition	Forme agglutinée	Traduction
Article défini	?' + 'ال' + 'ل' + "ل *l +Al + l?*	لل+ *ll+*	ل + ال + مسجد *l+Al+msjd*	للمسجد *llmsjd*	A la mosquée
			ل + ال + لجنة *l+Al+lajnat*	للجنة *llajnat*	Au paradis
Ta-Marbuta	ة- *-h* + pron	ت- *-t* +pron	مدرسة + هم *mdrsH+hm*	مدرستهم *Mdrsthm*	Leur école
Alif-Maksura	ى- *-y* +pron	ا- *-A* +pron	سقى + ه *sqY+h*	سقاه *sqAh*	Il l'a irrigué
	Exception	ي- -y +pron	على + ه *'lY+h*	عليه *'lyh*	Sur lui
Waw-de-Pluriel	وا- *-wA*+pron	و- *-w*	ضربوا + ه *DrbwA+h*	ضربوه *Drbwh*	Ils l'ont frappé
	تم- *-tm* +pron	تمو- -tmw	ضربتم + ه *Drbtm + h*	ضربتموه *Drbtmwh*	Vous l'avez frappé
Hamza	ء- *-'* +pron	ئ- *-*y+pron	سماء + ه *smA' +h*	سماؤه *smAwh*	Son ciel
		ؤ- -w+pron	سماء + ه *smA' +h*	سمائه *smAyh*	
		ء- *-'* +pron	سماء + ه *smA' +h*	سماءه *smA'h*	
Y-Shadda	ي +ي - *-y +y*	ي y	قاضي + ي qADy+y	قاضيّ *qADy*	Mon juge
N-Assimilation	من *mn + m/n*	م *m +m/n*	من + ما *mn+mA*	مما *mmA*	De
	عن 'n +*m/n*	ع ' +m/n	عن + من *'n+mn*	عمن *'mn*	De qui ?
	لا + أن *An+ lA*	ألّا *Ala*	أن + لا *An+lA*	ألّا *Ala*	Ne pas

Tableau 3. 4. Les règles de réécriture morphosyntaxique.

3.2.3. L'analyse syntaxique

Une phrase est une suite de mots permettant de véhiculer un sens. La majorité des théories linguistiques s'accordent sur le fait que les mots d'une phrase ne sont pas disposés de façon aléatoire, au contraire ils suivent un système d'organisation ou une structure assez rigide. Cette structure est appelée *la structure syntaxique de la phrase*. Il existe deux structure de représentation :

- *Structures syntagmatiques (PSG, de l'anglais Phrase Structure Grammar) :* elle décrit la façon dont les mots peuvent être groupés en des paquets de plus en plus gros en d'autres termes les mots se rassemblent en constituants et que chaque constituant doit avoir une tête.
- *Structure de dépendance (DG, de l'anglais Dependency Grammar) :* permet de mettre en avant les relations entre les mots d'une phrase en se basant sur le principe que les mots dans une phrase dépendent les uns des autres.

Les structures de dépendances (arbre de dépendance), auxquelles nous nous intéressons dans cette étude, sont plus anciennes que les structures syntagmatiques. En effet leur usage remonte à l'antiquité. Les grammairiens arabes du 8ème siècle, comme Sibawayh, distinguaient gouverneur et gouverné en syntaxe et utilisaient cette distinction pour formuler des règles d'ordre des mots et de rection (kahnane, 2001). Au 19ème siècle, les grammaires scolaires de l'anglais ont enseigné l'analyse de la phrase sous forme de diagramme basé sur la dépendance. Lucien Tesnière fut un des premiers à mettre en place dans les années 30 une théorie linguistique basée sur la dépendance, et fut publiée quelques temps après sa mort en 1959 sous le nom de « *Eléments de syntaxe structurale»*.

Un arbre de dépendance syntaxique est enrichi avec un étiquetage des dépendances par des fonctions syntaxiques. Cet étiquetage sert comme complément à l'arbre afin d'encoder l'organisation syntaxique des phrases. Une fonction ou relation syntaxique permet de distinguer les dépendants d'un même mot et de rassembler les dépendants qui ont un comportement syntaxique similaire. Par « relation », on réfère au lien entre gouverneur et dépendant et par « fonction », on réfère au rôle rempli par un dépendant dans le régime du gouverneur. La notion de fonction syntaxique est universelle mais sa déclinaison au niveau des langues donne des fonctions propres à chaque langue. Le recensement et l'énumération de ces fonctions reste à la charge des grammairiens, à ce sujet (Kahane, 2001) expose sur la difficulté de cette tâche :

« *L'une des principales difficultés pour décider combien de fonctions syntaxiques il est nécessaire de considérer est qu'on peut toujours attribuer une propriété particulière à la catégorie du dépendant ou du gouverneur (comme le font les grammaires syntagmatiques) plutôt qu'à l'étiquette de la relation de dépendance entre eux. Quitte à multiplier les catégories syntaxiques, il est formellement possible de limiter l'étiquetage des relations à un simple numérotage (il faut quand même garder un minimum pour distinguer entre eux les différents compléments du verbe). Il semble donc difficile d'établir des critères exacts pour décider si deux dépendances doivent ou non correspondre à la même fonction et il est nécessaire de prendre en compte l'économie générale du système en cherchant à limiter à la fois le nombre de catégories syntaxiques et le nombre de fonctions syntaxiques et à chercher la plus grande simplicité dans les règles grammaticales. On attribuera donc à la catégorie syntaxique les propriétés intrinsèques d'une lexie (c'est-à-dire qui ne dépendent pas de la position syntaxique) et à la fonction les propriétés intrinsèques d'une position syntaxique (c'est-à-dire qui ne dépendent pas de la lexie qui l'occupe)*. »

Les grammaires se basent sur des critères morphologiques, positionnels, catégoriels et sémantiques, afin de distinguer les différents types de fonctions syntaxiques. En français, la cliticisation, à titre d'exemple, est une des opérations les plus utilisées pour déterminer certaines fonctions syntaxiques. Par exemple, pour la phrase '*Ziyad mange la galette'*, on définit *la galette* comme un élément de la phrase remplissant la fonction syntaxique complément d'objet direct (COD) car il est remplaçable par le clitique objet *la*. Par contre, dans *Amel chante le soir*, on ne définit pas *le soir* comme un élément remplissant la fonction syntaxique COD car il n'est pas remplaçable par un pronom objet.

3.2.4. Les relations syntaxiques

L'annotation syntaxique, autrement dite *processus d'identification d'une relation syntaxique,* dans le cadre d'une analyse de dépendance implique les décisions suivantes : *l'attachement* et *l'étiquetage*. L'attachement concerne la détermination si deux mots sont connectés directement ou pas, en d'autre termes, c'est l'identification d'un lien direct de dépendance

syntaxique entre deux mots de la phrase par la mise en avant d'une relation syntaxique entre la tête (gouverneur) et le mot dépendant (régie). Quant à l'étiquetage, il consiste à regrouper les dépendants syntaxiques et annoter la relation identifiée par un nom référant à une famille (ou type) de constructions syntaxiques d'une langue donnée.

Dans cette section, nous procédons en deux étapes : 1) traiter les syntagmes nominaux, puis 2) présenter les relations sujet-verbe-complément. L'analyse effectuée est une analyse de dépendance, et comme nous utilisons le langage HTFST pour la partie implémentation, le fichier analysé en entrée n'est que ligne par ligne, et nous ne reconnaissons que des chaînes de caractères. Les relations ne sont pas représentées sous leur forme arborescente, mais elles sont « aplaties » et représentées sous forme de paires « tête-dépendant », auxquelles peuvent être ajoutés des éléments appelés *indications linguistiques*, tels que les déterminants, les prépositions, etc.

Chaque relation est typée selon des catégories choisies par les linguistes parmi lesquelles nous citons :

- **SV** pour les relations sujet-verbe;
- **VC** (verbe-complément) : cette relation regroupe à la fois les relations verbe-objet et les relations qu'entretient le verbe avec les compléments circonstanciels ;
- **GD**, uniquement dans les groupes nominaux, relie des éléments dont la tête est à gauche du dépendant ;
- **DG**, uniquement dans les groupes nominaux également, relie des éléments dont la tête est à droite du dépendant ;
- **CIRONSTANT** : relie les compléments circonstanciels à l'attribut d'une relation ATTRS.

Une étude linguistique spécifique de la langue arabe nous a permis de définir et d'écrire des règles dans le but d'établir des relations de dépendance (contiguës et non contiguës) entre les mots au sein du syntagme nominal dans le but de définir le rôle sémantique des mots. Ces relations permettent ensuite de reconnaître les mots composés présents dans une phrase. Nous avons passé en revue les différentes relations syntaxiques régies par le nom, l'adjectif et les mots outils que nous présentons par les relations syntaxiques suivantes :

3.2.4.1. Les relations syntaxiques gouvernées par le nom

Avant d'introduire les règles syntaxiques, rappelons qu'en arabe le nom prend les marques casuelles exprimées par des voyelles courtes. De ce fait, le nominatif est exprimé par le suffixe /u/, l'accusatif par le suffixe /a/ et le génitif par le suffixe /i/. De plus, un nom peut être défini (DEF) ou indéfini (INDEF), et en fonction du nombre du sujet qui lui est lié, il peut être au singulier, duel et au pluriel.

En considérant tous ces aspects, nous présentons quelques relations syntaxiques gouvernées par le nom. En particulier nous présentons les sept relations suivantes : la modification, la relation complément de nom, la relation complément d'objet indirect, l'apposition, la corroboration, la quantification numérale et la coordination.

3.2.4.1.1. La modification

La modification est la relation qui permet de lier un mot à un nom. Ce mot, désigné par *modifieur*, associera à travers cette relation une caractéristique au nom auquel il est rattaché. Ce rôle joué par le modifieur permet de déduire que le mot lié au nom est un adjectif. Nous pouvons représenter cette relation comme suit :

(N)-*modif*→(ADJ) …(1)

En général, le terme modifieur désigne une adjonction au nom et il est toujours placé après le déterminant. Il peut être *libre*, quand il est facultatif (ex : Kamel a acheté deux voitures ***blanches***), et il peut être *lié* quand il est obligatoire (ex : Michael Schumacher est dans un état ***critique***). Par défaut, un modifieur est un adjectif qui s'accord en genre (féminin, masculin), nombre (singulier, duel ou pluriel), définitude (défini ou indéfini) et cas (nominatif, accusatif ou génitif) avec le nom qu'il qualifie conformément au tableau suivant :

N (déterminant)
genre = g (fem,masc)
nombre = nb (SG, DL,PL)
cas = c (NOM,ACC,GEN)
définitude = d (DEF, INDEF)

Modification →

ADJ (modifieur)
genre = g
nombre = nb
cas = c
définitude = d

Afin de montrer cet accord, prenons l'exemple des phrases suivantes :

- ✓ 'نجح ولدٌ نجيبٌ' (un enfant brillant a réussi)
  ⇒ najaha waladun – modif → najiibun
  V(PASSE) (N,masc)SG+NOM+INDEF (ADJ,masc)SG+NOM+INDEF
- ✓ 'نجح الولدُ النجيبُ' (le brillant enfant a réussi)
  ⇒ najaha alwaladun – modif → annajiibun
  V(PASSE) (N,masc)SG+NOM+DEF (ADJ,masc)SG+NOM+DEF
- ✓ 'نجح ولدان نجيبان' (deux enfants brillants ont réussi)
  ⇒ najaha waladani – modif→ najiibani
  V(PASSE) (N,masc)DL+NOM+INDEF (ADJ,masc)DL+NOM+INDEF
- ✓ 'نجح الأولادُ النجباءُ' (les enfants brillants ont réussi)
  ⇒ najaha alawladu – modif→ annujabaou
  V(PASSE) (N,masc)PL+NOM+DEF (ADJ,masc)PL+NOM+DEF

L'opération de modification dans la grammaire arabe est aussi la fonction dite na't (نعت) ou sifaa (صفة). Elle peut être exprimée par un adjectif à valeur, un participe actif 'إسم الفاعل' (suivant le schème فَاعِل 'faa'il'), un participe passif 'إسم المفعول' (respectant le schème مَفْعُول'maf'uul'), un comparatif 'إسم التفضيل' (régi par le schème أَفْعَل 'aaf'al') ou encore un superlatif. Pour illustrer ces propriétés prenons les exemples suivants :

- ➢ 'كان محمدٌ رسولاً صادقاً' (Muhammad était un messager honnête)
  kaana muhammad+u+n rassul+a+n -modif → saadik+a+n
  (V)PASSE (N)+NOM+INDEF (N)+ACC+INDEF ($ADJ_{participe\ active}$)+ACC+INDEF
- ➢ 'كان محمدٌ رسولاً مَبْعُوثاً' (Muhammad était un messager envoyé)
  kaana muhammad+u+n rassul+a+n -modif → mab'uuth+a+n
  (V)PASSE (N)+NOM+INDEF (N)+ACC+INDEF ($ADJ_{participe\ passive}$)+ACC+INDEF
- ➢ 'كان محمدٌ رسولاً أَحْسَناً' (Muhammad était un messager le plus vertueux)

kaana muhammad+u+n rassul+a+n –modif → ʔhsan+a+n
(V)PASSE (N)+NOM+INDEF (N)+ACC+INDEF ($ADJ_{comparatif}$)+ACC+INDEF

- 'كان محمدٌ الرسولَ الأَفْضَلَ' (Muhammad était le meilleur messager)

kaana muhammad+u+n al+rassul+a –modif → al+ʔafdal+a
(V)PASSE (N)+NOM+INDEF DEF+(N)+ACC DEF+($ADJ_{superlatif}$)+ACC

3.2.4.1.2. Le complément de nom

Le complément de nom est un mot défini par un article (de définition) ou par annexion (nom propre ou un pronom clitique), mis au génitif. Nous pouvons représenter cette relation par la règle suivante :

(N)-*compN*→(N)GEN …(2)

Le complément de nom est caractérisé par les propriétés suivantes :

- Plusieurs compléments de noms peuvent s'enchaîner dans une phrase. Pour illustrer cet enchaînement prenons l'exemple suivant :
  خاتم ابنةِ صديقةِ أختي (la bague de la fille de l'amie de ma sœur)
  kaatam+u+Ø ***compN*** → *ibnati+Ø* ***compN*** → *sadiiqati+Ø* ***compN*** → *ʔukht+Ø#ii*
  (N)+NOM (N)+GEN (N)+GEN (N)GEN #(PRO)
- Le complément de nom peut être composé par une coordination comme c'est le cas du complément de la phrase أبهرتني قصةُ محمدٍ و كريمٍ (L'histoire de Mohammed et Karim m'a éblouie) où :
  abharat#nii *qisat+u*-**compN**→***Muhammad+i+n wa#kariim+i+n***
  (V)PASSE#(PRO) (N)+NOM (N)+GEN+INDEF (COORD)#(N)+GEN+INDEF

En plus de ces propriétés syntaxiques, un complément de nom peut appartenir aux types suivants :

a. Nom défini : par exemple :
   جاء وزيرُ الدولةِ (le ministre d'état est venu)
   jaa' ***waziir+u*** **–compN** → ***al+dawlat+i***
   (V)PASSE (N)+NOM DEF+(N)+GEN

b. Nom propre : prenons la phrase :
   أبهرتني قصةُ محمدٍ (l'histoire de Mohammed m'a éblouie)
   abharat#nii ***qissat+u*-compN** → ***muhammad+i+n***
   (V)PASSE#(PRO) (N)+NOM+INDEF (N)+GEN+INDEF

c. Pronom clitique : comme c'est utilisé dans la phrase :
   أبهرتني قصتُهُ(son histoire m'a éblouie)
   abharat#nii ***qissat+u*-compN** → ***#hu***
   (V)PASSE#(PRO) (N)+NOM #(PRO)

3.2.4.1.3. Le complément d'objet indirect

Ce type de relation est réalisé avec un constituant prépositionnel qui est suivie par un nom mis au génitif. Nous pouvons représenter cette relation par la règle suivante :

(N)-*PREP-compI*→(N)GEN …(3)

Cette relation possède les propriétés syntaxiques suivantes :

i. La structure est itérative car plusieurs compléments de nom peuvent être utilisés dans la même phrase comme dans la phrase suivante :

ترجمةٌ من الفرنسيةِ إلى العربيةِ (Une traduction du français vers l'arabe)

*tar*ʤ*amat+u+n-**PREP**→ mina-**compI**→**al+firinsijat+i -PREP**→ʔ*ilaa-**compI**→ **al+ʕarabijat+i***

(N)+NOM (PREP) DEF+(N)+GEN (PREP) DEF+(N)+INDEF

ii. Comme le complément d'objet direct, le complément indirect peut être aussi un constituant coordonné, et pour illustrer ce cas de figure, prenons l'exemple de la phrase :

ترجمةٌ من الفرنسيةِ والعربيةِ (Une traduction du français et de l'arabe)

*tar*ʤ*amat+u+n-**PREP**→ mina-**compI**→**al+firinsijat+i wa#al+ʕarabijat+i***

(N)+NOM+IND (PREP) DEF+(N)+GEN (COORD)#DEF+(N)+GEN

3.2.4.1.4. L'apposition

L'apposition est la relation permettant de rattacher un mot, considéré comme dépendant de la relation et appelé *appositif*, à un nom afin de lui apporter un complément d'information. Ce complément concerne une qualité ou une nature. Cette relation formalise en arabe le phénomène dit *albadal*, (' البدلla substitution'). Nous pouvons représenter cette relation comme suit :

(N)-appos→(N)DEF|NEUTRE …(4)

L'appositif peut avoir différents type :

- Nom défini : prenons l'exemple de cette phrase :
  محمدٌ الرسولُ (Mohammed le prophète)
  Muhammad+u+n -appos→ al+rassul+a+u
  (N)+NOM+INDEF DEF+(N)+NOM
- Nom indéfini : par exemple :
  أدهشنا محمداً صديقك (Mohammed ton ami nous a surpris)
  ʔ*adhachana **muhamad+a+n–appos→ sadiq+a#ka***
  (V)PRESENT (N)+ACC+INDEF (N)+ACC#(PRO)
- Démonstratif : comme nous l'illustrons dans la phrase suivante :
  الفتاة هذه (cette fille là)
  al+fatat+u* -appos → *haDihi
  DEF+(N)+NOM (DEI)
- Cardinal : comme c'est le cas de la phrase suivante :
  الفرد الثلاثون (le trentième individu)
  al+fard+u* -appos→*al+*ʈ*thalathuuna
  DEF+(N)+NOM DEF+(CARD)NOM

En terme de coordination, l'oppositif suit en nombre et en genre son gouverneur, par exemple :

'داويت عمراً و خالداً الجريحين' (J'ai soigné Omar et Khaled les blessés)

*daawajtu **'umar+a+n wa#khalid+a+n-appos→ al+jarihajn***

(V)PASSE (N,masc)ACC+INDEF (COORD)#(N,masc)+ACC+INDEF DEF+(N,masc)DUEL.ACC

Par ailleurs, la relation d'apposition est parfois complexe, car dans certains cas nous trouvons des noms coordonnés dont chacun apporte une identification différente, par exemple la phase :

جنبك الله أمرين فقراً و هرماً) ALLAH t'a épargné deux choses, pauvreté et vieillesse)

janabaka Allahu 'amrajn-**appos→** **[*faqr+a+n* *wa#ham+a+n*]**

(V)PASSE NP (N,**masc**)**DUEL**.ACC (N,**masc**)**SG**.ACC (COORD)#(N,**masc**)**SG**.ACC

Sur un autre registre, L'apposition possède plusieurs variantes selon l'étendu du sens qu'il apporte au nom qu'il suit. Ces variantes sont au nombre de trois : apposition du tout 'بدل الكل', apposition de la partie 'بدل الجزء', apposition d'inclusion 'بدل الإشتمال'.

- **Apposition du tout :** quand l'apposition désigne le nom suivi lui-même et l'égal au sens, et n'a pas besoin par conséquent d'un pronom la liant au nom suivi. Pour illustrer cette variante prenons la phrase :

  تولى الفاروقُ عمرُ الخلافةَ (Omar le juste a pris la succession)

  Tawalaa al+*faruwk+u*-**appos→** ***'umar+u*** *al+khilafat+a*

  (V)PASSE DEF+(N,masc)+NOM (N,masc)+NOM DEF+(N,masc)+ACC

  Remarque : dans le cas où l'appositif est un nom propre ou un prénom, il devient la tête de l'apposition et le nom qui le précède sera le dépendant de cette relation syntaxique.

- **L'apposition de la partie ou partitif :** cette variante concerne les appositifs qui réfèrent une partie matérielle des noms qu'ils suivent. Ces appositifs doivent être reliés à un pronom qui fait référence au nom suivi et qui s'accorde avec lui en genre et en nombre. A titre d'exemple, nous donnons la phrase suivante :

  ضاعت فلسطين أرضها (a été perdue la terre de la Palestine)

  Da'at ***falastin+u*** **-appos→** ***'arD+u#ha***

  (V)PASS (N)+<u>NOM</u>+INDEF N+<u>NOM#</u>**PRO**

  Nous remarquons bien dans cet exemple que la terre représente une partie matérielle de Palestine, et que l'apposition contient bien un pronom qui référence la Palestine.

- **Apposition d'inclusion** : l'appositif dans cette variante désigne une des caractéristiques, ou propriétés morales, liées au nom auquel il est lié. Comme dans la précédente variante, l'appositif doit être attaché à un pronom référençant le nom auquel il est lié l'appositif.

  ضاعت كرامتهافلسطين (a été perdue la dignité de la Palestine)

  Da'at ***falastin+u*** **-appos→** ***karamat+u#ha***

  (V)PASS (N)+<u>NOM</u>+INDEF (N)+<u>NOM#</u>**PRO**

  Dans cet exemple, nous remarquons que le mot dignité n'est pas une partie matérielle de la Palestine, mais c'est une propriété morale. De plus, nous voyons très bien que l'appositif est lié à un pronom référençant la Palestine.

3.2.4.1.5. La corroboration (al-tawabi' – al-tawkîd)

Nous disons que deux mots dans une phrase en arabe sont liés par une relation de corroboration si les deux mots se suivent et que le second est utilisé pour confirmer ou insister

sur le premier. Le deuxième mot s'appel dans le cadre de cette relation *le corroboratif*. Nous schématisons cette relation comme suit :

(N)- corrob→(N)DEF|NEUTRE …(5)

Il existe deux sortes de corroboration : formelle (lafzi : لفظي) et sémantique (ma'nawiy معنوي). La corroboration formelle est caractérisée par la répétition du mot, en d'autres termes le mot et son corroboratif sont les mêmes. Par exemple :

'هَيْهَاتَ هَيْهَاتَ لِمَا تُوعَدُونَ' (loin loin ce qu'on vous promet)

hayhAt+a **-appos→** ***hayhAt+a*** ***lima*** ***tu'ad+un+n***

(N)+ACC+INDEF (N)+ ACC+INDEF PART V(PREST)

La corroboration sémantique a la particularité d'utiliser l'une des unités lexicales suivantes : nafs (' نفسen personne') , ʔajn (' عينmême'), ʤamiiʕu (جميع 'tous'), 'amma (' عامةentier'), ʔadʒmaʕu kullu (كل 'tout'), killa (كلا ''), killta (كلتا ''), ainsi que leurs variantes morphologiques **possibles**. Nous précisons que les unités kilâ et kiltâ sont spécifiques à la corroboration du duel et sont fléchis à son cas. Il est obligatoire que ces unités lexicales se joignent à un pronom qui s'accorde, en genre et en nombre, avec le corroboré, sauf pour les variantes : ajma'un (أجمع 'tous'), jama'aa' (جمعاء 'tous'), 'ajma'un (أجمعون 'tous'), juma' (جُمَع 'tous'). Afin d'illustrer ce type de corroboration, voici quelques démonstrations :

- فاز المتسابق عينه(le concurrent a gagné en personne)
  faaʔza ***al+mutasabik+u -appos→'ayn+u#hu***
  (V)PASSE DEF+(N)**SG**+NOM (N)**SG**+NOM#(PRO)

- وصل الفائز نفسه (le gagnant est arrivé en personne)
  wassala ***al+fa'iz+u-appos→ nafs+u#hu***
  (V)PASSE DEF+(N)**SG**+NOM (N)**SG**+NOM#(PRO)

- وصل الفائزون أنفسهم (les gagnants sont arrivés en personne)
  wassala ***al+fa'izuuna-appos→ ʔanfus+u#humu***
  (V)PASSE DEF+(N)**PL**+NOM (N)**PL**+NOM#(PRO)

- تفوق المجتهدان كلاهما(les deux studieux ont gagné tous les deux)
  tafawaqa ***al+mujtahidaan*-appos→*kilaa#humaa***
  (V)PASSE DEF+(N)**DUEL**.NOM (N)**DUEL**.NOM#(PRO)

- سافر المعتمرون كلهم(les pèlerins sont partis tous)
  *sa*aʔ*fara* ***al+mu'tamur+u+un* -appos→*kull+u#humu***
  (V)PASSE DEF+(N)**PL**+NOM (N)**PL**+NOM#(PRO)

- حضر المدعون جميعهم(les invités sont arrivés tous)
  haDara ***al+mud'+u+un* -appos→*jamii'+u#humu***
  (V)PASSE DEF+(N)**PL**+NOM (N)**PL**+NOM#(PRO)

- استقبلنا الزائرين عامتهم(nous avons reçu les visiteurs en leur globalité)
  ʔistaqbal#na ***al+zaa'ir+iin*-appos→ *ʕaamat+u#hum***
  (V)PASSE#(PRO) DEF+(N)+NOM (N)+NOM#(PRO)

3.2.4.1.6. La quantification numérale

Cette relation syntaxique présente un cardinal suivi d'un nom singulier mis à l'accusatif indéfini. Dans ce cas, le gouverneur de cette relation est le nom et le dépendant est le *cardinal*. La quantification permet le repérage des mesures dans les textes. Nous pouvons présenter cette relation par la règle suivante :

(CARD)-*quant-num*→(N)INDEF …(6)

Le cardinal peut jouer les rôles suivants dans une phrase :

- Un cardinal peut être un sujet au nominatif, par exemple :
  جاء خمسةُ رجالٍ (Cinq hommes sont venus)
  'jaa' ***khamssat+u*-quant-num→*rijaal+i+n***
  (V)ACTIF.PASSE (CARD)**NOM**+INDEF (N)PL+GEN+INDEF

- Le cardinal peut aussi être un complément d'objet direct comme dans cette phrase :
  قتل المجاهدُ خمسةَ جُنُودٍ (le combattant a tué cinq soldats)
  jaa' *al+mujahid+u* ***khamssat+a*-quant-num→*junuud+i+n***
  (V)ACTIF.PASSE DEF+(N)+NOM (CARD)**ACC**+INDEF
  (N)**PL**+**GEN**+**INDEF**

Dans la grammaire arabe, en fonction du nombre véhiculé par le cardinal nous distinguons les trois cas de figure suivants :

- ✓ Si le cardinal représente un nombre compris entre 3 et 10, alors le gouverneur, qui est le nom dénombré (الْمَعْدُودُ), doit être au pluriel quel que soit son genre : masculin ou féminin. Le nombre prend différents cas suivant sa situation dans la phrase. De plus, il doit être indéfini et mis au génitif. Nous signalons aussi que le genre du nombre dans ce cas est opposé à celui du dénombré : si le dénombré est masculin alors le nombre doit être mis au féminin et vice versa. Nous utilisons les exemples suivant pour démontrer ces propriétés :

  - ➢ جاء سبعُ فتياتٍ (sept filles sont venues)
    Jaa' ***sab+'u*num→-quant-** ***fatayaat+in+***
    (V)PASSE (CARD)**MASC**+**NOM** (N)**FEM**+**SG**+**GEN**+**INDEF**

  - ➢ جاء سبعةُ رجالٍ (sept hommes sont venus)
    Jaa' sab'at+u-quant-num→ rijaal+i+n
    (V)PASSE (CARD)FEM+NOM (N) MASC +SG+GEN+INDEF

- ✓ Si le cardinal représente un nombre compris entre 11 et 99, alors le nom dénombré est au singulier, à l'accusatif et généralement à l'indéfini. Par ailleurs, le nombre prend différents cas suivant sa situation dans la phrase. La phrase suivante montre ces propriétés :
  رأى يوسفُ أحدَ عشَرَ كوكباً (Youssef a vu onze planètes)
  ra'a *yussuf+u+n* ***ahda 'achar+a*-quant-num→*kawkab+a+n***
  (V)ACTIF.PASSE (N)+NOM+INDEF (CARD)**ACC+NEUTRE**
  (N)**SG**+**ACC**+**INDEF**

- ✓ Si le cardinal représente un nombre compris entre 100 et 1000, alors le nom dénombré est toujours mis au singulier, généralement indéfini et mis au génitif. A titre illustratif, prenons cet exemple :

➢ عاش جدي مائة سنةٍ (mon grand-père a vécu cent ans)
'acha *jad+i#i* ***mi'at+a*-quant-num→** **sanat+*i*+*n***
V(PASSE) (N)+NOM#**PRO** (CARD)ACC (N)SG+**GEN**+INDEF

En plus des quantificateurs décrit jusque-là, il existe des unités lexicales qui indiquent le sens du nombre mais qui ne sont pas des nombres. Par conséquent, la quantité indiquée par ces entités est indéterminée. Nous notons que le nom dénombré par ces unités est mis au cas accusatif et avec certaines unités (bidh3o) a toujours un genre opposé à celui du cardinal. Par exemple :

➢ حضر الاحتفال كذا رجلاً (plusieurs hommes ont assisté à la cérémonie)
haDaara *a l+ihtifaal+a* **num→-quant-*kaðaa+Ø* *rajulaa+n+***
(V)PASSE DEF+(N)+ACC (CARD)ACC+NEUTRE (N)+ACC+INDEF

3.2.4.1.7. La conjonction de coordination

La conjonction est une relation qui permet de lier des éléments de la même classe. Les conjoints partagent le même trait de définitude et portent la même marque de cas. Mentionnons que dans ce type de relation, la tête et le dépendant sont reliés par l'élément *« indications linguistiques »* qui est la conjonction de coordination. Nous pouvons représenter cette relation comme suit :

(N1)-(CONJ_COOR)-*conj-coord*→(N2) …(7)

Pour illustrer cette définition, prenons la phrase suivante :

➢ جاء التلميذُ والأستاذُ (Ils sont venus l'élève et le professeur)
Jaa+a *al+tilmid+u* ***-COORD→*** ***waa#*** *al+'ustaaD+u*
V(PASSE) **DEF** +(N)+**NOM** (Coord) **DEF**+(N)+**NOM**

Cette relation est valable aussi entre deux cardinaux. Les cardinaux de 21 à 99 sont composés d'une manière analytique par une coordination suivant la règle suivante :

(CARD)-(CONJ_COOR)-*conj-coord*→(CARD) …(8)

Exemple :

➢ خمسةٌ وخمسون طالبا (cinquante-cinq étudiants)
khamsat+u+n*-COORD→** ***waa# ***khams+u+n*** *talib+n+n*
(CARD) (**COORD**) (CARD) (**N**)+ACC+INDEF

3.2.4.2. Les relations syntaxiques gouvernées par un adjectif

Rappelons qu'un adjectif est un mot qui associe à un nom, auquel il s'adjoint, une caractéristique ou une qualité. Il partage avec le nom les catégories grammaticales suivantes :

- **Le genre** : masculin (MASC) et féminin (FEM).
- **Le nombre**: singulier (SG), duel (DUEL) et pluriel (PL)
- **Le cas** : nominatif (NOM), accusatif (ACC) et génitif (GEN).
- **La définitude** : défini (DEF), indéfini (INDEF) et neutre (NEUTRE)

De part ces catégories et du fait que l'adjectif est joint à un nom, des règles d'accord morphologiques entre l'adjectif et le nom s'imposent. Bien entendu, l'adjectif reçoit le genre, le nombre, le cas et la définitude par le phénomène de l'accord du support auquel il se rapporte. Les règles d'accord sont diversifiées et complexes; par exemple si le nom est un

pluriel *brisé*, l'adjectif sera au féminin singulier même si le nom est masculin comme c'est le ca de cette phrase :

- العقاربُ السامة (les scorpions venimeux)
  al+'aqaarib+u *al+saamat+u*
  DEF+(N)**MASC**.**PL**+NOM DEF+(ADJ)**FEM**.SG+NOM

Nous décrivons sommairement les principales règles de l'accord entre le nom et l'adjectif comme suit :

- Si le nom est un cas de référents humains, l'adjectif s'accorde en genre et nombre avec lui, par exemple :
  - محمدٌ خفيفُ الروحِ (Mohammed a un esprit léger)
    Muhammad+u+n [***khafiif+u***] *al+ruuH+i*
    (NP)**MASC**.**SG**+**NOM**+INDEF (ADJ)**MASC**.**SG**+**NOM** DEF+(N)+GEN

  - مليكةٌ خفيفةُ الروحِ(Malika a un esprit léger)
    Maliikat+u+n [***khafiifat+u***] *al+ruuH+i*
    (NP)**FEM**.**SG**+**NOM**+INDEF (ADJ)**FEM**.**SG**+**NOM** DEF+(N)+GEN

  - الرجالُ أخِفّاءُ الروحِ les hommes ont un esprit léger)
    al+rijaal+u [***'akhifaa'+u***] *al+ruuH+i*
    DEF+(N)**MASC**.**PL**+**NOM** (ADJ)**MASC**.**PL**+NOM DEF+(N)+GEN

  - النساءُ خفيفاتُ الروحِ les femmes ont (un esprit léger)
    al+nissa'+u [***khafiifAt+u***] *al+ruuH+i*
    DEF+(N)**FEM.PL**+NOM (ADJ)**FEM**.**PL**+NOM DEF+(N)+GEN

- Si le nom est un cas de référents non humain et au singulier alors l'adjectif doit s'accorder en genre et en nombre avec lui, c'est le cas des exemples suivants :
  - القط كثير المواء (le chat qui miaule beaucoup)
    al+qiT+u [***kathiir+u***] *al+miwaa'+i*
    DEF+(N)**MASC**.**SG**+**NOM** (ADJ)**MASC**.**SG**+**NOM** DEF+(N)+GEN

  - القطة **كثيرة** المواء (la chatte qui miaule beaucoup)
    al+qiTat+*u* [***kathiirat+u***] *al+ miwaa'+i*
    DEF+(N)**FEM.SG**+**NOM** (ADJ)**FEM.SG**+**NOM** DEF+(N)+GEN

- Si le nom est un cas de référents nom humain et il est au pluriel alors dans ce cas, le genre de l'adjectif est au féminin et elle est au singulier

  - القطط **كثيرة** المواء (les chats qui miaulent beaucoup)
    al+qiTaT+u [***kathiirat+u***] *al+miwaa'+i*
    DEF+(N)**MASC**.**PL**+**NOM** (ADJ)**FEM**.**SG**+**NOM** DEF+(N)+GEN

  - القطات **كثيرة** المواء (les chattes qui miaulent beaucoup)
    al+qiTaat+*u* [***kathiirat+u***] *al+ miwaa'+i*
    DEF+(N)**FEM.PL**+**NOM** (ADJ)**FEM.SG**+**NOM** DEF+(N)+GEN

3.2.4.2.1. Les relations syntaxiques de surface contrôlées par la valence de l'adjectif

Nous présentons dans cette section les relations syntaxiques gouvernées par l'adjectif. Cinq

relations syntaxiques sont représentées : le complément de l'adjectif, le modifieur, le comparatif, le superlatif et la conjonction de coordination. Pour chaque relation, comme nous l'avons fait pour le nom, nous donnons le dépendant prototypique avec des exemples à l'appui sans aborder exhaustivement les propriétés syntaxiques.

3.2.4.2.2. La relation complément de l'adjectif

Le complément de l'adjectif, noté compAdj, est un nom défini et fléchi au génitif. Il suit directement son gouverneur. Nous pouvons schématiser cette relation comme suit :

(ADJ)NEUTRE-*compAdj*→(N) …(9)

Voici quelques exemples pour illustrer cette relation :

- المناطقُ طاردةُ العقولِ (Les lieux expulsant les esprits)
  al+manaatiq+u ***taridat+u+Ø* -compAdj** → ***al+'uquul+i***
  DEF+(N)+NOM (ADJ)+NOM+NEUTRE DEF+(N)+GEN
- فاطمةُ طيبةُ القلبِ (Fatima a un bon cœur)
  Fatimat+u+n ***tayyibat+u+Ø* - compAdj→ *al+qalb+i***
  (N)+NOM+INDEF (ADJ)+NOM+NEUTRE DEF+(N)+GEN
- محمدٌ حسنُ الخلقِ (Mohammad a de bonne manières)
  Muhammad+u+n ***hassunn+u+Ø* -compAdj→ *al+khulq+i***
  (N)+NOM+INDEF (ADJ)+NOM+NEUTRE DEF+(N)+GEN

Pour des raisons syntaxiques et morphologiques, l'omission du complément de l'adjectif entraîne un changement de détermination et conduit à un changement de sens dans l'information. Pour illustrer cette propriété, nous omettons le complément d'objet des exemples donnés dessus, ce qui donne les résultats suivants :

- المناطقُ طاردةٌ(Les lieux expulsifs)
  al+manaatiq+u ***taridat+u+n***
  DEF+(N)+NOM (ADJ)+NOM+NEUTRE
- فاطمةُ طيّبةٌ (Fatima est gentille)
  Fatimat+u+n ***tayyibat+u+n***
  (N)+NOM+INDEF (ADJ)+NOM+NEUTRE
- محمدٌ حسنٌ (Mohammad est joli)
  Muhammad+u+n ***hassunn+u+n***
  (N)+NOM+INDEF (ADJ)+NOM+NEUTRE

3.2.4.2.3. La relation modificative

Dans le cadre l'adjectif, le dépendant ou le *modifieur* de la relation est un adverbe, comme le mettons en avant dans la règle suivante :

(ADJ)-*modif*→(ADV) …(10)

Prenons cette phrase comme exemple pour illustrer cette relation :

- البحارُ عميقةٌ جداً (les mers sont très profondes)
  al+bihar+u ***'amiirqt+u+n* -modif** → ***ʤidan***
  DEF+(N)+NOM (**ADJ**)+NOM+INDEF (**ADV**)

Cette relation peut avoir des variantes. Une de ces variantes consiste à associer une particule, qui est le dépendant dans ce cas, à un adjectif afin de nier l'information véhiculée par ce dernier. Cette variante est généralement suivie par une conjonction de coordination pour

ajouter un autre adjectif. Considérons la phrase suivante pour illustrer cette variante :

- رجلٌ لا غنيٌ ولا فقيرٌ (un homme n'est ni riche ni pauvre)

  rajul+u+n ***laa*←modif-g*haniy+u+n*** *wa#laa* *faqir+u+n*

  (N)+NOM+INDEF (**ADV**) (**ADJ**) (COORD)#(ADV) (ADJ)

3.2.4.2.4. La relation comparative

Comme toutes les grammaires, la grammaire arabe possède des constructions syntaxiques permettant d'exprimer la comparaison entre deux entités. Cette comparaison est définie par un adjectif comparatif nécessitant l'utilisation d'un schème « أَفْعَلُ», par exemple : (جميل → أَجْمَلُ) ou (قليل → أَقَلُّ), et suivi de la préposition (من), qui correspond à la préposition en français 'que'. Dans ce cas, le dépendant de la relation syntaxique est la préposition *min* (من) suivi d'un nom fléchi au génitif. La règle suivante schématise la relation de comparaison :

(ADJ)-*comparative*→min+(N)GEN …(11)

Exemple :

- 'الطائرةُ أسرعُ من القطارِ' (L'avion est plus rapide que le train)

  al+Ta'irat+u *ʔasra'+u* *min* *al+qiTar+i*

  DEF+(N)+NOM+INDEF ($ADJ_{comparative}$)+NOM (PREP) DEF+(N)+GEN+INDEF

Toutefois, l'utilisation des superlatifs n'est pas toujours possible pour certains mots. C'est le cas de certains verbes ou des phrases où nous voulons exprimer un degré supérieur des adjectifs de couleurs ou de particularités physiques ayant déjà la forme d'un élatif (إسم التفضيل). Dans ce cas-là nous faisons appel à un élatif à sens vague (أكبر أشد، أقل، أكثر،) suivi d'un nom indéfini au cas accusatif, de la couleur ou de la particularité physique. Par exemple :

- كريم أقلُّ صمماً من أنيس (Karim est moins sourd que Aniss)

  Karim+u+n ***ʔaqall+u*** ***samam+a+n*** *min* *ʔanii+u+n*

  (NP)+NOM (ADJ)+NOM (ADJ)+ACC+INDEF (PREP) (NP)+GEN

- الثلجُ أشدُّ بياضاً من اللبنِ (la neige est plus blanche que le lait)

  Al+Talj+u ***ʔachad+u*** ***bayadh+a+n*** *min* *al+laban+i*

  DEF+(N)+NOM (ADJ)+NOM (ADJ)+ACC+INDEF (PREP) DEF+(N)+GEN

3.2.4.2.5. La relation superlative

Généralement la comparaison se fait entre deux entités, mais pour pouvoir faire la différence entre un groupe contenant de nombreuses entités nous faisons appel au superlatif. Ce dernier est utilisé pour désigner les extrêmes que nous exprimons par les mots : le meilleur, le premier, le pire, le dernier, etc. Dans le cadre d'une relation superlative, le dépendant est par défaut un nom mis au génitif indéfini. Nous résumons cette relation dans la règle suivante :

(ADJ)NEUTRE-*supertlatif*→(N) …(12)

Le dépendant d'une relation superlative est par défaut un nom mis au génitif indéfini. De plus ce dépendant est appelé dans cette relation élatif (إسم التفضيل) et il est invariable en genre et en nombre. Il peut être employé comme premier terme d'une annexion, et dans ce cas-là deux constructions sont alors possibles :

- Construction faisant appel à un complément de nom singulier indéterminé
  - بلالٌ أجملُ رجلٍ (Bilel est le plus beau des hommes)

bilal+u+n ***ʔajmal+u-supertlatif→*** ***raʤul+i+n***
(NP)+NOM (ADJcomparative)+NOM (N)+GEN+INDEF

- Construction faisant appel à un complément de nom déterminé au pluriel
  - ‘الرِّبَا أَعْظَمُ الكَبَائِرِ’ (l’adultère est le plus grand des péchés capitaux)
    Al+riba’ ***ʔaDam+u*** ***al+kaba’ir+i***
    DEF+(N)+NOM ($ADJ_{superlatif}$)+NOM (N)+GEN+INDEF

Enfin, nous notons qu’en plus des cas mentionnés dessus, il existe aussi deux adjectifs qui ne sont pas sous la forme de 'أَفْعَل', qui sont les mots خَيْرٌ(bien) et شَرٌّ(mal). Ces mots sont utilisés sous cette forme pour exprimer le comparatif et le superlatif. Pour illustrer ces cas, voici quelques phrases explicatives :

- ‘الصَّلاَةُ خَيرٌ منَ النَّومِ’ (la prière est meilleure que le sommeil)
  Al+salat+u khayr+u+n min al+nawm+i
  DEF+(N)+NOM ($ADJ_{comparative}$)+NOM (PREP) DEF+(N)+GEN
- ‘الشَّيطانُ شَرُّ الخلقِ’ (le diable est la pire des créatures)
  Al+shaytAn+u ***charr+u*** ***al+nawm+i***
  DEF+(N)+NOM ($ADJ_{superlatif}$)+NOM DEF+(N)+GEN

3.2.4.2.6. La relation conjonction de coordination

La conjonction est une relation, que nous considérons dans cette partie, permet de lier deux adjectifs appelés aussi *conjoints*. Les conjoints ont aussi les mêmes propriétés que celles des conjonctions des noms, à savoir partager le même trait de définitude et porter la même marque de cas. Notons que dans ce type de relation, la conjonction de coordination (appelé aussi *indications linguistiques*) lie la tête et le dépendant de la conjonction. La règle suivante donne un schéma global de cette relation :

(ADJ)-(CONJ_COOR)-*conj-coord*→(ADJ) …(13)

- أيامٌ صعبةٌ وحزينةٌ(des jours durs et tristes)
  'ayam+un *sa'bat+u+n-**COORD→*** *waa#* *haziinat+u+n*
  (**N**)+NOM+INDEF (**ADJ**)+**NOM**+**INDEF** (Coord)
  (**ADJ**)+**NOM**+**INDEF**

- كان بلال فَرحاً وسعيداً(Bilel était content et heureux)
  kana *Billel* *farih+a+n* ***COORD→*** *waa#* *sa'iid+a+n*
  V(PASSE) (NP) (**ADJ**)+**ACC**+**INDEF** (Coord)
  (**ADJ**)+**ACC**+**INDEF**

La grammaire de tradition arabe permet une coordination effectuée d'une façon asyndétique, autrement dit ; une succession des adjectives sans l'utilisation d'une coordination. Cette fonction permet de générer une chaîne d'adjectif. Voici une phrase où cette fonction est mise en œuvre :

- كان بلال فَرحاً سعيداً(Bilel était content heureux)
  kana *Billel* *farih+a+n* *sa'iid+a+n*
  V(PASSE) (NP) (**ADJ**)+**ACC**+**INDEF** (**ADJ**)+**ACC**+**INDEF**

3.2.4.3. Les relations syntaxiques gouvernées par les mots outils

Dans cette section, nous présentons les relations syntaxiques régies par les mots outils (appelés aussi *les lexèmes fonctionnels*). Ces mots outils représentent les unités lexicales

autres que les trois classes majeures, à savoir : verbe, nom et adjectif. Les relations concernées par la présentation de cette section sont : l'interjection, la préposition, la conjonction, la jonction et l'exception.

3.2.4.3.1. L'interjection d'appel

Dans la grammaire arabe nous énumérons sept interjections d'appel qui sont : (يا - أيا - هيا - أي - الهمزة - وآ - وا). L'utilisation de ces interjections se fait dans une construction se composant de l'interjection suivie d'un nom, neutre ou défini, fléchi au nominatif. Toutefois dans certain cas, le nom qui suit l'interjection peut-être mis à l'accusatif indéfini. C'est le cas où ce nom est une annexion au singulier. La construction obtenue est définie donc comme suit : « interjection d'appel + interjeté (appelé) ». D'après les grammairiens arabes, cette construction décrit une phrase verbale, dont le *munaada* 'appelé' est un cas de *maf'uul bih* 'المفعول به' (complément d'objet direct) et le verbe d'appel, *ʔunaadii ou ʔad'uu* 'j'appelle' (أنادي ou أدعو) est supprimé et remplacé par l'interjection. Nous résumons cette relation dans la règle suivante :

(INTERJ)-*interj-appel*→(N)…(14)

Voici quelques exemples d'utilisation de l'interjection :

- أيها الإنسانُ (Oh l'homme)
  ʔ*ajuhaa*–interj-appel→ *al+'insan+u*
  (INTERJ) DEF+(N)+NOM
- يا سائقَ العربةِ (Oh conducteur du véhicule)
  yaa *interj*-appel→ *saa'iq+a* *al+'arabat+i*
  (INTERJ) (N)+**ACC**+**INDEF** **DEF**+(N)+**GEN**

3.2.4.3.2. La préposition

Une préposition est une unité lexicale faisant partie du groupe appelé (حروف الجر). Elle a une fonction syntaxique qui consiste à mettre un nom au cas génitif. Cette relation peut être présentée par la règle suivante :

(PREP)-prép→(N)GEN …(15)

Ces prépositions peuvent être réparties dans les groupes suivants :

- **Prépositions usuelles** : بِ، تَ، لِ، فَ، إِلَى، حَتَّى، عَلَى، عَنْ، فِي، لَدُنْ، لَدَى، مَعَ، مِنْ، مُنْذُ، مُذْ
- **Quasi-préposition (de temps et de lieu) :** إِبَّانَ، أَثْنَاءَ، إِزَاءَ، أَمَامَ، بَدَلَ، بَعْد، بَيْنَ، تَحْتَ، تِجَاهَ، تِلْقَاءَ، رَيْثَ، رَيْثَمَا، شَمَال، ضِدَّ، عَبْرَ، عِوَض، فَوْر، قبَالَ، قَبْلَ، قُدَّام، قَرَابَة، قُرْبَ، نَحْوَ، وَرَاءَ، وَسَطَ، يَسَارَ، يَمِيْنَ، جَنْب، جِوَارَ، حَالَ، حَوْلَ، حِيْنَ، خَارِجَ، خَلْفَ، خِلَالَ، دَاخِلَ، جَنُوبَ،شَرْقَ، مُقَابِل، جِهَة ، ناحِيَة
- **Locution prépositionnelles usuelles (préposition + nom)** : إِلَى آخِرِهِ، إِلَى جَانِبِ، بِجِوَارِ، بِحَضْرَةِ، بِخُصُوصِ، بِخِلَافِ، بِدُونِ، بِسَبَبِ، بِشَأْنِ، بِفَضْلِ، بِمَثَابَةِ، بِوَاسِطَةِ، بَيْنَ يَدَي، بِإِنْتِظَارِ، عَلَى جَهْلِ، عَلَى حِسَابِ، عَلَى ضَوْءِ، عَلَى عِلْمِ بِ، عَلَى غِرَارِ، عَلَى قَدْرِ، عَلَى مَتْنِ، عَلَى مُسْتَوَى، عَلَى يَدِ، عَلَى طَرِيْقِ، فِي أَثْنَاءِ، فِي إِطَارِ، فِي حَالَةِ، فِي حُدُودِ، فِي خُصُوصِ، مِنْ خِلَال، فِي شَأْنِ، فِي غُضُونِ، فِي نَظَرِ، لِأَجْلِ، لِصَالِحِ، مِنْ أَجْلِ، مِنْ بَعْدِ، مِنْ قِبَلِ، مِنْ قَبْلِ
- **Locution prépositionnelles d'interrogation (préposition + particule d'interrogation)** : إِلَى مَتَى، إِلَى أَيْنَ، إِلَى مَنْ، إِلَامَ، بِكَمْ، بِمَاذَا، بِمَ،بِمَا، حَتَّامَ، حَتَّى مَاذَا، حَتَّى مَنْ، عَلَى مَاذَا، عَلَى مَنْ، عَلَامَ، عَمَّنْ، عَنْ مَاذَا، عَمَّ، فِي مَاذَا، فِيمَنِ،فِي مَنْ، لِمَنْ، فِيمَا، لِمَاذَا، لَم، مُنْذُ مَتَى، مِنْ أَيْنَ، مِمَّا،مِمَّنْ

Prenons quelques exemples de phrases utilisant ces prépositions :

- الإعجاز في القرآن (le miracle dans le coran)
  Al+'iajaaz+u ***fii* –prép→ *al+qur'aan+i***
  DEF+(N)+NOM (PREP) DEF+(N)+GEN
- Un autre exemple !!!

3.2.4.3.3. La conjonction de subordination

Dans le cas des mots outils, une conjonction de subordination est assimilée à une unité lexicale faisant partie du groupe appelé en arabe *h□uruf nas□bi wa#ʤ azmi almud□aariʕ i* (حروف نصب وجزم الفعل المضارع). Cette unité gouverne un verbe suivi d'une relation *de conjonction*. Le rôle de cette conjonction est de permettre au verbe de remplir les fonctions du verbe ou celle du nom. Nous pouvons présenter cette relation par la règle suivante :

(CONJ_SUB)-*conj-sub*→(V) …(16)

Le verbe qui suit ce type de conjonction doit être conjugué à l'inaccompli et son cas dépend du type de la conjonction :

- Si la conjonction est du groupe (حروف نصب) alors le verbe sera au subjonctif. Les particule faisant partie de ce groupe sont : لَن *(lan)*، حَتَّى *(hatta)*، أَنْ *(ann)*، كَي *(kay)* ، لام التعليل *(laam al-ta'lil)*، إذا *('iDana)* ، لام الجحود *(laam al-juHud)* ،فاء السببية *(faa al-sababiya)*. Exemple :
  - يريدُ أَنْ يتزوجَ (Il veut se marier)
    Yurid+u ***'an–Conj_Sub→*** ***yatazawaj+a***
    (V)+PRESENT+IND (CONJ_SUB) (V)+ **PRESENT**+**SUBJ**

- Si la conjonction est du groupe (حروف جزم) alors le verbe sera à l'apocope. Certain particules faisant partie de ce groupe, qui sont : **لَمْ** *(lam)* ، **لَمّا** *(lamma)*، **لام الأمر** *(laam al-amr)* ، **لا الناهية** *(laam al-nahiya)*, causent l'élision d'un verbe mais d'autres, en occurrence : إنْ ('in) ، مَنْ *(man)*، مَا *(maa)* ،متى *(mata)* ، مهما *(mahma)* ،أيّان *('ayyana)*، أين *('ayna)*، أنّى *('anna)*، حيثما *(hayTyma)*، أيّ *('ayy)*, causent l'élision de deux verbes. comme dans la phrase suivante :
  - لا تسافرْ
    laa–Conj_***Sub→*** ***tusaafir***
    (CONJ_SUB) (V)+ **PRESENT**+**APOC**

  - مَنْ يدرسْ ينجحْ (Celui qui étudies réussiras)
    Man –Conj_Sub→ yadrus ***yanʤah***
    (CONJ_SUB) (V1)+ **PRESENT**+**APOC** **(**V2)+ **PRESENT**+**APOC**

3.2.4.3.4. La conjonction de coordination

La jonction est la relation permettant d'établir un lien de coordination entre plusieurs entités. Elle est manifestée par l'utilisation d'une particule, parmi les particules de la coordination, qui se place entre le *coordonné à lui* (**مَعْطُوفٌ عَلَيْهِ**) et *le coordonné* (**مَعْطُوفٌ**). Notons que le *coordonné* s'accorde avec le *coordonné à lui* dans son cas : nominatif (**رَفْع**), accusatif (**نَصْب**), génitif (**جَرّ**) et apocope (**جَزْم**). La règle suivante présente cette relation :

(CONJ_COOR)-*conj-coord*→(V) …(17)

Les particules de coordination sont traditionnellement nommées *huruuf alʕatfi* (حروف العطف), et selon les grammairiens arabes il existe neuf particules de coordination :

- ***waa* (et) :** peut exprimer plusieurs sens : i) la successivité sans référence à un

intervalle temporel, et ii) la simultanéité.

- ***faa (ensuite)*** **الفـــاء** : exprime un ordre séquentiel sans intervalle temporel. il indique un enchaînement entre deux actions afin de mettre en avant l'ordre seulement.
- ***θumma (ensuite) :*** **ثُمّ** utilisé pour indiquer un ordre séquentiel avec intervalle temporel entre le *coordonné à lui* et le *coordonné.*
- ***hattaa (y compris, même)*** **حتّى** : utilisé pour faire la coordination dans le but d'exprimer l'objectif.
- ***ʔaw (ou)*** **'أوْ'** : c'est le connecteur standard de disjonction
- ***ʔam (ou exclusif)*** **'أَمْ'** : c'est aussi un connecteur de disjonction mais contrastif utilisé généralement pour lier des propositions interrogatives
- ***lakin (mais) :*** **'لكن'** utilisé pour coordonner des constituants non verbaux
- ***bal (plutôt)*** **'بَل'** : c'est un connecteur de rectification liant une proposition affirmative à une proposition négative

Voici quelques exemples de ces conjonctions :

- وزهيرٍ بخالدٍ مررتُ (*Je suis passé auprès de Khalid et Zuhaïr)*
  Marart+u bi# *khalid+i+n* ***COORD→waa#*** **zahiir+i+n**
  V(PASSE) PRE# (NP)+**GEN** Coord (NP)+**GEN**

- دخل الطلابُ فالأستاذُ (les élèves sont rentrés ensuite l'enseignant)
  dakhala *al+Tulab+u* ***COORD→*** ***faa*** ***al+'ustadh+u***
  V(PASSE) **DEF**+(N)MAS.PL+**NOM** Coord **DEF**+(N)MAS.PL+**NOM**

3.2.4.3.5. L'exception

L'exception est exprimé en arabe à travers l'utilisation de la particule 'إلا', qui signifie sauf ou hormis, dans une phrase affirmative gouverne en général le cas direct. L'emploi de l'exception nécessite l'engagement de deux éléments importants : l'entité exceptée (المستثنى) mise en générale à l'accusatif et le terme général (المستثنى منه). La relation d'exception peut être représentée par la règle suivante :

(EXCEP)-*excep*→(N) …(18)

Deux cas d'exception en arabe peuvent être distingués :

i. Cas où elle est mise à l'accusatif, quand la proposition est à l'affirmatif, et quand le terme général est mentionné.
   - جاء الأولادُ إلا محمداً (tous les enfants ne sont pas venus sauf Mohammed)
     Jaa *al+'awlad+u* ***EXCEP→*** ***'ilaa*** ***Mohammed+a+n***
     V(PASSE) **DEF**+(N)MAS.PL+**NOM** Excep (N)MAS +**ACC**

ii. Cas où elle est mise soit à l'accusatif, soit au même cas que le terme général, et cela quand la proposition est négative, et le terme général est mentionné, par exemple :
   - ما جاء الأولادُ إلا محمداً (tous les enfants ne sont pas venus sauf Mohammed)
     Maa *jaa* *al+'awlad+u* ***EXCEP→*** ***'ilaa*** ***Mohammed+a+n***
     PART V(PASSE) **DEF**+(N)MAS.PL+**NOM** Excep (N)MAS +**ACC**

   - ما جاء الأولادُ إلا محمدٌ (tous les enfants ne sont pas venus sauf Mohammed)
     Maa *jaa* *al+'awlad+u* ***EXCEP→*** ***'ilaa*** ***Mohammed+u+n***

PART V(PASSE) **DEF**+(N)MAS.PL+**NOM** Excep (N)MAS +**NOM**

3.2.4.4. Les relations syntaxiques gouvernée par le verbe

Rappelons que le verbe en arabe permet d'exprimer comme dans toute langue une action effectuée par un sujet. Cette action peut être effectuée sur un axe de temps allant du passé au futur en passant par le présent en prenant différentes formes : passive et active. Elle peut aussi être utilisée pour exprimer des ordres dans un mode impératif tout. De ce fait la conjugaison d'un verbe prend en compte plusieurs paramètres :

- Le temps (passé, présent, futur)
- Le mode (indicatif, subjonctif, apocope, impératif)
- La voix (passif, actif)
- Personne ($1^{ère}$, $2^{ème}$, $3^{ème}$)
- Genre (masculin, féminin)
- Nombre (singulier, duel, pluriel)

Nous pouvons schématiser ces paramètres dans le tableau suivant :

<table>
<tr><td colspan="8">Les grammèmes du verbe</td></tr>
<tr><td>Voix</td><td colspan="4">Actif</td><td colspan="3">Passif</td></tr>
<tr><td>Personne</td><td colspan="3">$1^{ère}$</td><td colspan="2">$2^{ème}$</td><td colspan="2">$3^{ème}$</td></tr>
<tr><td>Genre</td><td colspan="4">Masculin</td><td colspan="3">Féminin</td></tr>
<tr><td>Nombre</td><td colspan="3">Singulier</td><td colspan="2">Duel</td><td colspan="2">Pluriel</td></tr>
<tr><td rowspan="2">Mode</td><td rowspan="2">Jussif</td><td rowspan="2">Impératif</td><td rowspan="2">Subjonctif</td><td colspan="4">Indicatif</td></tr>
<tr><td>passé</td><td>présent</td><td>futur</td></tr>
</table>

Tableau 3. 5. Les paramètres de conjugaison d'un verbe.

En plus des aspects de conjugaison, il existe un concept, emprunté aux chimistes, et projeté sur les verbes qui est *la valence*. A l'origine, la valence correspond au nombre d'atomes avec lequel un atome donné peut se combiner à l'intérieur d'une molécule. C'est Tesnière qui a adapté ce concept au verbe (Tesnière, 1965), et cela en donnant la définition suivante : « le nombre d'actants qu'un verbe est susceptible de régir » en considérant les actants comme : « êtres ou les choses qui, à un titre quelconque et de quelque façon que ce soit, même au titre de simples figurants et de la façon la plus passive, participent au procès ». Pour les circonstants, il s'agit de : « circonstances de temps, lieu, manière, etc. dans lesquelles se déroule le procès ». En détails, cette définition oppose les actants aux circonstants et distingue parmi les actants, le prime actant, le second actant et le tiers actant que nous pouvons assimiler au sujet, l'objet et le complément d'objet (COI) respectivement.

Par ailleurs, en fonction du nombre d'actants régis par un verbe, Tesnière propose une typologie de valence pouvant elle-même considérée comme une classification de verbe comme suit :

- *Verbe avalent* : verbe n'ayant pas d'actants, donc sans aucune valence. Ces verbes sont plus souvent connus sous le nom de verbes impersonnels.
- *Verbe monovalent* : verbe ayant un seul actant et connus sous le nom de *verbes neutres* ou de *verbes intransitifs*.
- *Verbe bivalent* : verbe à deux actants appelé aussi verbe *divalent*.

- *Verbe trivalent* : représente la classe des verbes ayant trois actants.

Dans le reste de cette section nous détaillons les relations syntaxiques gouvernées par un verbe.

3.2.4.4.1. Relation Sujet {(V) -sujet→(N)}

Selon la théorie de valence présentée dessus, le sujet correspond au *prime actant.* Dans la grammaire arabe, le dépendant de la relation syntaxique est un nom mis au cas nominatif. Le type de relation construite est *'SV'* et suit la règle suivante :

(V) -sujet→(N) …(01)

Le sujet s'accorde en genre et en nombre avec un sujet pronominal (pronom précédant le verbe) et en genre seulement avec un sujet lexical. De ce fait, l'identification du pronom est déterminée par l'accord établi entre le sujet et le verbe. Aussi, le verbe est mis au duel ou au pluriel quand le sujet est un *pronom personnel*; et il reste au masculin singulier s'il est encadré par une lettre d'exclusion. Pour illustrer ces cas d'accord, prenons les exemples suivants :

- تخرَّج (il est diplômé)
  *takharraj**a*** **—sujet→** ***{huwa}***
  (V)PASSE.**MASC.SG** (PRO)**MASC.SG**
- تخرَّجوا (Ils sont diplômés)
  *takharraj**uu*** **—sujet→** ***{humu}***
  (V)PASSE.**MASC.PL** (PRO)**MASC.PL**
- تخرَّج الولدُ (Le garçon est diplômé)
  *takharraj**a*** **—sujet→** ***alwaladu***
  (V)PASSE.**MASC.SG** (N,**masc**)**SG**
- تخرَّج الأولادُ (Les garçons sont diplômés)
  *takharraj**a*** **—sujet→** ***alʔawlaadu***
  (V)PASSE.**MASC.SG** (N,**masc**)**PL**
- ما ذهب إلا البنات (Il n'est parti que les filles)
  *Ma Dahab**a*** ***'ilaa*—sujet→** ***alʔawlaadu***
  (NEG) (V)PASSE.**MASC.SG (PART**$_{exclusion}$**)** (N,**femi**)**PL**
- ذهبا (Ils sont partis)
  *Dahab**+A*** **—sujet→** ***{humA}***
  (V)PASSE.**MASC.DL** (PRO)**MASC.DL**

Sur un autre registre, le sujet peut prendre différents genres parmi ceux données dans ce tableau avec des exemples d'illustration :

Type de sujet	**Exemple**
Nom propre	تدرسُ فاطمةُ (Fatima étudie) tadrusu -**sujet→** ***faatimat+u+n*** (V)PRESENT (NP)+NOM+INDEF
Nom commun	تدرسُ الطالباتُ (Les étudiantes étudient) ***tadrusu*** **- sujet→** ***al+taalibaat+u*** (V)PRESENT DEF+(NC)+NOM
Pronom démonstratif	*كان هذا رائعا* (Cela est magnifique) ***kaana*-sujet→** ***haðaa*** *raa'i'a+a+n* (V)PRESENT (DEI) (ADJ)+ACC+INDEF
Cardinal	نجحَ خمسةُ طلابٍ (Cinq étudiants se sont réussis)

	*najaha***-sujet→** [*khamsat+**u*** *tulaab+i+n*] (V)PASSE.**MASC** (CARD)+**NOM** (N,**masc**)+GEN+INDEF
Pronom Relatif	يَوَدُّ **الَّذِينَ** كَفَرُواْ لَوْ كَانُواْ مُسْلِمِينَ (Ceux qui ne croient pas, veulent s'ils étaient musulmans) ***jawaddu-sujet→ allaðiina*** *kafaruu {humu}* *law* *kaanuu {humu}* *muslimiin* (V)PRESENT (**Pron-Rela**) (V)PASSE.MASC.PL (CONJ) (V)PASSE (N)MAS+PL+GEN

Notons que le sujet peut dans certains cas être omis, c'est le cas des verbes intransitif mis à la voix passive. Par exemple, les verbes نام (dormir) et نشأ (grandir) prennent en voix active un sujet et n'ont pas de complément d'objet direct, cependant ce sujet est omis lorsque ces verbes sont transformés à la voix passive comme suit :

نام (dormir)	
Forme Active	نام الطفلُ على السرير (L'enfant a dormi sur le lit) *Naama* **—sujet→** *al+Tifl+u* *'ala* *al+sarir+i* (V)**ACTIVE**.PASSE DEF+(N)+NOM (PREP) DEF+(N)+GEN
Forme Passive	نيم على السرير (Il a été dormi sur le lit) *Niima* *'ala* *al+sarir+i* (V)**PASSIVE**.*PASSE* (PREP) DEF+(N)+GEN
نشأ (grandir)	
Forme Active	نَشَأَ الولدُ على الطاعةِ (L'enfant a grandi sur la docilité) *Nacha'a* **—sujet→** *al+walad+u* *'ala* *al+taa'at+i* (V)**ACTIVE**.PASSE DEF+(N)+NOM (PREP) DEF+(N)+GEN
Forme Passive	نُشِئَ على الطاعةِ (Il a été grandi sur la docilité) *Nuchi'a* *'ala* *al+taa'at+i* (V)**PASSIVE**.PASSE (PREP) DEF+(N)+GEN

3.2.4.4.2. Le complément d'objet direct

Quand le verbe est transitif le complément d'objet représente principalement le 2ème actant sémantique, et il est appelé en arabe *maf'uul bih (مفعول به)*. Il est par défaut un nom sans préposition fléchi à l'accusatif. Le type de la relation est de la forme 'VC' que nous pouvons résumer dans la règle suivante :

(V)–COD→(N)ACC …(02)

Prenons la phrase suivante pour illustrer cette relation :

- نشرَ الصحفيُّ المقالَ (Le journaliste a publié un article)

nashara *al+sahafiyy+u* ***al+maqaal+a***

COD

(V)PASSE DEF+(N)+NOM DEF+(N)+ACC

Parfois un complément d'objet commute avec un pronom objet réalisé sous forme d'un clitique ou d'un élément détaché, la forme clitique étant plus fréquente en ASM. Nous illustrons cette propriété dans les phrases suivantes :

- ضربه الأولاد (les enfants l'ont frappé)
  Daraba **-coDir→#** ***hu*** *al+ʔawlaad+u*
  (V)PASSE (PRO) DEF+(N)+NOM

- 'نعبد إيّاه' (nous l'adorons en personne)
  *naʕbudu {nah□nu }***-coDir→** ***ʔija#hu***
  (V)PRESENT (PRO)

En arabe, un verbe transitif peut avoir besoin de plus d'un complément d'objet dans une phrase : certains verbes régissent deux compléments et d'autres trois. La structure des phrases utilisant ce type de verbe est généralement de la forme un verbe, un sujet, un complément d'objet 1 et éventuellement un complément d'objet II. Cette structure peut même être étendue à un troisième complément dans le cas des verbes : أنبأ *('anba'a – informer)* , نَبَّأ *(nabba'a – annoncer)* , أخبرَ *('akhbara – apprendre)* , خبّر *(khabbara - avertir)* ,حدّثَ *(haddatha – raconter)* , أعلمَ *('a'lama – faire part)* , أرى *('ara – montrer)*. Voici une phrase employant un verbe utilisant trois compléments :

- أريتُ سعيداً الأمرَ واضحاً (j'ai démontré à Said la question clairement)

'araytu ***Sa'id+a+n*** ***al+'amr+a*** ***waDih+a+n***

COD1

(V)PASSE (NP)+**ACC** DEF+(N)+**ACC** (N)+**ACC**+INDEF

Les verbes nécessitants deux compléments sont répartis dans différentes classes :

- ***Les verbes attributifs :*** cette catégorie a besoin de deux compléments où le second est appelé *attributif* et a pour rôle de compléter le sens de la phrase. Par conséquent, il n'est pas possible de supprimer l'un des deux compléments. Trois types sont considérés dans cette classe : pensée, changement d'état et certitude (ou de doute). Nous pouvons résumer la liste des verbes de cette classe dans le tableau suivant :

Verbe transitif		
Verbe de pensée	Verbe de changement d'état	Verbe de certitude (de doute)
ظن، خال، حسب، زعم، جعل *(zanna, khala, hasiba, za'ima, ja'ala)* *(croire, prétendre, supposer)*	صَيَّر، رَدّ، تَخِذَ، اتخذ، حَوَّل، جعل *(sayyara, radda, takhida), 'ittakhada, hawwala, ja'ala)* *(rendre, laisser, prendre, faire)*	رأى، علم، وجد، ألفى *('ara, 'alima, wajada, alfa)*

- ***Les verbes non-attributifs :*** sont des verbes ayant deux compléments mais ne comportant pas une phrase nominale sous-jacente. L'un des deux compléments est indispensable pour compléter le sens de la phrase ce qui n'est pas le cas de l'autre. Par conséquent, nous pouvons supprimer l'un des deux sans toucher à l'intégrité du sens de la phrase. Parmi les verbes de cette catégorie, nous citons : أعطى (*a'ta – donner*) ، سأل *(sa'ala – demander)*، منح *(manaha – octroyer)* ، منع *(mana'a – interdire)* ، ألبس/كسا *(kasa/'albassa – habiller)*، علّم *('allama – enseigner)*

Avant de détailler les différents types de complément considérés dans notre étude, nous terminons cette présentation par les exemples d'utilisation suivants :

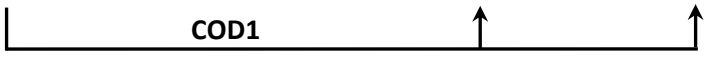
➢ اتخذ الله إبراهيم خليلا (Dieu a pris Ibrahim pour ami privilégié)

ItakhaDa *Allah+u* *ibrahi+m* *khalil+a+n*

(V)PASSE (NP)+NOM (NP)+ACC (N)+ACC+INDEF

➢ كسا اللهُ العظامَ لحماً (Dieu a revêtu les os de chaire)

Kassa *Allah+u* *al+'iDam+a* *lahm+a+n*



(V)PASSE (NP)+NOM DEF+(N)+ACC (N)+ACC+INDEF

Types du complément d'objet direct

Le complément d'objet direct ne peut pas être seulement un nom, il peut prendre d'autres valeurs syntaxiques comme l'adjectif, l'attribut, l'adverbe, etc. Dans notre étude nous avons considéré les types suivants :

- **Complément circonstanciel :**

Les compléments de ce type ont pour rôle la description des circonstances selon lesquelles ou dans lesquelles se déroule l'action décrite par le verbe qui le gouverne. Ils correspondent principalement à un adverbe circonstanciel de temps ou de lieu. Cette relation correspond généralement au phénomène dit dans la grammaire arabe *maf'uul fiih*, (**مَفْعُولٌ فِيهِ**). Nous distinguons deux classes de compléments circonstanciels :

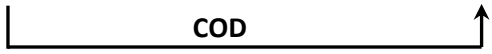
  - *Complément circonstanciel de temps (*pour exprimer le *: (ظَرْفٌ زَمَانٌ* contexte temporel dans lequel l'action véhiculée du verbe s'est exécutée, en : Par exemple .'*?quand*' e à la questiond'aure termes il permet de répondr
    - سَافَرَ الرّجلُ لَيْلاً (Il a voyagé de nuit)
  - *Complément circonstanciel de lieu (ظَرْفٌ مَكَانٌ) :* ce complément est utilisé pour situer où l'action du verbe s'est déroulée dans l'espace, ce qui constitue une réponse à a question '*où ?*'. C'est le cas de la phrase :
    - وَجَدَهُ خَارِجَ ٱلْبَيْتِ (Il l'a trouvé en dehors de la maison)

- **Complément de manière ou d'état**

Ce complément est utilisé pour exprimer les conditions ou les circonstances qui régissent le moment où l'action du verbe a eu lieu. Il décrit donc un état transitoire ou permanent dans un adjectif ou un nom au cas direct indéfini et accordé en genre et en nombre le verbe qu'il ne précède jamais. Ce phénomène est noté en arabe *ḥāl (حَالٌ)*, et en voici un exemple d'utilisation :

➢ سافر الولدُ مسروراً (Le garçon a voyagé content)

Safara *al+walad+u* ***masrur+a+n***



(V)PASSE DEF+(N)+NOM (ADJ)+**ACC**+**INDEF**

- **Complément absolu**

Ce complément joue pratiquement le même rôle que celui du complément de manière, la différence c'est que le complément absolu est un مصدر *(maṣdar)* du verbe employé afin de renforcer l'action et décrire la manière de son déroulement. Il est connu sous le nom de *maf'uul muTlaq* (مفعول مطلق). Il convient de mentionner, que la grammaire arabe ne possède pas d'adverbes proprement dit. La phrase suivante donne un exemple d'application pour ce

type de complément :

- فَهِمَ فَهْمًا (Il a parfaitement compris)

fahima ***fahma+a+n***

COD

(V)PASSE (N)+**ACC**+**INDEF**

- **Complément de but ou de cause**

Ce type de complément est utilisé pour exprimer la cause et l'objectif à travers un nom verbal (maṣdar) indéfini à l'accusatif, mentionné après le verbe qui le gouverne. Il permet de répondre à la question '*pourquoi ?*'. Dans la grammaire arabe, ce complément est connu sous le nom de maf'uul li'ajlihi (مَفْعُولٌ لِأَجْلِهِ). Voici un exemple d'utilisation de ce complément :

- صَاحَ أَلَمًا (*Il a crié à cause de la douleur*)

saaha ***alam+a+n***

COD

(V)PASSE (N)+**ACC**+**INDEF**

- هاجر طلبًا للأمنِ (*Il a voyagé en demandant la sécurité*)

haajara ***Talab+a+n*** ***l+al+amn+i***

COD

(V)PASSE (N)+**ACC**+**INDEF**

(PREP)+DEF+(N)+G

- **Complément de nature (spécificatif)**

C'est un substantif (nom ou masdar) utilisé pour spécifier et déterminer le terme ou la proposition qui le précède. C'est un terme indéfini et mis à l'accusatif et peut être exprimé au moyen de l'annexion ou de la préposition 'من'. Nous trouvons le spécificatif essentiellement après les entités lexicales suivantes :

- ✓ *Un verbe à sens vague* : le spécificatif permet de spécifier l'étendu ou le périmètre du verbe utilisé, par exemple :
  - فاضت العين دمعاً ('œil a débordé de larmes)

    faad□ati al+'ayn+u ***dam'+a+n***

    (V)PASSE DEF+(N)+NOM (N)+**ACC**+**INDEF**
- ✓ *Un élatif à sens général* : pour ce cas nous utilisons généralement le schème أَفْعَلٌ pour former l'élatif.
  - كان محمدٌ أكثر تواضعًا (Mohammed était le plus modeste)

    kaana *Mohammed+u+n* ***ʔakθar+a*** ***tawaaDu'+a+n***

    (V)PASSE (N)+NOM+INDEF (ADJ)+ACC (N)SG+ACC+INDEF
- ✓ *Un nom de mesure ou de poids :* comme son nom l'indique, ce spécificatif désigne une mesure ou un poids comme dans la phrase suivante :

➢ اِشْتَرَى رَطْلاً قمحًا (Il a acheté un demi-kilo de blé)
ʔishtara raTl+a+n qamh+a+n
(V)PASSE (N)+ACC+INDEF (N)+ACC+INDEF

✓ *Un nom de nombre* : nous illustrons cette utilisation dans la phrase suivante :
➢ جاء عشرون شخصاً (vingt personnes sont venus)
Jaa ʻichr+u+n chakhs+a+n
(V)PASSE (CARD)+NOM+INDEF (N)+ACC+INDEF

- **Complément d'objet direct sans verbe**

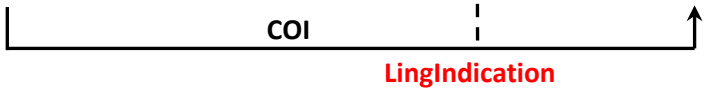
C'est des compléments utilisés dans des phrases et expressions utilisées couramment. Il s'agit d'expressions courantes dans lesquelles le verbe est sous-entendu. C'est le cas des phrases :
➢ شُكْرًا ! ⇔ Merci !
➢ مَهْلاً ! ⇔ Doucement !
➢ عَفْوًا ! ⇔ Pardon !
➢ أَهْلاً وَسَهْلاً ! ⇔ Bienvenu !

3.2.4.4.3. Le complément d'objet indirect

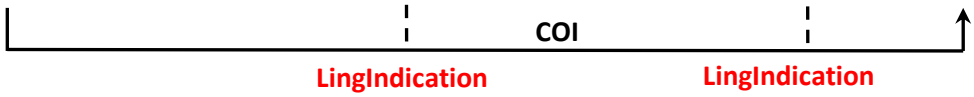
Si nous nous référons à la théorie de valence introduite au début de cette section, le complément d'objet indirect correspond au 3ème actant sémantique. Le dépendant de cette relation syntaxique est un nom fléchi au cas génitif relié avec le verbe (tête) par une constituant prépositionnel '(PREP)-prép→(N)GEN' appelé *linguistique Indication*. Ces prépositions peuvent aussi se succéder dans une phrase. Cette relation peut être présentée par la règle suivante :

(V)–COI→(PREP)|(N)GEN …(03)

Voici quelques exemples d'utilisation du complément d'objet indirect :
➢ صلّى المسلمون على الرسولِ (Les musulmans faisaient des prières pour le prophète)
Salla *al+muslimu+u+na* ***ʕalaa*** ***al+rasuul+i***
(V)PASSE DEF+(N)+NOM (PREP) DEF+(N)+GEN



➢ سافر الرئيس من دولةٍ إلى دولةٍ (le président s'est déplacé d'un pays à un autre)
saafara *al+ra'iis+u* ***mina*** ***dawlat+i+n*** ***ʕalaa*** ***dawlat+i+n***
(V)PASSE DEF+(N)+NOM (PREP) (N)+**GEN**+INDEF (PREP) (N)+**GEN**+INDEF



3.2.4.4.4. L'agent prépositionnel

Dans ce type de relation, on souligne un emploi des tournures modernes du passif qui expriment le complément d'agent, parmi ces locutions prépositionnelles : *min taraf+i, ʕalaa jad+i, min qibal+i* (' مِنْ طَرَفِ، مِنْ جَانِب، عَلَى يَد، مِنْ قِبَل ***de la part de'***). L'agent prépositionnel de cette relation syntaxique est le nom mis au génitif relié par l'une de ces locutions prépositionnelles. Il faut souligner que l'utilisation de ces types de tournures nécessite le changement des rôles sémantiques : le nom mis au génitif après la préposition est l'agent de

l'action représentant le premier actant sémantique du verbe, et le nom mis après le verbe est le complément d'objet correspondant au deuxième actant sémantique. L'usage de ce type de relation était limité dans l'arabe classique ce qui n'est pas le cas avec l'ASM où ces tournures sont très répandues dû probablement au contact et influence des langues indo-européennes (ref Dina). Nous pouvons formaliser cette relation par la règle suivante :

(V)PASSIF–agent-prép→(PREP) …(04)

Les phrases suivantes présentent des cas d'application de cette relation :

- أُعْتُقِلَتْ الفتاة **عَلَى يَد** الشرطة (La jeune fille a été arrêtée par la police)
  u'tuqilat *al+fatat+u* [***min*** *yad+i*] *al+churTat+i*
  (V)PASSIF.PASSE DEF+(N)+NOM [(PREP)**Loc_PREP** DEF+(N)+GEN

- يُموّل باريس سان جرمان **مِنْ طَرَفِ** الخليفي (Le Paris Saint Germain est sponsorisé par Al-Khelaïfi)
  Yumawwalu *Paris San Jirman* [***min*** *taraf+i*] *Al-Khelaïfi*
  (V)PASSIF.PASSE (NP)+NOM [(PREP)**Loc_PREP** (NP)+GEN

3.2.4.4.5. L'attribut

Par définition, une **copule** en linguistique correspond à un mot dont la fonction est de lier l'attribut au sujet d'une proposition. En arabe cette copule peut être assimilée à un exposant verbal, mis en tête de phrase. Ces verbes sont kana كان'Kana' et ses analogues (أصبح، أضحى، ظل، أمسى، بات، ما برح، ما انفك، ما زال، ما فتئ، ما دام، صار، ليس). Les phrases utilisant ces verbes sont incomplètes si nous nous contentons seulement d'un nom au nominatif, appelé aussi sujet (isme kana), en plus du verbe, elles nécessitent un autre élément, qui est *χabar* de *kaana*, pour assurer la cohérence grammaticale de la proposition. D'ailleurs, c'est pour cette raison que ces verbes sont appelés *verbes incomplets*. Le *χabar* de *kaana* correspond littéralement à l'information attribut du sujet (الخَبَر) et elle est un adjectif indéfini fléchi à l'accusatif. Il peut être aussi une préposition, proposition au présent de l'indicatif. Il s'accorde en genre et en nombre avec son sujet (المبتدأ). La règle suivante résume cette relation :

(V)–attr→(ADJ) …(05)

- **كان** الرجلُ قوياً (l'homme était fort)
  kaana *al+rajul+u* ***qawiy+a+n***
  (V)PASSE DEF+(N)**MAS.SIG**+NOM (ADJ)**MASC.SING**+ACC+INDEF

- كان الرجالُ أقوياءً (les hommes étaient forts)
  kaana *al+rijaal+u* ***'aqwiyaa'+a+n***
  (V)PASSE DEF+(N)**MAS .PL**+NOM (ADJ)**MASC.PL**+ACC+INDEF

- كان الرجلُ في المسجدِ (l'homme était dans la mosquée)
  a.***kaana*** *al+rajul+u* [*fii* ***al+masjid+I*** **]**
  (V)PASSE DEF+(N)MAS.SIG+NOM (PREP) DEF+(N)MASC.SING+GEN

- كان الرجلُ يصلي (l'homme était en train de prier)
  a.***kaana*** *al+rajul+u* [*fii* ***al+masjid+I*** **]**

(V)PASSE DEF+(N)MAS.SIG+NOM (PREP) DEF+(N)MASC.SING+GEN

Chapitre 4 Identification et typage des entités nommées

Introduction

Dans le but d'aborder la problématique d'extraction d'information un ensemble de conférences a été initié en 1987 sous l'intitulé *Message Understanding Coferences* (MUC). Ces conférences ont été financées par l'agence pour les projets de recherches avancées de défense DARPA (*Defense Advanced Research Projects Agency*). Le but de ces conférences est de rassembler le maximum d'efforts autour des problématiques d'extraction et de la compréhension automatique des messages, et notamment dans le domaine militaire, et d'évaluer les solutions proposées à travers l'organisation de compétitions entre les participants autours d'un corpus d'entrainement et un autre de test.

Lors des deux premières conférences, à savoir MUC 1 et MUC 2, l'objectif était d'explorer le terrain de recherche et d'aborder un certain nombre d'axes de recherche. Ces conférences ont abouti à la définition des principales tâches à faire dans le cadre d'une opération d'extraction. S'en suit trois autres conférences MUC 3, 4 et 5 qui ont mis l'accent sur les différentes tâches définies lors des précédentes conférences. Ces conférences ont contribué à un développement sophistiqué des différentes tâches d'analyse ce qui les a rendues plus complexes et a créé ainsi la nécessité de fragmenter chacune de ces tâches en des fonctionnalités indépendantes et plus maîtrisables. Les deux conférences MUC 6 et 7 ont repris ce besoins d'affinement des tâches en fonctionnalités indépendantes ce qui a donné naissance à de nouvelles tâches et à la transformation de certains modules impliqués dans le processus d'extraction en modules indépendants d'analyse de textes, ce qui a amené à la tâche *de reconnaissance des entités nommées (Named Entities)*. Ces conférences, et notamment la MUC 6, ont fait énormément de progrès au niveau du traitement de ce type d'entités avec des performances et des taux de précisions assez élevés lors de l'évaluation. D'autres conférences en parallèle autour de l'extraction des entités nommées ont eu lieu, comme la *Multilingual Entity Task* (MET) qui a fait émerger des systèmes de reconnaissance d'entités nommées pour l'espagnol, le japonais et le chinois.

Dans ce chapitre, nous nous intéressons au traitement des ENs en arabe. Un système de détection et de typage d'ENs pour l'arabe a été développé. Ce chapitre est consacré à la problématique de repérage et typage des entités nommées en arabe. La suite du chapitre est organisée comme suit.

Nous commençons par présenter la typologie des entités nommées ainsi que les principales applications qui utilisent les entités nommées dans les sections 4.1 et 4.2 respectivement. La section 4.3 est consacrée à exposer les particularités de la langue arabe liée à la détection des entités nommées. La section 4.4 est dédiée à présenter un aperçu sur les travaux réalisés sur les systèmes de reconnaissance des entités nommées en arabe. Notre approche de détection et de typage des entités nommées est décrite dans la section 4.5. La section 4.6 est consacrée a détailler la méthode de reconnaissance des noms propres (ENAMEX) de type personne, lieu et organisation. Nous présentons dans la section 4.7 la reconnaissance des expressions numériques (NUMEX).

4.1. Typologie des entités nommées

L'intérêt des entités nommées réside dans le fait qu'elles sont présentes et fréquentes dans tous les textes, tous types confondus, quel que soit le domaine. Elles constituent ainsi un aspect essentiel à prendre en compte dans le traitement et l'extraction de l'information contenue dans un texte. Dans la réalité, l'analyse d'un contenue de texte vise en général à détecter les actants ainsi que les coordonnées les événements relatés. C'est le cas des analyses de messages militaires ou des dépêches journalistiques portants sur des actes terroristes,

économiques, etc. Lors de la conférence MUC-6, l'extraction et la reconnaissance des entités étaient focalisées sur les trois types d'entités suivants :

- **NAMEX :** cette classe contient les noms propres qui peuvent être classé dans l'une des catégories suivantes :
  - *Personnes :* noms d'une personne comme جُون كنيدي *guwn kinydy* 'John Kennedy'
  - *Organisation :* raison sociale d'une société, banques, associations, universités, etc. à titre d'illustration nous citons يُونِيسْكُو*yuwniskuw* 'Unesco', etc.;
  - *Localisations :* cette catégorie concerne les toponymes tels que les noms de pays, villes, états, mers, océans, montagnes, fleuves, etc. Par exemple, الجَزَائر*Aljaza'ir* 'Algérie', بَارِيس *baâriys* 'Paris', البَحْر الأَبْيض المُتَوسّط *el bahr elaabyad elmutawassit* 'La mer méditerranée'.
- **NUMEX :** contient les entités formalisées dans des expressions numériques de pourcentage, taille, expressions monétaires, etc.
- **TIMEX :** concerne les entités exprimant le temps, la date ou une durée.

Nous pouvons résumer cette classification dans le schéma suivant :

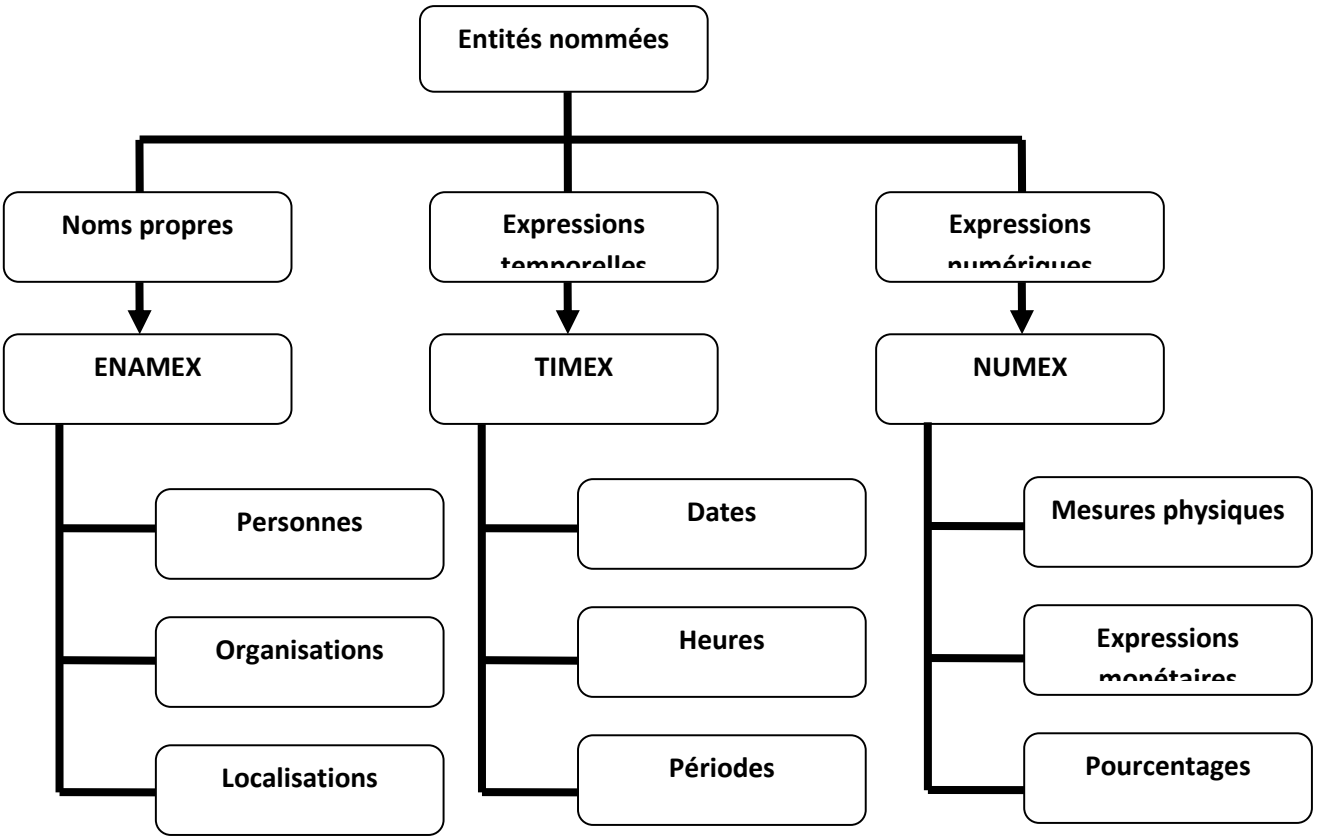


Figure 4. 1. Typologie des entités nommées (Mesfar, 2008)

4.2. Applications

L'utilisation de la reconnaissance des entités nommées diffère d'une application à une autre : elle est parfois utilisée comme un module interne d'un outil de TAL servant à d'autres modules, pour faire de l'analyse syntaxique ou de la désambigüisation lexicale par exemple ; comme elle peut être utilisée comme une partie d'une chaîne de traitement avec une application directe particulière. A titre d'illustration, voici quelques exemples d'applications employant l'analyse des entités nommées :

- ***Recherche d'Information (RI):*** RI est la tâche qui consiste à identifier et récupérer les documents pertinents depuis une base de données selon une requête utilisateur (Benajiba et al., 2009). L'utilisation de l'identification des entités nommées dans la RI doit être faite par la reconnaissance de l'EN au niveau de la requête et au niveau des documents à renvoyer. Les entités nommées sont extrêmement discriminantes et leur présence dans une question est gage de résultat précis.
- ***Traduction Automatique (TA).*** TA est l'opération permettant de traduire automatiquement un texte d'une langue naturelle source à une autre langue cible. Le traitement des ENs est requis afin de faire une traduction correcte. Ainsi, la qualité de la traduction des ENs devient une partie autonome qui améliore considérablement les performances du système de TA. A titre indicatif, certaines entités nommées et des mots possèdent la même forme orthographique mais ne jouent pas le même rôle dans la phrase ce qui signifie que pour les mots une traduction intégrale du mot est requise ce qui n'est pas le cas pour les ENs où seulement une translitération est nécessaire. Par exemple, si nous voulons traduire le mot arabe ' 'خالدen Français, s'il s'agit d'un mot il sera traduit en 'éternel' et s'il s'agit d'une EN il sera transcrit en 'Khaled'. Nous précisons que dans le cas où l'EN comprend un nom commun comme Mont Saint-Michel, dans ce cas Mont est traduit alors que Saint-Michel est translittéré.

- ***Question-Réponse (QR).*** Les systèmes de QR peuvent être considérés comme une application de recherche d'information avec des résultats sophistiqués. Ces systèmes prennent en entrée des questions en langue naturelle et tentent de renvoyer en sortie des réponses concises, précises et pertinentes. La reconnaissance des ENs peut être utilisée dans ce système afin de mieux analyser les questions et d'identifier les d réponses pertinents par rapport à la requêtes initiale. De ce fait, la prise en compte des ENs est essentielle pour la compréhension de la question et le calcul de la réponse. Par exemple, l'entité " الشرق الأوسط" Aš-šarq Al-awsat" Moyen-Orient" peut être considérée comme une Organisation (un nom de journal) ou comme un lieu en fonction du contexte.

- ***Clustering de texte (TC).*** L'utilisation de la reconnaissance des ENs dans ce type d'application peut être faite pour réaliser un classement des clusters générés en se basant sur l'association du taux des entités associées avec chaque cluster. Cela permet d'améliorer le processus d'analyse de la nature des clusters ainsi que l'approche de clustering en fonction des caractéristiques sélectionnées.

- **Analyse syntaxique.** Cette analyse est une étape primordiale dans n'importe quelle analyse de texte. Elle peut tirer profit de la reconnaissance des ENs (REN) à différents niveaux de l'analyse, considérant ainsi cette reconnaissance comme un module dans la chaine de traitement du texte. Au niveau de l'étiquetage morphosyntaxique et de la segmentation la REN peut être utilisée pour identifier certaines entités complexes contenant parfois des signes de ponctuation, ce qui entraîne un gain en temps et en précision. Cette REN peut permettre aussi de diminuer les erreurs au niveau de l'analyse syntaxique proprement dite, et notamment celles liées à la coordination des entités. Enfin, les relations grammaticales (dépendances syntaxiques) peuvent être enrichies en sémantique grâce aux ENs, par exemple dans la phrase 'Ils se sont rencontrés à Alger', la détection de l'entité Alger permet de construire la dépendance *Localisation* entre le verbe rencontrer et Alger à travers l'information géographique qu'elle contient.

4.3. Particularité de la langue arabe liée à la détection des entités nommées

Les systèmes de reconnaissance des ENs en arabe sont confrontés à plusieurs chalenges. La reconnaissance est d'autant plus difficile dans le cas de la langue arabe en raison de ses particularités rendant l'identification des ENs plus difficile que pour les langues latines. Nous décrivons ci-après, les principales caractéristiques menant à compliquer le traitement des ENs et donnant un trait particulier pour les systèmes de REN pour l'arabe :

- **Absence de la majuscule :** la langue arabe, comme toute langue sémitique, est caractérisée par le fait qu'elle ne dispose pas de la fonction de capitalisation, autrement dit au niveau de son écriture elle ne fait pas la distinction entre les lettres majuscules et les minuscules. Ce trait ne permet pas d'identifier facilement les ENs contrairement aux langues latines où les ENs commencent généralement par une majuscule. Cette absence de capitalisation en arabe n'empêche pas de l'utiliser lorsque nous transcrivons ou nous traduisions des ENs écrites en arabe vers une langue latine (comme le français ou l'anglais). Par conséquent, il faudra mettre la majuscule aux noms propres de personnes, organisations, lieux, etc.

- **Morphologie complexe et agglutination :** la langue arabe a une structure morphologique complexe. Elle est basée sur les systèmes de racines-schème et est considérée comme une langue très flexionnelle et fortement agglutinée dans le sens où un mot (lemme) peut être formé à partir d'une racine à laquelle nous pouvons ajouter des préfixes, suffixes et des clitiques. Cette problématique doit être traitée afin de faciliter le processus de détection des entités nommées dans les textes. A titre d'exemple, les clitiques attachés au mot comme le proclitique 'و' *waw* 'et' ou l'enclitique 'ب' *bi* 'avec' ou encore 'ل' *li* 'pour' doivent passer par une étape de prétraitement (analyse morphologique) pour segmenter les clitiques et extraire la racine du mot.

- **Absence des voyelles courtes et ambiguïté :** les signes diacritiques ou ce que nous appelons les voyelles courtes sont nécessaires pour la prononciation des mots en arabe. L'arabe moderne est caractérisé par l'absence de ces signes diacritiques au sein des textes ce qui est fréquent dans les articles de presse, livres, etc. Par conséquent une forme de mot en arabe peut être voyellée de multiples façons, avec des significations différentes en fonction du contexte où elle apparaît. Ce problème de la non-vocalisation des textes peut engendrer un haut degré d'ambiguïté affectant les systèmes de reconnaissance des entités nommées. En effet, les vocalisations acceptées pour une forme d'un texte peuvent désigner des mots (déclencheurs) introduisant différents types d'entités nommées. Par exemple, la forme non voyellée «منظمة» peut avoir les vocalisations suivantes avec des interprétations différentes :
  - ✓ «مُنَظَمَة» *munaDamat* 'l'organisation' : mot déclencheur d'un nom d'organisation;
  - ✓ «مُنَظِّمَة» *munaDDimat* 'l'organisatrice' : mot déclencheur d'un nom de personne.

- **Problème de délimitation et polysémie :** l'abondance dans les textes de mots inconnus du dictionnaire demandant un découpage par l'analyseur morpho-lexicale, c'est le cas des entités nommées, engendre des problèmes de délimitation des ENs. De plus, cette difficulté est accentuée par la présence de formes polysémiques dans une entité nommée. Pour illustrer ces propos, prenons la forme «أَكْرَمْ» (akram) qui peut désigner un prénom masculin, une forme verbale fléchie (il a honoré) ainsi qu'un

adjectif superlatif (le plus généreux). Par ailleurs, selon (Mesfar, 2008) il existe des cas d'ambiguïté où l'entité nommée peut être confondue avec un nom composé ou un fragment d'une phrase verbale. C'est le cas de la séquence «حافظ الأسد» (haâfiz al-asad) qui peut donner lieu aux trois analyses suivantes :

- ✓ *Le nom d'une personne politique :* 'الرئيس السوري حافظ الأسد' *al-rayiys al-suwriy haâfiz al-asad* 'le Président Syrien Hafedh Al-Asad';
- ✓ *Un nom composé :* 'نظّف حافظ الأسد القفص' *nazzafa haâfiz al-asad al-qafasa* 'Le gardien du lion a nettoyé la cage';
- ✓ *Un fragment d'une phrase verbale :* 'حافظ الأسد على هيبته' *haâfaza al-asad 'ala haybatihi* 'Le lion a préservé sa dignité'.

- **Variantes orthographiques :** un autre facteur se manifeste, rendant la tâche de la reconnaissance et typage des entités nommées plus difficile, réside dans l'absence d'une norme commune (normalisation de l'orthographe) et d'une stratégie arabe unifiée dans le domaine de la translittération des noms propres étrangers. Pour combler ce manque, la prononciation des dialectes vernaculaires est utilisée comme base pour la transcription des noms arabes. Par conséquent, un mot peut être transcrit en plusieurs formes ayant un même sens et référençant le même mot. Ces différentes formes créent bien évidemment des ambiguïtés supplémentaires pour le système des RENs. Par exemple le mot 'جرام(jrAm – *Gram*) peut être écrit aussi la forme غرام(*grAm*) tout en référençant toujours le même sens.

- **Le manque de ressources linguistiques :** la langue arabe se heurte au manque et limitation du nombre de ressources linguistiques qui sont libres et à des fins de recherche, et beaucoup de celles qui sont disponibles ne sont pas adaptées pour les tâches de reconnaissance des entités nommées. Cette inadéquation est due à l'absence d'annotation des entités nommées dans ces corpus ou à leur taille qui n'est pas parfois suffisamment grande. Il y a également le problème de la rareté des Gazeeters arabes qui sont généralement limités en taille. Pour contourner ces obstacles, les chercheurs ont tendance à construire leurs propres ressources linguistiques afin d'alimenter et évaluer les systèmes de reconnaissance des entités nommées en arabes.

4.4. État de l'art sur les systèmes de reconnaissance des entités nommées en arabe

Les travaux sur la reconnaissance des entités nommées s'articulent autour de deux axes : *la détection* de l'entité nommée (l'identification) et *l'extraction* de ces EN en les associant à différents types prédéfinis. Dans cette section, nous présentons les deux axes en mettant en avant pour la partie identification les indices manifestant la présence des ENs, ensuite nous décrivons les principaux systèmes d'extraction de ces entités.

4.4.1. Identification des entités nommées

Selon (MacDonald, 1996) la classification d'un nom propre fait émerger deux types de preuves complémentaires : *interne* (internal evidence) et *externe* (external evidence). Ces deux types découlent des exigences de la sensibilité au contexte et permettent de détecter une entité nommée dans un texte.

4.4.1.1. Les preuves internes

Elles sont dérivées de l'intérieur de la séquence de mots qui contiennent l'entité nommée. Ce sont des mots (ou groupe de mots) indices correspondant à des abréviations, des prénoms ou des sigles, appelés des « marqueurs lexicaux» ou « mots déclencheurs ». Ces indices accompagnent et entourent les entités nommées et permettent généralement de provoquer leurs catégorisation et prédire leurs présences. Elles peuvent être définies et contenues dans des listes appelées *gazetteers.*

Voici quelques exemples illustratifs de ce type de preuve interne :

- هدى سعدان(**Houda** Sâadane)
- محمد**III** (Mohamed **III)**
- عرفة **جبل**(le **Mont** Arafa)
- *سوسيته جنرال* **بنك**(la B**anque** Société Générale)
- لاس فيغاس **شارع**(l'**avenue** de Las-Végas)

4.4.1.2. Les preuves externes

La preuve externe est le critère de classification fournie par le contexte dans lequel le nom propre apparaît. Les noms propres sont des façons de faire référence à des individus d'un type spécifique (personne, église, groupe de rock, ...). Généralement, dans un entretien ou discours, les auteurs enrichissent leurs textes avec des informations complémentaires sur les personnes, lieux, organisations qu'ils citent afin d'aider les lecteurs et auditeurs à mieux identifier ces entités. Par conséquent, ces informations peuvent alimenter et faciliter un processus automatique de détermination de type d'un nom propre. C'est ainsi qu'un nom de personne est souvent accompagné d'un titre ou d'un grade, et un nom d'organisation d'un mot-clé de type classifiant comme c'est illustrer dans les exemples suivants :

- ✓ ***Mademoiselle*** Houda Saâdane
- ✓ Le ***professeur*** Mathieu Guidère
- ✓ ***Compagnie*** Air-Algérie

La preuve externe est nommée aussi *contexte droit* ou *contexte gauche* selon où elle se trouve par rapport au nom propre dans le texte (à droite ou à gauche). Par exemple :

- **مدينة باريس**(La ville de Paris – à droite du nom propre)
- **بوتفليقة الرئيس الجزائري**(Bouteflika le président algérien – à gauche du nom propre)

De ce fait, les preuves externes se basent sur les relations syntaxiques au sein d'une phrase pour attribuer la catégorie d'une telle entité. Cette catégorisation utilise les informations morphosyntaxiques fournies par l'étape d'analyse morphologique. Elles sont nécessaires pour de haute précision pour remédier au fait que les listes des mots prédéfinis ne peuvent jamais être complètes.

En conclusion, la prise en compte de ces preuves internes et externes peut aider un système de reconnaissance des entités nommées mais elle n'est pas suffisante. Un autre moyen de compléter ces informations pour un système est le recours à des lexiques. Ces lexiques sont des listes des mots auxquels sont associées des catégories sémantiques pour indiquer le type de l'entité nommée (personne, lieu ou une organisation).

4.4.2. Systèmes de reconnaissance des entités nommées

Après l'identification des ENs, il faudra les extraire. Trois *approches* sont couramment évoquées dans la littérature à savoir : l'approche à base de règle (appelé aussi linguistique ou symbolique), l'approche statistique (dite aussi à base d'apprentissage) et

l'approche hybride. Ces approches apportent des explications supplémentaires sur les systèmes de reconnaissance des entités nommées. Dans la suite de cette section nous passons en revue les différents systèmes ainsi que les principaux travaux réalisés pour la reconnaissance des entités nommées en arabe.

4.4.2.1. Les systèmes à base de règle

Cette approche est basée sur un ensemble de règles linguistiques et contextuelles construites et écrites manuellement. De ce fait elle repose sur l'intuition humaine. Ces règles prennent la forme de patrons d'extraction exprimés par une grammaire locale qui décrit les modèles de correspondance pour les ENs. Ces modèles de correspondance utilisent d'une part les preuves internes et externes fournies par le contexte où les ENs apparaissent, et d'autre part exploitent les annotations fournies par l'étiquetage morphosyntaxiques en plus des informations contenues dans des ressources comme les lexiques, dictionnaires ou encore '*gazetteers*'. Nous notons qu'il est nécessaire de savoir définir et attribuer les bonnes frontières aux entités nommées. Les règles utilisées sont généralement formulées par des des transducteurs à état finis (les expressions régulières).

Les premiers travaux sur la reconnaissance des ENs en arabe selon l'approche à base de règles, datent de 1998 où **(**Maloney et Niv, 1998) ont développé un outil baptisé *'TAGARAB'* qui repère les noms propres (Personne, Organisation, Lieu, Nombre et Heure) selon une technique combinant un module filtrage par motif *(pattern-matching)* avec un analyseur morphologique pour améliorer les performances. Les résultats des tests de cet outil sur un ensemble de données aléatoires, issues du journal AI-Hayat, montrent que la combinaison de la détection des ENs avec un analyseur morphologique permet d'améliorer significativement la précision de la reconnaissance des ENs.

(Abuleil, 2004) a développé de son côté un système d'extraction des noms propres en arabe fondé sur l'utilisation de règles écrites à la main et les déclencheurs. Le système commence par sélectionner les phrases qui peuvent inclure des noms propres, ensuite il construit des graphes qui représentent les mots de ces phrases et les relations entre eux et, enfin, les règles sont appliquées pour repérer et classer les noms propres avant de les enregistrer dans une base de données. Cette base de données peut servir au sein de systèmes de questions-réponses par exemple. Le système d'Abuleil a été évalué sur un corpus de 500 articles de presse du journal Alraya donnant lieux à une précision moyenne avoisinant les 92%.

(Traboulsi, 2006) a présenté un modèle de reconnaissance des entités nommées, appelé *NExtract*, utilisant la grammaire locale et les dictionnaires. Il a montré des résultats satisfaisant de l'application de son outil sur une petite échelle avec le corpus ***Reuters***. Cette approche a été améliorée dans (Traboulsi, 2009) en combinant cette fois-ci la grammaire locale avec des automates à état finis.

Les travaux de (Mesfar, 2007) ont permis la mise au point d'une composante arabe sous un environnement linguistique, dénoté *NooJ,* pour traiter des textes arabes et faire la reconnaissance des ENs. Cette composante effectue les traitements suivants : la tokenisation, l'analyse morphologique et la détection des ENs. Le détecteur des ENs exploite un ensemble de gazetteers et de listes d'indicateurs pour soutenir la construction de règles. Le système identifie les ENs de type: personne, lieu, organisation et expressions temporelles. Il utilise également les informations morphologiques pour extraire les noms propres inconnus et améliorer ainsi la performance globale du système.

L'approche à base de règles pour la REN est aussi adoptée dans les travaux de

(Shaalan et Raza, 2007) qui ont développé le système *PERA*. PERA est basé sur la grammaire qui est construite pour identifier les noms de personnes dans les textes arabe avec un degré élevé de précision. PERA est composé de trois éléments: des gazetteers, des grammaires locales et le mécanisme de filtration. Les listes blanches de noms de personne sont fournies dans le composant '*gazetteer*' afin d'en extraire les noms correspondants indépendamment des grammaires. Par la suite, le texte d'entrée est analysé par la grammaire donnant des expressions régulières pour identifier le reste des entités nommées de type Personne. Enfin, le mécanisme de filtrage est appliqué sur les ENs détectées par des règles grammaticales afin d'exclure celles qui sont invalides. PERA a donné des résultats satisfaisants lorsqu'il était appliqué sur les corpus *ACE* et *Treebank*.

Le système *NERA* (Shaalan et Raza, 2008; 2009) est une prolongation des travaux précédents permettant de reconnaître d'avantage de types d'ENs. Il est aussi fondé sur des règles et capable de reconnaître 10 types différents d'ENs : personne, localisation, organisation, date, heure, ISBN, prix, mesure, numéros de téléphone et les noms de fichiers. NERA a été mis en œuvre dans le cadre de la plateforme FAST ESP où le système comprend, comme PERA, trois composants ayant les mêmes fonctionnalités pour couvrir les 10 types d'ENs. De plus, les auteurs ont construit leur propre corpus de différentes ressources afin de disposer d'un nombre représentatif de cas pour chaque type d'EN.

(Elsebai et al., 2009) ont proposé un système de REN intégrant le filtrage par motif (en anglais *pattern matching*) associé avec l'analyse morphologique afin d'extraire les noms de personne à partir des textes arabes. Le moteur de filtrage par motif utilise des listes de mots-clés sans utiliser pour autant des listes prédéfinies de noms de personnes.

Les systèmes à base de règles ont été aussi investigués dans les travaux de (Zaghouani, 2012) qui a proposé le système *RENAR* pour extraire les entités nommées de type : personne, lieu et organisation. RENAR est composé de trois phases: 1) prétraitement morphologique, 2) la recherche des ENs connues et 3) l'utilisation de la grammaire locale pour extraire les ENs inconnues. Les expérimentations ont montré que RENAR dépasse les performances de ANERsys 1.0 (Benajiba et al., 2007), ANERsys 2.0 (Benajiba et Rosso, 2007) et LingPipe[7] pour l'extraction des entités nommées de type Lieu lorsqu'il est appliqué sur l' ensemble de données du corpus ANERcorp, tandis que LingPipe donne de meilleurs résultats que RENAR lorsqu'il s'agit de l'extraction des ENs de type personne et organisation.

4.4.2.2. Les systèmes Statistiques

L'objectif des systèmes à base d'apprentissage automatique (dits aussi statistiques) est de réaliser le développement, l'analyse et l'implémentation de modèles d'analyse automatisables par un processus d'apprentissage basé sur des volumes importants de données (corpus annoté). Parmi les modèles les plus utilisés pour la reconnaissance des ENs nous citons : l'entropie maximale (EM), les machines à vecteurs de support ou séparateurs à vaste marge (en anglais *Support Vector Machine*, SVM), les arbres de décision, les règles logiques, les modèles probabilistes, les chaînes de Markov cachées (HMM) ou encore les champs aléatoires conditionnels (*Conditional Random Field* : *CRF*). Par exemple, un système observant plusieurs fois la présence de l'abréviation «***Mlle***» devant un mot étiqueté comme *nom de personne* dans le corpus d'apprentissage pourra facilement en déduire un modèle d'analyse.

(Benajiba et al ., 2007) ont mis au point une première version d'un système de

[7] LingPipe est un logiciel libre disponible sur http://alias-i.com/lingpipe/.

reconnaissance des ENs pour l'arabe, appelé *ANERsys*. Ce système est basé sur une méthode d'apprentissage statistique qui utilise un étiquetage fondé sur le maximum d'entropie (*ME*). Les auteurs ont construit leurs propres ressources linguistiques qu'ils ont nommé *ANERcorp* (corpus annoté) et *ANERgazet* (gazetteers). Le système utilise des traits lexicaux et contextuels ainsi que des gazetteers. Il peut reconnaître quatre types d'ENs : personne, lieu, organisation et divers. L'apprentissage automatique embarqué dans ANERsys a été effectué sur un corpus de 125 000 mots. Dans le but d'améliorer les performances du système, l'approche adoptée a été combinée à un lexique qui a été construit manuellement à partir de plusieurs sites de nouvelles en ligne. Le lexique considéré comprend 1950 noms de lieux, 1920 noms de personnes et 262 noms d'organisations.

Cependant, cette version d'ANERsys présente des difficultés pour détecter les entités nommées qui sont composées de plus d'un token. Pour résoudre ces difficultés, (Benajiba et Rosso, 2007) ont développé une nouvelle version ANERsys 2.0, qui utilise un mécanisme de prédiction pour la reconnaissance des ENs. Ce mécanisme est effectué en deux étapes : 1) la détection des frontières (point du début et de la fin) de chaque EN en introduisant des catégories morphosyntaxiques (POS), et 2) classification des entités nommées détectées en précisant leurs types. (Benajiba et Rosso, 2008) ont introduit dans ANERsys l'application du CRF à la place de EM afin d'améliorer les performances. Ce nouveau système basé sur les CRF a permis d'explorer l'intégration de l'ensemble des traits dans un modèle unique et qui mène à des résultats plus élevés en termes de précision.

Un autre système basé sur les CRF a été proposé dans (Abdul-Hamid et Darwish, 2010) pour la reconnaissance de trois types d'ENs : personne, lieu et organisation. Il intègre un ensemble de traits intra-mots : *n-grammes*, la position des mots, la longueur des mots, la probabilité de *uni-gramme* des mots, les mots précédant et succédant les *n-grammes* et la probabilité des *n-grammes*. Cependant, le système ne tient pas compte de tout autre type de traits. Le système proposé a été évalué à l'aide des corpus ANERcorp et ACE 2005. Les résultats obtenus montrent que le système présente des précisions plus importantes que le système de reconnaissance des entités nommées basée sur les CRF proposé par (Benajiba et Rosso, 2008).

L'utilisation des SVM (Support Vector Machines) pour la reconnaissance des ENs a été proposé dans (Benajiba et al., 2008a). Le système proposé emploie des traits contextuels, lexicaux et morphologiques ainsi que des gazetteers, POS-tags et BPC. Il utilise également la nationalité et la capitalisation correspondante en anglais. Le système a été évalué en utilisant le corpus ACE et ANERcorp. Les meilleurs résultats sont obtenus lorsque tous les traits sont pris en considération, et met en avant l'efficacité d'un prétraitement des textes pour segmenter les différentes constituants d'un mot (proclitiques, lemmes et enclitiques).

Une autre approche combinant les deux méthodes d'apprentissage SVM et CRF a été proposée dans (Benajiba et al., 2008b). En outre, le système utilise des traits lexicaux, syntaxiques et morphologiques et une approche multi-classificateur où chaque classificateur est conçu pour marquer une classe d'EN séparément en utilisant une des techniques SVM ou CRF. Ce système a aussi été utilisé pour étudier la sensibilité des différents types d'EN par rapport à plusieurs types de caractéristiques. L'évaluation de cette approche a été faite sur des ensembles de données du corpus ACE et a obtenu une F-mesure de 83,5%. Un des principaux résultats obtenus est le fait que nous ne pouvons pas trancher sur la supériorité d'une technique sur une autre parmi celle utilisée (SVM et CRF) en matière de reconnaissance des ENs. D'autres études, en l'occurrence (Benajiba et al, 2009a; 2009b) ont confirmé ainsi

l'importance de tenir compte des caractéristiques spécifiques de la langue en arabe pour la reconnaissance des ENs.

Une autre étude comparative des techniques d'apprentissage type Machine Learning (ML) a été présentée dans les travaux de thèse de (Benajiba, 2009). Cette étude concerne la reconnaissance des ENs en arabe et compare les approches telles que l'entropie maximale (EM), Support Vector Machines (SVM) et Conditional Random Fields (CRF) en utilisant le système ANERsys. Cette étude a conclu qu'aucune approche ML n'est considérée comme meilleur que l'autre et que les meilleurs résultats ont été obtenus quand il a utilisé une approche multi-classificateur où chaque classificateur utilise la meilleure technique de ML pour la classe d'entité nommée spécifique.

Quant aux travaux de (AbdelRahman et al., 2010), ils ont intégré deux approches de systèmes statistiques pour traiter les ENs arabe incluant le CRF et la reconnaissance des formes d'amorçage. L'ensemble des caractéristiques utilisées avec le classificateur CRF inclut des spécificités au niveau des mots, des POS tag, des BPC, les gazetteers et des caractéristiques morphologiques. Le système est conçu pour extraire les 10 types d'EN : personne, lieu, organisation, le travail, dispositif, voiture, numéro de téléphone portable, devise, la date et l'heure. Les résultats des évaluations sur les données du corpus ANERcorp montrent que le système proposé présente des performances meilleures que celle obtenues par le système LingPipe.

4.4.2.3. Les systèmes Hybrides

L'approche hybride consiste à combiner les techniques des systèmes à base de règles et les techniques des systèmes statistiques. Cette combinaison a pour objectif de tirer profit des avantages des techniques présentées dessus et d'optimiser la performance globale du système (Petasis et al., 2001). Ces systèmes ont pour but l'enrichissement automatique des dictionnaires avec des corpus beaucoup plus petits que ceux dont ont besoin les systèmes statistiques. La direction du flux de traitements peut être du système à base de règles vers le système statistique ou vice versa. Nous considérons trois systèmes hybrides pour la reconnaissance des entités nommées mis au point récemment.

Le premier est développé par (Abdallah et al., 2012) offrant la capacité d'identifier les entités nommées de types suivants : personne, lieu et organisation. Ce système comporte deux composants : le premier est à base de règles qui est une ré-implémentation du système de NERA (Shaalan et Raza, 2008) utilisant l'outil GATE, le deuxième est une composante-ML utilisant des arbres de décision pour construire le classificateur des entités nommées. Chaque *token* est représenté par un vecteur de caractéristiques incluant les décisions issues des règles sous forme de propriétés. Les autres caractéristiques prise en compte sont : la taille du mot, POS tag, indice du nom (une fonction binaire utilisée pour tester si un POS tag est un nom ou pas), les gazetteers, marqueur de fin de proposition, les propriétés de préfixe et suffixe. Les résultats expérimentaux, en utilisant les données du corpus ANERcorp, montrent que le système hybride présente des performances meilleures que le système de reconnaissance des ENs basé sur les CRF et construit par (Benajiba et Rosso, 2008).

Le deuxième système hybride proposé par (Oudah et Shaalan, 2012) traite la problématique de la reconnaissance des ENs en largeur et en profondeur et nécessite des investigations supplémentaires pour améliorer l'étendu des traitements et la performance globale. Il est capable de reconnaître 11 types d'ENs dont : personne, lieu, organisation, date,

heure, prix, pourcentage, numéro de téléphone, mesure, ISBN et le nom d'un fichier avec un degré de précision assez élevé. Ce système utilise trois approches statistiques différentes, incluant les arbres de décision selon (Orphanos et al., 1999), SVM introduite dans (Vapnik, 1995) et la régression logistique présentée dans (Hastie et al., 2009). Ces approches s'appuient sur différents caractéristiques, incluant l'information contextuelle et morphologique, utilisées pour former différentes combinaisons afin de trouver les ensembles de traits avec des performances optimales.

Plus récemment (Gahbiche-Braham et al., 2013) ont aussi proposé un système de reconnaissance des entités nommées (NERAr). Le système distingue trois types d'EN : Personne, Lieu et Organisation. Il est basé sur des outils d'apprentissage automatique et utilise le modèle des champs markoviens conditionnels (CRF), tels qu'implémenté dans l'outil Wapiti. Cette implémentation permet d'utiliser de très gros modèles incorporant des centaines d'étiquettes et des centaines de millions de descripteurs et de sélectionner les descripteurs les plus utiles par le biais d'une pénalité L1.

Une représentation BIO (Begin, Inside, Outside) est utilisée pour limiter les frontières d'EN et le modèle développé prédit à chaque position une des dix balises différentes. Le système est capable de segmenter le texte, repérer les ENs et proposer des traductions de ces EN à partir de dictionnaires bilingues. Les expériences ont été réalisées sur le corpus ANER (benajiba et al., 2007). Une adaptation non supervisée de NERAr a été également explorée afin d'adapter l'outil de détection des ENs au type de données traitées.

À l'instar des autres systèmes de repérage des ENs à base de règles mentionnés précédemment, l'extraction et le typage des ENs avec notre système est fondée principalement sur un lexique, sous forme de dictionnaires, et sur un ensemble de règles de repérage, sous forme d'expressions régulières faites à la main. En plus de ces opérations conventionnelles, notre système effectue une analyse syntaxique supplémentaire afin de regrouper les éléments qui composent l'EN et à typer celle-ci. Cette étape exploite les relations syntaxiques de dépendance et sur les ENs simple typées. Elle implique les opérations suivantes : *l'attachement* et *l'étiquetage*. Nous rappelons que l'attachement concerne la détermination si deux mots (ou entités nommées) sont connectés directement ou pas, en d'autre termes, c'est l'identification d'un lien direct de dépendance syntaxique entre deux mots de la phrase par la mise en avant d'une relation syntaxique entre la tête (gouverneur) et le mot dépendant (régie). Quant à l'étiquetage, il consiste à regrouper les dépendants syntaxiques et annoter la relation identifiée par un nom référant à une famille (ou type) de constructions syntaxiques d'une langue donnée.

4.5. Approche proposée pour la reconnaissance des entités nommées

Afin de prendre en considération les contraintes imposées par les spécificités de la langue arabe, nous avons développé un système de reconnaissance d'entités nommées basé sur les automates à états finis pondérés. Ces automates effectuent plusieurs fonctions comme l'analyse morphologique, l'analyse syntaxique des textes ainsi que la désambiguïsation. L'architecture de notre système de reconnaissance est présentée dans la figure (4.2) comme suit :

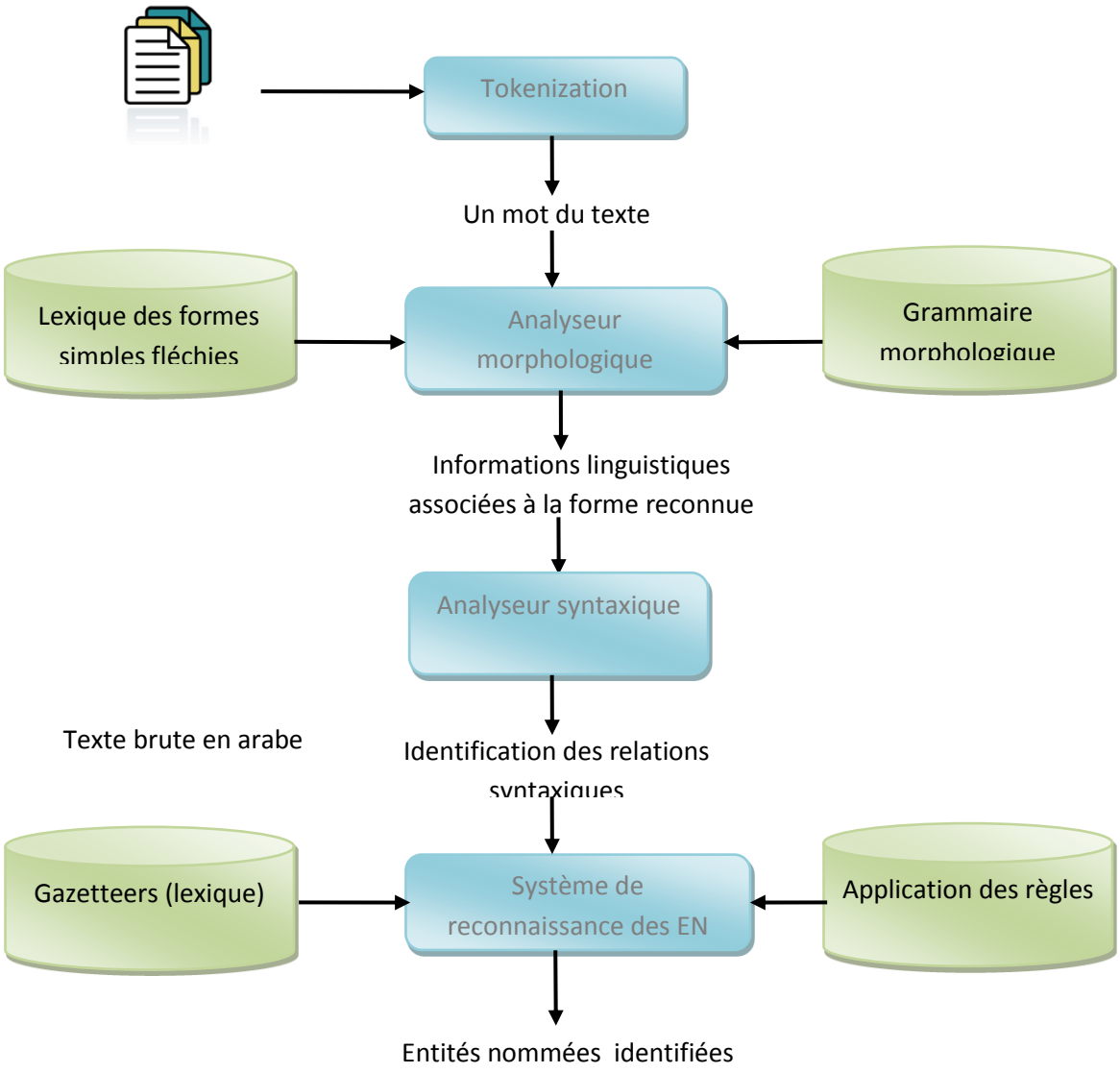


Figure 4. 2. Processus de reconnaissance des ENs

Le système commence par une phase classique de saisie du texte à analyser qui est introduit sous une forme brute. Le texte introduit subit ensuite les traitements suivants :

- ✓ une tokenization du texte
- ✓ une segmentation des formes agglutinée en morphèmes
- ✓ la désambiguïsation
- ✓ à un étiquetage morphosyntaxique.

Ces différentes opérations forment l'analyse morphologique que reçoit le texte en entrée, et renvoient en sortie des formes canoniques segmentées, à travers l'identification des proclitiques et des enclitiques rattachés à ces formes et ainsi que la forme normalisée de chaque mot du texte : par exemple les verbes sont normalisés dans leur forme à l'infinitif, les noms au singulier, les adjectifs au masculin singulier, etc. Ces formes sont fournies avec version étiquetée. L'étiquetage a pour objectif de produire *les catégories grammaticales* d'un mot ou d'un groupe de mots dans une phrase donnée (*noms, verbes, conjonctions)* en plus *des informations morphologiques* (*genre, nombre, personne)* associé à cette forme.

Ensuite, les formes produites par analyse morphologique passent dans notre système d'analyse syntaxique. Les automates syntaxiques établissent des relations syntaxiques de dépendance typées entre les mots, en s'appuyant surtout sur leurs catégories et sur leurs propriétés. En ce qui concerne, les entités nommées, ils mettent en évidence les liens entre les mots au sein des groupes nominaux, permettant ensuite d'identifier une entité nommée, même lorsque son annonceur est éloigné du nom propre. Nous avons spécifié quelques types de relations afin de mieux repérer les entités nommées. Par exemple :

- *PrenomNP* : elle désigne les relations entre un prénom et un nom de personne, avec la tête le nom de famille et le prénom étant le dépendant de la relation.

هُدَى سَعْدَان

Houda *Sâadane*

(Prenom) (NP)

- *AnnpNP* entre un annonceur et un prénom ou un nom propre, comme une tête le nom propre seulement s'il s'agit d'un nom de personne dans les autres cas c'est l'annonceur qui est considéré comme la tête.

الدُّكْتُورَة هُدَى(le docteur Houda)

al+dukturat+u *Houda*

DEF+(annp)+NOM (Prenom)

- *AnnpAdj* désigne les relations entre un annonceur et un adjectif, comme une tête l'annonceur et l'adjectif est le dépendant de la relation syntaxique.

  الرَّئِيسُ الجَزَائِرِيُّ (le président algérien)

  al+ra'iss+u ***al+jazaA'iriyy+u***

  DEF+(annp)+NOM DEF+(ADJ)+NOM
- *AnnpRelNom* (complément de nom) désigne une relation entre un annonceur et un nom (complément d'un nom), par exemple :

  وزيرُ الدولةِ(le ministre d'état)

  waziir+u –compN → al+dawlat+i

  (annp)+NOM DEF+(N)+GEN

Outre les informations morphosyntaxiques associées aux formes obtenues et aux relations syntaxiques détectées, notre système, à l'instar des autres systèmes de REN à base de règles, exploite deux types de ressources linguistiques :

- **Les Gazetteers (lexique) :** il s'agit d'un ensemble de dictionnaires proposant des listes de marqueurs lexicaux. Ces marqueurs présentent des preuves internes et externes, proposés par (McDonald, 1994), permettant l'identification d'une potentielle EN dans un texte. Dans notre analyse nous nous intéressons au deux types de marqueurs suivants :
  - *Les déclencheurs :* ce sont des mots ou des catégories provoquant la détection d'un nom propre. Ces déclencheurs sont définis dans notre système par une liste finie de mots ou catégories intégrable dans les règles. Par exemple les noms de familles sont considérés comme déclencheurs pour la détection des prénoms.
  - *Les annonceurs :* ce sont des mots suivant ou précédant un nom propre faisant partie d'une EN. Ils peuvent être des mots qui désignent un métier, le titre d'une personne, un type de lieu, d'organisation, etc. Ces annonceurs sont répertoriés dans des listes spécifiques et possèdent une catégorie grammaticale « annp ».

  Ces différents marqueurs sont utilisés pour identifier les types d'entités nommées

réparties dans des dictionnaires appropriés à ces catégories. Parmi ces dictionnaires, nous citons :

d. *Noms de personnes :* contient les prénoms arabes et prénoms étrangers transcrits;
e. *Noms de lieux :* stocke les noms de pays, villes, états, mers, océans, fleuves, etc.
f. *Nom d'organisation :* mémorise les noms d'organisations, d'associations internationales, d'universités, de télévisions, etc.
g. *Expressions monétaires :* dédié aux noms de monnaies et leurs subdivisions;
h. *Expressions temporelles :* contient les noms de jour, plusieurs listes de noms de mois, etc.

- **Les règles :** ce sont des règles écrites manuellement et décrites par des expressions régulières (Regular Expression). Elles sont utilisées pour la détection des ENs en se basant sur les marqueurs, déclencheurs et annonceurs, provenant des Gazetteers pour retourner en sortie des informations linguistiques comme le type de l'entité nommées identifiée (nom de personne, lieu, etc.) Permettre ensuite le typage des ENs et elles permettent aussi l'identification des bornes (dites aussi frontières) des ENs complexes. Ces règles regroupent aussi l'ensemble des éléments d'une même entité nommée, permettant de représenter des séquences de mots formant une EN.

La phase finale consiste à regrouper les éléments qui composent l'EN et à typer celle-ci. Cette étape repose sur l'exploitation des relations syntaxiques de dépendance typées au sein des syntagmes nominaux lors de l'analyse syntaxique et sur les ENs simple typées lors de la reconnaissance des entités nommées. Dans un premier temps, un automate repère un nom propre ou un annonceur dans la phrase donnée. Un autre automate parcourt toutes les relations qui ont comme tête ce nom propre ou cet annonceur en prenant en compte les types des relations syntaxiques et la position du nom propre ou de l'annonceur dans celle-ci. Nous les récupérons jusqu'à rencontrer un nom propre, un annonceur, ou bien une frontière de groupe nominal. Il convient de signaler que certains types de relations ne peuvent pas faire partie d'une entité nommée, comme les relations identifiées entre un sujet et un verbe.

Nous notons que l'ordre d'application de ces règles est très important afin de bien repérer les ENs simples ou complexes. Notre stratégie de repérage se base sur l'application d'abord des règles concernant les entités les plus longues et plus complexes, ensuite l'application de celles concernant les entités simples. Cette stratégie est motivée par le fait de détecter les ENs complexes en premier nous évite les cas de repérage partiel de ces ENs. L'ordre d'application concerne aussi le type des entités : par exemple nous détectons en premier les ENs de type personnes qui figurent déjà dans le dictionnaire de noms et de prénoms. Ce choix d'application est due au fait que trouver en même temps le nom et le prénom d'une personne dans le dictionnaire, nous donne la certitude que l'EN en question est bien détectée. Il y a aussi le problème de chevauchement de règles entre les règles des noms de personnes et celui des organisations qui justifie aussi notre approche d'application des règles. Enfin nous signalons que pour l'identification des entités de type numérique et temporel, nous les repérons dans une étape ultérieure séparée pour des raisons techniques liées à notre système de détection des ENs.

4.6. Reconnaissance des noms propres – ENAMEX

4.6.1. Les noms de personnes

La première catégorie des noms propres que nous présentons concerne les noms de personnes. Dans cette partie nous introduisons les différentes structures que peut prendre un nom de personne en arabe ensuite nous exposons notre technique de détection de ce type d'EN.

4.6.1.1. Structure des noms de personnes :

On sait à cet égard que le nom d'une personne contient plusieurs éléments en arabe. Il est constitué en principe de six composants principaux (Zaghouani, 2009; Saâdane et al., 2012):

- ✓ ***La « Sifa » (titre) :*** il s'agit d'un titre honorifique, par exemple Imam (إمام), Sheikh,(الشيخ), Lalla (لالة), Sidi (سيدي) etc..
- ✓ ***La « Kunya » (particule d'usage) :*** généralement composée de « Abou » (père de…), suivi du nom d'un enfant ou bien de « Oum » (mère de + nom d'un enfant de la famille). Exemple : « Abou Omar » (Père d'Omar), «Oum Mohamed» (Mère de Mohamed), etc.
- ✓ ***Le « Ism » (Prénom) :*** il peut être simple ou composé, par exemple, Omar, Ali, Mohamed, Khaled, Abd allah, etc. Il indique parfois l'origine ethnique ou confessionnelle de celui qui le porte : par exemple, « Omar » est un prénom typiquement sunnite ; « Rustam » est un prénom typiquement iranien ; « Arslan » est typiquement turc, etc.
- ✓ ***Le « Nasab » (particule généalogique) :*** chaque nom est précédé par « Ibn » ou «Bin/Ben» («Bint/Bent» pour les femmes). Il indique la filiation généalogique exacte de l'individu concerné. Les Arabes remontent parfois très loin dans l'indication des ancêtres pour éviter les confusions entre personnes : ex. Muhammad Bin Abdallah Bin Salih Bin Said, etc.
- ✓ ***La « Nisba » (suffixe d'origine) :*** ce suffixe renvoie en principe à la tribu ou au clan dans la généalogie ancienne mais aujourd'hui, il désigne surtout le lieu de naissance des individus : Maghribi (né au Maroc), Libi (né en Libye), Masri (né en Égypte), etc. La « Nisba » est toujours précédée de l'article « Al-» et se termine par le suffixe « i ». Elle indique la résidence territoriale initiale des personnes, ou encore leur nationalité. Il existe des règles de formation de la Nisba qui sont plus complexes comme dans le cas où les noms communs composés de deux ou trois lettres. Prenons les exemples suivants :
  - Le nom commun حي*Hay* 'vivant' se transforme en la Nisba الحيوي*Hayawi* 'le vivant'.
  - Le nom propre علي (' alyi) se transforme en la Nisba العلوي*Al-alawiy* 'celui qui appartient à la secte des Alaouites', avec l'ajout de la lettre *lwl* et la voyelle courte *li/.*
- ✓ ***Le « Laqab » (nom de famille)*** : C'est un mot attribué à une famille pour la distinguer parmi les autres familles. Dans la langue arabe, le Laqab réfère généralement, en plus du nom de famille, à une classe sociale ou simplement à une description physique ou morale d'une famille donnée. Par exemple le nom de famille الأكحل*Al-akHal* qui veut dire 'le noir' ou حافي راسو*HAfi-Rassou* qui signifie 'celui qui est le crâne rasé'.

4.6.1.2. Identification des noms de personnes

Nous rappelons qu'une bonne partie des systèmes de reconnaissance des entités nommées se basent sur l'utilisation de lexiques spécialisés. D'après (Fourour, 2002), ces lexiques sont composés d'éléments pouvant jouer un ou plusieurs rôles dans une phrase :

- *EN* : entité nommée connue comme *ONU, Djamel*, etc.
- *Mot déclencheur* : élément faisant partie de l'entité nommée à l'instar des mots *Organisation, Avenue*
- *Contexte* : représente l'élément appartenant au contexte gauche immédiat de l'EN, mais ne faisant pas partie de celle-ci, comme c'est le cas des mots *docteur, algérien, etc*.
- F*in d'EN* : c'est l'élément qui est la dernière forme composant l'entité nommée : *handball, national, etc.*
- *Eléments d'EN* : c'est tout élément lexical pouvant faire partie de l'EN, sans pour

autant permettre la délimitation ou la catégorisation de l'EN.

La première étape de notre approche consiste à construire les ressources d'EN en élaborant, manuellement et automatiquement, un dictionnaire à partir des ressources textuelles (pages web par exemple), des listes issues des gazetteers ANERgazet2[8] proposées par (Benajiba et Rosso, 2007), les ressources se trouvant dans la basse de données géographique GeoNames[9] qui contient des lieux géographiques contenant des noms de personnes, et le lexique proposé par Attia[10]. Pour les deux premières ressources textuelles, nous avons effectué une fouille manuelle des textes. Cette étape nous a permis de collecter et de sélectionner dans une liste de noms et de prénoms potentiels qui seront triés automatiquement par la suite afin d'éliminer les doublons. Nous avons utilisé aussi un translittérateur des noms propres pour translittérer les noms propres qui proviennent d'autres dictionnaires latins (essentiellement ceux utilisés au sein de l'entreprise GEOLSemantics). Cette ressource est construite afin de faire l'association de chaque entité à une ou plusieurs étiquette(s) sémantique(s) rendant compte de certaines caractéristiques du ou des référents possibles de l'entité. Par exemple associer à l'entité *Charles-de-Gaulle* les étiquettes suivantes : nom de personne, lieu, organisation, édifice. Cette étape, nous a permis l'élaboration d'un dictionnaire de prénoms et de noms de famille qui contiennent 9000 prénoms arabes et prénoms étrangers transcrits reconnaissables par le biais de l'étiquette <+Prenom> ou <+NP> respectivement. Les entrées de ce dictionnaire sont de la forme suivante :

- ✓ حُسْنِي حُسْنِي`+prenom'`+m'
- ✓ شَاهِيْن شَاهِيْن`+prenom'`+m'
- ✓ سَعْدَان سَعْدَان`+np'

Outre les entrées simple, ce dictionnaire contient aussi les formes composées telles que :

- ✓ نُور الدِّين نُور الدِّين`+prenom'`+m'
- ✓ عَبْد الرَّحمَن عَبْد الرَّحمَن`+prenom'`+m'

Pour la reconnaissance des compositions de prénoms, souvent présentes dans les prénoms arabes introduits par les éléments lexicaux tels que (ibn – le fils de), (bin – le fils de), etc., nous avons construit une règle qui identifie la particule (ibn, bin, etc.) suivi par le prénom afin d'extraire l'EN.

En ce qui concerne les annonceurs nous avons construit une liste de mots utilisés pour le repérage des noms de personnes tels que les noms de professions, les titres, etc. Ces listes sont utilisées pour la reconnaissance des noms de personnes ainsi que la catégorisation de celles-ci. Cette liste des mots déclencheurs a été créée manuellement sur la base de nos connaissances linguistiques et de nos observations faites sur des corpus. Par exemple, la présence d'une mention à une fonction politique tel que الوالي *Al-walyi* 'le préfet' avant un nom de personne nous confirme la présence d'un syntagme nominal désignant un nom de personne même en cas d'omission ou d'absence de son prénom qui lui correspond dans le dictionnaire des prénoms.

La détection des noms de personnes a nécessité le plus grand nombre de règles à écrire par rapport aux autres types d'EN. La raison de cette complexité est principalement due aux

8 Téléchargeable depuis le site : http://www1.ccls.columbia.edu/~ybenajiba/downloads.html

9 http://download.geonames.org/export/dump/

10 http://www.attiaspace.com/getrec.asp?rec=htmFiles/LexMWEs

nombreuses possibilités de combinaisons entre les différents annonceurs et déclencheurs. Les règles écrites décrivent aussi bien les contextes potentiels de droite que de gauche.

Les règles de détection exploitent aussi le dictionnaire des adjectifs de nationalité utilisés dans des expressions telles que الرئيس الجزائري عبد العزيز بوتفليقة *Alra'iys Al-gazaâ'iriy 'abd al-'aziyz buwtafliqah* 'le président algérien Abdelaziz Bouteflika'. Une nationalité isolée, se trouvant sans un prénom ou un nom propre, ne peut pas être utilisée pour identifier un nom propre.

Une étude effectuée par (Mesfar, 2008) sur les articles journalistiques du journal (*Le Monde Diplomatique*), a remonté les statistiques suivantes au sujet des noms de personnes comme suit :

1. 70% des noms de personnes sont accompagnés d'un contexte droit ou gauche, sous forme interne ou externe, contenant une civilité, un titre, un nom de profession ou un gentilé.
   - *L'entité nommée est accompagnée d'un contexte droit uniquement :* cette situation représente 60% des cas.
   - *L'entité nommée est accompagnée d'un contexte gauche uniquement :* à titre d'exemple : معمر القذافي؛ الزعيم الليبي'Mu'amar ghadhaffi; le leader libyen'.
   - *L'entité nommée est accompagnée d'un contexte droit et un contexte gauche*
2. 18% des noms de personnes n'ayant pas de contextes descriptibles contiennent un annonceur apporté par un prénom appartenant aux dictionnaires
3. 11% des noms de personnes sont sans contexte. Ces noms sont principalement ceux de personnes déjà citées dans le texte ou ceux de personnes très connues pour lesquels l'auteur du texte estime qu'il n'est pas nécessaire de préciser ni le prénom, ni le titre, ni la profession tel est le cas pour Picasso ou Mozart.

Nous illustrons dans ce qui suit quelques exemples de règles d'extraction des noms de personnes. Ces illustrations sont faites en utilisant les expressions régulières pour faciliter la lecture et la compréhension des règles.

- **$firstname+ $lastname+ → <en entype="pers">**
  Cette règle détecte les ENs qui commencent par un prénom suivi immédiatement par un nom de famille et syntaxiquement liés par une relation de dépendance de type *PrenomNP*. Le prénom et nom détectés sont référencés dans les dictionnaires des prénoms et noms de familles, et leur recherche est faite dans ces dictionnaires par l'appel des fonctions *firstname* et *lastname* respectivement.

- **$title $adj_nationality? [$ firstname* $lastname] → <en entype="pers">**
  Cette règle permet la détection des ENs complexes : l'EN identifiée par cette règle commence par un annonceur de fonction politique suivi d'un éventuel adjectif de nationalité et de zéro ou plusieurs prénoms et d'un nom de famille. Cette règle identifie des ENs comme celle de la phrase suivante : الرئيس الجزائري عبد العزيز بوتفليقة *Alra'iys Al-gazaâ'iriy 'abd al-'aziyz buwtafliqah* 'Le président algérien *Abdelaziz Bouteflika'*. Pour investiguer l'EN de cette phrase, la règle linguistique commence par chercher le prénom عبد العزيز(Abdelaziz) dans le dictionnaire des prénoms, ensuite cherche le nom suivant (*Bouteflika*) dans le dictionnaire des noms de famille. L'automate vérifie s'il existe un annonceur de personne dans la phrase. Si c'est le cas, l'automate parcourt la liste des mots de la phrase, pour récupérer toutes les relations syntaxiques qui ont comme tête cet 'annonceur'. Ces relations remontées ainsi que les

éléments liés avec l'annonceur sont parcourus après par l'automate jusqu'à ce qu'il trouve un nom propre (dans ce cas, il s'agit d'une EN), ou bien qu'il n'y ait plus de relation. L'analyse syntaxique de la phrase citée donne le schéma de la figure (4.3). Les relations détectées sont

- ✓ *AnnpAdj :* désigne la relation modificative entre l'annonceur (الرئيس - le président) et l'adjectif (الجزائري - algérien), tout en considérant l'annonceur comme la tête de la relation syntaxique et l'adjectif étant le dépendant.
- ✓ *PrenomNP :* relation entre le prénom عبد العزيز (AbdAziz) et le NP بوتفليقة (Bouteflika) avec le NP comme tête de la relation
- ✓ *AnnpNP :* désigne la relation syntaxique qui relie l'annonceur (président) avec le NP (Bouteflika) avec comme tête le nom Bouteflika

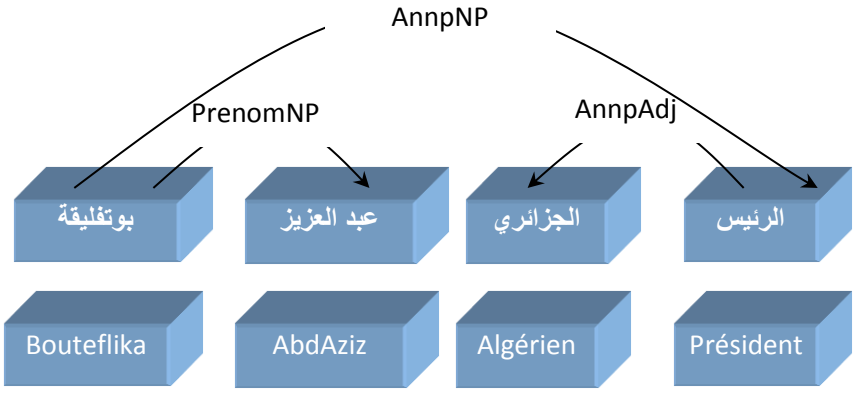


Figure 4. 3. Analyse syntaxique de la phrase الرئيس الجزائري عبد العزيز بوتفليقة

Le résultat de l'application de la règle est donné dans la figure (4.4).

- **$ {title} $ {adj_nationality?} ${ firstname } ${unknown} → <en entype="pers">**
  Cette règle est appelée dans le cas où le nom de famille ne figure pas dans le dictionnaire lors de l'application de la règle précédente. Par exemple, si dans la phrase : 'الرئيس الجزائري عبد العزيز بوتفليقة', le nom de famille بوتفليقة (Bouteflika) n'est pas détecté dans le dictionnaire alors cette règle est appelée et commence par détecter l'annonceur lexical de type titre (dans cette phrase c'est le mot الرئيس 'le président', ensuite la règle tente de détecter d'éventuel adjectif de nationalité (dans l'exemple c'est le mot الجزائري (algérien), après l'expression {firstname} permet de repérer un prénom connu comme le prénom عبد العزيزdans notre cas. Enfin, l'emploi de l'expression de repérage des mots inconnus {unknown}, permet de repérer le nom de famille Bouteflika et de terminer l'opération de repérage de cette EN.

```
<en entype="pers">
 <relation reltype="AnnpNP">
  <head>
   <posBeg>27</posBeg>
   <lemma>بُوتَفْلِيقَة</lemma>
   <catPos index="no">+np</catPos>
   <mCat>NP</mCat>
   <posEnd>35</posEnd>
  </head>
  <dept>
   <posBeg>0</posBeg>
   <lemma>رَئِيسّ</lemma>
   <catPos index="no">+annppers</catPos>
   <mCat>S</mCat>
   <prop index="no">+pers+ms+adjnom</prop>
   <posEnd>6</posEnd>
  </dept>
 </relation>
 <relation reltype="AnnpAdj">
  <head>
   <posBeg>0</posBeg>
   <lemma>رَئِيسّ</lemma>
   <catPos index="no">+annppers</catPos>
   <mCat>S</mCat>
   <prop index="no">+pers+ms+adjnom</prop>
   <posEnd>6</posEnd>
  </head>
  <dept>
   <posBeg>7</posBeg>
   <lemma>جَزَائِرِيّ</lemma>
   <catPos index="no">+adj</catPos>
   <mCat>J</mCat>
   <prop index="no">+ms+loc+nat</prop>
   <posEnd>15</posEnd>
  </dept>
 </relation>
 <relation reltype="PrenomNP">
  <head>
   <posBeg>27</posBeg>
   <lemma>بُوتَفْلِيقَة</lemma>
   <catPos index="no">+np</catPos>
   <mCat>NP</mCat>
   <posEnd>35</posEnd>
  </head>
  <dept>
   <posBeg>16</posBeg>
   <lemma>عَبْد العَزِيز</lemma>
   <catPos index="no">+prenom</catPos>
   <mCat>NP</mCat>
   <posEnd>26</posEnd>
  </dept>
 </relation>
</en>
```

Figure 4. 4. Les résultats d'application de règle d'extraction sur la phrase
الرئيس الجزائري عبد العزيز بوتفليقة

4.6.2. Identification des lieux

Pour la reconnaissance des noms de lieu, nous suivons la même stratégie que celle utilisée pour les noms de personnes. Tout d'abord, nous commençons par recueillir la liste des preuves internes (noms de lieux) en se basant sur les mêmes ressources déjà mentionnées. Nous notons que les ressources concernant les lieux géographiques dans le monde sont plutôt stables et généralement il convient de construire une liste des noms de lieux les plus connus, comme les noms de pays et ceux des principales villes dans le monde. Dans notre dictionnaire nous avons considéré en plus des noms de pays et de villes les noms de montagnes, de rivières, etc. En plus de cette liste de lieux, nous avons ajouté la liste de gentilés déjà utilisée pour la reconnaissance des noms de personnes. Voici quelques exemples de lieux issus de notre dictionnaire :

- ✓ *Les pays :* الجَزَائِر الجَزَائِر`+NP'`+LOC'`+COUNTRY'
- ✓ *Les villes :* عَنَابَة عَنَابَة`+NP'`+LOC"
- ✓ *Les mers* :البَحْر الأَبْيَض المُتَوَسِّط البَحْر الأَبْيَض المُتَوَسِّط`+NP'`+LOC'`+MER'

Ensuite, nous avons énuméré une liste de 85 annonceurs de lieux (mots déclencheurs) comme : دَوْلَة*dawlat* 'pays', مَدِينَة*madiynat* 'ville', شَارِع*šaAri'* 'Avenue', سَاحَة*saHat* 'Place', نَهْر*nahr* 'fleuve', جَبَل*jabal* 'mont', etc. Ces marqueurs lexicaux sont utilisés comme des éléments dans les règles de reconnaissance.

Les règles de reconnaissances des lieux, éditées manuellement, permettent d'identifier et typer les ENs quelle que soit la simplicité ou complexité de leur structure. Ces règles sont basées sur les ressources décrites ci-dessus en plus et des relations syntaxiques. Voici les cas d'EN pour lesquelles nous avons élaborées des règles de reconnaissance des ENs de lieux :

- Les noms de lieux avec preuve interne uniquement : tels que : فَرَنْسَا *faransa* 'France'.
- Les noms de lieux avec preuve externe : tels que : مَدِينَة وَهْرَان الجَزَائِرِيَّة *madiynat wahran al-jaza'iriya* 'la ville algérienne d'Oran', République Démocratique de Congo.
- Les noms de lieux accompagnés d'un point cardinal : tel que : جنوب شرق آسيا *januwb šarq AsyaA* 'Sud-Est de l'Asie'.
- Les noms de lieux accompagnés de noms de personnes : tels que : حي فضيلة سعدان *Hay FaDiylah SaadaAn* 'Cité de Fadhela Saâdane', شارع محمد البوعزيزي *šaAri' MuHamad al-bu3ziyziy* 'Avenue Mohamed Bouazizi'.
- Les noms de lieux accompagnés de dates 1945 ماي 8 سَاحَة *saHat 8 mai 1945* 'Place du 8 mai 1945'.

4.6.3. Les noms d'organisation

4.6.3.1. Structure des noms d'organisation

Les noms d'organisation représentent une partie assez importante de l'ensemble des ENs et sont caractérisés par leurs variétés et par leur durée d'utilisation (apparition et disparition) qui dépend de la situation dans le monde. Ces caractéristiques rendent l'identification de ces noms d'organisation assez difficile et par conséquent la tâche de reconnaissance des ENs de type organisation semble délicate. Différents facteurs se conjuguent pour rendre l'identification des ENs délicats, parmi lesquels nous citions :

- L'utilisation de ces noms d'organisation peut être avec ou sans annonceur. Ceci entraîne une alternance entre l'usage d'une forme longue et d'une forme courte de son

nom. Par exemple (mwunaZamat aalaaumam al muttahidal « Organisation des Nations Unies» qui est une forme longue, peut exister dans un autre texte avec une forme plus courte comme /alaaumam aalmuttahidal « les Nations Unies ».

- La structure des noms d'organisation en arabe, à l'instar des autres langues, peut être simple (contenant un seul mot) ou complexe (contenant deux mots ou plus).

- Les noms d'organisation en arabe peuvent combiner dans leur structure des mots arabes avec des mots en provenance d'autres langues (essentiellement du français ou de l'anglais). C'est le cas du nom de l'organisation رأس الخيمة سيراميكس *Raas aal khay-mat siramyiks*.

- Des noms d'organisation peuvent parfois être formés simplement du nom et du prénom d'une personne, ce qui crée une ambiguïté sur la nature de l'EN. Sans éléments contextuelles cette ambigüité est très difficile à résoudre.

Dans le tableau suivant, (Zaghouani, 2009) a résumé des cas d'utilisation des noms d'organisations dans les textes arabes.

Modèle du nom propre d'une organisation	Exemple avec translittération de l'arabe	Traduction littérale
Nom de personne	منى إبراهيم (muna ibrahiym)	Mona Ibrahim
Nom de personne + type de profession en anglais	المدني تايلورز (aal madanyi tay-lwurzl)	Les tailleurs Al Madani
Nom commun simple	تبريد (tab-ryid)	Refroidissement
Nom de personne + type de produit	رأس الخيمة سيراميكس (Raas aal khay-mat siramyiks)	Les céramiques Ras Al Khaima
Nom composé complètement en arabe	مطار دبي الدولي (mataar dubay aal dualyi)	L'aéroport international de Dubaï
Nom de personne + type de produit	محمد داوود بياس أوتو (Muhamm-id dawud biyaas aauTu)	Mohamed Daoud pièces auto
Usage de l'arabe et l'anglais en même temps	شركة حبة البركة و كو (sharikat habbat aal barakat aand kwu)	La société Habbat al baraka & compagnie

Tableau 4. 1. Illustration de quelques noms d'organisation en arabe

4.6.3.2. Identification des noms d'organisation

L'identification des noms d'organisations, des compagnies et des noms des gouvernements commence par l'élaboration d'un dictionnaire contenant environ 1000 noms d'organisations telles que سوناطراك *sounaAtraAk* 'Sonatrach' ou جامعة الدول العربية *jaAmi'at al-duwal al-'arabiya* 'Organisation des Nations Unies'. Ces noms sont reconnaissables par le bais de l'étiquette *<NP+ORG>*. La forme des références contenues dans notre dictionnaire sont comme suit :

- رويترز رويترز`+np'`+org'
- بي.بي.سي بي.بي.سي`+np'`+org'

La seconde étape consiste à recenser une liste de déclencheurs (au nombre de 48). Parmi ces déclencheurs nous citons : منظمة*munaDamat* 'organisation', مؤسسة*muwassasat*

‘compagnie’, شركة*šarikat* ‘société’, جمعية*jam'iyyat* ‘association’, etc. Ces déclencheurs sont utilisés pour la description des règles de reconnaissance. Parmi les cas identifiés par ces règles nous citons :

- les noms d'organisations avec une preuve externe simple tel que : شركة ألستوم *šarikat Alstom* ‘La compagnie Alstom’;
- Les noms d'établissement institutionnels (école, universités, instituts, facultés, etc.), tel que : كلية الطب *kulliyyat aT-Tibb* ‘Faculté de Médecine’ ;
- les noms de ministères et d'organisations internationales tel que : المنظمة العالمية للصحة al-*MunaDDamat al-‘aAlamiyyat lil-Sihat* ‘Organisation Mondiale de la Santé’;
- Les noms d'organisations accompagnés d'un nom de personne tel que : جامعة باجي مختار *jaAmi’at Baajiy Mokhtar* ‘Université de Badji Mokhtrar’ ;
- Les noms d'organisations accompagnés d'un nom de lieu tels quelconque comme dans : جَامِعَة بَارِيس*JaAmi'at baAriys* ‘Université de Paris’ ;
- Les noms d'organisations accompagnés d'un sigle tel que : مركز سي.آن.آر.آس *markaz al-siy.An.Ar.Ass* ‘le centre C.N.R.C : Centre National de la Recherche Nationale’.

4.7. Reconnaissance des expressions numériques – NUMEX

4.7.1. Identification des déterminants numériques

Les textes arabes sont caractérisés par l’utilisation de deux systèmes d’écriture des nombres : les chiffres arabes et les chiffres indiens. Dans les pays d'Afrique du Nord les chiffres arabes sont utilisés contrairement au pays arabes du Moyen-Orient en plus de l'Égypte et de l'Arabie Saoudite qui utilisent majoritairement les chiffres indiens. Toutefois quel que soit le système de chiffres, ces derniers s'écrivent de gauche à droite et se lisent de droite à gauche. Cette particularité doit être prise en compte lors des traitements automatiques de l’arabe sinon nous risquons d’avoir des difficultés lors de la construction des règles.

Type de chiffres	Transcription des chiffres
Chiffres arabes (Tunisie, Algérie, Maroc)	**0 1 2 3 4 5 6 7 8 9**
Chiffres indiens (Egypte, Arabie Saoudite, Moyen Orient)	٩ ٨ ٧ ٦ ٥ ٤ ٣ ٢ ١ ٠

Tableau 4. 2. Les différents systèmes numériques utilisés en arabe

Une autre forme de transcription des nombres en arabe consiste à les écrire en lettres et non pas en utilisant les systèmes décrits ci-dessus. Cette utilisation des lettres pour écrire les nombres complique leur identification. Les règles définies dans cette section traitent ce deuxième cas du moment que l’identification des nombres écrits dans les systèmes des chiffres est très simple. Les règles définies permettent d’identifier d’abord ces chiffres écrits en lettres et déterminer leur valeur correspondante. Par exemple le chiffre transcrit ‘ مائتان وسبعة وثلاثين’ *maA'ataAn wa-sab'atun wa-thalaAthyn* correspond au nombre ayant la valeur 237.

La reconnaissance des cardinaux écrits en toutes lettres est basée sur un lexique résumé dans le tableau suivant :

Chiffre écrit en lettres	Valeur
صِفْر, وَاحِد, إِثْنَان, ثَلَاثَة, أَرْبَعَة, خَمْسَة, سِتّة, سَبْعَة, ثَمَانِيَة, تِسْعَة	0, 1, 2, 3, 4, 5, 6, 7, 8, 9
عَشَرَة, عِشْرُون, ثَلَاثون, أَرْبَعُون, خَمْسُون, سِتّون, سَبْعُون, ثَمَانون, تِسْعُون	10, 20, 30, 40, 50, 60, 70, 80, 90
مَائَة, مَائَتَان, ثَلَاثَمَائَة, …	100, 200, 300, ….
أَلْف, مِلْيُون, مِلْيَار, بَلْيُون	1000, 1000000, 1000000000, 10^{12}

Ce lexique est stocké dans le dictionnaire utilisé par les règles qui contient en plus toutes les formes fléchies de ces nombres. A titre exemple, le cardinal اثنان*ithnaAni* 'deux') est représenté en إِثْنَان*ithnaAni* 'deux, au nominatif,' masculin, إِثْنَتَان*ithnataAni* 'deux, au nominatif, féminin', إِثْنَيْن*ithnayni* 'deux, accusatif, masculin' et إِثْنَتَيْن*ithnatayni* 'deux, accusatif, féminin'. En ce qui concerne les chiffres composés ne sont pas stockés dans le dictionnaire et leur reconnaissance est faite à l'aide de règles linguistiques. Voici quelques exemples des entrées de notre dictionnaire :

- صِفْر \0`+card'
- وَاحِد 1`+card'
- إِثْنَان 2`+card'
- ثَلَاثَة 3`+card'
- أَرْبَعَة 4`+card'
- أَلْف 1\0\0\0`+card'
- مِلْيُون 1\0\0\0\0\0\0`+card'
- مِلْيَار 1\0\0\0\0\0\0\0\0\0`+card'
- بَلْيُون 1\0\0\0\0\0\0\0\0\0\0\0`+card'

Pour la détection des nombres nous avons établi un ensemble de règles en fonction du type du nombre :

- *Cardinaux simples :* ces règles détectent les nombres unitaires par consultation du dictionnaire décrit dessus.
- *Les dizaines :* les règles développées pour cette catégorie concernent les déterminants compris entre 11 et 99. Lorsqu'il s'agit d'un nombre de dizaine ayant un chiffre d'unité non nul (dont la valeur n'est pas divisible par 10), une conjonction de coordination assure la liaison entre les deux. Pour exprimer les nombres de dizaines composés, nous faisons appel aux règles des cardinaux simples et nous concaténons les résultats tout en ignorant la conjonction.
- *Les centaines :* cette troisième catégorie concerne les déterminants compris entre 100 et 999. Cette règle fait appel aux autres règles précédentes pour déterminer les dizaines et les cardinaux simples.
- *Le reste des cardinaux :* cette catégorie inclue les cardinaux des milliers, millions et milliards. Les règles développées pour cette catégorie font appel aux règles des précédentes catégories tout en concaténant à chaque fois les résultats obtenus.

4.7.2. Identification des expressions numériques

Les expressions numériques sont généralement identifiables à l'aide de listes statistiques de mots déclencheurs tels que les unités de distances et de poids, ainsi que les noms de devises et leurs subdivisions potentielles, etc.

Les entités numériques incluent principalement les systèmes de mesures (poids, distance, volume, vitesse), les pourcentages, ainsi que les devises. La liste des entités numériques peut être plus longue

selon les définitions ; pour les besoins de ce -mémoire, nous nous contenterons des trois principales entités numériques, qui sont les systèmes de mesures, les devises et les pourcentages.

- **Les unités de mesure :** ayant le mètre comme unité de base. A ce niveau, nous avons répertorié les unités de mesure en regroupement tous les multiples ainsi que les sous-multiples tels que كِيلُومِتْر *kiyluwmitre* 'kilomètre, km', مِيلِيمِتْر *milliymitr* 'millimètre, mm', etc. Ces entrées lexicales servent pour la reconnaissance des distances simples (longueurs, largeurs, profondeurs, hauteurs, etc.) ainsi que les mesures composées telles que les mesures de volumes au moyen du mot clé مُكَعَّب *muka''ab* 'cube'. Notons aussi dans cet égard, que l'utilisation des abréviations des unités de mesure est fréquente, 5 كلمet 20 طن, etc.
- **Les unités de pourcentages :** Les règles de reconnaisse de ce type d'expressions sont les plus simples à mettre en œuvre puisqu'elles sont formées à l'aide d'un cardinal ou d'un nombre écrit en chiffre en arabe suivi du symbole de pourcentages « % » ou de la forme بالمائة*bil miyat* 'pour cent'.
- **Les unités monétaires :** En ce qui concerne ce type d'expressions, nous avons construit une liste des unités incluant des devises telles que دِينَار *diynar* 'Dinar' ou دُولَار *duwlaar* 'Dollar' ainsi que leurs subdivisons telles que سَنْتِيم *santiym* 'Centimes' ou مِيلِّيم *milliym* 'Millimes'. Dans les textes arabes, ces expressions monétaires sont caractérisées aussi par l'emploi des signes des symboles monétaires comme $ pour le dollar, ¥ pour le Yen et € pour l'euro.

Dans ce qui suit, nous illustrerons une des règles qui permet de repérer dans cet exemple une entité numérique.
$number $measure → <en entype = "mes">
Dans la règle ci-dessus, l'expression *number* permet de repérer un nombre précédant l'unité de mesure, tandis que l'expression *measure* entre accolades renvoi à la liste d'expressions de mesure que nous avons préalablement compilée. Cette règle simple permet de repérer systématiquement des expressions comme : 85 kg.

Voici un exemple illustratif de notre analyse. ثمانون ألف متر مربع 'quatre-vingt mille mètre carré'.

<en entype="mes"><relation reltype="mesure"><head><posBeg><POS=18></posBeg><lemma>مِتْر مُرَبَع</lemma><catPos index="no">+**unitmes**</catPos><mCat>S</mCat><posEnd><POS=26></posEnd></head><dept><posBeg><POS=6></posBeg><lemma>80000</lemma><catPos index="no">+**card**</catPos><mCat>Num</mCat><prop index="no">+adjnom</prop><posEnd><POS=17></posEnd></dept></relation>**</en>**

Partie III : Traitement des dialectes arabes

Chapitre 5 Analyse phonologique

Introduction

L'étude et la compréhension de la morphologie dialectale de la langue arabe, ou d'une autre langue, passe nécessairement par une bonne compréhension de sa phonologie. De plus, les différences phonologiques entre l'arabe dialectal et l'arabe standard portent essentiellement sur le système vocalique et consonantique de la langue. C'est la raison pour laquelle nous présentons et discutons dans cette section les préliminaires phonologiques qui s'appuient essentiellement sur des systèmes consonantiques et vocaliques. De ce fait, nous présentons et comparons dans la section 5.1 les systèmes consonantiques de l'arabe standard (MSA) et de l'arabe dialectal, ensuite nous mettons en avant leurs systèmes vocaliques qui doivent être distingués dans 5.2. Finalement, nous passons en revue dans la section 5.3, les alternances phonologiques, appelées aussi les variations ou dégradations phonologiques à savoir : l'assimilation, métathèse, l'emphase, épenthèse, élision, le raccourcissement.

5.1. Système consonantique

Rappelons que l'alphabet orthographique de l'arabe standard (MSA) comprend vingt-huit lettres qui représentent vingt-huit consonnes, mais trois d'entre elles sont également utilisées comme des voyelles. Ce n'est pas le cas généralement pour les dialectes, par exemple le dialecte égyptien (EA) ne comprend que vingt-six de ces consonnes : les interdentales du MSA, en l'occurence les / θ, / et /ð/, sont inexistantes dans l'EA, elles sont remplacées dans certains mots par les arrêts dentaires correspondant /t/ et /d/, respectivement, et dans d'autres mots par les correspondants alvéolaires fricatives /s/ et /z/, respectivement (voir tableau 5.1).

MSA	EA	Traduction
θaman(-un)	taman	Prix
ðahab(-un)	dahab	Or
θaabit(-un)	saabit	Fixe
ðakiyy(-un)	zaki	Intelligent

Tableau 5. 1. Exemple de changement des interdentales entre le MSA et le EA

Cet exemple illustre les modifications et les altérations que peut subir le système consonantique de l'arabe dans les dialectes. Ces modifications ne datent pas forcément d'aujourd'hui mais existent depuis longtemps et elles étaient même repérées par les grammairiens arabes dans les dialectes de leur temps. Toutefois, il est à noter aussi que certaines modifications sont dues aux récents progrès technologiques caractérisés par une utilisation des moyens de communications multi-langages, voir au dernières compagnes coloniales marquées par une influence de la culture et la langue du colonisateur (le français au Maghreb et l'anglais pour le Machrek) sur les dialectes des populations colonisées. Dans le reste de cette section, nous donnons une description détaillée de la prononciation des deux consonnes ق *qaf* 'q' et ج *jim* 'j' massivement utilisées dans les dialectes orientaux et maghrébins. Nous illustrons ces prononciations dans différents dialectes : égyptien, algériens, tunisiens, etc.

5.1.1. Prononciation de la consonne qaf

La consonne ق 'q' est l'un des sons qui présente une grande variété de prononciation dans les dialectes arabes. Ces variations peuvent être perçue entre les régions, les villes, et même entre les localités. Le son issue de la prononciation du « q » en arabe littéral, peut être perçu comme : [q,ʾa, k, ou g]. Cette consonne *occlusive* peut être *uvulaire sourde* «ق» 'q' dans certaines dialectes, comme ceux d'Alger, de Constantine, ou de Tunis ; *palatale sonore* «ڤ»

‘g’ dans d'autres dialectes, ce qui est le cas des dialectes d’Annaba, de Sétif, ou celui de Gafsa; ou *glottale sourde* ء (?,’), comme c’est le cas dans les dialectes égyptien et celui de Tlemcen. Notons dans le cas des dialectes n’utilisant pas la consonne *occlusive glottale sourde*, il existe quelques mots qui sont prononcés de la même façon quel que soit le dialecte, par exemple le mot «vache » est toujours prononcé بَقْرَة *bagra*.

Ces variations peuvent être aussi considérées, selon (Lajmi, 2009), comme une propriété qui traduit un clivage sociogéographique entre parler citadin et parler rural et encore parler bédouin. Selon ce clivage, (Cantineau, 1960) propose une classification des parlers pour les dialectes modernes comme suit :

- ***Les parlers sédentaires :*** les parlers dans lesquels l'ancien qâf est représenté par une sourde (q, k, ’). Nous pouvons géographiquement les répartir en trois groupes, suivant que le qâf est prononcé q, ', ou bien k :
    - Les parlers ayant un qâf *vélaire*, donc q, couvrent des surfaces assez importantes, notamment en Syrie et en Afrique du Nord : c'est le cas du Sahel tunisien, des villes de Tunis et de Constantine, Milla, et la majeure partie de Skikda. Cette prononciation est aussi utilisée à Alger, Cherchel, Dellys, Blida, Miliana, Média, Ténès voir dans l’ouest algérien à Mostaganem. Nous trouvons aussi cette prononciation dans une grande partie du Maroc.
    - Les parlers ayant un qâf *réduit à une simple occlusion glottale* ‘?’ sont surtout des parlers citadins comme les habitants d'Alep, Lattaquié, Hama, Homs et Damas en Syrie, Tripoli, Beyrouth, Saîda et quelques régions montagnardes au Liban, Safed, Haïfa, Jaffa, Jérusalem, Hébron et Ghaza en Palestine, Alexandrie et Le Caire en Egypte, Tlemcen en Algérie ; et Fès au Maroc.
    - Les parlers ayant un qâf prononcé *k postpalatal* sont ceux qui ont également une altération inconditionnée du kâf : par un processus tout à fait analogue d'avancement du point d'articulation, le qâf vélaire est devenu un k postpalatal. Ces parlers disent *kalb* ‘’cœur' (de qalb-), *kâl* ‘dire’ (de qâla), *kahwa* ‘café’ (de qahwat-), etc. Ces parlers sont ceux des sédentaires de Palestine, de l'oasis de Sukhne en Syrie, de la Petite Kabylie, Jijel, des Msirda et des Trara au Nord de Tlemcen en Algérie.
- ***Les parlers nomades :*** les parlers dans lesquels il est représenté par une sonore (g). Nous distinguons pour ces parlers plusieurs groupes comme suit :
    - Un premier groupe possède un gâf très en arrière, presque vélaire, mais non en toute position. Ces parlers sont assez rares on les retrouve dans l'Arabie du Nord et le Sud tunisien.
    - Un autre groupe a un *gâf post-palatal* en toute position. Ce groupe contient les parlers nomades d’Algérie, Maroc et Tunisie ; et en orient les populations nomades de l’ouest de l’Irak et l’est de la Jordanie ainsi qu’une majeur partie du Yémen et Oman.
    - Un troisième groupe, celui des *parlers de nomades nord-arabiques*, a un traitement du *gâf* absolument parallèle à celui du kâf, c’est-à-dire que le gâf se maintient au voisinage des voyelles postérieures u, o, a mais subit des altérations conditionnées au voisinage des voyelles antérieures i, e, ä, passant aux affriquées g (==dj) chez les petits nomades syro-mésopotamiens, et g (==dz) dans les grandes tribus arabiques. Ces affriquées sont senties comme des variantes combinatoires de g et forment avec lui un phonème unique.

L'avantage de cette répartition est qu'elle ne souffre pas de véritables exceptions. Si ces exceptions existent, comme pour certains mots des parlers nomades de l'Afrique du Nord ayant un qâf sourd : *qrâ* 'il a écrit' ou *bqâ* 'il est resté', elles paraissent des emprunts soit à la langue classique, soit à la langue des villes. De même les parlers sédentaires de la même région contiennent tous quelques mots ayant un gâf sonore, comme pour *gnîn* 'lapin' ou *gorba* 'outre', qui paraissent des emprunts aux parlers ruraux.

Les parlers maghrébins, tant de sédentaires que de nomades, ont en général, en face du qâf classique, deux phonèmes : un q vélaire sourd et un g post-palatal. Naturellement un seul de ces phénomènes : q chez les sédentaires, g chez les nomades, représente dans le dialecte en question l'évolution phonétique normale du qâf ancien; l'autre phonème n'apparaissant que dans des emprunts. De ce fait les prononciations q et g servent parfois à différencier deux sens d'un même mot formant ainsi pour ces mots des doublets ou des paires de mots, l'un ayant un q l'autre un g; c'est ainsi que nous aurons *begra* et *baqra* 'vache', *gubba* et *qobba* 'coupole de marabout; alcôve', *zreg* 'gris (chevaux)' et *zroq* 'bleu' ; *sherg* 'l'orient' *shorq* 'le pèlerinage', *gleb* 'vomir' et qleb 'renverser', *bgâ* 'être exténué de fatigue' et *bqâ* 'rester ', etc. Au contraire, dans les parlers orientaux, des doublets de ce genre ne se produisent pas.

Sur un autre registre, un qâf ancien peut se dissimiler en *k* devant un *t*. Par exemple, pour beaucoup de parlers, tant orientaux que maghrébins, le verbe 'tuer', cl. *qatala* est passé à *katal* au Maghreb ou *Ktäl*.

5.1.2. Prononciation de la consonne jim

Dans cette section nous nous intéressons à la prononciation de la lettre ج *jim* 'j' dans les dialectes arabes modernes, en introduisant avec des illustrations les différentes variantes à travers les régions du monde arabe. Les deux prononciations les plus fréquemment attestées sont la prononciation ǧ (= dj) et la prononciation ž (= j français). En plus des variantes déjà citées, une autre prononciation se trouve en Egypte où le ج est prononcée «g», et à titre d'exemple le mot *gabal* 'montagne' ou *negma* 'étoile' sont des prononciations des mots arabes 'جبل' et 'نجم' respectivement.

Au moyen orient, la variante ǧ est très répandue au Yémen, en Irak, dans le désert syrien, dans les campagnes palestiniennes, syriennes et transjordaniennes. La variante ž quant à elle est considéré comme une prononciation citadine très utilisée à Damas, Beyrouth, Haïfa, Naplouse, Jérusalem, Jaffa, Ghazza. Elle-même la plus utilisée de tout le Liban.

En Afrique du Nord, la prononciation ž est de loin la plus répondue : nous la trouvons à Tripoli, en Tunisie, au Maroc et une partie de l'Est algérien (Annaba, Guelma, Tebessa, Souk-Ahras, etc.). Elle est aussi utilisée par la plupart des nomades. Pour ce qui est de la prononciation ǧ, elle est attestée d'une façon régulière que dans une partie de l'Algérie : à Constantine, Sétif, Jijel, Barika, Bejaïa, Alger, tout le Tell, Oran, Mostaganem, Mascara, et enfin la ville de Tlemcen. Nous la trouvons aussi dans certains endroits, pour des cas de gémination, au Maroc comme à Tanger.

5.1.3. Prononciation des spirantes interdentales

Selon le principe suivant : *les spirantes interdentales dans les dialectes modernes de l'arabe sont conservées telles quelles : les t , d, ḍ, dans les parlers de nomades ou d'anciens nomades; passent aux occlusives correspondantes t, d, ḍ, dans les parlers sédentaires*.

L’application de ce principe donne les cas remontés dans les sections suivantes.

En Orient, dans les villes ayant un parler de sédentaires ce principe est particulièrement attesté. C’est le cas des villes : Le Caire, Alexandrie, Jérusalem, Damas, Alep, Bagdad. Cependant il est moins appliqué dans les campagnes : les parlers campagnards ayant une prononciation ‘q’ du ق, donc essentiellement sédentaires, ont conservé les interdentales. Nous trouvons cette tendance en Palestine, le sud du Liban; par contre l'inverse ne paraît pas être vrai, et aucun parler de nomades se semble avoir perdu la prononciation spirante des interdentales. Nous signalons aussi un phénomène de passage des spirantes interdentales aux spirantes labiodentales dans certaines communes comme c’est le cas à Palmyre, *felğ* au lieu de *ṯelğ* ‘neig’. Le phénomène inverse peut se produire : dans beaucoup de parlers orientaux, ‘la bouche’ (cl. Fum) se dit *ṯum* ,pluriel du *ṯmâm*.

Au Maghreb, les faits se présentent d'une manière analogue, c'est-à-dire que certains parlers de sédentaires peuvent avoir des interdentales, en dépit du principe posé ci-dessus, mais que l'inverse ne paraît guère se produire. C'est ainsi qu'en Tunisie, les parlers sédentaires du Sahel (type de Takrouna) ont des spirantes interdentales ainsi que la ville de Tunis. En Algérie, à Constantine, les spirantes interdentales sont souvent devenues occlusives dans toute la zone des parlers sédentaires qui couvre la commune de Collo, Skikda et Constantine, El-Milia, Jijel, Bougie.

Dans la wilaya d’Alger, les communes à parler sédentaires, à l’exception d’Alger, les spirantes interdentales sont conservées à cause probablement de l'influence des parlers de nomades, nous pouvons de ce fait qualifier cette conservation de restitution. Ce constat est mis en exergue dans les travaux de (Cantineau, 1960), et identifié dans les villes Cherchel, Blida, Médéa, Miliana et Ténès. Dans la ville d’Alger les spirantes interdentales sont occlusives. Dans la wilaya d'Oran les spirantes interdentales sont passées aux occlusives à Tlemcen seulement au Nord de la ville. C’est le cas aussi au Maroc où les parlers de sédentaires, citadins comme montagnards, font passer les spirantes interdentales aux occlusives. Il est à noter aussi que dans certains endroits en Algérie, comme pour les parlers nomades de la wilaya de Mostaganem, les spirantes interdentales passent aux spirantes labiodentales, à titre indicatif prenons les exemples suivants : ṯâni ‘aussi’ > fâni, ḏhab ‘or’ > vhab, ḍ̠alma ‘obscurité’ > valma.

Un autre fait important caractérisant les parlers sédentaires du Maghreb dans leur traitement des interdentales, c’est l'emphatique ḍ au lieu de passer ḍ s'assourdit en ṭ comme dans les mots ṭahro ‘son dos’; ṭlêla ‘ombre’, byaṭ ‘blanc’, mrêṭ ‘malade’, ṭofro ‘son ongle’, etc. Ce phénomène a une extension moins grande que la réduction des spirantes interdentales, et il n'est presque jamais réalisé complètement. Il est contraint par des limitations dues, soit à l’arabe classique, soit aux parlers de nomades avoisinants.

Nous pouvons aussi signaler une autre caractéristique des spirantes interdentales due à des altérations combinatoires de ces dernières. Elle consiste en une emphase de la sonore ḏ en ḍ̠ au voisinage d'une emphatique ou d'une vélaire. Cette caractéristique peut s’expliquer par des causes phonétiques régulières, l’influence des consonnes voisines, voir l’influence de la langue berbère. Cette caractéristique est très présente en Algérie où nous trouvons les exemples suivants : fḫaḍ̠ ‘cuisse’ (cl. fḫaḏ); ḫḍ̠a ‘prendre’ (cl. 'aḫaḏa), ḍ̠örwok ‘maintenant’ (ḏâl-waqt). (Marçais, 1908).

Enfin, dans certaines villes, comme Saïda en Algérie, les spirantes interdentales

s'assimilent très fréquemment à un t qui les suit, pour illustrer cette propriété nous citons les exemples suivants : ḫrôtt ‘j'ai labouré’ (<haraṯtu), gböttäh ‘je l'ai saisi’ (cl. qabadtuhu).

5.1.4. Traitement du hamza

La lettre hamza, peut être considérée comme un élément discriminant des deux groupes de parlers, ceux du Machrek et ceux du Maghreb. Nous donnons dans cette section l'évolution phonétique et les changements qu’a subit cette consonne chez les deux groupes de parlers.

Concernant les parler du Machrek, nous mentionnons les travaux de (Cantineau, 1960) sur ce sujet où l’auteur considère que : « le hamza, quoique affaibli, est resté un phénomène au sens phonologique du mot, un élément constitutif important du système consonantique de ces parlers. ». Donc, en fonction de la position de cette lettre dans un mot, elle peut avoir plusieurs états : inchangeable, modifiable ou supprimable. Nous illustrons dans les exemples suivants les différents cas selon la position de ce Hamza :

- A l'initiale du mot, le hamza est généralement conservé. Cette conservation affirme qu'elle garde généralement sa valeur d'une consonne radicale, par exemple : *'arnabe* ‘lièvre’, *'asba3* ‘doigt’ ; les pluriels de '*arâneb* et *'asâbe3*. Cependant, il existe des cas exceptionnels où il est changé en semi-voyelle w ou y comme c’est le cas des mots : *wallaf* ‘il plia bagage’, *waddab* ‘il corrigea’.
- A l'intérieur du mot, le hamza est, contrairement à la première position, rarement maintenu et souvent il a disparu pour faire place à un allongement de voyelle, comme dans les exemples suivants : *râs* ‘tête’ (cl.ra's), *bîr* ‘puits’ (cl.bi'r), *mara* ‘femme’ (cl.mar'at). Il passe aussi à w ou à y à l’instar des mots : *iTTâwab* ‘bâiller’ (cl.taTâ'aba), *lâyam* ‘convenir de’ (cl.lâ'ama), *malyân* ‘plein’ (cl. mal'ân), *Mîye* ‘cent’ (cl. mi'at). Il existe toutefois un cas démonstratif où cette lettre est maintenue, il s’agit du verbe *sa'al* ‘demander’.
- A la fin d’un mot le hamza peut avoir disparu parfois sans laisser de traces, ou être transformé, donnant ainsi plusieurs cas de figure comme suit :
  - *Le hamza supprimé* : par exemple *ghadâ* ‘déjeuner’ (cl.ghadâ'), *samâ* ‘ciel’ (cl.samâ')
  - *Le hamza remplacé par la semi-voyelle « y »* : c’est le cas des verbes à 3ème radicale hamza sont tous devenus des verbes à 3ème radicale y.
  - *Le hamza assimilé à une consonne précédente* : comme dans le mot *daww* ‘lumière’.

Quant aux parlers du Maghreb, le fait marquant est que le hamza a presque disparu et que les occlusives glottales, que nous pouvons entendre, n'apparaissent que dans des emprunts à la langue littéraire. Ainsi, dans les différents dialectes maghrébins, le hamza est soit tombé en complètement désuétude (disparu), soit remplacé comme dans les parlers du Machrek par une semi-voyelle w ou y. De ce fait le hamza subit différentes opérations en fonction de sa position dans le mot comme suit :

- A l'initiale, le hamza perd généralement toute valeur consonantique propre, générant par conséquent plusieurs cas de figure comme suit :
  - Le hamza est totalement tombé, prenons les exemples suivants : *bell* ‘chameaux’ (cl. 'ibil), *bra* ‘aiguille’ (cl. 'ibrat-), Nous pouvons admettre que dans ces mots, le hamza existe virtuellement; mais il n'est nullement prononcé. Selon (Marçais, 1902), lorsque l'accent portant sur une syllabe subséquente, la voyelle à laquelle était rattaché le hamza initial disparait aussi, qu'elle fut contenue dans une syllabe ouverte ou fermée : ابراهيم *brâhim*, *briq* ابريق ‘aiguière’, أمارة *mâra* ‘signe’. La conservation virtuelle du hamza sous forme de simple voyelle, bien

qu'il n'ait pas l'accent dans les mots : *islâm*, *imâm*, *amân*, *amer* (cf. sur l'allongement de a) s'explique par des influences de la langue littéraire. Dans un certain nombre de mots, il s'est réduit à une simple voyelle a, u, i; sous cette forme il s'est maintenu, là où il portait l'accent : أصل *Âsl* 'origine', أرض *ÂrD* 'terre', أنا *Âna* 'moi', أمان *Âmân* 'sécurité', أخرى *ukhra* 'autre', etc.

- Le hamza peut donner naissance à une semi-voyelle 'w' ou 'y' dans les mots où il portait l'accent. A titre illustratif prenons les mots suivants : وكّل *wukkel* 'faire manger' du verbe أَكَّلَ *Âkkal*, ولّف *wullef* 'habituer' du أَلَّفَ *Âllaf*, يبرة *yebra* 'aiguille' du إبرة *Ibrah,* ينس *yens* 'espèce humaine' du إنس *Ins,* يامس *yâmes* 'hier' du أمس *Âms*
- Le hamza est remplacé avec un 'l' initial dans une forme indéterminée dérivée d'une forme déterminée. Voici quelques exemples لَفْعَى *lef'a* 'vipère', لَنْجَاصْ *lenjâs* 'poire', لَرْضْ larD 'terre'.
- Le hamza est renforcé en 'h' comme dans les mots هَجَّالة *hajjâla* 'veuve' du mot أَجَّالة *Âjjâla* (de même dans tout le Maghreb), ou comme dans la locution conjonctive همّالا *hammâla* 'cependant', أمّالا *Âmmâla*.

- A l'intérieur du mot, le hamza disparaît pour laisser place à un allongement de voyelle comme pour les mots : فاس *fâs* 'pioche' pour فأس *fa's*, راس *râs* 'tête' du رأس *ra's*, d'oùle pluriel روس *rôs* du رؤوس *ru'ûs*, ذيب *dîb* 'chacal' pour ذئب *Di'b*, بير *bîr* 'puits' pour بئر *bi'r*; pluriel du بيار *byâr* du mot بئار *bi'ar* et مومن *mûmen* 'croyant' pour مؤمن *mu'min*, توام *twâm* 'jumeaux' du توأم *taw'am*, مليان *malyân* 'plein' du ملآن *mal'an*, فواد *fwâd* 'viscéres' du mot فؤاد *fu'âd*. Il peut être aussi renforcé en ه 'H' comme dans le mot زهر *zhôr* 'rugir' du mot زأر *za'ar*. Notons enfin qu'il existe une exception où le hamza est conservé. Cette exception est le mot قرآن *qor'ân* 'Coran' et possède une très curieuse prononciation proche de celle de l'arabe littéraire.

- En fin du mot, le hamza est soit tombé, comme pour le mot شُركاء *šorka* 'partenaires', soit réduit à une voyelle longue, par exemple : *brâ* 'guérir' du *bari'a*, *qrâ* 'lire' du verbe *qara'a*, *smâ* 'ciel' du *samâ'*, soit il s'assimile à une consonne précédente comme : *Daw(w)* 'lumière' du mot *Daw'*, *šay* 'chose' du mot *šay'*, ou donne un y qui finalement se déconsonnantise en y, c'est le cas du mot *bennây* du mot بنّاء *binnâ'*.

5.1.5. Autres cas de prononciation particulière :

Nous terminons cette partie concernant le système consonantique par la présentation d'un ensemble de cas de prononciation particulière :

- En Tunisie et dans l'est algérien et Tlemcen, la lettre غ *Ghin* 'ɣ' est remplacée par un 'kh' dans certains mots, par exemple les mots *ɣsal* (laver), ou *khsîl* (linge lavé).
- Dans beaucoup de dialectes, la lettre غ *ghin* est substituée par un ع *'ayn* 'ʿ' comme c'est le cas dans la racine du mot *ghamq* 'profond' du mot *'amiiq*.
- Dans certaines régions du nord-est du Sahara algérien, comme M'sila et Bou'Saâda, la lettre '*ɣ*' est remplacés par la lettre '*q*'. Par exemple les mots *ɣaliy* 'cher', *ɣmazli* 'm'a clignoté', *sɣayera* 'petite', sont prononcés respectivement : *qali*, *qmazli* et *sqayera*.
- La lettre ه 'h' disparaît fréquemment dans la conjugaison du verbe (râ) avec les pronoms (hu, hi hûm,) = (râhu, râhi, râhum); qui deviennent (râ, raî et raûm). Il disparaît aussi dans la locution adverbiale منّا *menna* 'par ici' au lieu de من هنا *min hounna*; dans le pluriel فواكي *fwâki* 'fruits' au lieu de فواكه *fawâkih*. Il y aussi le mot وجّ *wujj* 'visage' du mot وجه *wijh* est à rapprocher des égyptiens et syriens وجّ

wujj et وشّ *wušš*.

- Permutation du sin, Sad, Zad. Nous constatons qu'en tlemcenien des permutations des sifflantes sad, sin, zad existent. Nous en trouvons en arabe classique et dans la plupart des dialectes. Certaines sont dues à des causes phonétiques comme l'influence de la consonne voisine, par exemple فازدة *fâzda* 'corrompue' au lieu de فاسدة *fâsda*, زدم *zdam* 'heurter' au lieu de صدم *Sdem*, سدر *sder* 'poitrine' à la place de صدر *Sder*. Ce phénomène a été expliqué par des influences vocaliques secondaires comme pour le mot *Sêf* 'sabre' au lieu de *sîf*.

5.2. Système vocalique

A l'instar des autres phénomènes indiqués ci-dessus, nous décrivons dans cette partie le système vocalique des dialectes qui est en quelque sorte une évolution du système phonologique de l'arabe classique. Le système vocalique de l'arabe exhibe un système triangulaire simple constitué de trois voyelles courtes et trois longues. D'un point de vue de l'orthographe, les voyelles courtes sont représentées par des signes diacritiques au-dessus ou en dessous de la lettre, tandis que les voyelles longues sont représentées par les trois lettres أ /ʔ*alif*/, ي /*yaa*ʔ/ ainsi que واو /*waaw*/. Les voyelles longues sont prononcées deux fois plus longtemps que leurs homologues courts.

Concernant le système vocalique des dialectes, (Barkat, 2000) a établi une typologie dialectale fondée sur l'opposition *: parlers maghrébins* et *parler orientaux*. Cette étude a montré que l'espace vocalique des parlers maghrébins est plus centralisé que celui des parlers orientaux, avec une différence de durée entre voyelles brèves et longues. Cette étude confirme l'hypothèse soutenue dans plusieurs travaux de recherches qui est que le système vocalique de l'orient est plus enrichi de timbre vocalique que son homologue (du Maghreb) qui est composé de trois voyelles cardinales ainsi que le *schwa*.

De ce fait, le système vocalique des dialectes arabes modernes des parlers d'Orient est composé de huit voyelles : trois brèves */i, u, a/ et cinq longues */ī, ū, ē, ō, ā/. (ii, uu, ee, oo, aa). L'émergence de nouvelles voyelles intermédiaires longues illustrent bien que les anciennes diphtongues /ay/ et /aw/ ont évolué dans les langues arabes dialectales respectivement en /ē/ et /ō/.

Dans la même optique, il a été observé que "le vocalisme bref se réduit de façon croissante d'Est en Ouest" (Marçais, 1977) jusqu'à devenir - dans certains parlers - de simples points vocaliques ultra-brefs, aboutissant ainsi à des réalisations ultra-brèves des voyelles, c'est le cas des parlers marocains.

Le tableau (5.2) présente les voyelles utilisées dans le système vocalique de l'arabe standard AS et celui de l'arabe dialectal du l'Egypte AE :

	Courte			**Longue**		
	Avant	Central	Arrière	Avant	Central	Arrière
Haut	I		U	ii		uu
Milieu				ee+[11]		oo+
Bas		A			Aa	

Tableau 5. 2. Les voyelles dans MSA & AE

[11] (+) = Trouvé dans AE seulement.

Dans les dialectes arabes les séquences /ay/ et /aw/ sont transformées en /ee/ et /oo/, respectivement. Nous pouvons observer cette transformation dans les exemples suivants :

MSA	**AD**	**Traduction**
bayt(-un)	beet	maison
ShayTanat(-un)	SheeTana	diable
Shaykh(un)	Sheekh	vieux
naw3(un)	noo3	espèce
lawn(-un)	loon	couleur
lawH(-un)	looH	tableau/ plaque

Les exemples montrent que les deux timbres /ee/ et /oo/ proviennent d'un waw et d'un yay classique. A cet égard, nous soulignons que pour la majorité des dialectes arabes la diphtongue disparaît dans tout un paradigme d'unités au profit d'une voyelle longue. Il est important de signaler, qu'il existe quelques régions du Maghreb où cette diphtongue est conservée, c'est le cas par exemple des parlers de la ville d'Annaba de l'est de l'Algérie. En plus de cette exception, nous remarquons qu'il existe aussi deux cas d'utilisation où la diphtongue est conservée sans changement. Tout d'abord, lorsque la voyelle est suivie d'une gémination. Linguistiquement, ce phénomène est appelé "inaltérabilité des géminées" (Gadalla, 2000). Voici quelques exemples illustratifs de ces cas :

MSA	**AE**	**Traduction**
mayyit(-un)	mayyit	mort
bayyaD(-a)	bayyaD	à la chaux, à peindre
bawwaab(-un)	bawwaab	un portier
Sawwar(-a)	Sawwar	photographier

D'autre part, lorsque la séquence de voyelles et de glissement se trouve dans la syllabe initiale en forme de radical, comme en ces termes:

MSA	**AE**	**Traduction**
?awTaan(-un)	?awTaan	pays
mawluud(-un)	mawluud	nouveau-né
?aymaan(-un)	?aymaan	serments
Saydal-at(-un)	Saydal-a	science pharmaceutique

En dialecte tlemcenien (Algérie), la diphtongue provient du phénomène secondaire du ressaut qui plus général en tlemcenien que dans le Maghreb oriental, donne des groupements comme *qahhawti* 'mon café', *mešše**i**tek* 'ta marche' proviennent de *qahwa*, *mešya* respectivement. Ce phénomène n'est pas connu dans le tripolitain et le tunisien.

En ce qui concerne l'allongement vocalique, (Mejri et al., 2009) avance que « *Le système vocalique du dialectal se distingue par un enrichissement des degrés d'aperture. Si l'arabe littéral ne comporte que trois voyelles brèves doublées de leurs correspondantes longues, le dialectal tunisien connaît une extension de l'action de la durée vocalique dans ce sens qu'on assiste à l'émergence de nouvelles paires minimales fondées sur un allongement vocalique non réalisé dans le littéral*.». Toutefois, nous signalons que ce phénomène concerne à la fois des unités n'appartenant pas aux mêmes parties du discours, comme par exemple : سِرْ *sir*

‘secret’ qui a une catégorie grammaticale ‘*nom*’ et سِيرْ *siir* ‘marcher’ qui est *un verbe à l’impératif*, et des paires appartenant à la même partie du discours comme par exemple : les deux verbes يْسِلْ *ysil* ‘tirer’ et يْسِيلْ *ysiil* ‘couler’. Ce phénomène est fréquent dans les dialectes maghrébins.

Il est à noter que le système vocalique des dialectes diffère d'une région à une autre comme par exemple le dialecte Sfaxien qui se caractérise par rapport au dialecte sahélien par une voyelle finale longue dans des mots qui portent l'accent sur la dernière syllabe. Le tableau suivant illustre cette caractéristique :

Mot en français	**Prononciation en dialecte Sfaxien**	**Prononciation en dialecte Sahélien**
Ciel	«سْمَا » [sma]	«سْمَاء » [smA]
Eau	« مَا » [ma]	« مَاء » [mA]

Tableau 5. 3. Exemple de différence du système vocalique entre les régions**.**

Nous pouvons identifier plusieurs types d’allongement par l’accent, en voici quelques exemples :

- L'allongement de la voyelle terminale dans les mots provenant de racine défectueuse ou possédant la lettre hamza comme la dernière radicale, comme pour les mots : *rDâ* ‘s’est contenté’, *qrâ* ‘il a étudié’, *hlû* ‘sucré’, *jdî* ‘chevreau’, etc.
- L'allongement de la voyelle des impératifs de verbes concaves, que nous trouvons dans tous les dialectes maghrébin. A titre d’exemple nous citons : قول *qôl* ‘dis’, زيد *zîd* ‘continue’, بات *bât* ‘passe la nuit’. Nous trouvons cet allongement aussi dans le cas des mots provenant de racines assimilées ou ayant la première radicale hamza, comme تيقة *tîqa* ‘confiance’, جيهة *jîha* ‘côté’, نيف *nîf* ‘nez’, etc.
- L'allongement de la voyelle dans la dernière syllabe du parfait à la 3ème personne au féminin des verbes, quand s'y adjoignent les suffixes vocaliques. Par exemple ضرباتك *Darbâtek* ‘elle t'a frappé’.
- D'un allongement de voyelle brève ou de semi-voyelle déconsonnantisée, par contre-accent : *Amân* ‘sûreté’, يهود *Ihûd* ‘Juifs’.

Le système vocalique des dialectes arabes est caractérisé aussi par l’absence des voyelles brèves : elle consiste en une disparition des désinences casuelles dans les noms et des flexions finales dans les verbes. Les différents dialectes arabes négligent les voyelles courtes en particulier quand ils se trouvent à la fin d'une syllabe. Voici quelques exemples montrant la différence de la prononciation des mots en dialecte et en arabe standard :

Prononciation en MSA	**Prononciation en dialecte**	**Traduction**
طَاوِلَةٌ [TAwilapN]	طَاوْلَة [TAwlap]	table
سَرَقَ [saraqa]	سْرَقْ [sraq], saraq	voler
حَلَفَ [Halafa]	حْلِفْ [Hlif], حِلِفْ [Hilif]	jurer

Par ailleurs, des différences existent au niveau de l’allongement entre les dialectes maghrébins et orientaux. A cet effet, le système phonétique des dialectales du Maghreb possède une caractéristique intéressante pour la reconnaissance (Saâdane et al., 2013) : il présente une succession de deux consonnes au début du mot. Cette caractéristique est beaucoup moins marquée dans le système phonétique des dialectes orientaux. Ceci se traduit

par une particularité notable dans le schème verbal « fʿel » au Maghreb à la place de « faʿal» au Machrek par exemple :

« dreb » (frapper, algérien); « darab » (égyptien)

« sket » (se taire); « sakat »

« bʿed » (s'éloigner); « baʿad »

D'un autre côté, l'arabe dialectal diffère du standard par le fait que l'arabe standard permet d'avoir deux ou plusieurs voyelles longues en un mot phonologique, contrairement à l'arabe dialectal où elle ne permet d'en avoir qu'une seule. Par exemple, le mot 'clés' /*mafaatiih* > mafatiih/.

D'autres phénomènes caractérisent aussi l'évolution du système phonologique des voyelles dans le dialecte, il s'agit de la dénasalisation et la nasalisation des voyelles. Ces phénomènes sont en rapport avec les mots empruntés qui tiennent une place importante dans le vocabulaire des dialectes surtout au Maghreb. *En général, ces mots ont été facilement et naturellement incorporés dans leurs structures lexicales qui, elles, sont restées arabes de façon prédominante* (Guella, 2011). Ces emprunts sont toujours étudiés afin de déceler leur intégration dans les différents niveaux, *phonétique, morphologique et syntaxique* dans l'ensemble des dialectes maghrébins. Les phénomènes dénasalisation et de nasalisation dus à cet emprunt sont contradictoires et ont les propriétés suivantes :

- ***La dénasalisation :*** intervient sur le processus d'intégration des mots empruntés où nous assistons à un changement du mode d'articulation qui change la voyelle nasale en une voyelle orale. Pour illustrer ce propos, prenons les exemples suivants :
  - كَمْيُونَة *kamyounah* 'camion'
  - أَنْقْلِيزْ *lɛngliz* 'Angleterre'
  - قِيرَّة *girra* 'guerre'

- ***La nasalisation (ghounna) :*** est le résultat du voisinage de la nasale 'n' avec certains phonèmes. Ce phénomène met en avant le trait nasal comme un élément distinctif par rapport au littéral où les voyelles nasales n'ont pas de statut de phonème (ref : les unités de traitement dans les atlas linguistiques). L'illustration suivante montre ce phénomène : le mot قُنْبُلَةٌ *qunbulatun* en arabe MSA qui signifie 'une bombe' se transforme en قُنْبْلَة *qÕnbla* dans le dialecte tunisien.

5.3. Les alternances phonologiques (variations ou dégradations phonologiques)

Entre consonnes contiguës ou voisines nous pouvons identifier plusieurs phénomènes d'alternance qui résultent de la phonétique combinatoire. En arabe standard ou dialectal, cinq alternances phonologiques sont constatées : l'assimilation, métathèse, l'emphase, épenthèse, élision, le raccourcissement. Ces alternances peuvent se produire dans les morphèmes voir aux morphèmes et les limites des mots. La compréhension de ces processus est essentielle à l'étude de la morphologie de cette langue. L'étude de ces alternances n'a pas été faite de la même manière car les grammairiens arabes se sont peu occupés de la «métathèse» et de la « dissimilation», contrairement à «l'assimilation partielle» ou « accommodation » où une plus grande attention leur été accordée et ils les ont rangés parmi les différents phénomènes dénommés بدل *badal* ou إبدال *ibdal*, قلب *qalb* ou إقلاب *iqlâb* 'permutation de consonne'. Ils ont mis l'accent aussi sur 'l'assimilation complète', dite إدغام *iddgâm*.

5.3.1. Assimilation (الإِدْغَامْ)

En linguistique, le terme assimilation désigne un phénomène par lequel deux phonèmes tendent à devenir identiques ou à acquérir des caractères communs : par exemple -dt- > -tt-. Il existe deux types d'assimilation, الإِدْغَامْ *iddigâm* 'une assimilation complète' et 'partielle'. L'assimilation partielle est aussi appelée إقلاب *'qlâb* 'accommodation'.

Dans la langue arabe, l'assimilation complète des consonnes juxtaposées est manifestée dans certains cas. C'est le cas de l'assimilation complète de la consonne latérale /l/ de l'article défini, qui devient identique à la consonne initiale du mot si elle est l'une des lettres dites الحروف الشمسيّة *al-Huruf šamsiyya* 'lettres solaire', contrairement aux lettres dites consonnes الحروف القمريّة *al-Huruf qamariyya* 'lettres lunaires' où cette assimilation n'est pas réalisée. Ce cas est très répandu dans l'arabe standard.

L'assimilation de la lettre /l/ de l'article défini peut être formalisée par une règle appelée «*l-assimilation*» et représentée comme suit :

$$l\ [+def] \rightarrow Ci\ /\ \text{---} \begin{pmatrix} Ci \\ \\ +sol \end{pmatrix} \quad \textbf{..... (l-assimilation)}$$

Cette règle indique que la lettre /l/ de l'article défini est assimilée à la consonne suivante si elle est solaire. Pour voir illustrer l'application de cette règle et mettre en exergue la différence entre l'assimilation et la non assimilation de la lettre /l/, voici quelques exemples :

- /al+šuruq(-u)/ → /?aš-šuruq(-u)/ et non pas */?al-šuruq(-u)/ ''.
- /al+baHr(-u)/ → /?al-baHr(-u)/ et non pas */?ab-bHr(-u)/ ''.

Il existe aussi un autre type d'assimilations appelé *assimilation par contact*. Notons que le « l » de l'article, s'assimile non seulement avec les consonnes solaires, mais aussi à d'autres consonne. Dans cette optique, les grammairiens traitent le cas de l'assimilation ou non du « l » de l'article à la lettre ج *jim* 'j'. Quand le ج est prononcé comme une chuintante sonore j (considérée aussi comme une lettre de frontière entre lettre solaire et lunaire) elle n'assimile pas le «l» de l'article en MSA, mais l'assimile dans la plupart des dialectes. Nous nous référons aux travaux de (Marçais et Jellouli, 1933), où l'auteur signale que dans les parler d'El-Hamma de Gabès; l'assimilation n'est obligatoire que si la lettre j est l'élément initial d'un complexe consonantique : par exemple *ej-jbal* 'la montagne'; par contre, quand le «j» est la lettre initiale d'un mot déterminé est suivi d'une voyelle, l'assimilation du «l» de l'article devient facultative, par exemple, on trouve *ej-jar* à côté de *el-jar*.

Dans la même optique, nous soutenons les propos (Cantineau, 1960), qui signalent que l'assimilation de l'article ne se fait pas au Maghreb (voir rarement); contrairement au Machrek où elle paraît être la règle comme pour le mot *eg-gabal* 'la montagne'. (Gadalla, 2000) confirme dans son livre qu'en Égypte, le processus d'assimilation de l'article « l » avec les lettres solaires se produit avec l'ajout d'autre lettres comme la lettre /g/ ou très rarement la lettre /k/. Toutefois, cette assimilation du /l/ reste facultative. Pour illustrer ces propos, voici quelques exemples :

- /il+gabal/ → /?ig-gabal ~ ?il-gabal/ 'la montagne'.

- /il+kursi/ → /?ik-kursi ~ ?il-kursi/ ‘chaise’.

Cependant, (Cantineau, 1960) note que dans les villes, les gens instruits évitent de faire l'assimilation et disent par exemple *el-ğaw* ‘l'atmosphère’, *el-ğumle* ‘la somme’. En dehors de l'article, l'assimilation de la lettre «ğîm» se produit quand un mot contient un ğîm et une des sifflantes s, z, ṣ ou la chuintante š. Des accommodations se produisent quand le ğîm vient au contact des consonnes en question : ainsi un ancien *bi-l-ğizâf* ‘en bloc’ a abouti à l'algérien *bezzâf* ‘beaucoup’; un ancien *yağzi* a abouti en maghrébin à *yedzi*, voire à *yezzi* ‘cela suffi, assez’; le nom de l'île *ğazîra* a abouti dans les parler algérien à *dzîra* d'où le nom d'Alger *ed-Dzâir* qui à l’origine *al-ğazâ'ir*.

Une particularité caractérisant les dialectes maghrébins se manifeste dans une réduction du couple de lettres ‘dz’ en ‘zz’. Nous la trouvons par exemple dans l’interjection *yezzi*, assez! (à distinguer de *yedzi* ‘il suffit’ qui n'est pas une interjection).

D’un autre côté, quand un ğim tombe après un ‘s’, il est remplacé par un ‘y’, par exemple le mot du dialecte algérien *msyd* ‘école coranique’ vient du mot *masğîd*. Il y a également les assimilations de la lettre ğîm au voisinage d'une chuintante, qui existe dans plusieurs dialectes arabes et que nous pouvons classer et présenter comme suit :

- Dans les parlers de nomades tunisiens le ğim (prononcé ž : j) est transformé en un ‘z’ lorsqu’un mot contient un ‘z’ ou un ‘s’, par exemple le mot جزار *ğazzâr* ‘boucher’ devient ززار *zazzâr*, le mot عجوز *3ağûz* ‘vieille femme’ devient زوز *3zuz*, le mot زوج *zawğ* ‘paire’ devient زوز *zûz*, et le mot جبس *ğibs* ‘plâtre’ devient زبس *zebs*. Ces assimilations, par perte de chuintement, ont gagné les parlers sédentaires de la Tunisie, et ont pénétré assez loin vers l'est de l’Algérie, jusqu'à la ville de Constantine et jusqu'aux environs de Skikda.

- Dans les parlers de nomades du Sahara algérien, le ğim (prononcé ž) passe à ‘z’ quand le mot contient un z, un s, ou un š.

- Le ğim reste intact et c'est au contraire la sifflante qui devient chuintante : جوج *žûž* ‘deux’. Des faits de ce genre ont pénétré dans les parlers du Sahara oranais et au Maroc.

Au Maghreb, l'assimilation de la lettre « k » se produit seulement dans l'expression fréquente okkul (nous tous).

Contrairement à l’assimilation, la dissimilation consiste en un changement phonétique qui a pour objectif d’accentuer ou de créer une différence entre deux sons voisins mais non contigus. Elle se produit quand un « l » se trouve dans un même mot ou dans un même membre de la phrase au voisinage des lettres « r », « n », « m » ou « l » :

- Au voisinage d'un « r », la dissimilation se fait en « n » ou en « m » devant une labiale. Pour illustrer ces propos, prenons les exemples suivants : dans certains parlers du Machrek, nous attestons la forme بنّور *bennûr* ‘cristal’ en face du mot بلّور *ballûr*; البارح *Al-bârih* ‘hier’ qui est transformé en امبارح *embâreh* dans la plupart des parlers syro-palestiniens
- Au voisinage d'un « n » ou « m », la dissimilation peut se faire soit en « n », soit en « r ». Les exemples les plus célèbres de ce cas sont les noms du patriarche إسماعيل

Ismâ3îl ‘Ismaël’, qui est appelé dans beaucoup de parler إسماعين *Ismâ3în*, et l'ange ميخائيل *Mîkhâ'îl* ‘Michel’ qui est appelé aussi ميخائين *Mîkhâ'în*.

- Au voisinage d'un autre «l», la dissimilation se fait également en « n » ou en « r », par exemple le mot سلسلة *silsila* ‘une chaîne’, est prononcé dans une grande partie de l'Algérie et au Maroc سنسلة *sensla*, et c’est le même cas pour le mot زلزلة *zalzala* ‘un tremblement de terre’, est devenu زنزلة *zenzla*.

Un autre cas assez fréquent de la dissimilation est celui de la transformation de la lettre « n » en « l » au voisinage d'un autre « n ». Par exemple beaucoup de parlers prononcent le فِنْجَالْ *fingâl* ‘tasse à café’ qui est en MSA dit فِنْجَانْ *fingân*.

L'assimilation du ن *nûn* ‘n’ à une consonne sonante est un autre cas d'assimilation complète dans les limites des mots qui se produit régulièrement dans MSA et en arabe dialectal. La règle régissant cette transformation peut être formulée comme suit :

$$n \rightarrow Ci \,/\, ---\begin{pmatrix} Ci \\ +cons \\ +son \end{pmatrix} \quad \text{..... (n-assimilation)}$$

Dans (Mitchell, 1990) nous trouvons un autre exemple d’assimilation en MSA concernant la préposition من *min* ‘de’ qui peut être entendue en MSA sous les formes suivantes :

- *mir* si le mot qui le suit commence par la lettre r, exemple *mir rahmàti llάah (de la miséricorde de Dieu)*
- *mil* si est suivi par un mot qui commence par l, exemple *mil lúndun (de Londre)*
- *mim* si est suivi par un mot qui commence par m, exemple *mimmaa -<min ma- (de ce qui)*

Ce cas aussi s’applique de la même manière pour les prépositions (bin – fils) et (3an – selon), et il est présent aussi dans l'arabe dialectal, et les quelques exemples suivants en témoignent :

- /min + ramzi/ → /mir ramzi/ ‘de Ramzi’.
- /min + libnaan/ → /mil libnaan / ‘de Lebanon’.
- /min + mouhamed/ → /mim mouhamed/ ‘de Mohammed’

Par ailleurs, le « n » s'assimile souvent à un r ou à un l qui le suit, au Maghreb comme au Machrek, par exemple, nous prenons les cas suivants : *mellôz* du *men lôz* ‘d'olivier’, *wen râh* qui devient *werrâh* (où s'en est-il allé? »; *Bel3abbâs* qui était *Ben 3abbâs*. Dans cette optique, (Cantineau, 1960) souligne que l'assimilation peut se produire à distance avec un « l » qui précède; c'est ainsi que le surnom du célèbre saint de Bagdad : *Adb el-Qâder el-Gilâni* est devenu en Algérie et au Maroc « *ej-jîlâli* ». Dans (Marçais, 1908), l’auteur avance, pour le Sahara oranais, des exemples d'assimilation de « n » à un « t » à travers plusieurs exemples comme : *bett* du mot *bent* ‘fille de’, les pronoms personnels isolés comme, la 3ème personne masculin singulier تا *tta* ‘toi’ venant du pronom نتا *nta*, au féminin تي *tti* ‘toi’ qui provient du pronom نتي *nti*. De plus, la lettre « n » s'accommode en « m » devant une labiale, en particulier la lettre « b » : nous avons la tendance d'entendre fréquemment dans les différents dialectes les mots suivants :

- جمب *gamb* ‘côté’ pour le nom جنب *ganb,*
- مم بعد *mem ba3d* ‘après’ pour من بعد *men ba3.*

Nous avons aussi d’autres cas d’assimilation qui transforment le duo de lettre ‘sf’ en ‘ss’

comme dans le mot *noṣṣ* ou *nuṣṣ* ‘moitié, demie’ venant du terme *nuṣf*. C’est aussi pareil pour le duo de lettres ‘ft’ en ‘tt’, très répandu en Algérie dans les verbes *šott* ‘j'ai vu, tu as vu’ et *šottô* ‘je l'ai vu, tu l'as vu’. Dans la même optique, en Algérie et au Maroc nous trouvons l'accommodation de ‘mt’ en ‘nt’ dans les particules d'appartenance *ntâ3* tiré du mot *mtâ3*. Il y a aussi l'assimilation dans les parler syro-palestiniens du ‘b-’ de l'inaccompli au ‘m-’ préfixe de la 1ère personne au pluriel, comme dans le mot *mnektob* ‘nous écrivons’ qui vient du mot *b-nekto.*

Parfois les dialectes possèdent leurs propre assimilation comme pour la modification de la glotte /?/ à un glide /y/ quand il est suivie de la voyelle brève /i/ et précédé par une voyelle. Ce changement peut être présenté par la règle suivante :

$$V?i \rightarrow Vyi$$

Voici quelques exemples présentant cette transformation :

MSA	Dialectes	Traduction
نَائِم (naa?im)	نَايِم (naayim)	endormi
صَائِم (Saa?im)	صَايِم (Saayim)	qui jeûne
Haa?im	Haayim	perdu

5.3.2. Métathèses

C’est un processus phonologique très important qui s'applique dans les deux variétés. Elle est un phénomène par lequel deux phonèmes échangent leur place à l'intérieur d'une racine d'un mot, on dit aussi qu'il y a interversion. En d’autres termes, c’est l’action qui fournit des réalisations relativement importantes dont des permutations des consonnes. Par exemple Le nom du شمس *šεms* ‘soleil’ a continué en arabe dialectal la suite de ses métamorphoses. C'est ainsi que dans les parlers maghrébins, nous avons ordinairement par métathèse de chuintement, la forme سمش *semš*. Mais nous trouvons quelquefois, notamment dans les parlers du Maroc, une assimilation à distance de ‘s’ en ‘sh’ : شمش *šamš*; de sorte que la forme sémitique se trouve restituée.

Dans les parlers de nomades du Sahara algérien, le ğim (prononcé ž) est transformé en un ‘z’ lorsque le mot contient une des lettres suivantes : ‘z’, ‘s’ ou ‘š’, en même temps le ‘z’ ou le ‘s’ sont transformées respectivement en ‘ž’ ou à ‘š’. Ces transformations représentent une métathèse de chuintement, par exemple, le mot عجوز *3ağuz* ‘vieille femme’ passe à عزوج *3zûğ*, le mot *ğazza* ‘couper (la laine, le poil)’ passe à *zağğ*, et le mot جاز *ğaz* ‘passer’ devient زاج *zâğ*

Dans certains parlers du Moyens-Orient; جوز *juuz* ou *guuz* ‘époux’ pour زوج *zawj*. Une autre variante de la métathèse très importante s'applique dans les deux variétés (standard et dialectale), cette variante est la métathèse des consonnes identiques, qui opère dans les racines géminées quand elles sont suivies par une voyelle. La règle de métathèse des consonnes identiques, a été proposée par (Brame, 1970) et formalisée comme suit :

$$C_kVC_kV \rightarrow VC_kC_kV$$

Cette règle montre que lorsque deux consonnes identiques sont séparées par une voyelle et suivie par une autre voyelle, nous métathèsons la première consonne et la première voyelle de sorte que les consonnes se rejoignent avant les voyelles. Les exemples suivants illustrent cette opération :

Mot	MSA	EA	Traduction
madad(-a)	madd(-a)	madd	s'étirer
šadad(-a)	šadd(-a)	šadd	Tirer
masas(-a)	mass(-a)	Mass	Toucher

5.3.3. Epenthèse

Dans la tradition de la grammaire arabe, le système phonologique ne tolère pas d'avoir trois consonnes successives sans voyelle entre au moins deux consonnes. Il est donc impossible de trouver un groupement de trois consonnes successives, que ça soit dans la variété standard ou dialectale. Par conséquent, et afin d'éviter ce regroupement de consonnes en trois ou plus, un mot se terminant par deux consonnes est automatiquement suivi par un autre mot ou un suffixe commençant par une consonne, et vice-versa, et dans ce cas la voyelle haute brève /i/ est insérée à la jonction, autrement dit à la fin du premier mot. Nous pouvons aussi dire que cette opération consiste à insérer un phonème entre des consonnes. La voyelle est alors appelée voyelle d'anaptyxe' et ne sonne pas aussi clairement que les autres voyelles. La règle d'épenthèse en langue arabe peut être représentée comme suit :

$C+CC \rightarrow C^{V}CC$ …. (1)
$CC+C \rightarrow CC^{V}C$ …. (2)

L'épenthèse survenant ici n'est pas vraiment une conséquence des séquences de consonnes, mais plutôt de la syllabation des consonnes et ceci avec une seule grande différence entre les deux variétés : en MSA, la voyelle épenthétique survient après la première consonne, tandis qu'en arabe dialectal, elle survient après la deuxième consonne. Par conséquent, il sera plus précis de représenter l'épenthèse par les deux règles suivantes (où le point représente la jonction de syllabe)

Ø → i / C → C.C … (Médio-épenthèse en MSA)

Ø → i / C.C → C … (Médio-épenthèse en EA)

Dans les exemples de la partie (1) nous illustrons l'application de ces règles dans le MSA, alors que la partie (2) est consacrée pour l'application des règles dans l'arabe égyptien :

1. **MSA:**
   a. /Darab-at + al-bint(-a)/ →/Darab-at^{i} l-bint(-a)/ (elle a frappé la fille)
   b. /3an + al-bint(-i)/ →/3a.ni l-bint(-i)/ (de la fille)
   c. /kam + as-saa3-at(-u)/ → /ka.m^{i} s-saa3-at(-u)/ (quelle heure est-il?)
2. **EA:**
   a. /ba3d+bukra/ → /ba3.di.bukra/ (après-demain)
   b. /?ult+lak/ → /?ul.t^{i}.lak/ (Je vous ai dit)

Par ailleurs, il existe des cas spéciaux où le /u/ ou /a/ est inséré. En premier lieu, la voyelle /u/ est insérée avant un pronom ou un suffixe pronominal se terminant par [-um] dans les deux variétés, MSA et dialectale, et après les fractions 1/3 jusqu'à 1/9 en cas de construction avec /miyya/ (cent) pour former les centaines de 300 jusqu'à 900 incluses en AE. Des exemples de l'insertion de /u/ sont donnés en (3) pour le MSA et en (4) pour le dialecte égyptien.

3. **SA:**
    a. /?antum + al-mu3allim-uun/ →/?antu.mu l.mu3allim-uun/ (vous êtes les enseignants)
    b. /wa 3alaykum + as-salaam/ →/wa 3alayku.m^{u} s.salaam/ (et que la paix soit sur vous)
4. **EA:**
    a. /dars + hum/ →/dar.su.hum/ 'leur leçon'.
    b. /šuft + kum/ → /šuf.tu.kum/ 'Je vous ai vu (m)'.
    c. /xums + miyya/ → /xum.su.miyya/ 'cinq cent'.

En second lieu, la voyelle /a/ est insérée en MSA après /min/ quand c'est suivi par l'article défini [al-], comme c'est le cas en (5); et avant le suffixe pronominal [-ha] en AE comme illustré dans (6).

5. **MSA:** /min + al-bayt(-i)/ →/mi.na l.bayt(-i)/ 'de la maison'.
6. **EA:** /ism + ha/ →/?is.ma.ha/ 'son prénom'.

On peut justifier les cas spéciaux d'insertion de /u/ et /a/ par le fait que les mots et les suffixes concernés possèdent une voyelle harmonique "fantôme" ou "latente" (Zoll, 1996), ce qui n'apparaît pas lorsque ces mots ou suffixes sont prononcés individuellement les uns des autres, mais elle apparaît quand le discours est continu. Cette voyelle harmonique sera un /u/ lorsqu'elle est précédée ou suivie par un /u/. Cette configuration permet d'assurer l'harmonie des voyelles, comme c'est le cas pour les suffixes contenant un /u/ dans les deux variétés, ou pour les 'centaines' contenant des fractions sous la forme [Fu3L] en AE. Sinon, cette voyelle serait un /a/, comme on le trouve après /min/ en MSA et avant [-ha] en AE.

La voyelle fantôme peut être aussi appliquée au /a/ à l'article défini [al-] dans le cas où ce dernier est isolé ou se trouve en première position mais disparaît lors d'un discours continu en MSA. Lorsqu'il y a un conflit entre deux voyelles fantômes ou entre une voyelle fantôme et une voyelle épenthétique normale /i/, c'est la première de ces deux voyelles qui s'impose comme c'est le cas dans les exemples précédents, cf : (1), (3) et (5).

Pour éviter un groupement consonantique au début d'un mot, les deux variétés font appel à la "prosthèse" ou (épenthèse au début du mot) de la haute voyelle brève /i/. Ceci a été soutenu par (Broselow, 1976) pour la variété dialectale. Sa règle pour cette épenthèse peut être formalisée comme suit :

Ø → i / # ––– CC … (Epenthèse au début du mot)

Cette règle s'applique dans les deux variétés pour des impératifs avec un groupement consonantique initial dans la racine, ainsi que pour les verbes dérivés avec des groupements initiaux, entre autres. Des exemples de cette règle sont donnés dans le tableau suivant :

MSA	**EA**	**Traduction**
(?i)šrab	(?i)šrab	bois!
(?i)l3ab	(?i)l3ab	joue!
(?i)nkasar(-a)	(?i)nkasar	Il est cassé
(?i)sta3mal(-a)	(?i)sta3mal	Il a utilisé

Ces exemples représentent aussi l'insertion de l'arrêt glottal /?/ pour la protection de la structure phonotactique des mots : aucun mot ne doit commencer par une voyelle.

Lorsque le groupement consonantique ne se trouve plus au début du mot, l'épenthèse n'est plus applicable. Ainsi, les verbes dérivés en MSA possédant un /i/ épenthétique perdent cette voyelle lorsque le groupement est précédé par une voyelle au sein de la phrase, comme on peut le voir dans les exemples suivant :

7. **MSA:**
   a. /al-fariiq-u + (i)nhazam/ → /?al-fariiq-u nhazam/ (l'équipe a été défaite)
   b. /al-walad-u + (i)štagal/ → /?al-walad-u štagal/ (le garçon a travaillé)

De la même manière, l'épenthèse ne s'applique pas en arabe dialectal lorsque le groupement consonantique n'est plus initial, par exemple lorsque ce dernier est précédé par la marque de négation [ma-], comme dans les exemples suivants :

8. **EA:**
   a. /ma + (i)nhazam + š/ → /ma-nhazam-š/ (il n'a pas été battu)
   b. /ma + (i)štaghal + š/ → /ma-štaghal-š/ (il ne fonctionne pas)

5.3.4. Elision

Une différence remarquable entre le MSA et l'AE est que les hautes voyelles brèves /i/ et /u/ sont élidés dans des syllabes ouvertes médianes au sein d'une phrase dans l'arabe dialectal seulement. Ce fait a été représenté par une règle de suppression de haute voyelle proposée par (Broselow, 1976). (Mitchell, 1956) a fait référence à deux contextes où se produit l'élision de /i/ et /u/ dans la variété dialectale :

(1) si un suffixe commençant par une voyelle est attaché à un mot dont la dernière syllabe est du type / CiC / ou / CuC / et l'avant-dernière syllabe est ouverte, autrement dit se terminant par une voyelle, alors le /i/ ou /u/ de la dernière syllabe est presque constamment élidé. Par exemple,
   - /kaatib/ 'un écrivain' → /katb-a/ (une écrivaine)
   - /yaaxud/ 'il prend' → /yaxd-u/ (il le prend)

(2) Si les voyelles /i/ et /u/ se trouvant dans une syllabe courte et atone alors elles sont élidées lorsque cette syllabe devient médiane en une phrase phonologique, et que le mot précédent ou le préfixe se termine par une voyelle. Par exemple en dialecte égyptien nous avons :
   - /huSaan/ 'un cheval' → /?abu hSaan/ (l'homme avec un cheval)
   - /kitaab/ 'un livre' → /da ktaab/ (Il s'agit d'un livre)

Une autre différence majeure entre le MSA et l'arabe dialectal est que l'arrêt glottal à la fin d'un mot en MSA disparait en arabe dialectal. Ceci peut être expliqué par la règle suivante :

? → Ø / ——#… (Suppression de laglotte finale /?/)

Des exemples de l'application de cette règle sont donnés ci-dessous :

MSA	EA	Traduction
samaa?(-un)	sama	ciel
šitaa?(-un)	šita	hiver
mala?(-a)	mala	remplir
waraa?(-a)	wara	derrière

L'arrêt glottal disparait en AE également lorsqu'il constitue une partie ou bien la totalité de la coda d'une syllabe. Ce qui mène aux changements suivants :

- a? → aa
- i? → ii
- u? → uu

Des changements similaires peuvent être expliqués par la règle suivante :

? → V_i/V_i ——]$_\sigma$ … **(Allongement compensatoire en arabe dialectal)**

Cette règle indique que l'arrêt glottal devient similaire à la voyelle précédente si les deux se rencontrent à la fin d'une syllabe. Le tableau ci-après donne des exemples d'application de cette règle :

MSA	EA	Traduction
fa?r(-un)	faar	une souris
θa?r(-un)	taar	Vengeance
bi?r(-un)	biir	un puit

Notons qu'il y une exception pour le nom /bu?r-at(-un) > bu?r-a/ (foyer) qui reste inchangé en arabe dialectal. Nous affirmons également que le changement en MSA du /q/ en /?/ dans certaines variétés dialectale vient après les règles affectant le /?/ de base. Ainsi, l'allongement compensatoire ne s'applique pas aux mots comme /faqr(-un) > fa?r/ (pauvreté). Il existe un seul cas où cette règle peut être applicable en MSA : c'est le cas des verbes de la forme IV ayant un arrêt glottal au début.

5.3.5. Raccourcissement

Dans les deux variétés, une voyelle longue ne reste pas longue dans une syllabe fermée. Une syllabe fermée est une syllabe qui se termine par une consonne. Cependant, dans l'arabe dialectal, les syllabes en fin de mots se terminant par une seule consonne représente une exception qu'on discutera ci-après. Ceci est la conséquence d'un modèle limité de syllabe dans des positions non finales. Les voyelles longues sont interdites dans les syllabes fermées. La règle qui prend ceci en considération peut être nommée *Raccourcissement de la syllabe fermée* est formulée comme suit :

(V)VV → (V)V/ − C] $_\sigma$ … **(raccourcissement de la syllabe fermée)**

Cette règle indique qu'une voyelle extra-longue devient longue et qu'une voyelle longue en devient une courte dans une syllabe fermée. La seule différence entre les deux variétés dans ce sens et que la consonne finale en arabe dialectal n'intervient pas dans le calcul du poids de la syllabe mais intervient en MSA. Ceci peut s'expliquer par une condition

extramétrique spéciale en arabe dialectal qui exclut la consonne finale du calcul du poids de la syllabe.

Des exemples de raccourcissement de la syllabe fermée pour le MSA sont donnés en (1) et ceux de l'arabe dialectal sont donnés dans (2).

(1) MSA:

a. /(?i)nsadad(-a) → (?i)nsaadd(-a)/ → (?i)nsadd(-a)/ (bloquer)

b. /(?i)htajaj(-a) → (?i)htaajj(-a) → (?i)htajj(-a)/ (a contesté)

(2) EA:

a. /?uul + li/ → /?ul-li/ (Dis-moi!)

b. /kitaab + hum/ → /kitab-hum/ (leur livre)

Par ailleurs, les voyelles longues subissent un raccourcissement en passant à l'arabe dialectal quand elles se trouvent dans des positions finales conformément à la règle suivante :

VV ・ V / − # … **(raccourcissement de la voyelle finale en arabe dialectal)**

Le tableau suivant montre l'application de cette règle sur quelques exemples :

MSA	**EA**	**Traduction**
Sallaa	Salla	il priait
katab-uu	katab-u	ils ont écrit
ya-mšii	yi-mši	il marche

Toutefois, la règle de raccourcissement ne s'applique pas avant un suffixe commençant par une consonne, car la voyelle n'est plus définitive. En d'autres termes, les voyelles longues en MSA sont conservées longues dans EA avant les suffixes commençant par une consonne, comme dans les cas suivants:

MSA	**EA**	**Traduction**
?iksirii-h	?iksirii-h	casse le
ramaa-haa	ramaa-ha	il la jeta
?ilguu-hum	?ilguu-hum	annulez les

Une autre différence entre les deux variétés est qu'en cas de voyelles longues en arabe dialectal, qui survient souvent à cause d'une suffixation morphologique, la première voyelle subit un raccourcissement. Ceci est dû au fait qu'en arabe dialectal, une seule voyelle longue par mot est permise alors qu'en MSA, on peut en trouver plusieurs. Ceci est aussi une conséquence du raccourcissement des voyelles atones, étant donné qu'en arabe dialectal, toutes les voyelles atones doivent être courtes. Ceci peut être justifié par la règle nommée *raccourcissement atonique en arabe dialectal*. Donnée comme suit :

VV→V … **(raccourcissement atonique en arabe dialectal)**

Prenons les exemples suivants:

MSA	EA	Traduction
miizaán(-un)	mizaán	Balance
Taabuúr(-un)	Tabuúr	une file d'attente
Tuufaán(-un)	Tufaán	inondation
baa3uú-h	ba3uú-h	ils l'ont vendu
Saaduú-haa	Saduú-ha	ils l'ont aidé
kitaab-áyni	kitab-eén	deux livres
xabbaaz-iína	xabbaz-iín	boulangers

Dans quelques noms suivant le modèle nominal [CayCVVC], la séquence standard /ay/ est remplacée par /i/ dans l'arabe dialectal, qui constitue l'équivalent bref de /ee/, et dans certains cas moins fréquents par /a/. Nous illustrons ces particularités dans le tableau suivant :

MSA	EA	Traduction
maydaán(-un)	midaán	un terrain
šayTaán(-un)	šiTaán	un diable
rayhaán(-un)	rihaán	Myrte
zaytuún(-un)	zatuún	Olives
laymuún(-un)	lamuún	Citron

Le changement de [Fay3aaL(-un)] en MSA en [Fi3aaL] en arabe dialectal peut être justifié par l'hypothèse que la règle du raccourcissement atonique s'applique après la diphtongue en arabe dialectal. Ainsi le passage de /maydaán(-un)/ en MSA à /midaán/ en arabe dialectal se produit comme suit : *maydaán(-un)* → *meedaán* → *midaán*.

Chapitre 6 Analyse morphologique verbale

Introduction

Ce chapitre est consacré à la morphologie des verbes d'une part en arabe standard (MSA) et d'autre part en arabe dialectal : égyptien, tunisien et algérien. Nous notons qu'un effort important a été consacré pour cette étude. Une présentation sur les classes des verbes : les verbes trilitères et les verbes quadrilatères sera présentée en 6.1. La section 6.2 sera consacrée à présenter l'aspect et le mode de flexion. Enfin, nous terminons notre étude par présenter la voix de la flexion utilisée en MSA et en arabe dialectal dans la section 6.3.

6.1. Verbe, Stems et Classes

La base d'un verbe en arabe, appelée aussi radical ou thème, peut être classée selon leur structure morphologique en deux catégories : *simple (المُجَرَّدُ)* et *dérivée (المَزيدُ)*. Les verbes simples, référencés aussi comme *verbes radicaux*, sont les moins complexes de la langue car ils sont constitués de deux morphèmes : *une racine et un système vocalique*. Généralement ces verbes n'ont que trois lettres constituant le « radical ». Les verbes dérivés sont formés à partir des verbes radicaux auxquels nous rajoutons une, deux ou trois affixes dérivationnels. Selon (Mahadin, 1982), les verbes simples et dérivés sont traditionnellement appelés aussi non-augmentés et augmentés, respectivement. D'autant plus il suppose que : "La racine verbale peut être renforcée par un ou deux morphèmes de dérivation, ou il peut être non-augmentée. Les grammairiens traditionnels arabes ont utilisé des modèles (*?Awzaan* : الأوزان) pour identifier les formes augmentées et les formes non-augmentées" . Par ailleurs, nous distinguons aussi pour les verbes arabes deux autres formes par rapport à l'aspect du verbe: i) *perfectif* (*accompli*) exprimant une action terminée et accomplie, et ii) *imperfectif* (*inaccompli*) pour désigner une action inachevée.

En outre, les verbes écrits selon les normes MSA[12], TA[13] et EA[14] sont catégorisés en deux principales classes en fonction du nombre de consonnes dans leurs racines: verbe *tri-radical* (*trisyllabique*) et verbe *quadri-radical* (*quadri-syllabique*). Ces deux classes sont à leur tour divisées en sous-classes selon le type de consonnes dans leurs racines comme on le verra dans la suite de cette section où nous définissons dans les paragraphes suivants les caractéristiques de ces classes ainsi que leurs variétés.

6.1.1. Verbes tri-radicaux

Il existe dix formes de verbes tri-radicaux en arabe connues dans la littérature linguistique par leur énumération en chiffres romains : I à X. La première est considérée comme la forme primaire et le reste des classes sont les formes obtenues des verbes dérivés. (Schmidt, 1975) indique que les processus morphologiques impliqués dans cette dérivation sont : « de la plus simple forme de la racine sous-jacente, la forme I, et le reste des formes peut être obtenu par gémination d'une consonne, allongement de voyelle, préfixation, et infixation".

En fonction du type de consonnes dans leur racine, les verbes tri-radicaux peuvent être aussi subdivisés en quatre classes (Mahadin, 1982) :

i. ***Les verbes sonores :*** ce sont les verbes constitués de consonnes autres que /?/, /w/ ou /y/; ils sont aussi appelés verbes *forts* ou *réguliers* (السالِمُ),

[12]MSA : Modern Standard Arabic
[13]TA : Tunisian Arabic
[14]EA : Egyptian Arabic

ii. ***Les verbes géminés :*** dans ce type de verbe la deuxième et la troisième consonne de la racine sont les mêmes (la répétition de la même) ; ils sont également nommés verbes *doublés* ou *sourds* (المُضَعَّفُ),

iii. ***Les verbes glottalisés :*** qui ont le coup de glotte /?/ comme un radical; ils sont aussi appelés verbes *hamzé* (المَهْموزُ) (une consonne est un hamza, en 1°, 2° ou 3° position) ,

iv. ***Les verbes faibles :*** qui ont parmi leurs radicaux les semi-voyelles /w/ ou /y/ ou les deux. Il existe trois type en fonction de la position de la semi-voyelle : nous avons le المِثالُ (mithAl - *assimilés*) si cela concerne la première lettre, الأَجْوَفُ (al-Ajwaf - *concaves)* si cela concerne la deuxième et الناقِصُ (al-NaqiS - *défectueux)* si c'est la troisième.

6.1.1.1. Verbes sonores

Les verbes sonores sont ceux qui contiennent des consonnes arabes autres que /? /, /w/ ou / y /. Le Tableau (6.1) récapitule les dix formes canoniques des verbes sonores dans le MSA, tandis que le tableau (6.2) affiche leurs équivalents dans l'EA.

Classe	Forme	Exemple	Traduction
I	Fa3$\begin{pmatrix} a \\ u \\ i \end{pmatrix}$ L(-a)	*Darab(-a)*	frapper
		kabur(-a)	grandir
		hazin(-a)	devenir sévère
II	Fa33aL(-a)	faDDal(-a)	préférer
III	Faa3aL(-a)	haarab(-a)	guerroyer
IV	?aF3aL(-a)	?ahraj(-a)	embarrasser
V	taFa33aL(-a)	taharrak(-a)	bouger
VI	taFaa3aL(-a)	tanaaqaš(-a)	discuter
VII	(?i)nFa3aL(-a)	(?i)nhazam(-a)	perdre
VIII	(?i)Fta3aL(-a)	(?i)jtahad(-a)	cravacher
IX	(?i)F3aLL(-a)	(?i)hmarr(-a)	rougir
X	(?i)staF3aL(-a)	(?i)stagfar(-a)	demander pardon à Dieu

Tableau 6. 1. Les formes canoniques des verbes sonores dans l'arabe MSA

Classe	Forme	Exemple	Traduction
I	F$\begin{pmatrix} a \\ i \end{pmatrix}$ 3$\begin{pmatrix} a \\ i \end{pmatrix}$ L	Darab	frapper
		Kibir	grandir
II	Fa33$\begin{pmatrix} a \\ i \end{pmatrix}$L	faDDal	préférer
		?addim	présenter
IIIa	Faa3Il	haarib	guerroyer
IIIb	Foo3Al	soogar	enfermer bien
IV	?aF3Al	?ahrag	embarrasser
V	(?i)tFa33$\begin{pmatrix} a \\ i \end{pmatrix}$L	(?i)tharrak	bouger
		(?i)tgaddid	rénover
VI	(?i)tFaa3iL	(?i)tnaa?iš	discuter
VII	(?i)nFa3aL	(?i)nhazam	perdre
VIII	(?i)Fta3aL	(?i)gtahad	cravacher
IX	(?i)F3aLL	(?i)hmarr	rougir
X	(?i)staF3$\begin{pmatrix} a \\ i \end{pmatrix}$L	(?i)stagfar	demander pardon à Dieu
		(?i)sta3gil	être pressé

Tableau 6. 2. Les formes canoniques des verbes sonores dans l'arabe EA

Classe	Forme	Exemple	Traduction
I	F3 $\begin{pmatrix} a \\ u \\ i \end{pmatrix}$L	Drab	frapper
		GhloT	commettre une erreur
		Kbir	grandir
II	Fa33$\begin{pmatrix} a \\ i \end{pmatrix}$L	kassar	casser
		khammim	penser
III	Faa3$\begin{pmatrix} a \\ i \end{pmatrix}$L	waafak	accepter
		Saafir	voyager
IV	Ø	Ø	Ø
V	tFa33$\begin{pmatrix} a \\ i \end{pmatrix}$L	t3aTTal	retarder
		Tzayyin	
VI	tFaa3$\begin{pmatrix} a \\ i \end{pmatrix}$L	tnaaTaH	
		t3aarik	faire une bagarre
VII	Ø	Ø	Ø
VIII	(?i)Fta3$\begin{pmatrix} a \\ i \end{pmatrix}$L	(?i)rtaa7 ‘’ (?i)إِحْتَقَرَ	se reposer
		(?i)ntakhib	voter
IX	(?i)F3aaL	(?i)hmaar	rougir
X	(?i)staF3$\begin{pmatrix} a \\ i \end{pmatrix}$L	(?i)stagfar	demander la clémence
		(?i)sta3mil	utiliser

Tableau 6. 3. Les formes canoniques des verbes sonores dans l’arabe tunisien (AT)

La comparaison des tableaux (6.1), (6.2) et (6.3) montre que la majorité des formes des verbes sonores sont retenues dans le dialecte égyptien et que deux de ces formes, en occurrence la IV et la VII, sont supprimées dans le dialecte tunisien. Il existe aussi d’autres modifications à faire comme c’est illustré ci-dessous:

i. Les deux variétés sont similaires en utilisant les mêmes procédés morphologiques pour dériver les formes II-X de la simple forme I (fa3al) :

✓ ***La forme II*** (*fa33al*) : est dérivée par la gémination de la deuxième consonne de la racine,
✓ ***La forme III*** (*faa3al*) : est obtenue par l'allongement de la première voyelle,
✓ ***La forme IV*** (*?a-f3al*) : est constituée en faisant précéder la séquence [?a-] en position préfixale pour le dialecte égyptien. Le tableau (6.3) montre que cette forme n'existe pas en dialecte tunisien (et par extension dans les différents dialectes du Maghreb),
✓ ***La forme V*** (*ta-fa33al*) : caractérisé par le préfixe [ta-] avec une gémination de la deuxième consonne de la racine. En d'autres termes, la forme V présente le résultat de la concaténation du suffixe « t- » à la forme II. Le tableau (6.3) montre que le dialecte tunisien partage cette caractéristique en rajoutant le préfixe [t-] au lieu du [ta-] pour constituer la forme (t-fa33al).
✓ ***La forme VI*** (*ta-faa3al*) : c’est la forme utilisant le préfixe [ta-] et l'allongement de la première voyelle. La forme VI est le résultat de la concaténation du suffixe « t- » à la forme III. Nous précisons que le dialecte tunisien modifie légèrement cette forme en utilisant le préfixe [t-] au lieu du [ta-] donnant ainsi la forme (t-faa3al),
✓ ***La forme VII*** (*in-fa3al*) : nous obtenons cette forme en faisant précéder la racine par le morphème [n-]. Pour éviter la succession de deux consonnes au début du mot (chose que l’arabe classique ne tolère pas), une séquence prothétique « ?i » est insérée donnant lieu à la forme suivante : « ?infa3al ». Le tableau (6.3) montre que

cette forme est inconnue en dialecte tunisien, mais elle est très fréquente dans l'ouest de l'Algérie (Tlemcen, Oran...) connue sous la forme «?inf3al ».

✓ ***La forme VIII*** (*?ifta3al*) : elle est conçue par l'association du morphème [-t-], entre la première et la deuxième consonne de la racine, sans être dans une position préfixale. Un segment prothétique « ?i » est ajouté pour empêcher la succession de deux consonnes en début de mot, comme l'exige l'arabe classique : « (?i)fta3al ».

✓ ***La forme IX*** (*if3all*) : elle est formée par gémination de la troisième consonne de la racine et l'insertion d'une séquence prothétique « ?i » afin d'éviter la succession de deux consonnes au début de mot, ce qui donne la forme : « ?if3all ». Le tableau (6.3) met en avant une autre forme différente utilisée dans le dialecte égyptien et partagée dans différents dialectes maghrébins. Il s'agit de la forme (if3aal) qui est constituée par l'allongement de la deuxième consonne de la racine, et l'éventuel ajout d'une séquence prothétique « ?i » pour éviter la succession de deux consonne au début de mot.

Les cas où des variantes de ces formes sans la séquence prothétique « ?i » se présentent dans le dialecte maghrébin. Ceci est dû au fait que ce dialecte permet la succession de deux consonne au début du mot.

✓ ***La forme X*** *(sta-f3al)* : cette dernière forme est obtenue par l'ajout du préfixe [sta-] à la forme I. Elle a la particularité de réaliser un préfixe bi-consonantique : « st- ». Dans cette forme aussi, un segment prothétique « ?i » est ajouté afin d'éviter la succession de deux consonnes au début de mot : « ?istaf3al ». Cette forme existe dans les différentes variétés dialectales. Dans le dialecte algérien nous trouvons une autre forme supplémentaire différente de celle des autres dialectes, employée dans les parlers de l'ouest algérien qui utilisent la forme (ssa-f3al). A ce sujet (Marçais, 1902) indique la réduction de la séquence [st] classique à [ss]; que nous entendons fréquemment en un seul /s/.

ii. Il n'y a pas de changements dans les consonnes entre les formes MSA, EA et TA, mais il y a des différences de voyelles. Par exemple, la voyelle finale (-a) dans les modèles MSA ne figure pas dans EA (et TA) parce que ce dernier a en général perdu ses voyelles finales si elles sont utilisées à des fins flexionnels en MSA. La voyelle (-a) est, dans ce cas, utilisée comme un objet marqueur de la troisième personne du masculin singulier. C'est pourquoi il est disparu dans la variété dialectale.

iii. Dans la forme I, la première voyelle est toujours /a/ dans MSA. D'autre part, EA conserve cette voyelle dans certains verbes (par exemple / fataH(-a)> fataH / «ouvrir») et il se transforme en / i / dans d'autres verbe (par exemple / najaH (-a)> nigiH / «réussir»). Les dialectes Maghrébins ne conservent pas la première voyelle dans la forme I, elle est remplacée par Soukoun, par exemple pour les verbes fataH(-a) et najaH (-a) nous obtenons ftaH et njaH respectivement. Ainsi, il existe deux versions de cette forme dans l'arabe dialectal égyptien : [Fa3aL] et [Fi3iL]. Par ailleurs, EA n'a pas le /u/ comme deuxième voyelle. Il remplace le /u/ de MSA par un /i/. En d'autres termes, la racine perfective MSA [Fa3uL (-a)] est remplacée par la forme [Fi3iL] dans EA, par exemple, /kabur(-a) > kibir/ «grandir» . Il est à noter que l'EA tend à harmoniser la première voyelle avec la seconde voyelle. Contrairement au dialecte égyptien, le dialecte tunisien conserve le /u/ comme une deuxième voyelle (par exemple, /Dahar(-a)> Dhur / «apparaître», D3uf (devenir faible), GhluT (commettre une erreur), en plus du /a/ dans certains verbes (par exemple, / kasar(-a)> ksar / «briser») et le /i/ dans d'autres verbes (par exemple, / khasar(-a)> khsir / «perdre»). En résumé, il existe trois versions de la forme I dans l'arabe dialectal tunisien : [F3aL], [F3uL] et [F3iL]. Pour le

dialecte algérien, la voyelle dominante au perfectif concernant la deuxième consonne est le [ə] transcrit par la fatha (a). Le son /i/, habituel chez les parlers tunisiens et à l'Est et sud de l'Algérie n'est pas connu dans les autres villes de l'Algérie. D'un autre côté, le son [a] se rencontre régulièrement dans les verbes qui ont une 3ème radicale 3in, Ha' ou Hah comme dans le mot tla3 (il est monté) ou dans le mot fra7 (il s'est réjoui). Ce son est aussi fréquent dans les mots qui ont pour 2ème radicale une de ces lettres : 3in, Ha' ou Hah. C'est le cas des mots l3ab (il a joué), Dhak (il a ri) et Dhar (il a paru).

iv. En MSA, chacune des formes II, V et X possède deux contreparties différentes en EA, selon l'état de la voyelle pré-finale ou celle du stem. Bien que le MSA ne dispose que de la voyelle/ a / comme voyelle pré-finale dans ces formes, dans les deux variétés dialectales cette voyelle est un /a/ dans certains verbes et un /i/ dans d'autres. Nous signalons que le dialecte algérien ne possède qu'une seule forme comme le MSA, à l'exception des parlers du l'est et sud de l'Algérie qui suivent l'arabe dialectal tunisien.

v. La Forme III a subi un changement intéressant dans le dialecte égyptien comme le soutient (Carter, 1996) « en plus des mots familiers kaatib / yikaatib / mikatba, il existe aujourd'hui une nouvelle racine avec une longue première syllabe -oo-". La voyelle /oo/ ne peut pas être liée à / aw / par monophtongaison car il n'est pas un stem final. D'autres exemples de /aw/ dans la position de non stem final qui ne devient pas /oo/ sont : /mawrid / «une source» et /kawkab/ «une planète». Ce fait suggère que la voyelle /oo/ est sous-jacente dans les verbes de cette nouvelle forme qui ne possèdent pas de contreparties étymologiques dans MSA. Contrairement au MSA qui ne dispose que de la voyelle /a/ comme voyelle pré-finale dans ces formes, la forme III possède deux variantes différentes dans le dialecte tunisien dépendantes de l'état de la voyelle pré-finale ou de celle du stem. De ce fait, cette voyelle pré-finale est un /a/ dans certains verbes comme dans le verbe / waafaq(-a)> waafaq / «accepter» et un /i/ dans d'autres comme dans le verbe / Taalab(-a)> Taalib / «demander».

vi. Les formes dialectales [(?i)tFa33aL ~ (?i)tFa33iL] et [(?i)tFaa3aL ~ (?i)tFaa3iL] sont considérées comme des reflets des formes standard [taFa33aL (-a)] et [taFaa3aL(-a)], respectivement. Autrement dit, le préfixe dérivationnel MSA [ta-] est ré-analysé comme [t-] dans EA et TA, déclenchant l'ajout de (?i) par les règles de l'épenthèse .

En plus des différences dans la forme indiquée ci-dessus, il en existe aussi quelques-unes dans la fréquence d'utilisation. Par exemple, la forme IV se produit souvent dans EA ce qui n'est pas le cas pour le MSA où elle est moins utilisée, voire inexistante comme dans le dialecte tunisien. Cette forme IV est généralement remplacée par la forme II : par exemple /?adxal(-a)/ «apporter» est remplacé par /daxxal/ ; et parfois par la forme I, par exemple : /?ab3ad(-a)/ «à emporter» est remplacé par /ba3ad/. Deux raisons ont été avancées par (Malik, 1976) pour expliquer ce remplacement de la forme verbale IV dans la variété dialectale :

a. Les significations sémantiques des formes verbales II et IV se chevauchent souvent même en MSA
b. Les personnes semblent rencontrer des difficultés à prononcer le coup de glotte /?/ dont la prononciation implique une certaine tension dans le larynx.

Selon (Gadalla, 2000), bien que la première raison semble plausible, la seconde est discutable parce que les égyptiens ont tendance à changer la consonne /q/ du MSA en /?/ et le prononcent sans aucune difficulté. Alors, la vraie question ici est la tendance à perdre en EA le /?/ sous-jacent.

Nous renforçons notre précédente étude au sujet de la forme IV par les propos de Marçais dans son livre (Marçais, 1902) où il dit : *la forme IV ne s'est à proprement parler pas plus conservée en tlemecenien que dans les autres dialectes maghrébins. Caractérisée en arabe classique par la présence au parfait d'un Alif initial et au futur par de délicates nuances de vocalisation, elle a disparu dans un idiome qui comporte généralement l'aphérèse de l'Alif initial, et assourdit volontiers les voyelles colorées de la langue régulière. Généralement, elle a été remplacée par la II^{e} forme, voisine comme sens et nettement différenciée de la I^{ère} par le redoublement de la 2ème radicale; parfois aussi le verbe à la IV^{e} forme est simplement ramené à la I^{ère}.*

D'un point de vue sémantique, les formes MSA et leurs équivalentes en EA sont synonymes. Ceci peut être observé en examinant le sens des différentes formes morphologiques (Travis, 1979 et McGuirk, 1986) :

- Ia. [Fa3aL(-a) > Fa3aL ~ Fi3iL] : non transitive ou intransitive stative.
- Ib. [Fa3iL(-a) > Fi3iL]: état temporaire intransitif .
- Ic. [Fa3uL(-a) > Fi3iL]: état permanent intransitif.
- [Fa33aL(-a) > Fa33aL ~ Fa33iL]: causal, intensive, réitératif ou estimative.
- [Faa3aL(-a) > Faa3iL]: l'action fait ou d'une personne. Il désigne l'effort de faire quelque chose ou de quelqu'un. Il signifie aussi la réciprocité à l'égard de l'action indiquée par la forme I (Abdel-Malek, 1972).
- [?aF3aL(-a) > ?aF3aL]: causal ou factitif.
- [taFa33aL(-a) > (?i)tFa33aL ~ (?i)tFa33iL]: intransitif ou passive de la forme II.
- [taFaa3aL(-a) > (?i)tFaa3iL]: réciproque de la forme III.
- [(?i)nFa3aL(-a) > (?i)nFa3aL]: passive ou inchoatif de la forme I.
- [(?i)Fta3aL(-a) > (?i)Fta3aL]: réflexive de la forme I. Elle désigne le changement ou le développement. Elle est également réflexive de la forme IV (Moore, 1979).
- [(?i)F3aLL(-a) > (?i)F3aLL]: le développement d'une couleur ou d'un défaut.
- [(?i)staF3aL(-a) > (?i)staF3aL ~ (?i)staF3iL]: demander ou faire quelque chose pour soi-même. Il est également responsable de la Forme I. Il a une estimative ou préfixe désideratif qui signifie souvent «à envisager, à considérer (quelqu'un ou quelque chose) que" (Abdel-Malek, 1972).

Le tableau (6.4) présente les formes imperfectives des verbes sonores en MSA, le tableau (6.5) introduit leurs analogues en EA, tandis que le tableau (6.6) donne leurs correspondants en TA.

Classe	Forme	Exemple	Traduction
I	ya-F3$\begin{pmatrix} a \\ u \\ i \end{pmatrix}$L(-u)	ya-Drib(-u)	frapper
		ya-kbur(-u)	grandir
		ya-hzan(-u)	devenir triste
II	yu-Fa33iL(-u)	yu-kassir(-u)	casser
III	yu-Faa3iL(-u)	yu-haarib(-u)	guerroyer
IV	ya-taFa33aL(-u)	yu-hrij(-u)	embarrasser
V	ya-taFaa3aL(-u)	ya-taharrak(-u)	bouger
VI	ya-nFa3iL(-u)	ya-tanaaqaš(-u)	discuter
VII	ya-Fta3iL(-u)	ya-nhazim(-u)	être battu
VIII	ya-F3aLL(-u)	ya-gtahid(-u)	cravacher
IX	ya-staF3iL(-u)	ya-hmarr(-u)	rougir
X	yu-Fa33iL(-u)	ya-stagfir(-u)	demander la clémence

Tableau 6. 4. Les formes imperfectives des verbes sonores en MSA

Classe	Forme	Exemple	Traduction
I	yi-F3$\begin{pmatrix} i \\ u \\ a \end{pmatrix}$ L	yi-ktib	écrire
		yi-skut	se taire
		yi-Drab	frapper
II	yi-Fa33$\begin{pmatrix} a \\ i \end{pmatrix}$ L	yi-kassar	casser
		yi-?addim	présenter
IIIa	yi-Faa3iL	yi-haarib	guerroyer
IIIb	yu-Foo3aL	yi-soogar	enfermer bien
IV	yi-F3iL	yi-hrig	embarasser
V	yi-tFa33$\begin{pmatrix} a \\ i \end{pmatrix}$ L	yi-tharrak	bouger
		yi-tgaddid	renouveler
VI	yi-tFaa3iL	yi-tnaa?iš	discuter
VII	yi-nFi3iL	yi-nhizim	être battu
VIII	yi-Fti3iL	yi-gtihid	cravacher
IX	yi-F3aLL	yi-hmarr	rougir
X	yi-staF3$\begin{pmatrix} a \\ i \end{pmatrix}$ L	yi-stagfar(-u)	demander la clémence
		yi-sta3gil	être pressé

Tableau 6. 5. Les formes imperfectives des verbes sonores en EA

Classe	Forme	Exemple	Traduction
Ia	yi-F3$\begin{pmatrix} a \\ i \end{pmatrix}$ L	yi-rba7	gagner
		yi-skit	se taire
Ib	ya-F3aL	ya-3raf	savoir
Ic	yu-F3uL	yu-Khruj	sortir
II	y-Fa33$\begin{pmatrix} a \\ i \end{pmatrix}$ L	y-kassar	casser
		y-kammil	présenter
III	y-Faa3$\begin{pmatrix} a \\ i \end{pmatrix}$ L	y-waafaq	accepter
		y-chaarik	participer
IV	Ø	Ø	Ø
V	yi-tFa33$\begin{pmatrix} a \\ i \end{pmatrix}$ L	yi-t3aTTal	retarder
		yi-tzayyin	
VI	yi-tFaa3$\begin{pmatrix} a \\ i \end{pmatrix}$ L	yi-tnaaTaH ‘’ يِتْنَاطَحْ	
		yi-t3aarik	faire une bagarre
VII	Ø	Ø	Ø
VIII	yi-Fta3$\begin{pmatrix} a \\ i \end{pmatrix}$ L	yi-rtaa7 ‘’yi-7taram****	se reposer
		yi-ntakhib	voter
IX	yi-F3aaL	yi-hmaar	rougir
X	yi-staF3aiL	yi-stagfar	demander la clémence
		yi-sta3mil	utiliser

Tableau 6. 6. Les formes imperfectives des verbes sonores en TA

En comparant les tableaux (6 .4), (6.5) et (6.6), nous remarquons que la voyelle /a/ en MSA dans le préfixe imperfectif est changée en /i/ dans EA. Nous signalons que la forme [yi-

F3uL] peut être aussi prononcée dans certaines régions par la forme [yu-F3uL], par exemple le verbe Tabakha – yu-Tbukh (il a cuisiné). Quant au dialecte tunisien, nous identifions les cas suivants :

- Conservation de la voyelle /a/ dans le préfixe imperfectif
- Changement de la voyelle /a/ en /u/ dû à une harmonisation de la voyelle du préfixe avec celle de la 2ème radicale
- Changement de la voyelle /a/ en /i/
- Changement de la voyelle /a/ en /soukoun/

Les tableaux montrent également que dans les trois variétés étudiées (MSA, EA, TA), le changement de la voyelle du stem (c'est à dire que, avant la dernière radicale) de la Forme I entre le perfectif et l'imperfectif n'est pas systématique. Par conséquent, il doit être appris à partir d'un dictionnaire fiable. Cependant, quelques généralisations peuvent être proposées :

a. Les verbes en MSA de la forme [Fa3aL(-a)] ont deux voyelles (ou stems) imprévisibles, /u/ ou /i/, à savoir [ya-F3uL(-u) ~ ya-F3iL(-u)], par exemple : /sakat(-a), ya-skut(-u) / «se taire», ce qui n'est pas le cas pour le verbe /hamal(-a), ya-hmil(-u)/ « porter ». Il existe exceptionnellement des verbes dont le deuxième ou le troisième radical est une gutturale /3, ħ,?, h, x ou g/ qui ont tendance à avoir /-a-/ dans la forme imperfective. Par exemple prenons le verbe /fataħ(-a), ya-ftaħ(-u)/ « ouvrir ». Cependant cette exception ne s'applique pas pour le verbe /daxal (-a), yadxul (-u)/ « entrer » bien que la deuxième radicale est une gutturale. De même, les verbes en EA de la forme perfective [Fa3aL] ont trois formes imperfectives : [yi-F3aL ~ yi-F3iL ~ yi-F3uL(yu-F3uL)], par exemple /Darab, yi-Drab/ « frapper », /daras, yi-dris/ « apprendre » et /daxal, yi-dxul ou encore yu-dxul/ « entrer ». Les verbes en TA de la forme perfective [F3aL] possèdent quatre formes imperfectives : [ya-F3aL ~ yi-F3aL ~ yi-F3iL ~ yu-F3uL]. Prenons par exemple les verbes : /KhlaT, ya-xlaT/ « arriver au temps », /rba7, yi- rba7/ « gagner », /Tlab, yi-Tlib/ « demander » et /dxal, yu-dxul/ « entrer ». Le dialecte tunisien se caractérise aussi par le fait que la forme [F3al ~ yu-F3ul] peut être aussi remplacée dans certaines régions par la forme [F3al ~ yi-F3il], par exemple, le verbe /daxal/ « entrer », a deux formes à l'imperfectif /yu-dxul/ ou /yi-dxil/ (c'est le cas des parlers de Sfaxe). En Algérie, pour cette forme nous trouvons en plus de la forme [F3al ~ yu-F3ul], la forme [F3al ~ ya-F3al]. Il est à noter que les verbes ayant déjà un futur avec /u/ en arabe classique le gardent aussi en arabe dialectale. Dans certains régions de l'Algérie, trois formes imperfectives du dialecte tunisien sont reprises, il s'agit des formes [ya-F3aL ~ yi-F3aL ~ yu-F3uL]. A ces formes, ils rajoutent la forme [ya-F3ul] pour les parler de Tlemcen. Prenons quelques exemples : ya-skut, « il se taira », ya-tlob « il demandera », ya-skun, « il habitera », etc. Nous précisons que la forme [yi-F3aL] est utilisée seulement dans les parlers de l'est et le sud algérien.

Nous appuyons cette étude par quelques arguments avancés par (Marçais, 1902) : *le voisinage consonantique a une grosse influence sur la couleur et la nuance de la voyelle du futur. Le choix de cette voyelle n'est point au reste soumis à des règles aussi fixes que dans d'autres dialectes, en omani par exemple; et les variations sont nombreuses dans les prononciations individuelles.*

b. (b) les verbes en MSA de la forme [Fa3uL (-a)] à l'accompli (perfectif) ont la forme [ya-F3uL (-u)] à l'imperfectif, par exemple, /Kabur(-a), ya-kbur (-u)/ «grandir». Cependant, les verbes en MSA de la forme [Fa3iL (-a)] au perfectif ont la forme [ya-F3aL (-u)] à l'imperfectif, par exemple, /Hazin (-a), ya-hzan(-u)/ «devenir triste ».

D'autre part, les verbes en EA au perfectif de la forme [Fi3iL], qui est équivalente au deux formes [Fa3uL (-a)] et [Fa3iL (-a)] en MSA, prennent la forme [yi-F3aL] à l'imperfectif, par exemple, /Kibir, yi-kbar/ «grandir» et /hizin, yi-hzan/ «devenir triste ». Dans le dialecte tunisien, les verbes au perfectif de la forme [F3il] prennent la forme [yi-ktib] à l'imperfectif en gardant le /i/ de la deuxième consonne, par exemple, /qlib, yi-qlib/ «renverser», /skin, yi-skin/ « habiter ». Cependant les verbes de la forme [F3uL] au perfectif prennent la forme [yu-F3ul] à l'imperfectif comme pour le verbe /GhloT, yo-GhloT/ « commettre une erreur ».

Revenons maintenant aux tableaux (6.4) et (6.5), nous remarquons qu'en dehors de la voyelle du préfixe de l'imperfectif, le MSA et l'arabe dialectal suppriment à l'imperfectif de la forme I la première voyelle du radical au perfectif, c'est à dire celle qui suit la première consonne de la racine, y compris les préfixes de dérivation. Cela peut s'expliquer par la règle de « l'élision de la voyelle » proposée par (Brame, 1970) :

$$V \rightarrow \emptyset \,/\, V + C __ CV \ldots$$ **élision de voyelle**

Cette règle indique que nous supprimons la voyelle qui vient après la première consonne de la racine à l'imperfectif de la forme I. Le tableau (6.6) montre que cette caractéristique est partagée aussi par le dialecte tunisien et cela en raison de la suppression à l'origine de cette voyelle dans les dialectes maghrébins à la forme perfective.

Toutefois, cette voyelle a des comportements différents dans les différentes variétés dialectales dans d'autres formes. Tout d'abord, elle est conservée, c'est à dire elle reste identique à celle de MSA, sous trois formes: II, III et X, comme explicité dans les exemples suivants :

Forme	**MSA**	**EA**	**Traduction**
II	yu-qaddim(-u)	yi-?addim	présenter
III	yu-saafir(-u)	yi-saafir	voyager
X	ya-staslim(-u)	yi-staslim	se soumettre

Deuxièmement, elle est omise dans les deux formes: V et VI, comme une conséquence directe de la ré-analyse du préfixe [ta-] dans MSA tout comme celui du [t-] dans EA et TA. Le tableau suivant illustre quelques exemples :

Forme	**MSA**	**EA**	**Traduction**
V	ya-ta3allam(-u) ya-takallam(-u)	yi-t3allim yi-tkallim	étudier parler
VI	ya-tafaaham(-u) ya-tamaaraD(-u)	yi-tfaahim yi-tmaariD	comprendre simuler la maladie

Enfin, elle est remplacée par /i/ dans les deux formes VII et VIII dans le but de sécuriser l'harmonie de la voyelle avec la pré-finale /i/ dans le dialecte égyptien. Cette harmonie s'applique chaque fois que le /a/ est dans une syllabe lumière, comme dans les exemples suivants :

Forme	MSA	EA	Traduction
V	ya-ta3allam(-u) ya-takallam(-u)	yi-t3allim yi-tkallim	étudier parler
VI	ya-tafaaham(-u) ya-tamaaraD(-u)	yi-tfaahim yi-tmaariD	comprendre malinger

Le tableau (6.6) illustre bien que le dialecte tunisien ne possède pas la forme VII, contrairement au dialecte égyptien et au dialecte algérien (elle est très fréquente dans l'ouest algérien comme pour les parlers d'Oran ou de Tlemcen). Mettre un verbe transitif à la VII^{ème} forme est la façon habituelle d'en former le passif, comme pour le verbe /ya-nksar/ « a été cassé » dans le dialecte algérien et le verbe /yi-nkisir/ dans le dialecte égyptien. Dans le dialecte algérien nous constatons pour cette forme que la voyelle de la lettre /yaa/ de l'imperfectif est une voyelle /a/ comme en MSA mais la première racine du radical après /n/ reste sans voyelle (comme à la forme perfective) ainsi que la voyelle de la deuxième racine du radical est inchangée, elle reste /a/ contrairement au MSA et au dialecte égyptien.

6.1.1.2. Verbes géminés (sourds)

Un verbe géminé est caractérisé par le fait que dans sa racine, la deuxième et la troisième consonne forment une géminée. La forme de surface standard du verbe géminé est [Fa33 (-a)] dont le correspondant dialectal dans l'EA est représenté par la forme [Fa33]. Dans le dialecte TA ces verbes possèdent deux formes : [Fi33] et [Fu33]. Le tableau suivant montre quelques exemples de ses formes :

MSA	EA	TA	Traduction
šadd(-a)	Šadd	šadd	tirer
mall(-a)	Mall	mall	s'ennuyer
jarr(-a)	Jarr	jarr	traquer

Ces formes de verbes géminés ne sont pas similaires à celles des verbes sonores au perfectif. Cette différence est due à plusieurs facteurs que (Wright, 1967) tente de donner en avançant que : "*Lorsque la première et la troisième consonne de la racine ont des voyelles, la deuxième consonne rejette sa voyelle, et s'unit à la troisième, de manière à former une double lettre*". Ainsi, nous pouvons proposer que la forme [Fa33(-a)] est dérivée de la forme [Fa3a3(-a)] en suivant les règles appelées *Métathèse de Consonnes Identiques et Raccourcissement des Syllabes Fermées(MCI&RSF)*[15]. Ces règles s'appliquent à la fois à l'arabe standard et dialectal lors que la forme du verbe est suivie par une voyelle. Ces règles sont représentées comme suit :

Fa3a3(-a) → Faa33(-a) → Fa33(-a) … **MCI&RSF**

Dans certains cas ces règles ne s'appliquent pas : c'est le cas par exemple des verbes en MSA dont la deuxième consonne n'est pas suivie par une voyelle. Cependant, l'arabe dialectal (EA, TA, etc.) applique toujours ces règles parce qu'il insère la voyelle /ee/ en EA et le /i/ en dialecte maghrébin, entre la racine et les suffixes consonantiques. Cette différence traduit l'écart entre l'arabe dialectal et l'arabe standard au niveau de la suffixation consonantique à la forme perfective (accomplie) du verbe géminé. Par ailleurs, (Marçais, 1902) relate que « *la majorité des dialectes arabes ignore le dédoublement de lettre qui*

[15]Identical-Consonant Metathesis and Closed- Syllable Shortening

s'opère en arabe classique aux 1ère et 2ème personne du perfectif dans les verbes sourds. Il intercale à la fin du radical une voyelle longue (î) devant les suffixes de ces personnes; cette particularité se retrouve dans toutes les formes dérivées du verbe sourd ».

MSA	**EA**	**TA**	**Traduction**
šadad-tu	šadd-ee-t	šadd-î-t	j'ai tiré
šadad-naa	šadd-ee-na	šadd-î-na	nous avons tiré
šadad-ta	šadd-ee-t	šadd-î-t	tu as tiré
šadad-ti	šadd-ee-ti	šadd-î-ti	tu as tiré
šadad-tum	šadd-ee-tu	šadd-î-tu	vous avez tiré
šadad-tunna	šadd-ee-tu	šadd-î-tu	vous avez tiré

Par ailleurs, nous présentons dans les tableaux (6.7), (6.8) et (6.9) une comparaison des formes II –X dérivées des verbes géminés en MSA, EA et TA comme suit :

N°	**Forme**	**Exemple**	**Traduction**
I	Fa33(-a)	marr(-a)	passer
II	Fa33a3(-a)	harrar(-a)	libérer
IV	?aFa33(-a)	?amadd(-a)	fournir
V	taFa33a3(-a)	taharrar(-a)	se libérer
VII	(?i)nFa33(-a)	(?i)nsadd(-a)	se bloquer
VIII	(?i)Fta33(-a)	(?i)htajj(-a)	contester
X	(?i)staFa33(-a)	(?i)sta3add(-a)	être prêt

Tableau 6. 7. Les formes des verbes géminés en MSA au perfectif

N	**Forme**	**Exemple**	**Traduction**
I	Fa33	Marr	passer
II	Fa33 $\binom{a}{i}$ 3	Harrar	libérer
		?allil	diminuer
IV	?aFa33	?aSarr	insister
V	(?i)tFa33$\binom{a}{i}$ 3	(?i)tharrar	se libérer
		(?i)tgaddid	se renouveler
VII	(?i)nFa33	(?i)nsadd	se bloquer
VIII	(?i)Fta33	(?i)htagg	Contester
X	(?i)staFa33	(?i)sta3add	être prêt

Tableau 6. 8. Les formes des verbes géminés en EA au perfectif

N	**Forme**	**Exemple**	**Traduction**
I	F$\begin{pmatrix} a \\ i \\ u \end{pmatrix}$ 33	Habb	aimer
		Šidd	prendre
		quss	couper
II	Fa33$\binom{a}{i}$ 3	Qarrar	décider
		Xammim	penser
IV	ϕ	ϕ	ϕ
V	(?i)tFa33$\binom{a}{i}$ 3	(?i)tharrar	se libérer
		(?i)tjassis	éspioner

VII	ϕ	ϕ	ϕ
VIII	(?i)Fta33	(?i)htajj	contester
X	(?i)staFa33	(?i)st3add (?i)st7aqq	être prêt

Tableau 6. 9. Les formes des verbes géminés en TA au perfectif

Dans l'EA, la règle de *Métathèse de Consonnes Identiques* peut être utilisée pour comptabiliser les consonnes géminées dans les formes dérivée IV et X. Cette remarque s'applique aussi au dialecte tunisien pour la forme X. La Forme IV, par exemple, subit la dérivation suivante :

?aF3a3(-a) → ?aFa33(-a) > ?aFa33 … **MCI**

Nous remarquons aussi que dans le dialecte EA, les règles de *Métathèse de Consonnes Identiques et Raccourcissement des Syllabes Fermées* peuvent expliquer la dérivation des formes VII et VIII. Notons aussi que ce constat reste valable aussi pour le dialecte TA concernant la forme VIII. Par exemple, la forme VII est calculée comme suit:

(?i)nFa3a3(-a) → (?i)nFaa33(-a) → (?i)nFa33(-a) > (?i)nFa33 … **MCI&RSF**

L'analyse de tableaux (6.7) et (6.8) nous conduit à constater que les formes des verbes géminés au perfectif dans EA sont presque identiques à leurs homologues en MSA, mise à part la suppression de la flexion (-a) à la fin du mot. Il y a aussi l'inexistence des formes III, VI et IX dans l'arabe dialectal bien qu'elles soient présentent dans l'arabe standard.

Notons aussi que la forme IV, couramment utilisée en MSA, est souvent remplacée par la forme II dans l'arabe dialectal. Ceci est dû au fait que les deux formes sont sémantiquement similaires et signifient la causalité. C'est le cas par exemple du verbe /?atamm(-a)> tammim / « terminer ». En EA, ce remplacement est plus marquant, où la forme IV tend à disparaître des types de forme indépendants de l'EA (Gadalla, 2000).

Nous constatons à partir du tableau (6.9) que la forme IV n'existe pas dans le dialecte tunisien, et dans les dialectes maghrébins en général. La hamza est totalement tombée et la forme est remplacée par la forme I. Le tableau (6.9) donne aussi le même constat pour la forme VII. La conjugaison de ce type de formes se conjugue comme à la première forme des verbes géminés. Cependant, nous notons qu'il existe quelques exceptions pour cette forme, notamment à l'ouest de l'Algérie. C'est le cas du verbe (?i)ndégg (être pilé).

Il est important de signaler une caractéristique de la conjugaison des formes II et V à la forme perfective dans les dialectes maghrébins en général. Il existe un groupement de ($C_1VC_2C_2VC_3$, 't'$C_1VC_2C_2VC_3$) dans lequel le VC_3 disparaît devant les suffixes vocaliques ('it' et 'u') de la 2ème personne au féminin singulier et la 3ème personne au singulier et au pluriel. Par exemple, pour le verbe 'xammim' (réfléchir, se préoccuper) donne pour la 3ème personne au féminin singulier nous obtenons 'xammit' au lieu de 'xammamit' (en EA). Ce regroupement est valable aussi pour la forme imperfective pour la 2ème personne féminin singulier et au pluriel et la 3ème personne au pluriel.

En ce qui concerne les formes des verbes doublés à l'imperfectif, l'EA conserve les mêmes formes morphologiques que celles du MSA, comme le montre les tableaux (6.10) et (6.11). Notons seulement que cette remarque n'est pas valable pour le préfixe imperfectif.

N°	Forme	Example	Traduction
I	ya-F$\begin{pmatrix} a \\ u \\ i \end{pmatrix}$33(-u)	ya-mall(-u)	s'ennuyer
		ya-murr(-u)	Passer
		ya-qill(-u)	devenir moins
II	yu-Fa33i3(-u)	yu-harrir(-u)	se libérer
IV	yu-Fi33(-u)	yu-midd(-u)	Fournir
V	ya-taFa33a3(-u)	ya-taharrar(-u)	être libéré
VII	ya-nFa33(-u)	ya-nsadd(-u)	être bloqué
VIII	ya-Fta33(-u)	ya-htajj(-u)	Protester
X	ya-staFi33(-u)	ya-sta3idd(-u)	être prêt

Tableau 6. 10. Les formes des verbes géminés en MSA à l'imperfectif

N°	Forme	Example	Traduction
I	yi-F$\begin{pmatrix} a \\ u \\ i \end{pmatrix}$33	yi-mall	s'ennuyer
		yi-murr	Passer
		yi-?ill	devenir petit
II	yi-Fa33$\begin{pmatrix} a \\ i \end{pmatrix}$3	yi-harrar	se libérer
		yi-?allil	Diminuer
IV	yi-Fi33	yi-Sirr	Insister
V	(?i)tFa33$\begin{pmatrix} a \\ i \end{pmatrix}$3	yi-tharrar	être libérer
		yi-tgaddid	Renouveler
VII	yi-nFa33	yi-nsadd	être bloqué
VIII	yi-Fta33	yi-htagg	Protester
X	yi-staFi33	yi-sta3idd	se preparer

Tableau 6. 11. Les formes des verbes géminés en EA à l'imperfectif

N°	Forme	Example	Traduction
I	y-F$\begin{pmatrix} a \\ u \\ i \end{pmatrix}$33	y-habb	Aimer
		y-hukk	Gratter
		y-miss	toucher
II	y-Fa33$\begin{pmatrix} a \\ i \end{pmatrix}$3	y-qarrar	decider
		y-xammim	Penser
IV	ϕ	ϕ	ϕ
V	yi-tFa33$\begin{pmatrix} a \\ i \end{pmatrix}$3	yi-tharrar	être libérer
		yi-tjassis	espionner
VII	ϕ	ϕ	ϕ
VIII	yi-Fta33	yi-htajj	protester
X	yi-staFi33	yi-sta3idd	se preparer

Tableau 6. 12. Les formes des verbes géminés en TA à l'imperfectif

(Wright, 1967) explique la différence entre les verbes sonores et géminés comme suit : «*Si la troisième consonne a une voyelle contrairement à la première, la seconde consonne rejette sa voyelle sur la première, puis se combine avec la troisième, de manière à former une double lettre »*. En d'autres termes, l'arabe ne préfère pas que deux syllabes aient la même consonne. Ainsi, une voyelle entre deux consonnes identiques soit elle est supprimée, comme dans les formes perfectives déduites par les règles de *Métathèse de Consonnes Identiques et Raccourcissement des Syllabes Fermées*, ou transformée en métathèse de manière à venir avant les consonnes identiques, pouvant être obtenu en appliquant la règle de la *Métathèse de Consonnes Identiques*. Cette règle s'applique aux formes I, IV, VII, VIII et X à l'imperfectif. Par exemple, l'application de la règle MCI sur la forme X donne le résultat suivant :

ya-staF3i3(-u) → ya-staFi33(-u) > yi-staFi33… **MCI**

L'analyse des tableaux (6.10), (6.11) et (6.12), nous conduit à observer que les verbes géminés peuvent avoir l'une des trois formes suivantes à l'imperfectif : [ya-Fa33 (-u)], [ya-Fu33 (-u)] ou [ya-Fi33 (-u)]. Ces formes ont été retenues dans l'EA avec une tendance à changer la voyelle du stem dans la première forme en /i/. Cette remarque est valable pour le TA seulement pour les formes V, VIII, X. Pour le reste des formes (I-II), le TA effectue le changement en mettant le /soukoun/ au lieu du /i/. Nous donnons ci-après quelques exemples de ces transformations :

MSA	**EA**	**TA**	**Traduction**
ya-wadd(-u)	yi-widd	y-widd	Vouloir
ya-jurr(-u)	yi-gurr	y-jurr	Glisser
ya-hinn(-u)	yi-hinn	y-hinn	être nostalgique

Enfin, notons que certains parlers insèrent la séquence prothétique « ?i » pour éviter la succession de deux consonnes au début de mot. Par ailleurs, le tableau (6.12) montre que les deux formes IV, VII ont disparues dans le dialecte tunisien.

6.1.1.3. Verbes glottalisées (hamzé)

Ces verbes appartiennent à la classe des verbes glottalisés (*Mahmouz*) qui sont caractérisés par la présence du coup de glotte /?/ (*Hamza*) dans la racine trilitère. Le coup de glotte peut se produire dans n'importe quelle position dans le verbe standard: initiale, médiane ou finale. La fréquence d'occurrence de ces verbes est très élevée dans l'arabe MSA par rapport à l'arabe dialectal où cette fréquence est très faible voir dans certains cas nulle.
Le verbe initial-glottalisé de la Forme I se produit en très peu d'exemplaires dans l'EA. Il n'est pas aussi maintenu dans les dialectes maghrébins :

- Les verbes ayant la 2ème radicale (ء), ont été généralement ramenés à la classe des verbes concaves,
- Les verbes ayant la 3ème radicale (ء), ont été déclinés en des verbes défectueux,
- Les verbes à 1ère radicale (ء), dont les deux plus employés /?axað/ (prendre) et /?akal/ (manger) ont pris une conjugaison particulière qui ne se rapproche de celle du verbe hamzé qu'au futur. Les deux verbes : /?amér/ (ordonner), /?amén/ (croire), peuvent passer pour avoir conservé à peu près pure la conjugaison du verbe hamzé, mais sans que jamais la prononciation consonantique du (ء) y soit sensible.

Voici quelques exemples de verbes en MSA et leur homologue en EA et TA :

MSA	**EA**	**TA**	**Traduction**

?akal(-a)	?akal/kal	Klâ	Manger
?axað(-a)	?axad/xad	Xdâ	Prendre
?amar(-a)	?amar	?amar	Ordonner

À l'imperfectif, le coup de glotte est parfois remplacé par l'allongement de la voyelle du préfixe imperfectif en EA et TA. Ceci est accompli par un allongement compensatoire. Voici quelques exemples de ces cas :

MSA	**EA**	**TA**	**Traduction**
ya-?kul(-u)	yaakul	yaakil	Manger
ya-?xuð(-u)	yaaxud	yaaxid	Prendre
ya-?mur(-u)	yu?mur	yu?mur	Ordonner

Il est à noter aussi qu'à la forme imperfective la conjugaison des verbes (kla, xda) est celle de verbes hamzé. Ceci est appuyé par les résultats de l'enquête effectué par (Ouerhani, 2009) sur la réalisation du hamzé à la forme imperfective. Dans la totalité des réponses obtenues, la « hamza » n'est pas réalisée. Elle est en effet allégée en une voyelle amalgamée à la voyelle d'avant pour former une voyelle longue. Ainsi la première syllabe du verbe à l'accompli se transforme en CV à la place de CVC, avec une « hamza » en deuxième consonne.

Les formes dérivées du verbe initial-glottalisé préservent leur /?/ en EA, à l'exception de certains verbes de forme X, comme présenté dans les exemples suivants :

Forme	**MSA**	**EA**	**Traduction**
II	?axxar(-a)	?axxar	Retarder
	?ajjal(-a)	?aggil	Reporter
IV	?aanas(-a)	?aanis	Encourager
	?aaman(-a)	?aamin	Croire
V	ta?akkad(-a)	(?i)t?akkid	être sûre
X	(?i)sta?jar(-a)	(?i)sta?gar	Louer
	(?i)sta?ðan(-a)	(?i)sta?zin	demander autorisation
	(?i)sta?hal(-a)	(?i)staahil	Mériter

Les verbes de la forme IV /?aanas(-a) > ?aanis/ et /?aaman(-a) > ?aamin/ sont dérivés des formes de base /?a?nas(-a)/ et /?a?man(-a)/ en effectuant un allongement compensatoire. C'est le seul cas où cette règle s'applique en MSA. Le changement du verbe /(?i)sta?hal(-a) > (?i)staahil/ 'mériter' est aussi obtenu par allongement compensatoire. L'arrêt glottal est aussi conservé dans les formes imperfectives de ces verbes en EA comme donné dans le tableau suivant :

Form	**MSA**	**EA**	**Traduction**
II	yu-?axxir(-u)	yi-?axxar	Retarder
	yu-?ajjil(-u)	yi-?aggil	Reporter
IV	yu-?nis(-u)	yi-?aanis	Encourager

	yu-?min(-u)	yi-?aamin t	Croire
V	ya-ta?akkad(-u)	yi-t?akkid	être sûre
X	ya-sta?jir(-u)	yi-sta?gir	Louer
	ya-sta?ðin(-u)	yi-sta?zin	demander autorisation
	ya-sta?hil(-u)	yi-staahil	Mériter

L'usage de /?/ dans des positions médianes et finales est encore restreint en AE. Ainsi, les verbes suivants en MSA ne sont plus utilisés en arabe dialectal en général : /da?ab(-a)/ 'continuer', /sa?im(-a)/ 's'ennuyer' ou /Da?ul(-a)/ 'diminuer'. Il existe très peu d'exemples de verbes de la forme I glottalisés au milieu dans la variété dialectale :

MSA	**EA**	**TA**	**Traduction**
sa?al(-a)	sa?al	s?al	demander
ra?as(-a)	ra?as	r?as	se diriger
ra?af(-a)	ra?af	r?af	avoir pitié
ya?is(-a)	ya?is	y?is	perdre espoir

Ces verbes gardent leur /?/ dans les formes imperfectives dialectales comme suit :

MSA	**EA**	**TA**	**Traduction**
ya-s?al(-u)	yi-s?al	yi-s?al	Demander
ya-r?as(-u)	yi-r?as	yi-r?as	se diriger
ya-r?af(-u)	yi-r?af	yi-r?af	avoir pitié
ya-y?as(-u)	yi-y?as	yi-y?as	perdre espoir

Pour les verbes glottalisés au milieu, ils n'ont pas de formes dérivées dans l'arabe dialectal. Cependant, il y a quelques exceptions, comme pour la forme dialectale [(?i)tFaa3iL] qui est équivalente à la forme Standard [taFaa3aL(-a)], où le /3/ est un arrêt glottal. C'est le cas des verbes /tafaa?al(-a) > (?i)tfaa?il/ 'être optimiste' et /tašaa?am(-a) > (?i)tšaa?im/ 'être pessimiste'. Les formes imperfectives de ces verbes sont /ya-tafaa?al(-u) > yi-tfaa?il/ et /ya-tašaa?am(-u) > yi-tšaa?im/, respectivement.

Concernant les verbes glottalisés en position finale, ils perdent leur /?/ en arabe dialectal par suppression du /?/ final, comme dans le verbe /?ara, yi-?ra/ 'lire' qui constitue le reflet du verbe /qara?(-a), ya-qra?(-u)/ en MSA. La variété dialectale offre aussi des formes dérivées des verbes glottalisés en position finale ayant perdu leur /?/, comme c'est donné dans les exemples du tableau suivant :

MSA	**EA**	**TA**	**Traduction**
?aqra?(-a)	?arra	Qarra	faire lire
tabarra?(-a)	(?i)tbarra	Tbarra	Renier
(?i)btada?(-a)	(?i)btada	Bda	commencer

Toutefois, dans certain cas, l'arrêt glottal en position finale est conservé dans l'EA comme donné dans le tableau suivant :

MSA	EA	Traduction
tahayya?(-a)	(?i)thayya?	à préparer
(?i)stahza?(-a)	(?i)stahza?	se moquer
?axTa?(-a)	?axTa?	se tromper

Selon (Gadalla, 2000), une divergence morpho-phonémique entre l'arabe MSA et l'EA est à signaler à ce niveau : l'arrêt glottal à la fin des verbes glottalisés en position finale en MSA, avec la voyelle qui le précède, sont tous les deux remplacés par la voyelle longue /ee/ en EA. Ce replacement est effectué avant les suffixes pronominaux commençant par une consonne. Dans le dialecte maghrébin, cette règle est aussi appliquée mais au lieu d'utiliser la voyelle longue /ee/ nous utilisons la voyelle /î/. Voici quelques exemples d'illustration :

MSA	EA	TA	Traduction
qara?-tu	?aree-t	qrê-t	j'ai lu
xabba?-ta	xabbee-t	xabbî-t	tu as caché (msg)
hanna?-tum	hannee-tu	hannî-tu	vous avez félicité
tabarra?-ti	?itbarree-ti	tbarrî-ti	tu as renié (fsg)

Le changement qui se produit ici suit le sens suivant :

a? → ay → ee (EA)

(Gadalla, 2000) propose par ailleurs une solution mécanique pour répondre à la question de la transformation de /a?/ en /ay/ et non pas en un /aa/, par application de la règle de allongement compensatoire. Cette solution est représentée par une règle spéciale changeant /a?/ en /ay/ dans un contexte particulier comme suit :

a? → ay/ *dans les verbes glottalisés en position finale*

Cependant, une autre explication peut être avancée en mettant en avant que la hamza /?/ dans la racine des verbes est supprimée par la règle de suppression finale de /?/. Seulement, les parlers égyptiens considèrent le verbe comme un verbe faible final dont le /ee/ est dérivé de /ay/ par l'opération de la *monophtongaison*.

6.1.1.4. Les verbes faibles

Dans l'arabe traditionnel, les verbes contenant dans leurs radicaux l'une ou les deux glides suivantes /w/ et /y/ sont considérés comme des verbes 'faibles ou défectueux', nommés en arabe /*mu3tall-ah*/. Ces verbes possèdent différents types en fonction de la position de la glide: initiale, médiane ou finale; ou une combinaison de deux positions, à l'exclusion de l'initiale et la médiane. Nous retrouvons tous ces types dans l'arabe dialectal avec quelques modifications.

6.1.1.4.1. Les verbes ayant une glide initiale (appelés aussi verbes assimilé)

Ces verbes peuvent être classés en deux groupes : les verbes en [w] et les verbes en [y]. Nous retrouvons les verbes de ce premier groupe fréquemment que ça soit en MSA ou en arabe dialectal comme donné par les exemples du tableau suivant :

MSA	EA	TA	Traduction
wazan(-a)	wazan	wzin	Peser
waSal(-a)	wiSil	wsol	Arriver
waqa3(-a)	wi?i3	wqa3	tomber (EA), se passer ou avoir lieu (TA)
waqaf(-a)	wi?if	wqef	Arrêter

Dans la plupart des cas, les verbes faibles en [w] en position initiale perdent ce [w] à la forme imperfective en MSA, tandis qu'en EA et dans les autres dialectes maghrébins, la glide est conservée à chaque fois qu'elle est suivie par une consonne autre que /?/. Dans le cas où elle est suivie par un /?/ elle est remplacée par /u/ ou conservée dans le dialecte tunisien. Selon (Marçais, 1902), le verbe ayant la 1ère radicale [w] ou [y], conserve dans la plupart des dialectes ses glides avec l'annexion des préfixes du futur. La conjugaison de ces verbes est tout à fait similaire à celle des verbes réguliers. Selon toujours la même référence, la combinaison diphtonique de la voyelle du préfixe et du [w] ou du [y] se maintient régulièrement, par exemple : les verbes newsol, teyibes, dans le dialecte algérien se prononcent dans les dialectales orientaux et tunisien : nusol, tibes respectivement.

La règle suivante explique les changements qui se produisent en MSA en supprimant le /w/ après une voyelle et avant une consonne suivie par /i/. Cette règle est appelée par (Brame, 1970) "l'Occultation du w".

$w \rightarrow Ø / V — Ci$ **… Occultation du w en MSA**

Les changements de la règle équivalente en EA sont donnés comme suit :

1. $Vw? \rightarrow w? \rightarrow u?$
2. $VwC \rightarrow VwC$

… Occultation du w en AE

Nous signalons que le MSA possède deux formes pour les verbes à l'imperfectif en w, à savoir: [ya-3iL(-u)] et, moins fréquemment, [ya-3aL(-u)]. En revanche, l'EA en propose trois équivalents : [yi-w3iL], [yi-w3aL] et [yu-3aL]. En ce qui concerne le dialecte Tunisien, ce dernier possède les formes suivantes [yu-3iL], [yu-3uL], [yi-3aL]. Nous trouvons aussi ces formes dans le dialecte tlemcénien et algérois qui en rajoutent en plus la forme yé-w3al. Voici quelques illustrations :

MSA	EA	TA	Traduction
ya-zin(-u)	yi-wzin	yu-Wzin	Peser
ya-Sil(-u)	yi-wSal	yu-wSul	Arriver
ya-qa3(-u)	yu-?a3	yu-wqa3	tomber (EA), se passer ou avoir lieu (TA)
ya-qif(-u)	yu-?af	yu-wqaf	arrêter / stand up (debout)

Il est à noter que la règle *d'occultation du w* s'applique sur les verbe en MSA /ya-qa3(-u)/ 'tomber' même s'il n'y a pas de /i/ après la lettre /q/. Ceci est dû au fait que cette règle s'applique avant une autre règle qui change le /i/ en /a/ aux alentours d'un laryngale (larynx), la /3/ dans notre cas. Cette règle a été proposée par (Brame, 1970), nommée '*l'Assimilation Laryngale en MSA*'.

$i \rightarrow a / \begin{Bmatrix} L\ — \\ —\ L \end{Bmatrix}$/imperfectif **… l'Assimilation Laryngale en MSA**

*Où "L" fait référence à une laryngale : /3, h, H ou ?/

.

Concernant les verbes primaires en [y] avec glide initiale, leur nombre est extrêmement faible dans l'arabe standard et dialectal. Ils se comportent comme des verbes sains et restent conformes à leurs formes dans les deux variétés, en voici quelques exemples :

MSA	EA	TA	Traduction
ya?is(-a)	yi?is	y?is	désespérer (pf)
yabis(-a)	yibis	Ybis	sécher (pf)
ya-y?as(-u)	yi-y?as	yi-y3is	désespérer (impf)
ya-ybas(-u)	yi-ybas	yi-ybis	sécher (impf)

La même remarque est valable aussi pour le nombre des formes dérivées des verbes assimilés qui reste assez faible à la fois dans l'arabe standard et dialectal. Seulement, nous pouvons mettre en avant une distinction majeure qui concerne la Forme VII. Cette forme est formée à partir des racines initiales /w/ dans l'EA, ce qui est impossible en MSA. Voici un tableau comparatif des verbes dérivés en MSA, EA et le TA :

Forme	MSA	EA	TA	Traduction
II	wadda3(-a)	wadda3	Wadda3	pour dire adieu
"	yabbas(-a)	yabbis	Yabbis	sécher
III	waafaq(-a)	waafi?	Waafaq	accepter
			Chaarik	participer
IV	?awqa3(-a)	wa??a3	wazza3	à faire tomber
"	?awqaf(-a)	wa??af	Waqqaf	arrêter
V	tawarraT(-a)	(?i)twarraT	TwarraT	s'être mêlé
VI	tawaajah(-a)	(?i)twaagih	tqaatil	faire face l'un l'autre
VII	ϕ	(?i)nwazan	ϕ	à peser
"	ϕ	(?i)nwaga3	ϕ	à ressentir de la douleur
VIII	(?i)ttaSal(-a)	(?i)ttaSal	?ittaSil	contacter
X	(?i)stawda3(-a)	(?i)stawda3	ϕ	à laisser, dépôt

Ces exemples montrent clairement que l'EA utilise la Forme II pour remplacer la Forme IV, ce qui indique que la dernière forme commence à disparaitre de cette variété. Nous pouvons aussi observer que la glide initiale est remplacée par un /t/ dans la Forme VIII dans les différentes variétés. Ceci peut être justifié par la règle spéciale suivante :

$$w \rightarrow t \,/\, \text{—}\ \text{t dans la Forme VIII}$$

Nous remarquons aussi que la forme VII ne s'applique pas en arabe MSA à ces verbes. Il en est de même dans la plupart des dialectes. Cependant en EA, tout comme en Tlemcenien en Algérie, il arrive que ce 'nif'al' moderne soit adapté à des racines, assimilées pour leur donner une signification passive.

Par ailleurs, les dialectes maghrébins ramènent à /w/ la première lettre d'un bon nombre de verbe à première radical hamza, comme dans le verbe wânes du verbe Anas (tenir compagnie), wâlef, du verbe alaf (devenir familier).

Pour la forme X, (Marçais, 1902) souligne qu'elle existe dans le dialecte tlemcenien et que : la diphtongaison due à la combinaison de la voyelle du préfixe formatif et de la première radicale [w] est maintenue comme pour les verbes : [?ista-wjab – ss-wjab et s-wjab]. Par ailleurs, nous remarquons dans le tableau que cette forme existe dans l'arabe MSA et EA, mais elle n'est pas utilisée dans le TA et la majorité des dialectes maghrébins.

Ceux-ci sont les équivalents imparfaits des verbes dérivés ci-dessus :

Forme	MSA	EA	TA	Traduction
II	yu-waddi3 (-u)	yi- wadda3	y-wadda3	pour dire adieu
	yu-yabbis(-u)	yi-yabbis	y-yabbis	sécher
III	yu-waafiq(-u)	yi-waafi?	y-waafaq	accepter
			y-chaarik	participer
IV	yuuqi3(-u)	yi-wa??a3	y-wazza3	à faire tomber
	yuuqif(-u)	yi-wa??af	y-waqqaf	arrêter
V	ya-tawarraT(-u)	yi-twarraT	y-twarraT	s'être mêlé
VI	ya-tawaajah(-u)	yi-twaagih	yi-twaagih yi-tqaatil	faire face à l'autre
VII	ϕ	yi-nwizin	ϕ	à peser
	ϕ	yi-nwigi3	ϕ	à ressentir de la douleur
VIII	ya-ttaSil(-u)	yi-ttiSil	yi-ttaSil	contacter
X	ya-stawdi3(-u)	yi-stawdi3	ϕ	à laisser, dépôt

6.1.1.4.2. Les verbes concaves (creux)

Ce sont les verbes qui ont un /aa/ superficiel dans certaines formes et /w/ ou /y/ dans d'autres formes, au milieu de leurs racines. Les tableaux suivants (6.13), (6.14) et (6.15) présentent les formes dérivées des verbes creux dans l'arabe MSA, EA et TA respectivement.

No	Forme	Exemple	Traduction
I	FaaL(-a)	gaab(-a)	s'absenter
II	$\mathrm{Fa}\binom{ww}{yy}\mathrm{aL(-a)}$	Sawwar(-a)	photographier
III	$\mathrm{Faa}\binom{w}{y}\mathrm{aL(-a)}$	3ayyan(-a)	designer
		haawal(-a)	essayer
		3aayan(-a)	inspecter
IV	?aFaaL(-a)	?abaad(-a)	éradiquer
V	$\mathrm{taFa}\binom{ww}{yy}\mathrm{aL(-a)}$	taSawwar(-a)	imaginer
		taxayyal(-a)	imaginer
VI	$\mathrm{taFaa}\binom{w}{y}\mathrm{aL(-a)}$	ta3aawan(-a)	collaborer
		ta3aayaš(-a)	cohabiter
VII	(?i)nFaaL(-a)	(?i)mbaa3(-a)	être vendu
VIII	(?i)FtaaL(-a)	(?i)xtaar(-a)	choisir
IX	$(?i)\mathrm{F}\binom{w}{y}\mathrm{aLL(-a)}$	(?i)swadd(-a)	devenir noir
		(?i)byaDD(-a)	devenir blanc
X	(?i)staFaaL(-a)	(?i)stafaad(-a)	bénéficier

Tableau 6. 13. Les formes perfectives des verbes concaves en MSA

No	Forme	Exemple	Traduction
I_a	FaaL	Gab	s'absenter
I_b	Fa$\binom{w}{y}$aL	Dawaš	prendre une douche
		Xayal	distraire
II_a	Fa$\binom{ww}{yy}$aL	Sawwar	photographier
		3ayyaT	Pleurer
II_b	Fa$\binom{ww}{yy}$iL	Kawwin	Former
		3ayyin	designer
III	Faa$\binom{w}{y}$iL	Haawil	Essayer
		3aayin	inspecter
IV	?aFaaL	?abaad	éradiquer
V_a	(?i)tFa$\binom{ww}{yy}$aL	(?i)tSawwar	imaginer
		(?i)txayyal	imaginer
V_b	(?i)tFa$\binom{ww}{yy}$iL	(?i)tkawwin	se former
		(?i)t3ayyin	être designé
VI	(?i)tFaa$\binom{w}{y}$iL	(?i)t3aawin	coopérer
		(?i)t3aayin	être désigné
VII_a	(?i)nFaaL	(?i)mbaa3	être vendu
VII_b	(?i)nFa$\binom{w}{y}$aL	(?i)nxawat	être harcelé
		(?i)ndayan	s'endetter
VIII	(?i)FtaaL	(?i)xtaar	Choisir
IX	(?i)F$\binom{w}{y}$aLL	(?i)swadd	devenir noir
		(?i)byaDD	devenir blanc
X_a	(?i)staFaaL	(?i)stafaad	bénéficier
X_b	(?i)staF$\binom{w}{y}$aL	(?i)stabwax	se moquer
		(?i)stašyax	prétendre être un imam
X_c	(?i)satF$\binom{w}{y}$iL	(?i)stamwit	prétendre être mort
		(?i)sta3yib	désohonorer

Tableau 6. 14. Les formes perfectives des verbes concaves en EA

No	Forme	Exemple	Traduction
I	FaaL	Qaal	dire
II_a	Fa$\binom{ww}{yy}$aL		
		Qayyal	faire une sieste
II_b	Fa$\binom{ww}{yy}$iL	Sawwir	photographier
		Tayyib	cuisiner
III	Faa$\binom{w}{y}$iL	Haawil	essayer
		3aayin	inspecter
IV	ϕ	ϕ	ϕ
V_a	(?i)tFa$\binom{ww}{yy}$aL	(?i)tSawwar	se prendre en photo
		(?i)txayyal yyil?	imaginer
V_b	(?i)tFa$\binom{ww}{yy}$iL	(?i)tkawwin	se former
		(?i)t3ayyin	être désigné
VI	(?i)tFaa$\binom{w}{y}$iL	(?i)t3aawin	coopérer
		(?i)t3aayin	être désigné
VII	ϕ	ϕ	ϕ

VIII	(?i)FtaaL	(?i)xtaar	choisir
IX	F$\binom{w}{y}$aaL	swaad	devenir noir
		byaaD	devenir blanc
X_a	(?i)stFaaL	(?i)stfaad	bénéficier
X_b	(?i)satF$\binom{w}{y}$iL	(?i) staHwiD	
		ϕ	ϕ

Tableau 6. 15. Les formes perfectives des verbes concaves en TA

En comparant les tableaux (6.13), (6.14) et (6.15), nous voyons bien que les verbes concaves se ressemblent au niveau des consonnes en MSA, TA et EA. Cependant, ils divergent un peu au niveau des voyelles, par exemple : la voyelle avant le dernier radical est toujours /a/ en MSA alors qu'en EA et TA elle est changée en /i/ dans les deux formes III et VI et dans certains verbes des formes II et V. Ces tableaux révèlent aussi que les formes avec /aa/ au milieu du radical sont I, IV, VII, VIII et X pour le MSA et EA. Ce fait est valable aussi pour le TA sauf pour les formes IV et VII qui n'existent pas dans ce dialecte ; et aussi pour la forme (?i)staFaaL (X) est remplacée par (?i)stFaaL. Nous constatons aussi dans ces tableaux que les formes ayant /w/ ou /y/ sont les formes II, III, V, VI et IX.

Selon (Gadalla, 2000), les glides médianes sont conservées dans certaines formes et remplacées par une voyelle longue dans d'autres formes. Ceci est dû au fait que ces formes suivent la généralisation donnée par (Thackston, 1984) : "La règle de base est la suivante : toute glide entourée par des voyelles brèves est ignorée accompagnée de la voyelle qui la suit, tandis que la voyelle précédente est allongée avec compensation si possible". Autrement dit, la non-application de cette règle pourrait créer une syllabe très lourde. De ce fait, les règles responsables de la modification d'une glide en une longue voyelle dans les verbes creux en arabe standard et dialectal sont :

1. CVGV → CVV **… Elision de la glide**
2. CGV → CVV **… Assimilation vocoïde anticipatoire**

De plus, l'élision de la glide s'applique aux Formes I, VII et VIII, mais ne s'applique ni aux formes II et V en raison de l'inaltérabilité des géminées, ni à la forme III puisque la glide [y] est précédée par une voyelle longue. Pour sa part, l'Assimilation vocoïde anticipatoire s'applique aux formes IV et X, mais ne s'applique pas à la forme IX car sans cette application le résultat serait une syllabe extrêmement lourde. En d'autres termes, [(?i)FGVLL(-a)] ne devrait pas se changer en [(?i).FVVL.L(-a)] qui contient une syllabe super-lourde non finale. De ce fait, nous pouvons affirmer que l'exigence que /CGV/ devienne /CVV/ est subordonné à la condition qu'il n'y ait pas de syllabes super-lourdes non finales. Cette règle aussi est présente dans la forme primaire qui se voit remplacer la séquence [VGV] par une longue voyelle via l'Élision de la glide en arabe standard et dialectal, et devient ainsi [FaaL(-a)], comme illustré dans les exemples suivants:

Racine	**MSA**	**EA**	**Traduction**
q-w-l	qaal(-a)	?aal	dire
S-w-m	Saam(-a)	Saam	jeûner
g-y-r	gaar(-a)	Gaar	être jaloux
b-y-3	baa3(-a)	baa3	vendre

Cependant, lorsqu'un suffixe commençant par une consonne est attaché à un verbe primaire creux, la séquence [VGV] est changée en /u/ pour les verbes en w, et en /i/ pour les verbes en y, dans les deux variétés (MSA et dialecte). La raison derrière ce changement est que la séquence [CVGVC] donne naissance à une syllabe super-lourde quand elle subit la règle /CVGV → CVV/. Par conséquent, la meilleure façon d'éviter les conséquences de l'application de la règle /CVGV → CVV/, est de penser que les formes /CVwV/ et /CVyV/ sont en général impossibles, et qu'il devrait y avoir des 'réparations' afin d'éviter ces séquences. L'une de ces réparations implique les règles suivantes : /awa → aa/ et /aya → aa/, mais son application est soumise au fait qu'il ne faut pas obtenir au final une syllabe superlourde, comme dans /kawan-tu → *kaan-tu/ 'J'étais'. Dans ces cas une réparation alternative doit être choisie entre /awa → u/ et /aya → i/, comme dans les verbes suivants :

MSA	**EA**	**TA**	**Traduction**
qul-tum	?ul-tu	qul-tu	vous avez dit
ju3t-u	gu3-t	ju3-t	je suis devenu faim
gir-ta	gir-t	gir-t	tu es devenu jaloux
bi3-naa	bi3-na	ba3-na	nous avons vendu

(Thackston, 1984) a identifié que les verbes /xaaf(-a)/ 'avoir peur', /naam(-a)/ 'dormir' et /maat(-a)/ 'mourir' représentent des exceptions aux règles présentées précédemment, puisqu'ils se manifestent sous la forme /xif-/, /nim-/ et /mit-/ lorsqu'ils sont suivis par des suffixes commençant par une consonne. Cette situation est devenue possible puisque la forme de base de ces verbes est /xawif(-a)/, /nawim(-a)/ et /mawit(-a)/, ainsi que le choix de /i/ comme second vocalisme est permis. Nous notons que l'EA et le TA diffèrent du MSA par le fait qu'ils possèdent les formes /xuf-t/ et /mut-t/ qui sont attendues de /xawif(-a)/ et /mawit(-a)/ et qui ont été régularisés en /xawaf(-a)/ et /mawat(-a)/. La seule exception qui persiste encore en EA est le verbe /nim-t/.

Nous signalons aussi une divergence phonologique entre les deux variétés qui se manifeste dans le fait que la voyelle longue /aa/ est raccourcie en EA et TA par *raccourcissement atonique*, et ce lorsque le verbe creux est suivi d'une voyelle longue, comme le montrent les exemples suivants :

MSA	**EA**	**Traduction**
qaal-uu lii	?al-uú-li	Ils m'ont dit
baa3-uu laka	ba3-uú-lak	Ils t'ont vendu
(?i)xtaar-uu-haa	(?i)xtar-uú-ha	Ils l'ont choisi

Une autre différence entre l'EA et TA d'un côté et le MSA d'un autre côté, réside dans la tendance de l'arabe dialectal pour les glides ou les faibles consonnes /w/ et /y/ à être de plus en plus traitées comme des consonnes fortes. Ceci a été suggéré par (Carter, 1996) : "*Il existe maintenant plusieurs cas où les semi-voyelles w et y fonctionnent comme des consonnes normales, tout en créant (ou recréant?) de nouveaux paradigmes*". Comme le montrent les tableaux (6.13), (6.14) et (6.15), les glides apparaissent dans les Formes II, III, V, VI et IX dans les deux variétés. La nouveauté est que dans l'EA et TA ces consonnes commencent à apparaitre aussi dans les Formes I, VII et X. D'autres exemples de verbes en EA avec des glides médianes sont donnés dans (Gadalla, 2000) comme suit :

Forme	Exemple	Traduction
I	hawag	Avoir besoin
	rawaš	Perturber
	ziwir	Etouffer
VII	(?i)nhawal	Développer un strabisme
	(?i)nhawag	Avoir besoin
X	(?i)sta3wa?	Considérer tard
	(?i)stabya3	Agir imprudemment
	(?i)staxwin	Se trahir
	(?i)staxyib	Se ridiculiser

Ces exemples montrent que l'EA a développé ses propres formes en /w/ et / y / qui n'existent pas en MSA. Pour expliquer cela, (Gadalla, 2000) avance que les règles de l'élision de la glide (dans les formes I et VII) et l'assimilation vocoïde anticipatoire (dans la forme X) deviennent faibles dans l'AE. Par conséquent, les formes creuses traditionnelles et les nouvelles formes fortes existent maintenant côte à côte dans l'AE. En projetant cette analyse, nous pouvons nous attendre à ce que dans l'avenir, l'EA pourrait également développer une nouvelle forme IV, à la place ou côte à côte avec la forme [?aFaaL] en retenant simplement la forme de base [?aFGaL]. Cette projection reste valable si seulement si la forme IV survive, car il existe plusieurs indications montrant que cette forme commence à disparaître.

Les tableaux (6.16), (6.17) et (6.18) montrent les formes imperfectives des verbes défectueux médians dans l'MSA, l'EA et le TA.

No	Forme	Exemple	Traduction
I	ya-F$\begin{pmatrix} aa \\ uu \\ ii \end{pmatrix}$L(-u)	ya-naam(-u)	dormir
		ya-quul(-u)	dire
		ya-giib(-u)	absenter
II	yu-Fa$\begin{pmatrix} ww \\ yy \end{pmatrix}$iL(-u)	yu-Sawwir(-u)	photographier
		yu-3ayyin(-u)	inspecter
III	yu-Faa$\begin{pmatrix} w \\ y \end{pmatrix}$iL(-u)	yu-haawil(-u)	essayer
		yu-3aayin(-u)	inspecter
IV	yu-FiiL(-u)	yu-biid(-u)	éradiquer
V	ya-taFa$\begin{pmatrix} ww \\ yy \end{pmatrix}$aL(-u)	ya-taSawwar(-u)	imaginer
		ya-taxayyal(-u)	imaginer
VI	ya-taFaa$\begin{pmatrix} w \\ y \end{pmatrix}$aL(-u)	ya-ta3aawan(-u)	coopérer
		ya-ta3aayaš(-u)	cohabiter
VII	ya-nFaaL(-u)	ya-mbaa3(-u)	être vendu
VIII	ya-FtaaL(-u)	ya-xtaar(-u)	choisir
IX	ya-F$\begin{pmatrix} w \\ y \end{pmatrix}$aLL(-u)	ya-swadd(-u)	devenir noir
		ya-byaDD(-u)	devenir blanc
X	ya-staFiiL(-u)	ya-stafiid(-u)	Bénéficier

Tableau 6. 16. Les formes imperfectives des verbes défectueux médians en MSA

No	Forme	Exemple	Traduction
$\mathbf{I_a}$	yi-F$\begin{pmatrix} aa \\ uu \\ ii \end{pmatrix}$L	yi-naam	dormir
		yi-?uul	dire
		yi-giib	absenter
$\mathbf{I_b}$	yi-F $\begin{pmatrix} w \\ y \end{pmatrix}$iL	yi-dwiš	déranger
		yi-xyil	distraire
$\mathbf{II_a}$	yi-Fa$\begin{pmatrix} ww \\ yy \end{pmatrix}$aL	yi-Sawwar	photographier
		yi-3ayyaT	pleurer
$\mathbf{II_b}$	yi-Fa$\begin{pmatrix} ww \\ yy \end{pmatrix}$iL	yi-kawwin	former
		yi-3ayyin	désigner
III	yi-Faa$\begin{pmatrix} w \\ y \end{pmatrix}$iL	yi-Haawil	essayer
		yi-3aayin	inspecter
IV	yi-FiiL	yi-biid	éradiquer
$\mathbf{V_a}$	yi-tFa$\begin{pmatrix} ww \\ yy \end{pmatrix}$aL	yi-tSawwar	imaginer
		yi-txayyal	imaginer
$\mathbf{V_b}$	yi-tFa$\begin{pmatrix} ww \\ yy \end{pmatrix}$iL	yi-tkawwin	former
		yi-t3ayyin	être désigné
VI	yi-tFaa$\begin{pmatrix} w \\ y \end{pmatrix}$iL	yi-t3aawin	coopérer
		yi-t3aayin	être inspecté
$\mathbf{VII_a}$	yi-nFaaL	yi-mbaa3	être vendu
$\mathbf{VII_b}$	yi-nFi$\begin{pmatrix} w \\ y \end{pmatrix}$iL	yi-nxiwit	être harcelé
		yi-ndiyin	s'endetter
VIII	yi-FtaaL	yi-xtaar	choisir
IX	yi-F$\begin{pmatrix} w \\ y \end{pmatrix}$aLL	yi-swadd	devenir noir
		yi-byaDD	devenir blanc
$\mathbf{X_a}$	yi-staF$\begin{pmatrix} aa \\ ii \end{pmatrix}$L	yi-stafaad	bénéficier
		yi-stafiid	bénéficier
$\mathbf{X_b}$	yi-satF$\begin{pmatrix} w \\ y \end{pmatrix}$iL	yi-stabwix	se moquer
		yi-sta3yib	désohonorer

Tableau 6. 17. Les formes imperfectives des verbes défectueux médians en EA

No	Forme	Exemple	Traduction
I	y-F$\begin{pmatrix} aa \\ uu \\ ii \end{pmatrix}$L	y-naam	Dormer
		y-quul	Dire
		y-giib	absenter
$\mathbf{II_a}$	y-Fa$\begin{pmatrix} ww \\ yy \end{pmatrix}$aL	y-dawwar	Tourner
		y-bayyaD	Blancher
$\mathbf{II_b}$	y-Fa$\begin{pmatrix} ww \\ yy \end{pmatrix}$iL	y-Tayyib	Cuisine
		yi-3ayyin	désigner
III	y-Faa$\begin{pmatrix} w \\ y \end{pmatrix}$iL	y-Haawil	Essayer
		y-3aayin	inspecter
IV	ϕ	ϕ	ϕ
$\mathbf{V_a}$	yi-tFa$\begin{pmatrix} ww \\ yy \end{pmatrix}$aL	yi-tSawwar	imaginer
		yi-txayyal	Imaginer
$\mathbf{V_b}$	yi-tFa$\begin{pmatrix} ww \\ yy \end{pmatrix}$iL	yi-tkawwin	Former
		yi-t3ayyin	être inspecté
VI	yi-tFaa$\begin{pmatrix} w \\ \end{pmatrix}$iL	yi-t3aawin	Coopérer

		yi-t3aayin	être inspecté
VII	ϕ	ϕ	ϕ
VIII	yi-FtaaL	yi-xtaar	Choisir
IX	yi-F$\binom{w}{y}$aLL	yi-swadd	devenir noir
		yi-byaDD	devenir blanc
X_a	yi-staF$\binom{aa}{ii}$L	yi-stafaad	Bénéficier
		yi-stafiid	Bénéficier
X_b	yi-satF$\binom{w}{y}$iL	yi-staHwiD	‘’
		yi-**staryiD**	‘’

Tableau 6. 18. Les formes imperfectives des verbes défectueux médians en TA

La comparaison des tableaux (6.16), (6.17) et (6.18), tout comme celle faite pour les tableaux (6.13), (6.14) et (6.15), montre que l’arabe standard et dialectal sont traités de manière similaire à cause de l’applicabilité des deux règles, régissant le changement de la glide médiane en une voyelle longue, aux formes imperfectives des verbes creux.

En effet, l’élision de la glide s’applique aux formes I, VII et VIII, tandis que l’Assimilation vocoïde anticipatoire s’applique aux formes IV et X. Les raisons d’application ou de non application de ces deux règles aux différentes formes, ont déjà été données dans la discussion autour des changements qui s’opèrent sur les formes imperfectives. Nous notons aussi que les formes IV et VII n’existent pas dans le dialecte tunisien, et maghrébin en général. Il y aussi une différence remarquable entre l’arabe MSA et l’EA au niveau des formes I, VII et X : l’assimilation vocoïde anticipatoire s’applique toujours à tous les verbes des formes I et X en MSA alors qu’elle ne s’applique qu’à quelques verbes de ces formes en AE, alors que les autres verbes de ces formes conservent leurs formes de base, à savoir : [yi-FGiL] et [yi-staFGiL]. Et pour la forme VII, l’élision de la glide s'applique à tous les verbes imperfectifs en MSA, ce qui n’est pas le cas dans l’EA sauf pour quelques verbes de manière exceptionnelle. Ces comparaisons confirment quelque part, l’hypothèse avancée par (Gadalla, 2000) stipulant que le champ d'application de ces règles est en diminution apparente dans l'AE. Ces comparaison confirment aussi cette hypothèse dans l’arabe TA.

La forme primaire des verbes creux en MSA suit l’une des formes imperfectives suivantes : [ya-FaaL(-u)], [ya-FuuL(-u)] et [ya-FiiL(-u)]. Ces formes sont conservées dans les dialectes EA et TA avec quelques changements particuliers dû à la suppression du suffixe flexionnel et le changement du préfixe imperfectif en [yi-] pour l’EA et en [y-] pour le TA. Voici quelques exemples illustratifs :

Racine	**MSA**	**EA**	**TA**	**Traduction**
x-w-f	ya-xaaf(-u)	yi-xaaf	y-xaaf	avoir peur
m-w-t	ya-muut(-u)	yi-muut	y-muut	mourir
b-y-3	ya-bii3(-u)	yi-bii3	y-bii3	vendre
3-y-š	ya-3iiš(-u)	yi-3iiš	y-3iiš	vivre

Il est perceptible dans les exemples, présentés ci-dessus, que les formes imperfectives [ya-FaaL(-u)$_{MSA}$, yi-FaaL$_{EA}$, y-FaaL$_{TA}$] et [ya-FuuL(-u)$_{MSA}$, yi-FuuL$_{EA}$, y-FuuL$_{TA}$] sont utilisées pour les verbes en [w]. D’autre part, la forme [ya-FiiL(-u)$_{MSA}$, yi-FiiL$_{EA}$, y-FiiL$_{TA}$] est la seule forme utilisée pour les verbes en [y]. Par ailleurs, les formes imperfectives de deux verbes /baat(-a)/ ‘passer la nuit’ et /baan(-a)/ ‘apparaitre’ en arabe dialectal (EA ou TA), ne sont pas conformes à leurs homologues en MSA où ils possèdent les formes imperfectives

/ya-biit(-u)/ et /ya-biin(-u)/ respectivement. Leurs formes respectives dans l'arabe dialectal sont /yi-baat/ et /yi-baan/. Il existe également une autre exception, c'est le verbe 'venir' qui a la forme /gaa?(-a), ya-gii?(-u)/ en MSA cependant, il possède la forme /ga ou gih, yiigi/ en EA et /ja ou jaa, yji/ en TA.

6.1.1.4.3. Les verbes défectueux

Un verbe défectueux en arabe est un verbe qui possède une glide ou une voyelle longue dans la troisième lettre de son radical. Les formes perfectives de verbes défectueux dans l'arabe MSA, EA et TA sont données dans les tableaux suivants :

No	Forme	Exemple	Traduction
I_a	Fa3aa	bakaa	pleurer
I_b	Fa3iy(-a)	nasiy(-a)	oublier
II	Fa33aa	ghannaa	chanter
III	Faa3aa	naadaa	appeler quelqu'un
IV	?aF3aa	?aghraa	séduire
V	taFa33aa	tasammaa	être nommé
VI	taFaa3aa	tafaadaa	éviter
VII	(?i)nFa3aa	(?i)mbaraa	être affûté
VIII	(?i)Fta3aa	(?i)štaraa	acheter
X	(?i)staF3aa	(?i)stagnaa	se passer de

Tableau 6. 19. Les formes perfectives de verbes défectueux en MSA

No	Forme	Exemple	Traduction
I	$F\binom{a}{i}3\binom{a}{i}$	bakaa	pleurer
		nisi	oublier
II	Fa33a	Ganna	chanter
III	Faa3a	Naada	appeler quelqu'un
IV	?aF3a	?agra	séduire
V	(?i)tFa33a	(?i)tsamma	être nommé
VI	(?i)tFaa3a	(?i)tfaada	éviter
VII	(?i)nFa3a	(?i)mbara	être affûté
VIII	(?i)Fta3a	(?i)štara	acheter
X	(?i)staF3a	(?i)stagna	se passer de

Tableau 6. 20. Les formes perfectives de verbes défectueux en EA

No	Forme	Exemple	Traduction
I	F3 *a*	Bkaa	pleurer
II	Fa33a	ghanna	chanter
III	Faa3a	'aana	souffrir
IV	ϕ	ϕ	ϕ
V	(?i)tFa33a	(?i)t'ašša	dîner

VI	(?i)tFaa3a	(?i)tdaawa	se soigner
VII	ϕ	ϕ	ϕ
VIII	(?i)Ft3a	(?i)štha	avoir envie
X	(?i)staF3a	(?i)stagna	se passer de

Tableau 6. 21. Les formes perfectives de verbes défectueux en TA

Nous pouvons déduire de l'analyse tableaux (6.19), (6.20) et (6.21) plusieurs enseignements :

- La glide /y/ en position finale dans la racine d'un verbe en MSA est supprimée en EA et TA. Nous remarquons que dans le tableau (6.19) il existe en MSA deux variantes de la forme I, I_a et I_b, ce qui n'est pas le cas du EA et TA où il existe seulement une seule forme I. Cette différence est due au fait que la voyelle /i/ qui précède la glide /y/ a été remplacée par /a/ et la glide a été substituée par la hamza. Plus précisément, le changement de la forme [Fa3iy(-a)] en MSA à la forme [F3a] en TA est réalisée en effectuant les opérations suivantes :
  a) la forme [Fa3iy(-a)] est changée en [Fa3ia] par élision
  b) le changement morphologique de la forme [Fa3ia] puisqu'il implique le remplacement de /i/ par /a/ pour la voyelle de la racine et on obtient [Fa3aa]
  c) la forme devient [fa3a] par raccourcissement de la voyelle finale de la racine
  d) la suppression de la voyelle /a/ de la première consonne de la racine et nous obtenons [F3a].
- La comparaison montre aussi que le /aa/ en position finale d'un mot en MSA est régulièrement raccourci lors du passage en EA ou AT.
- Les tableaux montrent aussi que la forme IX des verbes défectueux n'existe pas dans l'arabe MSA, EA et TA, à l'exception d'un seul verbe dans l'EA qui est : /(?i)hlaww/ 'devenir beau. Cependant nous notons que dans l'EA cette forme est en train d'émerger.
- Nous remarquons aussi que dans le TA, tableau (6.21), les deux formes IV et VII sont disparues. Ceci est valable pour tous les dialectes maghrébins. Il ressort aussi de ce tableau que la forme VIII a subi une modification sur la racine où nous remarquons la suppression de la voyelle /a/ du préfixe /(?i)Fta/.

Par ailleurs, il existe deux modèles pour la forme primaire des verbes défectueux en MSA : la forme [Fa3aa] dont la fin du radical à l'origine pourrait être /w/ ou /y/ et la forme [Fa3iy(-a)]. En AE, et TA la première forme devient [Fa3a] et [F3a] respectivement, et cela par raccourcissement de la voyelle finale dans les verbes ayant les glides /y/ ou /w/. Il est aussi à noter que la forme [Fi3i] dans l'EA est réservée pour les verbes avec /y/ seulement. Cette forme n'a pas de correspondant dans le dialecte tunisien et maghrébin en général. La seconde forme est assimilée à la première forme dans le TA, alors qu'est réduite dans l'EA à [Fa3a] dans certains verbes, et à [Fi3i] dans d'autres, comme le montrent les exemples suivants :

Racine	**MSA**	**EA**	**TA**	**Traduction**
d-3-w	da3aa	da3a	d3a	Inviter
r-m-y	ramaa	rama	Rma	Jeter
m-š-y	mašaa	miši	Mša	Marcher
l-q-y	laqiy(-a)	la?a	Lqa	Trouver
r-D-w	raDiy(-a)	riDi	rDa	se contenter

Les deux premiers exemples montrent que la forme [Fa3aG(-a)] devient [Fa3aa] en MSA par élision de la glide. Cette forme passe à [Fa3a] par raccourcissement de la voyelle finale en EA et TA. Dans le TA la suppression de la voyelle /a/ de la première racine du radical est effectuée en plus du raccourcissement de la voyelle finale. Le changement de la forme [Fa3aa] en MSA, issue de la forme [Fa3ay(-a)], en [Fi3i] en EA est un changement morphologique puisqu'il inclut le changement du vocalisme de /a/ en /i/. Par exemple : /mašay(-a) → mašaa > mišiy-Ø → miši/ 'marcher'. Le changement de la forme [Fa3iy(-a)] en MSA en la forme [Fa3a] en EA est aussi de nature morphologique puisqu'il implique le remplacement de la voyelle /i/ de la racine par la voyelle /a/, c'est le cas du verbe /laqiy(-a) > la?a/ 'trouver'.

Les verbes en EA ayant le modèle perfectif [Fi3i] sont à base de la forme [Fi3iy] avec un changement de la première voyelle du modèle MSA correspondant. Ils deviennent [Fi3ii] par application de l'assimilation vocoïde persévérative puis [Fi3i] par raccourcissement de la voyelle finale, comme dans le verbe : /raDiy(-a) > riDi/ 'être content'. Néanmoins, le changement de la première voyelle de /a/ en /i/ en EA reste aussi un changement morphologique.

Du point de vue phonologique, l'EA et le TA diffèrent du MSA dans le cas de la suffixation pronominale des verbes défectueux. Bien que les séquences /aw/ et /ay/ sont employées avant un pronom commençant par une consonne en MSA, elles sont remplacées par la voyelle longue /ee/ en EA et par /î/ en TA. Ces voyelles sont créées à partir de /ay/ par monophtongaison et est ensuite fractionnée en tant qu'affixe indépendant à appliquer dans d'autres contextes, tels que la suffixation consonantique des verbes géminés et les verbes avec arrêt glottal final. Dans le tableau suivant nous donnons quelques exemples de ces formes :

Forme	**MSA**	**EA**	**Traduction**
I	da3aw-tu	da3ee-t	J'ai invité
II	Sallay-naa	Sallee-na	Nous prions
III	naaday-ti	nadee-ti	Tu as appelé
IV	?agray-tum	?agree-tu	Vous avez séduit
V	tahadday-ta	(?i)thaddee-t	Tu as défié

Dans le premier exemple concernant la forme I, le /w/ est changé en /y/ avant que l'application de la monophtongaison. Cependant, si le verbe défectueux se termine par un /i/ dans l'arabe dialectal, ce dernier est simplement rallongé (c.à.d. non raccourci) avant tous les suffixes, à l'exception du marqueur de la troisième personne du féminin singulier comme le montre les exemples suivants :

MSA	**EA**	**Traduction**
raDii-tu	riDii-t	Je suis satisfait
saxii-ta	sixii-t	Tu es devenu généreux
qawii-naa	?iwii-na	Nous sommes devenus fort

Dans le cas où il existe une suffixation avec le marqueur de la troisième personne du féminin singulier [-it], la forme de base en EA serait [Fi3iy-it], puis le deuxième /i/ serait syncopé[16] en une syllabe ouverte, comme dans le verbe /maša-t > mišy-it/ 'elle a marché'.

[16] Syncope : phénomène linguistique qui fait disparaître un ou plusieurs phonèmes au sein d'un même mot

Notons que la règle d'assimilation vocoïde anticipatoire ne s'applique pas ici sur la forme dialectale, ce qui veut dire que la syncope devrait être classée après cette règle. Si nous appliquons la syncope en premier lieu, le résultat serait sur l'exemple précédent /mišiy-it → mišy-it → *mišiit/. Les formes imperfectives des verbes défectueux dans le MSA et l'EA sont données dans les tableaux (6.22), (6.23) et (6.24).

No	Forme	Exemple	Traduction
I	ya-F3 $\begin{pmatrix} uu \\ ii \\ aa \end{pmatrix}$	ya-rjuu	implorer
		ya-rmii	jeter
		ya-nsaa	oublier
II	yu-Fa33ii	yu-ghannii	chanter
III	yu-Faa3ii	yu-naadii	appeler quelqu'un
IV	yu-F3ii	yu-grii	séduire
V	ya-taFa33aa	ya-tasammaa	s'appeler
VI	ya-taFaa3aa	ya-tafaadaa	éviter
VII	ya-nFa3ii	ya-mbarii	être affûté
VIII	ya-Fta3ii	ya-štarii	acheter
X	ya-staF3ii	ya-stagnii	se passer de

Tableau 6. 22. Les formes imperfectives de verbes défectueux en MSA

No	Forme	Exemple	Traduction
I	yi-F3$\begin{pmatrix} u \\ i \\ a \end{pmatrix}$	yi-rgu	implorer
		yi-rmi	jeter
		yi-nsa	oublier
II	yi-Fa33i	yi- ghanni	chanter
III	yi-Faa3i	yi-naadi	appeler quelqu'un
IV	yi-F3i	yi-gri	séduire
V	yi-tFa33a	yi-tsamma	s'appeler
VI	yi-tFaa3a	yi-tFaada	éviter
VII	yi-nFi3i	yi-mbiri	être affûté
VIII	yi-Fti3i	yi-štiri	acheter
IX	yi-F3aww	yi-hlaww	devenir beau
X	yi-staF3i	yi-stagni	se passer de

Tableau 6. 23. Les formes imperfectives de verbes défectueux en EA

No	Forme	Exemple	Traduction
I	yi-F3$\begin{pmatrix} i \\ a \end{pmatrix}$	yi-rmi	jeter
		yi-nsa	oublier
II	yi-Fa33i	yi- ghanni	chanter
III	yi-Faa3i	yi-'aani	souffrir
IV	ϕ	ϕ	ϕ
V	yi-tFa33a	yi-t'ašša	dîner
VI	yi-tFaa3a	yi- tdaawa	se soigner
VII	ϕ	ϕ	ϕ
VIII	yi-Ft3a	yi-štha	avoir envie

X	yi-staF3a	yi-stagna	se passer de

Tableau 6. 24. Les formes imperfectives de verbes défectueux en TA

Le tableau (6.22) montre bien que toutes les voyelles longues finales dans les verbes défectueux découlent de l'application de la règle de l'élision de la glide. De la même façon, les tableaux (6.23) et (6.24) exhibent l'application du raccourcissement de la voyelle finale ultérieur en EA et TA respectivement. Ce raccourcissement donne aux voyelles finales brèves les formes montrées dans ces tableaux.

L'analyse approfondie des formes perfectives, données dans les tableaux (6.19), (6.20 et (6.21); et des formes imperfectives, données dans les tableaux (6.22), (6.23) et (6.24) montre que les verbes défectueux ont deux consonnes dans leur structure de surface même s'ils possèdent trois consonnes dans leurs structures de base. Dans la grammaire arabe, nous pouvons déterminer l'identité et l'origine d'un glide apparaissant dans la troisième consonne à partir de la forme imperfective, car la voyelle longue en MSA (qui est raccourcie en EA et TA) est réalisée sous la forme d'une séquence de voyelles brèves, similaire à la voyelle longue de la forme imperfective, et d'une glide apparentée à cette voyelle. En d'autres termes, la séquence /uu/ devient /uw/, la séquence /ii/ devient /iy/ et la séquence /aa/ devient /VG/ où le /G/ est une glide non spécifiée selon (Mahadine, 1982).

Enfin, le tableau (6.24) montre qu'il y a une différence remarquable entre le MSA et le TA concernant les deux formes VIII, X. A ce stade, la voyelle longue /a/ reste /a/ et elle ne change pas en /i/ comme montré pour le MSA dans le tableau (6.22).

Nous rappelons que dans la tradition de la grammaire arabe, les verbes défectueux disposent de trois modèles pour les formes primaires de l'imperfectif en MSA à savoir :

a) *Le modèle [ya-F3uu] :* pour les verbes avec /aa/ dans la forme perfective dont le radical de base se termine par la voyelle longue /w/. Par exemple le verbe da'aa – ya-d'uu.
b) *Le modèle [ya- F3ii] :* pour les verbes ayant un /aa/ dans la forme perfective dont le radical de base se termine par la voyelle longue /y/, comme c'est le cas du verbe bak**aa** – ya-bk**ii**.
c) *Le modèle [ya-F3aa] :* pour les verbes dont la forme au perfectif est [Fa3iy(-a)], par exemple le verbe nasiya – ya-nsaa.

Le premier modèle (a), possède deux homologues morphologiques en EA et TA. Il s'agit des modèles *[yi-F3i] et [yi-F3a]*. En revanche le deuxième modèle (b) n'en possède qu'un seul équivalent qui est le patron *[yi-F3i]*. Il est obtenu par le raccourcissement de la voyelle finale. Le troisième et dernier modèle (c) possède un seul équivalent représenté par [yi-F3a]. Cet équivalent est aussi obtenu par le raccourcissement de la voyelle finale. Ceci met en avant une différence majeure entre le MSA et les dialectales EA et TA. Cette différence réside dans le fait que la voyelle longue /uu/ apparait en dernière position de la forme imperfective des verbes en MSA, mais elle est remplacée par /i/ ou /a/ lors du passage en EA et AT. Cette caractéristique est illustrée dans les exemples suivants :

Racine	MSA	EA/ TA	Traduction
d-3-w	ya-d3uu	yi-d3i	inviter
3-l-w	ya-3luu	yi-3la	s'élever
S-f-w	ya-Sfuu	yi-Sfa	se purifier
r-m-y	ya-rmii	yi-rmi	jeter
m-š-y	ya-mšii	yi-mši	marcher
l-q-y	ya-lqaa	yi-l?a/ yi-lqa	trouver
r-D-w	ya-rDaa	yi-rDa	être satisfait

Les trois premiers exemples montrent que l'EA innove au niveau de certains verbes ayant /w/ en position finale de leurs racines, en leur permettant de posséder des imperfectifs avec /a/ tandis que d'autres ont des imperfectifs en /i/. De ce fait, le développement suivant se produit en EA :

a. d-3-w: yi-d3iw → yi-d3iy → yi-d3i
b. 3-l-w: yi-3law → yi-3laa→ yi-3la

Ceci est cohérent avec le constat d'avoir des contrastes nouveaux dans le second vocalisme en EA tandis qu'il n'en existe aucun en MSA.

6.1.1.4.4. Les verbes doublement faibles

Il existe aussi dans la grammaire arabe, des verbes faibles qui contiennent deux glides. Ces verbes sont nommés *"verbes doublement faibles"(lafife).* Ce sont en général des verbes faibles ayant une autre glide supplémentaire au début ou au milieu du radical. La plupart de ces verbes sont des verbes dont la seconde et la troisième radicale de la racine sont [w] ou [y]. Ils peuvent être classés en deux groupes :

a) *Lafife makroune :* ce sont des verbes ayant deux glides successives au niveau de la deuxième et troisième lettre de leur racine. Par exemple le verbe rawa : raconter.
b) *Lafifie mafrouke :* ce sont des verbes dont la première et la troisième lettre de leur racine sont des glides. C'est le cas des verbes : wakaa : protéger et wa3a : être conscient.

Ces verbes possèdent les mêmes caractéristiques que les verbes assimilés et/ou des verbes défectueux. En ce qui concerne les transitions du MSA vers l'arabe dialectal, les verbes doublement faibles illustrent les mêmes changements que ceux qui sont attendus de la combinaison des comportements montrés pour les verbes simplement faibles. Par exemple, le verbe /wašaa > waša/ 'moucharder' est dérivé de [wašay(-a)] par élision de la glide dans les deux variétés, puis par raccourcissement de la voyelle finale en AE, comme présenté dans la règle (a) ci-après. La forme imperfective pour ces verbes est /ya-šii/ obtenue par occultation du /w/ de la forme /ya-wšii/ utilisée dans le MSA, comme illustré dans (b). Cependant, le radical initial ne subit aucun changement dans l'EA /yi-wši/, comme donné dans la règle (c) :

a. wašay-a → (MSA) wašaa → (EA) waša
b. ya-wšiy-u → ya-wšiu → ya-wšii → (MSA) ya-šii
c. yi-wšiy → yi-wšii → (EA) yi-wši

D'autres exemples de verbes doublement faibles sont donnés dans le tableau suivant :

Racine	MSA	EA	Traduction
w-f-y	wafaa	wafa/waffa	Tenir ses promesses
n-w-y	nawaa	nawa	Avoir l'intention

q-w-y	qawiy(-a)	?iwi to	Devenir fort
h-y-y	hayiy(-a)	hiyi	Survivre

Par ailleurs, les formes imperfectives des verbes donnés dans le tableau précédent sont illustrées comme suit :

Racine	MSA	EA	Traduction
w-f-y	ya-Fii	yuufi	remplir
n-w-y	ya-nwii	yi-nwi	avoir l'intention
q-w-y	ya-qwaa	yi-?wa	devenir fort
h-y-y	ya-hyaa	yi-hya	survive

6.1.2. Les Verbes quadrilitères

Les verbes quadrilitères sont des verbes ayant quatre lettres dans leur racine radicale. Ils sont répartis en deux grandes classes de verbes en arabe : les verbes sains et les verbes dupliqués. Les verbes sains sont formés de quatre lettres différentes tandis que les verbes dupliqués possèdent deux lettres dupliquées (redoublée). En ce qui concerne l'origine de ces verbes, (Fleisher, 1956) avance qu'ils sont formés par 'la répétition d'un élément bilitère' qui consiste à répéter les deux premières consonnes d'une racine trilitère.

Dans le système arabe classique, les quadrilitères possèdent quatre formes. Certaines de ces formes ne sont plus utilisées aujourd'hui même dans le MSA, c'est le cas de la forme III donnée par [(?i)F3anL1aL2(-a)] qui existe dans l'Arabe Classique, comme pour le verbe /(?i)slanTah(-a)/ 'être mis à plat', elle n'est plus utilisée ni en MSA ni dans l'arabe dialectal. De ce fait, le nombre possible des formes des verbes quadrilitères sains est limité à trois, que ce soit en MSA ou en AE, dont la première est la forme primaire et les deux autres sont les formes dérivées. Du point de vue sémantique :

- La forme I est non-stative qu'elle soit transitive ou intransitive (comme la Forme I trilitère)
- La Forme II est passive ou intransitive de la Forme I (comme la Forme V trilitère)
- La Forme IV consiste à développer un état (comme la Forme IX trilitère).

En TA et au Maghreb en général, le verbe quadrilitère possède deux formes : i) la forme I qui correspond à la forme II du verbe régulier et ii) la forme II qui correspond à la forme V du verbe régulier.

Les tableaux (6.25), (6.26) et (6.27) donnés ci-après, présentent ces différentes formes pour les verbes quadrilitères sains, en considérant que L_1 et L_2 sont deux consonnes différentes.

No	Forme	Exemple	Traduction
I	Fa3L1aL2(-a)	zaxraf(-a)	décorer
II	taFa3L1aL2(-a)	tašayTan(-a)	se diaboliser
IV	(?i)F3aL1aL2L2(-a)	(?i)Tma?ann(-a)	se rassurer

Tableau 6. 25. Les formes perfectives des verbes quadrilitères sains en MSA

No	Forme	Exemple	Traduction
I	Fa3L1$\binom{a}{i}$L2	Zaxraf	décorer
		targim	traduire
II	(?i)tFa3L1 $\binom{a}{i}$L2	(?i)tšayTan	se diaboliser
		(?i)tša3lil	s'emporter
IV	(?i)F3aL1aL2L2	(?i)Tma?ann(-a)	se rassurer

Tableau 6. 26. Les formes perfectives des verbes quadrilitères sains en EA

No	Forme	Exemple	Traduction
I	Fa3L1$\binom{a}{i}$L2	garba'	produire un son
		xarbich	griffonner
II	(?i)tFa3L1 $\binom{a}{i}$L2	(?i)tfarhad	se défouler
		(?i)tkarbis	se tourner

Tableau 6. 27. Les formes perfectives des verbes quadrilitères sains en TA

L'analyse comparative des trois tableaux, (6.25), (6.26) et (6.27), nous a conduits aux remarques suivantes :

a) Dans l'arabe MSA, les voyelles finales à la fin des verbes sont considérées comme des marques d'inflexion pour la troisième personne du masculin singulier. Cependant, ces voyelles sont négligées dans l'arabe dialectal. Ainsi, le (-a) final de la forme perfective est omis lors du passage en arabe dialectal.
b) Le /i/ pré-final de la forme dialectale [Fa3L$_1$aL$_2$ ~ Fa3L$_1$iL$_2$] est considéré en tant que variante de /a/, tandis que son équivalent standard possède seulement le /a/ comme voyelle pré-finale.
c) La seconde forme [(?i)tFa3L$_1$aL$_2$] en arabe dialectal représente le reflet de la forme [taFa3L$_1$aL$_2$(-a)] en MSA. Ceci est le résultat de la transformation du préfixe [ta-] du MSA en [t-] dans les versions dialectales EA et TA avec l'ajout de l'épenthétique (?i) au début du verbe.
d) La forme IV n'existe pas dans le dialecte tunisien et dans les dialectes du Maghreb en général.

Le tableau (6.28) montre les formes imperfectives des verbes quadrilitères en MSA, et les tableaux (6.29) et (6.30) donnent leurs équivalents en EA, TA respectivement :

No	Form	Exemple	Traduction
I	yu-Fa3L1iL2(-u)	yu-zaxrif(-u)	décorer
II	ya-taFa3L1aL2(-u)	ya-tašayTan(-u)	se diaboliser
IV	ya-F3aL1iL2L2(-u)	ya-Tma?inn(-u)	se rassurer

Tableau 6. 28. Les formes imperfectives des verbes quadrilitères sains en MSA

No	Forme	Exemple	Traduction
I	yi-Fa3L1$\binom{a}{i}$L2	yi-zaxraf	décorer
		yi-targim	traduire

II	yi-tFa3L1 $\binom{a}{i}$L2	yi-tšayTan	se diaboliser
		yi-tša3lil	s'emporter
IV	yi-F3aL1iL2L2	yi-Tma?inn(-a)	se rassurer

Tableau 6. 29. Les formes imperfectives des verbes quadrilitères sains en EA

No	**Forme**	**Exemple**	**Traduction**
I	yi-Fa3L1$\binom{a}{i}$L2	yi-garba'	produire un son
		yi-xarbich	griffonner
II	yi-tFa3L1 $\binom{a}{i}$L2	yi-tfarhad	se défouler
		yi-tkarbis	se tourner

Tableau 6. 30. Les formes imperfectives des verbes quadrilitères sains en AT

Nous notons que les verbes quadrilitères dupliqués comprennent une ou deux consonnes dupliquées. Ils possèdent deux formes en MSA et autant en EA et TA, comme le montrent les tableaux (6.31), (6.32) et (6.33) respectivement. Seulement, nous remarquons que les deux formes possèdent différentes variantes dans chaque type d'arabe considéré (MSA, EA et TA).

No	**Form**	**Example**	**Traduction**
I_a	Fa3Fa3(-a)	dagdag(-a)	chatouiller
I_b	Fa3FaL(-a)	samsar(-a)	agir en tant qu'intermédiaire
II	taFa3Fa3(-a)	tazalzal(-a)	être secoué

Tableau 6. 31. Les formes perfectives des verbes quadrilitères dupliqués en MSA

No	**Form**	**Example**	**Traduction**
I_a	Fa3F$\binom{a}{i}$3	SahSah	être alerté
		Zalzil	secouer
I_b	Fa3F$\binom{a}{i}$L	Samsar	agir en tant qu'intermédiaire
		Dardiš	discuter
I_c	Fa3L1$\binom{a}{i}$L1	za?TaT	être de bonne humeur
		3aknin	déranger
II	(?i)tFa3Fi3	(?i)tzalzil	être secoué

Tableau 6. 32. Les formes perfectives des verbes quadrilitères dupliqués en EA

No	**Form**	**Example**	**Traduction**
I_a	Fa3F$\binom{a}{i}$3	SahSah	être alerté
		Zalzil	secouer
I_b	Fa3F$\binom{a}{i}$L	Samsar	agir en tant qu'intermédiaire
		Farkiš	saboter
I_c	Fa3L1$\binom{a}{i}$L1		
		daskir	Faire le DJ
I_d	Fa3Fi3	Karkib	tourner
II_a	(?i)tFa3Fi3	(?i)tzalzil	être secoué

II_b	(?i)tFa3FaL	(?i)tmahmas	

Tableau 6. 33. Les formes perfectives des verbes quadrilitères dupliqués en TA

L'analyse des tableaux (6.31), (6.32) et (6.33) montre que les formes quadrilitères ayant une duplication de la troisième lettre existent seulement en AE. De la même manière, le dialecte tunisien présente une autre variante de la forme II qui n'existe pas dans l'arabe MSA et EA. Cette forme concerne la forme quadrilitères ayant une duplication de la première lettre. Les formes imperfectives des verbes quadrilitères en MSA, EA et TA sont données dans les tableaux (6.34), 6. (35) et (6.36), respectivement, donnés comme suit :

No	**Forme**	**Exemple**	**Traduction**
I_a	yu-Fa3Fi3(-u)	yu-dagdig(-u)	chatouiller
I_b	yu-Fa3FiL(-u)	yu-samsir(-u)	agir comme un intermédiaire
II	ya-taFa3Fa3(-u)	ya-tazalzal(-u)	être secoué

Tableau 6. 34. Les formes imperfectives des verbes quadrilitères doublés en MSA

No	**Forme**	**Exemple**	**Traduction**
I_a	yi-Fa3Fi3	yi-zalzil	secouer
I_b	yi-Fa3F$\binom{a}{i}$L	yi-samsar	agir comme un intermédiaire
		yi-dardiš	discuter
I_c	yi-Fa3L1$\binom{a}{i}$L1	yi-za?TaT	être de bonne humeur
		yi-3aknin	déranger
II	yi-tFa3Fi3	yi-tzalzil	être secoué

Tableau 6. 35. Les formes imperfectives des verbes quadrilitères doublés en EA

No	**Forme**	**Exemple**	**Traduction**
I_a	y-Fa3Fi3	y-zalzil	Secouer
I_b	y-Fa3F$\binom{a}{i}$L	y-Samsar	agir en tant qu'intermédiaire
		y-Farkiš	Saboter
I_c	y-Fa3L1$\binom{a}{i}$L1		
		y-daskir	Faire le DJ
I_d	y-Fa3Fi3	y-karkib	Tourner
II_a	yi-tFa3Fi3	yi-tzalzil	être secoué
II_b	(?i)tFa3FaL	(?i)tmahmas	

Tableau 6. 36. Les formes imperfectives des verbes quadrilitères doublés en TA

6.2. L'aspect et le mode de la flexion:

L'arabe standard et l'arabe dialectal expriment deux aspects du verbe en employant des dispositifs morphologiques : le perfectif et l'imperfectif. Le premier est utilisé pour l'action finie ou terminée. Le dernier désigne une action inachevée, y compris l'action habituelle, en cours ou future. Certains linguistes, comme (Eisele, 1990) affirment que ces formes correspondent à une distinction entre le passé et le non-passé, alors que d'autres disent

qu'il n'y a pas de correspondance un-à-un entre l'aspect et le temps. C'est ce que soutient (Radwan, 1975), qui affirme que : « *l'aspect et le temps doivent être traitées comme deux catégories indépendantes… Les deux termes sont utilisés pour désigner deux caractéristiques différentes de modèles verbaux. Le terme 'Aspect' couvre les plages sémantiques d'achèvement contre le non-achèvement et de la continuation contre la non-continuation, tandis que 'le temps' couvre la référence temporelle».*

Morphologiquement parlant, la forme perfective est obtenue par l'ajout de suffixes seulement, alors que la forme imperfective est obtenue par l'addition d'affixes : il s'agit d'un ensemble de combinaisons de préfixes et de suffixes. De plus de leur inflexion pour l'aspect, les verbes arabes sont fléchis aussi par le mode. En effet, les verbes en MSA sont conjugués par un seul mode dans l'aspect perfectif - l'indicatif - et par trois modes dans l'imperfectif: l'indicatif, le subjonctif et le jussif; en plus de l'impératif. La classification croisée (Gadalla, 2000) des verbes en MSA par l'aspect et le mode peut être effectuée comme suit :

	Indicatif	**Subjonctif**	**Jussif**
Perfectif	+	–	–
Imperfectif	+	+	+

Tableau 6. 37. La classification croisée des verbes en MSA pour l'aspect et le mode.

L'arabe dialectal pour sa part ne considère qu'un seul mode, il s'agit de l'indicatif que ce soit pour l'aspect perfectif ou imperfectif. Il en est de même pour l'impératif. Ce constat reste valable à la fois pour les dialectes du Maghreb et ceux du Machrek.

La conjugaison par l'aspect et le mode implique la fixation de marqueurs d'accord à la racine du verbe. Ces marqueurs sont des suffixes et des préfixes permettant de matérialiser l'accord du verbe avec le sujet en personne, genre et nombre. En d'autres termes, nous pouvons dire que la flexion verbale de la langue arabe est basée sur la concaténation de préfixes et suffixes (affixes) aux lemmes verbaux. Le tableau (6.38) illustre les marqueurs du perfectif de l'indicatif utilisés avec le verbe /šakar(-a)/ 'remercier' dans le MSA et les dialectes égyptien (EA) et tunisien (TA) :

Referent	**MSA**	**EA**	**TA**
1sg (m & f)	šakar-tu	šakar-t	škar-t
1pl (m & f)	šakar-naa	šakar-na	škar-na
2msg	šakar-ta	šakar-t	škar-t
2fsg	šakar-ti	šakar-ti	škar-ti
2d (m & f)	šakar-tumaa	šakar-tu	škar-tu
2mpl	šakar-tum		
2fpl	šakar-tunna		
3msg	šakar-a	šakar-Ø	škar-Ø
3fsg	šakar-at	šakar-it	šakr-it
3md	šakar-aa	šakar-u	škar-u
3fd	šakar-ataa		
3mpl	šakar-uu		
3fpl	šakar-na		

Tableau 6. 38. Les marqueurs du perfectif de l'indicatif en MSA, EA et TA

Le tableau (6.38) montre que le duel (masculin ou féminin) est supprimé dans les différentes variétés dialectales (EA et TA), il est remplacé, voir englouti, par le pluriel masculin. Il est de même pour le pluriel féminin qui est substitué par le pluriel masculin. Le tableau illustre aussi une différence morpho-phonémique entre le MSA et les dialectes considérés. Cette différence concerne la règle de suppression de la haute voyelle, appliquée dans les dialectes, qui sont à l'origine de la suppression du /i/ du radical médian lorsque le suffixe perfectif commence par une voyelle. Notons que le dialecte tunisien ajoute à cette règle une opération de permutation entre le 1^{ier} et le 2ème radical afin d'obtenir la même forme que celle du dialecte égyptien. Cette modification est justifiée pour des raisons phonétiques. Ce type d'élision de voyelle est commun dans la forme du verbe [Fi3iL] en EA qui est l'équivalent de deux formes en MSA [Fa3iL(-a)] et [Fa3ul(-a)], comme le montre les exemples suivants :

MSA	**EA**	**Traduction**
šarib-at	širb-it	Elle a bu
šarib-uu	širb-u	Ils ont bu
kabur-at	kibr-it	Elle a grandi
kabur-uu	kibr-u	Ils ont grandi

De la même manière, ce type d'élision de voyelle est commun dans la forme du verbe [F3iL], [F3uL] en TA qui est l'équivalent de deux formes en MSA [Fa3iL(-a)] et [Fa3ul(-a)]. Le même phénomène est aussi présent lorsque le verbe est attaché à un pronom affixe. De ce fait, nous obtenons une modification du schème verbal lorsque le verbe à la 3ème personne masculin au singulier est attaché aux pronoms (u/ah/ak) de la 3ème personne masculin du singulier. Par exemple, pour le verbe Darab+ak 'il t'a frappé' en MSA nous obtenons dans le TA : /Dr**a**b/+/ak/ → /D**a**rbak/.

La même fonction morpho-phonémique implique la suppression de la voyelle haute dans le croisement du pronom /?ana/ 'Je' et d'un verbe de la forme [Fi3iL] dans l'AE. Par exemple, en MSA nous avons l'expression /?anaa šarib-tu/ 'j'ai bu', son équivalent en EA est : /?ana šrib-t/ avec la suppression de la voyelle initiale/i/. Selon (Gadalla, 2000), la raison de cette élision est que la voyelle se trouve dans une syllabe médiane ouverte.

Dans la tradition de la grammaire arabe, la forme imperfective du verbe est obtenue en ajoutant l'un des quatre préfixes suivant : [?a-, na-, ya- ou ta-] à la racine imperfective du verbe conformément au référent. Par exemple, la racine imperfective de la forme I, possède le modèle [-CCVC-]. La variété dialectale pour sa part, a innové dans la formation de l'imperfectif. Les différences entre l'arabe MSA et l'arabe dialectal au niveau de la construction de la forme imperfective, peuvent être remarquées dans les exemples donnés dans le tableau suivant :

Référent	MSA			EA	TA
	Indicatif	Subjunctif	Jussif	Indicatif	Indicatif
1sg (m & f)	?a-škur-u	?a-škur-a	?a-škur-Ø	(?)a-škur-Ø	nu-škur-Ø
1pl (m & f)	na-škur-u	na-škur-a	na-škur-Ø	nu-škur-Ø	nu-škr-u
2msg	ta-škur-u	ta-škur-a	ta-škur-Ø	tu-škur-Ø	tu-škur-Ø
2fsg	ta-škur-iina	ta-škur-ii	ta-škur-ii	tu-škur-i	tu-škr-i
2d (m & f)	ta-škur-aani	ta-škur-aa	ta-škur-aa	tu-škur-u	tu-škr-u
2mpl	ta-škur-uuna	ta-škur-uu	ta-škur-uu		
2fpl	ta-škur-na	ta-škur-na	ta-škur-na		
3msg	ya-škur-u	ya-škur-a	ya-škur-Ø	yu-škur-Ø	yu-škur-Ø
3fsg	ta-škur-u	ta-škur-a	ta-škur-Ø	tu-škur-Ø	tu-škur-Ø
3md	ya-škur-aani	ya-škur-aa	ya-škur-aa	yu-škr-u	yu-škr-u
3fd	ta-škur-aani	ta-škur-aa	ta-škur-aa		
3mpl	ya-škur-uuna	ya-škur-uu	ya-škur-uu		
3fpl	ya-škur-na	ya-škur-na	ya-škur-na		

Tableau 6. 39. Les marqueurs de l'imperfectif en MSA, EA et TA

Nous notons également, que la forme imperfective du dialecte tunisien TA diffère de celle du MSA, d'un point de vue morpho-phonémique, au niveau de l'application de la règle de suppression de la haute voyelle qui supprime le /u/ du radical médian lorsque le suffixe commence par une voyelle. Ce type d'élision de voyelle est appliqué dans le cas de verbes ayant les formes [F3iL], [F3aL] et [F3uL] en supprimant à chaque fois les voyelles /i/, /a/ et /u/ respectivement. Il est important de mentionner que dans la même optique, le dialecte algérien dispose des mêmes formes pour certains parlers, mais possède en plus d'autres formes concernant le pluriel, par exemple nous avons pour le verbe škar (remercier) les formes suivantes : nu-ššukr-u (nous remercions), yu-ššukru (ils remercient), tu-ššukr-u (vous remerciez). Ces nouvelles formes sont obtenues par la gémination de la première radicale et le déplacement de la voyelle de la 2ème radicale vers le 1ier radical.

Comme pour la forme perfective, la comparaison entre les marqueurs imperfectifs des variantes de l'arabe EA, TA et MSA montre que les marqueurs du duel (féminin et masculin) ont disparu lors du passage en EA et TA. Ces marqueurs disparus sont remplacés par ceux du masculin pluriel. D'autant plus, le suffixe indicatif [-u] en MSA disparait lors du passage en arabe dialectal. Nous signalons également que les suffixes dissyllabiques indicatifs ont perdu la seconde syllabe [-na] et se sont vu écourtés leur longues voyelles par la règle de raccourcissement de la voyelle finale, autrement dit les syllabes [-iina] et [-uuna] ont été réduites en [-i] et [-u] respectivement.

Par ailleurs, l'EA a conservé les préfixes imperfectifs du MSA, cependant il remplace les voyelles par /i/ à l'exception du préfixe [(?)a-] où la voyelle est conservée. Le dialecte tunisien conserve aussi les préfixes imperfectifs du MSA, à l'exception du préfixe de la première personne au singulier qui est remplacé par le préfixe [n-]. De manière générale, les dialectes du Maghreb sont caractérisés par l'utilisation de la particule « n » à la première

personne du singulier, à l'inaccompli, alors que cette particule est presque absente dans les dialectes du Machrek (l'égypte dans ce cas).

Cependant, dans quelques verbes, et particulièrement ceux ayant /u/ comme voyelle dans la racine, la voyelle imperfective est changée en /u/, à l'exception encore une fois de [(?)a-] pour l'EA et de [nu] pour l'AT, comme illustré dans le tableau suivant :

MSA	**EA**	**TA**	**Traduction**
ya-xruj(-u)	yu-xruj	Yu-xruj	il sortira
na-Tlub(-u)	nu-Tlub	nu-Tlub	nous demanderons
ta-xnuq(-u)	tu-xnu?	tu-xnug	elle étrangle
?a-ktub(-u)	?a-ktib	ni-ktib	j'écrirai

Lorsque le préfixe imperfectif en MSA contient un /u/, comme c'est le cas dans les formes II et III, perceptible dans les modèles de verbe [Fa33aL(-a), yu-Fa33iL(-u)] et [Faa3aL(-a), yu-Faa3iL(-u)], il est remplacé par un /i/ en AE. Dans le dialecte TA le préfixe imperfectif est substitué par un /i/ dans la forme [yi-Faa3iL] et par un soukoun dans la forme [y-Fa33il]. Les exemples suivants illustrent les effets de ces changements :

MSA	**EA**	**TA**	**Traduction**
yu-?axxir(-u)	yi-?axxar	yi-waxxir	il retarde
tu-?akkil(-u)	ti-?akkil	yi-ttaakil	elle se nourrit
nu-3allim(-u)	ni-3allim	n-3allmu	nous enseignons
yu-ðaakir(-u)	yi-zaakir		il étudie
tu-haarib(-u)	ti-haarib	t-haarib	tu te bats
nu-saafir(-u)	ni-saafir	n-saafru	Nous voyageons

Nous signalons aussi que les préfixes imperfectifs de tous les verbes, quel que soit l'arabe considéré, contiennent des voyelles courtes. Toutefois, certaines exceptions existent dans les dialectales et sont présentées par Abdel-Malek dans (Abdel-Malek, 1972) comme suit : « *Il existe seulement trois verbes dont la forme au présent contient de longues voyelles dans le préfixe : ces verbes sont /ga/ 'venir', /axad/ 'prendre' et /akal/ 'manger'. La forme /aagi/ 'je viens' possède /aa/ dans le préfixe; toutes les autres formes du présent du verbe /ga/ ont /ii/ dans le préfixe. Les formes du présent des verbes /axad/ et /akal/ ont /aa/ dans le préfixe si la racine commence avec une consonne ; si celle-ci commence par deux consonnes, la voyelle du préfixe serait /a/ ».*

Cependant nous relevons quelques remarques au sujet de ces exceptions : les verbes */axad/ et /akal/* sont considérés comme exceptionnels par le fait qu'ils ont un /a/ dans le préfixe plutôt qu'un /i/. Ainsi, nous pouvons proposer sur la voyelle longue /aa/ au début de la forme imperfective du verbe /?axad/ et /?akal/ dérivée du préfixe /a?/ par raccourcissement compensatoire, comme montré dans le processus (a). Le fait que la voyelle du préfixe soit /a/ dans certains contextes peut être justifié par l'application des règles de raccourcissement au résultat du raccourcissement compensatoire, comme dans l'enchaînement (b). Enfin, nous remarquons que le verbe /ga/ est une véritable exception ne nécessitant pas de règles spécifique, comme illustré pour les deux exceptions précédentes.

(a)	(b)
ya-?xud → yaaxud 'he takes'	ya-?xud-u → ya-?xd-u (par suppression de la haute voyelle) → yaaxd-u (par allongement compensatoire→ yaxd-u (par raccourcissement de la syllabe fermée) 'il le prend'
na-?kul → naakul 'we eat'	na-?kul-ha → naakul-ha (par allongement compensatoire) → nakul-ha (par raccourcissement atonique) 'Nous le mangeons (f)'

Rappelons que la forme perfective peut faire référence au présent ou au futur. Elle peut être utilisée avec l'adverbe de temps présent standard /?al?aan/ 'maintenant' ou avec son homologue dialectal /dilwa?ti/ (en EA), /tawwa/ (en TA), /dabba/ (dialecte marocain) ou /dork/ (dialecte algérien); pour produire un sens similaire au présent perfectif de l'anglais. Elle peut aussi renvoyer au futur lorsqu'elle est utilisée avec des phrases conditionnelles. De plus, la forme imperfective fait référence au présent habituel ou progressif. Néanmoins, l'EA diffère du MSA à cet égard. Le premier devance les préfixes imperfectifs par un préfixe additionnel [bi-] pour signaler les actions 'progressives' ou 'habituelles', comme dans l'exemple suivant :

Type de l'arabe	Phrase	Structure	Traduction
MSA	?al-walad-u ya-šrab-u 1-laban	le-Nom-garçon boir-imper-3msg le-lait	"le garçon boit / est en train de boire le lait"
EA	?il-walad bi-yi-šrab il-laban	the-boy impf-impf-drink the-milk	"le garçon boit / est en train de boire le lait"

Nous pouvons affirmer en EA, que le préfixe [bi-] est combiné avec le préfixe imperfectif [(?)a-], pour la première personne du singulier, afin de former un morphème [ba-]. Par exemple, /bi-?a-ftah/ 'j'ouvre / je suis en train d'ouvrir' est réduit à /ba-ftah/. Ceci indique que l'opération d'insertion d'arrêt glottal dans EA insère la glotte [?a-] dans la surface de la structure seulement. Si la forme de base du premier préfixe singulier est [a-], alors nous pourrions proposer que le verbe /bi-?a-ftah/ devient /baa-ftah/ en appliquant l'*Assimilation vocoïde anticipatoire*, ensuite nous appliquons sur le résultat obtenu l'opération de *raccourcissement de la syllabe* fermée afin d'obtenir à la fin /ba-ftah/.

La forme imperfective peut aussi porter sur le futur. L'EA se distingue du MSA, à cet égard, par l'utilisation de marqueurs non-standards. Les marqueurs de futur en MSA sont le préfixe [sa-] et la particule /sawfa/ 'aller faire' qui sont déposés avant la forme imperfective, alors qu'en EA, ils sont remplacés par [ha- ou Ha-] qui sont, à leur tour, préfixés à la forme imperfective. Selon (Robertson, 1970) et (Gadalla, 2000), ces préfixes se sont développés à partir du verbe /raah(-a)/ 'partir'. Pour la première personne du masculin singulier, [ha-] est combiné avec [?a-] pour constituer un morphème [ha-]. Cette particule est aussi utilisée dans certains dialectes du Maghreb comme celui des parlers d'Annaba. Enfin, la comparaison de /sa-?a-ktub-u/ en MSA avec /ha-ktib/ en EA confirme le fait que même si l'arrêt glottal est encore basique en MSA, il ne l'est plus en AE. Voici quelques exemples de cette utilisation du futur en MSA et en EA.

Type de l'arabe	Phrase	Structure	Traduction
MSA	sa-?a-ktub-u d-dars	fut-impf-écrire-1sg la-leçon	"J'écrirai la leçon".
	sawfa ?aakul-u 1-3inab	fut-impf.manger-1sg les-raisins	"Je mangerai les raisins"
EA	ha-ktib id-dars	fut-impf-écrire-1sg la-leçon	"J'écrirai la leçon"
	haakul il-3inab	fut-impf.manger-1sg les-raisins	"Je mangerai les raisins"

Dans le dialecte tunisien, l'utilisation du futur passe par l'emploi de la particule indépendante [bA$], qui n'est plus préfixée à la forme imperfective et qui se situe avant le verbe. Quant au dialecte du Maroc, le futur est exprimé par l'introduction de la particule [ka] qui à son tour préfixé à la forme imperfective. Le dialecte algérien emploie l'auxiliaire rayəh رايح (-a, -in) avant le verbe à la forme imperfective.

Au sujet du mode impératif, sa formation est liée à la discussion autour de la conjugaison de l'aspect/le mode. Selon certains linguistes, le mode impératif en MSA est dérivé à partir du jussif, comme le dit (Schramm, 1962), qui a établi les règles de cette dérivation comme suit : « *la règle empirique traditionnelle pour la dérivation de l'impératif à partir du jussif procède comme suit : Les deux premiers phonèmes de la seconde personne du jussif sont soustraits, et si le résidu commence par un groupement de consonnes, le /?/ en plus d'une voyelle seront préfixés. En général, si la voyelle du résidu est /u/, la voyelle préfixée est /u/, sinon c'est /i/. ».*

Il existe bien sûr une différence entre 'une règle empirique' et une règle grammaticale réelle. La première est un outil indispensable pour un apprenti, tandis que la dernière est une hypothèse sur la façon avec laquelle les locuteurs natifs savent inconsciemment quelle forme produire. On ne sait pas encore s'il est linguistiquement justifié de dire que l'impératif est dérivé du jussif, bien qu'ils partagent plusieurs propriétés similaires. Ceci peut être justifié par le fait que dans les dialectes en général, il n'y a pas de jussif. Ainsi, contrairement à l'hypothèse de (Schramm, 1962), nous supposons que le jussif en MSA en plus de l'impératif en arabe standard et dialectal, sont tous dérivés de la racine imperfective du verbe. Considérons les exemples suivants:

MSA		EA		Traduction
impf	**imper**	**impf**	**imper**	
ya-nzil(-u)	?i-nzil	yi-nzil	?i-nzil	Descendre
ya-dfa3(-u)	?i-dfa3	yi-dfa3	?i-dfa3	Payer
ya-Drib(-u)	?i-Drib	yi-Drab	?i-Drab	Frapper

Nous remarquons à travers les exemples du tableau précédent que lorsque les racines imperfectives sont identiques en arabe standard et dialectal, les formes impératives sont aussi identiques, et de la même manière lorsque les formes imperfectives sont différentes, les impératives le sont aussi. Ceci est en faveur de l'hypothèse faisant état que l'impératif est dérivé de l'imperfectif. Dans la grammaire arabe, l'impératif est formé en faisant les actions suivantes :

- Détacher le préfixe de la racine imperfective

- Insérer une voyelle haute et un arrêt glottal pour réaliser la syllabisation. La voyelle haute est choisie en fonction de cette règle spéciale :

$$Impératif\ V \rightarrow \begin{cases} u\ /\ si\ C.Cu \\ i\ /\ sinon \end{cases}$$

Cette règle montre que la voyelle du préfixe impératif est toujours /i/ sauf si la racine du radical est la voyelle /u/. Dans ce dernier cas la voyelle /u/ est utilisée pour garantir l'harmonie des voyelles. Un exemple avec un /u/ dans le radical imperfectif dans le MSA et l'arabe dialectal est le verbe /rasam(-a) → ya-rsum(-u) > yu-rsum/ 'dessiner'. Le tableau suivant présente une comparaison entre les formes impératives de ce verbe dans les deux variétés :

Référent	**MSA**	**EA**	**TA**
2msg	?ursum	?ursum	?ursum
2fsg	?ursum-ii	?ursum-i	?ursm-i
2du (m & f)	?ursum-aa	?ursum-u	?ursm-u
2mpl	?ursum-uu		
2fpl	?ursum-na		

L'analyse du tableau précédent affirme bien l'hypothèse de formation de la forme impérative à partir de la forme imparfaite. De plus, dans ce tableau il est montré que pour le dialecte tunisien l'élision de la voyelle dans la 2ème personne au féminin singulier et au pluriel à cause de raisons phonétiques. Nous retrouvons cette même caractéristique dans le dialecte algérien, cependant ce dernier possède aussi une autre variante pour la forme impérative pour les deux personnes mentionnées dessus. Cette forme consiste à géminer la première racine et déplacer la voyelle de la 2ème consonne du radical vers le 1ière consonne. Par exemple, pour le verbe rasama (dessiner), nous pouvons dire à la fois pour la 2ème personne au féminin singulier /?ursm-i/ et /?uruusm-i/ et 2ème personne au pluriel nous disons en dialecte algérien /?ursm-u/ et /?uruusm-u/.

Par ailleurs, le dialecte tunisien contient aussi des situations particulières. C'est le cas où il utilise la voyelle du préfixe impératif (a), au lieu de /i/ ou /u/, lorsque la racine du radical est la voyelle /a/. Cette situation nous permet de déduire la règle suivante pour l'impératif dans le tunisien :

$$Impératif\ V \rightarrow \begin{cases} u\ /\ si\ C.Cu \\ a\ /\ si\ C.Ca \\ i\ /\ sinon \end{cases}$$

Voici quelques exemples de l'application de cette règle pour les cas de la voyelle du préfixe impératif /a/ :

TA		**Traduction**
impf	**Imper**	
ya-fraH	?a-fraH	se réjouir
ya-qra	?a-qra	étudier
ya-HbaT	?a-HbaT	descendre
ya-xlaT	?a-xlaT	arriver

L'introduction de la règle concernant la voyelle /a/ est aussi présente dans le dialecte de l'est de l'Algérie et du Sahara. Dans les autres régions algériennes, le dialecte courant utilise le préfixe impératif /a/ pour la majorité des verbes utilisant le /i/ dans le dialecte tunisien ou égyptien.

Par ailleurs, nous avons identifié une nouvelle règle dans le dialecte tunisien pour connaître la voyelle du préfixe impératif à partir de celle du préfixe imparfait du verbe. Pour illustrer cette règle prenons les exemples suivants :

Impf	**Imper**	**Traduction**
ya-Tla3	?a-Tla3	monter
ya-3Si	?a-3Si	désobéir
yi-zni	?i-zni	forniquer
yi-nisa	?i-nsa	Oublier
yu-dxul	?u-dxul	Entrer
yu-qtul	?u-qtul	Mourir
ya-s3al	?a-s?al	demander
y-kassar	kassar	Casser
y-quum	quum	se lever

Un autre cas est aussi montré par les exemples du tableau, il s'agit de la chute de la hamza lorsque le préfixe imparfait du verbe est soukoun (i.e. le verbe se lever). Dans la grammaire arabe, il existe des verbes où le préfixe impératif n'est pas utilisé, en particulier, selon (Malik, 1976) : "*lorsque, après l'omission du préfixe imperfectif, la racine commence par une consonne vocalisée* ". Ces verbes peuvent appartenir à l'une des catégories suivantes :

i. les verbes géminés
ii. les verbes creux
iii. les verbes sains de la Forme II
iv. les verbes sains de la Forme III
v. les verbes quadrilitères

Voici quelques exemples illustratifs donnés dans le tableau suivant :

MSA	**EA**	**TA**	**Traduction**
jurr	gurr	jurr	tirer!
šudd	šidd	šidd	tirer!
xaf	xaaf	xaaf	avoir peur!
qum	?uum	qumm	lève-toi!
Tir	Tiir	Tiir	vole-toi!
3allim	3allim	3allim	enseigner!
saafir	saafir	saafir	voyager!
Zaxrif	Zaxrif	zaxrif	décorer!

Le tableau montre une autre caractéristique de la forme impérative dans les dialectes, notamment tunisien et égyptien. Il s'agit de la conservation des voyelles longues des verbes creux. Cette caractéristique représente un contraste avec leur quantité réduite en MSA. Selon

(Gadalla, 2000), ceci peut être expliqué en proposant que la règle de raccourcissement des syllabes fermées s'applique seulement en MSA car dans le dialecte égyptien, et par extension pour le dialecte tunisien et les dialectes maghrébins en général, la consonne finale ne compte pas dans le poids de la syllabe comme ceci a été soutenu par la condition extra métrique en AE.

Sur un autre point, les verbes glottalisés et plus particulièrement les deux verbes /ʔaxað/ (prendre) et /ʔakal/ (manger) présentent une conjugaison particulière qui ne se rapproche de celle du verbe hamzé au futur en dialecte tunisien, et plus généralement les dialectes maghrébins. Nous constatons aussi la chute de la syllabe entière pour ces verbes. Pour expliquer cette particularité, [Ouerhani, Bachir,], propose d'utiliser le modèle (le schème) et la réalisation suivants pour les verbes concernés :

Accompli	**Inaccompli**	**Impératif**
cacaca	ya-ccucu	ʔu-ccuc
ʔakala	ya-ʔkulu	ʔu-ʔku

A l'impératif en MSA, la «Hamza » de la racine consonantique est en voisinage direct, dans la même syllabe, avec une deuxième «hamza» nécessaire à la conjugaison de l'impératif. Cette proximité entraîne la chute de la syllabe entière présentée par la règle morpho-phonologique suivante : ʔuʔkul → kul. Cette chute de la " hamza " est une caractéristique de la conjugaison de tout verbe « mahmouz» ayant un [u] comme 2ème voyelle à l'inaccompli. Cette règle de la chute de la hamza est aussi homogène et présente dans les dialectes arabes. Cependant la structure syllabique obtenue par cette règle dans les dialectes marque une différence par rapport à son application dans le MSA. Cette différence consiste en un allongement systématique de la 2ème voyelle, par exemple : le verbe kuul (mange, 2ms), kuuli (mange, 2fs), kuul-u (mangez, 2mp).

Nous mentionnons aussi un autre cas exceptionnel dans le dialecte tunisien concernant le verbe jaa (venir). Ce verbe suit un modèle préétablie pour sa conjugaison à la forme impérative qui change la structure du schème verbale comme suit :

Perfectif	**Imperfectif**	**Impératif**		
jaa	y-jii	?ija (2ms)	?ij-i (2fs)	?ij-u (2mp)

Ce cas possède un autre traitement différent dans le dialecte algérien[17]. En effet, l'impératif du verbe jaa dans le dialecte algérien est exprimé, non pas en utilisant le radical du verbe, mais plutôt en substituant le radical par un autre radical ayant la même sémantique. Ce radical utilisé est celui du verbe raaH. Le tableau suivant montre cette exception :

Perfectif	**Imperfectif**	**Impératif**		
jaa	y-jii	?arwaaH (2ms)	?arwaaH-i (2fs)	?arwaaH-u (2mp)

Enfin, nous signalons que pour exprimer les invitations et les suggestions dans le dialecte égyptien, un préfixe imperfectif distinctif [ma-] est employé avant la forme imperfective. Par exemple, prenons les verbes /ma-taakul/ 'Allez, mange', /ma-til3ab-u/ 'Allez, jouez' et /ma-tiigi/ 'Allez, viens'.

[17] Voir le même traitement pour l'égyptien

6.3. La voix de la flexion:

Dans la grammaire arabe, il existe deux voix pour la flexion des verbes : la voix active et la voix passive. La voix active préserve la forme usuelle du verbe, quant à la voix passive, elle est dérivée *avec un simple changement du timbre de la mélodie vocalique* de la forme active (Arbaoui, 2010), en utilisant un ensemble de modèles de voyelles qui est *une séquence vocalique interne contrastant avec celle des verbes actifs* : **[u-i]** ou **[u-a]**. Cette dérivation est aussi appelée formation par une *apophonie.*

Le choix du modèle des voyelles dépend de la valeur des traits morphologiques représentant l'aspect du verbe qui peut être perfectif ou imperfectif :

- *Verbe perfectif* : pour ces verbes la flexion est effectuée en utilisant le modèle [u-i]. C'est le cas par exemple des verbes :
  - /jama3(-a)/ 'il a ramassé' devient à la voix passive /jumi3(-a)/ 'il a été ramassé'
  - /šarib(-a)/ 'il a bu' devient à la voix passive /šurib(-a)/ 'il a été bu'.
- *Verbe imperfectif* : la flexion pour les verbes ayant ce trait suit le modèle **[u-a]**. Pour illustrer cette utilisation, voici quelques exemples :
  - /ya-jma3(-u)/ 'il ramasse' est changé dans la voix passive en [yu-jma3(-u)/ 'il est ramassé' et /ya-šrab(-u)/
  - 'il bois' devient dans la voix passive /yu-šrab(-u)/ 'il est bu'

Par ailleurs, la flexion possède de formes particulières en fonction du type des verbes. Dans le cas des verbes creux, la forme perfective passive possède le modèle [FiiL(-a)] qui est obtenu à partir du modèle de base [FuGiL(-a)] en appliquant les règles d'élision de la glide et d'assimilation vocoïde anticipatoire. Par exemple, le verbe /qaal(-a)/ 'il a dit' devient /qiil(-a)/ 'il a été dit' dans la voix passive. La forme imperfective passive possède quant à elle le modèle de surface [yu-FaaL(-u)] basée sur le modèle [yu-FGaL(-u)] et obtenue par application à ce modèle de base l'opération de l'assimilation vocoïde anticipatoire, comme pour le verbe /ya-bii3(-u) / 'il vend' qui est exprimé dans la voix passive par /yu-baa3(-u)/ 'il se vend'.

Le tableau (6.40) donne les équivalents passifs pour toutes les formes actives des verbes trilitères en MSA.

No	**Perfectif**		**Imperfectif**	
	Active	Passive	Active	Passive
I	Fa3aL(-a)	Fu3iL(-a)	ya-F3i/a/uL(-u)	yu-F3aL(-u)
II	Fa33aL(-a)	Fu33iL(-a)	yu-Fa33iL(-u)	yu-Fa33aL(-u)
III	Faa3aL(-a)	Fuu3iL(-a)	yu-Faa3iL(-u)	yu-Faa3aL(-u)
IV	?aF3aL(-a)	?uF3iL(-a)	yu-F3iL(-u)	yu-F3aL(-u)
V	taFa33aL(-a)	tuFu33iL(-a)	ya-taFa33aL(-u)	yu-taFa33aL(-u)
VI	taFaa3aL(-a)	tuFuu3iL(-a)	ya-taFaa3aL(-u)	yu-taFaa3aL(-u)
VII	(?i)nFa3aL(-a)	Ø	ya-nFa3iL(-u)	Ø
VIII	(?i)Fta3aL(-a)	?uFtu3iL(-a)	ya-Fta3iL(-u)	yu-Fta3aL(-u)
IX	(?i)F3aLL(-a)	Ø	ya-F3aLL(-u)	Ø
X	(?i)staF3aL(-a)	?ustuF3iL(-a)	ya-staF3iL(-u)	yu-staF3aL(-u)

Tableau 6. 40. Les formes passives des verbes trilitères en MSA

Le tableau (6.40) montre que les formes VII et IX ne possèdent pas de formes passives. Ceci est dû au fait que ces verbes sont des verbes *inaccusatifs* ou *intransitif* : ces verbes possèdent un seul argument qui est le sujet pour accomplir le sens de la phrase.

La voix passive existe aussi dans la variété dialectale mais avec quelques différences significatives par rapport à cette même voix dans le MSA. Sur ce sujet, (Al-Toma, 1969) affirme que "*la forme passive résultant d'un changement interne des voyelles du verbe a disparu des dialectes modernes*". Ce constat s'applique bien évidemment à l'EA et le TA, ainsi aux autres variétés dialectales qui abandonnent l'usage des modèles de voyelles cités au-dessus pour la formation de la voix passive. Cette voix est ainsi formée par l'introduction de nouveaux morphèmes, souvent le [t-] et moins fréquemment le morphème [n-] dans l'EA et le dialecte algérien, qui sont préfixés à la forme perfective et infixés à la forme imperfective. Par exemple, le dialecte tunisien marque la voix passive du verbe exprimé en MSA par كُتِب [kutiba] "il est écrit", par تِكتِب [**ti**ktib].

Cependant, il existe quelques exceptions à la règle citée ci-dessus dans certains dialectes. Dans certains dialectes algériens, comme par exemple certains parlers d'Alger ou les parlers de Djidjel et du Nord oranais (Tlemcen), cette voix peut être exprimée avec les mêmes formes moyennant un amalgame de ces deux préfixes : /nte-/, /ten-/ ou /tten/, ce qui donne des formes comme /ntedreb/ (il a été frappé) tendreb, ttendreb, (نتضرب، تتضرب).

Dans le dialecte égyptien, une voyelle initiale et un arrêt glottal sont insérés seulement dans les formes perfectives, et cela selon les règles phonologiques régulières d'épenthèse de début de mot et d'insertion d'arrêt de glotte. Un examen plus attentif de ces faits indique que les morphèmes [t-] et [n-] ne sont pas nouveaux puisqu'ils sont déjà utilisés dans les formes V, VI et VII. Par conséquent, selon (Gadalla, 2000), la voix passive est en train de perdre du terrain dans l'EA et se voit remplacer par ces formes. En d'autres termes, l'EA utilise des formes réflexives pour indiquer la voix passive. Ces formes sont [(?i)tFa3aL, (?i)tFa33aL ~ (?i)tFa33iL, (?i)tFaa3iL et (?i)nFa3aL] au mode perfectif et [yi-tFi3iL, yi-tFa33aL ~ yi-tFa33iL, yi-tFaa3iL et yi-nFi3iL] à l'imperfectif.

Ceci montre que [t-] est plus fréquemment utilisée par rapport à [n-], étant donné que le premier est utilisé avec trois formes, tandis que le dernier est utilisé dans une seule forme. De plus, le premier morphème est utilisé pour former l'équivalent de trois formes passives en MSA : I, II et III, alors que le second est utilisé pour former l'équivalent d'une seule forme en MSA, à savoir la forme I. Le tableau suivant illustre quelques exemples d'utilisation de cette voix dans le dialecte égyptien :

MSA		**EA**		**Traduction**
Perfectif	**Imperfectif**	**Perfectif**	**Imperfectif**	
šurib(-a)	yu-šrab(-u)	(?i)tšarab	yi-tširib	Il a été bu
kussir(-a)	yu-kassar(-u)	(?i)tkassar	yi-tkassar	Il a été cassé
quubil(-a)	yu-qaabal(-u)	(?i)t?aabil	yi-t?aabil	Il a été rencontré
Durib(-a)	yu-Drab(-u)	(?i)nDarab	yi-nDirib	Il a été frappé

Dans certains verbes dialectaux, le [t-] formant la voix passive est assimilé, selon (Robertson, 1970) aux consonnes suivant le morphème. Cette assimilation est appliquée si les consonnes sont : dentales, palatales ou vélaires, comme le montrent les verbes suivants :

MSA	EA	Traduction
dummir(-a)	(?i)d-dammar	Il a été détruit
šujji3(-a)	(?i)š-šagga3	Il a été encouragé
suriq(-a)	(?i)s-sara?	Il a été volé
kunis(-a)	(?i)k-kanas	Il a été balayé
jumi3(-a)	(?i)g-gama3	Il a été collecté

(Gadalla, 2000) souligne que les marqueurs de voix passive en dialecte [t-] et [n-] sont aussi utilisés pour former les verbes intransitifs ou inchoatifs. En d'autres termes, les verbes ayant ces marqueurs dans leurs structures surfaciques ne peuvent pas être tous considérés comme des transformations passives de verbes actifs. Ils peuvent être des verbes de l'une des formes V, VI ou VII, qu'on appelle généralement "passif d'état" selon (Wise, 1975). Par conséquent, nous nous retrouvons avec deux classes de verbe : passif et passif d'état, et ces marqueurs peuvent être considérés comme un facteur sémantique pour faire la différence entre ces deux classes. Cependant, la vérification de l'appartenance à l'une de ces classes dépend de la présence d'un agent externe, par exemple :

- 'il3asaakir itlammu fi ilmidaan' 'les soldats se sont rassemblés dans la cours' : forme passive d'état suite à la présence de l'agent 'il3asaakir' (les soldats)
- 'ilfirawla itlammit' 'les fraises ont été cueillies' : forme passive.

Cette ambiguïté ne pose pas de problème dans le MSA puisque les seules formes qui peuvent être des passives d'état sont les formes V, VI et VII et que le passif possède une forme incluant les modèles de voyelles [u....i/a]. Ainsi, nous pouvons trouver des verbes en dialecte égyptien possédant deux équivalents différents en MSA: un équivalent passif et un autre passif d'état comme montré dans le tableau d'exemples suivant :

MSA Passive	MSA Pseudo-Passive	EA
nubbih(-a) 'il a été alerté'	tanabbah(-a)'il est devenu alerté'	(?i)tnabbah
nuuqiš(-a) 'il a été discuté'	tanaaqaš(-a) 'il a discuté avec'	(?i)tnaa?iš
kusir(-a) 'il a été cassé'	(?i)nkasar(-a) 'il a cassé'	(?i)nkasar

Chapitre 7 Analyse morphologique nominale

Introduction

Ce chapitre traitera la morphologie des noms d'une part en arabe standard (MSA) et d'autre part en arabe dialectal (égyptien et algérien). Les noms seront divisés en deux classes: les noms primaires qui sont directement dérivés de la racine et les noms déverbaux qui sont eux, dérivés des verbes. Les formes des racines des noms primaires seront exposées dans (7.1). Ensuite, les modèles des noms déverbaux seront discutés dans (7.2). La différence entre les noms définis et indéfinis sera indiquée dans (7.3). Après cela, l'inflexion des noms pour le cas, le genre et le nombre sera traitée dans (7.4; 7.5 et 7.6), respectivement.

7.1. Les noms simples

Comme indiqué dans l'introduction, les noms simples en arabe, appelé aussi primaires, sont les noms dérivés de la racine directement. Il existe diverses classifications pour ces noms selon plusieurs critères. Dans la grammaire arabe, il existe trois classes morphologiques pour stratifier les formes des radicaux des noms simples : les racines, les schèmes (modèles) et les affixes :

- Une racine représente une séquence de deux, trois, quatre ou cinq lettres (massivement composé de trois consonnes) définissant une notion abstraite. Par exemple, les trois consonnes -F-T-H-, dans cet ordre, forment une racine pouvant être dérivée dans divers schèmes pour former des noms.
- Un schème appelé aussi un patron syllabique est un modèle vocalique permettant de définir comment placer les radicaux dans la racine. La structure ma-R1R2ûR3- forme un schème dans lequel R1, R2 et R3 représentent les phonèmes radicaux de n'importe quelle racine tri-consonantique. Dans le cas où nous considérons la racine F-T-H-, R1, R2 et R3 correspondent aux lettres F, T et H respectivement, ce qui donne pour le schème ma-R1R2ûR3- le nom ma-FTûH (ouvert).
- Les affixes sont des morphèmes liés qui s'ajointent au radical d'un mot en arabe. Dans la grammaire arabe, la morphologie dérivationnelle désigne par le terme '*affixes*' tous les morphèmes s'ajoutant aux radicaux, tandis que la morphologie flexionnelle les appelle '*marques flexionnelles*' ayant pour fonction de marquer le genre et le nombre.

  Par ailleurs, les affixes dérivationnels possèdent à leur tour des cas spécifiques en fonction de la fonction de dérivation utilisée. Ils peuvent être :
  - ✓ Des morphèmes vocaliques, comme c'est le cas pour l'allongement des voyelles brèves,
  - ✓ Des morphèmes consonantiques comme pour la gémination des consonnes simples,
  - ✓ Des morphèmes vocalo-consonantiques qui se présentent sous forme de préfixes, suffixes ou infixes :
    - [CV-] : les préfixes qui sont placé avant le radical, c'est le cas du préfixe [ma-] dans le mot maktab(-un) [مَكْتَبٌ, un bureau]
    - [-VVC] : les suffixes qui sont placés après le radical, comme dans l'exemple [-aan] dans /gufraan(-un) [غُفْرَانٌ, pardon].
    - [-C-] : les infixes qui sont placés à l'intérieur du radical, par exemple l'infixe [-n-] dans /fannaan(-un) [فَنَّانٌ, un acteur].

Il existe aussi par ailleurs une classification des noms simples, que ce soit en arabe standard (MSA) ou en arabe dialectal, selon le nombre de consonnes constituant leur racine. Il s'agit des classes : bilitères, trilitères et quadrilitères.

Les noms bilitères possèdent une seule racine de base qui est [F-3-], comme dans les mots /fam(-un) (فَمٌ, une bouche), dam(-un) [دَمٌ, un sang].

En ce qui concerne les noms trilitères, leurs racines consonantiques possèdent différents types pouvant être présentés comme suit :

Type	Description	Exemple
F-3-L	radical sain ayant trois différentes consonnes.	batn(-un) [بَطْنٌ, un ventre], jabal(-un) [جَبَلٌ, une montagne]
F-33	radical double qui comporte deux consonnes identiques à la deuxième et troisième position.	wubb(-un) [حُبٌّ, amour], fakh(-un) [فَخٌّ, un piège]
w/y-3-L	radical à initial-défectueux, c'est à dire ayant une consonne faible au début.	wa3d(-un) [وَعْدٌ, une promesse], yusr(-un) [يُسْرٌ, aisonce]
F-w/y-L	racine médiane-défectueuse, c'est à dire ayant une consonne faible au milieu	fawz(-un) [فَوْزٌ, une réussite], bayD(-un) [بَيْضٌ, œufs]
F-3-w/y	racine finale-défectueuse, c'est à dire ayant une consonne faible à la fin.	ra?y(-un) [رَأيٌ, une opinion], qabw(-un) [قَبْوٌ, une cave]

Pour ce type de noms il existe plusieurs modèles (patrons) vocaliques que nous pouvons trouver dans les racines nominales. Selon (Gadalla, 2000), ces modèles peuvent être présentés comme suit :

1. [-V--], comme dans le mot : nahr(-un) [نَهْرٌ, une rivière].
2. [-V-V-], comme dans : rajul(-un) [رَجُلٌ, un homme].
3. [-V-VV-], comme dans : kitaab(-un) [كِتَابٌ, un livre].
4. [-VV-V-], comme dans waalid(-un) [وَالِدٌ, un père].

Quant aux noms quadrilitères, ces derniers ont une seule racine de base, à savoir la forme [F-3-L1-L2] où L1 et L2 sont deux consonnes différentes, comme dans les mots /?arnab(-un) (أَرْنَبٌ, un lapin), dirham(-un) [دِرهَمٌ, un Dirham].

7.1.1. Les racines bilitères

Les noms dans l'arabe standard ayant des racines bilitères, par conséquent composés de deux lettres, constituent un vocabulaire limité. Ce type de nom dans l'arabe dialectal présente quelques spécificités différentes de celles du MSA, comme le souligne Marçais dans (Marçais, 1902) : « *les vieux bilitères ont repris la trilitarité par le redoublement de leur deuxième consonne* ». En d'autres termes, les noms bilitères n'existent pas dans les dialectes car ils sont transformés en racine trilitère. Cette transformation est effectuée par une opération morphologique qui consiste en un redoublement de la deuxième consonne de leur racine bilitère d'origine. (Marçais, 1902) souligne par ailleurs que cette trilitarité est particulièrement mise en relief par l'apparition de diminutifs comme idida (petite main), dmiyem (un petit sang) et de pluriel comme dmûm. Le tableau suivant présente une comparaison de cette forme de nom entre le MSA, l'arabe égyptien (AE) et l'arabe algérien (AA).

MSA	AE	AA	Traduction
dam(-un)	damm	demm	Sang
fam(-un)	ϕ	fumm	Bouche

yad(-un)	(?i)ydd	yedd	Main

Tableau 7. 1. Une comparaison des noms bilitères entre le MSA, AE & AA

Le tableau montre que le dialecte égyptien ne possède pas le mot fam(-un), il est remplacé par un autre mot qui est « bou' ». Le tableau montre aussi que le dialecte égyptien rajoute l'épenthèse '?i' au début du mot (?i)ydd afin d'éviter la succession de deux consonnes. Nous constatons aussi que pour le dialecte algérien il existe un changement de la voyelle (a) par la voyelle (e) pour les mots sang et main, et par (u) pour le mot bouche.

Nous terminons cette sections par la citation suivante de (Fleish, 1961) au sujet des mots bilitères, ce dernier souligne que : « *toutes les langues sémitiques sont à base de trilitéralité (mis à part le petit vocabulaire bilitère, analysé pour l'arabe, et par ailleurs les quadrilitères). Donc le sémitique commun, d'où ces langues sont issues, était lui aussi à base de trilitéralité prédominante*. »

7.1.2. Les racines trilitères

Dans la grammaire arabe, une immense majorité des noms ont des racines trilittères. La plupart des racines trilitères des noms en (MSA) sont conservées en arabe dialectal (EA ou AA). Autrement dit, elles sont identiques dans les deux variétés, comme le montre le Tableau (7.2).

No	Racine	Exemple	Traduction
1	Fa3L(-un)	qalb(-un)	un cœur
2	Fa3L-at(-un)	dast-at(-un)	une douzaine
3	Fi3L(-un)	3ilm(-un)	Science
4	Fi3L-at(-un)	fikr-at(-un)	une idée
5	Fu3L(-un)	furn(-un)	un four
6	Fu3L-at(-un)	furS-at(-un)	une chance
7	Fa3aL(-un)	haram(-un)	une pyramide
8	Fa3aL-at(-un)	barak-at(-un)	bénédiction
9	Fa3iL(-un)	malik(-un)	le roi
10	Fa3iL-at(-un)	malik-at(-un)	la reine
11	Fu3uL(-un)	3unuq(-un)	un cou
12	Fa3aaL(-un)	salaam(-un)	la paix
13	Fa3aaL-at(-un)	salaam-at(-un)	la sécurité
14	Fa3uuL(-un)	rasuul(-un)	un messager
15	Fa3uuL-at(-un)	3aruus-at(-un)	une mariée
16	Fu3aaL(-un)	guraab(-un)	un corbeau
17	Fu3aaL-at(-un)	fukaah-at(-un)	Humour
18	Fu3uuL(-un)	muruur(-un)	le trafic
19	Fu3uuL-at(-un)	hukuum-at(-un)	un gouvernement
20	Fi3aaL-at(-un)	risaal-at(-un)	une lettre
21	Faa3aL(-un)	3aalam(-un)	le monde
22	Faa3iL(-un)	Taabi3(-un)	un tampon

23	Fa33aaL(-un)	bahhaar(-un)	un marin
24	Fa33aaL-at(-un)	dabbaas-at(-un)	une agrafeuse
25	Fi33aaL-at(-un)	Sinnaar-at(-un)	un crochet
26	Fu33aaL-at(-un)	kurraas-at(-un)	Agenda
27	FaaL(-un)	baab(-un)	une porte
28	FaaL-at(-un)	saa3-at(-un)	une montre
29	FiiL(-un)	fiil(-un)	un éléphant
30	FiiL-at(-un)	ziin-at(-un)	décoration
31	FuuL(-un)	nuur(-un)	une lumière
32	FuuL-at(-un)	Suur-at(-un)	une photo
33	Fa3at(-un)	sanat(-un)	une année
34	Fu3at(-un)	lugat(-un)	une langue
35	?uF3uuL(-un)	?usbuu3(-un)	une semaine

Tableau 7. 2. Les racines trilitères identiques des noms en MSA & EA

Cependant, nous signalons deux remarques importantes qui font la différence entre le MSA et les variantes dialectales de l'arabe au niveau de ces formes :

1. Le signe ou la marque de cas (-un) n'existe pas dans l'arabe dialectal (cf. Chapitre 5, Section 5.2).
2. Le /t/ final dans le MSA est omis en arabe dialectal à cause d'une règle morphologique régulière que nous avons nommée '*la suppléance [-a ~ -t]*'. Pour définir ce phénomène, considérons la définition de (Guella, 2010) : «*la suppléance est un procédé linguistique qui consiste à remplacer ou à substituer un élément ou terme ou proposition précédemment mis en contexte.* ». Ce procédé s'applique afin de répondre à des besoins conversationnels d'individus ou de communautés, ce qui lui donne un caractère flexible et évolutif basé sur l'expérience créatrice. Il s'applique aussi à certains domaines de langue comme la création lexicale. En arabe dialectal, la règle de la suppléance [-a ~ -t] est formulée comme suit :

$$\begin{cases} -\text{at} \rightarrow -\text{a} \,/\, \text{à la fin d'une phrase} \\ -\text{at} \rightarrow -\text{t} \,/\, \text{ailleurs} \end{cases}$$ **… règle de la suppléance en arabe dialectal**

Le tableau (7.2) montre bien que, hormis l'effet de la règle de suppléance, les modèles suivants sont utilisés pour les radicaux trilitères dans l'arabe standard et dialectal : [CVCC(a)], [CVCVC(a)], [CVCVVC(a)], [CVVCVC(a)], [CVCCVVC(a)], [CVVC(a)], [CVCa] et [?uCCVVC]. Le dialecte égyptien présente une exception qui est le nom /kubbaaya/ (un verre), qui est équivalent en (MSA) au nom /kuub(-un)/ suivant la forme [FuuL(-un)] conservée dans l'arabe dialectal en général. Au niveau du dialecte algérien, nous recensons aussi une autre exception au niveau du dernier modèle [?uCCVVC] qui est remplacé par le modèle [CCVVCa], comme pour le mot smaana (semaine).

D'autres formes de racines ont subi des changements phonologiques réguliers dans l'arabe dialectal, et ces formes sont données dans le tableau (7.3).

No	Forme			Exemple			
	MSA	**EA**	**AA**	**MSA**	**EA**	**AA**	**Traduction**
1	FayL(-un)	FeeL	Fi:L	Sayf(-un)	Seef	Si:f	l'été
2	FayL-at(-un)	FeeL-a	Fi:L-a	layl-at(-un)	leel-a	li:l-a	Nuit
3	Fay3aaL(-un)	Fi3aaL	Fi3aaL	maydaan(-un)	midaan	midaan	a square
4	FawL(-un)	FooL	Fo:L	ðawq(-un)	zoo?	ðo:q	le goût
5	FawL-at(-un)	FooL-a	Fo:L-a	zawj-at(-un)	zoog-a	zooj-a	une épouse
6	Fa?L(-un)	FaaL	FaaL	fa?s(-un)	faas	faas	une pioche
7	Fi?L(-un)	FiiL	FiiL	ði?b(-un)	ziib	ðiib	un loup
8	Faa?(-un)	Fayya	Faa	maa?(-un)	mayya	maa	l'eau
9	Faa3uuL(-un)	Fa3uuL	Fa3uuL	Saaruux(-un)	Saruux	Saruux	une fusée
10	Fii3aaL(-un)	Fi3aaL	Fi3aaL	miizaan(-un)	mizaan	mizaan	une balance
11	Fuu3aaL(-un)	Fu3aaL	ϕ	duulaab(-un)	dulaab	ϕ	une armoire
12	Faa3iL-at(-un)	Fa3L-a	Fa3L-a	jaami3-at(-un)	gam3-a	jam3-a	université
13	Fu3Liyy(-un)	Fu3Li	Fu3Li	kursiyy(-un)	kursi	kursi	une chaise
14	Fa3iyy(-un)	Fa3i	Fa3i	nabiyy(-un)	nabi	Nabi	un prophète
15	Fu3Laa	Fu3La	Fu3La	Dunyaa	dunya	denya	le monde
16	Fi3aa?(-un)	Fi3a	F3aa	šitaa?(-un)	šita	Sta	Hiver
17	Fii3aa?(-un)	Fii3a	Fii3a	miinaa?(-un)	miina	miina	le port

Tableau 7. 3. Les formes de racines nominales en MSA ayant subi des changements phonologiques dans les dialectes égyptien et algérien

Le tableau (7.3) montre les modèles caractérisant les racines trilitères en (MSA) qui ont subi des changements phonologiques réguliers en arabe dialectal (AE et AA). Ces modèles sont : [CawC(a)], [CayC(a)], [CayCaC], [CayCaaC], [CuCayC], [CV?C], [Caa?], [CVVCVVC], [CaaCiCa], [CVCCiy], [CaCiy], [CuCCaa], [CVCaa?], [CVVCVV?]. Ces changements peuvent être expliqués par l'application de certaines règles morphologiques ou phonologiques comme suit :

- ✓ La règle de la "Suppléance [-a~ -t]" caractérisant les dialectes est responsable de la disparition du /t/ du MSA à la fin des racines en arabe dialectal dans les cas identifiés par les formes : (2), (5) et (12).
- ✓ Les règles phonologiques appliquées dans les formes remontées du tableau (7.3) sont :
  - o La diphtongue s'applique aux racines des cas (1) à (5)
  - o L'allongement compensatoire est utilisé dans les cas (6) et (7)
  - o Les formes (8), (16) et (17) ont subi la suppression du /?/ en position finale
  - o Le raccourcissement atonique apparaît dans les racines (9) à (11)
  - o Le raccourcissement de la voyelle en dernière position est appliqué dans les racines (8) et les cas de (15) à (17). Cette règle est aussi présente dans les racines (13) et (14) car les linguistes considèrent /iy/ comme équivalent à /ii/ dans le contexte phonologique. A titre d'exemple, ce changement de /iy/ en /ii/ est présent devant un suffixe consonant en MSA, comme c'est le cas du mot nabiyy(-un) [نَبِيٌّ, un prophète] en MSA qui est transformé à nabii-na [نَبِينَا, notre prophète] en dialecte.

- ✓ En ce qui concerne le dialecte algérien : les mêmes transformations sont effectuées sauf pour les exceptions suivantes :
  - La forme (16) est caractérisée par la suppression de la hamza /?/ en position terminale du mot et aussi par l'élision de la voyelle initiale /i/.
  - Pour la forme (12), le dialecte algérien présente une exception concernant le mot faakih-at(-un) (fruit) ayant comme équivalent en MSA la forme Faa3iL-at(-un) et en arabe égyptien Fa3L-a. La forme caractérisant ce mot est par conséquent, en dialecte algérien, la forme Fa3y-a donnant le mot faky-a. Cette nouvelle forme est caractérisée par la chute de la lettre (h) substituée de ce fait par la lettre (y). La même transformation morphologique existe dans le mot du MSA wajh(-un) (visage). Après la chute de la lettre (h) à la fin du mot en arabe dialectal, le mot obtenu à une racine bilitère est transformé en trilitère avec le redoublement de la deuxième consonne. Nous obtenons en arabe algérien wajj, et wass en arabe égyptien. Il est important de souligner que le pluriel de ce mot, nous reprenons la forme classique du pluriel en MSA ce qui donne les pluriels suivants : (wjuuh) et (wguuh) en AA et EA respectivement.

    Dans la même optique, il est important de signaler que les mots trilitères en MSA qui après la chute d'une de leurs racines en dialecte, reprennent leur trilitarité par redoublement d'une consonne. Le nom *(?a)had(-un)* (dimanche), après la chute de la hamza en dialecte algérien, le mot est transformé en une racine trilitère par le redoublement de la dernière racine (hadd).

Sur un autre registre, certaines racines nominales du MSA subissent deux changements phonologiques dans les dialectes égyptien et algérien. Ceux-ci incluent les racines des formes (3, 16-17), par exemple : dans la forme (3), l'un des deux changements phonologiques n'apparaît pas à la surface. La forme du radical [Fay3aaL (-un)] subit d'abord une diphtongue, créant ainsi la forme [Fee3aaL]. Cette dernière forme subit à son tour un raccourcissement atonique, donnant lieu la forme [Fi3aaL]. Nous rappelons que le /i/ est considéré comme un court homologue du /ee/ en EA et comme un homologue du /î/ en AA.

De plus, il existe d'autres formes des radicaux subissant plus de deux changements phonologiques comme c'est le cas des formes (8) et (12).

En ce qui concerne la forme (8), [Faa?(-un)] en MSA, elle subit les transformations suivantes dans le dialecte égyptien (Gadalla, 2000):

i. La suppression du /?/ en dernière position,
ii. Le raccourcissement de la dernière voyelle : cette opération concerne la voyelle longue finale /aa/ et donne lieu à la forme du mot /ma/. Cette forme obtenue n'existe pas dans le dialecte égyptien, vu les conditions requises minimales pour un radical selon (McCarthy & Prince 1990a). De ce fait, comme la forme /ma/ ne contient qu'une seule more, elle est considérée fausse (mauvaise) selon la contrainte minimale du radical puisqu'une seule more est insuffisante selon cette contrainte, donc elle nécessite d'autres transformations, dont la compensation.
iii. La compensation de la more perdue, en raison du résultat de la transformation ii, le dialecte fait donc appel à l'ajout d'une autre more pour compenser celle perdue afin de

satisfaire le minimum bimorique, ou ce qui est communément appelé contrainte minimale du radical. Il est donc logique d'avoir une more de même nature phonologique que celle supprimée. L'équivalent phonologique le plus proche de /a?/ est /aa/, comme dans la règle de l'allongement compensatoire. La forme résultante serait donc /maaa/ qui dépasse la longueur maximale d'une voyelle. Il manque ainsi un début pour le dernier /a/.

iv. L'ajout d'un début pour le /a/ qui est par l'occurrence /y/ : l'ajout de la glide /y/ aboutit à la forme /maya/. Mais étant donné que /aya/ est équivalent à /aa/ comme nous le voyons dans le changement régulier /baya3(-a) → baa3(-a)/ 'vendre', ce résultat serait aussi inexact et nécessite une dernière transformation,

v. Le remplacement de la voyelle longue /aa/ par une voyelle et une glide /ay/ : pour obtenir la forme exacte, le dialecte égyptien opte pour le remplacement de la voyelle longue avant la glide par une voyelle brève et une glide, c'est à dire /aa/ → /ay/ et nous obtenons alors le mot /mayya/.

Ces changements appliqués au mot maa?(-un) en arabe égyptien sont résumés comme suit :

maa?(-un) →[i] maa →[ii] ma →[iii] maaa →[iv] maaya →[v] mayya

Pour ce qui est de la forme (12), à savoir la forme [Faa3iL-at(-un)], elle a aussi subi une transformation morphologique et deux autres phonologiques décrites comme suit :

i. La suppression du /t/ en dernière position : c'est une opération morphologique appliquant la règle de suppléance.
ii. L'élision d'un /i/ dans la syllabe médiane ouverte qui est une opération phonologique
iii. Le raccourcissement de /aa/ dans la syllabe médiane fermée, c'est l'opération phonologique permettant de garder toujours une syllabe de la forme /CVVC/ à la fin du mot

A l'issue de ces transformations, cette forme [Faa3iL-at(-un)] devient [Fa3L-a] dans le dialecte égyptien. Nous pouvons aussi citer d'autres exemples de ce type de transformation, avec un tableau d'exemples illustratifs comme suit :

- ✓ Faa3iL-at → Faa3iL-a (par la suppléance [-a ~ -t] en AE),
- ✓ Faa3iL-a → Faa3L-a (par suppression de la haute voyelle en(AE)), et
- ✓ Faa3L-a → Fa3L-a (par raccourcissement de la syllabe fermée en AE).

MSA	**EA**	**Traduction**
jaami3-at(-un)	gam3-a	Une université
naaZir-at(-un)	naZr-a	Un proviseur
qaaniy-at(-un)	sany-a	Une seconde

Cependant, même si cela concerne très peu de mots dans le dialecte, la forme standard [Faa3iL-at (-un)] est conservée, à l'exception du rejet du /t/ final, comme dans /?aanis-at(-un) → ?aanis-a/ 'Demoiselle' utilisé pour s'adresser à une femme non mariée, et /Taalib-at(-un) → Taalib-a/ 'une étudiante'. Ces formes sont particulières à l'élision du /i/. Le

raccourcissement de /aa/ ne sera d'ailleurs plus nécessaire car il n'est pas dans une syllabe médiane fermée.

Le tableau (7.4) présente quelques formes de racines trilitères en MSA possédant deux équivalents en AE : l'un est identique à celui du MSA et l'autre présente une nouvelle forme résultat d'un changement phonologique. Le tableau recense par ailleurs l'équivalent de ces formes dans le dialecte algérien qui en présente qu'une seule.

N°	Forme			Exemple			Traduction
	MSA	EA	AA	MSA	EA	AA	
1	Fi3aaL(-un)	Fi3aaL	F3aaL	hisaab(-un)	hisaab	hssab	Compte
		Fu3aal		hiSaan(-un)	huSaan	hSaan	un cheval
2	Fa3iiL(-un)	Fa3iiL	F3iiL	hadiid(-un)	hadiid	hdiid	Fer
		Fi3iiL		ragiif(-un)	Rigiif	rgiif	une miche
3	Fa3iiL-at(-un)	Fa3iiL-a	F3iiL-a	jariid-at(-un)	gariid-a	jriid-a	Journal
		Fi3iiL-a		daqiiq-at(-un)	di?ii?-a	dqiiq-a	Minute
4	Fa3iyy-at(-un)	Fa3iyy-a	F3iyy-a	Sabiyy-at(-un)	Sabiyy-a	Sbiyy-a	une jeune fille
		Fi3iyy-a		hadiyy-at(-un)	hidiyy-a	hdiyy-a	Cadeau
5	Fu33aL(-un)	Fu33aL	Fu33uL	sukkar(-un)	sukkar	sukkur	Sucre
		Fi33iL	Fa33uL	sullam(-un)	sillim	sallum	Echelle
6	Fu33aaL(-un)	Fu33aaL	Fu33aaL	duxxaan(-un)	duxxaan	duxxaan	Fumée
		Fi33aaL	Fa33aaL	šubbaak(-un)	šabbaak	ϕ	Fenêtre
7	Fi33(-un)	Fi33	Fi33	sinn(-un)	Sinn	sinn	Age
		Fu33	Fa33	qiTT(-un)	?uTT	qaTT	chat (m)
8	Fi33-at(-un)	Fi33-a	Fa33-a	Sihh-at(-un)	Sihh-a	Sahh-a	Santé
		Fu33-a	Fa33-a	qiTT-at(-un)	?uTT-a	qaTT-a	chat gumma
9	Fa3aa?(-un)	Fa3a	F3a	samaa?(-un)	Sama	sma	Ciel
		Fi3a	ϕ	masaa?(-un)	Misa	ϕ	Soirée

Tableau 7. 4. Les formes nominales de racines trilitères en MSA avec deux équivalents en EA et AA

Le tableau (7.4) montre que les radicaux trilitères en (MSA) possédant deux homologues en (EA), excluent toujours le [-t] final suivant une des formes suivantes : [CVCVVC(a)], [CVCVyya], [CVCCVC], [CVCCVVC], [CVCC(a)] et [CVVC]. Nous notons aussi que ces formes subissent des transformations pour produire leur équivalent en AA et EA. Parmi ces transformations nous citons la suppression du /?/ en dernière position pour la forme (9) qui est un changement régulier.

De plus, le tableau (7.4) montre que les autres radicaux trilitères en (MSA) possèdent un seul homologue en arabe algérien (AA), à l'exception des cas (5), (6), (7) et (8). Cette unicité d'équivalence peut s'expliquer par le fait que le dialecte algérien, et de manière globale les dialectes maghrébins, sont caractérisés par la succession de deux consonnes au début du mot, une caractéristique absente dans les dialectes du Machrek, comme l'égyptien.

De même, l'existence de deux équivalents en (AE) pour quelques racines nominales en (MSA) pourrait être expliquée par le processus de « la diffusion lexicale » qui a été définie par (Trask, 1996) comme : « *Le processus par lequel un changement phonologique commence par s'appliquer uniquement à certains mots et se propage ensuite progressivement à d'autres mots phonétiquement similaires. Dans certains cas, la diffusion lexicale s'arrête à un moment donné, laissant tous les autres mots non-affectés en permanence; dans d'autres cas, le processus atteint son terme en affectant tous les mots restants.* ». Autrement dit, la diffusion lexicale est une modélisation des changements phonétiques par et à travers le lexique, qui met l'accent sur sa propagation dans l'espace.

Par exemple, nous pouvons proposer le changement du /a/ en /i/ au niveau des formes MSA a été appliqué dans un premier temps sur un ensemble de quelques mots en dialecte égyptien. Les mots considérés ne commencent pas par une gutturale. Rappelons que les consonnes gutturales comprennent les fricatives vélaires /x/ et /g/, les pharyngales /h/ et /3/ et les laryngales /?/ et / h/. A partir de cette application, nous pouvons faire la généralisation suivante :

$$a \rightarrow i \,/ \begin{bmatrix} +\, cons \\ -\, gutturale \end{bmatrix}$$

La transformation de /a/ à /i/ après des consonnes non gutturales peut être observé dans les mots suivants du modèle [Fa3iiL(-un)] en MSA et leur équivalent en arabe égyptien :

MSA	**EA**	**Traduction**
ragiif(-un)	Rigiif	une miche
jamiil(-un)	Gimiil	une faveur
šariiT(-un)	širiiT	une bande
sariir(-un)	Siriir	un lit
zamiil(-un)	Zimiil	un collègue

Dans le cas des mots contenant des consonnes gutturales, la transformation du /a/ n'a pas lieu d'être en dialecte égyptien, comme c'est montré dans les exemples suivants:

MSA	**EA**	**Traduction**
xabiir(-un)	xabiir	un expert
gasiil(-un)	gasiil	la lessive
habiib(-un)	habiib	un amour
3aSiir(-un)	3aSiir	jus
?amiir(-un)	?amiir	un prince

Toutefois, il existe quelques exceptions où après une gutturale, la transformation du /a/ est réalisée, comme pour le mot commençant par une gutturale /hadiyy-at(-un)→ hidiyy-a/ 'un cadeau'.

Par ailleurs, il est important de souligner qu'au stade actuel de la langue, les noms en (AE) suivant le modèle [Fa3iiL] ne commencent pas tous par une gutturale. A noter que cette forme est transformée en [F3iiL] en dialecte algérien et en dialecte maghrébin.

MSA	EA	Traduction
bariid(-un)	bariid	courrier
mariiD(-un)	mariiD	patient
waziir(-un)	waziir	ministre
daliil(-un)	daliil	guide
safiir(-un)	safiir	ambassadeur

7.1.3. Les racines quadrilitères

Tous les radicaux des noms quadrilitères en MSA sont conservés en arabe dialectal, comme le montre le tableau (7.5).

No	Forme	Exemple	Traduction
1	Fa3L_1aL_2(-un)	kawkab(-un)	Une planète
2	Fu3 L_1u L_2 (-un)	bulbul(-un)	un rossignol
3	Fi3 L_1a L_2 (-un)	dirham(-un)	une unité de la monnaie
4	Fa3 L_1a L_2-at(-un)	falsaf-at(-un)	une philosophie
5	Fa3 L_1a L_2-i(yy-un)	3askar-i(yy-un)	un soldat
6	Fu3 L_1aa L_2 (-un)	fustaan(-un)	une robe
7	Fi3 L_1aa L_2 (-un)	simsaar(-un)	un courtier
8	Fa3 L_1ii L_2 (-un)	kabriit(-un)	les allumettes
9	Fi3 L_1ii L_2 (-un)	?injiil(-un)	la bible
10	Fu3 L_1uu L_2 (-un)	3uSfuur(-un)	un moineau

Tableau 7. 5. Les formes de racines des noms quadrilitères identiques en MSA &EA

Nous remarquons que L_1 et L_2 représente deux consonnes différentes.
Le tableau (7.5) montre que les radicaux des noms quadrilitères, dans les deux variétés possèdent l'un des modèles suivants: [CVCCVC], [CVCCVCa], [CVCCVCiy] et [CVCCVVC].

7.2. Les noms déverbaux

En linguistique, le concept *nom déverbal* désigne les noms dérivés des verbes en opposé aux *noms dénominaux* qui sont issus des noms primaires, qui eux, résultent directement de la racine. Selon (Al-Ghulayaini, 2010), il existe neuf types de noms déverbaux en arabe, chacun d'entre eux correspond à une relation sémantique entre le verbe et le déverbal. Le tableau suivant donne la liste de ces types.

Numéro	Type	Traduction
1	اسم الفاعل	Participe actif
2	اسم المفعول	Participe passif
3	مصدر	Nom verbal
4	اسم المكان	Nom de lieu
5	اسم الزمان	Nom de temps
6	اسم الآلة	Nom d'instrument
7	صفة مشبهة	Adjectif comparatif
8	اسم التفضيل	Adjectif superlatif
9	صيغة المبالغة	Substantif d'exagération

Dans cette section, nous présentons les différents types donnés dans le tableau ci-dessous à l'exception des adjectifs comparatif et superlatif qui seront développés dans le chapitre (Analyse Morphologique des adjectifs).

7.2.1. Les noms verbaux (المصدر)

Dans la grammaire arabe, un nom verbal, communément appelé المصدر *masdar*, est un nom abstrait dérivé à partir de la même racine que le verbe auquel il est associé et exprime le même contenu sémantique que lui. Cependant, il ne réfère à aucune notion de temps, d'aspect, de modalité, de personne, ni même de voix. D'un point de vue sémantique, il véhicule une action, un état ou un processus reflété par le verbe auquel il est associé. Il remplit par ailleurs toutes les fonctions syntaxiques d'un nom.

Les modèles des noms verbaux varient en fonction du type des verbes auxquels ils sont liés. Par exemple, il existe plus de quarante modèles de noms verbaux pour les verbes trilitères primaires. Ces modèles peuvent être parfois liés des sens spécifiques ou des formes de verbes spécifiques. Les modèles standards liés au sens des verbes trilitères primaires sont présentés dans la liste suivante :

a. ***Le modèle [Fi3aaL-at(-un)] {فِعَالَةٌ} :*** ce modèle est utilisé pour faire référence à un métier, souvent manuel, ou une occupation, comme dans زِرَاعَة *ziraa3-at(-un)* 'culture' dérivé du verbe زَرَعَ *zara3(-a)* cultiver, وِلَايَةٌ *wilaay-at(-un)* 'gouvernorat' dérivé du verbe وَلِيَ *waliy(-a)* 'gouverner' et وِلَادَةٌ *wilaadat(-un)* 'naissance' dérivé du verbe وَلَدَ *walad-a* 'naître'.

b. ***Le modèle [Fa3aLaan(-un)] {فَعْلَانٌ} :*** ce modèle est employé pour former des mots exprimant la perturbation et le mouvement (violent ou continu) comme dans غَلَيَانٌ *galayaan(-un)* 'bouillonnement' de غَلَى *galaa* 'bouillir', خَفَقَانٌ *xafaqaan(-un)* 'palpitation' dérivé du verbe خَفَقَ *xafaq(-a)* 'palpiter' et طَيَرَانٌ *Tayaraan(-un)* 'vol' dérivé du verbe طَارَ *Taar(-a)* 'voler'.

c. ***Le modèle [Fi3aaL(-un)] {فِعَالٌ} :*** ce modèle est utilisé pour construire les noms signifiant le refus ou l'abstention, comme dans les mots : ذِهَابٌ *ðihaab(-un)* 'départ' dérivé de ذَهَبَ *ðahab(-a)* 'partir', صِيَامٌ *Siyaam(-un)* 'jeûne' dérivée du verbe صَامَ *Saam(-a)* ' jeûner' et غِيَابٌ *giyaab(-un)* 'absence' dérivé de غَابَ *gaab(-a)* 's'absenter'.

d. ***Le modèle [Fu3aaL(-un)] {فُعَالٌ} :*** ce modèle est utilisé pour désigner les maladies, comme زُكَامٌ *zukaam(-un)* 'rhume' de زُكِمَ *zukim(-a)* 's'enrhumer', صُدَاعٌ *Sudaa3(-un)* 'mal de tête' dérivé du verbe صَدَعَ *Sada3(-a)* 'avoir mal à la tête' et عُطَاسٌ *3uTaas(-un)* 'éternuement' dérivé du verbe عَطَسَ *3aTas(-a)* 'éternuer'.

e. ***Le modèle [Fu3aaL(-un)] {فُعَالٌ} ou [Fa3iiL(-un)] {فَعِيلٌ} :*** ce modèle permet d'avoir des noms exprimant le son, comme pour les mots نُبَاحٌ *nubaah(-un)* 'aboiement' de نَبَحَ *nabah(-a)* 'aboyer', زَئِيرٌ *Za'iir(-un)* 'rugissement' dérivé du verbe زَأَرَ *za'ar(-a)* 'rugir' et صُرَاخٌ *suraax(-un)* 'hurlement' dérivé du verbe صَرَخَ *sarax(-a)* ' hurler'.

f. ***Le modèle [Fu3L-at(-un)] {فُعْلَةٌ} :*** ce modèle est utilisé pour désigner les couleurs , comme حُمْرَةٌ *humr-at(-un)* 'rougeur' de حَمِرَ *Hamir(-a)* 'rougir', صُفْرَةٌ *Sufr-at(-un)* 'jaunisse' de أَصْفَرَ *?aSfar(-u)* 'jaune'. Nous signalons tout de même que ce modèle prend en considération le fait que les adjectifs de couleurs peuvent être traités comme des verbes d'état.

Si un nom verbal ne désigne aucune des significations précédentes, il suivra l'un des six modèles suivants, qui sont le plus communs parmi les modèles des noms verbaux, et cela en fonction de la classe du verbe à partir de laquelle il est dérivé:

a. Le modèle [Fa3L(-un)] {فَعْلٌ} : utilisé pour les verbes d'action transitifs de la forme [Fa3aL(-a)] {فَعَلَ} et [Fa3iL(-a)] {{فَعِلَ}}, comme dans les mots فَتْحٌ *fatH(-un)* 'ouverture' dérivé du verbe فَتَحَ *fataH(-a)* 'ouvrir' et سَمْعٌ *sam3(-un)* 'audition' de سَمِعَ *sami3(-a)* 'entendre'. Ce modèle possède la variante [Fa33(-un)] pour les verbes géminés, par exemple : شَدٌّ *šadd(-un)* 'tir' de شَدَّ *šadd(-a)* 'tirer'.

b. ***Le modèle [Fa3aL(-un)] {فَعَلٌ}*** : il est dédié pour les verbes indiquant un état passager ou intransitifs de la forme [Fa3iL(-a)] {فَعِلَ}, comme pour les mots مَرَضٌ *maraD(-un)* 'maladie' dérivé du verbe مَرِضَ *mariD(-a)* 'être malade' et تَعَبٌ *ta3ab(-un)* 'fatigue' dérivé du verbe تَعِبَ *ta3ib(-a)* 'être fatigué'.

c. ***Le modèle [Fu3uuL(-un)] {فُعُولٌ} :*** ce modèle est employé pour formuler les noms verbaux à partir des verbes intransitifs ayant la forme [Fa3aL(-a)] {فَعَلَ}, comme dans les mots خُرُوجٌ *khuruuj(-un)* 'sortie' de خَرَجَ *kharaj(-a)* 'sortir' et دُخُولٌ *duxuul(-un)* 'entrée' du verbe دَخَلَ *daxal(-a)* 'entrer'.

d. ***Le modèle [Fu3uuL-at(-un)] {فُعُولَةٌ} et [Fa3aaL-at(-un)] {فَعَالَةٌ} :*** utilisés pour dériver des noms à partir des verbes indiquant un état durable intransitifs de la forme [Fa3uL(-a)], comme les mots سُهُولَةٌ *suhuul-at(-un)* 'facilité' de سَهُلَ *sahul(-a)* 'devenir facile' et شَجَاعَة *shajaa3-at(-un)* 'courage' de شَجُعَ *shaju3(-a)* 'devenir courageux'. Vu que les verbes de la forme [Fa3uL(-a)] sont d'habitude des verbes d'état, nous pouvons inférer que les noms issus de la forme [Fu3uuL-at(-un)] sont habituellement des noms abstraits désignant un état.

e. Le modèle [Fi3al(-un)] {فِعَلٌ} : utilisé pour former des noms depuis des verbes intransitifs indiquant un état durable et ayant la forme [Fa3iL(-a)] {فَعِلَ }. Par exemple,

nous avons les mots كِبَرٌ *kibar(-un)* ‘ vieillesse’ dérivé du verbe كَبِرَ *kabir(-a)* ‘vieillir’ et صِغَرٌ *Sigar(-un)* ‘jeunesse’ dérivé du verbe صَغِرَ *sagir(-a)* ‘rajeunir’.

f. ***Le modèle [Fa3aal-at(un)] {فَعَالَةٌ} :*** ce modèle concerne les noms issus des verbes intransitifs indiquant un état durable de la forme [Fa3uL(-a)] {فَعُلَ}, comme pour les mots /naDaafat(-un) [نَظَافَةٌ, propreté] dérivé du verbe نَظُفَ *naDuf(-a)* ‘nettoyer’ et جَزَالَةٌ *jazaalat(-un)* ‘fermeté’ dérivé du verbe جَزُلَ *jazul(-a)* ‘être ferme’.

La plupart des modèles standards des noms verbaux trilitères simples sont conservés dans les dialectes Egyptien et Algérien, comme indiqué dans le tableau (7.6).

N°	Modèle	MSA&EA	AA	Traduction
1	Fa3L(-un)	Darb(-un)	Darb	frappe
2	Fa3L-at(-un)	rahm-at(-un)	rahm-a	miséricorde
3	Fi3L(-un)	hifZ(-un)	hifD	Préservation
4	Fi3L-at(-un)	xidm-at(-un)	Hirfa	Service
5	Fu3L(-un)	šukr(-un)	Sukr	Remerciment
6	Fu3L-at(-un)	ru?y-at(-un)	ru?y-a	Vision
7	Fa3aL(-un)	karam(-un)	Karam	générosité
8	Fa3aL-at(-un)	šafaq-at(-un)	šafaq-a	Compassion
9	Fa3aaL(-un)	fasaad(-un)	Fsaad	Corruption
10	Fa3aaL-at(-un)	Daxaam-at(-un)	Dxaam-a	Immensité
11	Fi3aaL(-un)	hisaab(-un)	hsaab	Calcul
12	Fi3aaL-at(-un)	kitaab-at(-un)	ktaab-a	Ecriture
13	Fu3aaL(-un)	su?aal(-un)	su?aal	questionnement
14	Fa3uuL(-un)	qabuul(-un)	qbuul	Acceptation
15	Fu3uuL(-un)	duxuul(-un)	dxuul	Entrée
16	Fu3uuL-at(-un)	buruud-at(-un)	bruud-a	Froideur
17	Fa3iiL(-un)	rahiil(-un)	rhiil	Départ
18	Fi3Laan(-un)	nisyaan(-un)	nisyaan	oubli
19	Fu3Laan(-un)	gufraan(-un)	gufraan	Pardon
20	Fa3aLaan(-un)	xafaqaan(-un)	xafqaan	palpitation
21	maF3aL(-un)	maqtal(-un)	maqtal	meurtre
22	maF3iL-at(-un)	ma3rif-at(-un)	ma3rif-a	Savoir
23	maFa33-at(-un)	mawadd-at(-un)	mawadd-a	convivialité
24	maFaaL(-un)	manaam(-un)	mnaam	rêve

Tableau 7. 6. Les modèles des noms verbaux trilitères identiques en MSA, EA et AA

Le tableau (7.6) montre que les noms verbaux trilitères possèdent l’un des modèles suivants : [CVCC-at], [CVCVC-at], [CVCVVC-at], [CVCC-aan], [CVCVC-aan], [maCCVC-at], [maCVCC-at] and [maCVVC]. Seulement il existe quelques différences à signaler entre les dialectes (EA et AA en particulier) et le MSA. D’abord, le signe ou la marque de cas (-un) n’existe pas dans l’arabe dialectal et la terminaison (-at) est changée en (-a) par application de la règle de suppléance. Nous signalons aussi la suppression de la première voyelle du radical

des deux modèles (9), (10), (11), (12), (14), (15), (16) et (17) et la suppression de la deuxième voyelle du modèle [Fa3aLaan] dans le dialecte algérien. Ensuite, le modèle [Fa3aLaan] est plus utilisé dans les dialectes qu'en MSA. Ce modèle est utilisé comme forme primaire pour les verbes géminés et des verbes défectueux avec une voyelle en dernière position à la place du modèle [Fa3L(-un)]. Par exemple, le verbe /jarr(-a)/ 'tirer' en MSA possède seulement nom verbal qu'est /jarr(-un)/ alors qu'en Egyptien, ce même mot possède deux possibilités, à savoir /garr/ ou /gararaan/, alors qu'en AA ce même mot possède aussi deux possibilités qui suivent un autre modèle [Fa33aan], à savoir /jarr/ ou /jarraan/.

Nous signalons aussi que pour le cas des verbes défectueux en MSA ayant une voyelle finale, comme pour le verbe /jaraa/ 'courrir', nous avons un seul nom verbal qui est /jary(-un)/, alors qu'en dialecte nous possédons deux possibilités avec quelques différences morphologiques entre l'EA et l'AA : en EA pour le mot précédent nous avons les deux possibilités /gary/ et /garayaan/, suivant les modèles [Fa3L] et [Fa3aLaan] et leurs correspondantes en AA sont /jary/ et /jaryaan/ formatées selon les modèles [Fa3L] et [Fa3Laan].

Sur un autre registre, le dialecte algérien est caractérisé par la fréquence d'utilisation des noms verbaux du modèle [F3iiL]. Ce modèle est déjà utilisé en arabe classique pour beaucoup de verbes indiquant un son ou un mouvement. Ce modèle est très présent au Maghreb appliqué à des verbes indiquant l'idée d'une besogne ou d'une violence physiques comme pour les mots : *hriis* 'action de casser', *khsiil* 'action de laver', *dliik* 'action de masser', etc.

Enfin, le nom verbal respectant le modèle [F3uuLa], s'applique à des verbes qui ont déjà cette forme dans le MSA, comme pour les mots : *bruuda* 'fraîcher', *mlûha* 'degré de salaison'. Ce modèle donne en outre, en dialecte AA, les substantifs abstraits des noms de couleurs, de difformités, etc. Par exemple : *byuuDa* 'blancheur', *khuula* 'noirceur' et *truucha* 'surdité', etc.

Cependant, certains de ces modèles ne sont pas identiques dans les variétés dialectales. Le tableau (7.7) donne un aperçu de ces modèles comme suit :

No	**MSA**		**EA**		**AA**		**Traduction**
	Modèle	Exemple	Modèle	Exemple	Modèle	Exemple	
1	Fa3Laa	da3waa	Fa3La	da3wa	Fa3La	da3wa	appel
2	Fi3aL(-un)	Sigar(-un)	Fu3L	Sugr	Fu3L	Sugr	petitesse
3	Fa3iL(-un)	Dahik(-un)	Fi3L	Dihk	Fa3L	Dahk	rire
4	Fa3iLat(-un)	sariq-at(-un)	Fi3L-a	sir?-a	Fa3L-a	sarq-a	vol
5	FawL(-un)	nawm(-un)	FooL	noom	FooL	Noom	sommeil
6	FayL-at(-un)	gayr-at(-un)	FiiL-a	giir-a	FiiL-a	giir-a	jalousie
7	Fa3aaLiy-at(-un)	karaahiy-at(-un)	Fa3aLiyy-a	karahiyy-a	Fa3aLiyy-a	karahiyy-a	haine
8	Fi3aa	ginaa	Fi3a	gina	F3a	Gna	richesse

Tableau 7. 7. Les modèles des noms verbaux trilitères simples changés phonologiquement en MSA, EA & AA

Le tableau (7.7) montre que les modèles des noms verbaux en MSA issus des verbes primaires trilitères, et qui ont subi des changements phonologiques en dialecte, suivent l'un des modèles suivants : [CaCCaa], [CVCVC-at], [CVCC-at], [CVCaaCiy-at] and [CVCaa].

Certains de ces changements sont phonologiquement réguliers tandis que d'autres ne le sont pas. Sur ces modèles, les changements suivants sont effectués :

a. Le raccourcissement de voyelle finale en (EA et AA) peut être utilisé pour expliquer les changements dans les modèles (1 et 8).
b. Le rehaussement de la voyelle en EA peut être utilisé pour expliquer les changements dans (3 et 4).
c. La diphtongue explique le changement en (5).
d. Le raccourcissement atonique survient dans le modèle (7). Cependant, la géminée y est irrégulière.
e. Le changement en (2) est irrégulier. Celui en (6) l'est aussi puisqu'on s'attendrait normalement à ce que [ay > ee].

Les modèles des noms verbaux quadrilitères dans le MSA et les dialectes sont comparés dans les tableaux (7.8), (7.9) et (7.10).

No	Forme Verbale	Nom Verbal	Exemple	Traduction
I	Fa3L1aL2(-a)	Fa3L1aL2-at(-un)	zaxraf-at(-un)	Décoration
		Fi3L1aaL2(-un)	zilzaal(-un)	tremblement de terre
II	taFa3L1aL2(-a)	taFa3L1uL2(-un)	tadahwur(-un)	Détérioration
IV	(?i)F3aL1aL2L2(-a)	(?i)F3iL1L2aaL2(-un)	(?i)Tmi?naan(-un)	tranquillité

Tableau 7. 8. Les modèles de noms verbaux de verbes quadrilitères en MSA

No	Forme Verbale	Nom Verbal	Exemple	Traduction
Ia	Fa3L1aL2	Fa3L1aL2-a	laxbaT-a	Confusion
Ib	Fa3L1iL2	Fa3L1aL2-a	falfal-a	saisonnement
		Fi3L1aaL2	zilzaal	tremblement de terrer
II	(?i)tFa3L1aL2	Fa3L1aL2-a	margah-a	Balançoire
IV	(?i)F3aL1aL2L2	(?i)F3iL1L2aaL2	(?i)šmi?zaaz	Dégoûtant

Tableau 7. 9. Les modèles de noms verbaux de verbes quadrilitères en AE

No	Forme Verbale	Nom Verbal	Exemple	Traduction
I	Fa3L1aL2	tFa3L1iiL2	txarbiich	Griffonner
		tFa3L1iiL2-a	txarbiich-a	
		Fa3L1aaL2	zalzaal	
II	tFa3L1aL2	tFa3L1iiL2	tkharbiit	
		tFa3L1iiL2-a	tkharbiit-a	

Tableau 7. 10. Les modèles de noms verbaux de verbes quadrilitères en AA

La comparaison des tableaux (7.8), (7.9) et (7.10), montre que le dialecte EA a perdu le modèle des noms verbaux quadrilitères de la forme II, à savoir [taFa3L1uL2]. Sauf pour les emprunts du MSA, EA utilise le modèle de la forme I, [Fa3L1aL2-a], pour certains verbes de la forme II. Pour les autres verbes de cette forme (II), il emploie un nouveau modèle, [tiFa3L1iiL2], qui a été décrit par (Carter, 1996) comme : "*un nouveau modèle de nom verbal est indéniablement en train d'apparaître, comme on le voit dans itfabrik / tifabriik, et il est devenu véritablement productif et est utilisé avec une large gamme radicaux (quadrilittère pur, augmenté, redoublée)* ". Quant au dialecte algérien, le tableau (7.10) montre qu'il a perdu le modèle des noms verbaux quadrilitères de la forme IV, à savoir [(?i)F3iL1L2aaL2(-un)]. Le

même tableau illustre aussi que le modèle des noms verbaux habituel issus des verbes quadrilatères est [tFa3L1iiL2], et nous signalons que les formes des noms verbaux de ces même verbes sont employées dans le MSA sont rarement utilisées.

7.2.2. Les substantifs d'exagération (صيغ المبالغة)

Le substantif de l'exagération est un nom dérivé d'un verbe trilitère transitif afin de désigner une caractéristique d'une personne. Cette caractéristique concerne la personne ayant effectué l'action du verbe dont elle est issue, et met en exergue la répétitivité, l'abondance et l'exagération dans la réalisation de l'action concernée. Ce substantif est obtenu via cinq modèles de dérivation, valables à la fois dans le MSA et l'arabe dialectal. Ces modèles sont :

1. [***Fa33aaL(-un)] de la forme classique (فَعَّالٌ) :*** ce modèle est utilisé, dans le MSA et les variétés dialectales, pour exprimer en général les noms de métier. Par exemple, /kaððaab(-un) > kaddaab/ 'un menteur, qui ment souvent', /xabbaaz(-un) > xabaaz/ 'un boulanger'. Dans ce modèle, les diphtongues /aw/ et /ay/ sont conservés, par conséquent pas changées en /oo/ et /ee/ respectivement, dans l'arabe dialectal. Cette conservation est justifiée par le fait que la glide en question constitue une partie d'une consonne géminée et est donc soumise à l'inaltérabilité géminée, et aussi en raison de la production des diphtongues uniquement à la fin des radicaux. Par exemple, /xawwaaf/ 'peureux', /šayyaal/ 'un porteur'.
2. ***[mi-F3aaL(-un)] de la forme classique (مِفْعَالٌ) :*** comme dans /mi-qwaal(-un) > mi-qwaal/ 'parleur'.
3. ***[Fa3uuL(-un)]de la forme classique (فَعُولٌ):*** comme dans /šakuur(-un)/ 'qui remercie beaucoup' et /katuum(-un)/ 'dissimulé'.
4. ***[Fa3iiL(-un)] de la forme classique (فَعِيلٌ) :*** comme dans /samii3(-un)/ 'auditeur' et /xabiir(-un)/ 'expert'.
5. ***[Fa3iL(-un)] de la forme classique (فَعِلٌ) :*** comme dans /farih(-un)/ 'joyeux' et /yaqiZ(-un)/ 'éveillé'

Ces modèles sont aussi appliqués pour dériver des substantifs d'exagération à partir verbes trilitères dérivés. Cependant, ces cas restent limités et considérés comme des exceptions, à l'instar des mots /bašiir(-un)/ 'annonceur de bonnes nouvelles' de /baššar(-a)/ 'annoncer une bonne nouvelle' /naðiir(-un) > naziir/ 'avertissement' de /?anðar(-a)/ 'avertir' et /mi-qdaam(-un)/ 'courageux' de /?aqdam(-a)/ 's'aventurer'.

7.2.3. Les noms de lieu et de temps (أسماء المكان والزمان)

Les noms de lieu et de temps sont des déverbaux d'un verbe utilisés pour désigner, respectivement, le lieu et le temps de survenue de l'action indiquée par le verbe, en d'autres termes où et quand se produit l'événement exprimé par le verbe. Dans la grammaire arabes, ces déverbaux suivent le même processus de dérivation et ne se distinguent que par le contexte dans lequel ils se produisent. Ils se forment à partir des verbes trilitères primaires dans la forme imperfective, en remplaçant la lettre يَ 'ya' du future par la lettre مَ 'ma'. Nous soulignons par ailleurs que ces noms ne sont généralement pas issus de verbes quadrilitères que ce soit dans le MSA ou dans l'arabe dialectal. Ces noms sont dérivés selon l'un des modèles suivant (Al-Toma, 1969) :

1. ***[ma-F3aL(-un)] de la forme مَفْعَلٌ :*** ce modèle est utilisé pour les verbes ayant une voyelle /a/ ou /u/ dans leurs radicaux en (MSA). Cette forme est aussi celle qui est utilisée pour former ces noms dans le dialecte. Techniquement, les noms de ce modèle sont obtenus à partir de la forme imperfective du verbe, en changeant le préfixe (ya)

avec la séquence (ma) et si la voyelle de l'avant dernière lettre est (u) on la change en (a). Comme le montre les exemples ci-dessous :

Verbe imperfectif	Nom de lieu/temps	Traduction
ya-3mal-u	ma-3mal(-un)	un laboratoire
ya-Sna3-u	ma-Sna3(-un)	une usine
ya-ktub-u	ma-ktab(-un)	un bureau
ya-dxul-u	ma-dxal(-un)	une entrée

Il convient de signaler que la grammaire arabe exclut de cette règle les douze noms suivant, qui changent la voyelle (u) de la forme imperfective en avant dernière position en (i) à savoir : /ma-sjid(-un)/ 'mosqué', /mašriq(-un)/ 'l'orient', /maghrib/ 'le couchant, le Maghreb', /marfiq/ 'le lieu où l'on appuie ses coudes', /manbit/ 'le lieu où l'on recueille les herbes', /masqiT/ 'le lieu de la chute', /mankhir/ 'le lieu de la respiration, les narines', /mansik/ 'le lieu de sacrifice', /maTli3/ 'le lieu de l'ascension', /mafriq/ 'le lieu où les cheveux se séparent sur la tête', /maskin/ 'habitation', /majzir/ 'l'endroit où l'on égorge les chameaux'.

Cependant, il existe une exception pour le nom de lieu /ma-xzan/ 'un entrepôt' qui est dérivé d'un verbe ayant /i/ comme voyelle imperfective en avant dernière position. Ce verbe est /ya-xzin(-u) > yi-xzin/ 'stocker'.

Pour *les verbes à consonne double*, la variante de [ma-F3aL(-un)] est [ma-Fa33(-un)]. Cette dernière est obtenue en appliquant la règle générale de métathèse de la consonne double, comme dans les mots : /ma-marr(-un)/ 'un passage' et /ma-hall(-un)/ 'un magasin'.

Pour *les verbes avec voyelle médiane*, les noms de lieu et de temps suivent le modèle [ma-FaaL(-un)] dérivé de [ma-FwaL(-un)] pour les verbes en 'w' et [ma-FyaL(-un)] pour ceux en 'y' et ce à cause de l'assimilation vocoïde anticipative ex. /ma-kaan(-un)/ 'une place' et /ma-Taar(-un)/ 'un aéroport'.

Pour *les verbes avec voyelle finale*, le noms de lieu et de temps suivent le modèle [ma-F3a(n) > ma-F3a] issu [ma-F3ay(-un)] par application de l'élision de la glide et l'assimilation vocoïde persévérative, suivi du raccourcissement de la syllabe fermée en (MSA) ou du raccourcissement de la voyelle en position finale en dialecte. Par exemple, le mot /ma-lha(n) > ma-lha/ 'un cabaret'.

2. ***[ma-F3iL(-un)] (مَفْعِلٌ) :*** ce modèle est utilisé pour les verbes ayant une voyelle initiale faible, et les verbes sonore ayant la voyelle /i/ comme voyelle du radical à la forme imperfective en (MSA). Ce même modèle existe dans la variante dialectale aussi avec quelques modifications dans certains dialectes, comme c'est le cas dans le dialecte algérien où la voyelle (i) est changée en (a).

Verbe imperfectif MSA	Nom de lieu/temps	Traduction
ya-3id(-u)	ma-w3id(-un)	un rendez-vous
ya-rid(-u)	ma-wrid(-un)	une source
ya-hbiT(-u)	ma-hbiT(-un)	piste de descente
ya-jlis(-u)	ma-jlis(-un)	un conseil

Dans certains cas, l'AE utilise le modèle [ma-F3aL] pour les noms de lieu et de temps qui suivent en MSA le modèle [ma-F3iL(-un)], comme pour le mot /ma-wqif(-un) > ma-w?af/ 'un arrêt, une station'. Dans d'autres cas, dans le même dialecte, nous utilisons le modèle [ma-F3iL] à la place de [ma-F3aL(-un)]. Par exemple, le mot /ma-rkab(-un) > ma-rkib/. Le choix de la vocalité dans ces modèles en (AE) semble arbitraire (Gadalla, 2000). Dans les verbes avec une voyelle médiane faible, le modèle [ma-F3iL (-un)] a la variante [ma-Fiil (-un)] de [ma-FGiL (-un)] obtenu par l'assimilation vocoïde anticipative dans les deux variétés, MSA et dialecte, comme dans le mot /ma-Siir(-un)/ 'un destin'. Par ailleurs, il est important de signaler, que cette forme [ma-F3iL(-un)] est inexistante dans le dialecte algérien, sauf dans l'est algérien, où ils utilisent dans certain cas la voyelle /i/ comme dans le dialecte tunisien contrairement aux autres régions, où il n'utilise pas la voyelle /i/ pour transcrire le schewa qui correspond à la lettre /a/ en dialecte algérien.

3. ***[ma-F3aL-at(-un)] (مَفْعَلَة) :*** ce modèle est souvent appliqué aux verbes ayant la voyelle /u/ dans le radical imperfectif en (MSA). Il représente les noms qui prennent le suffixe [-at]. Il est également emprunté en dialecte, même si la voyelle du radical est changée en /i/. Les dialectes maghrébins effectuent quelques modifications sur ce modèle en supprimant la voyelle du radical afin de former le modèle [ma-F3L-at]. Selon (Thackston, 1984), le genre du nom de lieu /temps, suivant les modèles [ma-F3aL(-un)] ou [ma-F3aL-at(-un)], est imprévisible. Voici quelques exemples illustratifs :

Verbe imperfectif MSA	**Nom de lieu/temps**	**Traduction**
ya-drus(-u)	ma-dras-at(-un)	une école
ya-ktub(-u)	ma-ktab-at(-un)	une librairie
ya-qbur(-u)	ma-qbar-at(-un)	un cimetière
ya-hkum(-u)	ma-hkam-at(-un)	un tribunal

Pour les verbes géminées, la variante de ce modèle est [ma-Fa33-at(-un)] issue du modèle [ma-Fa33-at(-un)] par application de la métathèse la consonne double. C'est le cas du mot مَحَطَّةٌ *ma-hatt-at(-un)* 'une station'. En ce qui concerne les verbes défectueux, le nom de lieu est formé selon le modèle [ma-Faal-at(-un)] issue du schème [ma-FGaL-at(-un)] par assimilation vocoïde anticipative, comme dans le mot مَغَارَةٌ *ma-gaar-at(-un)* 'une cave'.

4. ***[ma-F3iL-at(-un)]:*** ce modèle concerne les verbes ayant un /i/ en avant dernière position dans le radical défectueux en (MSA). Ce modèle est aussi emprunté dans la variante dialectale avec parfois quelques modifications. C'est le cas du dialecte algérien, et plus généralement les dialectes maghrébins, où la voyelle /i/ est supprimée.

Verbe imperfectif MSA	**Nom de lieu/temps**	**Traduction**
ya-nzil(-u)	ma-nzil-at(-un)	un rang
ya-nTiq(-u)	ma-nTiq-at(-un)	une zone

Les noms de lieu et de temps qui sont issus des verbes trilitères dérivés ont la forme [*mu-...aC*] mais ils sont moins utilisés que ceux des verbes primaires, particulièrement dans les dialectes. Les noms de ce modèle sont obtenus à partir de la forme imperfective du verbe, en changeant le préfixe يَ *'ya'* avec la séquence مُ *'mu'* avec la voyelle (a) dans l'avant dernière lettre du radical. Voici quelques exemples :

Verbe imperfectif MSA	Nom de lieu/temps	Traduction
ya-jtami3(-u)	mu-gtama3(-un)	une société
ya-stahill(-u)	mu-stahall(-un)	l'entame
ya-stawdi3(-u)	mu-stawda3(-un)	un dépôt
ya-stašfii	mu-stašfa(n)	un hôpital

7.2.4. Les noms d'instruments (أسماء الآلة)

Le nom d'instrument est un nom issu d'un verbe et utilisé pour désigner l'instrument par lequel l'action du verbe est effectuée, en d'autres termes, c'est un nom qui désigne l'instrument dont on se sert pour exécuter l'action exprimée par le verbe. En général, dans la grammaire arabe, il existe quatre modèles standards pour ce type de noms : [mi-F3aL(-un), مِفْعَلٌ], [mi-F3aL-at(-un), مِفْعَلَةٌ], [mi-F3aaL(-un), مِفْعَالٌ] et [Fa33aaL-at(-un), فَعَّالَة]. Ces modèles sont également présents dans l'arabe dialectal avec quelques modifications, par exemple le dialecte égyptien remplace le préfixe de l'instrument [mi-] dans les deux premiers modèles par [ma-], et dans le troisième modèle, dans certains cas, par [mu-]. Le dialecte Algérien possède aussi ses particularités pour les noms d'instruments, en effet il remplace le préfixe [mi] dans les différents modèles par [ma]. Le tableau ci-après donne quelques exemples comparatifs des noms d'instrument :

MSA	EA	AA	Traduction
mi-dfa3(-un)	ma-dfa3	ma-dfa3	un canon
mi-jhar(-un)	ma-ghar	ma-jhar	un microscope
mi-Syad-at(-un)	ma-Syad-a	ma-Syd-a	un traquenard
mi-nqal-at(-un)	ma-n?al-a	ma-nql-a	un rapporteur
mi-qyaas(-un)	mi-'yaas	ma-qyaas	une mesure
mi-ftaah(-un)	mu-ftaah	ma-ftaah	une clef
tallaag-at(-un)	tallaag-a	Tallaaga	un réfrigérateur
gassaal-at(-un)	gassaal-a	gassaal-a	une machine à laver

Par ailleurs, les modèles des noms d'instrument pour les verbes géminés sont différents de ceux présentés ci-dessus. Pour la variante du modèle [mi-F3aL(-un)] (exprimé en EA et AA par [ma-F3aL]) est [mi-Fa33(-un)] (exprimé en EA par [ma-Fa33], et en AA par [m-Fa33]). Cette forme est obtenue par l'application de la règle de métathèse de la consonne double au modèle d'origine [mi-F3a3(-un) > ma-F3a3] comme dans /mi-qaSS(-un) (MSA) > ma-?aSS (EA) > m-qaSS (AA)/ 'une paire de ciseaux'. Pour les verbes commençant avec une voyelle faible, la variante est [mii3aaL(-un) > mi3aaL] de la forme [mi-G3aaL(-un)] par assimilation vocoïde anticipative dans les deux variétés, puis par raccourcissement atonique en arabe dialectal, ex. /miizaan(-un) > mizaan/ 'une balance'. Pour les verbes se terminant par une voyelle, la variante est [mi-F3aat(-un) > ma-F3a] issue de [mi-F3aG-at(-un)] par élision de la glide dans les deux variétés, puis par raccourcissement de la voyelle finale en AE, ex. /mi-kwaat(-un) > ma-kwa/ 'un fer à repasser'.

7.3. Noms définis vs noms indéfinis

Selon (Aboul-Fetouh, 1969), la définition des noms est une des catégories flexionnelles qui joue un rôle important dans l'inflexion des noms vers une autre catégorie. Cette définition des noms est faite de la même manière que celle des adjectifs en MSA, et elle donne lieu à deux types de noms : les noms définis (اَلْمَعْرِفَةُ) et indéfinis (اَلنَّكِرَةُ). Elle est réalisée par l'intermédiaire d'un morphème préfixé qui est l'article défini [al-] 'le' et assimilée à une initiale c'est à dire une consonne sonore, et ce par la règle de l'assimilation. Il convient de rappeler qu'un arrêt glottal est inséré après la pause par l'insertion de l'arrêt glottal. En arabe dialectal (AE ou AA), nous utilisons une forme réduite de ce marqueur morphologique, à savoir [l-]. Ce marqueur est parfois accompagné dans certains dialectes avec certaines opérations supplémentaires. Par exemple, l'EA effectue deux opérations supplémentaires qui sont : i) l'insertion de la lettre /i/ par l'application de l'opération d'épenthèse au début du mot, et ii) l'ajout de la lettre /?/ par application de la règle d'insertion d'arrêt glottal, le dialecte AA quant à lui effectue les même opérations avec les différences suivantes : 1) pour la première opération, dans certains cas l'AA insère la lettre /e/ au lieu du /i/, et 2) l'opération de l'ajout de l'arrêt glottal est quasi inexistante. En ce qui concerne l'indéfini en arabe, c'est un état qui concerne les mots désignant une chose non précise, par exemple: رَجُلٌ rajul-un 'un homme', كِتابٌ kitaab-un 'un livre'. Cet état est manifesté par quelques indications comme la présence d'un signe diacritique fusionné au signe de la voyelle courte appelé tanwiin, qui est mis uniquement sur les noms singulier et ceux au pluriel interne quel que soit le genre. Nous notons que ce signe diacritique est perdu lorsque le nom est défini, il se termine ainsi par une voyelle courte. Dans l'arabe dialectal le tanwine n'est pas utilisé pour les mots indéfinis du fait que ces dialectes sont caractérisés par la suppression des désinences finales. Ci-dessous des exemples de la définition des noms en MSA et en arabe dialectal (EA et AA) :

MSA	EA	AA	Traduction
(?)al-kitaab(-u)	(?i)l-kitaab	el-ktaab	le livre
(?)al-qamar(-u)	(?i)l-?amar	el-qmar	la lune
(?)aT-Taalib(-u)	(?i)T-Taalib	eT-Taaleb	l'étudiant
(?)aš-šams(-u)	(?i)š-šams	eš-šams	le soleil

Par ailleurs, nous soulignons que les noms peuvent aussi être définis lorsqu'ils sont suivis par l'un des éléments suivants :

- ***Un complément déterminant*** : qui peut être le second nom dans une phrase de construction, ou un pronom possessif suffixé au nom. Dans ce cas, ils ne nécessitent pas un d'article pour être définis. Seul le second nom peut porter l'article défini, même si les deux peuvent être définis. Quelques exemples du MSA: /kitaab-u ?ahmad(-a)/ '*le livre d'Ahmad*', /kitaab-u 1-walad(-i)/ '*le livre du garçon*' et/kitaab-u-hu/ '*son livre*'. Ceci montre que la définition n'est pas très importante en arabe dialectal, que ça soit égyptien ou algérien (maghrébin), car les désinences ont disparu dans cette variété.
- ***Les noms propres*** : sont aussi définis sans ajout d'articles, et ce dans les deux variétés, par exemple : /3adil/ 'Adil' et /Huda/ 'Houda'.

7.4. Le cas de flexion

Les noms arabes standards subissent une déclinaison (la flexion casuelle ou les cas de flexion) à trois cas: nominatif (مَرْفُوع – marfu3), accusatif (مَنْصُوب - mansub) et génitif (مَجْرُور - majrur). Ce marquage diversifié caractérise le système flexionnel des noms en arabe et est réalisé en fonction du rôle du mot dans la phrase. A cette règle nous avons quelques

exceptions et cas particuliers. Par ailleurs la déclinaison est matérialisée d'un point de vue graphique par l'ajout d'un élément à la fin des formes nominales. Cette terminaison est appelée *cas* qui est défini par (Crystal, 1985) comme étant "*Une catégorie grammaticale utilisée dans l'analyse des classes des mots afin d'identifier les différentes relations syntaxiques existants entre les mots d'une phrase*".

Dans la grammaire arabe, il est généralement admis que le cas nominatif est utilisé pour marquer les noms suivants :

- le sujet d'une phrase verbale,
- les deux parties d'une phrase équationnelle,
- le nom du sujet du verbe /kaan(-a)/ 'être' et ses soeurs, et
- le nom de l'attribut ou prédicat de/?inna/ 'effectivement ou en réalité' et ses sœurs.

Quant à l'accusatif, il est utilisé pour indiquer les noms suivants :

- l'objet d'un verbe transitif,
- le nom de sujet de /?inna/ ' effectivement' et ses sœurs, et
- le prédicat du verbe /kaan(-a)/ 'être' et ses sœurs.

Le génitif de son côté marque les noms suivants :

- le deuxième nom dans une construction possessive (le complément d'un nom),
- l'objet d'une préposition, et
- l'objet d'une nominalisation.

Le système de déclinaison et de désinence du MSA est complexe compte tenu du fait que les désinences ne sont pas toujours applicables à tous les noms ou aux mêmes noms dans des conditions différentes. De ce fait, les noms en MSA, tout comme les adjectifs, sont classés en deux groupes selon leur degré de déclinaison : les noms *totalement déclinés* et les noms *semi-déclinés*. Les grammairiens arabes nomment la première classe المصروف *?al-maSruuf(-u)* 'la nunable' et la seconde classe الممنوع من الصرف *?al-mamnuu3-u min As-Sarf(-i)* 'la non-nunable'. Cette distinction est liée au phénomène de التنوين *tanwiin(-un)* 'la nunation' qui fait référence, rappelons-le, à la présence d'un signe diacritique [-n] à la fin d'une désinence.

Les noms totalement déclinés, appelé aussi les triptotes, prennent en MSA les deux ensembles de terminaisons de cas suivants :

1. Ce premier ensemble est constitué des noms définis, incluant le nominatif [-u], l'accusatif [-a] et le génitif [-i],
2. Cet ensemble contient les noms indéfinis, incluant le nominatif [-un], l'accusatif [-an] et le génitif [-in]. En d'autres termes, les éléments de cet ensemble sont ceux où la présence (ou l'absence) de la [-n] est déterminée par le fait que le nom en question est défini ou pas. Dans le cas où ce dernier est indéfini, le suffixe [-n] est forcément utilisé.

Quant aux noms semi-déclinés, ou diptotes, ils se distinguent comme indiqué dessus par l'absence de nunnation. La plupart des noms de cette classe possèdent trois marqueurs dans la forme définie : [-u] pour le nominatif, [-a] pour l'accusatif et [-i] pour le génitif.

Cependant, ils ne possèdent que deux marqueurs dans l'indéfini : [-u] pour le nominatif et [-a] pour l'accusatif et le génitif. Selon (Al-Toma, 1969), cette classe contient les catégories de noms suivantes :

i. ***Pluriels brisés :*** les noms de cette catégorie sont considérés comme des diptotes en raison du fait que ces noms possèdent trois syllabes, dépassant ainsi le nombre maximum de syllabes. Le tableau suivant illustre quelques noms de cette catégorie avec leur modèle :

Modèle MSA	**Exemple**	**Traduction**
Fa3aaL1iL2(-u)	qanaabil(-u)	bombes
Fa3aaL1iiL2(-u)	3aSaafiir(-u)	oiseaux
Fa3aa?iL(-u)	qabaa?il(-u)	tributs
Fawaa3iL(-u)	3awaaTif(-u)	émotions
?aFaa3iL(-u)	?amaakin(-u)	places
?aFaa3iiL(-u)	?ahaadiiθ (-u)	discussions
maFaa3iL(-u)	maSaani3(-u)	usines
maFaa3iiL(-u)	maSaabiih(-u)	lampes
Fu3aLaa?(-u)	šu3araa?(-u)	poètes
?aF3iLaa?(-u)	?aqribaa?(-u)	proches
Fa3aaLaa	Yataamaa	orphelins
Fa3Laa	marDaa	patients

Nous signalons que le dernier modèle [Fa3Laa] est composé de deux syllabes sous la forme de surface, mais il dispose de trois syllabes sous la forme sous-jacente [Fa3Lay(-u)]. Cette forme sous-jacente a subit une élision de glide et une assimilation vocoïde persévérative pour produire [Fa3Laa]. Les pluriels brisés des autres modèles sont considérés comme complètement déclinés et sont traités, par rapport au cas, comme le singulier.

ii. ***Les noms (non radicaux) avec un [-aa ou -aa?] en position finale***, par exemple /bušraa/ 'une bonne nouvelle' et /Sahraa?(-u)/ 'un désert'.

iii. ***Les noms propres d'origine étrangère, ou qui se terminent par [-at], ou un [-aan] non radical ou un [-aa?]*** : cette catégorie contient les noms suivant le modèle [Fu3aL(-u)], les noms des formes composées et la plupart des noms propres féminins. Par exemple, prenons les noms suivants : /?idriis(-u)/, /xaliif-at(-u)/, /marwaan(-u)/, /najlaa?(-u)/,/?ahmad(-u)/, /niyu yoork(-u)/ et /faaTim-at(-u)/.

Concernant les dialectes (AE ou AA), nous signalons la disparition des désinences casuelles (les marqueurs de cas) pour les noms et aussi l'absence de la nunation pour former les noms indéfinis. Nous rappelons que le processus responsable de la disparition des terminaisons de cas dans l'arabe dialectal est connu parmi les linguistes par le nom "perte". Ce terme a été définie par (Langacker, 1977) comme "*la disparition d'un élément ou d'un outil grammatical d'une langue quelconque*". Voici quelques exemples comparatifs :

MSA	**katab-a**	**1-walad-u**	**risaAl-ah**
	écrire.pf-3msg	le-enfant-Nom	lettre
EA	?i1-walad	katab	risaAl-a
	le-enfant	écrire.pf-3msg	lettre
Traduction	“l’enfant a écrit une lettre”		

Certains linguistes, comme (Gadalla, 2000), soulignent que la perte des cas de flexion est l’un des éléments qui a influencé l’ordre des mots dans une phrase en arabe dialectal. Rappelons que le MSA préfère la structure ‘VSO’ (Verbe-Sujet-Objet) à ‘SVO’ (Sujet-Verbe-Objet), l’arabe dialectal quant à lui (que soit ce soit égyptien ou algérien) préfère la structure ‘SVO’ à ‘VSO’. C’est pour cette raison que le sujet en MSA dans l’exemple précédent est déplacé vers le début en dialecte.

7.5. Le genre dans la flexion

Dans la grammaire arabe, quelle que soit la variante standard ou dialectale, deux genres de noms sont considérés : le masculin et le féminin. Ces genres sont appliqués aussi pour les pronoms personnels, relatifs et démonstratifs. Les noms masculins n'ont pas d'indice et sont représentés par un morphème zéro. Tandis que les noms féminins sont souvent marqués par un suffixe. Toutefois, il existe en arabe quelques exceptions, soit de noms féminins qui n'ont pas de suffixe ou de noms masculins ayant des suffixes féminins. Par exemple, les noms qui renvoient sémantiquement aux femelles sont féminines sans suffixes, comme /bint(-un)/ ‘une fille’ et /?uxt(-un)/ ‘une sœur’, tandis que des noms comme /xaliif-at(-un)/ ‘un calife’ et /hamz-at(-u)/ ‘nom propre masculin’ sont au masculin même s’ils se terminent par des suffixes féminins. Dans la plupart des cas, l’arabe dialectal suit l'arabe standard en faisant la distinction de genre.

Traditionnellement, dans la grammaire arabe il existe deux classes de suffixes féminins :

1. ة [-at] : qui est remplacé par [-a(h)] dans les formes pausales. Il convient de signaler que le suffixe [-at] est le moyen le plus utilisé pour indiquer le genre féminin dans les deux variétés, comme nous pouvons le voir dans les exemples suivants:

Masculin	**MSA Fem**	**EA Fem**	**Traduction**
mudarris(-un)	mudarris-at(-un)	mudarris-a	Professeur
jamiil(-un)	jamiil-at-(-un)	gamiil-a	Beau
ϕ	majall-at(-un)	magall-a	Magasine
ϕ	daraj-at(-un)	darag-aa	Degré
nahr(-un)	ϕ	ϕ	Rivière
3alam(-un)	ϕ	ϕ	Drapeau

Les exemples donnés montrent que, dans certains cas, il existe un nom féminin singulier correspondant au nom masculin singulier, mais dans d’autres, les noms féminins ne possèdent aucun équivalent masculin. Idem pour certains noms masculins singuliers n’ayant pas de féminins correspondants.

Sur un autre registre, dans le MSA, le morphème [-at] possède l'allomorphe [-ah] dans la forme pausale des noms. En d'autres termes, il y'a deux allomorphes de terminaison féminine en MSA qui sont : [-ah #] et [-at ailleurs]. En arabe dialectal, la marque du féminin a probablement deux allomorphes: [-a #] et [-t ailleurs]. Cela peut s'expliquer par la règle de "suppléance [-a ~ t] en arabe dialectal". L'allomorphe [-a] est utilisé dans les positions finales, comme dans / madras-a/ 'une école' et /ward-a/ 'une rose' tandis que l'allomorphe [t] est utilisé dans des positions non finales, cette situation se produit par exemple lorsque le substantif féminin est suivi par un nom génitif de définition ou par un suffixe pronominal. Comme dans les exemples suivants : /madras-t il-walad / 'l'école du garçon' et /madras-t-u/ 'son école'.

2. اَء [-aa?] et اَى [-aa] : qui sont tous remplacés par un seul suffixe, la lettre [-a], en arabe dialectal par application de la suppression du /?/ final et du raccourcissement de la voyelle finale. Voici quelques noms dans les deux variétés pour illustrer cette classe :

MSA	EA	Traduction
samaa?(-un)	sama	un ciel
miinaa?(-un)	miina	un port
da3waa	da3wa	une invitation
fatwaa	fatwa	opinion religieuse

Il convient de souligner que le /?/ final est maintenu en AE en position médiane, par exemple /mina?-een <miinaa?-aani/ 'deux ports', ce cas reflète la présence de l'arrêt glottal dans la forme sous-jacente de la marque féminine dialectale. En outre, les noms qui se terminent par [-a < -aa] sont dupliqués en AE par l'utilisation du /t/, par exemple, /fatwit-een <fatway-aani/ 'deux opinions religieuses', ce qui indique que la terminaison [-aa] est remplacée par [-at] dans l'AE. Ces cas sont valables à la forme duale qui est présente en EA et absente dans les dialectes Maghrébins.

Comme indiqué auparavant, il existe certaines exceptions aux règles morphologiques formulées ci-dessus. Par exemple, le MSA propose un nombre important de noms prenant deux genres, par exemple, /ruuh(-un)/ 'une âme', /suuq(-un)/ 'un marché', /kabid(-un)/ 'un foie' et /Tariiq(-un)/ 'une route' (Al-Toma, 1969). Cependant, l'arabe dialectal associe toujours un seul genre pour les noms, même pour les exceptions citées précédemment. Par exemple le nom /suug/ 'un marché' est toujours masculin en AA, et /rooh/ 'une âme' est toujours féminin en EA. Enfin il convient de signaler que certains noms peuvent avoir deux genres différents entre le singulier et le pluriel, c'est le cas par exemple de /burtu?aan-a hilw-a/ 'une orange sucrée' qui est au féminin singulier qui donne un pluriel masculin /burt?aan hilw/ 'des oranges sucrées'.

7.6. Le nombre dans la flexion

Dans la grammaire arabe, selon le nombre des sujets exprimés il existe trois types qui sont : le singulier, le duel et le pluriel. Le singulier est la forme simple non marquée correspondant à un sujet seulement. Le duel et le pluriel sont formés par suffixation. De plus le pluriel peut être aussi obtenu par un changement dans la structure interne du nom. Au niveau des dialectes ces types varient d'un dialecte à un autre, par exemple le dialecte égyptien partage les mêmes caractéristiques que le MSA alors que le dialecte algérien (et maghrébin en général) n'en possède que deux avec une utilisation rare pour quelques cas exceptionnels.

7.6.1. Le singulier

Dans les deux variétés discutées, le singulier est considéré comme étant la forme non-marquée des noms. Les formes du singulier ont été introduites lors des sections précédentes où tous les modèles ont été présentés au singulier.

7.6.2. Le duel

Le duel en MSA est formé en raccordant au singulier un suffixe qui peut prendre l'une des deux formes, selon le cas du substantif dans la phrase : [-aan(i)] au nominatif et [-ayn(i)] à l'accusatif et au génitif. La différence majeure et remarquable entre le MSA et l'arabe dialectal réside dans l'utilisation limitée et restreinte, voire nulle, du duel dans certaines variétés dialectales. Dans l'AE par exemple nous utilisons la forme suivante pour le suffixe du duel: [-een]. Il semble que l'AE n'a maintenu du MSA que [-ayn] qui est devenu [-een] par diphtongue.

En ce qui concerne le dialecte algérien, voir dans les dialectes maghrébins en général, l'emploi du duel est beaucoup plus rare par rapport à l'égyptien. Cette utilisation rare est présente dans les cas suivants :

- Dans la définition des parties doubles du corps comme pour les mots suivants : rajliin 'deux pieds', yaddiin 'deux mains'. Nous signalons ici qu'il existe quelques parties doubles du corps pour lesquelles le duel n'est pas conservé dans la majorité des dialectes maghrébins (l'oranais par exemple le conserve) comme pour zuj sigan 'deux jambes' et zuj rkâbi 'deux genoux'. Les duels des noms de parties doubles du corps sont devenus de ce fait des sortes de pluriels et sont couramment employés comme tels.
- Dans la définition des noms de mesures qui regroupe entre autre les catégories suivantes :
    - *nom de nombre* comme alfiin 'deux mille'
    - *nom de temps* à l'instar de dqiiqtiin 'deux minutes', marrtiin 'deux fois'
    - *noms de longueur* afin d'exprimer des mesure de taille ou de distance comme pour le mot mitartiin 'deux mètres'
    - *nom de poids* comme par exemple : qontaariin 'deux quintaux'

Par ailleurs, le dialecte algérien, ainsi que les autres dialectes maghrébins (tunisien et marocain), expriment en général les formes duales en mettant le mot زوج *zuwj* /zu:dj/ qui

signifie ‘deux’ avant le nom au pluriel. Ce procédé est décliné en tunisien et marocain par l’utilisation du mot زوز *zuwz* /zu:z/ et جوج *juwj* /ju:j/ respectivement.

Les phrases suivantes montrent comment le mot /al-bint(-u)/ ‘la fille’ sont formées en MSA, EA et AA :

Phrase MSA	**?al-bint-aani**	**fi-1-madras-ah**
Décomposition	la-fille-Nom.duel	à-le-école
Phrase EA	?il-bint-een	fi-1-madras-a
Décomposition	la-fille-Nom.duel	à-le-école
Phrase AA	zu:j bnaAt	f-1-madras-a
Décomposition	deux fille-Plur	à-le-école
Traduction	les deux filles sont à l'école	

Il convient de signaler les remarques suivantes concernant le duel des noms :

- Dans les phrase de constructions où un nom duel représente le premier élément de la phrase ou quand il est suivi d’un suffixe pronominal, nous notons que la partie [-ni] de la marque du duel en MSA est supprimée. C’est le cas de la phrase : /kitaab-aa-ni/ ‘deux livres’ s’est vu ôter son [ni] dans /kitaab-(a) l-walad(i)/ ‘les deux livre du garçon’ tout en appliquant le raccourcissement de la syllabe fermée sur /aa/ et /kitaab-aa-hu/ ‘ses deux livres’. En arabe dialectal, cette construction est conservée avec quelques particularités comme l’occultation de la relation possédé-possesseur. Ainsi la phrase précédente devient en égyptien /kitaab-een il-walad(i)/ et /kitaab-een-u/, respectivement. Il existe aussi en EA une autre alternative pour exprimer cette construction. Cette alternative consiste à séparer le duel du nom suivant ou le suffixe pronominal par le pronom possessif /bituu3/ pluriel de /bitaa3/ ‘appartenant à’ et d’ajouter l’article défini [l-] au nom du duel possédé. Le dialecte algérien quant à lui applique aussi la même alternative mais en utilisant le pronom possessif /taa3/ ou /dyaAl/ à la place de /bituu3/ du EA. Ainsi, pour les exemples ci-dessus, nous obtenons avec cette alternative : /?il-kitaab-in *bituu3* il-walad/ (avec réduction de /ee/ en /i/ par raccourcissement de la syllabe fermée) et /?il-kitaab-een *bituu3-u*/ en EA et /zouj ktuub *dyaAl* il-wlad/ et /zouj ktuub *dyaAl-u*/ en AA.

- En MSA, l’allomorphe [-at] du signe du féminin est utilisé avant le signe du duel. Dans l'arabe dialectal l’allomorphe [-t] est utilisée au lieu du [-at]. Par exemple : /madras-at-ayn > madras-t-een/ ‘deux écoles’ et /daqiiq-at-ayn > di?ii?t-een > dqiqt-iin/ ‘deux minutes’.

Enfin, nous signalons que l’AE a une tendance générale à simplifier le processus de formation du nom de duel. Cette simplification est due à la perte des cas où l’AE adopte un seul signe pour le duel qui peut être accordé à n'importe quel nom singulier, contrairement au MSA qui utilise deux signes différents selon les cas de la forme au singulier. Une autre différence en EA, mettant un écart supplémentaire du processus morphologique de la

formation des duels en MSA, est aussi à signaler. Elle réside dans l'expression du duel dans le cas des noms de monnaies, les poids et les commandes d'aliments et de boissons, où le mot /?itneen/ qui signifie 'deux' est mis avant le nom au singulier à l'instar des exemples suivants : /?itneen gineeh/ 'deux livres', /?itneen kiilu/ 'deux kilos', /?itneen kabaab/ 'deux (plats de) kebab' et /?itneen šaay/ 'deux (verres de) thé' (Gadalla, 2000).

7.6.3. Le pluriel

La grammaire arabe dispose de deux types de pluriel, traditionnellement connus par le pluriel *'sain'* (externe, régulier) et le pluriel *'brisé'* (interne, irrégulier). Le dialecte arabe utilise les mêmes types de pluriels utilisés en MSA. Le pluriel sain est formé via suffixation au nom singulier sans faire de changement dans le nom lui-même, que ça soit pour le masculin ou pour le féminin ce qui justifie l'utilisation du terme 'sain'. Quant au pluriel brisé, il est formé par changement interne du modèle de voyelles, autrement dit, il prend un certain nombre de modèles qui introduisent une modification dans la structure interne du nom, justifiant ainsi l'utilisation du terme 'brisé'. Dans le reste de cette section nous détaillons d'avantage ces différents types.

7.6.3.1. Le pluriel masculin sain

Traditionnellement, les grammairiens arabes utilisent deux suffixes selon le cas du nom pour former le pluriel masculin sain des noms. Ces suffixes sont :

- *[-uun(a)] :* utilisé dans le cas nominatif
- *[-iin(a)] :* utilisé dans les deux cas : accusatif et génitif.

Les exemples suivants illustrent l'utilisation de ces flexions avec l'adjectif «fanaAn(-un) » 'acteur'.

Phrase	**jaA-a**	**l-fanaAn-uuna**	**l-Hafla**
Décomposition	venir.pf-3msg	le-artiste-Nom.m.pl	la-fête
Traduction	Les artistes sont venus à la fête		
Phrase	ra?ay-na	l-fanaAn-iina	fi-l-Hafla
Décomposition	voir.pf-3pl	le- artiste-Acc.m.pl	dans-la-fête
Traduction	Nous avons vu les artistes à la fête		
Phrase	taHadaT-naa	ma'a l-fanaAn-iina	fi-l-Hafla
Décomposition	parler.pf-3pl	avec-le- artiste-Gen.m.pl	dans-la-fête
Traduction	Nous avons parlé avec les artistes à la fête		

Pour ce qui est de l'arabe dialectal (AA, EA, etc.), le suffixe [-iin] est généralement le seul suffixe à utiliser pour former le pluriel sain des noms quel que soit le cas du nom considéré (nominatif, accusatif et génitif). Nous montrons dans les exemples suivants, les équivalents dialectaux (en dialecte algérien) des flexions données précédemment.

Phrase	**l-fanaAn-iin**	**jaA-w**	**l-Hafla**
Décomposition	le-artiste.m.pl	venir.pf-3mpl	la-fête
Traduction	Les artistes sont venus à la fête		
Phrase	šuf-naa	l-fanaAn-iin	fi-l-Hafla
Décomposition	voir.pf-3pl	le- artiste.m.pl	dans-la-fête
Traduction	Nous avons vu les artistes à la fête		
Phrase	hdar-naa	ma'a l-fanaAn-iin	fi-l-Hafla
Décomposition	parler.pf-3pl	avec-le- artiste.m.pl	dans-la-fête
Traduction	Nous avons parlé avec les artistes à la fête		

Il convient de rappeler que dans les phrases de constructions où un nom pluriel représente le premier élément de la phrase ou quand il est suivi d'un suffixe pronominal, la partie [-na] de la marque du pluriel en MSA est supprimée. C'est le cas par exemples de, /fanaAn-uuna/ 'des artistes' s'est vu ôter son [na] dans /fanaAn-u l-masraH(i)/ 'les artistes du théâtre' (avec raccourcissement de /uu/ par la règle de raccourcissement de la syllabe fermée) et /fanaAn-uu-ha/ 'ses artistes'.

En arabe dialectal, l'état de la construction est exprimé sans montrer de relation morphologique entre le possédé et le possesseur. Par conséquent, les phrases équivalentes à celle données ci-dessus sont /fanaAn-in l-masraH/ et /fanaAn-in-ha/. Il existe une autre alternative pour l'arabe dialectal qui consiste à séparer le possédé du possesseur par le biais d'un pronom possessif signifiant 'appartenant à' et un article défini ajouté au début du nom possédé au pluriel. Ce pronom en EA est /bituu3/ pluriel de /bitaa3/ et en AA est /taa3/ ou /dyaAl/. Par conséquent, nous exprimons les exemples précédents en dialecte EA et AA respectivement comme suit : /?il-fanaAn-in bituu3 l-masraH/ et /?il-fanaAn-in bitu3-ha/; /?il-fanaAn-in dyaAl l-masraH/ et /?il-fanaAn-in dyaAl-ha/.

Nous signalons aussi une autre différence entre le MSA et le dialecte concernant l'utilisation d'un signe spécial pour le pluriel sain qui ne se trouve pas en MSA. Il s'agit du suffixe [-iyya] qui est employé pour le pluriel d'un petit nombre de noms de métiers se terminant par [-gi] ou [-i]. Par exemple nous avons : قهواجي *makwa-gi* 'repasseur' qui a un pluriel قهواجيّة *makwag-iyya,* حرامي *haraam-i 'un voleur'* qui a un pluriel حراميَّة *haram-iyya*.

7.6.3.2. Le pluriel féminin sain

La formation du pluriel féminin sain des noms en MSA est réalisée par l'ajout du morphème 'ات' [-aat] au singulier féminin après la suppression de la lettre 'ة', ensuite nous ajoutons l'un des deux marqueurs de cas suivants :

- *la désinence [-u(n)] :* qui est utilisée pour le cas nominatif
- *la désinence [-i(n)] :* qui est utilisée à la fois pour les deux cas accusatif et génitif.

Nous rappelons par ailleurs que dans le cas des phrases de construction, la partie [-n] des terminaisons de cas est supprimée avant les suffixes pronominaux et avant le second nom. C'est les cas des exemples suivants :

Phrase	**naAl-at**	**il-mutasaAbiq-aat-u**	**1-jaA'iza**
Décomposition	avoir.pf-3fsg	le-concurrent-fpl-Nom	la-récompense
Traduction	Les concurrentes ont eu la récompense		
Phrase	?a3Tay-naa	il-mutasaAbiq-aat-i	1-jaA'iza
Décomposition	dicerner.pf-3pl	le-concurrent-fpl-Acc	la-récompense
Traduction	Nous avons discerné la récompense aux concurrentes		
Phrase	?a3Tay-naa	1- jaA'iza	li-l- mutasaAbiq-aat-i
Décomposition	dicerner.pf-3pl	la-récompense	au-le-concurrent-fpl-Gen
Traduction	Nous avons discerné la récompense aux concurrentes durant la fête		

Quant au dialecte arabe, EA ou AA, le pluriel féminin sain est formé par l'utilisation du même suffixe [-aat] quel que soit le cas du nom. De plus, aucune terminaison de cas n'est associée à ce pluriel dans le dialecte. Les exemples ci-après illustrent le pluriel féminin sain en arabe dialectal égyptien correspondant aux exemples du MSA donnés précédemment :

Phrase	**?il-mutasaAbiq-aat**	**xad-u**	**1-gaAyza**
Décomposition	le-concurrent-fpl	avoir.pf-3pl	la-récompense
Traduction	Les concurrentes ont eu la récompense		
Phrase	3aT-eenaa	il-mutasaAbiq-aat	1-gaAyza
Décomposition	dicerner.pf-3pl	le-concurrent-fpl	la-récompense
Traduction	Nous avons discerné la récompense aux concurrentes		
Phrase	3aT-eenaa	1-gaAyza	li-l-mutasaAbiq-aat
Décomposition	dicerner.pf-3pl	la-récompense	au-le-concurrent-fpl
Traduction	Nous avons discerné la récompense aux concurrentes durant la fête		

Enfin, nous remarquons que le suffixe [-aat], en plus du fait qu'il est le signe du pluriel par défaut, il est maintenant devenu le signe le plus populaire pour mettre au pluriel les mots empruntés, comme pour les mots /tilifizyoon(-un) > tilifizyoon →tilifizyoon-aat(-un) > tilifizyun-aat/'des télévisions' et /faaks(-un) > faks →faaks-aat(-un) > faks-aat/ 'des fax'.

7.6.3.3. Le pluriel brisé

Le pluriel brisé, nommé aussi 'pluriels internes', implique une modification interne dans la structure du nom singulier, contrairement à la formation du pluriel sain où le nom singulier reste inchangé. Cette modification interne suit toutefois des modèles qui orientent les changements internes du nom singulier. Un compte rendu détaillé de ces modèles est donné dans (Gadalla, 2000). Dans la suite de la section nous présentons trois tableaux, (7.11), (7.12) et (7.13) qui présentent des comparaisons entre le pluriel brisé du MSA et celui du dialecte EA.

Dans le tableau (7.11) nous montrons les modèles des pluriels brisés identiques en MSA et en AE, avec des exemples présents dans les deux variétés comme suit :

N°	Racine	Exemple	Traduction
1	Fu3uL(-un)	sufun(-un) > sufun	navires
2	Fu3aL(-un)	guraf(-un) > guraf	chambres
3	Fi3aL(-un)	minah(-un) > minah	subventions
4	Fi3aL-at(-un)	dibab-at(-un) > dibab-a	ours
5	Fa3aL(-un)	šajar(-un) > šagar	arbres
6	Fa3aL-at(-un)	sahar-at(-un) > sahar-a	magiciens
7	Fu3uuL(-un)	?usuud(-un) > ?usuud	lions
8	Fi3aaL(-un)	hibaal(-un) > hibaal	cordes
9	Fi3aaL-at(-un)	hijaar-at(-un) > higaar-a	pièrres
10	Fu3aat(-un)	quDaat(-un) > quDaah	juges
11	Fu3Laan(-un)	fursaan(-un) > fursaan	chevaliers
12	Fu33aaL(-un)	hurraas(-un) > hurraas	Gardiens
13	taFaa3iL(-u)	tajaarib(-u) > tagaarib	Expériences
14	maFaa3iL(-u)	madaaris(-u) > madaaris	écoles
15	aF3uL(-un)	asTur(-un) > asTur	Lignes
16	aFaa3iL(-u)	akaabir(-u) > akaabir	Personnalité
17	Fawaa3iL(-u)	3awaaSif(-u) > 3awaaSif	Tempêtes
18	Fa3aaL1iL2(-u)	sanaabil(-u) > sanaabil	Epis

Tableau 7. 11. Les modèles de pluriel brisé identiques en MSA et EA.

Le tableau (7.11) montre que les modèles suivant sont utilisés pour former le pluriel brisé dans le MSA et l'arabe dialectal comme l'EA : [CVCVC(a)], [CVCVVC(a)], [CVCVVt], [CVCCaan], [CVCCVVC], [taCVVCVC], [maCVVCVC], [aCCVC], [aCVVCVC] et [CVCVVCVC]. Le tableau (7.12) quant à lui montre les modèles des pluriels brisés en MSA possédant deux homologues en EA : l'une analogue à celle du MSA et l'autre subi un changement phonologique.

	MSA	EA	MSA	EA	Traduction
1	Fa3iiL(-un)	Fa3iiL	3abiid(-un)	3abiid	Esclave
		Fi3iiL	hamiir(-un)	himiir	Anes
2	?aF3aaL(-un)	?aF3aaL	?aShaab(-un)	?aShaab	amis
		?iF3aaL	?amšaaT(-un)	(?i)mšaaT	peignes
3	?aF3iL-at(-un)	?aF3iL-a	?a3mid-at(-un)	?a3mid-a	colonnes
		?iF3iL-a	?argif-at(-un)	(?i)rgif-a	pains

Tableau 7. 12. Les modèles des pluriels brisés en MSA possédant deux homologues en EA

Le Tableau (7.12) montre que les noms pluriels brisés en MSA qui possèdent deux homologues en AE suivent l'un des trois modèles suivants : [CVCVVC], [?VCCVVC] et [?VCCVC-at(-un)]. L'existence de deux homologues en AE pour le modèle MSA [Fa3iiL(-un)] peut être expliqué par le processus de diffusion lexicale. L'existence de [(?i)F3aaL(-un)] comme homologue de [?aF3aaL(-un)] et de [(?i)F3iL-a] en tant que variante de [?aF3iL-at(-

un)] peut être attribué à la perte du [?a-] en AE, suivie par l'épenthèse du /i/ par la règle de l'épenthèse du début du mot et l'insertion du / ?/ par la règle de l'insertion d'un arrêt glottal et ce comme suit :

a. (MSA) ?aCCVVC → CCVVC → iCCVVC → ?iCCVVC (EA).
b. (MSA) ?aCCVC-at →CCVC-a → iCCVC-a → ?iCCVC-a (EA).

Enfin le tableau (7.13) énumère les formes de racines ayant subi des changements phonologiques dans l'arabe dialectal comme suit :

Modèle			**Exemple**		
N°	**MSA**	**EA**	**MSA**	**EA**	**Traduction**
1	Fa3Laa	Fa3La	marDaa	marDa	Patients
2	Fa3aaLaa	Fa3aaLa	Sahaaraa	Sahaara	Déserts
3	Fa3aaLii	Fa3aaLi	karaasii	karaasi	Chaises
4	?aFaa3iiL(-u)	?aFa3iiL	?asaaTiir(-u)	?asaTiir	Mythes
5	Fawaa3iiL(-u)	Fawa3iiL	tawaabiit(-u)	tawabiit	Cercueils
6	maFaa3iiL(-u)	maFa3iiL	mafaatiih(-u)	mafatiih	clés
7	Fa3aaL1iiL2(-u)	Fa3aL1iiL2	3aSaafiir(-u)	3aSafiir	moineaux
8	FiiLaan(-un)	FiLaan	niiraan(-un)	niraan	Feu
9	Fu3aLaa?(-u)	Fu3aLa	šuhadaa?(-u)	šuhada	Martyrs
10	?aF3iLaa?(-u)	?aF3iLa	?anbiyaa?(-u)	?anbiya	Prophètes
11	Fa3aa?iL(-u)	Fa3aayiL	3ajaa?ib(-u)	3agaayib	merveilles
12	Fa3aa?iL-at(-un)	Fa3ayL-a	malaa?ik-at(-un)	malayk-a	anges
13	?aFaa3iL-at(-un)	?aFa3L-a	?asaatið-at(-un)	?asatz-a	professeurs
14	Fu?uuL(-un)	FuuL	ru?uus(-un)	ruus	têtes

Tableau 7. 13. Les formes de racines de pluriel brisé en MSA ayant subi des changements phonologiques dans le dialecte égyptien

Le tableau (7.13) montre que les modèles du pluriel brisé en MSA qui sont morphologiquement modifiés dans l'EA suivent l'un des modèles suivants : [CVCCaa], [CVCVVCaa], [?aCVVC], [CVVC], [CVCVVCVVC], [maCVCVVC], [CVVCaan], [CVCVCaa?], [?aCVCaa?], [CVCVV?VC(a)], et [CV?VVC]. Les règles phonologiques utilisées pour expliquer les changements effectués dans les formes des radicaux nominaux singuliers pourraient également être utilisées pour expliquer les changements qui se produisent dans les modèles pluriels brisés. La liste suivante donne les règles de transformation appliquées par modèles :

- Les modèles (1-3) ont subi un raccourcissement de la voyelle finale.
- Les modèles (4-8) ont subi un raccourcissement atonique.
- Les modèles (9-10) sont obtenu par application de la suppression du /?/ final, puis le raccourcissement de la voyelle finale.
- Le modèle (11) [Fa3aa?il(-u) > Fa3aayiL] est obtenu en faisant une assimilation intervocalique du /?/.

- Le modèle (12) [Fa3aa?iL-at(-un) > Fa3ayL-a] est le résultat de l'application de trois changement phonologiques en plus de la suppression du /t/ final qui est une règle morphologique. Les changements phonologiques appliqués sont :
    - le changement de /?/ en /y/ par assimilation intervocalique du / ?/,
    - l'élision du /i/ par la suppression de la voyelle longue, et
    - le raccourcissement de la voyelle non finale par raccourcissement de la syllabe fermée.

  Ainsi, les changements peuvent être résumés, avec un exemple d'illustration, comme suit :
    - ✓ Fa3aa?iL-at(-un) → Fa3aa?iL-a → Fa3aayiL-a → Fa3aayL-a → Fa3ayL-a
    - ✓ malaa?ik-at(-un) → malaa?ik-a → malaayik-a → malaayk-a → malayk-a

- Le modèle (13), [Fu?uul> FuuL] quant à lui est le fruit de l'application des deux processus phonologiques suivants : l'élision du /?/ et le raccourcissement d'une voyelle extra-longue. Par conséquent, les changements qui se produisent dans ce modèle peuvent être schématisé, avec un exemple explicatif, comme suit :
    - ✓ Fu?uuL → FuwuuL → FuuL
    - ✓ ru?uus → ruuus → ruus

Chapitre 8 Analyse morphologique adjectivale

Introduction

La distinction entre la classe des noms et celle des adjectifs suscite une polémique puisque la majorité des membres de la première classe peuvent bien être utilisés comme des membres de la deuxième classe et vice-versa. Une tentative pour régler cette controverse proposer par (Gadallah, 2000), est de prendre la flexion pour le degré comme un paramètre de distinction entre les adjectifs et les noms: tandis que les adjectifs subissent une flexion pour le degré, (soit morphologiquement soit syntaxiquement) les noms ne la subissent pas.

Ce chapitre traite la morphologie des adjectifs en MSA et AE et AA. Dans la section 8.1, nous présentons les formes des racines adjectivales. La section 8.2 est consacrée pour présenter la différence entre les adjectifs définis et indéfinis. Ensuite, la flexion des adjectifs pour le cas, le genre et le nombre est indiquée en 8 .3, 8.4 et 8.5, respectivement. De plus, un aperçu sur le degré de flexion est donné en 8.6. Finalement, nous présentons les adjectifs relationnels dans la section 8.7.

8.1. Les formes des racines adjectivales

Selon plusieurs chercheurs linguistes, de nombreuses formes de racines adjectivales en MSA sont conservées dans la variété dialectale comme l'EA ou l'AA. Autrement dit, elles sont identiques dans les deux variétés, comme le montre le tableau (8.1).

N°	Stem Forme	Exemple	Traduction
1	Fa3L(-un)	sahl(-un) > sahl	facile
2	Fayy(-un)	hayy(-un) > hayy	vivant
3	Fu33(-un)	hurr(-un) > hurr	Libre
4	Fa3aL(-un)	baTal(-un) > baTal	courageux
5	Fa3aaL(-un)	jabaan(-un) > gabaan	Lâche
6	Fa3uuL(-un)	Sabuur(-un) > Sabuur	patient
7	Fu3aaL(-un)	šujaa3(-un) > šugaa3	brave
8	Fa33aaL(-un)	kaDDaab(-un) > kazzab	monteur
9	Fi33iiL(-un)	sikkiir(-un) > sikkiir	alcoolique
10	Fu33uuL(-un)	qudduus(-un) > qudduus	saint / sacré
11	Fa3Laan(-un)	ta3baan(-un) > ta3baan	fatigué
12	?aF3aL(-u)	?abyaD(-u) > ?abyaD	Blanc

Tableau 8. 1. Les formes de racines adjectivales identiques en MSA, EA & AA

Le tableau (8.1) montre que les formes des racines adjectivales qui sont identiques en AE et MSA suivent l'un de ces six modèles : [CVCC], [CVCVC], [CVCVVC], [CVCCVVC], [CVCCaan], ou [?aCCVC]. En ce qui concerne le dialecte algérien, nous notons que les formes des racines adjectivales suivent l'un des modèles du dialecte égyptien et sont identiques en MSA, sauf qu'il rajoute un nouveau modèle qui est [CCVC], comme dans l'adjectif /fHal/ 'brave', à la place de la première forme qui n'existe pas dans l'AA. Il est important de signaler un point important qui marque la différence entre le MSA et les variantes dialectales au niveau de ces formes, il s'agit de la perte du signe (marque) du cas (-un) dans les formes de l'arabe dialectal.

D'autres formes ont subi des changements phonologiques, dont certains sont réguliers et d'autres irréguliers, comme l'indique le tableau (8.2).

N°	Stem Forme		Exemple			Traduction
	MSA	EA	MSA	EA	AA	
1	Fa3Laa?(-u)	Fa3La	hamraa?(-u)	hamra	Hamra	rouge (f)
2	FayLaa?(-u)	FeeLa	bayDaa?(-u)	beeDa	bayDa	blanche (f)
3	FawLaa?(-u)	FooLa	sawdaa?(-u)	sooda	kahla	noire (f)
4	Faa?iL(-un)	Faayil	gaa?ib(-un)	gaayib	gaayib	absent
5	Fa3?aan(-u)	Fa3yaan	mal?aan(-u)	malyaan	malyaan	rempli
6	Faa3i?(-un)	Faa3i	haadi?(-un)	haadi	haadi	calme
7	Fa3iyy(-un)	Fa3i	ganiyy(-un)	gani	Gani	riche
8	Faa3iL-at(-un)	Fa3L-a	maahir-at(-un)	mahr-a	maahr-a	intelligent (f)
9	?aF3aa	?aF3a	?a3maa	?a3ma	?a3ma	Aveugle
10	Faa3ii	Faa3i	baaqii	baa?i	Baaqi	Restant
11	FawLaan(-u)	FaLaan	jaw3aan(-u)	ga3aan	jii3aan	Faim
12	Fa3iL(-un)	Fi3iL	natin(-un)	nitin	Naatan	Puant
13	Fu3w(-un)	Fi3w	hulw(-un)	hilw	Hlaw	Sucré
14	Fu3yaan(-un)	Fi3yaan	3uryaan(-un)	3iryaan	3aryaan	Nu

Tableau 8. 2. Les formes de racines adjectivales en MSA ayant subi des changements phonologiques en EA &AA

Le tableau (8.2) montre les modèles en (MSA) qui ont subi des changements phonologiques en arabe dialectal (AE et AA). Après l'analyse de certaines formes, nous sommes arrivés au même constat qu'a déduit (Gadalla, 2000). Ce constat est sur le fait que les alternances phonologiques, et même morphologiques, susceptibles de se produire dans les formes de racines adjectivales, sont contrôlées par des règles phonologiques et morphologiques similaires à celles régissant les changements de même nature dans les formes des racines nominales. Ces changements peuvent être expliqués par l'application de certaines règles comme celles décrite ci-après :

- ✓ *La règle morphologique de la "Suppléance [-a~ -t]"* : cette règle caractérise les dialectes et elle est responsable de la disparition de la lettre /t/, utilisée dans les formes se terminant par [-at] dans le MSA, des racines adjectivales en arabe dialectal. La forme (8) représente un cas d'application de cette règle.
- ✓ *Les règles phonologiques* appliquées dans les formes remontées du tableau (8.2) sont :
  - Le signe (marque) de cas (-un) n'existe pas dans les formes dialectales
  - La diphtongue s'applique sur les formes des racines adjectivales (2-3). Ce sont principalement des adjectifs de couleurs et de défauts corporels.
  - L'allongement compensatoire, appelé aussi assimilation intervocalique du /?/, est illustrée dans les formes (4-5)
  - La suppression du /?/ final est présente dans les formes de racines adjectivales (1-3 et 6).
  - Le raccourcissement de la voyelle en dernière position est appliqué dans les formes de racines (1-3 et 9-10). Il est aussi appliqué à la forme de racine (7) étant donné que la syllabe /iy/ est équivalente à /ii/ d'un point de vue phonologique.

Les formes des racines (11-14) ont subi des changements phonologiques irréguliers. Etant irréguliers, les raisons de ces changements restent inconnues. Par ailleurs, nous remarquons que certaines formes de racines adjectivales en MSA subissent deux changements phonologiques lors du passage en dialecte AE et AA. Parmi les formes subissant ce type de changements nous signalons les formes (2-3). En effet, la forme de racine adjectivale [Faa3iL-at(-un)] subi les mêmes changements que subit son homologue nominale, à savoir la suppression du /t/ du suffixe[-at] par la règle morphologique de suppléance, la suppression de la voyelle /i/ par syncope, et le raccourcissement du /a/ résulte de l'interdiction d'avoir des syllabes médianes super lourdes. Nous notons que le troisième changement n'est appliqué que dans le cas du dialecte EA. Cependant le dialecte AA autorise les syllabes médianes super lourdes comme le MSA. Voici d'autres exemples illustratifs :

MSA	**EA**	**AA**	**Traduction**
maalih-at(-un)	malh-a	maalh-a	salée
saabir-at(-un)	Sabr-a	Saabr-a	patiente
naajih-at(-un)	nagh-a	naajh-a	réussie
baarid-at(-un)	bard-a	baard-a	froide
saabi3-at(-un)	sab3-a	saab3-a	septième (f)

Sur un autre registre, il existe quelques formes de racines adjectivales en MSA possédant plus qu'un seul équivalent en AE et AA : l'un est semblable à l'original en MSA et les autres ont subi des changements phonologiques, comme le montre le tableau (8.3).

N°	**Forme**			**Exemples**			**Traduction**
	MSA	**EA**	**AA**	**MSA**	**EA**	**AA**	
1	Fa3iiL(-un)	Fa3iiL	F3iiL	rahiim(-un)	rahiim	rhiim	miséricordieux
		Fi3iiL	ϕ	kabiir(-un)	kibiir	ϕ	Grand
		Fu3ayyaL	F3ayyaL	qaSiir(-un)	?uSayyar	qSayyar	Court
		Fu3ayyiL	F3ayyiL	qaliil(-un)	?ulayyil	qlayyil	Peu
		ϕ	F3uyyuL	ϕ	ϕ	Sguyyur	Petit
2	Faa3iL(-un)	Faa3iL	Faa3aL	jaamid(-un)	Gaamid	gaamad	Solide
		Fu3L	F3uun	saaxin(-un)	suxn	sxuun	Chaud
3	Faa33(-un)	Faa33	Faa33	jaaff(-un)	Gaaff	jaaff	Sec
		Fa33	Fa33	haarr(-un)	harr	haarr	Chaud
4	FayyiL(-un)	FayyiL	FayyiL	jayyid(-un)	gayyid	jayyid	Bon
		FayyaL	FayyaL	Dayyiq(-un)	dayya?	Dayyaq	Etroit

Tableau 8. 3. Les formes de racines adjectivales en MSA avec plus d'un équivalent en EA & AA

Selon (Gadalla, 2000) trois formes de racines adjectivales sont spécialement utilisées dans le dialecte égyptien : [Fa33iiL], comme dans /hassiib/ 'doué en comptabilité', [Fa33uuL], comme /dalluu3/ 'gâté' ou /habbuub/ 'beaucoup aimé', et [Fa3aL], ex. /šala?/ 'rude'. Le tableau (8.3) montre que la forme de la racine adjectivale en MSA [Fa3iiL(-un)] est conservée dans certains adjectifs en dialecte. Il montre également que cette forme est phonologiquement modifiée dans d'autres. Néanmoins, il n'existe pas de règles définies

concernant la conservation de cette forme ou son altération en d'autres formes de racine nominales. Par ailleurs, et de manière générale, la forme [Fa3iiL(-un)] est conservée dans les adjectifs commençant par une consonne gutturale, comme c'est le cas pour les exemples suivants :

MSA	EA	AA	Traduction
xafiif(-un)	xafiif	xfiif	léger
gariib(-un)	gariib	griib	étrange
hadiiT(-un)	hadiis	hdiiT	moderne
3ariiD(-un)	3ariiD	3riiD	Large

Ces exemples illustrent que le dialecte algérien est caractérisé par la perte de la voyelle /a/ de la première consonne, même pour les adjectifs commençant par une consonne gutturale. Par conséquent la forme est devenue [F3iiL] en AA, alors qu'elle est [Fa3iiL] en dialecte égyptien et en MSA. Les adjectifs de la forme [Fa3iiL(-un)] qui ont été conservés en AE, avec la perte de la dernière voyelle (-un), ne commencent pas tous forcément par une consonne gutturale, comme c'est le cas des adjectifs suivants:

MSA	EA	AA	Traduction
jamiil(-un)	gamiil	Gmiil	beau
kariim(-un)	kariim	Kriim	généreux
našiiT(-un)	našiiT	nšiiT	actif
rahiim(-un)	rahiim	Rhiim	miséricordieux

D'autre part, le tableau ci-après montre des exemples d'adjectifs dont la forme [Fa3iiL(-un)] devient [Fi3iiL] dans le dialecte égyptien, en raison de changements phonologiques :

MSA	EA	AA	Traduction
jadiid(-un)	gidiid	jdiid	nouveau
kabiir(-un)	kibiir	kbiir	grand
naZiif(-un)	niDiif	nDiif	propre
raxiiS(-un)	rixiiS	rxiiS	pas cher
samiin(-un)	simiin	smiin	gros
ba3iid(-un)	bi3iid	b3iid	loin
šadiid(-un)	šidiid	šdiid	fort
qaqiil(-un)	ti?iil	Tqiil	Lourd

Les exemples précédents montrent que la forme en MSA [Fa3iiL(-un)] possède un seul homologue en arabe algérien (AA) qui est [F3iiL]. Cette unicité d'équivalence peut s'expliquer par le fait que le dialecte algérien, et de manière globale les dialectes maghrébins, a une caractéristique marquante qui le distingue à la fois du MSA et des dialectes orientaux, représenté par l'égyptien dans notre étude. Cette caractéristique est le fait de commencer des mots par des consonnes 'neutres', sans voyelles et avec un sukun. Ceci est concrétisé aussi par la succession de deux consonnes au début du mot ce qui n'existe pas dans les dialectes orientaux. Les exemples montrent aussi l'existence de deux équivalents à la forme en (AE) pour la forme [Fa3iiL(-un)] en MSA. Ce fait pourrait être expliqué par l'application du processus de «la diffusion lexicale». Il est aussi à signaler que dans les adjectifs désignant la

rareté ou la petitesse, la forme standard [Fa3iiL(-un)] est changée en [Fu3ayyaL] lors du passage en AE, et dans très peu de cas en [Fu3ayyiL]. Le mot /nahiif(-un)/ 'mince' constitue une exception puisqu'il est conservé sans changements en AE. Dans la même optique, la même forme est changée parfois en [F3ayyaL] ou [F3uyyuL] lors du passage en AA. Nous illustrons ce cas par les exemples suivants :

MSA	EA	AA	Traduction
qaSiir(-un)	?uSayyar	qSayyar	court
Sagiir(-un)	Sugayyar	Sgayyar	petit
		Sguyyur	
qaliil(-un)	?ulayyil	Qluyyul	peu

8.2. Adjectifs définis vs indéfinis

Nous rappelons, que dans la grammaire arabe, les adjectifs s'accordent en genre et en nombre avec les noms qu'ils qualifient. Il existe deux types d'adjectifs en relation avec le processus de définition : les adjectifs définis et indéfinis. La définition des adjectifs est faite de la même manière que la définition des noms en MSA. Elle est faite par l'intermédiaire de l'article défini [al-] 'le' qui est inséré avant l'adjectif comme étant un proclitique. Cette insertion engendre des phénomènes de transformation morphologique qui affectent les mots en fonction de la nature de leur lettre initiale. Autrement dit, si le mot contient l'article [al-] (ال), il faut faire la distinction entre les lettres «solaires» et les lettres «lunaires». Pour les lettres solaires, le «l» n'est pas prononcé et la lettre qui le suit est dédoublée à la fois dans la prononciation et dans l'écriture. A l'inverse, dans le cas des lettres lunaires, le «l» de l'article se prononce et la lettre qui le suit n'est pas dédoublée ni dans la prononciation ni dans l'écriture (Saâdane et Semmar, 2013). L'arabe dialectal (AE ou AA) utilise la forme réduite de ce marqueur morphologique, c-à-d. [l-]. Pour l'EA deux opérations supplémentaires sont appliquées : i) insertion de la lettre /i/ par l'application de l'opération d'épenthèse au début du mot, et ii) ajout de la lettre /?/ par application de la règle d'insertion d'arrêt glottal. Le dialecte algérien quant à lui, insère seulement le /e/ par épenthèse du début du mot au lieu du /i/ inséré dans le dialecte EA. Des exemples d'adjectifs définis dans les deux variétés sont donnés dans le tableau suivant :

MAS	AE	AA	Traduction
(?)al-baxiil(-u)	(?i)l-baxiil	l-bxiil	l'avare
(?)al-ganiyy(-u)	(?i)l-gani	l-gani	le riche
(?)aT-Tawiil(-u)	(?i)T-Tawiil	T-Twiil	le long
(?)as-sarii3(-u)	(?i)s-sarii3	s-srii3	le rapide

Concernant les adjectifs commençant par un arrêt glottal /?/ en MSA, l'opération d'épenthèse et l'insertion de l'arrêt glottal ne s'appliquent pas à l'article défini de ce type d'adjectifs dans l'AE et l'AA, comme c'est illustré dans les exemples suivants :

MSA	EA&AA	Traduction
(?)al-?ahmar(-u)	l-ahmar	le rouge
(?)al-?asmar(-u)	l-asmar	le noir
(?)al-?a3raj(-u)	l-a3rag	boiteux
(?)al-?a3maa	l-a3ma	aveugle

Dans la grammaire arabe, les adjectifs peuvent être aussi définis lorsqu'ils décrivent des noms suivis par un complément déterminant. Cependant, bien que les noms, dans ce cas, ne portent pas d'article défini, les adjectifs doivent obligatoirement porter ce marqueur. C'est le cas des exemples en MSA suivants : /kitaab-u Houda (?)al-?ahmar-u/ 'le livre de Houda le rouge', /kitaab-u-haA (?)al-?ahmar-u/ 'son livre (le) rouge'. Les équivalents dialectaux de ces exemples sont /kitaab Houda l-ahmar/ et /ktaab-haA 1-ahmar /, respectivement.

Pour le dialecte AA, il existe un article indéfini qui est constitué du nom de nombre وَاحَد *waahad* 'un', suivi d'un nom défini par l'article défini. Cependant, bien que les noms, dans ce cas, portent d'article défini, les adjectifs doivent obligatoirement ne pas porter ce marqueur. Il est important de souligner que cet article indéfini est invariable pour le masculin, le féminin et le pluriel. Voici quelques exemples illustratifs :

Phrase			**Traduction**
wahd	el-ktaab	?aSfar	Un livre jaune
un	le-live	jaune	
wahd	el-mrâ	Hniina	Une femme douce
Une	la-femme	Douce	
wahd	en-naas	Taybiin	Des gens gentils
Des	les-gens	Gentils	

8.3. Les Cas de flexion

Dans la grammaire traditionnelle arabe, la flexion casuelle d'un mot en général, nommée al-'3raab (الإعراب) en arabe, présente la terminaison qui est liée à sa catégorie grammaticale, voire même sa fonction syntaxique dans la phrase. Cette flexion casuelle peut être une voyelle brève dans le cas du singulier ou une séquence de consonnes et de voyelles dans les cas du duel et du pluriel. Il existe trois cas de flexion : le nominatif, l'accusatif et le génitif.

La déclinaison, appelée aussi la flexion casuelle ou les cas de flexion, des noms en arabe est concrétisée en trois principaux cas : nominatif (مَرْفُوع - *marfuw3*), accusatif (مَنْصُوب - *mansuwb*), génitif (مَجْرُور - *majruwr*). Ces déclinaisons sont faites en fonction du rôle du mot dans la phrase, à l'exception de certains cas particuliers. Elles sont par ailleurs traduites d'un point de vue graphique par un élément adjoint à la fin des formes nominales. De plus, la déclinaison est influencée par la forme du nom (simple 'triptotes' ou diptotes). Les adjectifs, comme mentionné dessus, s'accordent avec les noms qu'ils qualifient dans leur flexion dans ces trois cas. Autrement dit :

- un adjectif serait au nominatif si le nom qu'il qualifie est au nominatif
- un adjectif serait à l'accusatif si le nom est à l'accusatif
- un adjectif serait au génitif si le nom est au génitif

Comme ceci est le cas pour les noms, les adjectifs en MSA sont classés en deux groupes selon leur degré de déclinaison : les adjectifs totalement déclinés et les adjectifs semi-déclinés. Nous les appelons aussi 'triptotes' et 'diptotes', respectivement. Les triptotes prennent deux ensembles de terminaisons : le premier ensemble, employé pour les adjectifs définis, inclut le nominatif [-u], l'accusatif [-a] et le génitif [-i], le second ensemble, employé pour les adjectifs indéfinis, inclut le nominatif [-un], l'accusatif [-an] et le génitif [-in]. Quant aux adjectifs semi-déclinés ou diptotes, ils se distinguent par l'absence de nunnation. Les

adjectifs de cette classe possèdent trois marqueurs dans la forme définie, c.à.d. [-u] pour le nominatif, [-a] pour l'accusatif et [-i] pour le génitif. Cependant, ils ne possèdent que deux marqueurs dans l'indéfini : [-u] pour le nominatif et [-a] pour l'accusatif et le génitif. Les catégories suivantes sont incluses dans cette classe :

i. **Pluriels brisés :** avec l'un de ces modèles présentés dans le tableau ci-après. La raison pour laquelle ces modèles en particulier sont des diptotes est qu'ils possèdent trois syllabes dépassant de ce fait le nombre maximum de syllabes.

MSA Pattern	Exemple	Traduction
fawaa3iL(-u)	?awaa?il(-u)	les premiers
maFaa3iil(-u)	masaakiin(-u)	les pauvres
fu3alaa?(-u)	buxalaa?(-u)	les radins
?aF3ilaa?(-u)	?asqiyya?(-u)	Vilains

ii. **Les adjectifs masculins suivant la forme de racine [Fa3Laan(-u)]** : ces adjectifs prennent au féminin la forme [Fa3Laa], comme dans les cas : /gaDbaan(-u) ⇨ gaDbaa/ 'en colère' et /jaw3aan(-u) ⇨ jaw3aa/ 'avide'. Cependant, si un adjectif possède la forme [Fa3Laan(-un)] ayant le [Fa3Laan-at(-un)], il sera totalement décliné, comme dans /farhaan(-un) ⇨farhaan-at(-un)/ 'heureux'.

iii. **Les adjectifs de la forme [?aF3aL(-u)] :** qui prennent au féminin la forme [Fa3Laa?(-u)] comme dans les cas : /?axDar(-u) ⇨ xaDraaa?(-u)/ 'vert' et /?a3raj(-u) ⇨ 3arjaaa?(-u)/ 'boiteux'.

En ce qui concerne l'arabe dialectal (AE ou AA), la flexion casuelle est caractérisée par la disparition des désinences casuelles (les marqueurs de cas) pour les adjectifs, et n'utilise pas de la nunnation (tanwin) pour former les adjectifs indéfinis.

8.4. Le Genre de la flexion

Comme nous l'avons déjà mentionné, les adjectifs s'accordent avec les noms qu'ils qualifient en genre. Ainsi, deux genres d'adjectifs peuvent être distingués en arabe, MSA ou dialectes, à savoir le masculin et le féminin. Toutefois, il existe quelques adjectifs qui n'ont pas de genre précis surtout dans la variété dialectale. Ces adjectifs incluent entre autre les adjectifs : /baladi/ 'natif' et /miiri/ 'militaire' en plus de certains adjectifs de couleur tels que /burtu?aani/ (EA) et /tchini/ (AA) 'orange', /bunni/ (EA) et /qahwi/ (AA) 'marron' voir /ramaadi/ 'gris'.

Les adjectifs masculins n'ont pas d'indice et sont représentés par un morphème zéro. Tandis que les adjectifs féminins sont souvent marqués par un suffixe. L'arabe dialectal ainsi que le MSA font le même usage de la suffixation afin de former les adjectifs féminins. Traditionnellement, dans la grammaire arabe, trois suffixes féminins sont utilisés pour la formation des adjectifs féminins : ـَة [-at], qui est remplacé par [-a(h)] dans les formes pausales, ـَاء [-aa?] et ـَى [-aa] qui sont tous remplacés par un seul suffixe, la lettre [-a], en arabe dialectal. De plus, le suffixe [-at] est le moyen le plus commun pour former les adjectifs féminins, comme on peut le voir dans les exemples suivants:

Adj Masc	Adj Fem	Traduction
qaadir(-un)	qaadir-at(-un)	capable
jamiil(-un)	jamiil-at(-un)	beau
naa3im(-un)	Naa3im-at(-un)	doux

Tayyib(-un)	Tayyib-at(-un)	gentil/ bon

Nous présentons brièvement dans ce qui suit, les processus de changement de formation des adjectifs féminins et le remplacement des trois suffixes du MSA vers un seul en arabe dialectal. Ces changements sont justifiés par la simplification de la syntaxe de l'arabe dialectal par rapport à l'arabe standard.

- *le [-t] final* du suffixe du féminin singulier dans l'arabe dialectal, l'AE ou l'AA, n'apparait jamais dans les adjectifs.

- *le suffixe [aa?]* est fréquemment utilisé pour les adjectifs féminins en MSA exprimant les couleurs ou les défauts corporels, et dont la forme du masculin est [?aF3aL(-u)]. Dans la variété dialectale, ce marqueur de féminin est en général réduit à [-a] par application de la suppression du /?/ final et du raccourcissement de la voyelle finale. Voici quelques adjectifs dans les deux variétés pour illustrer ce cas :

Masc	**MSA Fem**	**EA Fem**	**Traduction**
?aHmar(-u)	Hamr-aa?(-u)	Hamr-a	rouge
?a3raj(-u)	3arj-aa?(-u)	3arg-a	boiteux
?a3maa	3amy-aa?(-u)	3amy-a	aveugle

- *Le suffixe [-aa]* est utilisé en MSA seulement pour former les adjectifs féminins comparatifs/superlatifs. Par exemple, /?akbar(-u)/ 'plus vieux' possède la forme féminine /kubr-aa/ et /?aSgar(-u)/ 'plus jeune' a la forme /Sugr-aa/. Dans l'arabe dialectal, la forme comparative féminine est abandonnée et le suffixe [-a] est employé. Par exemple : /?al-bint-u l-kubr-aa > ?il-bint il-kibiir-a/ 'la grande fille' et /?al-bint-u S-Sugr-aa > ?il-bint iS-Sugayyar-a/ 'la petite fille'.

- Enfin, nous mentionnons que certains adjectifs masculins en MSA de la forme [Fa3Laan(-u)] possèdent la forme féminine [Fa3Laa], comme pour l'adjectif /3aTšaan(-u)/ 'assoiffé' qui devient /3aTš-aa/ 'assoiffée'. La forme féminine de ce genre d'adjectifs en arabe dialectal est formée en suffixant [-a] à la forme masculine, comme dans /ta3baan/ 'fatigué' qui devient /ta3baan-a/ 'fatiguée'.

8.5. Le Nombre de la flexion

La flexion des adjectifs en arabe est influencée aussi par le nombre en accord avec les nombres des mots qu'ils qualifient : le singulier, le duel et le pluriel. En ce qui concerne la flexion des adjectifs en arabe dialectal (égyptien, algérien, tunisien...), nous constatons la disparition de l'emploi du duel. Par conséquent, le paradigme de flexion du dialecte arabe est pauvre par rapport au MSA et s'appuie seulement sur une inflexion pour deux nombres seulement : le singulier et le pluriel.

8.5.1. Le Singulier

Dans les deux variétés discutées, le singulier est considéré comme étant la forme non-marquée des adjectifs. Tout ce qui a été indiqué dans les sections précédentes se ramène aux adjectifs singuliers.

8.5.2. Le Duel

Les adjectifs en MSA possèdent deux suffixes selon le cas de l'adjectif: [-aan(i)] au

nominatif et [-ayn(i)] à l'accusatif et au génitif. Une différence majeure et remarquable entre le MSA et l'arabe dialectal émane de l'utilisation limitée, voire nulle, du duel dans la dernière variété. En effet, le système morphosyntaxique des dialectes arabes est moins riche que son homologue du MSA. Dans celui des dialectes, nous rappelons, comme mentionné auparavant, la disparition totale du duel pour les adjectifs, les verbes et les pronoms avec une utilisation restreinte du duel pour les noms. Toutefois, il existe une exception dans l'EA qui est l'utilisation de l'adjectif /?asasiyyit-een/ 'essentiels (du)' pour décrire /hagt-een/ 'deux choses'. (Gadalla, 2000) considère cette exception comme un emprunt du MSA et le résultat du mélange du dialecte avec cet emprunt. La forme standard du duel des adjectifs est remplacée par les formes plurielles pour qualifier les noms au duel. Voici quelques exemples illustratifs :

Variante	Phrase			Traduction
MSA	?al-mahal-aani	**kabir-aan**		Les deux magasins sont grands
	le-magasin-Nom.duel	grand-Nom.duel		
EA	?il-mahal-een	**Kubaar**		
	le-magasin-duel	grand.pl.brisé		
AA	Zoudj	mhal-aat	**kbaar**	
	deux-Nbr	magasins-pl	grand.pl.brisé	
MSA	?al-bint-aani	**naajih-at-aan**		Les deux filles ont réussi
	la-fille-Nom.duel	réussie-f-Nom.duel		
EA	?il-bint-een	**nagh-iin**		
	la-fille-duel	réussie-pl		
AA	Zoudj	bn-aat	**nagh-iin**	
	deux-nbr	filles-pl	réussie-pl	

8.5.3. Le pluriel

La langue arabe dispose de deux types de pluriel à la fois pour les noms ou les adjectifs qui sont le pluriel sain (externe) et le pluriel brisé (interne). Le dialecte arabe utilise les mêmes types de pluriels utilisés en MSA. Comme c'est le cas pour les noms, le pluriel sain est formé via suffixation, alors que le pluriel brisé est formé par changement interne du modèle de voyelles. De plus, le MSA possède deux genres pour le pluriel sain des adjectifs, en l'occurrence le masculin et le féminin alors que l'arabe dialectal (AE ou AA) n'en possède qu'un seul genre qui est le pluriel sain.

8.5.3.1. Le Pluriel Masculin Sain

Les grammairiens arabes traditionnels utilisent deux suffixes selon le cas de l'adjectif pour former le pluriel masculin sain des adjectifs. Ces suffixes sont :

- *[-uun(a)] :* utilisé dans le cas nominatif
- *[-iin(a)] :* utilisé dans les deux cas : accusatif et génitif.

Les exemples suivants illustrent l'utilisation de ces flexions avec l'adjectif **« ma3ruf(-un)** » 'célèbre'.

Phrase	**jaA-a**	**l-fanaAn-uuna**	**l-ma3ruf-uuna**	**l-Hafla**
Décomposition	venir.pf-3msg	le-artiste-Nom.m.pl	le-célèbre-Nom.m.pl	la-fête
Traduction	Les artistes célèbres sont venus à la fête			
Phrase	ra?ay-na	l-fanaAn-iina	**l-ma3ruf-iina**	fi-l-Hafla
Décomposition	voir.pf-3pl	le- artiste-Acc.m.pl	le-célèbre-Acc.m.pl	dans-la-fête

Traduction	Nous avons vu les artistes célèbres à la fête			
Phrase	taHadaT-naa	ma'a l-fanaAn-iina	**l-ma3ruf-iina**	fi-l-Hafla
Décomposition	parler.pf-3pl	avec-le- artiste-Gen.m.pl	le--Gen.m.pl	dans-la-fête
Traduction	Nous avons parlé avec les artistes célèbres à la fête			

Pour ce qui est de l'arabe dialectal, le suffixe [-iin] est généralement utilisé pour former le pluriel sain des adjectifs dans tous les cas (nominatif, accusatif et génitif). Nous montrons dans les exemples suivants, les équivalents dialectaux (en dialecte algérien) des flexions données précédemment.

Phrase	**l-fanaAn-iin**	**l-ma3ruf-iin**	**jaA-w**	**l-Hafla**
Décomposition	le-artiste-Nom.m.pl	le-célèbre-Nom.m.pl	venir.pf-3msg	la-fête
Traduction	Les artistes célèbres sont venus à la fête			
Phrase	šuf -naa	l-fanaAn-iina	**l-ma3ruf-iina**	fi-l-Hafla
Décomposition	voir.pf-3pl	le- artiste-Acc.m.pl	le-célèbre-Acc.m.pl	dans-la-fête
Traduction	Nous avons vu les artistes célèbres à la fête			
Phrase	hdar-naa	m'a l-fanaAn-iina	**l-ma3ruf-iina**	fi-l-Hafla
Décomposition	parler.pf-3pl	avec-le- artiste-Gen.m.pl	le-célèbre-Gen.m.pl	dans-la-fête
Traduction	Nous avons parlé avec les artistes célèbres à la fête			

8.5.3.2. Le Pluriel Féminin Sain

La formation du pluriel féminin sain des adjectifs en MSA est réalisée par l'ajout du morphème 'ات' [-aat] au singulier féminin après la suppression de la lettre 'ة'. Ensuite nous ajoutons l'un des deux marqueurs de cas suivants :

- *la désinence [-u(n)] :* qui est utilisée pour le cas nominatif
- *la désinence [-i(n)] :* qui est utilisée pour les deux cas accusatif et génitif.

Dans les exemples suivants nous montrons les différentes déclinaisons de l'adjectif « **ma3ruf-at(-un)»** 'célèbre (f)' :

naAl-at	**il-mutasaAbiq-aat-u**	**n-naajih-aat-u**	**1-jaA'iza**	
avoir.pf-3fsg	le-concurrent-fpl-Nom	le-gagnant-fpl-Nom	la-récompense	
Les concurrentes gagnantes ont eu la récompense				
?a3Tay-naa	il-mutasaAbiq-aat-i	n-naajih-aat-i	1-jaA'iza	
dicerner.pf-3pl	le-concurrent-fpl-Acc	le-gagnant-fpl-Acc	la-récompense	
Nous avons discerné la récompense aux concurrentes gagnantes				
?a3Tay-naa	1- jaA'iza	li-l- mutasaAbiq-aat-i	n-naajih-aat-i	fi-1-hafl-ah
dicerner.pf-3pl	la-récompense	au-le-concurrent-fpl-Gen	le-gagnant-fpl-Gen	durant-la-fête
Nous avons discerné la récompense aux concurrentes gagnantes durant la fête				

En ce qui concerne le dialecte arabe, AE ou AA, l'utilisation de la forme des adjectifs féminins pluriels sains a totalement disparue. Toutefois, le suffixe [-iin], utilisé pour former le masculin pluriel sain des adjectifs, est également utilisé pour former le pluriel féminin sain.

De ce fait, le suffixe [-iin] est utilisé en dialecte arabe comme suffixe de pluriel sain, quel que soit le genre, plutôt qu'un simple suffixe de masculin pluriel. Les exemples ci-après illustrent le pluriel féminin en arabe dialectal égyptien correspondant aux exemples du MSA donnés précédemment :

?il-mutasaAbiq-aat	**in-nagh-iin**	**xad-u**	**1-gaAyza**	
le-concurrent-fpl	le-gagnant-pl	avoir.pf-3pl	la-récompense	
Les concurrentes gagnantes ont eu la récompense				
3aT-eenaa	il-mutasaAbiq-aat	in-nagh-iin	1-gaAyza	
dicerner.pf-3pl	le-concurrent-fpl	le-gagnant-pl	la-récompense	
Nous avons discerné la récompense aux concurrentes gagnantes				
3aT-eenaa	1-gaAyza	li-l-mutasaAbiq-aat	in-nagh-iin	fi-1-hafl-ah
dicerner.pf-3pl	la-récompense	au-le-concurrent-fpl	le-gagnant-fpl-Gen	durant-la-fête
Nous avons discerné la récompense aux concurrentes gagnantes durant la fête				

8.5.3.3. Le Pluriel Brisé

Le pluriel brisé des adjectifs est formé par un changement interne de la structure des formes singulières. Selon (Kihm, 2013), ces pluriels, nommés aussi pluriels internes, posent un défi à la théorie morphologique : ils ne mettent pas ouvertement en jeu l'affixation d'un morphème de pluralité, mais semblent plutôt reposer sur un contraste de formes qui, si elles partagent la même racine consonantique, diffèrent, entre autres, pour la vocalisation. La plupart des adjectifs ayant un pluriel brisé dans les variétés, MSA et dialectes, se classent en l'une des catégories suivantes :

1) Les adjectifs de la forme [?aF3aL(-u)] possèdent au pluriel la forme [Fu3L(-un)] dans le MSA et l'AE. En ce qui concerne le dialecte algérien, nous signalons que cette forme a subi une modification sur la racine où une opération de permutation de voyelle entre le 1er et le 2ème radical afin d'obtenir la forme [F3uL]. Par exemple, prenons l'adjectif 2) /?ahmar(-u)/ 'rouge' qui a les pluriels suivants : MSA /humr(-un)/ → EA /humr/ → AA /hmur/. Toutefois les grammairiens arabes définissent trois exceptions en MSA, qui ont été reprises dans le EA, pour les adjectifs de cette catégorie. Cette pluralisation exceptionnelle est due au fait que ces adjectifs possèdent une glide au milieu du radical dans la représentation de base. Ces exceptions sont :
    a. /?abyaD(-u)/ 'blanc' qui possède la forme plurielle /biiD(-un)/. Cet adjectif a subit au pluriel les règles phonologiques régulières de changement du u-en-i et d'assimilation vocoïde persévérative comme suit : buyD(-un) → biyD(-un) → biiD(-un).
    b. /?a3war(-u)/ 'borgne' qui devient au pluriel /3uur(-un)/. Pour le pluriel de cet adjectif, seul l'assimilation vocoïde persévérative est appliquée comme suit : suwd(-un) → suud(-un).
    c. /?aswad(-u)/ 'noir' qui devient /suud(-un)/. Comme le cas précédent, cet adjectif subit seulement l'assimilation vocoïde persévérative pour former son pluriel : 3uwr(-un) → 3uur(-un).

Quant au dialecte algérien, ce dernier garde les formes du pluriel de base pour les deux premières exceptions, à savoir /byuD/, /3wur/ respectivement. Pour le troisième adjectif, le dialecte algérien emploie une nouvelle forme pour le mot /?aswad(-u)/ 'noir' qui est /?akhal(-u)/ qui a le pluriel /khul/. Nous notons par ailleurs que le pluriel des couleurs en dialecte algérien possède aussi les deux formes suivantes : [F3uuL-a] et [Fuu3aL], par exemple le mot /?abyaD/ possède les formes plurielles suivantes : /byuD/, /byuuD-a/ et /buuyaD/.

2) Les adjectifs de la forme [Fu3aaL(-un)] possèdent au pluriel la forme [Fu3Laan(-un)] dans les deux variétés, par exemple le mot /šujaa3(-un)/ 'courageux' qui devient en EA /šugaa3/ possède le pluriels suivants en MSA et EA respectivement /šuj3aan(-un)/ et /šug3aan/.

3) Les adjectifs de la forme [Fa3iiL(-un) (MSA) → Fi3iiL (EA) et F3iiL (AA)] deviennent au pluriel [Fi3aaL(-un)] ou [Fu3alaa?(-u)] en MSA, et se transforment en [Fu3aaL] et [Fu3ala] en AE et en [F3aaL] et [Fu3ala] en AA. Le changement de la forme [Fu3aLaa?(-u)] en MSA en [Fu3aLa] lors du passage à l'arabe dialectal peut être justifié par la suppression du /?/ final et le raccourcissement de la voyelle finale. A titre illustratif, prenons les exemples suivants :

MSA		**Variante EA**		**Variante AA**		**Traduction**
Adjectif	**Pluriel**	**Equivalent**	**Pluriel**	**Equivalent**	**Pluriel**	
Tawiil(-un)	Tiwaal(-un)	Tawiil	Tuwaal	Twiil	Twaal	Long
Kabiir(-un)	Kibaar(-un)	Kibiir	Kubaar	Kbiir	Kbaar	grand
?amiin(-un)	/?umanaa?(-u)	?amiin	?umana	?amiin	?umana	honnête

4) Les adjectifs de la forme [Fa3iyy(-un)] ont pour pluriel le modèle [?aF3iyaa?(-u)] en MSA, qui possèdent en EA deux formes équivalentes : [?aF3iya] et [Fu3aay]. La première forme est obtenue via suppression du /?/ final et le raccourcissement de la voyelle finale, comme pour l'adjectif /taqiyy(-un)/ 'pieux' (MSA) qui devient en EA /ta?i/, qui a pour pluriel en MSA le mot /?atqiyaa?(-u)/ et en EA le mot /?at?iya/. Pour la deuxième forme nous avons l'exemple de l'adjectif /Tariyy(-un)/ 'moelleux' (MSA) qui devient en EA /Tari/ possède les pluriels /?aTriyaa?(-u)/ et /Turaay/ pour le MSA et le EA respectivement. Quant au dialecte algérien, nous notons le changement de la forme [Fa3iyy(-un)] qui devient [F3ay] ou [F3iy], qui possède au pluriel les deux formes équivalentes : [?aF3iya] et [F3yyiin] comme pour les adjectifs /Tariyy(-un)/ 'moelleux' et /qawiyy(-un)/ 'fort' en MSA, qui sont en AA /Tray/ et /qwiy/ au singulier et /Trayyiin/ et /qwiyyiin/ au pluriel respectivement.

5) Les adjectifs de la forme [Fa3aaL(-un)] ont le modèle [Fu3aLaa?(-u)] au pluriel dans le MSA, qui est réduit ensuite au modèle [Fu3aLa] par suppression du /?/ final et raccourcissement de la voyelle finale en AE, AA. Par exemple, l'adjectif /jabaan(-un)/ (MSA) > /gabaan/ (EA) 'lâche' qui a les pluriels suivants : /jubanaa?(-u)/ (MSA) et /gubana/ (EA).

Le tableau (8.4) suivant résume les modèles des adjectifs pluriels brisés identiques dans le MSA et l'AE.

Id	Modèle	Exemple	Traduction
1	Fu3L(-un)	humr(-un) > humr	rouge
2	Fu3Laan(-un)	šuj3aan(-un) > šug3aan	courageux
3	FiiL(-un)	biiD(-un) > biiD	blanc
4	FuuL(-un)	suud(-un) > suud	noir
5	Fa3aa?iL(-u)	?awaa?il(-u) > ?awaa?il	premier
6	(?)aF3aaL(-un)	(?)ahraar(-un) > (?)ahraar	libre
7	Fu33aaL(-un)	šuTTaar(-un) > šuTTaar	intelligent

Tableau 8. 4. Les modèles des adjectifs pluriels brisés identiques dans le MSA et l'AE

Le tableau (8.4) montre que les adjectifs pluriels brisés en MSA et qui sont conservés sans changement après le passage en AE suivent l'un de ces modèles : [CVCC], [CVCCaan], [CVVC], [CVCVV?VC], [aCCVVC] et [CVCCVVC]. Pour le dialecte algérien, ce dernier dispose des modèles suivants :

- *F3uL :* c'est l'équivalent en AA des modèles Fu3L(-un), FiiL(-un) et FuuL(-un) comme pour l'adjectif humr(-un) > hmur 'rouge'
- *FaL3iin :* qui correspond au modèle Fa3aa?iL(-u) du MSA comme pour l'adjectif ?awaa?il(-u) > lawlliin 'premier'
- *(?)aF3aaL :* équivalent du modèle (?)aF3aaL(-un). Par exemple l'adjectif (?)ahraar(-un) > (?)ahraar 'libre'
- *Faa3Liin :* c'est l'équivalent du modèle Fu33aaL(-un) comme pour l'adjectif šuTTaar(-un) qui devient šaaTriin 'intelligent'

Le tableau (8.5) illustre quant à lui les changements phonologiques que subissent certains modèles d'adjectifs aux pluriels brisés en passant du MSA à l'AE et AA. La plupart de ces changements sont réalisés par des règles phonologiques régulières, bien que quelques-unes soient phonologiquement irrégulières.

Id	Pattern			Exemple			Traduction
	MSA	EA	AA	MSA	EA	AA	
1	Fu3aLaa?(-u)	Fu3aLa	Fu3aLa	kuramaa?(-u)	kurama	kurama	généreux
		Fu3aaL	F3aaL	Zurafaa?(-u)	Zuraaf	Draaf	drôle
2	Fi3aaL(-un)	Fu3aaL	F3aaL	kibaar(-un)	kubaar	kbaar	grand
3	(?)aF3iLaa?(-u)	(?)aF3iLa	(?)aF3iLa	(?)agniyaa?(-u)	(?)agniya	?)agniya	riche
		Fa3aayiL	F3aayiL	(?)aqribaa?(-u)	?araayib	qraayib	proche
		Fu3aaL	ϕ	(?)ašqiyaa?(-u)	šu?aay	ϕ	vilain
4	Fu3uL(-un)	Fu3aaL	F3uL	judud(-un)	gudaad	jdud	nouveaux
5	Fa3Laa	Fa3La	Fa3La	marDaa	marDa	marDa	maalde
			F3aaL	ϕ	ϕ	mraaD	maalde
6	maFaa3iiL(-u)	maFa3iiL	maFa3iiL	masaakiin(-u)	masakiin	masakiin	pauvre
7	Fa3aaLaa	Fa3aaLa	ϕ	kasaalaa	kasaala	ϕ	fainéant

Tableau 8. 5. Les formes de racines adjectivales aux pluriels brisés en MSA ayant subi des changements phonologiques en EA &AA

Le tableau (8.5) montre que les adjectifs pluriels brisés qui subissent des changements phonologiques en AE et AA suivent l'un des modèles suivants : [CVCVCaa?], [CVCVVC], [aCCVCaa?], [CVCVC], [CVCVV], [maCVVCVVC] et [CVCVVCVV]. Ces changements peuvent être expliqués par l'application de certaines règles phonologique, décrite par modèle comme suit :

- Forme 1 → suppression du /?/ final + raccourcissement de la voyelle finale
- Forme 2 → changements phonologiques irréguliers
- Forme 3 → suppression du /?/ final + raccourcissement de la voyelle finale
- Forme 4 → changements phonologiques irréguliers
- Forme 5 → raccourcissement de la voyelle finale
- Forme 6 → raccourcissement atonique
- Forme 7 → raccourcissement de la voyelle finale

Nous remarquons seulement que pour les formes 2 et 4, les changements sont irréguliers car nous ne connaissons pas encore les conditions exactes dans lesquelles se produisent ces changements.

Il est à noter aussi que certains modèles d'adjectifs pluriels brisés sont spécifiquement dialectaux comme le modèle [Fa33iiL-a], par exemple l'adjectif en EA /hassiib-a/ 'experts en comptabilité' et le modèle [Fi3Laan], comme pour l'adjectif /gid3aan/ 'braves'.

Nous signalons aussi qu'il existe une similitude notable entre le MSA et l'arabe dialectal. Il s'agit de l'utilisation d'un adjectif féminin singulier afin de qualifier un nom pluriel non humain. Toutefois, l'arabe dialectal égyptien étend cette option même pour le pluriel des noms féminins de référence humaine (Gadalla, 2000). Voici quelques exemples illustratifs :

Phrase MSA	**abwaab-un**	**maftuuh-ah**
Décomposition	porte-Nom.pl	ouvert-f.sg
Phrase EA	abwaab	maftuuh-a
Décomposition	porte.pl	ouvert-f.sg
Traduction	Les portes ouvertes	
Phrase MSA	banaat-un	šaaTir-aat
Décomposition	fille.br.pl-Nom	clever-fpl
Phrase EA	banaat	šaTr-a
Décomposition	fille.br.pl	intelligent-fsg
Traduction	Des filles intelligentes	

Il est à mentionner aussi qu'il existe une différence essentielle entre le MSA et l'arabe dialectal. Cette différence concerne le fait que l'arabe dialectal utilise les adjectifs masculins pluriels sains ou brisés afin de qualifier les noms féminins pluriels sains. En d'autres termes, le suffixe féminin pluriel [-aat] n'est pas utilisé pour les adjectifs dans la variété dialectale, comme le montrent les exemples comparatifs suivants :

Phrase MSA	**il-mutasaAbiq-aat-u**	**n-naajih-aat-u**
Décomposition	le-concurrent-fpl-Nom	le-gagnant-fpl-Nom
Phrase EA	?il-mutasaAbiq-aat	in-nagh-iin
Décomposition	le-concurrent-fpl	le-gagnant-mpl
Traduction	Les concurrentes gagnantes	

Phrase MSA	?al-lawh-aat-u	1-jamiil-aat
Décomposition	le-tableau-fpl-Nom	le-belle-fpl
Phrase EA	?il-lawh-aat	ig-gumaal
Décomposition	le-tableau-fpl	le-belle-br.pl
Traduction	Les beaux tableaux	

Le MSA et l'AE se distinguent aussi sur une autre différence qui réside au niveau de la flexion des adjectifs et les accords avec le genre et le nombre. En effet, tous les adjectifs en MSA sont fléchis pour le nombre et le genre, cependant une minorité d'adjectifs dialectaux présentent une anomalie morphologique et n'acceptent pas de ce fait d'affixes pour le nombre ou le genre. Ce sont spécialement des adjectifs relationnels désignant des couleurs ou des choses relatives à un groupe de personnes. Par exemple, nous avons les couleurs /bunn-i/ 'marron', /ramaad-i/ 'gris' et /burtu?aan-i/ 'orange', et pour les caractéristiques de groupes nous citons /rigaal-i/ 'd'homme', /Subyaan-i/ 'de garçon' et /banaat-i/ 'de filles' (Hussein, 1973).

8.6. Le degré de la flexion:

Pour les adjectifs, les grammairiens ont défini des degrés de comparaison, dits aussi degrés de signification, qui décrivent l'intensité de la notion exprimée, attribuant de ce fait des gradations aux adjectifs. Ces degrés sont exprimés de manière analytique en utilisant la syntaxe, ou de façon synthétique à travers la flexion. Dans notre étude nous nous intéressons aux degrés de la flexion où nous distinguons classiquement trois types :

- **Le positif :** qui est une forme de base non marquée
- **Le comparatif :** ce degré établit une hiérarchie entre deux éléments
- **Le superlatif :** ce degré est utilisé afin d'exprimer le plus haut niveau d'une qualité, pour formuler la supériorité, ou le plus bas niveau d'un qualificatif en cas d'expression d'infériorité.

En arabe, les adjectifs possèdent une forme spéciale appelée *élatif* qui peut exprimer les diverses valeurs de comparatif et du superlatif. Par exemple كبير *kabiir* 'grand' a pour *élatif* أكبر *akbar* qui peut être traduit 'plus grand' (comparatif) ou 'le plus grand' (superlatif) selon le contexte. Il est à noter aussi qu'en arabe il est possible que le superlatif soit parfois formé en ajoutant un pronom pluriel, comme affixe, à la forme comparative.

8.6.1. Le Degré Positif

Le degré positif des adjectifs inclut les formes non-marquées de ces adjectifs. Ce degré a été présenté avec ses différentes formes dans la section (8.1).

8.6.2. Le Degré Comparatif

Comme toutes les grammaires, la grammaire arabe possède des constructions syntaxiques permettant d'exprimer la comparaison entre deux entités. Quel que soit la variante de l'arabe, standard ou dialectal, l'expression de la forme comparative est faite à partir des adjectifs dans la forme positive en utilisant le modèle «أَفْعَلُ » [?aF3aL(-u)], où le [?a-] représente le préfixe d'inflexion, suivi de la préposition « من » /min/ 'que'. Il est à noter aussi que dans les deux variétés, la forme comparative n'éprouve aucune variation de genre et de nombre. Elle est de ce fait invariable et généralement exprimé au singulier masculin, comme pour la phrase suivante : ?al-Ta'irat-u ʔasra'-u min ?al-sayaArat-i (l'avion est plus rapide que la voiture), qui exprime une comparaison entre deux entités au singulier et qui donne pour ces entités au pluriel la phrase : ?al-Ta'iraAt-u ʔasra'-u min ?al-sayaAraAt-i (les

avions sont plus rapides que les voitures).

Le tableau suivant montre quelques exemples d'illustration des formes positives et leur correspondant au comparatif.

Adj. Positif	**Adj. Comparatif**
sahl(-un) 'facile'	?ashal(-u) 'plus facile'
kabiir(-un) 'grand'	?akbar(-u) 'plus grand'
jabaan(-un) ' lâche '	?ajban(-u) 'plus lâche'
šujaa3(-un) 'brave'	?ašja3(-u) 'plus brave'
mufiid(-un) 'utile'	?afyad(-u) 'plus utile'
munaasib(-un) 'adapté'	?ansab(-u) 'plus adapté'

Les exemples du tableau montrent aussi que certains changements morpho-phonémique, en MSA, sont impliqués dans la formation du modèle comparatif [?aF3aL(-u)] à partir d'adjectifs positifs. Ces changements incluent sont le résultat des opérations suivantes :

- La suppression de la première voyelle de la forme radicale
- Le remplacement de la seconde voyelle par un /a/
- L'ajout du [?a-] comme préfixe d'inflexion au début du mot

Ces changements sont aussi appliqués dans la variété dialectale, avec en plus une autre opération qui consiste à supprimer la dernière voyelle du schème [?aF3aL(-u)], en occurrence le (-u), ce qui donne la forme [?aF3aL]. De plus, dans certaines régions de l'ouest Algérien (comme Tlemcen) et du Maroc, ce schème passe à la forme [F3aL] qui résulte de la suppression du préfixe [?a] du début du mot. Par exemple, /?atqal/ 'plus lourd' passe à /tqal/, l'adjectif /?ahlâ/ 'plus doux' passe à /hlâ/.

Nous signalons aussi quelques cas particuliers en fonction de la nature du radical de l'adjectif. Pour les adjectifs à radicaux doublés, qui possèdent la même lettre à la deuxième et troisième position, la variante du schème [?aF3aL(-u)] devient [?aFa33(-u)] par application de la règle de Métathèse des consonnes identiques. En ce qui concerne les adjectifs de radical défectueux, se terminant par une glide, la variante du schème [?aF3aL(-u)] devient [?aF3aa] en MSA et [?aF3a] en dialecte. Cette transformation est le résultat de l'application de la règle d'élision de la glide et Assimilation vocoïde persévérative dans les deux variétés, puis dans le dialecte, de la règle de raccourcissement de la voyelle finale. Voici quelques exemples de ces adjectifs :

Positif	**MSA Comp.**	**EA Comp.**	**Traduction**
xafiif(-un) > xafiif > xfiif	?axaff(-u)	?axaff	plus léger
šadiid(-un) > šidiid > šidiid	?ašadd(-u)	?ašadd	plus fort
qaliil(-un) > qylayl> qliil	?aqall(-u)	?aqall	moins nombreux
ganiyy(-un) > gani > gani	?agnaa	?agna	plus riche
3aali(n) > 3aali > 3aali	?a3laa	?a3la	plus haut
gaali(n) > gaali > gaali	?aglaa	?agla	plus cher
hulw(-un) > hilw > hlaw	?ahlaa	?ahla	plus doux

(Gadalla, 2000) avance dans son livre, qu'il existe en plus du modèle normal, une

nouvelle construction syntaxique alternative utilisée dans le dialectal égyptien pour exprimer la comparaison. Elle est utilisée dans un registre assez bas et consiste à placer la préposition /3an/ après l'adjectif qui reste inchangé. La même construction est utilisée dans les dialectes maghrébins à l'exception de l'emploie de la préposition /3la/ à la place de /3an/. Voici un exemple qui exprime la phrase بلال أَكْبَر مِنْ كمال *Billel-u ?akbar-u min Kamel* 'Bille est plus âgé que Kamel' en MSA et ses équivalents en dialecte égyptien et algérien:

- *MSA:* Billel-u ?akbar-u min Kamel
  Billel-Nom plus-agé-Nom que Kamel
- *EA:* ? Billel kibiir 3an Kamel
  Billel agé que Kamel
- *AA:* ? Billel kbiir 3la Kamel
  Billel agé que Kamel

Certains adjectifs, en particulier ceux exprimés sous la forme de participes actifs ou passifs, ne peuvent pas être fléchis pour la comparaison. Leur forme comparative dans ce cas est obtenue dans le MSA en utilisant un nom ou un nom verbal indéfini au cas accusatif après un adjectif comparatif à sens vague, comme أَكْثَرُ *?akθar(-u)* 'plus', أَكْبَرُ *?akbar(-u)* 'plus grand', أَشَدُّ *?ašadd(-u)* 'plus fort', أَقَلُّ *?aqall(-u)* 'moins', etc. Il est à noter que l'utilisation de ces adjectifs doit être cohérente avec le nom ou le nom verbal utilisé ainsi que le contexte de la comparaison. Les formes comparatives des adjectifs de couleurs et de défauts corporels, qui possèdent [?a-] dans leurs forme masculine au singulier sont formés aussi de la même manière. Par ailleurs, dans l'arabe dialectal, les deux alternatives citées ci-dessus sont utilisées sans aucun changement sur les adjectifs. Voici quelques exemples illustratifs de ce cas :

Phrase MSA	**?al-burj-u**	**?akθar-u**	**rtifaa3- an**	**min al-3imaar-ah**
Décomposition	la-tour-Nom	plus-Nom	haut-Acc	que le-bâtiment
Phrase EA	?il-burg	murtafi3		3an il-3imaar-a
Décomposition	la-tour	haut		que le-bâtiment
Traduction	La tour est plus haute que le bâtiment			
Phrase MSA	Billel-u	?aqall-u	htimaam-an	min Kamel
Décomposition	Billel-Nom	moins-Nom	intéressé-Acc	que-Kamel
Phrase EA	Billel	muhtamm	?a?all	min Kamel
Décomposition	Billel	intéressé	moins	que-Kamel
Traduction	Billel est moins intéressé que Kamel			
Phrase MSA	haaðihi 1-ward-at-u	?ašadd-u	hmiraar	min tilk
Décomposition	cette-fleure-Nom	plus-intense	rouge	que celle-ci
Phrase EA	?il-ward-a di	hamra	?aktar	min di
Décomposition	la-fleure cette	rouge	plus	que celle-ci
Traduction	Cette fleure est plus rouge que celle-ci			

Enfin, nous notons que certains adjectifs ne subissent, en aucun cas, le degré de flexion, et ce pour les deux variétés. Parmi ces adjectifs nous citons les adjectifs numériques (ex. /ar-raabi3(-u)/ 'le quatrième') et les adjectifs quantitatifs (par exemple. كُل *kul* 'tous' et مُعْظَم *mu3Zam* 'la plupart').

8.6.3. Le degré Superlatif:

Nous faisons appel au superlatif afin de faire une comparaison entre un groupe

d'entités et un individu appartenant au même groupe, ce qui n'est pas le cas dans la comparaison classique faite généralement entre deux entités seulement. Le superlatif est utilisé pour désigner les extrêmes que nous exprimons par les mots : le meilleur, le premier, le pire, le dernier, etc. Dans le MSA et l'arabe dialectal, les adjectifs possèdent la même forme pour le degré comparatif et le superlatif. Ils se distinguent seulement par l'affixation ou par des moyens syntaxiques. L'affixation d'un pronom pluriel à la forme comparative fait partie des moyens morphologiques de formation du superlatif, comme c'est le cas dans la phrase '*cette fille est la plus belle d'entre elles*' qui est exprimée en MSA et en dialecte égyptien comme suit :

- *MSA:* haaðihi 1-bint-u ?ajmal-u-hunna
  cette la-fille-Nom plus belle-Nom-3fpl
- *EA:* ?il-bint di ?agmal-hum
  la-fille celle-ci plus belle-3pl

Un autre moyen morphologique pour exprimer le superlatif est l'ajout d'un article défini aux adjectifs comparatifs, pour illustrer ce moyen prenons l'exemple de la phrase '*Salah est le plus petit*' qui a les équivalents suivants en MSA et en EA respectivement :

- *MSA:* Salah-u huwa 1-?aqSar
  Salah-Nom il le-plus-petit
- *EA:* Salah huwwa 1-a'Sar
  Salah il le-plus-petit

D'un point de vue syntaxique, les adjectifs comparatifs sont transformés en des superlatifs en étant suivi d'un des éléments suivants :

i. Un nom singulier indéfini, comme pour la phrase '*Jéricho est la plus ancienne cité dans le monde*' qui est exprimée en MSA et en EA comme suit :
   MSA: ?ariiha ?aqdam-u madiinat(-in) fiy l-'aAlam
   Jéricho-Nom plus-ancienne-Nom cité-Gen dans le monde
   EA : ?ariiha ?a'dam madiinat fiy l-'aAlam
   Jéricho plus-ancienne cité dans le monde

ii. Un nom pluriel défini, comme pour la phrase 'l'usure est le plus grand des péchés capitaux' qui est exprimée en MSA et en EA comme suit :
   MSA: ?al-ribA ?a'Dam-u ?a-kaba'ir-i
   L'usure-Nom plus-grand-Nom péché.pl-Gen-Def
   EA: il-ribA ?a'Zam ?a-kaba'ir
   L'usure plus-grand péché.pl-Def

iii. Le pronom indéfini /waahid-at(-un)/ en MSA qui a comme équivalent en dialecte le pronom /wahd-a/ 'un'. Prenons pour ce cas l'exemple de la phrase 'Salah est le plus doux' qui est exprimée en MSA et EA comme suit :
   MSA: Salah-u ?aHan-u waahid(-in)
   Salah-Nom plus-doux-Nom un-Gen
   EA: Salah ?aHan waahid
   Salah plus-doux un

Enfin, nous signalons qu'il existe, en plus des formes superlatives et comparatives décrites ci-dessus, trois adjectifs ne suivant pas le schème [?aF3aL] qui sont : خَيْرٌ 'bien', شَرٌّ 'mal' et حَبّ 'mieux'. Ces mots sont utilisés sous cette forme pour exprimer le comparatif et le superlatif.

8.7. Les Adjectifs Relationnels (Nisbas)

L'adjectif relationnel est un adjectif dérivé d'un nom utilisé pour relier une personne ou une entité à ce nom. La règle générale pour la formation de tels adjectifs dans les deux variétés est l'ajout de l'un de ces trois suffixes relationnels au nom : [-iyy], [-aaw-iyy] ou [-aan-iyy] où le dernier est le moins fréquemment utilisé. Dans l'arabe dialectal (AA ou EA), ces suffixes sont raccourcis en position finale, dans quel cas, la consonne géminée finale est perdue. Cette perte de consonne est due à un changement phonologique régulier dans lequel une règle de raccourcissement de la voyelle finale s'applique à deux reprises comme suit :

[-iyy = -iii → -ii → -i]

De ce fait, les suffixes deviennent en arabe dialectal : [i], [-aaw-i] ou [-aan-i]. L'arabe dialectal utilise en plus de ces suffixe un quatrième marqueur distinctif emprunté de la langue turque qui est [-gi]. Nous détaillons dans ce qui suit l'utilisation de ces différents marqueurs.

a. Afin de voir l'usage du suffixe [-iyy > -i#] dans la formation des adjectifs relationnels, observons les adjectifs suivants exprimés dans les deux variantes :

Nom	MSA Adj	EA Adj	Traduction
?asyuuT	?asyuuT-iyy(-un)	asyuuT-i	de Assiut
Faransaa	farans-iyy(-un)	farans-i	français
Taariix	taariix-iyy(-un)	tariix-i	historique
3ilm	3ilm-iyy(-un)	3ilm-i	scientifique

Dans le cas des noms féminins marqués par [-at], ce marqueur ne fait pas partie de la base et donc il ne doit plus exister lorsque le marqueur est ajouté, comme pour les adjectifs suivants :

Nom	MSA Adj	EA Adj	Traduction
handas-at(-un)	handas-iyy(-un)	handas-i	géométrique
haqiiq-at(-un)	haqiiq-iyy(-un)	ha?ii?-i	réel
Tabii3-at(-un)	Tabii3-iyy(-un)	Tabii3-i	naturel
ziraa3-at(-un)	ziraa3-iyy(-un)	ziraa3-i	agricole

b. Le suffixe [-aawiyy > -aawi#] est utilisé dans les deux variétés pour former les adjectifs relationnels à partir de noms féminins se terminant par [-aa], [-aa?] ou, rarement, [-a(t)]:

Nom	MSA Adj	EA Adj	AA Adj	Traduction
qinaa	qinaaw-iyy(-un)	?inaaw-i	qinaaw-i	de Qena
samaa?(-u)	samaaw-iyy(-un)	samaaw-i	smaaw-i	bleu ciel
Sahraa?(-u)	Sahraaw-iyy(-un)	Sahraaw-i	Sahraaw-i	Saharien

c. Le troisième marqueur relationnel qui est utilisé dans les deux variétés est [-aaniyy> -aan-i#]. Il est plus commun en dialecte qu'en MSA. Il est employé avec des formes nominales, des formes adjectivales et certaines formes adverbiales comme soutenu dans (Ghaly, 1960). Voici quelques exemples d'utilisation de ce marqueur :

Mot	MSA Adj	EA Adj	Traduction
haqq(-un)	haqqaan-iyy(-un)	ha??aan-i	droit

?asmar(-u)	?asmaraan-iyy(-un)	?asmaraan-i	grisâtre
?awwal(-u)	?awwalaan-iyy(-un)	?awwalaan-i	premier
taht(u)	tahtaan-iyy(-un)	tahtaan-i	inférieur
fawq(u)	fawqaan-iyy(-un)	fu?aan-i	supérieur

d. Comme mentionné dessus, la variété dialecte possède un marqueur relationnel emprunté du Turc, en l'occurrence [-gi]. Il s'adjoint à un certain nombre de mots pour former des noms de métiers, comme illustré dans les mots suivants :

Nom en dialecte		**Adjectif relationnel**	
Nom	**Traduction**	**Adj**	**Traduction**
sufra	salle à manger	sufra-gi	garçon
?ahwa	café	?ahwa-gi	garçon de café
3arabiyya	voiture	3arba-gi	voiturier
saA'a	montre	Sa'aAji	horloger
HammaAm	bain	HammaAmji	patron de bain

Dans les deux variétés, un adjectif relationnel féminin, arabes ou turques, se forme en dédoublant le [i] final, donnant le suffixe [-iyy-at(-un)] où le [-yy] apparait en arabe dialectal car le contexte suscitant le raccourcissement n'est plus le même : l'affixation du [-a] évite le raccourcissement de la voyelle finale, comme dans les exemples suivants :

msg (MSA)	**MSA fsg**	**EA fsg**	**Traduction**
?ajnab-iyy(-un)	?ajnab-iyy-at(-un)	?agnab-iyy-a	étranger
jazaA'ir-iyy(-un)	jazaA'ir-iyy-at(-un)	gazayr-iyy-a	algérienne
Sahraaw-iyy(-un)	Sahraaw-iyy-at(-un)	Sahraw-iyy-a	Saharien
tahtaan-iyy(-un)	tahtaan-iyy-at(-un)	tahtan-iyy-a	inférieur

Nous notons que dans le dialecte algérien, la présence de très nombreux noms à forme d'adjectifs relationnels féminins. Ces formes existent déjà dans la langue arabe classique et se rencontrent en AA avec une fréquence particulière. Ils sont les substantifs abstrait d'adjectif qualificatifs comme : 3ozuubiyya 'célibat, hommes', shbuubiyya 'beauté', hmuuriyya, 'rougeur', etc.

Par ailleurs, lorsqu'un adjectif relationnel est au pluriel, la forme [-iyy] en MSA est conservée en arabe dialectal, car encore une fois, l'affixation d'un suffixe pluriel sain évite l'application du raccourcissement de la dernière voyelle, comme dans les exemples suivants :

MSA pl	**EA pl**	**Traduction**
suur-iyy-uun	sur-iyy-iin	syrien
farans-iyy-uun	farans-iyy-iin	français
haqqaan-iyy-uun	ha??an-iyy-iin	droit

Il est à noter aussi que certains adjectifs relationnels ont une forme en pluriel brisé dans les deux variétés. Les mots du tableau suivant illustrent cette propriété :

MSA sg	**EA sg**	**Plural**	**Traduction**
?injliiziyy(-un)	?ingliizi	?ingliiz	anglais

3arabiyy(-un)	3arabi	3arab	arabe
?iiTaaliyy(-un)	?iTaali	Talyaan	italien
?amriikiyy(-un)	?amriiki	?amriikaan >?amrikaan	americain

Au sujet du pluriel féminin, sa formation pour les adjectifs relationnels passe par l'utilisation du suffixe féminin [-aat] dans le MSA et l'AA. Quant à l'AE, ce pluriel est formé en employant le suffixe masculin [-iin]. Ci-après nous donnons quelques mots à titre illustratif :

Fsg	**MSA fpl**	**EA pl**	**Traduction**
libiyy-at(-un)	libyy-aat(-un)	libiyy-iin	libyen
tunusiyy-at(-un)	tunusiyy-aat(-un)	tunusiyy-iin	tunisien
tahtaaniyy-at(-un)	tahtaaniyy-aat(-un)	tahtaaniyy-iin	inférieur

Enfin, nous signalons une exception en AE qui est l'adjectif /maSriyy-aat/ 'égyptiennes' qui se produit optionnellement avec l'adjectif /maSriyy-iin/ comme pluriel de l'adjectif féminin /maSriyy-a/. Ceci peut être considéré comme un emprunt direct à partir du MSA comme l'a argumenté McGuirk dans (McGuirk, 1986).

Partie IV : Contexte et matériel / Génération automatiques des ressources

Chapitre 9 Création des lexiques

Introduction

Le développement d'outils de traitement automatique pour l'arabe dialectal se heurte au manque de ressources linguistiques. Afin de combler cette carence de ressources, nous avons procédé à la constitution des lexiques en arabe dialectal rédigé à la fois en écriture arabe et latine. Les dialectes auxquels nous nous sommes intéressés sont essentiellement ceux d'Algérie, du Maroc, de Tunisie et de l'Egypte.

Dans ce chapitre, nous allons décrire notre démarche de création des lexiques. La section 9.1 présente la méthode de construction des dictionnaires dialectaux. Nous présentons dans la section 9.2 les démarches suivies pour la constitution des dictionnaires verbaux puis la méthode suivie pour la constitution des dictionnaires des noms et des mots-outils dans la section 9.3. La section 9.4 est dédiée à présenter le cas de l'emprunt. Finalement, nous présentons dans la section 9.5 la méthode de la génération automatique des formes fléchies.

9.1. Construction des dictionnaires dialectaux

Avant de présenter la méthode de construction de notre lexique, il convient de signaler que notre lexique arabe (MSA ou dialectal) est stockés dans des dictionnaires électroniques. Chaque dictionnaire contient des entrées lexicales utilisant des caractères arabes et codées en Unicode / UTF-8. Chaque entrée lexicale présente une liste d'associations contenant le couple lemme/information.

De ce fait, il est important d'avoir une phase préliminaire lors de l'élaboration du noyau de base du dictionnaire. Cette phase consiste à consulter des sources susceptibles de contenir des données à insérer. Cette phase peut s'avérer coûteuse en traitements et calculs en raison de la diversité et de la volumétrie des sources analysables. Pour assouplir ces traitements nous mettons l'accent dans un premier temps sur la création d'une liste de lemmes de base. Ces lemmes incluent les termes les plus fréquemment utilisés dans la langue, et pourront être mis à jour avec des éléments issus des étapes d'analyse ultérieures. Cette phase de création de liste de lemmes, appelée aussi *lemmatisation*, permet d'introduire des entrées flexionnelles de la liste de lemmes comme suit :

- la forme conjuguée à la 3ème personne du singulier de l'accompli et à l'inaccompli actif pour les verbes ;
- la forme au masculin singulier pour les noms variables selon le genre ;
- la forme au singulier pour les noms invariant en genre.

En revanche, nous signalons que les entrées non fléchissables seront listées telles qu'elles apparaissent dans le lexique. De plus, chaque entrée est associée clairement à une catégorie grammaticale désignée par un code. Les codes utilisés sont formulés à partir de lettres, et dans la liste suivante nous présentons quelques éléments de l'ensemble des codes que nous utilisons :

Catégorie	Code	Exemple
Verbe	VERB	خْرَج *xraj* 'sortir'
Nom	NOM	بَاب *baab* 'porte'
Adjectif	ADJ	مَزْيَان *mazyaan* 'beau'
Adverbe	ADV	
Préposition	PREP	عْلَى *3laa* 'sur'
Pronom personnel	PRON_PRES	أَنْتُومَا *Antuwmaa* 'vous'
Pronom Relatif	PRON_REL	اللي *alliy* 'qui'

Pronom Démonstratif	PRON_DEM	هَاذُو *haaDuw* ‘ceux-ci’

Tableau 9. 1. La liste des catégories grammaticales

Lors de la création des lemmes, il est important de prendre en considération les caractéristiques linguistiques de la langue arabe. Dans ce registre, nous convenons de faire une distinction entre les mots fléchis et non fléchis, dénommées aussi les formes canoniques. Dans ce contexte (Mesfar, 2008) avance les propos suivants : « *il s'agit d'une différenciation formelle très importante; d'un point de vue morphologique cela nous aidera à définir deux sous-ensembles spécifiques du lexique d'une langue, avec des caractéristiques distinctives qui seront en conséquence formalisées par des algorithmes de flexions également distinctifs.* ». Les sections à venir seront consacrées à la description de la méthodologie suivie pour la construction de nos dictionnaires

9.1.1. Préliminaires

Dans cette partie nous introduisons quelques concepts et éléments de base pour décrire la morphologie des verbes et des noms. Ces concepts et éléments sont nécessaires pour la description des processus que nous avons développés pour élaborer nos dictionnaires. Les concepts concernés sont : la racine, le schème et le lemme.

i. *La racine (الجِذْر) :* c’est une unité lexicale constituée d'une séquence de deux à cinq lettres, massivement composée de trois consonnes, utilisée pour désigner un sens général ou une notion abstraite. Elle forme la base du mot et elle est très utilisée dans les langues dérivationnelles. De ce fait, les racines correspondent à un champ sémantique et via des modèles, appelés aussi schèmes, nous pouvons générer une famille de mots appartenant à ce champ sémantique. Par exemple, la racine « د ر س » [d r s] peut engendrer plusieurs mots autours de la notion d’étudier tels que مُدَرِّس *mudarris* ‘enseignant’, مَدْرَسَة *madrasat-un* ‘école’, دِرَاسَة *diraasat-un* ‘étude’, etc.

ii. *Le schème (الوَزْن) :* appelé aussi gabarit ou patron (pattern), est un modèle composé de trois consonnes ف [f], ع [‘], et ل [l], (dont le sens est faire) qui sont vocalisées et pouvant être augmenté d’autres lettres sous le forme de préfixe, suffixe et infixe. Il est parfois composé de chiffre et de lettres définissant le format d’un lemme. Le rôle du schème est très important dans le cadre de la génération des formes dérivées à partir d’une racine. Le schème est utilisé dans ce processus comme le modèle de génération où chaque lettre ou chiffre du schème est remplacé par la lettre correspondante dans la racine du mot considéré. Reprenons l’exemple du lemme verbal دَرَسَ *darasa*, il est obtenu à partir de la racine « د ر س » [d r s] et le schème *1a2a3a* en remplaçant, les chiffres 1,2 et 3, par les lettres correspondantes de la racine.

En d’autres termes, nous pouvons considérer que les schèmes sont des modèles avec différentes structures qui sont appliquées à la racine pour créer un lemme. L’application de différents modèles ou schèmes sur la même racine conduit à la construction de lemmes ayant des significations différentes. Il convient de signaler qu’il existe deux catégories de schème : les schèmes verbaux et les schèmes nominaux. Ainsi, à partir d’une racine, on peut générer des noms et des verbes selon la catégorie du schème utilisé, par exemple à partir de la racine «د ر س » [d r s] nous pouvons générer le verbe دَرَسَ *darasa* par le schème verbal *1a2a3a* ou فَعَلَ *fa‘ala* et le nom مَدْرَسَة *madrasat-un* ‘école’ par le schème nominal *ma-12a3at-un* ou مَفْعَلَةٌ *ma-f‘alat-un*.

iii. *Le lemme :* il s'agit d'une suite de caractères formant une unité sémantique pouvant constituer une entrée lexicale dans un dictionnaire. Il est entièrement vocalisé et généralement issu à partir d'une racine et d'un schème en s'appuyant sur un ensemble de règles de transformations et de concaténations. Ceci concerne notamment les verbes, les noms et quelques particules. Nous pouvons aussi signaler que chaque mot est rapporté à son lemme qui représente ainsi sa forme canonique qui dépend toujours de la catégorie grammaticale de ce mot : si c'est un nom il doit être au singulier et si c'est un verbe il doit être à l'accompli avec la 3ème personne du singulier (Khemakhem, 2006).

Catégorie grammaticale	Lemme	Mot
Verbe	مَدْرَسَة	مَدَارِسٌ
Nom	دَرَسَ	دَرَسْنَا
Particule	فِي	فِي

Les lemmes sont utilisés dans l'analyse des textes et sont généralement formés par des mots simples mais parfois composés. Enfin, un lemme peut regrouper les mots ayant la même racine pour certaines catégories lexicales, ce qui diminue le nombre des entrées d'un dictionnaire.

Enfin nous terminons cette section par l'exemple suivant qui montre la relation entre les différents concepts introduits à travers le processus de dérivation à partir de la racine « دَ رَ سَ » [d r s].

Racine	دَ رَ سَ	**[d r s]**	دَ رَ سَ	**[d r s]**
Schème	فَ عَ لَ	[fa ʿa la]	مَ فْ عَ لَ ةٌ	ma-fʿalat-un
Lemme	دَرَسَ	Darasa	مَدْرَسَة	ma-drasat-un
Traduction	Il a étudié		Ecole	

9.1.2. Présentation de la méthode

Nous rappelons que le TAL des dialectes connait une rareté au niveau du développement et de la disponibilité des ressources nécessaires pour les traitements. Afin de faire face à cette carence nous faisons recours à deux méthodes :

- *Exploitation des ressources du MSA* : le principe de cette méthode est de s'appuyer sur les ressources déjà disponibles et suffisamment matures du MSA. Elle permet de transformer et de convertir les lemmes en MSA et développer leurs correspondants dans les quatre dialectes étudiés : algérien, tunisien, égyptien et marocain. Cette projection effectuée passe par une conversion des lemmes, déjà construits pour l'analyse morphologique du MSA, vers les quatre dialectes considérés. L'extraction des lemmes est faite en utilisant un analyseur morphologique développé au sein de l'entreprise GEOLSemantics. Cet outil permet de réaliser l'ensemble des opérations de l'étape de lemmatisation, comme décrite ci-dessus. Nous avons appliqué cette méthode sur les verbes, les noms (et les adjectifs) et les mots-outils. De plus, nous avons enrichi les listes générées avec d'autres informations associées à ces lemmes comme la transcription en écriture latine ou la traduction en français et en anglais. Cette méthode a été aussi utilisée pour d'autres dialectes arabes comme le levantin (Chiang et al., 2006) et même pour certains des dialectes étudiés comme l'égyptien avec (Mohamed et al.,

2012) et plus récemment le tunisien dans (Boujelbane et al., 2014). La figure suivante montre le processus de construction des dictionnaires MSA vers l'arabe dialectal inspirée des travaux effectués par (Boujelbane et al., 2014).

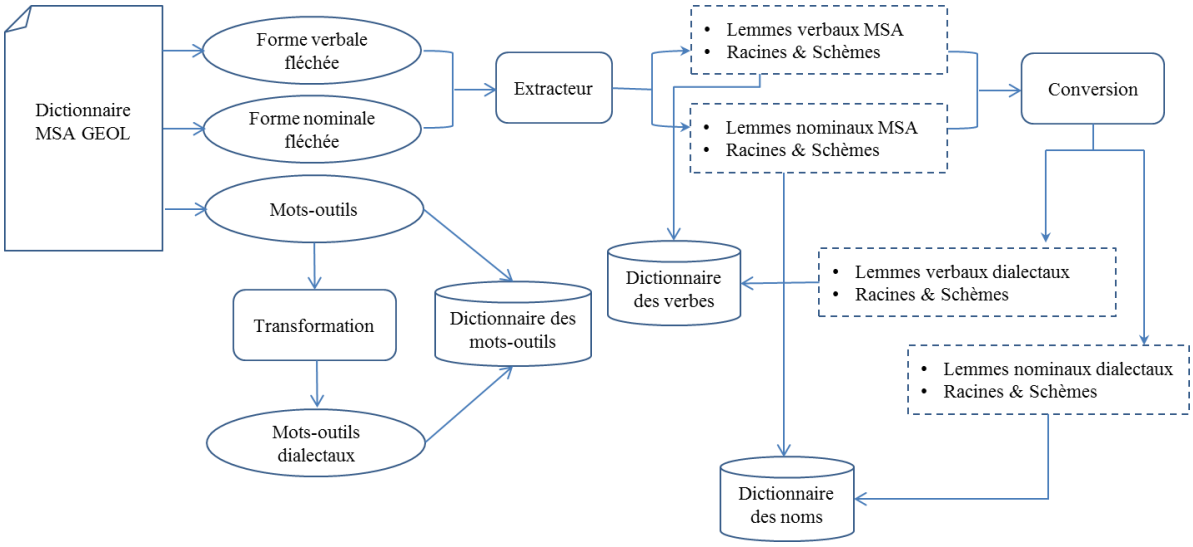


Figure 9. 1. Le processus de construction de dictionnaires.

- *Transcription des mots écrits en caractères latins vers l'écriture arabe* : la deuxième méthode s'appuie sur la translittération des lemmes contenus dans différents dictionnaires contenant des lemmes dialectaux transcrits en écriture latine. Parmi ces dictionnaires nous citons : *Dictionary of Egyptian Arabic (Arabic English)* développé par (Hinds et Badawi, 1986), *al-ʿāmmīyah al-Miṣrīyah = Ánistuuná : Egyptian colloquial* proposé par (Awni, 1999), le dictionnaire *le Karmous* du Tunisien élaboré par (Abdellatif, 2010) ; et le *Moroccan Arabic textbook* dont l'auteur est (Boujnab, 2011). En plus de ces dictionnaires nous avons aussi exploité des ressources internes comme les corpus transcrits en caractères latins que nous avons développés dans le cadre de cette thèse (cf. Chapitre construction des corpus). Les mots transcrits en caractères latins contiennent des voyelles ce qui facilite la tâche de transformation et voyellisation en caractères arabes, ce qui représente un avantage considérable pour cette méthode. Nous signalons tout de même qu'au niveau de cette méthode, les dictionnaires et les corpus ne respectent pas une norme de transcription, ce problème est beaucoup plus présent lors de la transcription des lemmes extraits des corpus où les locuteurs ou les écrivains les transcrivent différemment. Ce problème nous a imposé d'étudier le système phonologique de chaque dialecte afin de proposer les tables de correspondance du latin vers l'arabe. La validation des résultats est manuelle. Elle est faite par des linguistes du domaine et des locuteurs natifs. Afin d'enrichir nos dictionnaires de base et de maîtriser les situations d'ambiguïté, nous avons rajouté pour chaque lemme la catégorie grammaticale ainsi que sa traduction en MSA, en français et en anglais. La figure ci-après montre le processus de construction adopté dans cette deuxième méthode :

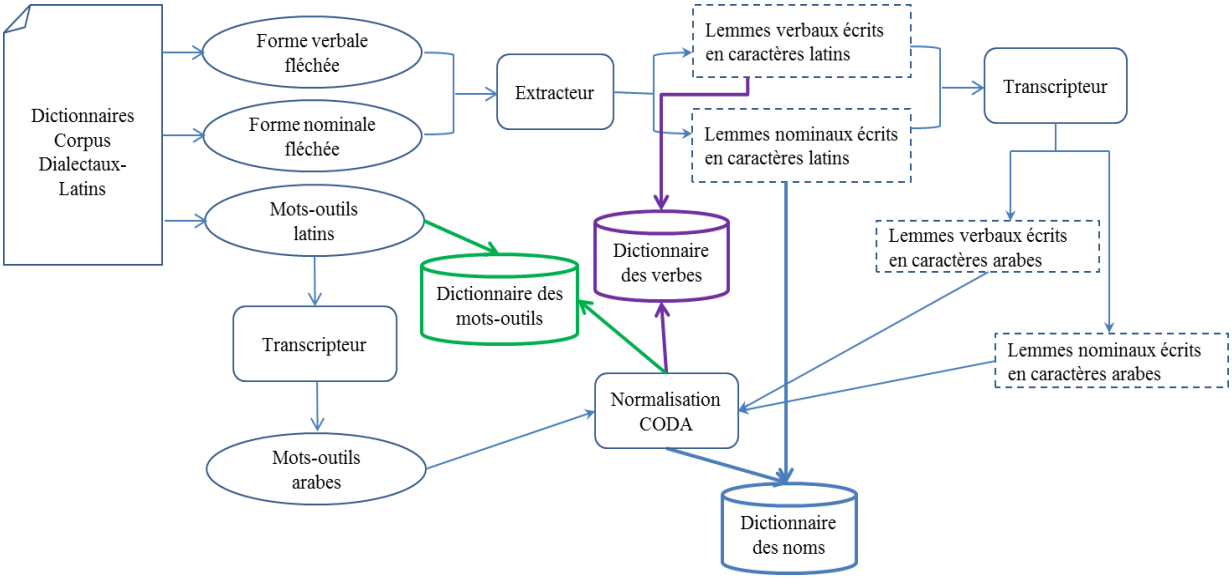


Figure 9. 2. Le processus de construction de dictionnaires latins vers l'arabe.

Dans le processus présenté, nous remarquons une étape de normalisation effectuée avant l'intégration finale des lemmes translittérés dans les dictionnaires. Nous avons développé la norme CODA pour les différents dialectes concernés. Nous mentionnons que CODA est une convention orthographique ayant une variante par dialecte : dialecte algérien (Saâdane et Habash, 2015), dialecte tunisien (Zribi et al., 2014) et pour le dialecte égyptien (Habash et al., 2012). En ce qui concerne le marocain nous avons proposé des règles en étudiant le système phonologique, morphologique, lexical et orthographique. Cette norme sera détaillée dans la section suivante.

9.1.3. La convention CODA

CODA, Convention Orthographique du Dialecte Arabe, est une convention qui a pour but principal l'élaboration d'un standard pour la transcription du dialecte arabe. Ses principes et objectifs sont introduits et décrits dans (Habash et al, 2012). Nous pouvons résumer les principaux éléments de cette convention comme suit :

- CODA est une convention de cohérence interne pour l'écriture de l'arabe dialectal.
- CODA est créée à des fins de calculs.
- CODA utilise l'alphabet arabe.
- CODA est conçu comme un cadre unifié pour l'écriture de tous les dialectes arabes.
- CODA vise à trouver un équilibre optimal entre le maintien d'un niveau d'unicité dialectal et l'établissement d'une convention basée sur les similitudes MSA-AD.

Par ailleurs, CODA est conçue en respectant de nombreux principes. Parmi ces principes nous citons :

1. CODA est une convention ad hoc qui utilise uniquement les caractères arabes, y compris les signes diacritiques utilisés pour l'écriture du MSA.
2. CODA est cohérente car elle associe à chaque mot AD une forme orthographique unique qui représente sa phonologie et sa morphologie.

3. CODA utilise et étend les décisions orthographiques de base du MSA (les règles, les exceptions et les choix ad hoc). Par exemple, l'utilisation de la Shadda pour la gémination phonologique ou encore l'orthographe morphologique de l'article défini.
4. CODA conserve généralement la forme phonologique des mots dialectaux en prenant en considération seulement les règles phonologiques de chaque dialecte (comme le raccourcissement de la voyelle), et les limites de l'écriture arabe (par exemple, l'utilisation d'un signe diacritique et d'un glissement conforme à l'écrire d'une voyelle longue).
5. CODA préserve la morphologie et la syntaxe des dialectes arabes.
6. CODA est facile à apprendre et à écrire.
7. Les principes de CODA sont les mêmes pour tous les dialectes, mais chaque dialecte aura ses propres règles de correspondance avec la convention CODA. Ces règles uniques de correspondance respectent la phonologie et la morphologie du dialecte considéré.
8. CODA n'est pas une représentation purement phonologique. Le texte écrit sous la convention CODA peut être facilement lu dans un dialecte sous réserve de connaître les règles de correspondance de CODA avec le dialecte considéré.

Nous terminons cette section par un exemple de CODA pour le dialecte algérien illustré dans le tableau (9.2) comme suit :

Raw Text	مرحبا بكم في بلاتو حصة برنامج الخط لحمر لنهار اليومة والي يتزامن مع عيد المرأة. إنشاء الله قاع النساء الي راهم يشوفو فينا إنشاء الله أيام سعيده وجميلة فحياتهم. إنشاء الله يتهناو ب ماليهم، ب والديهم وولادهم. قبل منروحو للموضوع نتاع اليومة والي خصصناه للمرأة ف الجزاير وكيفاش راهي عايشة خلونا نرحبو بالضيوف تع لبرنامج.
	mrHbA bkm fy plAtw HSħ brnAmj AlxT lHmr lnhAr Alywmħ wlly ytzAmn mç çyd AlmrÂħ. ĂnšA' Allh gAç AlnsA' Aly rAhm yšwfw fynA ĂnšA' Allh ÂyAm sçydh wjmylħ fHyAthm. ĂnšA' Allh ythnAw b mAlyhm, b wAldyhm wwlAdhm. qbl mnrwhw llmwDwç ntAç Alywmħ wAly xSSnAh llmrÂħ f AljzAyr wkyfAš rAhy çAyšħ xlwnA nrhbw bAlDywf tç lbrnAmj.
CODA	مرحبا بكم في بلاتو حصة برنامج الخط الحمر لنهار اليوم واللي يتزامن مع عيد المراة. انشا الله قاع النسا اللي راهم يشوفوا فينا انشا الله ايام سعيدة وجميلة فحياتهم. انشا الله يتهناوا بماليهم، بوالديهم وولادهم. قبل ما نروحوا للموضوع تاع اليوم واللي خصصناه للمراة فالجزاير وكفاش راهي عايشة خلونا نرحبوا بالضيوف تاع البرنامج.
	mrHbA bkm fy blAtw HSħ brnAmj AlxT AlHmr lnhAr Alywm wAlly ytzAmn mç çyd AlmrAħ, AnšA Allh qAç AlnsA Aly rAhm yšwfwA fynA AnšA Allh AyAm sçydħ wjmylħ fHyAthm. AnšA Allh ythnAwA bmAlyhm, bwAldyhm wwlAdhm. qbl mA nrwhwA llmwDwç tAç Alywm wAlly xSSnAh llmrAħ fAljzAyr wkfAš rAhy çAyšħ xlwnA nrhbwA bAlDywf tAç AlbrnAmj.
English	Hello everyone, in « The Red Line » daily show, which coincides with the Women's Day. God willing, for all the women who watch this show, they may have happy and beautiful days in their lives. God willing, and they will rejoice in their families, parents and children. Before addressing the topic of the day, where we focus on women in Algeria and how they are living, let's welcome to our program's guests.

Tableau 9. 2. Un exemple d'une phrase en dialecte algérien

9.2. Construction des dictionnaires verbaux

Nous avons adopté une méthode basée sur la définition des lemmes, des schèmes et des racines verbaux afin de créer les tables de flexions pour alimenter les dictionnaires verbaux, et cela toujours pour les dialectes maghrébins et égyptien. La méthode utilisée a été aussi suivie dans les travaux de (Boujelbane et al., 2014) qui avaient pour objectif le développement des dictionnaires pour le dialecte tunisien. Dans le reste de cette section nous détaillons chacune des définitions liées à la méthode utilisée.

9.2.1. Construction des lemmes

En s'inspirant des méthodes de construction de dictionnaire exposées dans les sections (9.1.1) et (9.1.2), nous avons mis en place deux étapes pour la définition des lemmes comme suit :

1 Lors de la première étape, nous construisons manuellement les lemmes dialectaux correspondant aux 1000 lemmes recueillis de ressources MSA. Nous signalons à ce niveau que pour chaque lemme extrait, nous élaborons son ou ses correspondants dans chacun des dialectes considérés (AA, AT, AE et AM). Les résultats obtenus ont été validés par trois locuteurs natifs pour chaque dialecte.

2 La translittération des lemmes constitue l'essentiel de la deuxième étape. En effet, à cette étape les lemmes verbaux, extraits à partir des dictionnaires identifiés dans la section 1.2 ainsi que les corpus transcrits en caractères latins que nous avons créés, sont translittérés en écriture arabe. Cette translittération présente un avantage qui réside dans le fait que les mots transcrits en caractères latins contiennent des voyelles. Dans ce registre, nous signalons que les dictionnaires et les corpus ne respectent pas une norme ou un standard de transcription. Ce problème est beaucoup plus présent lors de la transcription des lemmes extraits à partir des corpus où les locuteurs ou les écrivains transcrivent différemment. Par ailleurs, nous avons généralement collecté 500 lemmes différents de ceux utilisés lors de la première étape. Cet effort pour avoir un minimum de lemmes différents de ceux de la première étape nous a permis aussi de faire une validation supplémentaire de l'étape 1. De plus, nous remarquons à ce niveau que les lemmes construits pour l'égyptien et le tunisien sont plus nombreux que ceux de l'algérien et du marocain. Cette distinction est due au fait que les dictionnaires concernant l'EA et le TA sont plus riches que ceux des autres dialectes. Les résultats obtenus ont été d'abord normalisés selon la norme CODA ensuite validés manuellement. La validation a été faite par des linguistes du domaine et des locuteurs natifs.

9.2.2. Construction des schèmes

La création des tables de flexion est basée sur la construction des schèmes verbaux afin de proposer les formes fléchies correspondantes. Dans la grammaire arabe, chaque verbe en MSA appartient à un schème ou une classe permettant de regrouper les verbes qui ont les mêmes représentations et structures morphologique. En d'autres termes, les schèmes permettent d'avoir des représentations morphologiques communes à un grand nombre de verbes. En se basant sur cette définition, nous avons étudié et analysé la morphologie verbale des dialectes étudiés (AA, TA, AM, AE) afin de construire les schèmes correspondants à chaque dialecte étudié.

Il est important de rappeler que pour l'arabe moderne standard (MSA) des efforts ont été fournis depuis des années afin de donner naissance au développement d'une quantité importante de ressources. Ceci a permis de perfectionner les outils de traitement automatique des langues en les alimentant avec des sources riches et plus élaborées. Vu que les dialectes

arabes sont une variante de l'arabe et qu'il existe une proximité entre ces deux variétés, notre approche vise comme les autres travaux à exploiter et adapter les outils existants pour le MSA aux dialectes arabes. Pour cela, nous avons étudié les correspondances ainsi que les différences qui existent entre les schèmes MSA et les verbes en dialectes. Cette étude est détaillée dans le chapitre analyse morphologique des verbes.

Nous soulignons par ailleurs qu'il existe trois niveaux de classification des schèmes verbaux dialectaux comme suit :

- classification selon le modèle du verbe dialectal
- classification selon la voyelle de la deuxième consonne du schème
- classification selon l'aspect imperfectif du verbe, autrement dit selon la voyelle de la marque à la forme imperfective (inaccomplie)

Plus précisément, notre démarche vise à adapter les contenus des ressources MSA aux dialectes en essayant de trouver les correspondances entre les schèmes MSA et les verbes en dialecte. Pour ce faire, nous avons d'abord réalisé la construction des schèmes verbaux des dialectes à travers la classification selon le modèle du verbe dialectal en se basant sur les classes ou schèmes en MSA. Ensuite, nous avons défini, selon la voyelle de la deuxième consonne du schème, des sous classes pour chaque classe identifiée. Enfin, nous avons affiné les sous-classes avec une stratification selon la voyelle de la marque de l'inaccompli.

Dans le reste de cette section, nous donnons plus de détails sur les opérations que nous avons effectuées pour chacune des classifications des schèmes dialectaux données ci-dessus.

9.2.2.1. Classification selon le modèle du verbe

Pour cette classification, nous avons commencé, dans une première étape, par identifier et marquer certains verbes dialectaux. Les verbes concernés sont ceux qui ne changent pas lors du passage MSA/dialecte tout en gardant le même modèle. Comme nous l'avons déjà présenté, la grammaire arabe définit dix schèmes pour les racines tri-radicaux (trilitères). Le tableau suivant résume ces schèmes :

Numéro	Classe de schème
I	فعل Fv3vL
II	فعّل Fv33vL
III	فاعل Fvv3vL
IV	أفعل ?vF3vL
V	تفعّل tvFv33vL
VI	تفاعل tvFvv3vL
VII	انفعل (?i)nFv3vL
VIII	افتعل (?i)Ftv3vL
IX	إفعلّ (?i)F3vLL
X	إستفعل (?i)stvF3vL

Tableau 9. 3. Les classes de schème en MSA

Les exemples suivants illustrent cette première étape :

❖ **Exemple 1 :**
- ***MSA :*** كَبُرَ *kabur(-a)* 'grandir' qui suit le modèle [Fv3vL] correspond à la présentation (C1VC2VC3) appartenant à la classe MSA-I.

- ***AE :*** كِبِرْ *kibir* qui suit le modèle [Fv3vL] correspond à la présentation(C1VC2VC3) et qui appartient à la classe EGY-I.
- ***AT :*** كُبِرْ *kobir* qui suit le modèle [Fv3vL] correspond à la présentation (C1VC2VC3) et qui a pour classe TUN-I.

❖ **Exemple 2 :**

- ***MSA :*** سَافَرَ *saafar* 'voyager' qui suit le modèle [Fvv3vL] correspond à la présentation (C1vAC2vC3) qui appartient à la classe MSA-III
- ***EA && AT :*** سَافِرْ *saafir* [Fvv3vL] correspond à la présentation (C1vAC2vC3) appartenant à la fois aux classes EGY-III et TUN-III.

La deuxième étape consiste à chercher les verbes qui changent complètement leurs formes en passant du MSA vers les dialectes. Nous avons aussi vérifié s'il existe des ressemblances morphologiques et phonétiques entre eux et ceux qui sont déjà classés afin d'éviter au maximum les redondances au niveau des racines. Voici quelques lemmes pour illustrer cette étape :

❖ **Exemples EA :**

- Le verbe دَلَلَ *dalal* 'gâter' en MSA possède un lemme différent en dialecte égyptien qui est la forme دَلَّعْ *dalla3*. Cette forme suit le modèle [Fv33vL] C1vC2C2VC3 et ressemble au lemme verbal en AE قَدَّمْ ?*addim* 'présenter' qui lui aussi suit à son tour le même modèle.
- Comme le verbe en AE قَدِّمْ ?*addim* appartient au schème EGY-II par conséquent le verbe égyptien دَلَّعْ *dalla3* appartient aussi au même schème.

❖ **Exemples TA :**

- Le verbe بَحَثْ *baHaT* 'chercher' en MSA possède un lemme différent en dialecte tunisien qui est la forme لوّج *law~aj*. Cette forme لوّج *law~aj* suit le modèle [Fv33vL] C1vC2C2VC3 qui ressemble au lemme verbal en TA كَسَّرْ *kassar* suivant le même modèle.
- Le verbe en TA كَسَّرْ *kassar* appartient au schème TUN-II par conséquent le verbe tunisien لوّج *law~aj* appartient à son tour à TUN-II.

La troisième et dernière étape consiste à définir les formes pour certains schèmes pour lesquels les deux premières étapes n'ont pas associé des schèmes. Par exemple, les dialectes maghrébins en général ne disposent pas des lemmes sous les deux classes IV, VIII et IX. Par exemple, pour la classe IX nous avons constaté que tous les verbes MSA appartenant à ce schème possèdent le même modèle en se traduisant en arabe dialectal maghrébin. C'est la raison pour laquelle nous avons classé les verbes des dialectes maghrébins concernés, suivant le même schème, dans cette classe mais avec une nouvelle présentation. Par exemple :

- إِحْمَرَّ *(?i)hmarr(-a)* 'rougir' en MSA est transformé en حْمَارْ *hmaar* en dialecte. Nous notons que certains l'écrivent اِحْمَارْ *(?i)hmaar*.

9.2.2.2. Classification selon la voyelle de la deuxième consonne

Selon (Ouerhani, 2009), la voyelle de la deuxième consonne est l'un des éléments les plus systématiques dans la morphologie verbale en arabe. Cette voyelle, à l'accompli et à l'inaccompli, est considérée comme le premier critère de classement des verbes. Dans la grammaire arabe, la même 2ème voyelle génère à l'accompli pour le premier schème 3 types représentés comme suit :

- ***1ier type :*** فَعَلَ *Fa3aL(-a)* ~ C1aC2aC3a
- ***2ème type :*** فَعِلَ *Fa3iL(-a)* ~ C1aC2iC3a

- ***3ème type*** : فَعُلَ *Fa3uL(-a)* ~ C1aC2uC3a

A la forme inaccomplie, la 2ème voyelle génère aussi un sous-classement au sein de chacun des types. De ce fait, nous rappelons que seul le premier schème (I) peut avoir six sous schèmes différents selon la variation de la voyelle de la deuxième consonne de la racine à la forme accomplie et à la forme inaccomplie comme suit :

Accomplie	**Inaccomplie**
Fa3aL(-a) ~ C1aC2aC3a	ya-Fa3iL(-a) ~ ya-C1aC2iC3a
Fa3aL(-a) ~ C1aC2aC3a	ya-Fa3uL(-a) ~ ya-C1aC2uC3a
Fa3aL(-a) ~ C1aC2aC3a	ya-Fa3aL(-a) ~ ya-C1aC2aC3a
Fa3iL(-a) ~ C1aC2iC3a	ya-Fa3aL(-a) ~ ya-C1aC2aC3a
Fa3iL(-a) ~ C1aC2iC3a	ya-Fa3iL(-a) ~ ya-C1aC2iC3a
Fa3uL(-a) ~ C1aC2uC3a	ya-Fa3uL(-a) ~ ya-C1aC2uC3a

Tableau 9. 4. La variation de la 2ème voyelle du schème I.

Pour les dialectes arabes, l'étude de leurs morphologies et leurs schèmes nous a permis de constater que même si le lemme verbal ne change pas, les voyelles de la deuxième consonne changent à la fois dans les deux formes accomplie et inaccomplie. De ce fait, nous déduisons que même pour les dialectes cette voyelle reste un critère fiable pour le classement des verbes. Prenons par exemple le verbe يَشْرَبُ / شَرِبَ *šarib/yašrab* 'boire' en MSA donné à la forme accomplie et à l'inaccompli respectivement. Ce verbe suit le modèle C1aC2iC3a/yaC1oC2aC3 et traduit en dialecte algérien par يُشْرُبُ / شْرَبْ *šarab/yušrub* qui suivent les modèles C1aC2aC3a et yuC1oC2uC3 respectivement. Ces représentations montrent que le verbe شَرِبَ en MSA appartient au sous-schème MSA-I-ia du schème I sous la forme (Fa3iL(-a)/ ya-F3aL(-u)). Cela signifie que la voyelle de la deuxième consonne des verbes appartenant à ce sous-schème prend un [i] à la forme accomplie et un [a] à la forme inaccomplie. En faisant la correspondance avec le dialecte, le verbe penche vers le schème ALG-I-au sous la forme (F3aL/yu-F3uL) ayant les caractéristiques suivantes pour la deuxième voyelle : elle prend un [a] à la forme accomplie et un [u] à la forme inaccomplie.

En ce qui concerne le reste de schème en MSA, nous signalons qu'il n'existe pas une variation au niveau de la voyelle de la deuxième consonne, autrement dit au sein d'un même schème tous les verbes ont la même voyelle de la deuxième consonne de leur racine. Cela est traduit par l'existence d'un seul schème au sein des autres modèles (à part le premier schème). Ceci nous a conduits à constater qu'au sein d'un même schème tous les verbes ont la même voyelle de la deuxième consonne de leur racine.

En ce qui concerne les dialectes arabes, nous avons constaté qu'au niveau des verbes appartenant à ces schèmes (les autres modèles), il existe une variation vocalique de la deuxième consonne. Voici quelques exemples illustratifs :

❖ **Schème** V :

- ***Exemple 1:***
  - *MSA :* يُجَرِّبَ/جَرَّبَ *jar~ab/yu-jar~ib* 'essayer' suit le modèle

C1aC2C2aC3/yuC1aC2C2iC3
  - *AE :* جَرَّبْ / يِجَرَّبَ *gar~ab/yi-gar~ab* suit le modèle C1aC2C2aC3/yiC1aC2C2aC3
- ***Exemple 2 :***
  - *MSA :* قَدَّمَ/يُقَدِّمُ *qad~am/yu-qad~im* 'présenter' suit le modèle C1aC2C2aC3/yuC1aC2C2iC3
  - *AE :* قَدَّمْ / يِقَدِّمْ *?ad~im/yi-?ad~im* suit le modèle C1aC2C2iC3/yiC1aC2C2iC3

❖ **Schème III:**

- ***Exemple 1 :***
  - *MSA :* سَافَرَ/يُسَافِرُ *saAfar/yu-saAfir* 'voyager' qui suit le modèle C1aAC2aC3/yuC1aAC2iC3.
  - *TA :* سَافِرْ /يْسَافِرْ *saAfir/y-saAfir* qui le modèle C1aAC2iC3/yC1aAC2iC3.
- ***Exemple 2 :***
  - *MSA :* وَافَقَ/يُوَافِقُ *waAfaq/yu-waAfiq* 'accepter' suit le modèle C1aAC2aC3a/yuC1aAC2iC3.
  - *TA :* وَافَقْ / يْوَافَقْ *waAfaq/yi-waAfaq* suivant le modèle C1aAC2aC3/yC1aAC2aC3.

Cette variation vocalique au niveau de la deuxième consonne a nécessité un traitement supplémentaire pour les dialectes arabes. Par conséquent, nous avons défini des sous-classes pour chaque classe (ou chaque modèle) si cela existe. Dans ce cas voici les sous schèmes des exemples que nous avons illustrés :

- ***EA :***
  - جَرَّبْ *gar~ab* appartient au sous-schème EGY-III-aa
  - قَدَّمْ *?ad~im* appartient au sous-schème EGY-III-ii
- ***TA :***
  - سَافِرْ *saAfir* appartient au sous-schème TUN-III-ii
  - وَافَقْ *waAfaq* appartient au sous-schème TUN-III-aa.

Le tableau suivant montre quelques exemples des sous-classes que nous avons pu créer et identifier pour décrire les spécificités des verbes en TA et AE.

	Schème						
MSA	II: فَعَّلَ	III: فَاعَلَ	IV: أَفْعَلَ	V: تَفَعَّلَ	VI: تَفَاعَلَ	VII: اِنْفَعَلَ	VIII: اِفْتَعَلَ
TA	II-aa	II-aa	ϕ	V-aa	VI-aa	ϕ	VIII-ii
	II-ai	II-ii		V-ii	VI-ii		
	II-ii						
EA	II-aa	II-ii	IV-ai	V-aa	VI-ii	VI-ai	VIII-ai
	II-ii			V-ii			

Tableau 9. 5. Des exemples des sous-classes des schèmes en TA && EA.

9.2.2.3. Classification selon la marque de l'inaccompli

Dans la grammaire arabe, les verbes se caractérisent par la stabilité et le non-changement de la marque à la forme inaccomplie au sein du même schème. A titre d'exemple, au niveau du schème I, les verbes à l'inaccompli commencent toujours par le préfixe يَ *ya* comme dans les verbes : يَضْرِبُ *ya-Drib(-u)* 'frapper', يَخْرُجُ *ya-khruj(-u)* 'sortir' et يَحْزَنُ *ya-hzan(-u)* 's'attrister', de même au niveau du schème III les verbes commencent par le préfixe يُ *yu* comme par exemple يُحَارِبُ *yu-haarib(-u)* 'guerroyer', etc.

Cette caractéristique n'est pas valable pour les dialectes car cette marque varie même au sein d'un même schème. C'est la raison pour laquelle ce facteur a été considéré dans notre étude comme étant un élément essentiel pour la classification des verbes dialectaux. De ce fait, à l'instar des travaux de (Boujelbane et al., 2014), nous avons enrichi d'avantage les sous-schèmes déjà créés, lors des classifications précédentes, afin d'en proposer des nouveaux prenant en considération cette variation vocalique de la forme à l'inaccomplie. Prenons par exemples, les deux verbes رْبَحْ *rbaH* 'gagner' et عْرَفْ *3raf* 'savoir' du schème I, d'après la deuxième classification (selon la voyelle de la deuxième consonne à la forme inaccomplie), nous avons constaté que les deux verbes ont la voyelle 'a' par conséquent nous les avons classés sous le même sous-schème : TUN-I-aa. Cependant, si nous considérons le troisième facteur de classification, nous constatons que le premier verbe possède la forme inaccomplie يِرْبَحْ *yi-rbaH* et le deuxième possède la forme يَعْرَفْ *ya-3raf*, ceci peut être traduit par une différence dans la marque de l'inaccompli. Ainsi, nous avons créé les deux sous-schèmes **[TUN-I-aa-i]** et **[TUN-I-aa-a]** qui seront liés pour l'exemple cité pour les verbes يِرْبَحْ *yi-rbaH* et يَعْرَفْ *ya-3raf* respectivement.

9.2.3. Construction des racines

Les verbes en MSA sont catégorisés en deux principales classes en fonction du nombre de consonnes dans leurs racines: verbe tri-radical (trisyllabique) et verbe quadri-radical (quadrisyllabique). Ces deux classes sont à leur tour divisées en sous-classes selon le type et la nature des consonnes dans leurs racines. Ces nouvelles sous-catégorisations sont dues à des transformations ou des remplacements engendrés par l'application des règles phonologiques nécessaires après la modification des voyelles ou l'ajout d'un suffixe. A ce sujet, nous identifions plusieurs niveaux de classifications, en s'appuyant sur des critères comme : la présence ou non des consonnes défectueuses et la lettre hamza, la position de ces consonnes et leur nombre dans la racine, la duplication des consonnes, etc. ...

Dans le cas des dialectes, la définition des racines dialectales reste un sujet d'actualité et une problématique assez large attirant de plus en plus d'attention. Cette problématique se traduit par l'absence d'un standard stable pour définir la racine des verbes dialectaux. Cette absence se manifeste lorsque le verbe dialectal ne change pas seulement les voyelles mais lorsqu'il change complètement la racine en passant du verbe du MSA vers le dialecte. Afin d'étudier cette problématique, nous avons effectué des travaux en deux étapes. Nous avons commencé d'abord par définir et agréger les racines des verbes qui restent inchangeables en passant du MSA vers les dialectes, tel que le verbe كَتَبَ *katab-a* en MSA qui devient كْتِبْ *ktib* en dialecte. De ce fait, nous avons gardé la même racine MSA pour le verbe en dialecte, à savoir la racine «ك ت ب» 'k t b'. Ensuite, nous avons défini les racines des verbes qui changent totalement en passant du MSA vers les dialectes.

Nous rappelons que jusqu'à présent aucun standard n'est défini pour les racines dialectales, c'est pour cette raison que nous avons adopté une méthode déductive afin de déterminer les racines en dialectes. Dans ce registre, nous rappelons que dans la grammaire arabe, nous avons la règle suivante pour la définition d'un lemme :

Racine + Schème = Lemme … (1)

Pour les dialectes, nous avons élaboré pour chaque dialecte les lemmes et schèmes associés. Pour les racines, elles sont déduites à partir des lemmes et schèmes défini, en appliquant la règle (1), où nous voyons clairement que la racine peut être déduite à partir d'un schème et d'un lemme. Par exemple, nous avons classé le lemme إِسْتَنَّى *?istan~a* sous le schème X, إِسْتَفعل *?istaf3al*, en substituant le lemme et le schème dans la règle (1), nous obtenons : la

racine (?) + إسْتَفعل *?istaf3al* = إسْتَنَّى *?istan~a*, ce qui donne la valeur ننى *nnY* pour la racine recherchée.

Toutefois, nous signalons que la racine trouvée n'est pas unique, car nous pouvons aussi trouver d'autres racines si nous classons le lemme dans un autre schème. Pour l'exemple précédent, nous pouvons aussi classer le lemme إسْتَنَّى *?istan~a* dans le schème VIII افتعل ?ifta3al, ce qui donne la racine سنن *snn*. Nous pouvons aussi obtenir la racine quadrilatère سنني *snnY* si nous classons le lemme إسْتَنَّى *?istan~a* sous le schème افتعلل *?ifta3lal*. Afin de simplifier cette situation et ne pas générer plusieurs racines pour le même lemme, nous avons opté pour la sélection d'une seule racine obtenue à partir de la plus simple classification que nous pouvons obtenir du lemme considéré.

9.3. Construction des noms et des mots outils

Suivant le même processus que pour les verbes, nous avons tout d'abord construit manuellement une base de 1 500 lemmes traduits et translittérés dans les quatre dialectes. Nous pouvons répartir les noms dialectaux comme les noms MSA en trois catégories selon le système morphologique comme suit :

1. **Les primitifs :** sont les noms qui constituent le glossaire fondamental de la langue arabe, et représentent les noms qui ne peuvent pas être rattachés à une racine verbale. Cette catégorie inclue aussi les noms propres, les noms communs et les racines bilitères. Par exemple, nous citons فاس (faas - pioche), محمد (Mohammed) et فم (fum – bouche).
2. **Les dérivés :** sont les noms formés à partir d'une racine verbale et d'un schème. Le statut de cette dernière détermine la nature et le nombre de ces formes. Nous trouvons dans cette catégorie les participes actifs (ضَارِب – celui qui frappe), les participes passif (مضروب – frappé), les noms de lieux ou de temps (مَضْرِب – lieu de frappe), le nom d'une fois (ضربة – une frappe), etc.
3. **Les nombres :** ce sont les numéros simples représentant les unités (zéro- à neuf), les dizaines et les centaines, etc ; et les numéros composés comme les cardinaux, par exemple 'سطاش' – seize.

Après l'obtention des lemmes, nous effectuons plusieurs traitements pour la construction des schèmes nominaux dialectaux.

- Regroupement dans le même schème, des lemmes ayant les mêmes caractéristiques morphologique au singulier (modèle + voyelle). (classement des lemmes qui ont le même comportement au singulier dans le même schème. Ces lemmes sont réguliers lors de la flexion et acceptent un seul pluriel sain.
- Agrégation des lemmes qui n'acceptent pas un pluriel sain ou qui possèdent plus d'une forme au pluriel. Ensuite, nous avons créé des sous-schèmes qui décrivent le comportement morphologique au pluriel de ces lemmes
- Définition pour chaque schème du genre étant donné que certains mots acceptent le masculin et le féminin à la fois alors que d'autres n'acceptent qu'un seul genre.

9.4. Le cas de l'emprunt

Le vocabulaire des dialectes est caractérisé par une présence assez importante de mots issus d'autres langues non arabes. L'incorporation de ces mots dans le vocabulaire dialectal est facilement et naturellement réalisée au niveau des structures lexicales qui, elles, sont restées globalement arabes. Cette introduction de mots étrangers, nommée aussi *emprunt*, contribue significativement au dynamisme et à la vitalité des dialectes.

Selon (Ouerhani, 2009) « *l'emprunt est une composante importante du système du dialectal d'une manière générale. Par ailleurs, certains travaux ont mis l'accent sur l'emploi fréquent en tunisien de verbes empruntés notamment au français qui sont les moins intégrés ou en voie d'intégration dans le système dialectal via le moulage des schèmes. Une fois franchie l'étape de l'intégration phonétique par le biais de la substitution d'un phonème inexistant, soit par le biais de la conservation de son sens d'origine, la matière consonantique du verbe emprunté est versée dans un schème arabe lui permettant de se conjuguer* ».

Cependant, les termes ou mots empruntés aux langues étrangères, comme le français ou l'anglais, sont majoritairement fondus dans un schème arabe, que ce soit en arabe MSA ou en arabe dialectal. Par exemple, en AE nous trouvons le verbe *kansel* 'annuler', emprunté au verbe anglais 'to cancel' et ramené à une racine quadrilitère. Nous avons aussi le verbe *chargaA* en dialecte AA emprunté au verbe français 'charger' et ramené aussi à une racine quadrilitère.

Différents efforts ont été consacrés pour l'étude des emprunts, et de notre côté nous avons mis en place une approche originale qui traite les lemmes empruntés des langues étrangères comme le français pour les dialectes maghrébins (algérien, tunisien et marocain) et l'anglais pour le dialecte égyptien. Dans le cadre de cette approche, nous avons d'abord construit les lemmes verbaux et nominaux empruntés en s'appuyant sur les corpus dialectaux construits et transcrits en écriture arabe et latine. Ensuite, nous avons agrégé les verbes empruntés qui gardent le même modèle en passant du français/anglais vers le dialecte. En d'autres termes vérifier s'il existe des correspondances de schèmes entre les verbes empruntés et les verbes dialectaux. Puis, nous avons extrait les verbes qui n'ont pas de modèles correspondants en dialectes. Ceci nous a mené à développer par la suite des nouveaux schèmes pour les nouveaux lemmes. La même procédure a été suivie pour la classification de ces nouveaux schèmes empruntés.

Par exemple, le verbe فَاصَى *faaSa* 'effacer' qui suit le verbe نَادَى *naada* 'appeler' correspondant au modèle [III-aa]. Quant au verbe مَتْرِزْ *matriz* 'avoir sa maîtrise' suit le schème Fa3L1iL2 de la forme I des verbes quadrilittères tout comme le verbe xarbis 'griffonner'. Nous pouvons aussi citer le verbe بْرُوفِيتَى *pruufiitaa* 'profiter' qui suit un nouveau schème du type F3uuL1iiL2aa (C1C2uuC3iiC4aa). Il y a aussi le verbe أَكْسِيرِيلَى *(?a)ksiiriila* 'accélérer' suivant un nouveau schème du type (?a)F3iiL1iiL2aa (?aC1C2iiC3iiC4aa). Pour ces deux derniers exemples, nous avons défini des nouveaux modèles en s'appuyant sur la même procédure utilisée pour les lemmes verbaux dialectaux afin de définir les tables de flexion correspondantes. Nous notons enfin que ces verbes cités sont ceux du dialecte AA.

9.5. Génération automatique des formes fléchées

Dans cette section nous présentons comment nous avons construit les formes fléchies. Notre objectif est de mettre en place un système de génération automatique de toutes les formes fléchies potentielles. Cette génération est fondée sur des paradigmes flexionnels préalablement assignés à chaque entrée lexicale. Nous prenons en considération dans notre approche les particularités de la langue arabe ainsi que des mots empruntés afin de définir

suffisamment de description formalisée. Ces descriptions sont basées sur un ensemble de règles de transformation morphologiques et phonologiques nécessaires pour le passage de la forme de base (le lemme) à toutes les formes fléchies qui y sont associées.

Nous rappelons que la création des lexiques de formes fléchies est basée sur les racines, les schèmes et les lemmes que nous avons définis. Pour ce faire, nous avons établi des paradigmes de flexion qui représentent une description détaillée des étapes de définition, construction et génération de toutes les formes fléchies d'un lemme donné. Il est à noter que les paradigmes se distinguent par le nombre de consonne de la racine (trilitères ou quadrilitères) et les schèmes.

Notre construction prend en considération les classiques de l'arabe MSA en matière de flexion. En grammaire arabe, la flexion d'un lemme verbal est très régulière. Elle est basée essentiellement sur la concaténation de préfixes et suffixes (affixes) aux lemmes verbaux. La détermination des affixes repose sur les valeurs des traits morphologiques suivants :

- ***Aspect :*** en MSA nous avons d'un point de vue morphologique les trois aspects du verbe suivant : le perfectif (accompli), l'imperfectif (inaccompli) et l'impératif.
- ***Mode :*** les verbes en MSA sont conjugués par un seul mode dans l'aspect perfectif 'l'indicatif" et par trois modes dans l'imperfectif: l'indicatif, le subjonctif et le jussif; en plus de l'impératif.
- ***La personne, le genre et le nombre du sujet :*** l'arabe standard distingue trois personnes (1ère, 2ème et 3ème personne) et deux genres (le masculin et le féminin). Cependant en arabe nous avons trois valeurs pour le nombre à savoir le singulier, le duel et le pluriel.

Pour les dialectes, nous rappelons qu'ils se caractérisent par une morphologie pauvre par rapport à celle du MSA. Cette pauvreté se manifeste par exemple au niveau du mode qui n'est pas marqué ou au niveau des valeurs du nombre qui sont réduites à deux seulement : singulier et pluriel. Les études effectuées ont montré que le duel (masculin ou féminin) est supprimé dans les différentes variétés dialectales, et qu'il est remplacé, voir englouti, par le pluriel masculin. Quant au genre, il n'est spécifié que lorsqu'il s'agit de la troisième personne du singulier dans certains dialectes comme le tunisien (Hamdi et al., 2013). Nous notons aussi que la formation de la voix passive dans les dialectes est faite via l'introduction des nouveaux morphèmes, qui sont généralement les préfixes [t-] et [n-], contrairement au MSA où cette voix est formée par un changement interne des voyelles du verbe. Par exemple, la forme MSA passive قُتِلَ *qutila* 'il a été tué' devient en dialecte تِقْتَلْ *tiqtal*, اِتْقَتَلْ *?itqatal* ou encore اِنْقَتَلْ *?inqatal*.

9.5.1. Les règles morpho-phonémiques et orthographiques

La conjugaison en arabe se base sur un principe simple appliqué avec une réalisation d'un ensemble de règles qui sont généralement d'ordre morpho-phonémique et orthographiques. Ces règles morpho-phonémiques et orthographiques sont appelées aussi les règles de réécriture ou les règles de transformation. Elles prennent en considération les contraintes morphologiques et orthographiques caractérisant la grammaire arabe (standard et dialectale). Parmi ces règles nous avons : l'ajout, la suppression ou la substitution de lettres ou de voyelles.

9.5.1.1. Les règles morpho-phonémiques

Ces règles se basent sur un ensemble de transformations morphologiques telles que l'ajout de lettres, la suppression, la substitution, etc. Nous citons ci-après quelque unes de ces règles :

- ***Transformation des voyelles :*** cette règle transforme les voyelles de manière récurrente et peut toucher la plupart des voyelles du lemme lors de la génération d'une forme fléchie. L'outil que nous utilisons, HTFST, possède une opération *substitute* qui permet de remplacer une voyelle par une autre tout en mentionnant la position de l'opération et la nouvelle voyelle remplacée.
- ***Ajout d'un préfixe ou d'un suffixe :*** cette opération d'ajout d'un préfixe ou d'un suffixe est étroitement liée aux traits morphologiques de la forme fléchie considérée. Nous notons à ce niveau qu'un suffixe peut être une voyelle ou un pronom personnel affixé.
- ***Suppression d'un hamza wasliya ou de la dernière voyelle :*** l'opération de suppression de la dernière voyelle est fréquente. Elle est toujours suivie par l'ajout d'un suffixe. Cependant, la suppression de hamza wasliya est appliquée seulement avec les verbes commençant par "اِسْتَ" [*?ista*] et elle est toujours suivie dans ce cas par l'ajout d'un préfixe.
- ***Transformation de la lettre taa marbuuta :*** cette règle est appliquée lors de la description des paradigmes de déclinaisons nominales. Elle permet de vérifier si la dernière consonne du nom à décliner est une ة (h -ta' marbuuta) ou pas. Dans le cas positif, elle remplace dans certaines règles flexionnelles ou dérivationnelles la lettre ة par ت (t- tâ' maftuha).

9.5.1.2. Les règles orthographiques

Les règles orthographiques permettent de réécrire seulement la représentation orthographique. Elles prennent en compte le changement de l'orthographe de certaines lettres. La complexité morphologique de la langue arabe s'exprime dans l'utilisation des règles phonologiques au cours du calcul d'une forme fléchie. Pour éviter l'apparition de formes erronées, nous limitons les altérations qui peuvent être induites par la lettre hamza, les lettres défectueuses ou le signe šadda au cours de la conjugaison. En plus, nous avons besoin de fusionner la première consonne du suffixe avec la dernière consonne de la racine lorsqu'elles sont identiques. Voici un exemple de quelques règles orthographiques :

- ***Transformation de la graphie de hamza :*** la graphie de la hamza dépend en général de sa position, de sa voyelle et de la voyelle de la lettre qui la précède. Nous avons déjà évoqué que le calcul des formes fléchies nécessite la transformation des voyelles. Cette transformation provoque des changements pouvant être avant ou après la lettre hamza, ce qui nécessite l'application ensuite de règles phonologiques et de transformer la graphie de cette lettre. Par exemple, si nous changeons la voyelle du hamza de 'ـَ' (a – fatha) à 'ـُ' (u – damma), cette hamza prend la graphie suivante ؤ.
- ***Transformation des lettres défectueuses :*** nous signalons d'abord que des altérations importantes peuvent survenir au cours de la génération des formes fléchées en raison de la présence de lettres défectueuses dans la racine. Ces lettres peuvent avoir la forme *alif* dans le lemme, alors que lors de la conjugaison elles reprennent leur forme originale. Ces altérations sont dues au changement des voyelles qui nécessite l'application d'une règle phonologique. Par exemple, la racine "م و ت " [m w t] avec le schème فَعل *Fa3aLa* et selon la règle phonologique considérée donne le lemme مَاتْ

maAt ‘mourir’. De plus pendant le calcul de la forme fléchie à l’inaccompli actif, nous obtenons la forme يْمُوتُ *ymuwt*, par le remplacement du ‘ا’ *alif* par la forme originale de la racine qui est ‘و’ *w* car la voyelle précédente ‘َ’ (a – fatha) est remplacée par ‘ُ’ (u – damma).

- ***Transformation de šadda :*** cette règle concerne un phénomène realtif au signe *šadda* qui peut figurer dans les lemmes des verbes dupliqués. La consonne contenant ce signe peut être remplacée lors de la conjugaison par deux consonnes identiques.
- ***Fusion du suffixe avec la dernière consonne de la racine ن ou ت :*** cette règle est utilisée pour l'unification des descriptions flexionnelles des verbes comme *katab* ‘écrire’ et *Tabat* ‘s’avérer’. Cette règle effectue deux transformations différentes pour donner la conjugaison au passé. En effet, lors de sa flexion à la première personne du singulier à l'inaccompli actif, le premier verbe *katab* nécessite la suppression de la dernière voyelle et l'ajout de la séquence تْ pour obtenir la forme fléchie *katabt* ‘j'ai écrit’ et pour le second verbe la même conjugaison est accompagnée de l'ajout de la marque de redoublement ou Shadda et une dernière voyelle (u)(Tabatt- je me suis avéré). Pour ce qui est des verbes se terminant avec ‘ن’ des transformations sont effectuées lors de la conjugaison des verbes au passé , à la première personne du pluriel. C’est le cas des verbes *katab* ‘écrire’ et *dahan* ‘peindre’ où le premier verbe devient *katabnaA* ‘nous avons écrit’ et le deuxième devient *dahannaA* ‘nous avons peint’.

9.5.2. Présentation des règles utilisées

La génération morphologique de notre système est réalisée par l’outil HTFST. Ce dernier est un système à base de règles qui permet de décrire les systèmes morphologiques des différentes variétés de l’arabe (MSA et dialectes) et de les compiler sous la forme d’un transducteur d’états finis.

Notre outil dispose un ensemble d'algorithme permettant la génération de l'ensemble des formes fléchies en s'appuyant sur un appel aux descriptions flexionnelles attribuées à chaque entrée du dictionnaire. Ces descriptions peuvent être décrites sous la forme d'expressions régulières. Elles reposent sur des opérateurs morphologiques effectuant des transformions au sein des lemmes en entrée. Ces transformations sont basées sur l'utilisation de certaines commandes prédéfinies :

<L> : déplacement vers la gauche (Left arrow)
<R> : déplacement vers la droite (Right arrow),
<B> : suppression du dernier caractère (Backspace),
<S> : suppression du caractère courant (Suppr),
<D> : duplication du caractère courant (Duplicate)
Ces commandes peuvent être associées à deux types d'argument :
- un nombre : par exemple :

<B2> : suppression des deux derniers caractères
<L3> : déplacement à gauche, trois fois
<R4> : déplacement à la droite, quatre fois
<S5> : suppression des 5 caractères qui suivent

Le processus de génération d’un lemme consiste à remplacer chaque chiffre du schème par la lettre correspondante dans la racine.

شْكَرْ , V+Tr C1 = š, C2 = k C3= r, pour avoir la forme nous suivons le schème *yuC1C1uC2oC3uWA* qui est équivaut à l'ensemble des opérations suivantes

- *<LW> :* positionner le curseur (|), initialement placé à la fin du mot (|شْكَرْ - škar|), à la tête du lemme par un déplacement vers la gauche → *(- |C1C2C3);*
- insérer le préfixe à l'inaccompli (يُ - *yu*) à la tête de la forme → (يُ | - *C1C2C3*);
- <D> : dupliquer la première consonne → يُ C1C1C2C3;
- insérer les voyelles (ُ - u) après la première consonne → (yuC1C1uC2C3*);*
- <R> : sauter une lettre vers la droite → (يُشُّكَ|رْ - yu*ššuk|ar);*
- insérer la voyelle (ْ - o) après la deuxième consonne → yuC1C1uC2oC3 (يُشُّكْ|رْ - yu*ššuko|r);*
- <R> : sauter une lettre vers la droite → (|يُشُّكْرْ - yu*ššukor|o);*
- insérer la voyelle (ُ - u) après la troisième consonne → yuC1C1uC2oC3u (|يُشُّكْرُ – yu*ššukru|);*
- <RW> : aller à la fin du mot ;
- insérer les suffixes (وا - *wA*) à la fin de la forme → yuC1C1uC2oC3uwA (يُشُّكْرُوا - *yuššukruwA*);
- Remplacé la lettre C1 par '*š*' et la consonne C2 par la lettre 'k' et la troisième par 'k' : (يُشُّكْرُوا - *yuššukruwA*)

Cette commande permet de générer la forme suivante : يُشُّكْرُوا - *yuššukruwA 'ils remerciement'*. Cette forme sera associée aux informations flexionnelles : V+Tr+A+A (P)+3+m+P, c'est-à-dire : verbe transitif direct (V+Tr) conjugué au masculin (m) pluriel (pl), troisième personne (3), l'accompli (présent), à la voix active (A).

Quant à la flexion des noms :

La forme canonique دْزِيرِيَّة *dziyriyyah* 'algérienne' on veut obtenir la forme دْزِيرِيَّاتْ *dziyriyyaAt'*
«LW » <R> <S> <RW> اتْ /f+pl
«LW » <R> <S> <RW> At /f+pl
Cette commande inclut les opérations morphologiques suivantes :

- <LW> : positionner le curseur (|), initialement placé à la fin du mot (| دْزِيرِيَّة - *dziyriyyah*|), à la tête du lemme par un déplacement vers la gauche → (دْزِيرِيَّة| - |*dziyriyyah*);
- <R6> : sauter six lettres vers la droite → (ة | دْزِيرِيَّ - *dziyriyya|h*);
- <S> : supprimer la lettre suivante → (دْزِيرِيَّ - *dziyriyya);*
- <RW> : aller à la fin du mot;
- insérer (اتْ - *At*) à la fin de la forme → (دْزِيرِيَّاتْ - *dziyriyyaAt*)

Chapitre 10 Translittération des noms propres arabes

Introduction

Les évolutions rapides des nouvelles technologies d'information et de communication sont accompagnées d'un essor important de la quantité et la diversité d'information générée et manipulée notamment celle disponible sur le web. Cette dernière, étant destinée à un public large et varié, est transcrite dans différentes langues ce qui a fait émerger la nécessité d'internationaliser les contenus afin de permettre un partage de données le plus large possible, entre des utilisateurs manipulant des langues différentes. Ainsi, les techniques de translittération trouvent tout leur intérêt afin de rendre cette perspective de partage possible.

La transcription consiste à substituer à chaque son ou à chaque phonème d'un système phonologique, un graphème ou un groupe de graphèmes d'un système d'écriture, tandis que la translittération consiste à substituer à chaque graphème d'un système d'écriture, un autre graphème ou un groupe de graphèmes d'un autre système d'écriture, indépendamment de la prononciation.

Dans le premier cas (transcription), l'objectif est de reconstituer la prononciation originale à partir de l'écriture cible; dans le second (translittération), l'objectif est de retrouver l'écriture cible à partir du système d'écriture source.

Pour atteindre ces objectifs, il existe des systèmes de transcription phonologique et des normes de translittération graphématiques. Mais ces systèmes et ces normes conventionnels sont multiples et complexes, d'autant plus complexes que les langues mises en contact sont éloignées.

Nous présentons dans la section 10.1 les différents aspects du notre sujet. Ensuite, nous présenterons dans la section 10.2 un état de l'art dans le domaine de la translittération suivi d'une description des approches que nous avons utilisées pour développer notre système de translittération automatique des noms arabes voyellés et non voyellés vers les différentes transcriptions possibles en écriture latine et inversement dans les sections 10.3 et 10.4. Nous validons notre technique dans la section 10.5 en présentant des expérimentations utilisant des moteurs de recherche de référence.

10.1. Les différents aspects du sujet

10.1.1. Aspect linguistique

Mais les problématiques soulevées par la transcription et la translittération ne relèvent pas tant de la nature de la convention adoptée que de la relation entre oralité et scripturalité lors du passage d'un système linguistique à un autre. En effet, l'oral et l'écrit obéissent à des règles différentes : l'un a un matériau sonore, l'autre a un matériau visuel, et chaque matériau possède une dynamique interne et des contraintes propres.

Parmi ces contraintes, citons les phénomènes de transformation morphologique qui affectent les mots en fonction de la nature de leur lettre initiale. Ainsi, si le mot contient l'article « Al » (ال), il faut faire la distinction entre les lettres «solaires» et les lettres «lunaires». Avec les premières (solaires), le «L» ne se prononce pas et la lettre qui le suit est dédoublée dans la prononciation et dans l'écriture. A l'inverse, avec les lettres lunaires, le «L» de l'article se prononce et la lettre qui le suit n'est pas dédoublée ni dans la prononciation ni dans l'écriture. Mais ces règles de la langue arabe ne sont pas toujours respectées dans la translittération usuelle comme en témoignent les exemples concernant la translittération des noms de journaux arabes en écriture latine : الزمان(Al Zaman) en Iran au lieu de (AZZAMAN), الصباح(Al-Sbah) en Palestine au lieu de (AS-SABAH), الشروق(echorouk) en

Algérie au lieu de (Ech-Chorouk), الزمن(Azamn) au lieu de (Az-Zaman) à Oman...

On constate un phénomène analogue d'écart par rapport à la norme phonologique du système d'origine dans la translittération de la lettre «T-attachée» (ة), dite en arabe « ta-marbouta ». Celle-ci se prononce (t) à l'état d'annexion exclusivement. Mais là encore, certaines translittérations de noms de journaux rendent compte de la graphie du mot arabe et non pas de sa prononciation effective : par exemple, الثورة(Al-tawra) au lieu de (ATH-THAWRAH).

Il est par conséquent important de mener un questionnement, préalable au traitement automatique, concernant des aspects qui peuvent paraître évidents au premier abord mais qui méritent une analyse approfondie. Citons les aspects suivants comme prioritaires pour le traitement automatique de la transcription et la translittération:

- ✓ Degré d'adéquation de la transcription à l'oral (phonème);
- ✓ Degré d'adéquation de la translittération à l'écrit (graphème);
- ✓ Degré d'adéquation de la notation symbolique à l'usage (social).

10.1.2. Aspect cognitif

On oublie également que les phonèmes et les graphèmes possèdent souvent une valeur symbolique qui se révèle de diverses manières :

- ✓ La réaction psychologique des utilisateurs face aux symboles utilisés pour noter l'oral ou l'écrit (tolérance ou rejet) ;
- ✓ La revendication d'une tradition orale liée à l'histoire ou à des valeurs propres au groupe source (par exemple, la notion d'honneur liée à la parole donnée dans les pays arabes) ;
- ✓ La pression politique concernant la prééminence d'un idiome sur les autres, en l'occurrence de l'arabe comme langue coranique teintée de sacralité, sur les dialectes d'un côté et sur les langues étrangères de l'autre.

Ces considérations signifient que le spécialiste en traitement automatique doit tenir compte, en amont de son traitement, d'un certain nombre de phénomènes qui ne ressortent pas directement du TAL mais qui doivent être pris en considération dans la phase d'analyse et de modélisation afin que les solutions proposées ne soient pas déconnectées de la réalité linguistique ou de la demande sociale.

Ainsi, en ce qui concerne l'arabe, proposer un système de transcription des noms propres ne peut se faire sans une prise en compte minimale des spécificités morphologiques et de la valeur symbolique du système de nomination (Roman, 2007). Par exemple, un grand nombre de prénoms arabes sont formés en intégrant l'un des noms d'Allah (il y en a 99) et la combinatoire de ces formations n'est pas aléatoire mais strictement réglée, ce qui suppose une cohérence interne lors de la transcription ou de la translittération, ainsi que la définition d'un certain nombre de règles contextuelles qui prennent en compte cet aspect, lequel aspect s'avère parfois symboliquement important pour les personnes dénommées.

Considérons le prénom arabe «عبد الله», prénom du roi de Jordanie et du roi d'Arabie Saoudite. Ce prénom est très fréquent en arabe et peut être translitteré de plusieurs façons en fonction de la volonté du scripteur de faire ou non ressortir le sens original du prénom puisque «Abd» signifie littéralement «Serviteur de». On trouve ainsi: {Abdallah, AbdAllah, Abd Allah, Abd-Allah, Abdullah,...». Et le même phénomène peut être observé pour d'autres prénoms composés avec l'un des 99 noms d'Allah, comme par exemple Abdelkader « عبد

القادر» ou encore Abderrahim «عبد الرحيم».

Dans certains cas, cette dimension symbolique constitue le cœur même du système de nomination. En effet, dans des organisations comme Al-Qaïda, il a été démontré par (Guidère, 2006) que le « nom de guerre » porté par chaque membre –comme par exemple celui d'Abou Moussab Al-Zarqawi – était non seulement significatif mais avait surtout une valeur stratégique[18].

10.1.3. Aspect dialectologique

A cette dimension symbolique dans laquelle intervient la sémantique du nom, s'ajoute une dimension phonologique qu'il est nécessaire de prendre en considération lors du traitement automatique et qui tient à la situation linguistique du monde arabe. En effet, la langue arabe présente aujourd'hui une situation caractérisée par une polyglossie complexe. Il existe ainsi une diversité de réalisations de l'arabe littéraire (classique, moderne, moyen, etc.) et une pluralité de dialectes (variétés d'arabe, régionales ou locales).

Dans cette optique, la dialectologie arabe distingue deux grandes familles de dialectes, celle du Maghreb (Maroc, Algérie, Tunisie et Libye) et celle du Machrek (Egypte, Syrie et Moyen-Orient). Mais à l'intérieur de ces familles de géolectes, on trouve aussi bien des dialectes nationaux (natiolectes) que des dialectes régionaux (régiolectes) ou encore des dialectes locaux (topolectes), parlés sur un espace limité (village, localité).

Ainsi, lorsque l'on se propose de développer un translittérateur des noms arabes, l'on se trouve confronté aux spécificités phonologiques de ces variantes dialectales, car le locuteur-auditeur prononce différemment le même nom en fonction de son dialecte et reconnaît la variation dialectale en fonction de la prononciation qu'il entend. Cela est d'autant plus vrai que chaque dialecte arabe a subi l'influence, au cours de l'histoire moderne, d'autres langues comme le français, l'italien et l'espagnol (pour le Maghreb) ou encore de l'anglais (pour le Machrek). Cela fait qu'un même nom ou prénom en arabe peut avoir plusieurs prononciations différentes dans les dialectes et diverses translittérations en fonction des spécificités phonologiques et graphématiques des langues cibles.

Ainsi par exemple, le nom du journaliste irakien Muntadhar al-Zidi, qui possède une orthographe unique en arabe (منتظر الزيدى) mais plusieurs prononciations et accentuations en fonction des dialectes, est transcrit en écriture latine par différentes formes possibles, parmi lesquelles : Mountazer al-Zaïdi, Mountacer Al-Zaidi, Mountasser Al Zaïdi, Mountazer El-Zaïdi, Muntazer al Zaidi, Muntader al-Zaidi, Mountadher al-Zaïdi, Muntader El Zaidi, Muntazer az-Zaidi, Muntazir Az-Zaydi, …

Cette multiplicité des formes ne manque pas de poser problème tant au niveau de la recherche d'information pour une entité nommée (ici le nom d'un journalise) que pour l'enrichissement interlingue de données concernant un sujet particulier. En effet, le fait de ne pas répertorier toutes les formes disponibles à un moment donné pour un même nom peut être préjudiciable à l'efficacité de la recherche.

---

[18] Guidère (2006) s'appuie sur une analyse morphologique et sémantique du système de nomination employé à l'intérieur des organisations islamistes pour inférer l'intention des utilisateurs et la valeur stratégique du nom en contexte. L'article est disponible à l'adresse suivante : http://www.c4ads.org/files/defense_concepts_I.3.pdf

10.2. État de l'art sur la translittération et la transcription

Le problème de la transcription et de la translittération a intéressé les spécialistes dans plusieurs langues, mais cet intérêt est relativement récent et concomitant de l'essor de l'Internet et de la recherche d'information interlingue. Nous citons à titre d'exemple les travaux de (Jiang et al., 2007) pour la translittération des entités nommées (ENs) du chinois vers l'anglais, qui utilisent un modèle d'entropie maximale pour déterminer la translittération candidate, en se basant sur la similarité phonétique avec l'EN dans la langue source. Ces méthodes fonctionnent bien avec les entités nommées qui sont traduites phonétiquement, mais ce n'est pas toujours le cas. Pour ces types d'ENs, il est plus recommandé d'explorer les similitudes sémantiques entre les ENs dans différentes langues. Ce constat a été approuvé dans les travaux de (Huang et al., 2004) qui combine les similitudes sémantiques et phonétiques. Les expérimentations effectuées montrent que cette approche réalise une précision de 67%. Par ailleurs, (Huang et al., 2003) ont travaillé sur l'extraction des paires d'ENs (hindi-anglais) grâce à l'alignement d'un corpus parallèle. Des paires chinois-anglais sont d'abord extraites à l'aide d'une programmation dynamique. Ce modèle chinois-anglais est alors adapté à l'hindi-anglais de manière itérative, en utilisant les paires hindi-anglais d'entités nommées déjà extraites pour l'amorçage du modèle. On trouve aussi des propositions de systèmes visant à attribuer une seule translittération à un nom donné : c'est le cas du modèle génératif proposé pour les noms d'origine anglaise écrits en japonais (Katakana) vers le système d'écriture latin (Knight et Graehl, 1997). Cette approche a été adaptée par (Stalls et al., 1998) à la façon dont un nom anglais écrit en arabe est transcrit en anglais. Le système de génération de translittérations s'appuie sur un dictionnaire d'apprentissage et ne prend pas en compte les prononciations non répertoriées ou inconnues du dictionnaire. Cela a conduit certains chercheurs à pallier cette carence par un recours au modèle non supervisé. C'est le cas du système de translittération des noms anglais vers l'arabe proposé par (Abduljaleel et al., 2003). Mais celui-ci a montré également ses limites parce qu'il est basé sur le calcul de la forme la plus probable, censée être la forme correcte, ce qui n'est pas vrai pour tous les pays arabes ni pour tous les dialectes.

Pour contourner la difficulté de la prononciation et le problème des variantes dialectales, (Al-Ghamdi, 2005) a proposé un système de translittération en écriture anglaise des noms arabes voyellés. Ce système est basé sur un dictionnaire de noms arabes dans lequel la prononciation est réglée au moyen de voyelles ajoutées aux noms répertoriés, avec indication en vis à vis de leur équivalent en écriture anglaise. Mais cette approche cumule les inconvénients des deux précédentes : non seulement elle ne prend pas en compte les prononciations non répertoriées dans le dictionnaire, mais en plus elle est normative par le fait qu'elle ne propose qu'une seule translittération pour un nom donné. L'objectif de l'auteur semblerait être de favoriser l'adoption d'un standard de translittération, mais cela ne peut être le résultat d'une initiative individuelle et isolée.

En réalité, l'état actuel de la recherche dans ce domaine ne rend pas compte de la complexité du problème de la transcription et de la translittération, lequel touche autant à l'oralité qu'à la scripturalité dans deux ou plusieurs systèmes linguistiques en même temps. En effet, transcrire un nom ou un prénom d'un système linguistique source vers un système d'écriture cible, est une tâche délicate qui nécessite un certain nombre d'opérations exigeant de prendre en considération un ensemble de propriétés morphologiques, phonologiques et sémantiques. Ces opérations sont nécessaires pour assurer un processus de translittération robuste, notamment pour des applications de sécurité, de vérification d'identité, ou encore de recherche d'informations sur Internet.

Or, très peu d'études prennent en considération le lien :

- entre phonologie comparée et transcription interlingue,
- entre graphématique comparée et translittération multilingue,
- entre dialectologie arabe et systèmes de translittération latins.

Les rares études qui proposent une solution prenant en compte partiellement l'une de ces problématiques, sont dédiées à l'identification automatique de l'origine du locuteur à partir de son dialecte. C'est le cas notamment des travaux de (Guidère, 2004) et de (Barkat-Defradas et al., 2004).

Dans le cadre de nos travaux, notre objectif est de proposer un système automatique de translittération qui tient compte de l'ensemble des aspects sus-indiqués, à savoir le lien entre phonologie, graphématique et dialectologie, dans la transcription des noms et des prénoms arabes vers l'écriture latine. Pour ce faire, nous définissons un certain nombre de règles, issues d'une étude expérimentale, et qui rendent compte de la complexité du domaine. Il existe, en effet, une multitude de cas de figure qu'il convient de traiter en fonction du niveau auquel l'on se situe. Nous récapitulons ces cas dans le tableau (Table 1) suivant, établi à partir des situations observées expérimentalement :

Type de traitement automatique	**Unité de traitement source**	**Unité de traitement cible**
Translittération	graphème d'arabe standard : ex. خ	graphème latin : kh (FR); kh (EN), j. (ES)...
	graphème latin : kh (FR); kh (EN), j. (ES)...	graphème d'arabe standard : ex. خ
Transcription	phonème arabe (en fonction des dialectes) : ex. ق	phonème latin : k (FR); q (EN), k (ES)...
	phonème latin : ex. g (FR) / r (grasseyé)	phonème arabe (en fonction des dialectes) : ex. ر ; غ

Tableau 10. 1. Les liens entre graphèmes et phonèmes arabes et latins

10.3. Translittération de noms arabes en écriture latine

Le système d'écriture de la langue arabe est constitué d'un alphabet de 28 lettres. Cet alphabet contient 25 consonnes et 3 voyelles longues. Il existe aussi des voyelles courtes (tableau 10.2) qui sont généralement présentes uniquement dans les textes religieux (Coran, Hadith, etc.) ou les manuels scolaires pour enfants. Cette particularité est l'une des principales sources d'ambiguïté pour les systèmes de translittération.

Voyelle courte	**Transcription**	**Nom**
ـَ	A	Fatha
ـُ	U	Damma
ـِ	I	Kasra
ـْ	E	Sukun
ـّ	doublement	Shadda
ـً	Aa	Fathatan

ـٌ	Uu	Dammatan
ـٍ	Ii	Kasratan

Tableau 10. 2. Les voyelles courtes en arabe.

10.3.1. Normes de translittération pour l'arabe

Il existe plusieurs normes de translittération, dont EI (1960), ISO/R 233 (Système international pour la translittération des caractères, 1961), UNO (Organisation des Nations unies : Groupe d'experts des Nations Unies pour les noms géographiques, 1972), DIN-31635 (Deutsches Institut für Normung, 1982), ISO 233 (Organisation internationale de normalisation, 1984) ainsi que la norme ALA-LC (America Library Association, 1997). Parmi ces normes deux sont reconnues et utilisées internationalement par la communauté scientifique : DIN-31635 et la norme adoptée par l'Encyclopédie de l'Islam (EI).

10.3.2. Correspondances proposées pour la translittération des lettres arabes vers le latin

Afin de renvoyer la totalité des cas possibles de la translittération d'un nom arabe en écriture latine, nous nous sommes intéressés aux questions et aux problèmes liés à la translittération basée sur le système phonétique de l'arabe littéraire ainsi que sur la majorité des familles de dialectes (Algérie, Égypte, Maroc, Tunisie, Liban), en prenant en compte des nombreuses variantes régionales et locales, car la prononciation de certaines lettres et la vocalisation de certains mots changent selon les provinces et les villages. Nous avons commencé par recenser les translittérations existantes pour chaque lettre de l'alphabet arabe standard depuis les normes et usages observés sur le Web et sur les dictionnaires de lieux géographiques de GeoNames. Nous avons constaté qu'au sein du même dictionnaire géographique un nom propre peut avoir plusieurs translittérations différentes. Cette investigation empirique est basée sur un corpus de textes qui a été recueilli dans les différentes langues cibles visées par le translittérateur. Elle a permis de constituer une librairie des équivalents graphématiques utilisés dans les écrits utilisant l'alphabet latin. Nous faisons figurer dans le tableau suivant (tableau 10.3) les équivalences graphématiques établies à partir de cette étude sur corpus :

Lettre	**Transli DIN-31635**	**Transli EI**	**ISO 233**	**ISO/R 233**	**UN**	**ALA-LC**	**Ajouts des translittérations observées dans les corpus**
ء	ʾ	ʾ	ʿ	ʾ	ʾ	ʾ	A
ا	A / â	ā / â	ʾ	—	—	—	a / ä / á / e / ê
ب	B	B	B	b	B	b	
ت	T	T	T	t	T	t	
ث	ṯ	Th	ṯ	ṯ	Th	th	
ج	Ǧ	Dj	ǧ	ǧ	J	j	G
ح	ḥ	ḥ	ḥ	ḥ	ḥ	ḥ	H
خ	ḫ / ẖ	Kh	ḫ	ḫ	Kh	kh	
د	D	D	D	d	D	d	

ذ	ḏ	Dh	ḏ	ḏ	Dh	dh	d / ḍ
ر	R	R	R	r	R	r	
ز	Z	Z	Z	z	Z	z	ẓ
س	S	S	S	s	S	s	
ش	S	Sh	S	š	Sh	sh	Ch
ص	ṣ	ṣ	ṣ	ṣ	S	ṣ	Ṣ / ş
ض	ḍ	ḍ	ḍ	ḍ	ḍ	ḍ	d
ط	ṭ	ṭ	ṭ	ṭ	T	ṭ	t / ţ
ظ	ẓ	ẓ	ẓ	ẓ	ẓ	ẓ	z / dh /d
ع	‘ / ‘	‘ / ‘	‘	‘	‘	‘	' / a / â
غ	G	Gh	Ġ	ḡ	Gh	Gh	G
ف	f	F	F	f	F	f	Ph
ق	Q	q	Q	q	Q	q	k / c
ك	K	K	K	k	K	K	C
ل	L	L	L	l	L	l	
م	M	m	M	m	M	M	
ن	N	N	N	n	N	n	
ه	H	h	H	h	H	H	
ة	H / t	a / at	ẗ	h / t	h / t	h / t	
و	W	w	W	w	W	w	ou / o / u / ô / û / ū / ú / ü
ي	Y	Y	Y	y	Y	y	i / ï / î / ī
ى	A	A	ỳ	—	Y	Y	

Tableau 10. 3. Une table de translittération des caractères arabes vers le latin.

Certaines lettres arabes sont transcrites en chiffres. Cette translittération constitue la norme dans le langage SMS en Europe et au Moyen Orient. Cela est très utile pour supposer quelle est l'origine sociale de la personne qui écrit et quelle est l'origine géographique des données extraites (géolocalisation). Le tableau (10.4) suivante récapitule ces chiffres spéciaux :

Lettre	ء	ح	خ	ص	ض	ط	ظ	ع	غ	ق
Équivalence alphanumérique	2	7	'7	9	'9	6	'6	3	'3	8

Tableau 10. 4. Les Equivalences alphanumériques dans les textes écrits en alphabet latin.

Ainsi, en combinant ces deux types de représentation symbolique, on peut rencontrer dans les textes des translittérations qui illustrent ces différentes équivalences pour des noms et des prénoms courants dans le monde arabe (tableau 10.5) :

Nom en arabe	هدى	عديل	حسيبة	طاهر
Equivalents en écriture latine	Houda ou Hoda...	Adil ou 3adil...	Hassiba ou 7assiba...	Taher ou 6ahar...

Tableau 10. 5. Les Equivalences pour des noms et des prénoms courants dans le monde arabe

Cette variation dans les usages translittérationnels, source d'ambiguïté lors du traitement automatique et de la recherche d'information, s'explique par trois types de raisons.

Tout d'abord, des raisons historiques puisque certains pays arabes ont été colonisés ou placés sous mandat français ou britannique pendant une période plus ou moins longue selon les pays et ont, par conséquent, gardé de cette période des traces dans leur vocabulaire, dans leur prononciation et dans la manière dont ils ont tendance à translittérer les noms et les prénoms. Ainsi, l'influence du système linguistique et graphématique du français est perceptible dans les usages translittérationnels des pays du Maghreb, de manière plus ou moins forte selon les pays. Il en est de même des pays du Proche et du Moyen-Orient par rapport à l'influence britannique ou américaine.

Ensuite, pour des raisons politiques puisqu'il n'existe pas de norme commune ni de stratégie unifiée dans le domaine de la translittération pour ce qui est de la langue arabe. Cela a conduit chaque écrivain ou scripteur à s'appuyer sur la prononciation dialectale qui lui était la plus familière pour transcrire les noms arabes. L'exemple le plus célèbre est celui de Laurence d'Arabie qui, pour transcrire le nom de la ville de Djeddah (جدة) en Arabie saoudite, utilise : 25 fois l'orthographe « Jeddah », 6 fois l'orthographe « Jidda », et 1 fois l'orthographe « Jedda », et cela dans le même ouvrage. Laurence d'Arabie justifie cette variation dans la translittération de la manière suivante : « On ne peut pas transcrire correctement et de la même façon un nom arabe à cause des consonnes qui diffèrent des consonnes latines et des voyelles dont la prononciation diffère d'une région à une autre. » (Alsalman et al., 2007). Cela est d'autant plus vrai que les différentes orthographes données par Laurence d'Arabie diffèrent de l'usage actuel en Arabie Saoudite pour la transcription du nom de cette même ville : « Jaddah ».

Enfin, pour des raisons dialectologiques puisqu'il existe une telle variété de parlers régionaux et locaux dans le monde arabe qu'il est impossible de retrouver la même prononciation d'un pays à l'autre et d'une région à l'autre. Ainsi par exemple, l'un des prénoms arabes les plus répandus, celui du Prophète Muhammad (محمد) – transcrit en français Mahomet depuis le Moyen-Âge – possède une dizaine de prononciations – et donc de transcriptions – différentes. Citons notamment : Mohamed, Mouhammad, Muhamed, Mhamed, M'Hamed, Muhammad, etc. Même lorsque ce prénom est voyellé (مُحَمَّدْ), il présente plusieurs translittérations dans les textes : Muhamad, Mouhamad, Mohamad, Mehammad, Mehammade.

Cette variation dans les translittérations possibles selon les dialectes est parfois accompagnée par l'utilisation de caractères spéciaux dans certaines régions ou pays arabes. Citons comme exemples les noms suivants qui présentent des formes non conventionnelles en écriture latine : Mu`ammar, Mabrūk, Musţafá, Ismā`īl, Hâdî.

Tous ces phénomènes nécessitent une observation fine en amont du traitement pour identifier les cas problématiques et construire des règles efficaces permettant l'automatisation du processus de translittération des noms arabes.

10.3.3. Approche proposée pour la translittération de noms propres arabes en écriture latine

Le module de translittération de l'écriture arabe vers l'écriture latine est fondé sur les automates d'états finis pondérés de type transducteurs. Nous avons utilisé l'outil HTFST qui est constitué d'une interface basée sur la librairie open-source OpenFst. Cet outil sert à créer les automates de règles morphologiques, syntaxiques, et autres, et les appliquer ensuite à des textes. HTFST possède aussi une syntaxe propre aux "règles de replacements parallèles et contextuelles" offrant les mêmes possibilités que celles de XFST «Xerox Finite State Tool» (Beesley et Karttunen, 2003) implémentées selon l'approche de FOMA (Hulden, 2009).

Ces hypothèses de transcription peuvent être simplifiées pour la recherche sur Internet, car on sait que, par exemple, les moteurs de recherche éliminent les diacritiques. Cela permet de diminuer considérablement le nombre de requêtes à effectuer.

A la première étape de notre travail, nous avons tenté de vérifier nos hypothèses de translittération en utilisant les règles syntaxiques et contextuelles et sans recours ni à un moteur de recherche, ni à des dictionnaires. Parmi les règles syntaxiques que nous avons considérées dans notre translittération, le fait que le nom arabe ne prend pas en compte la dernière voyelle courte ou tanwin (marqueur du cas) à la fin du mot. Par exemple : زار بلال محمداً, le prénom - محمداًest transcrit par Mohammed et non pas Mohammedan. Le module de translittération de l'écriture arabe vers l'écriture latine tient compte du lien entre la phonologie, la graphématique et la dialectologie en utilisant un certain nombre de règles, issues d'une étude expérimentale.

Ce module permet de générer toutes les formes latines possibles à partir d'un nom propre arabe. Les variantes latines couvrent le plus de langues latines possibles, dont le français et l'anglais.

```
Fonction Algorithme Principal ( NOMS : Mots) :
Chaîne de caractères
  Résultat : Liste de chaînes de caractères
  Pour tout NOM entré faire
    Si (NOM) est non voyellé Alors
      Appliquer les règles contextuelles de la
       translittération.
      Traitement du nom en écriture latine.
      Résultat ← Liste pondérée des noms latins.
     Sinon
      Supprimer les voyelles courtes
      Appliquer les règles contextuelles de la
      translittération.
      Traitement du nom en écriture latine.
      Résultat ← Liste pondérée des noms latins.
    Fin Si
  Fin Pour
  Retourner Résultat
Fin
```

Figure 10. 1. Algorithme de la translittération des noms arabes en écriture latine

Le fonctionnement de notre approche de translittération est déterminé par la nature du mot fourni en entrée : l'automate passe d'état en état suivant les transitions, à la lecture de

chaque lettre arabe de l'entrée (Sâadane et al, 2012). Il peut être décrit par l'organigramme suivant (Figure 10.2):

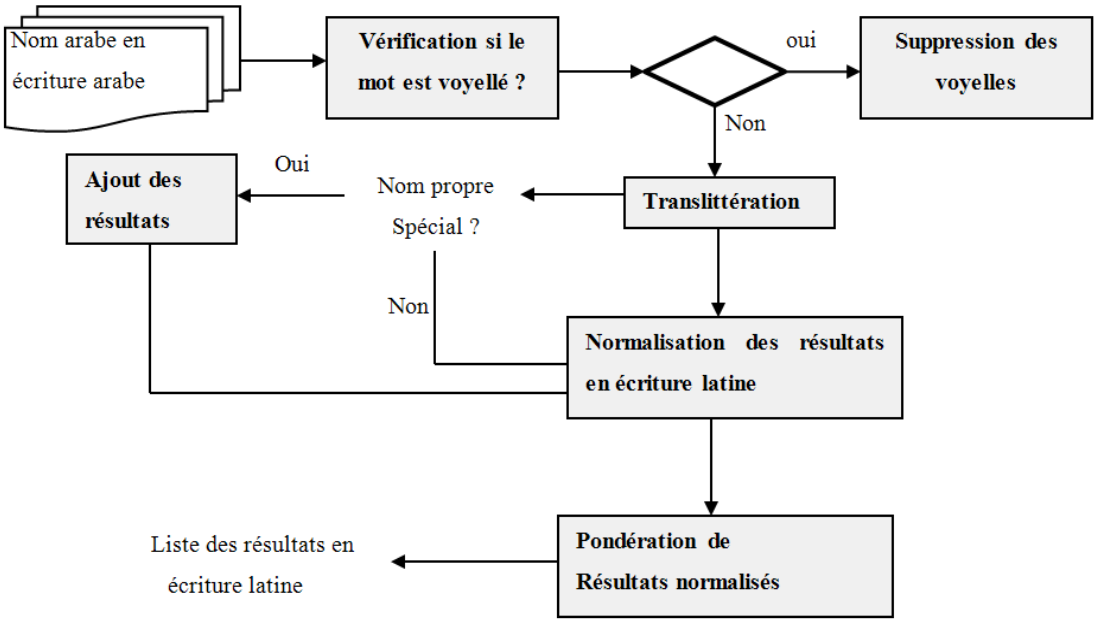


Figure 10. 2. Organigramme du fonctionnement du translittérateur de l'arabe vers le latin

A l'issue de la lecture, un premier automate traite l'entrée de la manière suivante: si l'entrée est voyellée, il supprime les voyelles avant de translittérer le nom; si l'entrée est non-voyellée, il procède directement à la translittération du nom. Nous supprimons les voyelles afin de générer toutes les translittérations françaises et anglaises possibles. Ceci est dû à l'influence des dialectes sur les voyelles où les translittérations des mots issus du dialecte du Macherek sont orientées vers la translittération anglaise et celles du dialecte du Maghreb sont plus orientées vers la translittération française. Enfin, le module produit en sortie une liste triée de noms arabes écrits en caractères latins.

Le cœur du système de translittération est constitué de règles contextuelles. Ces règles permettent le remplacement des lettres arabes en lettres latines ainsi que l'ajout des voyelles latines, en prenant en compte les lettres situées devant et/ou derrière la lettre à ajouter ou remplacer.

Prenons, par exemple, le prénom عبد qui peut être translittéré de plusieurs façons différentes. Nous attribuons un poids pour chaque translittération, sachant que le poids le plus bas indique la solution la plus probable.

R = ((عبد) .x. (((\` A b d <3000>) | A b d | (A b e d <1000>) | (3 (a|A) b d <2000>))).*;

Cette règle indique que, lorsqu'un mot débute par عبد, il est transcrit le plus souvent par « Abd », ou bien moins souvent par « Abed ». Plus rarement, il sera transcrit « 3abd » ou « 3Abd », et dans quelques cas il sera transcrit par « `Abd ».

Les règles contextuelles visent aussi à rendre compte de la manière la plus précise possible des formes observées en entrée : s'agit-il d'une « kunya » ? d'un nom précédé d'un article ? ou bien d'un prénom seul ?

On sait à cet égard que le nom d'une personne contient plusieurs éléments en arabe. Il est constitué en principe de quatre composants principaux, décrits dans le chapitre 4, la section (4.7.1.1)

Selon la forme d'entrée, on applique d'abord des règles adéquates pour transcrire la partie qui ne constitue pas le nom à proprement parler (particules), puis on applique les règles pour la translittération des noms eux-mêmes.

Les règles pour la translittération des noms s'appliquent à leur tour selon le nombre de consonnes du nom considéré, et dans un ordre de priorité déterminé. Par exemple, si le mot est composé par Abd (عبد) + Al (ال) + Nom (رشيد), le système procède de la manière suivante :

- Translittération de la particule عبد« Abd »;
- Translittération de l'article ال« Al »;
- Concaténation de la particule « Abd » et de l'article « Al » en les reliant au nom par un trait d'union ou en insérant un blanc entre les deux : Abd Al Rachid (عبد الرشيد)
- Génération de toutes les formes de translittération possibles pour ces trois éléments (tableau 10.6):

Nom propre arabe	**Translittérations**
عبد الرشيد	Abd Al-Rachid
	Abdul Rashid
	abd al-Rashid
	3abd El Rachid
	abd Al Rashid
	Abdar-Rashid
	Abd Arrashîd
	Abdel Rachid

Tableau 10. 6. Quelques translittérations pour le nom **عبد الرشيد**

Une étape intermédiaire s'ajoute afin de procéder à d'autres traitements, pour ne pas occulter l'un des problèmes très difficile de la transcription, comme la transcription de certains noms propres qui changent totalement phonétiquement pour des raisons religieuses ou autres : c'est le cas de Moussa qui est traduit par Moïse, Yussuf par Josef, Yaakoub par Jackoub, Hawa par Eve, etc. Cette étape consiste à fournir ces transcriptions dans une liste.
Après la génération de la liste triée des noms translittérés, deux types de traitements sont lancés:

- Normalisation de la liste des noms en écriture latine : cette phase consiste à effectuer certains traitements sur la sortie du nom en écriture latine tels que la suppression des caractères spéciaux (diacritiques et chiffres) et l'ajout de la majuscule au début de nom propre, étant donné que les majuscules n'existent pas dans l'écriture arabe des noms. Cette notion de majuscule est conservée seulement dans le cas d'une utilisation dans des bases de données, mais elle n'est pas ajoutée pour les moteurs de recherche usuels, qui ne considèrent pas la casse comme pertinente;
- Pondération de la liste des noms en écriture latine : cette étape consiste à attribuer un

poids aux règles qui ont servi à la génération de la liste, afin de pouvoir afficher les résultats en sortie du plus probable vers le moins probable, ou inversement. Pour réaliser cette pondération, nous utilisons le moteur de recherche Google en notant à chaque fois le nombre d'occurrences pour chaque forme générée du nom propre : par exemple pour le prénom arabe جمال(jamal), le système génère trois translittérations distinctes et attestées dans les textes (Djamel, Jamel, Gamel) et dont le calcul de fréquences via le moteur de recherche Google donne respectivement 4000000, 5500000 et 500000. Du point de vue de la pondération, cet exemple permet de constater que la lettre arabe (ج) est transcrite, en termes de fréquence, majoritairement par la lettre (J), puis par la graphie (Dj), puis par la lettre (G). Cette procédure a été appliquée à toutes les formes de translittération des caractères arabes. Elle a permis d'établir une liste d'équivalences pondérée au niveau des graphèmes, qui sert à afficher les résultats en sortie du plus probable vers le moins probable.

10.4. Transcription des noms arabes en écriture latine vers l'arabe

Après la translittération des noms arabes vers le latin, nous nous sommes intéressés à la translittération des noms arabes écrits en latin vers l'écriture arabe. Cet intérêt est motivé essentiellement par deux raisons pratiques : d'une part, le besoin de garder le même clavier lors de la recherche de noms de personnes arabes, pour des arabophones; et d'autre part, le besoin de faciliter la recherche d'information interlingue et le recueil de données multilingues à partir de noms arabes écrits en caractères latins.

10.4.1. Les problèmes de la transcription du latin vers l'arabe

Parmi les problèmes que nous avons rencontrés, nous pouvons citer (Saâdane et al., 2012):

- ✓ Les différentes translittérations des noms qu'on observe en situation réelle ne respectent pas les règles standards proposées pour la translittération des noms arabes en latin. Ceci influence aussi la translittération des noms arabes écrits en latin vers l'écriture arabe. Par exemple, citons le problème de l'article (ال), avec la non distinction entre les lettres « L Solaire » et « L Lumière ». Pour mieux comprendre, prenons quelques exemples : le nom (Azzamane) est transcrit par (الزمان), par contre le nom (Azziz) est transcrit par (عزيز). Si on prend le nom (Annahar), il est transcrit par (النهار), contrairement au nom (Annwar) qui est transcrit par (أنوار).... Comme le montrent les exemples cités, ce problème se rencontre à chaque fois qu'on a un nom qui commence par la lettre (A) suivie de l'une des lettres Solaires. On constate cette difficulté de transcription surtout si le « L Solaire » n'est pas suivi par un blanc ou un tiret pour la distinguer des lettres (أ) ou (ع), comme le montrent les exemples précédents.
- ✓ La lettre (a), lorsqu'elle apparaît à la fin du mot, peut être transcrite par plusieurs lettres qui sont {ء, ا, ى, ة}. Par exemple, Houda est transcrite par هدىet Karima (au lieu de Karimah) est transcrite par كريمة, ainsi que Wafa (au lieu de Wafaa) qui est transcrite par وفاء.
- ✓ La translittération des noms arabes en écriture latine est touchée par le phénomène de « pluriglossie ». On constate que la translittération des noms arabes écrits en latin vers l'écriture arabe se caractérise aussi par le fait que certaines lettres latines peuvent être transcrites par différentes lettres en arabe. En voici quelques exemples:
    - o La lettre arabe (ث) transcrit en écriture arabe l'une des graphies latines

suivantes : (Th) en Tunisie et en Irak ; (Th) ou (T) en Algérie, selon les régions; (T) au Maroc; (T) ou (S) au Moyen-Orient, selon les pays.

- La lettre arabe (ج) transcrit en écriture arabe l'une des graphies latines suivantes : (J) au Maroc et en Tunisie; (Dj) en Algérie et dans certaines régions du Moyen-Orient; (g) en Égypte. A noter que cette dernière translittération (en Egypte) est source d'ambiguïté puisqu'elle se confond avec la transcription du son dialectal (ڤ), également translitteré par la lettre (g), ainsi qu'avec la translittération de la lettre arabe (ق), prononcée pareillement que le (g) dans certains dialectes.
- Le graphème «aa» est translittérée en arabe par plusieurs lettres en fonction des pays : {, اء}. Ex. dans le prénom «Wafaa».

➢ **Exemple :** Ambiguïté dans la translittération arabe des consonnes latines

Nom en écriture latine	Transcription en arabe
Mu**h**amad	محمد ou مهمد
Houda	هدى ou حدى
Hind	هند ou حند
Ni**h**ad	نهاد ou نحاد
Sa**h**ar	سحر ou سهر

Tableau 10. 7. Les ambiguïtés dans la translittération arabe des voyelles latine.

حامد	هامد
حميد	هميد
حاميد	هاميد
حمد	همد

Tableau 10. 8. Les résultats obtenus pour le prénom Hamid

Outre le phénomène d'ambiguïté, ces exemples illustrent l'absence de stratégie unifiée de translittération en écriture arabe des noms arabes. En effet, les formes translittérées dépendant des usages propres à chaque région ou pays arabe. Elles reflètent l'interaction complexe, au sein du système général de la langue arabe, entre l'arabe littéraire et l'arabe dialectal d'un côté, et entre l'arabe standard et les langues étrangères de l'autre, en l'occurrence le français et l'anglais essentiellement.

10.4.2. Le fonctionnement du module de translittération du latin vers l'arabe

Le module de translittération traite en entrée des noms arabes écrits en caractères latins et produit en sortie une liste de noms propres arabes translittérés en écriture arabe.

Les formes générées en arabe sont des variantes orthographiques qui ne sont pas toutes attestées dans l'usage, étant donné que certains caractères latins peuvent être translittérés par

différents caractères arabes. Mais la confrontation des résultats aux textes du corpus permet de résoudre cette surgénération et de réduire de façon drastique le nombre des formes pour ne retenir que les formes pertinentes pour la recherche de documents.

Nous avons mis en œuvre deux stratégies pour résoudre l'ambiguïté associée à la multiplicité des formes générées pour un même nom.

10.4.2.1. Stratégie 1 : Désambiguïsation par la racine consonantique

Dans cette stratégie, la désambiguïsation est fondée sur la position relative des lettres constitutives du nom, en ayant recours à la racine du mot. En effet, l'ordre des consonnes dans les racines arabes permet de distinguer deux noms ayant les mêmes lettres et de générer ainsi une équivalence probable de chaque lettre du nom en fonction de sa position dans la racine.

Pour établir une telle équivalence, on commence par supprimer les voyelles courtes et longues (و ، ي، ا), si elles existent, pour obtenir la racine du mot (racine consonantique). On crée ensuite une liste des noms en fonction de la position de la lettre dont on veut la translittération exacte. Si la lettre considérée est en première position dans l'ordre des consonnes, la racine est stockée dans la première liste. En revanche, si la lettre en question n'est pas en première position, la racine est stockée dans la deuxième liste.

- **Exemple1 :**

Considérons en entrée le nom «Hatem» (nom arabe translittéré en français; variante : Hatim en anglais).

La translittération de la lettre H est source d'ambiguïté dans le cas présent, car ce nom peut être translittéré en écriture arabe par هاتمou par حاتم.

Pour résoudre cette ambiguïté, le système commence par supprimer la lettre arabe (ا), afin de générer la racine consonantique du nom considéré (حتم).

Il recherche ensuite la racine de deux consonnes (حت) ou la racine de trois consonnes (حتم) dans la première liste.

Réponse : Cette racine existe dans la liste.

Sortie : Accepter seulement la translittération (حاتم) et rejeter la translittération (هاتم).

- **Exemple2 :**

Considérons en entrée le nom «Wahab» (nom arabe translittéré en français : variante: (Waheb).

La translittération de la lettre H est source d'ambiguïté dans le cas présent, car ce nom peut être translittéré en écriture arabe par وهابou par وحاب.

Pour résoudre cette ambiguïté, le système commence par supprimer la lettre arabe (و) et la lettre (ا), afin de générer la racine consonantique du nom (حب). Comme le nom ne commence pas par la lettre (H), le système doit chercher la racine (حب) dans la deuxième liste.

Réponse : Cette racine n'existe pas dans la liste.

Sortie : Accepter seulement la translittération (وهاب) et rejeter la translittération (وحاب).

- **Limites de la stratégie :**

Cette stratégie de désambiguïsation par la racine consonantique a des limites. Il existe, en effet, des contre-exemples dans chacune des listes établies, puisqu'on peut trouver pour chaque racine des noms dont la lettre initiale peut être translittérée par deux lettres arabes différentes tout en ayant un résultat pertinent (le nom produit existe bel et bien mais ne correspond pas à l'original).

Par exemple, pour le nom «Houda», on peut avoir deux translittérations arabes pertinentes : (هدى) mais aussi (حودة). Pour ce type de cas, il a fallu imaginer une stratégie complémentaire de désambiguïsation.

10.4.2.2. Stratégie 2 : Désambiguïsation par le contexte phonologique

Cette solution est basée sur l'identification du contexte phonologique dans lequel apparaissent les lettres que l'on souhaite translittérer.

Les règles inférées permettent de faire la distinction entre deux graphèmes pour une même lettre d'origine.

Considérons l'exemple déjà cité de la lettre H qui peut être translittérée en arabe par les lettres (ه) ou (ح). Parmi les règles inférées à partir d'une analyse du contexte phonologique, citons les suivantes à titre d'exemple :

- ✓ Si la suite phonologique est la lettre {h,s}, le phonème est translitéré en arabe par le graphème (ح).
- ➢ **Exemple** : Houssin : حسين, Houssam : حسام
- ✓ Si la suite phonologique est {h,... i / ee / y /... C}, le phonème (h) est translitéré en arabe par le graphème (ح).
- ➢ **Exemple** : Hamid : حميد, Habiba : حبيبة
- ✓ Si la suite phonologique est {S / M,... h}, le phonème (h) est translitéré en arabe par le graphème (ح).
  - ➢ **Exemple** : Masbah : مصباح, Samah : سماح
- ➢ **Limites de la stratégie :**

Malgré sa robustesse, cette stratégie de désambiguïsation par le contexte phonologique a des limites. En effet, le nombre important de règles nécessaires pour la désambiguïsation rend le système quelque peu lourd et lent dans la phase de traitement, notamment sur de grands volumes de données. Ainsi, si l'application des règles contextuelles permet d'améliorer le fonctionnement du système, elle n'élimine pas pour autant toutes les formes d'ambiguïté constatées. Cela est dû notamment au fait qu'il existe une grande variation dans la translittération des noms arabes même entre les langues utilisant le même alphabet latin.

10.5. Validation des résultats

Au stade actuel du développement du système, le processus de validation des hypothèses et des résultats a été réalisé essentiellement avec des moteurs de recherche sur Internet.

Grâce aux stratégies mises en place, le moteur de recherche permet de récupérer tous les documents pertinents, et cela quelle que soit la langue du document d'origine et quelle que soit la forme du nom propre employée.

Le processus de recherche de l'existence de la translittération par un moteur de recherche (Google par exemple) peut être décrit par l'algorithme suivant :

```
Fonction Rech-Google( NOMS : Mots) : Entier
  Résultat : Entier
  Ouverture de la connexion vers la machine distante
  www.google.fr, dans la socket S.
  Envoie de la requête de recherche (suite de mots) à
   travers la socket S.
  Lecture de la page Web résultat depuis la socket S
  dans un buffer B.
  Si (la chaîne de caractères "Aucun résultat trouvé"
  existe dans B) Alors
     Résultat ← 0;
   Sinon
     Résultat ← N : le nombre de pages Web
       renvoyées  par Google dans le buffer B
  Fin Si
  Retourner Résultat (les pages web)
Fin
```

Figure 10. 3. Algorithme de la recherche des translittérations via *google*

Cette vérification se fait concernant la corrélation entre le mot translitteré (requête) et les documents récupérés pour ce nom. Une translittération est considérée comme pertinente si le résultat des requêtes pour une forme translittérée n'est pas nul, et si, pour chaque forme translittérée, le moteur de recherche récupère au moins une réponse à chaque fois pour une même personne. Considérons l'exemple suivant : requête sur le nom du président algérien.

Nom en entrée : بوتفليقة

Sortie : formes translittérées indiquées en gras dans les documents récupérés à partir d'une dizaine de langues différentes.

1. KING_ SADDRESS AT ARAB SUMMIT IN ALGIERS-By:IMRA
algiers, march 23 (petra-jordan news agency)--his majesty king abdullah ii said that the roadmap peace plan is the only available means to settle the palestinian ... thanks and appreciation for his excellency president abdul aziz **botafliqah** and to the algerian people for their kind hospitality ...
israpost.com/Community/articles/show.php?articleID=5361>Cached
2. The Angry Arab News Service " As'ad
the angry arab news service. a source on politics, war, the middle east, arabic poetry, and art by as'ad abukhalil **...** posted by as'ad at 6:52 am 04/10/09. **butufliqa** wins. the **algerian** president wins re-election with 99.99% of the **...**
angryarab.net/author/falastin>Cached

3. كونا: Arab League congratulates Algeria"s Boutfalika...
cairo, april 11 (kuna) -- the arab league congratulated saturday president abdelaziz boutaflika for his win in the **algerian** presidential elections, hoping that he would continue the

development process in the north african nation. arab league
kuna.net.kw/NewsAgenciesPublicSite/...&Language=en>Cached

4. Times of Oman
it comes in implementation of the directives of his majesty sultan qaboos bin said and president abdulaziz **boutafliqah** aimed at cementing bilateral relations. later at a press conference, alawi said that the two sides signed a number of agreements and mous. **...**
timesofoman.com/innercat.asp?detail=33983>Cached

5. Abdelaziz **Bouteflika** - Wikipedia, the free encyclopedia
bouteflika lived and studied in **algeria** until he joined the front de libération nationale **...** on boumédienne's unexpected death in 1978, **bouteflika** was seen as one of the two main **...**
en.wikipedia.org/wiki/Abdelaziz_Bouteflika>Cached

6. Maliki : If Sultan Hsahim is not executed I will resign
kurdish aspect covers issues related to kurds and kurdistan within the larger context of middle eastern concerns. the website offers readers a **...** he revealed to them that in the opec meeting the **algerian** prime minister abd-al-aziz **botafliqa** had asked him: are you from an **iranian** origin? **...**
kurdishaspect.com/doc020208AWENE.html>Cached

7. ...النهار الجديد- ثورة في عالم الإعلام - لأول مرة ..أسرار عن
. أضف 66 - 1 | عرض: 66المجموع: 0. i love you **boutafli8a** **...** .لأول مرة ..أسرار عن بوتفليقة الرجل تعليقك. اسمك: أضف تعليقاتك: اقرأ أيضا فى: الوطني. وزارة التربية الوطنية تلغي التسجيل في بكالوريا النظام القديم. المتهم الرئيسي في مقتل سارة يواجه تهم الخلوة**...**
ennaharonline.com/ar/national/29309.html>Cached

8. Butaflika fires Benflis, brings back Oyahya, due to Algerian ...
butaflika fires benflis, brings back oyahya, due to algerian presidential elections, algeria, politics. arabicnews.com - your source for daily news about the arabic world. **...** **algerian** news agency quoted benflis as saying after a meeting with **butfalika** that he did not take part in taking the decision of **...**

9. YouTube - viva l'algerie , algeria is back HMD algeria mon amour
algeria is back no matter what **algeria** is still standing ya rab hamdolah , maghrab united ya **botaflika** , les marocain khawatna toujour m3a jazair
youtube.com/watch?v=IAF1AOmnQlM>Cached

10. The Angry Arab News Service " As'ad
the angry arab news service. a source on politics, war, the middle east, arabic poetry, and art by as'ad abukhalil **...** posted by as'ad at 6:52 am 04/10/09. **butufliqa** wins. the **algerian** president wins re-election with 99.99% of the **...**
angryarab.net/author/falastin>Cached

11. Saylac | Somalia News and Information
botofliqa ayaa lagu soo warramayaa inuu si weyn ugala dooday siyuu masfen ciidamada itoobiyaanka ah ee soo **...** mr: musfin waxa uu intaasi ku daray inuu guddoonsiiyay madaxweynaha dalka **algeria** c/caziiz **botofliqa** farriin qoraal ah oo uu uga siday ra'isul

wasaaraha dowladda itoobiya melas zenawi in **...**
saylac.com/news/warJan1907.htm>Cached

12. Boutifliqa the only candidate for Algerian elections
earlier, **boutafliqa** opposed the participation of foreign observers in the elections in order to investigate the "honesty" of any voting process **...** in an interview with **french** television, **boutafliqa** said: "i am a committed nationalist, **...**
arabicnews.com/ansub/Daily/Day/990415/1999041510.html>Cached

13. Answers.com - Algeria Questions including "Do you need a visa ...
algeria questions including "do you need a visa to go to algeria" and "approximately how far apart are the capital of us and algeria" **...** abd al-aziz **botafliqa** عبد العزيز بوتفليقةabdelaziz bouteflika is the president of **algeria**, having taken the...
wiki.answers.com/Q/FAQ/2837>Cached

14. النهار الجديد- ثورة في عالم الإعلام - نقل ابنة يزيد زرهوني نقل ابنة يزيد زرهوني للعلاج بالخارج في حالة خطيرة
... et la bonne santé et je dit aussi mabrouk à tous les algeriens et les algeriennes avec cette victoir ce qui est .**boutoflika** .rabi yoslah halo **...**

15. 葡萄牙每周信息(2006年9月15-22日)
15
使及德国常驻葡大、利亚澳大、爱沙尼亚、斯洛文尼亚、总统分别接受中国尔瓦日席。马其顿等非常驻葡大使递交的国书、菲律宾、黑山、赞比亚、卢旺达、亚利比亚总统尔及利阿boutoflika总统席尔瓦致信葡,拥有的共同战略视角肯定两国,协调两国政治强愿加,。动双方相互合作更加深入推 16日
赞成明年社会党和社民党1。进行全民公决月全国就堕胎合法化
别税总局透露葡海关和特,缉毒行动中在两天的,共逮捕3贩并缴获名毒34。公斤海洛因
财长表示,为使2007财赤降至占年gdp的3.7%,继续紧缩经济明年将,财政开支削减5%,达23.9亿欧元,占gdp的1.5 ...
pt.china-embassy.org/chn/ptyshx/t273560.htm>Cached

16. maliweb.net :: De la rébellion au terrorisme : Ibrahim...
il existe bien une jonction entre le bandit ibrahim bahanga et le réseau al qaïda par le biais du groupe salafiste pour la **...** bahanga avait exécuté le chef des salafistes sur ordre de **boutouflika**. **..**

Chapitre 11 Constitution des corpus

Introduction

Avec la croissance du Web2.0, les gens expriment et partagent de plus en plus leurs opinions à travers les médias sociaux.

Dans le monde Arabe, tandis que l'arabe standard moderne est généralement utilisé dans des contextes écrits formels, les gens utilisent de plus en plus l'arabe dialectal, la langue d'usage quotidien, afin de commenter des artilces, des vidéos ou d'interagir avec la communauté sur des sites. L'ensemble de ces commentaires générés par les utilisateurs offrent une riche source de phrases émanant de différents pays du monde arabe. Ce chapitre présente un corpus arabe multi-dialecte de large échelle collecté à partir des commentaires des utilisateurs sur les vidéos, des journaux, réseaux sociaux, etc. Notre corpus couvre différents groupes de dialectes : l'algérien (AA), le tunisien (TA), le marocain (MA) et l'égyptien (EA). Nous effectuons une analyse empirique sur le corpus et nous démontrons que notre méthode proposée basée sur les mots propres à chaque dialecte ainsi la géolicolisation est efficace pour l'étiquetage des dialectes. L'originalité de notre travail réside dans la constitution des corpus dialectaux écrits en caractères arabes et latins.

Dans la section 11.1, nous présentons un éventail de travaux ayant comme focus la constitution des corpus pour les dialectes arabes. La section 11.2 est consacrée à présenter notre méthode de constitution de corpus de l'arabe dialectal rédigé à la fois en écriture arabe et latine. Lors de cette section, nous détaillons les étapes suivies pour la constitution des corpus, nous avons commencé par étudier les liens que nous allons exploiter pour la constitution des corpus, les outils utilisés pour le téléchargement des pages HTML ainsi que les étapes d'extraction des données.

Dans la section 11.3, nous présentons la méthode utilisée pour l'annotation des corpus et l'identification des dialectes. Nous commençons par présenter les difficultés de l'identification des dialectes. Puis nous présentons les applications de l'identification des dialectes, ensuite nous détaillons l'approche suivie pour l'annotation des textes arabes après l'annotation des textes Arabizi. Cette annotation est faite en deux étapes, la première consiste l'annotation au niveau du mot et la deuxième consiste l'annotation au niveau des textes. La section 11.4 est dédiée à présenter l'interface d'annotation que nous avons développé afin de faciliter la validation des résultats de notre analyse linguistique d'une part et d'annoter manuellement les mots hors vocabulaire afin d'enrichir nos dictionnaires initiaux d'autre part. Enfin, dans la section 11.5, nous exposons quelques traites extraits pour la reconnaissance automatique des dialectes arabes.

11.1. État de l'art sur la constitution des corpus

Depuis plus d'une décennie, la constitution des corpus dialectaux constitue un champ d'investigation très animé qui a attiré l'attention de plusieurs chercheurs. Cette tâche a pour objectif de pallier la carence des ressources en arabe dialectal nécessaires pour le développement d'outils de traitement automatique des langues. Dans la présente section nous présentons un éventail de travaux ayant comme focus la constitution des corpus pour les dialectes arabes.

D'abord, dans (Habash et al., 2008), un guide est présenté pour l'identification du contenu dialectal dans un contenu arabe. Ce guide accorde une attention particulière au phénomène de l'alternance codique (code switching). Ce travail exhibe aussi les résultats d'annotation sur un petit corpus de 59 documents (environ 19 k mots). Les annotations sont faites à la fois au niveau des mots et au niveau des phrases comme suit :

- *Annotation au niveau du mot :* ils ont défini quatre niveaux d'annotation au niveau du mot :

- *Niveau du mot 0 :* utilisé pour annoter les mots qui sont du pur MSA
- *Niveau du mot 1 :* fait référence à un mot en MSA écrit avec une orthographe non standard
- *Niveau du mot 2 :* annote un mot en MSA ayant une morphologie dialectale
- *Niveau du mot 3 :* pour référencer un mot en dialecte.

o *Annotation au niveau des phrases :*

- *niveau du segment 0:* est défini pour des phrases parfaitement MSA
- *niveau du segment 1 :* annote des phrases en MSA mais qui contiennent des phénomènes linguistiques dialectaux comme l'existence d'accord sujet-verbe incorrecte.
- *niveau du segment 2 :* défini des phrases en MSA avec un basculement total vers le dialecte : à ce niveau la phrase peut être formulée en MSA comme elle peut être formulée en dialecte.
- *niveau du segment 3 :* utilisé pour référencer des phrases en dialecte contenant des mots ou des citations issues du MSA.
- *niveau du segment 4 :* référence des phrases purement dialectales.

Nous citons aussi le projet des *Croix-Lingual arabes blogs Alertes –Colaba-* introduit dans (Diab et al., 2010). Ce projet représente un autre effort à grande échelle pour créer des ressources arabes dialectales et des outils pour le traitement associé. Les initiateurs de ce projet se sont basés sur les sources en ligne comme les blogs et les forums, et ont utilisé des tâches de recherche d'information pour mesurer leur capacité à traiter correctement le contenu de l'arabe dialectal. Colaba démontre par ailleurs l'importance de l'utilisation du contenu dialectal lors de la formation et la conception des outils qui traitent l'arabe dialectal. Dans ce projet, quatre dialectes arabes différents sont ciblés : l'égyptien (EGY), l'irakien (IRQ), le levantin (LEV) et le marocain (MOR). Nous remarquons seulement que le focus était mis sur les trois premiers dialectes, alors que l'effort pour le marocain était moins important que les autres dialectes.

(Chiang et al., 2006) ont étudié la construction d'un analyseur syntaxique pour le Levantin sans utiliser une quantité significative de données dialectales, en essayant d'adapter des ressources en MSA pour traiter le Levantin. Ils utilisaient un lexique levantin-MSA disponible, mais qui n'est pas analysé correctement. Leur travail illustre la difficulté d'adapter les ressources de MSA pour une utilisation dans un domaine dialectal.

(Zbib et al., 2012) s'intéressaient à l'intégration de données d'apprentissage dans un système de traduction automatique statistique. Ils ont montré que cette intégration améliore grandement la qualité de la traduction de phrases de dialecte par rapport à un système formé uniquement sur une parallèle MSA-Anglais. Lors de la traduction des jeux de tests égyptiens et du Levant, formés sur 150 millions de mot arabe-anglais (plus de 100 fois la quantité de données de leur dialecte parallèle), il a été constaté qu'un système de traduction MT (Machine Translation) de l'arabe dialectal surpasse un système de traduction du MSA.

La constitution des corpus est aussi importante dans le domaine de la reconnaissance de la parole, où (Lei and Hansen, 2011) et (Biadsy, Hirschberg et Habash 2009) ont mené des recherches sur l'identification de l'arabe dialectal dans ce domaine. (Lei et Hansen, 2011) ont construit des modèles de mélange gaussiens pour identifier trois dialectes qui sont : Golf, Levantin et l'Egyptien. La précision de leurs résultats a atteint un taux d'exactitude de 71,7% en utilisant environ 10 heures de données de parole pour l'apprentissage. Quant à (Biadsy, Hirschberg et Habash, 2009), ces derniers ont utilisé un ensemble de données beaucoup plus

grand, 170 heures de données de parole, et ont adopté une approche basée sur la reconnaissance de la voix et la modélisation du langage. Ce travail s'est intéressé aux quatre dialectes suivants : Golf, Irakien, Levantin et l'Egyptien ainsi que le MSA. Les auteurs ont proposé une méthode de classification atteignant un taux de précision de 78,5%. Nous notons qu'au niveau de ces travaux, utilisant des données de parole, l'identification du dialecte se fait au niveau de l'enceinte et non pas le niveau de la phrase.

Les ressources les plus stables de texte arabe dialectal sont les données en ligne, qui sont formées de manière plus individuelle et moins institutionnalisée, et qui sont donc plus susceptibles de contenir des contenus dialectaux. Les sources possibles de texte dialectal comprennent les blogs, les forums et les transcriptions de conversations. A cause de la nature informelle de conversations en arabe dialectal, la langue est souvent mélangée avec le MSA. Par conséquent, la ligne qui sépare ce qui est dialecte de ce qui MSA est floue. Ainsi, il peut être plus approprié de prendre en considération la tâche d'identification de dialecte au niveau des mots comme dans les travaux de (Elfardy et Diab, 2012a).

Le travail de (Zaidan et Callison-Burch., 2011) s'inscrit dans cette optique et présente une construction de corpus basée sur des ressources web. En effet, les auteurs ont utilisé les sites web de trois journaux arabes : *Al-Ghad* de la Jordanie, *Al-Riyadh* de l'Arabie Saoudie et *Al-Youm Al-Sabe'* de l'Égypte, qui utilisent respectivement les dialectes suivants : Levantin, Golfe et l'Égypte. Ces sites sont parcourus et les commentaires des lecteurs sur les articles ont été extraits et utilisés pour construire un ensemble des données arabes en ligne. Dans ce travail, une intention particulière a été accordée au problème de l'alternance des codes entre le MSA et l'arabe dialectal au niveau de l'inter-phrase. Les auteurs ont rassemblé un large ensemble de commentaires en ASM-AD. Puis ils ont utilisé Amazon Mechanical Turk, via un travail collaboratif de masse, afin d'annoter l'ensemble au niveau des phrases. Ensuite, ils ont utilisé une approche de modélisation de la langue afin de prédire la classe (MSA ou DA) pour une quelconque phrase. Concernant l'annotation des corpus, (Tratz et al., 2013) a introduit des efforts afin d'améliorer l'annotation des corpus arabes, à travers la création d'un outil conçu spécialement pour faciliter l'annotation de données de réseaux sociaux.

Il existe plusieurs autres corpus de dialectes arabes connus comme celui de (Al-Sabbagh et Girju., 2012) qui a créé un corpus d'arabe égyptien avec des annotations et des classifications humaines. Seulement, nous notons qu'un petit sous-ensemble a été humainement annoté afin de former un classificateur pour annoter automatiquement le reste du corpus. Il y aussi l'initiative pour créer un ensemble de données d'arabe dialectal afin de remédier au manque de ressources qui est présentée dans le papier (Cotterell et Callison-Burch, 2014). Les auteurs ont recueilli une quantité importante de données dialectales issues des commentaires des utilisateurs de journaux électroniques et de Twitter, en proposant une extension des travaux de (Zaidan et Callison-Burch, 2011). En effet, les auteurs utilisent une méthodologie similaire à celle de (Zaidan et Callison-Burch, 2011) pour la collecte des données et la classification. Les auteurs ont choisi par ailleurs cinq journaux arabes pour l'extraction des commentaires. Ces journaux sont : le journal égyptien Al-Youm Al-Sabe', Al-Riyadh de l'Arabie Saoudite, Al-Ghad de la Jordanie, Ech Chorouk El Youmi de l'Algérie et Al-Wefaq de l'Irak. Ces cinq journaux permettent de traiter respectivement les dialectes suivants : l'Egyptien, le Golfe, le Levantin, l'Algérien et l'Irakien. Il est à noter toutefois que ce travail se distingue par le nombre de dialectes considérés et l'utilisation de l'annotation humaine. Chaque phrase dans le corpus a été annotée par des humains sur le site Amazon Mechanical Turk. Cette annotation est en nette contraste avec les travaux de (Al-Sabbagh et

Girju, 2012) où seulement une petite partie du travail a été annotée par des êtres humains et le reste par un classificateur. En plus des données, les auteurs fournissent des résultats de la tâche d'identification du dialecte arabe qui sont meilleurs que ceux rapportés dans (Zaidan & Callison-Burch, 2011).

Le travail de (Elfardy et Diab, 2012c) traite aussi de la construction des corpus en fournissant les grandes lignes directrices pour la formation de larges corpus de ressources arabes mixtes à code alterné. En complément du travail précédent, les auteurs de (Elfardy et Diab, 2012b) ont introduit le système AIDA (Automatic Identification and glossing of Dialectal Arabic), pour l'identification, la classification et l'interprétation des dialectes que ce soit au niveau des mots ou au niveau des phrases. Dans AIDA, certaines analyses statistiques et morphologiques sont appliquées pour l'alternance des codes entre le MSA et l'arabe dialectal (AD) dans une même phrase. Suivant le contexte, chaque mot de la phrase est labélisé suivant qu'il est en MSA ou en AD. Le processus de labélisation dépend généralement de l'approche de modélisation du langage, cependant, lorsqu'un mot est inconnu de cette modélisation, sa labélisation se fait à travers le système MADAMIRA qui est un désambiguïseur morphologique. Dans la continuité des travaux sur AIDA, les auteurs de (Elfardy et Diab, 2013) ont présenté une approche supervisée pour l'identification des phrases dialectales et ont aussi étudié les effets des techniques de prétraitement sur la précision des classificateurs.

Il existe par ailleurs des corpus arabes parallèles multi-dialectes comme celui développé dans le travail de (Bouamor et al., 2014). Les auteurs ont construit un corpus qui couvre cinq dialectes : l'égyptien (EG), Le tunisien (TN), le syrien (SY), le jordanien(JO) et le palestinien (PA) en plus de l'Arabe Standard Moderne (MSA) et l'anglais (AN). Afin de construire le corpus, ils ont demandé à quatre traducteurs (locuteurs natifs du Palestinien, Syrien, Jordanien et du Tunisien) de traduire 2,000 phrases de l'égyptien vers leur dialecte. Le choix de l'égyptien comme point de départ était justifié par les auteurs par le fait que c'est le dialecte le plus compréhensible et le plus utilisé à travers le monde arabe. L'industrie médiatique égyptienne a souvent joué un rôle prédominant dans le monde Arabe. Un très grand nombre de productions cinématographiques, de séries dramatiques ou comiques ont depuis longtemps familiarisé l'auditoire Arabe au dialecte égyptien. Un cinquième traducteur, (qui se trouvait être Égyptien) avait pour rôle de traduire les mêmes phrases en MSA. Les auteurs ont ensuite demandé à chaque traducteur de : (a) lire les phrases attentivement et les traduire simplement sans ajouter aucune information; (b) éviter la traduction mot par mot; et (c) avoir une certaine régularité par rapport aux choix orthographiques et d'écriture. Ils ont demandé aux traducteurs d'être cohérents en épelant les mots puisqu'il n'existe pas encore d'orthographe standard disponible pour les dialectes arabes et que les auteurs ont voulu éliminer les parcimonies non utiles.

Dans le même registre, (Salama et al., 2014) ont présenté le *YOUDACC*, un corpus arabe multi-dialecte de large échelle annoté automatiquement et collecté à partir des commentaires des utilisateurs sur les vidéos de YouTube. Ce corpus couvre différents groupes de dialectes écrits en caractères arabes et définis par (Habash, 2010) : L'Egyptien (EG), Le Golfe (GU), l'Irakien (IQ), le Maghrébin (MG) et le Levantin (LV). Dans cette étude, les auteurs ont simplement sélectionné un ensemble de mots clés arabes à l'aide desquels ils ont effectué une recherche sur YouTube. Cette recherche s'intéresse aux vidéos commentées par des utilisateurs du monde arabe. Ensuite les auteurs ont extrait les commentaires des utilisateurs spécifiques à chacun des cinq dialectes étudiés. Pour chaque dialecte, la liste des mots clés est fournie par un locuteur natif décrivant les vidéos qu'il regarde souvent sur

YouTube. Ce qui limite la recherche à la région où un dialecte donné est parlé. Pour chaque mot clé, la vidéo renvoyé par YouTube, ainsi que l'ID de la vidéo et l'URL de la page ont été sauvegardés. A partir de l'URL d'une vidéo donnée, le titre ainsi que les premiers mille commentaires (lorsqu'ils sont disponibles) sont récupérés avec les informations sur l'auteur, l'horaire de publication et les notes des commentaires. Seuls les commentaires des utilisateurs écrits en caractères Arabes sont conservés. L'annotation de chaque phrase/commentaire extrait avec sa classe dialectale correspondante a été réalisée. Cette annotation exploite certaines fonctionnalités de YouTube concernant les profils et ainsi que la liste des mots clés fournis par le locuteur natif.

Dans le même objectif de créer des ressources dialectales pour les outils du traitement automatique des langues, (Masmoudi et al., 2014) ont construit un corpus pour le dialecte tunisien, baptisé TARIC (*Tunisian Arabic Railway Interaction Corpus*). La tâche principale de ce corpus audio tunisien est la demande d'informations sur les services de chemin de fer dans une gare en dialecte tunisien. Ces demandes correspondent aux types de train, ses horaires, sa destination, le prix et la réservation des billets. La création de ce corpus a été réalisée en trois étapes : i) la production d'enregistrements sonores, ii) la transcription de ces enregistrements et, iii) la normalisation. Le nombre final d'heures est de 20 heures d'enregistrements audio avec plus de 4,662 dialogues, 18,657 d'énoncés et 71,684 de mots. Une fois que les enregistrements étaient prêts, la transcription est faite manuellement par trois personnes. En raison de cette transcription manuelle les auteurs ont dû utiliser une orthographe standard car jusqu'à présent, l'arabe tunisien, et le dialecte en général, n'a pas orthographes standards. Cette orthographe est normalisée en faisant recourt à une orthographe conventionnelle pour l'arabe dialectal. Cette orthographe conventionnelle suit les normes décrites dans CODA dédié au dialecte tunisien (Conventional Orthography for Dialectal Arabic : Tunisian Arabic) présentée dans (Zribi et al., 2014).

Les travaux introduits jusqu'à présent concernent la constitution de corpus écrits en caractère arabe. Il existe aussi d'autre corpus conçus pour l'arabe dialectal écrit en caractère latin, dit aussi Arabizi. Seulement, l'Arabizi a la particularité de l'alternance des codes où nous trouvons un mélange entre du texte dialectal écrit en latin avec du texte français (dialectes Maghrébins) ou anglais (dialectes du Machrek). L'Arabizi est très répandu sur Internet et dans les échanges SMS et de ce fait plusieurs utilisateurs utilisent ce type de données latinisées pour la communication médiatisée par ordinateur à travers le monde. Par conséquent, d'avantage d'informations seront générées dans ces conditions, et il est essentiel pour les futurs systèmes de traitement automatique des langues de pouvoir traiter les données produites. Cependant, ces données créent un problème pour les outils standards de la PNL qui sont formés sur le langage avec une orthographe standard. Plusieurs travaux sont à mentionner à ce niveau, ainsi les corpus que nous présentons sont élaborés dans cette direction. Nous notons que dans le cadre de la présente thèse, la pratique langagière de l'arabe latinisé Arabizi, a eu une attention plus particulière, et dès 2013 nous avons développé un corpus qui à notre connaissance, est un précurseur dans la construction des corpus multi-dialectes à code alterné. Ce corpus développé traite les dialectes suivants : l'algérien, le tunisien, le marocain et l'égyptien.

Le travail de (Cotterell et al., 2014) fourni un corpus algérien *Franco-Arabe* à code alterné romanisé (Arabizi) et annoté pour l'identification du langage au niveau du mot. Pour construire ce corpus, les auteurs ont examiné un journal quotidien algérien, en occurrence *Echorouk* qui est le deuxième journal le plus consulté de tous les journaux arabes algériens, et ont extrait les commentaires des des nouvelles. Ils ont pu feuilleter 598,047 pages. Ces

commentaires sont riches en contenu d'arabe dialectal et français. Le corpus contient des discussions sur un large ensemble de problèmes, dont la politique intérieure, les relations internationales, la religion et les événements sportifs. Les auteurs ont extrait 6,949 commentaires, contenant 150,000 mots en total. Ils ont séparé la section des commentaires de l'article principal dans chaque page. Les métadonnées ont été éliminées afin de préserver l'anonymat. Ensuite, les auteurs ont séparé tous les commentaires, dans lesquels plus de la moitié des caractères d'espaces non-blancs sont écrits en alphabet romain, pour déterminer les caractères qui devraient être romanisés. Aucun traitement supplémentaire n'est effectué. Le dernier ensemble de données qu'ils ont obtenu contient 339,504 commentaires avec une longueur moyenne de 19 tokens, après séparation des espaces blancs et de la ponctuation. Environ 1,000 commentaires sont annotés suivant les lignes directrices dans (Elfardy et Diab, 2012a). L'ensemble des données pourrait annoter chaque point de l'alternance des codes avec ces modèles. Cependant, dans ce travail, les auteurs se sont concentrés que sur les points d'alternance des codes. Les annotateurs ont reçu des messages et ont leur a demandés de marquer chaque mot, séparés des espaces blancs et de la ponctuation. Ils ont pu choisir entre l'Arabe (A), le Français (F) et d'autres langues (O). Ils ont exclu la ponctuation de l'annotation.

Dans le même état d'esprit, (Lui et al., 2014) ont proposé un système qui fait l'identification de la langue dans les documents multilingues, en utilisant un modèle génératif de mélange qui est basé sur des algorithmes de modélisation supervisés. Ceci est analogue au travail de (Eskander et al., 2014) en termes d'identification des alternances de codes. Néanmoins, leur système traite l'Arabizi, qui a une orthographe non-standard ayant une grande variabilité ce qui rend la tâche d'identification encore plus difficile et complexe.

Pour ce qui est des outils TALN pour l'Arabizi, (Darwish 2014) a publié une mise à jour (Darwish, 2013), similaire au travail présenté dans (Eskander et al., 2014) car il identifie l'anglais dans le texte en Arabizi et translittère aussi les textes arabes de l'Arabizi en arabe. Pour identifier les mots non-arabes en Arabizi, (Darwish, 2013) utilise des fonctionnalités au niveau des mots et des séquences avec une modélisation CRF (Conditional Random Fields ou "champs markoviens conditionnels") ; tandis que (Eskander et al., 2014) utilisent les arbres de décisions et des Machines à Vecteur Support (SVM). (Darwish, 2013) identifie par ailleurs trois labels : arabe, étranger et autres (pour désigner les adresses E-mail ou les URLs). Le travail de (Eskander et al., 2014) identifient quant à lui un plus grand ensemble d'élément : arabe, étranger, nom propre, son, ponctuation et émoticône. De plus, Darwish (2013) utilise à peu près 5K mots pour former ses labels et 3.5K mots pour les tester; Ceci est considérablement plus limité que l'ensemble d'apprentissage et test constitué par 113K et 32K mots, respectivement dans (Eskander et al., 2014).

Notre travail de constitution des corpus que ce soit en arabe ou Arabizi est proche de l'approche proposée par (Zaidan and Callison- Burch, 2011a) du fait que nous utilisons une méthodologie similaire pour la collecte des données. En ce qui concerne l'identification des dialectes et le traitement des codes alternés, notre approche est similaire à celle des travaux (d'Elfardy et Diab 2012a). Cependant, pour notre travail il est possible d'identifier le type de dialecte associé à un mot donné : dialecte algérien, dialecte tunisien, dialecte égyptien ou marocain. De plus, les labels contiennent des informations sur la catégorie grammaticale des mots à annoter. Au sujet de l'identification de l'Arabizi, notre traitement est aussi proche de celui de (Darwish, 2013) et de (Eskander et al., 2014) où nous utilisons aussi des fonctionnalités au niveau des mots et des séquences en s'appuyant sur des modèles de langage. Les deux travaux identifient seulement l'anglais dans les textes contrairement à notre

travail où nous identifions en plus de l'anglais, le français. A ce niveau, nous identifions l'ensemble des labels suivants : arabe-latinisé, étranger (français ou anglais), nom propre, autres (les adresses E-mail ou les URLs, chiffres, son, ponctuation et émoticône). Nous notons, que chaque mot étranger peut contenir une catégorie grammaticale.

11.2. Constitution des corpus

Le développement d'outils de traitement automatique pour l'arabe dialectal se heurte au manque de ressources linguistiques. En particulier, il n'existe pas de corpus d'arabe dialectal tel que pratiqué en écriture latine dans les réseaux sociaux. Afin de combler cette carence de ressources, nous avons procédé à la constitution de corpus de l'arabe dialectal rédigé à la fois en écriture arabe et latine. Les dialectes auxquels nous nous sommes intéressés sont essentiellement ceux d'Algérie, du Maroc, de Tunisie et de l'Egypte. Ces corpus développés s'alimentent essentiellement de ressources issues d'Internet, qui est à notre égard un gisement important d'informations; ainsi que d'autres ressources générées par des applications comme les transcripteurs de paroles.

La communication en ligne est un domaine de la communication écrite dans laquelle le MSA et l'arabe dialectal sont tous deux couramment utilisés. Ceci est dû à la nature de la communication qui est plus individuelle et moins institutionnalisée que d'autres lieux. Cela rend un dialecte beaucoup plus susceptibles d'être la langue d'usage de l'utilisateur, ce qui est traduit par une forte présence de l'arabe dialectal dans les blogs, les forums, et des commentaires d'utilisateurs / lecteurs, des sites d'information (commentaires des articles et de l'actualité), des réseaux sociaux (Facebook, Twitter, etc), des diffuseurs de vidéos (Youtube, Dailymotion, etc.) et des transcriptions de conversations. Par conséquent, les données en ligne sont une ressource précieuse de texte arabe dialectal, et la récolte de ces données est une étape indispensable pour les linguistes informaticiens pour la création de grands ensembles de données pour être utilisé dans l'apprentissage statistique et la constitution de dictionnaires et de corpus. La figure (11.1) montre l'importance d'internet dans le monde arabe à travers les proportions d'utilisation d'internet au sein des populations arabes. Par ailleurs, il y a aujourd'hui un intérêt crucial à développer des nouveaux outils afin d'exploiter intelligemment cette source inestimable d'information dialectale.

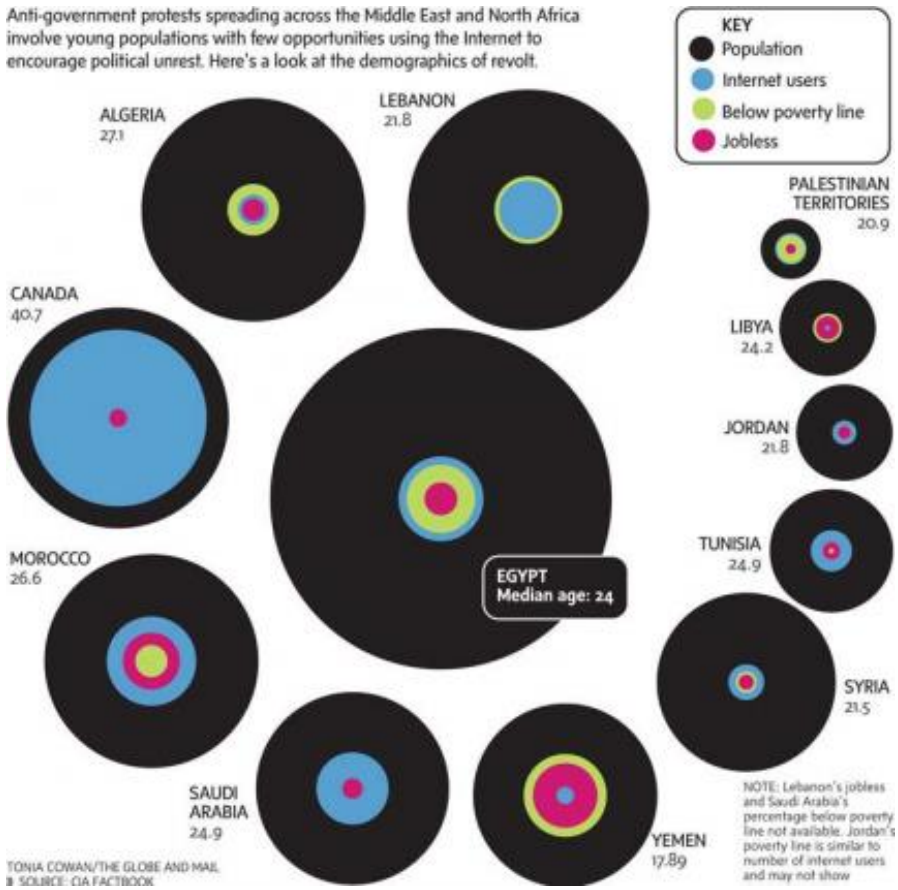


Figure 11. 1. La proportion d'utilisateurs d'Internet dans le monde arabes.

Selon (Zaidan et al., 2011), la constitution de corpus à partir de ressources en ligne en général, et les commentaires des utilisateurs en particulier, présente les avantages suivants :

- Grosses quantités de données avec des fréquences de génération élevée et une disponibilité quotidienne.
- Les données sont publiquement accessibles avec un format cohérent et structuré.
- Extraction facile de ces données.
- Ces données contiennent des échanges dominés par l'utilisation de l'arabe dialectal.

De plus, les données en lignes constituent un lieu d'échange où certaines pratiques langagières se manifestent. C'est le cas de la transcription de l'arabe dialectal en caractère latin (Arabizi), ou le phénomène du code swatching (alternance codique) où deux langues sont utilisées au sein du même texte, par exemple l'utilisation dans le même texte du MSA et de l'arabe dialectal en écriture arabe ou le français, voir l'anglais, avec l'arabe dialectal écrit en caractères latins.

La constitution de ces corpus est faite selon les étapes suivantes :

1. Obtention des liens des sources à utiliser
2. Téléchargement des codes HTML des pages vers lesquelles pointent les liens identifiés
3. Extraction des données à partir des codes téléchargés et création des corpus
4. Annotation des corpus créés
5. Résolution des alternances codiques

Ces étapes seront détaillées dans les sections suivantes et peuvent être résumées dans le schéma suivant :

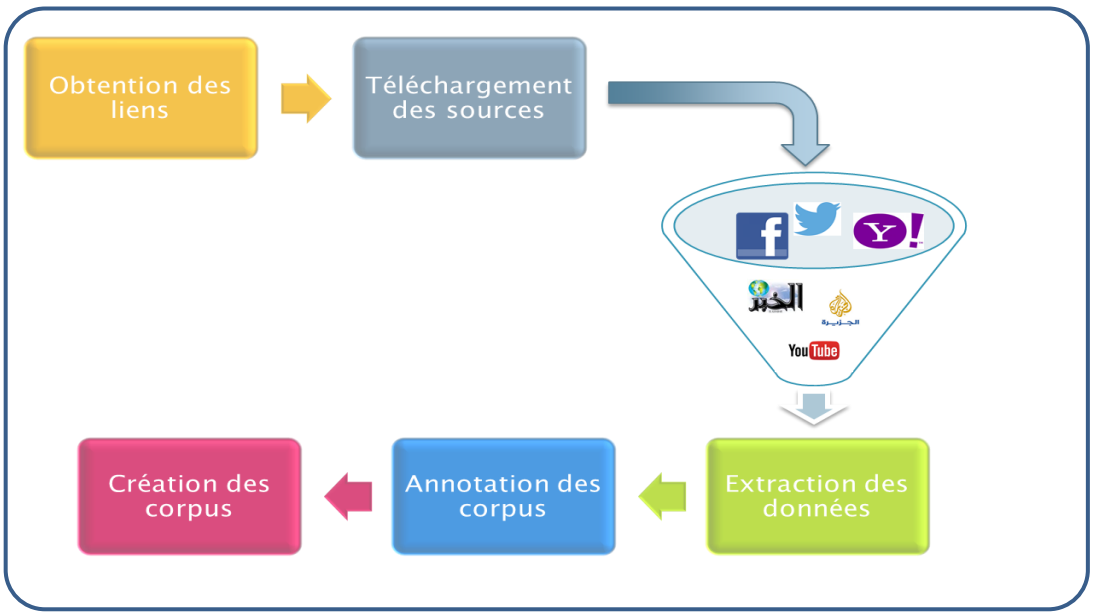


Figure 11. 2. Le schéma d'élaboration des corpus pour les dialectes arabes.

11.2.1. Obtention des liens

A l'instar de Marcel Cohen au début de ses instructions d'enquête linguistique qui considère que « la première chose à faire est de rechercher et délimiter », et de (Meunier, 2009) qui visait et recherchait les liens des sites web alimentant ses corpus oraux ; nous avons défini la première étape de notre processus de construction de corpus comme étant l'identification des objets sources de notre corpus. Ces objets sont majoritairement sur le web, ce qui résume leur identification à l'obtention du lien du site web où se trouvent les objets concernés. Pour la recherche, la collecte et l'identification de ces données à partir du web, nous avons ciblé des sites web susceptibles de contenir des échanges et textes formulés avec le dialecte tel que les journaux, les réseaux sociaux, les forums, etc. Ces liens pointeront vers des contenus à utiliser dans la construction des corpus caractérisant chaque type de dialecte.

Pour réaliser cette étape, nous avons utilisé des moteurs de recherche. Plus précisément, la recherche des liens est faite en sélectionnant et utilisant, à titre de requêtes, un ensemble de mots clés arabes écrits en caractères arabes et latins. Ces mots clés sont des mots dialectaux, fournis par de locuteurs natifs et spécifiques à chacun des quatre dialectes étudiés, à savoir l'algérien, le tunisien, le marocain et égyptien. Pour chaque mot nous avons effectué une recherche sur le web avec un moteur de recherche ensuite les résultats obtenus sont analysés afin d'isoler les sites contenant le plus de commentaires exprimés en dialecte, quel que soit sa transcription en caractère arabe ou latin. Nous précisions aussi que le type des sites issus lors de la recherche (journaux, forums, etc.) n'est pas déterminant dans l'identification du lien. Ces étapes de recherche sont résumées dans le schéma suivant :

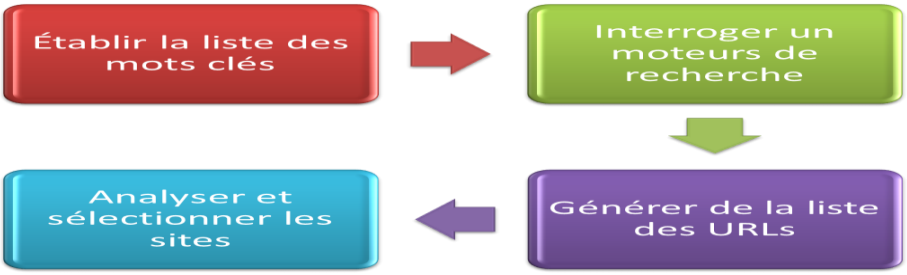


Figure 11. 3. Le processus de recherche et sélection des sites

Par ailleurs, nous signalons que les sources censées alimenter notre corpus concernent divers sujets : politique, sport, économie, etc. Ceci nous permet de brasser différentes populations et d'étudier le maximum possible de pratiques langagières des internautes. L'ensemble des contenus concernés par cette étape de construction est présenté comme suit :

11.2.1.1. La presse

La presse, comme tous les domaines, a été influencée par la révolution numérique et technologique que connait notre monde. De ce fait, les pratiques langagières associées sont devenues complexes et font appel à des connaissances interdisciplinaires pour avoir un angle d'approche des réalités relatées à travers les vecteurs numériques. Cette évolution de la presse fait émerger de nouvelles problématiques linguistiques suscitant de plus en plus l'intérêt des chercheurs pour les langues dans le domaine médiatique. Selon (Chachou 2013), ces problématiques impliquent la diversité linguistique dans la presse et la publicité qu'elle véhicule, la combinaison des acquis des sciences du langage et des sciences de l'information et de la communication, et l'analyse de la pertinence des contenus des corpus nécessitant parfois la connexion avec d'autres disciplines.

Pour appuyer ce constat, prenons l'exemple de la presse en Algérie : cette dernière est un terrain très intéressant qui fait émerger les problématiques citées ci-dessus. En effet, elle essaye de s'adapter aux différentes mutations socioéconomiques et politiques qui régissent le pays depuis quelques années. Cette adaptation se manifeste par une expansion des espaces dédiées à la publicité qui étaient très réduits à l'époque du socialisme mais ont vite augmenté après l'ouverture sur l'économie de marché. Ajouté à cela, le bilinguisme de cette presse, aligné sur celui de l'état, a permis, surtout dans les années 2000, de développer de nouvelles formes linguistiques et des pratiques innovantes spécifiques au genre publicitaire : par exemple l'utilisation du dialecte dans la formulation des slogans comme '*flash flash fi eddeniya kifou makache*' (Flash Flash n'a pas d'équivalent dans la vie).

De manière générale, la langue officielle utilisée dans les journaux arabophones est l'arabe standard. Toutefois, l'arabe dialectal est employé dans certains messages publicitaires, les caricatures, les chroniques, les discours rapportés et dans certains titres d'articles ou de rubriques. Ces domaines d'utilisation sont pratiquement les même dans le dialecte Algérien (Chaouche), Marocain (Catherine Miller, 2012)[19] et égyptien (Ibrahim, 2010). Cette tendance

[19] « La montée de l'usage de l'arabe dialectal : Bien qu'il n'existe pas encore d'étude comparative sur ce phénomène, on constate que l'usage de l'arabe dialectal dans la presse quotidienne ou hebdomadaire arabophone reste, dans son ensemble relativement circoncis. On note ainsi que de nombreux journaux ont tendance à mettre des titres ou sous-titres en dialectal alors que le corps de l'article restera largement MSA » (Cf. Ibrahim 2010 pour certains journaux égyptiens, Miller, sous presse pour les journaux marocains) (Miller, 2012).

d'utiliser le dialecte a pour objectif de se rapprocher davantage des populations arabes pour qui, dans la majorité d'entre elles, l'arabe dialectal représente la langue maternelle et d'usage quotidien contrairement à l'arabe standard réservé au études et aux communications officielles. Bien que cette tendance soit justifiée, elle occupe des espaces réduits dans la presse écrite, contrairement à la presse électroniques, où cette utilisation du dialecte est favorisée notamment par l'usage des commentaires d'actualités, souvent rédigés en dialecte. Par exemple dans le dialecte algérien, des mots du dialecte sont carrément utilisés à l'intérieur des articles écrits en arabe standard, notamment dans la presse sportive. C'est le cas par exemple des mots '*Hogra*'(abus de pouvoir), '*Harraga*'(immigrants clandestins), '*Trabendo*' (trafic de bande), voir des phrases mélangeant à la fois arabe standard et dialectal.

La publicité est le domaine de la presse où l'arabe dialectal est utilisé à grande échelle. Ceci est du au fait que la publicité a pour objectif de communiquer des messages au maximum possible de personnes, et il est évident de parler avec la langue de la masse qui est le dialecte. Outre le dialecte, nous trouvons aussi le MSA, le français, l'anglais et certains phénomènes d'alternances codiques. La presse arabe en général utilise, comme mentionné ci-dessus, un mélange de l'arabe standard (voir même du français ou anglais) et dialectal. Ce mélange représente le phénomène d'alternances codiques. Ce phénomène a fait naître le concept d'arabe médian qui concerne les conversations orales qui est devenu fréquent dans le domaine audiovisuel et la presse écrite, mettant en avant de nouvelles pratiques langagières. Selon (Miller, 2008, 386) : « *cette variété d'arabe emprunte sa phonologie, sa syntaxe et sa morphologie principalement au dialecte, mais une partie de son lexique (et quelques traits phonologiques et grammaticaux) a l'arabe moderne standard* ». Cette variante est très présente actuellement chez les animateurs et journaliste des chaînes de télévision et des radios.

Par ailleurs, cet arabe médian est caractérisé par des traits linguistiques, d'ordre phonologique, lexical et grammatical, permettant de l'identifier. Par exemple, le pronom interrogatif '*waqtâsh*' et '*imtta"* (quand) sont produits en arabe médian en usage dans l'arabe algérien, tunisien et marocain pour le premier et l'égyptien pour le second. Le verbe '*nal'ab*' est emprunté du MSA mais soumis à la conjugaison de l'arabe dialectal maghrébin où nous notons l'utilisation du préfixe 'n' au lieu de 'a' utilisé en MSA pour indiquer la première personne du singulier.

11.2.1.2. Réseaux sociaux (Facebook)

Par définition un réseau social est un ensemble de sites permettant à des communautés d'amis ou de personnes de communiquer efficacement. En plus de la communication, ces réseaux fournissent à leurs utilisateurs des outils et des interfaces d'interaction facilitant la communication entre les membres du réseau et offrant de nouvelles exploitations des informations échangées sur le plan personnel (outils de recommandation) et public (marketing). Les informations échangées concernant donc des aspects personnels, voir intimes, des membres du réseau et des aspects plus génériques brassant différents domaines : économie, culture, politique, etc.

En ce qui concerne le monde arabe, ce type de réseau est très répandu, notamment Facebook, et y a joué un rôle très important dans les changements politiques qu'a connu cette région du monde depuis 2011. De ce fait, les échanges effectués sont écrits avec le dialecte local pour ce qui est des communautés du même pays et avec un mélange arabe standard et dialecte pour les échanges entre pays pour contaminer les autres pays avec le printemps arabe. Quantitativement, le nombre d'utilisateurs arabes des réseaux sociaux a atteint 43 millions en

2012 pour seulement le réseau Facebook. La figure ci-après donne la distribution de ce nombre sur les différents pays arabes en 2011 :

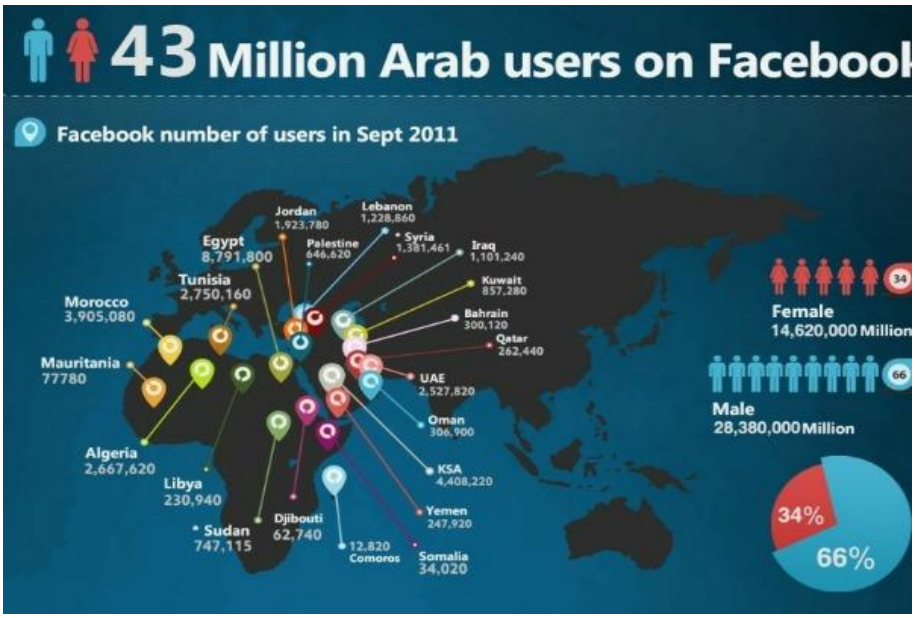


Figure 11. 4. La répartition des utilisateurs arabes de Facebook

Ces chiffres ne viennent que confirmer la place des réseaux sociaux et leur pénétration dans le monde arabe. Les contenus échangés dans ces réseaux sont des images, vidéos et surtout des commentaires et messages écrits avec le dialecte, ce qui représente pour nous une source cruciale de contenus pour l'alimentation de nos corpus dialectaux. Cette prise en compte de ces réseaux sociaux dans la constitution des corpus peut être justifiée par les éléments suivants :

- La quantité des échanges contenus dans ces réseaux
- La diversité des sujets abordés dans ces échanges
- L'utilisation du dialecte arabe dans les échanges avec plusieurs pratiques langagières

11.2.1.3. Youtube

La vidéo est un moyen efficace de communication, d'expression d'opinions et de transmission de message et d'information. De ce fait, ce moyen est très utilisé dans les réseaux sociaux associés avec une nouvelle génération de services internet. Youtube fait partie de cette génération de services en offrant une plateforme de partage de vidéo, généralement courtes, avec un accès relativement rapide. De plus, Youtube donne la possibilité d'associer aux vidéos des informations supplémentaires sous forme de commentaires et d'évaluation. Selon les dernières statistiques, Youtube est le troisième site le plus visité au monde avec un capital d'utilisateur atteignant 1 milliard. Le volume des vidéos stockées sur ce site ne cesse de croître avec une vitesse qui a atteint les 300h de vidéos publiées par minute, occupant ainsi 20% de l'ensemble du trafic HTTP, ou à peu près 10% de l'ensemble du trafic sur internet (Cheng et al., 2007). De ce fait, Youtube est l'un des médias participatifs les plus larges et les plus populaires dans l'environnement en ligne.

Plus précisément, Youtube offre la possibilité aux utilisateurs connectés de fournir des tags de catégories ou ajouter des commentaires, écrits dans plus de 61 langues, comme

réponse à une vidéo. Ces possibilités ont contribué à augmenter la cote de cette plateforme ainsi sa croissance dans le domaine du partage de contenus en ligne. Youtube est devenu même un outil journalistique utilisé pour informer les personnes là où les moyens classiques ne peuvent pas accéder. Cette dernière exploitation est largement perçue dans les mouvements de la révolution arabe, où Youtube est devenu carrément une arme de propagande entre les mains des protagonistes des différentes révolutions syrienne, libyenne, tunisienne, etc. Cette utilisation est favorisée par l'absence de censure assurée par Google qui est le propriétaire de Youtube.

En effet, les commentaires et les tags ajoutent une dimension supplémentaire à la vidéo, enrichissent son contenu, et créent autour d'elle un espace d'échange mettant en relation différentes communautés d'utilisateurs. Dans le monde arabe, ces espaces d'échanges sont caractérisés par une utilisation massive du dialecte au détriment de l'arabe standard qui est généralement réservée pour les contextes écrits (ex., les communications gouvernementales, les cours académiques, etc.). Et vue la quantité de vidéos visionnées et le nombre d'utilisateurs arabes connectés, Youtube constitue une source riche de contenus écrits en arabe dialectal de différents pays : Algérie, Tunisie, Maroc, Egypte, etc. Selon (Salama et al., 2014), l'examen des pratiques de partage sur YouTube montre que la communauté Arabe est très active en ce qui concerne la création des chaînes, le partage et les commentaires des vidéos. La figure 4 illustre la distribution des commentaires arabes sur YouTube par pays.

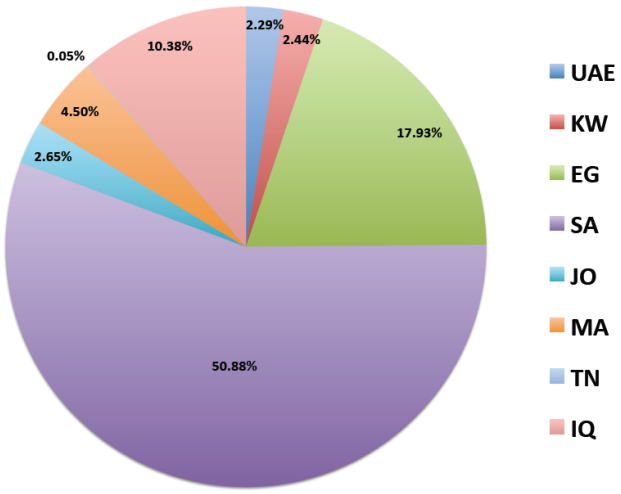


Figure 11. 5. La distribution des commentaires arabes sur YouTube par pays

Par ailleurs, les commentaires de la vidéo sont généralement structurés comme suit :

- Titre de la vidéo
- Un contenu
- L'utilisateur ayant généré le commentaire avec possibilité d'accéder à son profil qui contient plusieurs informations sur l'utilisateur, comme sa profession, sa localisation géographique, etc.

Comme les commentaires sur les vidéos peuvent être issus de plusieurs communautés d'utilisateurs, ils peuvent être rédigés ainsi en plusieurs dialectes et langues. Pour faire la distinction entre les différents dialectes, l'utilisation des informations sur la localisation géographique des utilisateurs contenues dans les profils constitue une première piste pour identifier le type de dialecte.

11.2.1.4. Forums de discussion

Un forum de discussion est un media de communication permettant à un ensemble d'utilisateurs (internautes) d'échanger des messages de manière asynchrone. Les échanges effectués concernent un sujet bien identifié, qui est utilisés pour classer les messages. Le forum est généralement composé de différents fils de discussion (sujet de discussion), où le premier message donne une introduction au sujet abordé dans le fil.

Les fora sont considérés comme l'une des applications principales d'internet et offrent un moyen d'information et d'échange efficace. De plus, ils permettent de créer des espaces de communication pour des groupes de personnes partageant les mêmes centres d'intérêt et avec la langue ou le dialecte qui leur convient. De ce fait, les forums de discussion sont une source riche pour constituer un corpus idéal afin d'analyser les pratiques langagières des internautes. Les raisons suivantes (Khadraoui, 2010) renforcent notre prise en considération des forums lors de la constitution des corpus :

- La consultation est permise pour chaque visiteur du forum tandis que la participation nécessite généralement une inscription
- L'enregistrement automatique des messages des forums facilite la constitution du corpus.
- Ce type de corpus est caractérisé par son homogénéité pour sa mise en mémoire et par le dispositif qui assure cette mise en mémoire
- La liberté des utilisateurs dans la proposition des sujets. Toutefois, il est impératif de respecter la catégorie appropriée au type du sujet ;
- La possibilité de collecter toutes les discussions faites sur le même sujet

Dans le monde arabe, ces espaces d'échanges sont très utilisés par les internautes arabes pour parler de différents sujets comme le sport, le cinéma voir la politique. Les échanges sont généralement effectués en dialecte transcrit en caractères arabe et/ou latin avec parfois l'introduction de chiffres pour certaines lettres. Ces échanges se sont multipliés ces derniers temps avec l'avènement du printemps arabe afin de permettre aux différentes communautés d'avoir des informations utiles, comme l'a fait Anonymous pour informer les tunisiens sur les moyens et les méthodes pour contourner la censure, et partager des expériences militantes. La quantité des messages contenus dans ces forums est très importante et constitue de ce fait un gisement intéressant pour notre corpus de dialecte arabe.

11.2.2. Téléchargement des pages HTML

A l'issue de l'étape précédente, nous avons obtenu une liste d'URLs. Chacune d'elle est donnée en paramètre à un aspirateur de site Web afin de télécharger et de traiter les documents vers lesquels pointent les liens obtenus. L'objectif de ce traitement est d'obtenir les commentaires et les pages structurées afin de les insérer dans un corpus.

Dans cette section, nous décrivons l'outil utilisé pour collecter des données linguistiques pour les quatre dialectes. Nous avons choisi les sites au fort contenu rédactionnel pour collecter ces données. Il s'agit par exemple des sites Web de journaux en ligne, réseaux sociaux, forums, etc. Ensuite, nous avons collecté et extrait les ressources depuis ces sites repérés. Pour réaliser cette tâche, nous avons choisi l'outil HTTrack qui est un aspirateur de site Web open source permettant de copier tout le contenu d'un site sur un support local. Il permet de récupérer la structure originale du site ainsi que tous les fichiers (HTML, images, sons, etc) constituant le site analysé. Dans notre contexte, nous avons utilisé cet outil afin de télécharger les pages Web des différents sites arabes.

11.2.2.1. Présentation du logiciel HTTrack

HTtrack est un aspirateur de site web, c'est-à-dire un logiciel qui sert à télécharger tout le contenu d'un site internet, selon les droits d'accès accordés aux fichiers du site, pour en avoir une "copie" sur son ordinateur et ne plus avoir besoin d'internet pour y accéder. L'utilisation de cet outil est faite selon les étapes suivantes :

1. ***Lancement du l'outil :*** consiste à choisir la langue de travail ainsi que le démarrage d'un nouveau projet d'aspiration

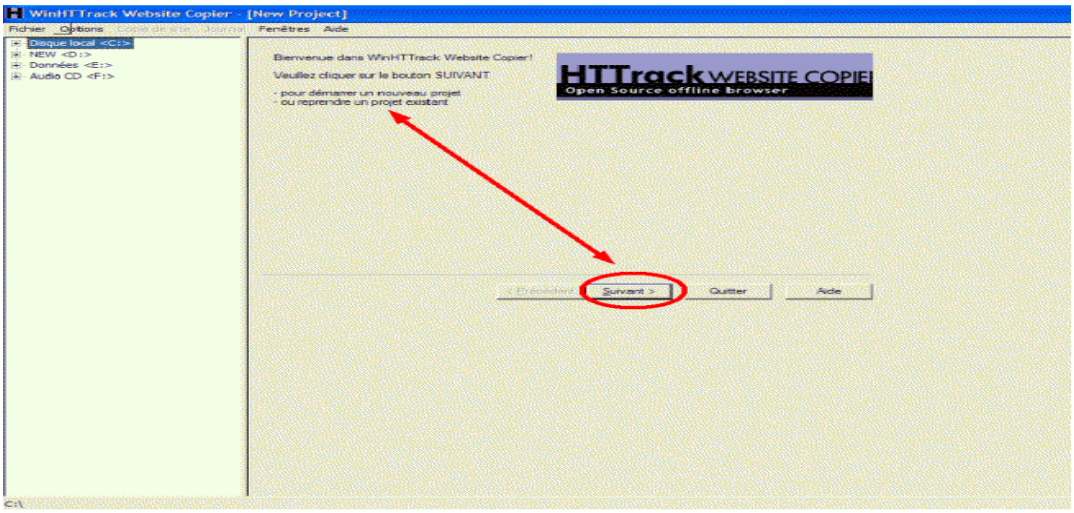


Figure 11. 6. Le lancement du l'outil.

2. ***Paramétrage du projet :*** dans cette étape nous affectons à notre projet un nom et une catégorie avant de spécifier le chemin ou l'emplacement de l'enregistrement des fichiers du site aspiré sur l'ordinateur cible.

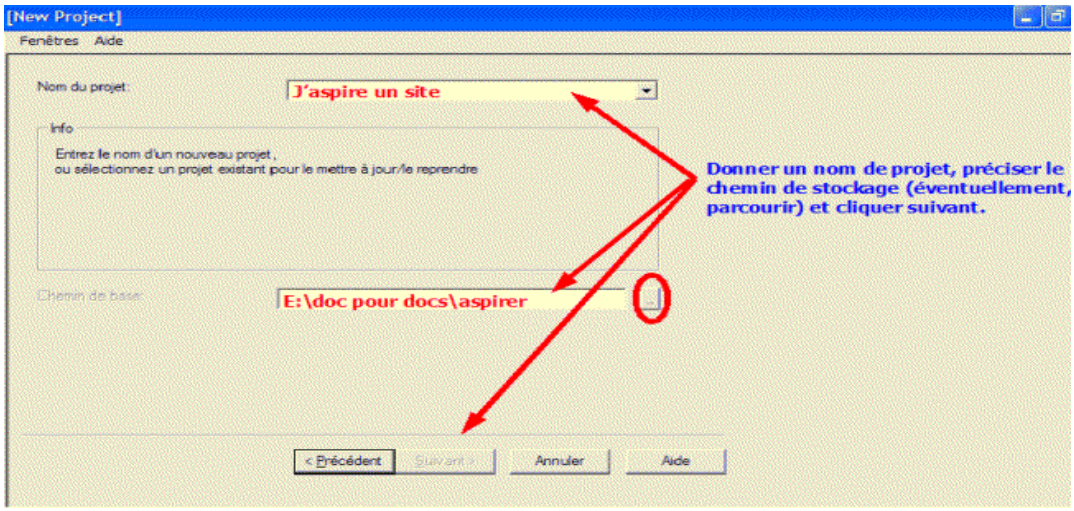


Figure 11. 7. Le paramétrage du projet.

Le résultat de cette étape est la création d'un répertoire portant le même nom que le projet définit, dans l'emplacement spécifié, avec un fichier nommé 'index.html'

contenant la liste des projets aspirés.

3. ***Définition de l'URL de la source :*** cette étape consiste à définir le(s) lien(s) du site à aspirer ainsi que les paramètres de copie du site. Parmi ces paramètres, nous définissons les règles de filtrage

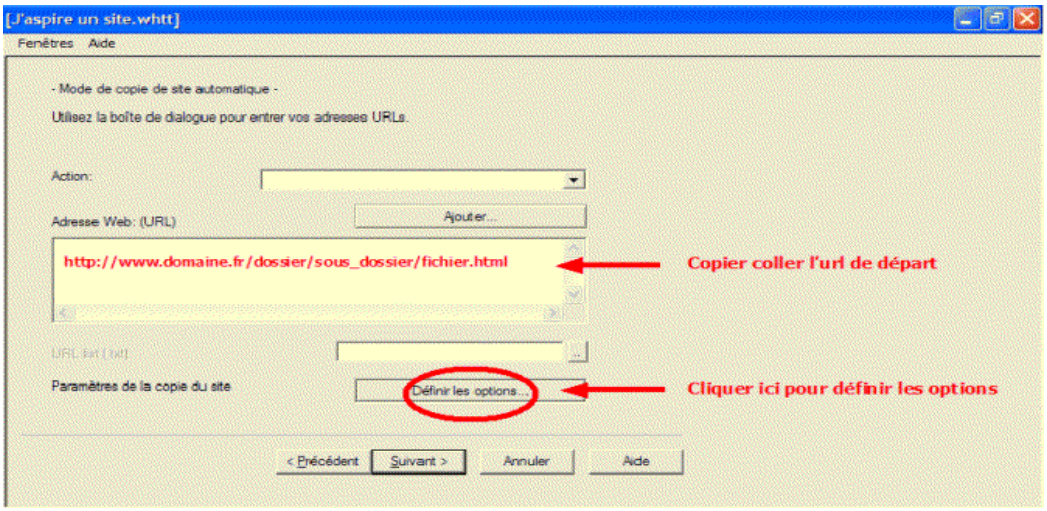


Figure 11. 8. La définition de l'URL de la source.

Par défaut, tous les types de fichiers sont aspirés, cependant dans certains cas nous pouvons exclure ou ajouter d'autres types en fonction des contenus ciblés. Dans l'exemple ci-dessous, le logiciel aspirera les fichiers mp3 (+*.mp3) et exclue les fichiers exe (-*.exe). Nous pouvons également inclure ou exclure des sous-dossiers, des liens ou encore définir des mots-clés à inclure ou exclure.

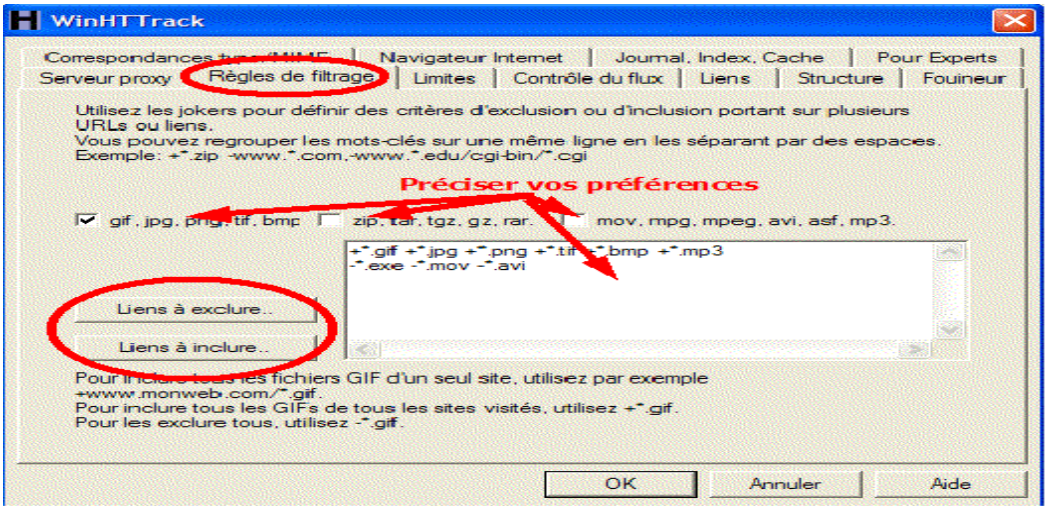


Figure 11. 9. La définition des règles de filtrage

11.2.3. Extraction des données

La construction de nos corpus multi dialecte, comme indiqué dessus, est basé sur le traitement des données contenues dans les pages html issues de l'exploitation de l'activité des utilisateurs et leurs interactions. Ces interactions sont capturées à travers les commentaires postés sur les différentes sources et contenus considérés dans notre étude, à savoir les journaux, les forums, les vidéos, etc. Les corpus construits couvrent différents groupes

dialectaux où chaque message est munie d'un label indiquant le dialecte correspondant.

Un exemple d'exploitation des données contenus dans les sources de nos corpus, prenons les sites d'information. Sur ces sites, un commentaire possède un titre, un contenu et un auteur. Chaque auteur a un profil qui contient plusieurs informations sur son identité, ses loisirs, sa profession, sa localisation géographique, etc. Dans notre étude, la localisation géographique de l'utilisateur peut être exploitée pour attribuer à chaque commentaire un label indiquant sa classe dialectale qui correspond généralement à la zone géographique depuis laquelle le commentaire a été saisi.

De manière générale, l'opération d'extraction consiste d'abord à analyser les codes HTML des pages web considérées afin d'isoler les métadonnées et les commentaires des utilisateurs contenus dans ces pages. Cette extraction est faite grâce à un ensemble d'algorithmes que nous avons développés en Java. Chacun des algorithmes développés est spécifiques à un site web analysé car les données et informations nécessaires pour notre corpus ne sont pas formatées de la manière dans le code HTML des sites web considérés. Après cette étape, les codes HTML isolés seront normalisés afin de respecter une structure uniforme et cohérente pour tout le corpus. Le contenu du corpus est créé lors de cette étape en analysant le code HTML des contenus téléchargés afin d'extraire les informations suivantes si elles existent :

- séparation des messages des différents rédacteurs,
- indication du message auquel le message courant répond,
- l'URL et le nom du site téléchargé,
- la date et l'heure du commentaire,
- l'ID de l'auteur, le nom ou le pseudonyme,
- le sous-titre,
- l'adresse mail,
- la localisation géographique de l'auteur,
- le contenu du message.

Notons que selon l'architecture du site web traité, nous pouvons être ramenés à traiter plusieurs fichiers HTML contenus dans plusieurs répertoires. La figure (11.10) résume le processus d'extraction des données.

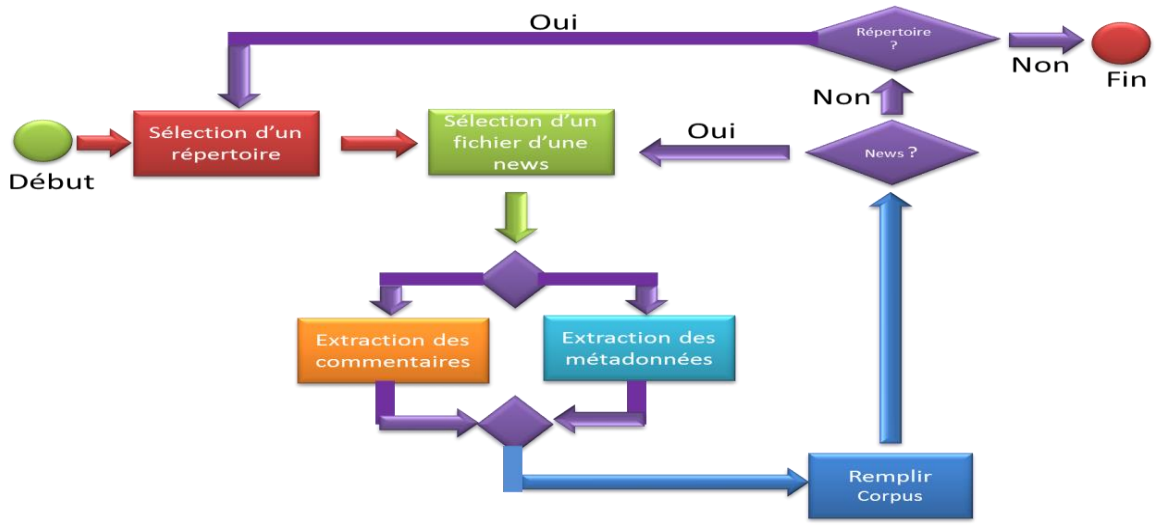


Figure 11. 10. Le processus d'extraction des données.

Par exemple, prenons le contenu HTML d'un message récupéré d'un fichier téléchargé du journal algérien El Khabar :

```
…. CURRENT_URL:
http://www.elkhabar.com/ar/autres/athman_snadjki/240186.html ...
<div class="comment_holder">
  <a name="comment_46854"> </a>
  <div class="comment_header">1 - RABIE</div>
  <div class="comment_header_pays">MARSEILLE</div>
  <div class="comment_header_time">2011-01-0113:31 مر على</div>
   <div class="comment_body_holder">
     <div class="comment_body">
```

Voici le contenu du message obtenu après l'application des programmes de récupération des informations sans annotation mais avec les métadonnées liées à la récupération des données:

```
<doc docid="elkhabar_comment1"
articleURL="http://www.elkhabar.com/ar/autres/athman_snadjki/240186.html" author="1-
RABIE" pays="MARSEILLE" date="2011-01-01" time="13:31">
<comment> ALLAH YARHMEK. INA LILLAH WA INA ILAYHI RAJI3OUN </comment> </doc>
```

11.3. Annotation des corpus et identification des dialectes

Dans la communauté de la PNL, le traitement du texte informel est devenu un domaine d'investigation extrêmement populaire chez les chercheurs (Yang and Eisenstein, 2013). L'identification du dialecte est une des tâches, réalisées dans ce domaine, qui consiste à identifier si une phrase écrite en arabe contient du contenu dialectal. Cette tâche peut être sophistiquée en recherchant la variété du dialecte si ce dernier est identifié dans une phrase, ceci nécessite un niveau d'analyse plus fin. Par ailleurs, l'identification des dialectes peut être considérée comme étant un cas difficile de l'identification des langues où selon (Zaidane et al., 2011) elle est appliquée à un groupe de langues étroitement apparentées qui partagent un ensemble de caractères communs. Cette identification des dialectes est compliquée par l'utilisation des codes alternés au niveau du mot et du texte. Pour la communauté linguistique, une analyse de l'alternance des codes basée sur un corpus offre la possibilité de tester différentes hypothèses sur une large quantité de documents.

Par ailleurs, l'identification de la langue ou du dialecte est précédée par une étape d'annotation pour identifier le type de transcription utilisé pour le dialecte analysé, ex : arabe standard, arabe latinisé, français, etc. Cette annotation est aussi compliquée par l'utilisation des codes alternés, ceci engendre des situations où plusieurs codes linguistiques peuvent exister au niveau du texte voir du même mot. Les noms de lieux en forment un simple exemple : (Paris) est un mot en MSA qu'on trouve dans la plupart des dictionnaires arabes, mais qui est clairement d'origine française. D'autre part, (vidéo) est un exemple des emprunts récents des mots européens qui devraient être considérés comme des noms arabes. Afin de simplifier la décision, nous avons suivi les grandes lignes directrices d'annotation des

dialectes Arabes fournies dans (Elfardy et Diab., 2012) et dans (Eskander et al., 2014).

11.3.1.1. Les difficultés de l'identification des dialectes

Distinguer et séparer automatiquement les dialectes n'est pas une tâche facile car toutes les variétés arabes utilisent le même jeu de caractères, et beaucoup de vocabulaire est partagé entre les différentes variétés. Autrement dit, identifier le dialecte dans une phrase n'est pas simplement une question de construire un dictionnaire du vocabulaire du dialecte et ensuite détecter si une phrase donnée contient un ou plusieurs mots appartenant à ce dictionnaire. Cette source d'ambiguïté au niveau du mot est provoquée par plusieurs facteurs (Zaidan et al., 2011) :

- Une phrase dialectale peut être entièrement composée de mots qui sont utilisés dans tous les variétés arabes y compris le MSA.
- Certains mots sont utilisés dans les variétés avec des fonctions différentes. Par exemple, Tyb est utilisé dialectalement comme interjection, mais comme un adjectif dans le MSA.
- Un mot en arabe ou en dialecte est généralement caractérisé par l'omission des voyelles courtes. Cette caractéristique peut engendrer des situations d'ambiguïté dans lesquelles un mot dialectal peut avoir la même orthographe qu'un mot de MSA avec une toute autre signification, formant des paires d'hétéronymes. Cela comprend des mots fortement dialectaux tels que *dwl* : ce mot est prononcé *dowl* en égyptien et signifie 'ceux' alors qu'en MSA il est prononcé *duwal* et signifie 'pays'.
- La structure et l'ordre des mots dans une phrase de l'arabe dialectal rajoutent un degré supplémentaire de complexité pour l'identification.

11.3.1.2. Applications de l'identification des dialectes

L'identification des dialectes peut avoir plusieurs applications d'un point de vue purement linguistique et expérimental. Nous citons entre autres :

- *Création de corpus de données dialectales monolingue :* cette création est possible grâce à la distinction des contenus dialectaux des contenus non-dialectaux. Ces corpus sont utiles pour aider de nombreux systèmes de TAL traitant les dialectes. Par exemple, former un modèle de langue pour un système de reconnaissance de la parole de dialecte arabe. L'identification des contenus dialectaux peut aussi aider à la création d'un ensemble de données parallèles utile pour la traduction automatique.
- *Ciblage des utilisateurs :* la reconnaissance de contenus dialectaux dans les textes produits par des utilisateurs peut aider à caractériser ces derniers et les communautés qu'ils créent ainsi que leurs attributs associés comme l'origine, le profil (francophone, anglophone, etc) voir la localisation des utilisateurs (du Maghreb ou du Macherek, de l'Algérie, de l'Égypte, etc.). En d'autres termes, cette identification peut alimenter le *profilage linguistique*.
- *Amélioration de l'efficacité des systèmes de traduction automatiques :* quand un système de traduction rencontre un mot hors vocabulaire (inconnu) soit il le rejette, soit il le transcrit. L'identification des dialectes peut contribuer à diminuer considérablement ces mots hors vocabulaire : identifier les mots dialectaux permet aux systèmes de traduction de tenter de trouver les mots MSA équivalent qui sont traités correctement et plus facilement.

11.3.1.3. Annotation des textes arabes

Après la séparation du texte d'entrée en écriture arabe et Arabizi, les mots du texte arabe considéré subissent un ensemble de décisions d'annotation. Cette annotation est appelée *la labélisation des mots en arabe*. Dans le cadre de nos travaux, toutes les annotations en

arabe ont été initialement effectuées en utilisant un traitement automatique du MSA (Saâdane, 2013) et des dialectes réalisé au sein de l'entreprise *GEOLSemantics*. Ces traitements sont suivis par une correction et validation manuelles. Cette labélisation des mots en arabe affecte à chaque mot l'un des cinq labels suivants :

- *Lang1 :* correspond à un mot du MSA. Ce mot reçoit le tag <MSA>. par exemple
- *Lang2 :* correspond à un mot de l'arabe dialectal. Ce mot est marqué comme arabes dialectal. Nous notons, que nous donnons toutes les possibilités d'appartenance à un des quatre dialectes étudiés : algérien <DZ>, tunisien <TN>, marocain <MA> et égyptien <EG>. Nous avons précisé le type du dialecte pour faciliter la reconnaissance et l'identification d'un dialecte au sein d'une phrase par la suite.
- *Entités nommées (EN) :* les entités nommées sont marquées comme tel, puis sont ultérieurement manuellement corrigés.
- *Ambig :* Correspond à un mot où la classe ne peut être déterminée étant donné le courant contexte, pourrait être soit *Lang1* ou *Lang2*. Par exemple : la phrase كله تمام *klh tmAm* signifie 'tout est bien' est ambigu si le contexte est absent, car il peut être utilisé à la fois dans le MSA et l'arabe égyptien. Dans notre cas, nous avons choisi de donner les deux tags pour ce mot <MSA> et <AD>, en d'autres termes il est labélisé *Lang1* et *Lang2*.
- *Autres :* correspond aux ponctuations, chiffres, sons et émoticônes
    - *Punct :* les signes de ponctuation sont un ensemble de signes conventionnels utilisés pour faciliter l'interprétation en indiquant la division du texte en phrases, clauses, etc. Parmi les exemples des signes de ponctuation nous trouvons les points d'explication ';', le point d'exclamation '!' et l'accolade '}', point d'interrogation ' ?', de suspension, etc.
    - *Les émoticônes :* les émoticônes peuvent être composés d'une simple suite de signes de ponctuation, de chiffres ou de lettres concaténés utilisés pour exprimer des sentiments. Les émoticônes ne sont pas considérés comme des signes de ponctuation et leur détection est exigée avant de marquer la ponctuation.
    - *Son :* Les sons sont une liste d'interjections qui n'ont aucun sens grammatical, mais qui miment les sons émis par les humains. Ces sons représentent souvent les émotions. Parmi les exemples des sons on trouve *hahaha* (rire), *hmmm* (réflexion), *umm* (pause) et *eww* (dégout). Il est commun d'étendre les sons afin de les accentuer et les rendre plus forts afin d'exprimer des émotions plus intenses. Par exemple, *hmm* peut être étendue à *hmmmmm* pour exprimer une réflexion profonde.

11.3.1.4. Annotation des textes Arabizi

Le texte Arabizi subit un ensemble de décisions d'annotation après avoir été séparé du texte en entrée. Ces annotations sont appelées *la labélisation des mots en Arabizi*, et elles ont été initialement effectuées en utilisant des traitements automatiques du français et de l'anglais, suivi par une correction et validation manuelles. Ces traitements ont été réalisés au sein de l'entreprise *GEOLSemantics*. Dans le cadre de cette labélisation, chaque mot en Arabizi reçoit l'un des quatre labels suivants :

- *lang1* : ce label concerne tous les mots étrangers à l'arabe issus d'autres langues comme le français ou l'anglais. Ces labels sont marqués <English> ou <French> dans le cas où le mot annoté conserve la même forme orthographique après traduction dans sa langue d'origine. Cette langue d'origine est souvent l'anglais dans le corpus égyptien et le français dans les corpus maghrébins. Selon (Eskander et al., 2014), ces mots sont dits 'étranger' et correspondent aux emprunts non altérés qui sont écrits de

la même manière que dans langue d'origine. Les autres mots non-arabes qui sont des emprunts comprenant des affixes arabes suite à des inflexions suivant la morphologie arabe appliquées à l'emprunt, ne sont pas marqués étrangers même si la racine est écrite comme dans la langue d'origine, comme par exemple Almobile.

- *lang2* : tous les autres mots non *étrangers* sont marqués comme arabe dialectal, puis sont ultérieurement manuellement corrigés.
- *Entités nommées (EN)* : les entités nommées sont marquées comme tel, puis sont ultérieurement manuellement corrigés.
- *Autres* : correspond aux ponctuations, chiffres, URLs, sons et émoticons

11.3.1.5. Approche d'annotation au niveau du mot

Nous rappelons que dans notre approche d'annotation, nous suivons les grandes lignes directrices d'annotation des dialectes arabes fournies dans (Elfardy et Diab., 2012) et dans (Eskander et al., 2014) et cela afin de simplifier la décision. Nous utilisons aussi une variante du système pour identifier le tag de chaque mot dans une phrase arabe (écrite en caractère arabe et latin).

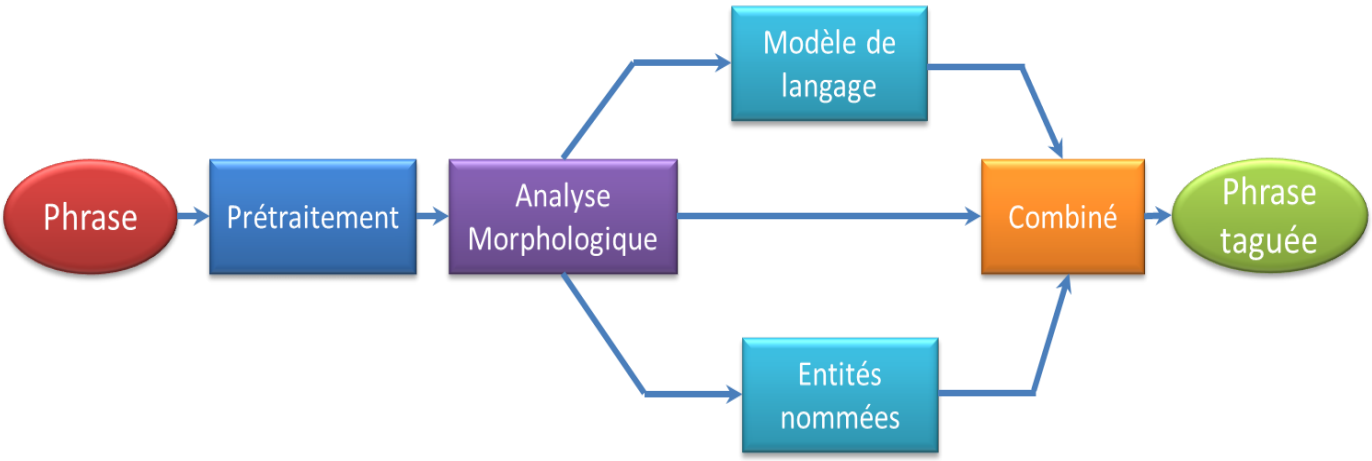


Figure 11. 11. L'Approche d'annotation au niveau du mot

Par ailleurs, l'approche que nous proposons repose sur des modèles de langue et des analyseurs morphologiques pour attribuer des tags à des mots dans une phrase donnée. Cette approche est résumée dans la figure 11 et décrite dans les étapes suivantes :

11.3.1.5.1. Prétraitement

Dans cette étape, nous effectuons un nettoyage du texte permettant de séparer la ponctuation et les nombres attachés aux mots, de normaliser les effets de l'allongement des lettres, de détecter les URLs et les nombres et de labéliser enfin la ponctuation, les émoticônes et les sons.

Concernant le processus de labélisation, ce dernier est amorcé par la détection des trois types de classes de mots suivants : Ponctuation, Sons et Emoticônes. Nous utilisons des expressions régulières pour la détection de leur occurrence dans le texte. Ainsi, ces expressions sont appliquées au texte original en Arabizi et en arabe, mot par mot. Cette application nécessite que le texte soit entièrement écrit en minuscules, c'est la raison pour laquelle les textes en entrée sont formatés en minuscules. Ce prétraitement est justifié par le fait que les émoticônes et les sons peuvent tous les deux contenir des lettres majuscules.

De plus, l'ordre de la détection est important surtout entre les émoticônes et la

ponctuation : nous rappelons que les émoticônes peuvent être composés d'une simple suite de signes de ponctuation concaténés, de ce fait leur détection est exigée avant celle de la ponctuation. Une fois détectées, les émoticônes sont remplacées par #. Ensuite, les signes de ponctuation sont détectés. Si un mot (non-émoticône) est composé de signes de ponctuation seulement, alors il sera annoté 'Punct'. Après la ponctuation, nous travaillons sur la détection des sons. Un mot sera marqué 'Son' s'il correspond à l'expression de détection des sons.

11.3.1.5.2. Analyseurs morphologiques

Pour l'annotation des textes et des mots ; et la détection de la langue, nous avons utilisé plusieurs analyseurs morphologiques arabes, français et anglais. Ces analyseurs nous ont permis de faire les opérations suivantes :

A. Marquages des mots arabes

Nous rappelons que les annotations en arabe ont été initialement effectuées en utilisant un traitement automatique du MSA (Saâdane, 2013) en plus d'un analyseur morphologique des dialectes suivi par une correction et validation manuelles. Nous utilisons ces analyseurs pour chaque mot donné dans une phrase afin de le segmenter, lemmatiser et étiqueter. Les analyseurs utilisés dans cette étape sont :

- **GEOLAR** : est un système d'analyse morphologique de l'arabe standard. Un mot est considéré comme un mot MSA si ce dernier est reconnu par l'analyse. Chaque mot analysé par ce système est considéré comme un *token* et il est retourné avec certaines informations sur sa forme comme la catégorie grammaticale. Le mot est dans ce cas labélisé <MSA>.

- **GEOLARD** : à l'instar de GEOLAR, GEOLARD est un analyseur morphologique pour l'arabe dialectal, permettant de détecter si un mot est un mot dialectal ou pas. Dans le cas où un mot dialectal est détecté, l'analyseur envoie des informations sur sa forme et le mot est labélisé <DA>. Nous notons que cette analyseur permet de préciser aussi l'origine du dialecte en intégrant les labels suivants <DZ>, <TN>, <MA> et <EG> correspondant aux dialectes algérien, tunisien, marocain et égyptien respectivement.

B. Marquage des mots étrangers

L'objectif de cette étape est d'identifier les mots étrangers dans le texte original en Arabizi. Selon (Eskander et al., 2014) 10% de l'ensemble du texte Arabizi est étranger, généralement issu du français dans les textes maghrébins et issu de l'anglais pour ce qui est des textes du Macherek (égyptien par exemple). Le marquage de ces mots étrangers est difficile en raison de l'ambigüité engendrée par l'interprétation de ces mots qui est parfois à cheval entre l'arabe (en arabizi) et une langue étrangère comme le français ou l'anglais. Par exemple, le mot Arabizi mesh peut faire référence au 'cadrage' en anglais ou au mot Arabe 'non'. De ce fait, la recherche dans le dictionnaire n'est pas suffisante pour déterminer si un mot est arabe ou étranger. Les annotations des mots étrangers ont été effectuées aussi en utilisant des analyseurs morphologiques des textes français et anglais suivi par une correction et validation manuelles. Les analyseurs utilisés dans cette étape sont :

- **GEOLFR** : est un système d'analyse morphologique du français. Ce système reconnait les mots du français, et pour chaque mot reconnu, considéré comme token dans l'analyseur, il envoie des informations sur la forme du mot comme la catégorie grammaticale et assigne au mot le label 'FR'.

- **GEOLEN** : à l'instar de GEOFR, cet analyseur effectue le même traitement mais pour les mots en anglais. En d'autres termes, il détecte si un mot est un mot anglais ou pas.

Dans le cas positif, il renvoie des informations sur la forme du mot considéré et labélise ce dernier avec 'EN'.

C. Fonctionnement de l'analyseur morphologique

Dans notre approche de labélisation, la détection d'un mot et l'identification de son label passent d'abord par une étape d'analyse morphologique ensuite par une phase de désambiguïsation basée sur les modèles de langage. Cette analyse morphologique consiste à retrouver la forme de surface d'un mot stocké dans le lexique à partir de la forme canonique (lemmatisation) de ce dernier (infinitif du verbe, masculin singulier d'un adjectif, etc...). Elle permet aussi d'attribuer à ces unités lexicales, simples ou complexes, divers types d'informations à partir de deux types d'étiquettes :

- ✓ Étiquette syntaxique qui concerne les catégories grammaticales (nom, verbes, etc.)
- ✓ Étiquette morphologique qui traduit les traits morphologiques (genre, nombre, la voix, le mode, ...etc.).

Par ailleurs, l'interprétation du mot est essentielle pour l'analyseur morphologique. La fonction permettant de récupérer les différentes interprétations possibles pour un mot (en arabe MSA ou standard, en français ou en anglais) peut être décrite comme suit:

1. ***Détection des expressions idiomatiques :*** avant de cherche dans les dictionnaires des formes canoniques et fléchis, les expressions idiomatiques sont repérées grâce à un dictionnaire d'expressions et des règles les identifient auparavant. Deux types d'expressions sont à distinguer éventuellement lors de cette étape, à savoir : i) les expressions *absolues*, qui sont repérées à ce niveau seulement, et ii) les expressions *non absolues*, qui sont analysées en tant que mots séparés ensuite la meilleure solution, expression ou mot, est choisie par les modèles du langage n-grammes.
2. ***Recherche dans les dictionnaires :*** la recherche à ce niveau traite les cas suivants :
   - ✓ Le mot existe tel quel dans le dictionnaire : nous récupérons toutes les entrées correspondantes,
   - ✓ Le mot n'existe pas dans le dictionnaire : le mot et les dictionnaires sont d'abord "désaccentués" : en arabe elle consiste en la suppression des voyelles courtes, la transformation de certaines lettres, etc. ; et en français elle représente la suppression des diacritiques. Deux situations sont à traiter dans ce cas :
     - o Le mot existe dans le dictionnaire désaccentué : nous récupérons toutes les entrées correspondantes
     - o Le mot n'existe pas dans le dictionnaire désaccentué : en fonction de la forme du mot, nous lui attribuons un ensemble de catégories par défaut. Si le mot est constitué de lettres alors il peut s'agir d'un nom, verbe, adjectif ou adverbe ; et s'il est formé de chiffres alors le mot sera considéré dans une autre catégorie comme nombre ou date. Si aucun des cas précédents ne se manifeste pas alors il s'agit d'une ponctuation ou autres.

3. ***Segmentation des formes agglutinées*** : c'est un traitement additionnel pour les textes arabes qui consiste à segmenter et séparer les formes agglutinées afin d'identifier et de traiter correctement ces formes. Ce traitement a aussi pour objectif de reconnaitre toutes les formes canoniques du mot et les différents affixes et clitiques qui lui sont collés. Les formes considérées dans cette étape sont les formes canoniques et les formes fléchies avec des proclitiques et/ou des enclitiques. Ces formes ne sont pas présentes dans le dictionnaire des formes fléchies. Nous notons que l'arabe dialectal présente un degré de cliticisation plus complexe que celui du MSA. Par exemple, le

dialecte égyptien définit les degrés de cliticisation suivant : [**cnj+** [**prt+**[**TENSE** [**BASE +PRN_D**]] **PRN_I**][**prt-neg**]], où :

- **cnj :** le proclitique de conjonction comme و *w* + 'et'.
- **prt :** la classe de proclitiques de particule de négation ما+ *ma*+ 'ne'.
- **TENSE :** la classe de proclitiques de particule (verbale) de future, comme ح+ *H*+.
- **BASE +PRN_D :** la base peut avoir un membre de la classe des enclitiques pronominaux (directe), par exemple : هُمْ *houm* 'eux'.
- **PRN_I :** la classe des pronoms indirects contient des enclitiques de particule de préposition ل+ *l*+ 'à / pour' et la classe des enclitiques pronominaux, par exemple : هُمْ houm 'eux'.
- **prt-neg :** la classe des enclitiques de particule de négation ش+ *sh*+ 'pas'.

Nous rappelons que le processus de segmentation des formes agglutinées se déroule de la manière suivante :

i. Chercher toutes les compositions possibles entre les clitiques (proclitique, enclitique) et le radical en utilisant les dictionnaires des proclitiques, enclitiques et formes fléchies.
ii. Chercher chaque radical dans le dictionnaire des formes fléchies. Si ce radical n'existe pas dans le dictionnaire, des transformations morphologiques sont appliquées avant leur suffixation en se basant sur des règles morphosyntaxiques (règles de réécriture).
iii. Chercher le radical issu de l'étape précédente dans le dictionnaire des formes fléchies.

4. ***Règle de réécriture :*** ces règles ont pour objectif la réalisation de la correspondance entre un radical traité non reconnu, et un mot du dictionnaire. Cette correspondance est effectuée par un ensemble de règles de réécriture à appliquer au radical ou à la segmentation afin d'obtenir une forme fléchie dans le dictionnaire. Par conséquent, la consultation du lexique des formes du dictionnaire est nécessaire tout au long du processus de la transformation. Nous avons proposé plusieurs règles de réécriture qui prennent en considération les contraintes morphologiques et orthographiques de la grammaire arabe. Ces règles sont détaillées dans le Chapitre 3. Pour illustrer ces règles, considérons la forme agglutinée «حكتب» (j'écrirai) et le clitique inclus dans cette forme (ح). Le radical récupéré «كتب» est la 1ère personne au singulier à la forme imperfective mais cette forme n'existe pas dans le dictionnaire égyptien des formes fléchies. Mais après l'application de la règle de réécriture rajoutant la lettre «أ» au début de mot, le radical modifié «أكتب» (j'écris) est trouvé dans le dictionnaire des formes fléchies et la forme agglutinée «حكتب» est découpée en proclitique + radical comme suit : حكتب = ح + أكتب (j'écrirai).

5. ***Modèle du langage :*** une fois que toutes les analyses possibles du mot sont récupérées, nous utilisons un algorithme basé sur les Modèles de Langage (ML) qui utilisent des matrices de bi-grammes et trigrammes de catégories morphosyntaxiques, établies à partir des corpus d'apprentissage composés d'articles de journaux (pour l'arabe, français et anglais). Ces corpus sont étiquetés et désambiguïsés manuellement. Ces n-grammes sont établis à partir du corpus, et permettent d'attribuer une pondération aux séquences de catégories afin de calculer la catégorie la plus probable d'un mot en contexte. Nous notons que les catégories grammaticales utilisées ont été conçues spécifiquement pour ce type d'algorithmes, car nous utilisons des catégories

positionnelles ("adjectif postérieur"). Ces catégories positionnelles permettent d'agrandir fictivement les contextes, car, elles donnent des indications sur le contexte gauche ou droit. A chaque trigramme de catégories est associé un poids, calculé selon le corpus d'apprentissage. Grâce à ce poids, l'algorithme va choisir le chemin le plus probable, c'est-à-dire la suite de catégories la plus probable selon ces poids.

6. ***Règle correctives :*** nous avons ajouté quelques règles correctives, pour les cas fréquents que l'algorithme ne peut pas gérer, comme c'est le cas de l'analyse du mot français "été" qui est souvent analysé comme un nom alors que dans les textes il est plus souvent verbe. Nous avons aussi effectué des traitements pour l'élimination de la répétition de la lettre dans un mot comme كبيبيبير *kbbbbbbir 'grand'(*transformé en كبير *kbir),* ou dans les mots qui sont collés aux chiffres afin de trouver la forme présente dans les dictionnaires. La répétition des lettres est utilisée pour indiquer un stress ou une émotion.

7. ***Liste des entités nommées et règles d'extraction:*** cette étape est dédiée pour la construction des ressources d'entités nommées (ENs) par une élaboration, manuelle et automatique des éléments suivants :
    a. Un dictionnaire à partir des ressources textuelles (pages web par exemple)
    b. Des listes issues des gazetteers ANERgazet[20] proposées par (Benajiba, Rosso, 2007)
    c. Des ressources se trouvant dans la basse de données géographique GeoNames[21] qui contient des lieux géographiques contenant aussi des noms de personnes
    d. Le lexique proposé par Attia[22].

    Pour les deux premières ressources textuelles, nous avons effectué une fouille manuelle des textes. Cette fouille nous a permis de collecter et de sélectionner, dans une liste, des noms et des prénoms potentiels qui seront triés automatiquement par la suite afin d'éliminer les doublons. Nous avons aussi utilisé un translittérateur des noms propres pour translittérer les noms propres qui proviennent d'autres dictionnaires latins, essentiellement ceux utilisés au sein de l'entreprise *GEOLSemantics*. Lors de cette étape, nous avons élaboré :
    - Un dictionnaire de prénoms et de noms de famille qui contient 9000 prénoms arabes et prénoms étrangers transcrits
    - Un dictionnaire de lieux qui contient 8464 noms de lieux, ville, etc.
    - Un dictionnaire d'organisation qui contient 1000 noms d'organisation

    Règles de détection : les règles de reconnaissances des lieux, éditées manuellement, permettent d'identifier et typer les ENs quelle que soit la simplicité ou complexité de leur structure. Ces règles sont basées sur les ressources décrites ci-dessus et des relations syntaxiques. Pour plus de détails sur l'élaboration des dictionnaires et les règles de détection des entités nommées, nous renvoyons le lecteur au chapitre 4.

11.3.1.5.3. Combinaison :

L'étape de combinaison permet d'agréger plusieurs composants, y compris des dictionnaires et des modèles de langue afin d'effectuer la reconnaissance d'entités nommées et l'identification de langue du texte d'entrée. Chaque mot de la phrase d'entrée peut obtenir

[20] Téléchargeable depuis le site : http://www1.ccls.columbia.edu/~ybenajiba/downloads.html
[21] http://download.geonames.org/export/dump/
[22] http://www.attiaspace.com/getrec.asp?rec=htmFiles/LexMWEs

différentes étiquettes des étapes précédentes, de ce fait, l'étape de combinaison, en se basant sur les étiquettes générées, utilise un ensemble de règles de décision pour affecter le tag final à chaque mot de la phrase d'entrée. Les règles de décision utilisées sont :

1. Si le mot contient des numéros ou des signes de ponctuation, alors il est associé à la balise *Autre* (Ponct, NUM, etc)
2. Si le mot est présent dans l'un des dictionnaires ou si l'analyseur GEOL assigne la balise *entité nommée*, alors le mot est étiqueté comme Entité nommées <EN>
3. Si le mot est identifié par le modèle de langage (ML) que ce soit Lang1 ou Lang2, le mot est alors associé à l'étiquette correspondante
4. Si le mot identifié par le modèle du langage (ML) est associé à la fois à Lang1 et Lang2, alors nous attribuons au mot les balises Lang1 et Lang2. Toutefois ce cas introduit une certaine ambiguïté.
5. Si le ML n'étiquettepas le mot, c'est le cas par exemple où le mot est considéré comme un mot hors vocabulaire par le ML, alors nous associons la balise <UNK> au mot analysé.

Le tableau suivant donne les correspondances entre les balises utilisées et les textes :

Texte	**Tag**	**Correspondance**
Textes arabes	Lang 1	MSA
	Lang2	AD (arabe dialectal)
Textes Arabizi	Lang1	Langues étrangères français <French> ou anglais <English>
	Lang2	Arabizi <arabe-latin>

Tableau 11. 1. Les correspondances entre les balises et les textes.

Nous avons développé une interface d'annotation afin de valider les résultats et les tags réalisés par l'analyseur morphologique et tagguer les mots hors vocabulaire afin de compléter les dictionnaires correspondants.

Voici un exemple d'annotation au niveau du mot pour le texte Arabizi en dialecte égyptien suivant : « *ya Houda lel asaf e7na mogtama3 mot5alef bgd, w kollena benedfa3 taman da kol youm* »

Arabizi	**Sorry**	**Houda**	**e7na**	**mogtama3**	**mot5alef**	**bgd**	**,**	**kollena**	**benedfa3**	**taman**	**da**
Tag	English	EN	AD	AD	AD	AD	Autre	AD	AD	AD	AD
Traduction	Pardon	Houda	nous	société	Arriérée	vraiment	,	Tous	payons	prix	ça

11.3.1.6. Approche d'annotation au niveau des textes

L'annotation pour l'identification de la langue au niveau du texte à codes alternés est une tâche difficile et alimente beaucoup de sujets de travaux au niveau de la communauté de la PNL. Rappelons que les locuteurs arabes dans les médias sociaux, mais aussi dans les forums de discussion, les messages SMS et le chat en ligne, utilisent des caractères non standards. Ceci a engendré au niveau de la communication en arabe dans les médias sociaux une utilisation d'une variété d'orthographes et de systèmes d'écriture, incluant les écritures arabe et latine, le français, l'anglais et un mélange de code alterné romanisé. Les conversations dans le monde arabe se caractérisent aussi par un phénomène linguistique d'alternance des codes dans lequel les locuteurs commutent entre deux ou plusieurs langues

lors de la conversation. Pour prendre en charge cette problématique, nous avons développé des programmes pour la séparation des segments dans différentes langues au sein d'un même texte. Ces programmes effectuent les opérations suivantes :

1. Tagguer le segment analysé en alphabet arabe ou latin,
2. Annotation pour l'identification de la langue au niveau du mot en effectuant les actions suivantes :
   a. Utiliser les analyseurs du français et de l'anglais de GEOLSemantics pour tagguer les mots écrits en caractères latins et identifier si les mots sont écrits en français, anglais ou en arabe latinisé (Arabizi), en respectant les étapes d'annotation au niveau du mot décrite dans la section précédente
   b. Utiliser les analyseurs morphologiques du MSA et ceux de l'arabe dialectal pour tagguer les mots écrits en arabe et identifier si les mots appartiennent au MSA ou à l'arabe dialectal
3. Déterminer et identifier la langue des segments:
   a. Si la phrase contient un mot dialectal alors nous découpons le segment et lui attribuons la balise *Lang2*
   b. Si la phrase contient plus de trois mots consécutifs en français/anglais, alors la phrase est en français/anglais et se voit attribuée la balise *Lang1*
   c. Si la chaîne contient moins de trois mots consécutifs en français/anglais, alors la phrase est en arabe écrit en latin et annotée par conséquent par la balise *Lang2*.
4. Valider les frontières pour chaque segment,
5. Ajouter dans les métadonnées concernant l'origine du message en introduisant les codes ISO-2 suivants :
   - DZ : correspond au code de l'Algérie ;
   - TN : correspond au code de la Tunisie ;
   - MA : correspond au code du Maroc ;
   - EG : correspond au code de l'Égypte.

   L'ajout de l'origine du message est basé sur la combinaison des informations suivantes si elles sont présentes :
   - la source des journaux (algérienne, tunisienne, etc)
   - la localisation géographique de l'auteur
   - le nombre des mots dialectaux détectés au niveau du message.

L'extrait suivant d'un commentaire de blog (Yahoo) montre un exemple de découpage des frontières concernant le phénomène de l'alternance des codes (code-switching) au niveau du texte :

<doc docid="Yakhoo_comment3" articleURL="http://z4.invisionfree.com/Yakhoo/ar/t73.htm" author="3assimi" date="October 4, 2007" time="11:32" subtitle="M3askri ra7 yekhtob (Yakhoo)">
<comment> hahahaha .. et ca me rappelle une autre... wahed kal el yemah zawjini, mais choufili wahda chaba, c tout ce que je demande.. yemah kaletlou ya wlidi ezine 3omrou ma bna dar ... alors kalelha .. choufili maçon alors :P
</comment>
</doc>

Le résultat de l'annotation et l'identification de la langue est donné comme suit :

<doc docid="Yakhoo_comment3" articleURL="http://z4.invisionfree.com/Yakhoo/ar/t73.htm" author="3assimi" date="October 4, 2007" time="11:32" subtitle="M3askri ra7 yekhtob (Yakhoo)" lang="DZ">
<comment>
<French> hahahaha .. et ca me rappelle une autre... **</French>**
<Arabizi> wahed kal el yemah zawjini, mais choufili wahda chaba, **</Arabizi>**
<French> c tout ce que je demande.. **</French>**
<Arabizi> yemah kaletlou ya wlidi ezine 3omrou ma bna dar ... alors kalelha .. choufili maçon alors :P **</Arabizi>**
</comment>
</doc>

11.3.1.7. Description des balises et des attributs

Avant de finir cette section, voici un aperçu des méta-informations contenues dans les corpus d'apprentissages constitués. Ces informations sont encapsulées par les balises <doc> … </doc> et renseignent les métadonnées extraites à partir du site web. Parmi ces métadonnées nous citons les attributs suivants :

- docid=" " : cet attribut contient le nom de la source (journal, vidéo, Youtube, etc.) suivi d'un séparateur (_) et du numéro du commentaire. Par exemple, docid="elkhabar_comment1" renseigne le premier commentaire du journal *Elkhabar*.
- articleURL=" " : contient l'URL de l'auteur et le nom du site téléchargé ;
- author =" " : cet attribut contient l'ID de l'auteur, le nom le pseudonyme ;
- pays =" " : cet attribut représente la localisation géographique de l'auteur ;
- date =" " : cet attribut contient la date du commentaire ;
- time =" " : cet attribut contient l'heure du commentaire ;
- subtitle= " " : cet attribut contient le titre de la publication ou le commentaire ;
- lang=" " : cet attribut contient l'origine du message en utilisant les codes ISO-2 ;
- <comment> … </comment> : cette balise contient le contenu du message ;
- <French> ... </French> : cette balise contient le message écrit en français ;
- <English> ... </English> : cette balise contient le message écrit en anglais ;

<arabe-latin> … </arabe-latin> : cette balise contient le message en arabe dialectal écrit en caractères latins.

11.4. Interface d'annotation

Nous avons développé une interface d'annotation afin de faciliter la validation des résultats de notre analyse linguistique d'une part et d'annoter manuellement les mots hors vocabulaire afin d'enrichir nos dictionnaires initiaux d'autre part. Cette interface sert aussi d'environnement semi-automatique d'enrichissement des dictionnaires bilingues dialecte-MSA. Cette interface utilise la sortie d'un analyseur linguistique. Cet analyseur prend en argument un fichier de texte arabe (MSA et de chaque dialecte) en caractères arabes non annoté et renvoie un fichier XML dans lequel il attribue à chaque mot dans le fichier d'entré les informations suivantes :

- la catégorie grammaticale du mot (nom, verbe, adjectif ou autre)
- les tags dialectaux ou standard appropriés (MSA, DZ, TN, MA, EG). le/les dialectes/MSA auxquels le mot appartient.

Il est important de signaler que l'analyse traite les mots avec ou sans voyelles, qu'ils aient des clitiques ou qui se présentent comme une forme canoniques.

L'interface, codée en html, affiche le fichier XML contenant le résultat qui est un corpus annoté avec l'association de couleurs différentes en fonction du dialecte dans lequel le mot est présent. Dans un onglet séparé les mots qui ne sont présent dans aucun des dictionnaires, un annotateur humain prendra en charge la validation ou non de ces mots pour les ajouter aux dictionnaires existants. L'interface permet à l'annotateur d'afficher les mots regroupés par jeu de dialectes dans lequel ils sont présents. Quant au découpage fait en interne par l'analyseur linguistique, ce dernier sera affiché en mode *tooltip* comme une information supplémentaire pour aider l'annotateur dans sa mission.

Figure 11. 12. L'interface d'annotation

Le fait d'avoir un environnement semi-automatique dans lequel l'annotateur peut voir les mots selon le jeu de dialectes dans lequel ils sont présents, permet à ce dernier de choisir de modifier l'annotation des résultats de l'étape automatique. Cette modification est due à une carence de mots dans un dictionnaire dialectal par rapport à un autre, et peut consister soit en une suppression d'un tag de dialecte, soit en un ajout d'un. En ce qui concerne les mots hors vocabulaires qui existent dans le corpus et non pas dans les dictionnaires, l'interface permet aux annotateurs de prendre en charge l'annotation manuelle de ces mots inconnus par l'analyseur linguistique.

Nous avons évoqué que l'interface contient des onglets permettant de montrer les mots selon le jeu de dialectes dans lequel ils sont présents, c'est-à-dire qu'ils existent :

i. Dans un seul dialecte ou dans le MSA (ALG, EGY, MAR, MSA, TUN)

Figure 11. 13. Les mots appartenant à un seul dialecte.

ii. Dans deux dialectes ou dans un dialecte et dans MSA (ALG|EGY, ALG|MAR, ALG|MSA, ALG|TUN, EGY|MAR, EGY|MSA, EGY|TUN, MAR|MSA, MAR|TUN, MSA|TUN)

Figure 11. 14. Les mots appartenant à deux dialectes.

iii. Dans trois dialectes ou deux dialecte et MSA : ALG|EGY|MAR, ALG|EGY|MSA, ALG|EGY|TUN, ALG|MAR|MSA, ALG|MAR|TUN, ALG|MSA|TUN, EGY|MAR|MSA, EGY|MAR|TUN, EGY|MSA|TUN, MAR|MSA|TUN

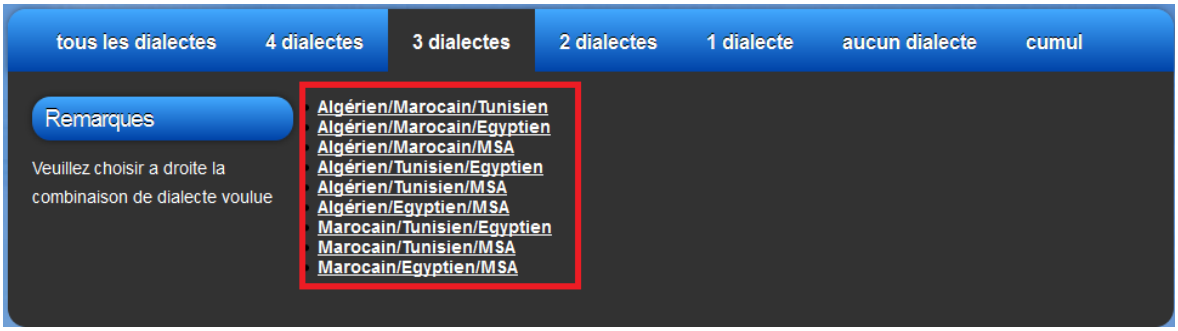


Figure 11. 15. Les mots appartenant à trois dialectes.

iv. Dans quatre dialectes ou trois dialectes et MSA : ALG|EGY|MAR|MSA, ALG|EGY|MAR|TUN, ALG|EGY|MSA|TUN, ALG|MAR|MSA|TUN, EGY|MAR|MSA|TUN

Figure 11. 16. Les mots appartenant à quatre dialectes.

v. Dans tous les dialectes et dans MSA : ALG|EGY|MAR|MSA|TUN. Ce cas de figure nous donne aussi une indication sur le dialecte dans lequel nous utilisons le plus le MSA.

Figure 11. 17. Les mots appartenant à tous les dialectes et au MSA.

vi. Aucun dialecte : cet onglet contient les mots hors vocabulaires et qui ne sont pas présents dans aucun des dictionnaires, MSA ou des dialectes.

Figure 11. 18. Les mots n'appartenant à aucun dialecte.

Sur un autre registre, l'onglet 'aucun dialecte' donne accès à la fonction d'affichage des mots hors vocabulaires et non reconnus par l'analyse linguistique. Ces mots sont annotés avec une couleur prédéfinie. Nous avons choisi d'afficher tout le texte autour afin de faciliter le travail de l'annotateur en lui renvoyant le contexte dans lequel le mot a été employé.

L'annotation humaine des mots hors vocabulaires est réalisée via le bouton « ***commencer l'annotation*** ».

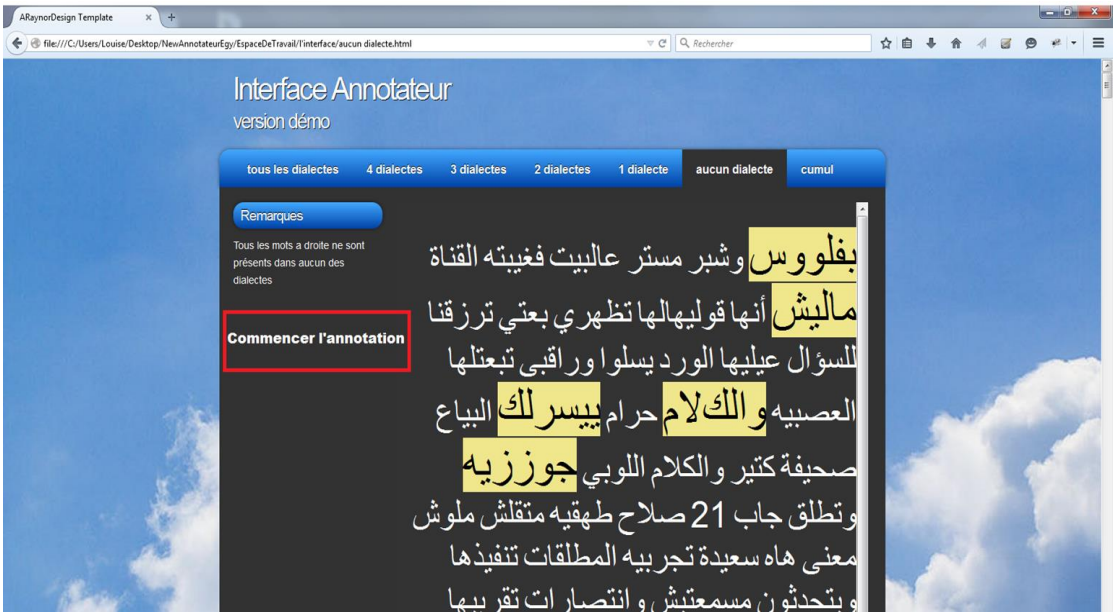
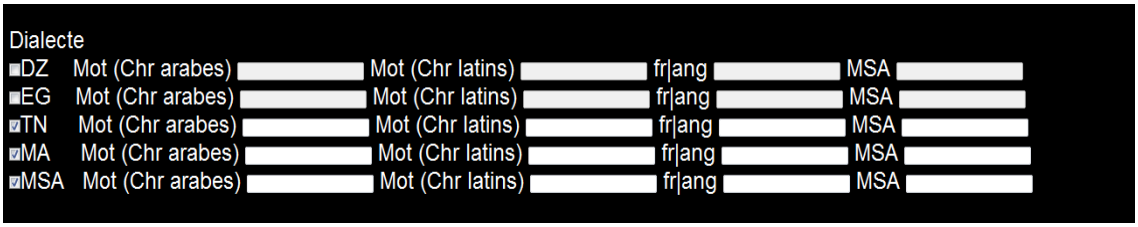


Figure 11. 19. Le bouton pour commencer l'annotation.

Cette fonction permet de parcourir rapidement les mots inconnus, ainsi accéder aux tags <unk> un par un et ensuite pouvoir traiter le mot directement en cliquant dessus. Un clic sur un mot lance une interface de formulaire dans laquelle l'annotateur devra préciser en premier la catégorie grammaticale du mot. Nous notons que l'annotateur aura aussi la possibilité de signaler une faute d'orthographe.

Figure 11. 20. Signaler une faute d'orthographe.

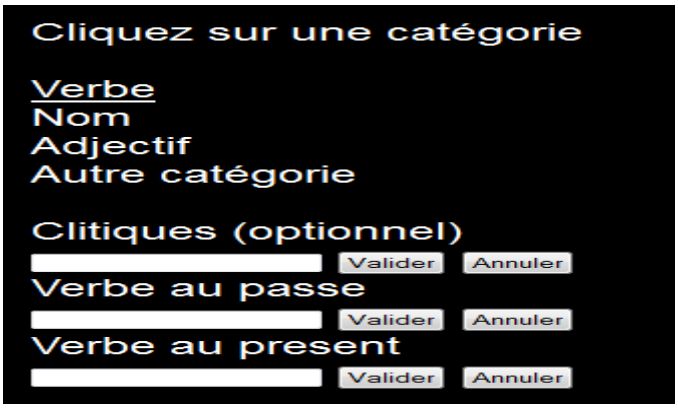
Si nous choisissons le lien 'Faute d'orthographe', ceci nous permet d'extraire le mot et l'intégrer dans une feuille html qui regroupera les fautes et sera intégrée à l'interface de l'annotateur. Le choix mot valide renverra un formulaire qui contient tous les attributs composant les dictionnaires : translittération latine, mot écrit en écriture arabe, traduction français, traduction MSA et catégorie, et cela en fonction du nombre et l'origine du dialecte. Par ailleurs, l'annotateur pourra changer l'orthographe selon le dialecte. Dans le choix du dialecte l'utilisateur peut décider qu'un mot soit présent dans plusieurs dialectes en cochant plusieurs cases à la fois, ce qui renverra à l'annotateur les champs qui représentent les valeurs du mot à remplir.



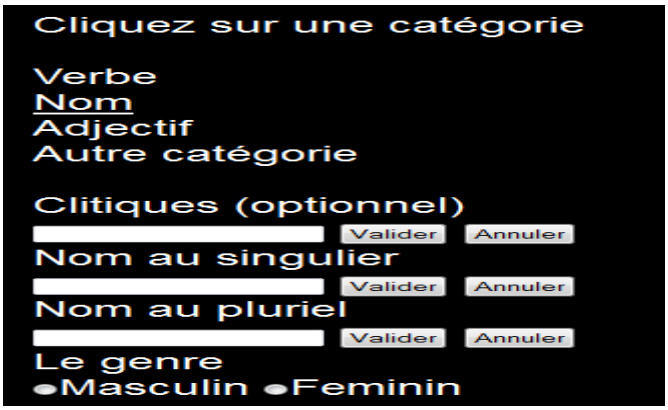
Le choix de la catégorie : verbe, adjectif ou nom nous renverra vers un autre formulaire, en plus de celui du dessus. Ce formulaire nous permet de faire les tables de flexion sachant que selon la catégorie du mot, le formulaire est différent et contient les informations de flexion à renseigner selon la catégorie considérée. Ces informations seront stockées dans des fichiers séparés. Par ailleurs, le genre du nom et de l'adjective nous aide à faire un fichier des exceptions.

Si l'annotateur trouve un clitique, alors une case à remplir supplémentaire est rajouté. Si le mot contient un enclitique et un proclitique l'annotateur les sépare avec une virgule.

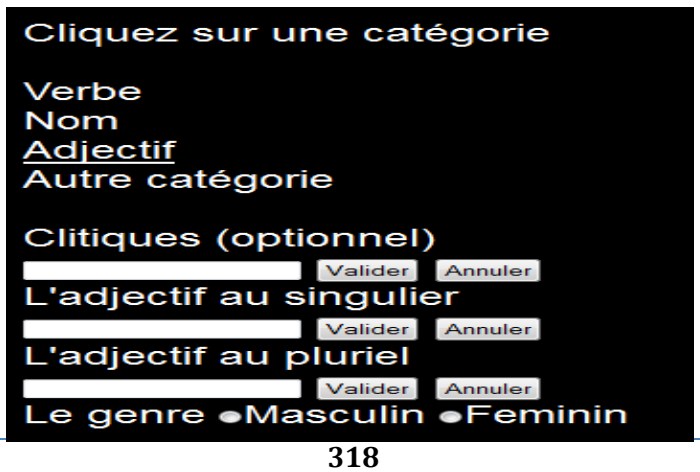
- ***Verbe :***



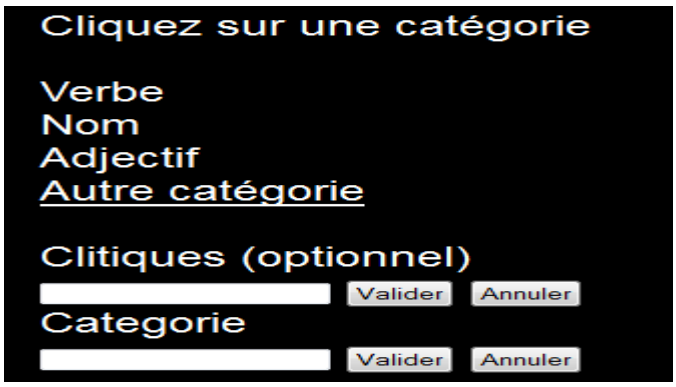
- ***Nom :***



- ***Adjectif :***



- ***Autre catégorie***



Une fois annoté, le tag « unk » devra disparaître et le tag du mot devra disparaître dans le jeu de dialectes auquel le mot appartient, pour ne pas avoir à annoter le même mot à chaque fois que nous ouvrons l'annotation. Pour cela, les fichiers *html* dans lesquels sont stockés les résultats doivent être dynamiques.

11.5. Extraction des traits de reconnaissance automatique des dialectes arabes

L'approche que nous avons développée ne vise pas une description exhaustive de chaque dialecte mais seulement la mise en évidence de traits linguistiques qui lui sont spécifiques et qui sont susceptibles d'être intégrés à un module de reconnaissance automatique de l'écrit dialectalisé.

L'objectif d'un tel traitement est le « criblage linguistique » qui consiste à passer un texte dialectalisé nouvellement recueilli au crible d'une base de données sémantique. Ce type d'opération vise à préciser les caractéristiques linguistiques et sociologiques de la production écrite par rapport aux données de la base.

11.5.1. Détection des dialectales par le biais des pronoms personnels isoles

Les pronoms personnels isolés se prononcent de façon différente selon les dialectales. Par exemple, la troisième personne du singulier en dialecte marocain est prononcée «hûwa» (masculin) et «hîya» (féminin), alors qu'en dialecte libanais c'est «huwweh» (masculin) et «hiyyeh» (féminin). De même, la troisième personne du pluriel en marocain est prononcée «hûma», alors qu'en libanais c'est «hinneh». Dans ce cas, les deux pronoms n'ont pratiquement plus de points communs, ce qui constitue un trait distinctif de ce dialecte dans le corpus.

Pour illustrer le caractère opérationnel de ce type d'indices, nous faisons figurer ci-après un tableau récapitulatif des usages du pronom personnel dans les dialectales du Yémen (Guidère, 2004). Celui-ci montre qu'on peut affiner la reconnaissance jusqu'au niveau «local» dans ce type d'étude dialectologique.

Régions du Yémen	Hadra Mawt	Shabwa	Mukeyras	Lahej	Dhâlef	Yâfif	ʾAden
Pro. 1er pers. Sg m.f	Ana	Ana	Ana	Anî	Ana	Ani	Ana/Ani

11.5.2. Détection des dialectes par les pronoms et les adverbes interrogatifs

La base contient également une série de pronoms interrogatifs qui permettent de distinguer les dialectes entre eux. Par exemple, le pronom «âsh» (que? Quoi?) peut former, en étant combiné à d'autres particules, des adverbes variés dans les dialectales du Maghreb : *lâsh* (à quoi), *gaddâsh* (combien), *ʿalâsh* (pourquoi), etc.

Dans les dialectales du Machrek, ces adverbes sont beaucoup moins fréquents, sauf pour les interrogatifs «combien, pourquoi», «édesh, lesh», mais les pronoms se prononcent et se transcrivent différemment : par exemple, en dialectale libanais, le pronom se prononce et se transcrit comme une voyelle fermée «*é*»; c'est pourquoi on retiendra le suffixe «esh» au lieu de «ash» comme trait de distinction de ce dialecte.

11.5.3. Détection des dialectales par les pronoms personnels suffixes

Les locuteurs des dialectales du Maghreb (algérien, tunisien, marocain) prononcent la 2ème et la 3ème personne du singulier (masculin) différemment par rapport au dialecte du Machrek. Par exemple : le radical verbal « na /si /ya » (oublier), on dira dans le dialecte maghrébin « nsitek », « nsiteh », tandis qu'en dialecte du Machrek, on dira « nsitak », « nistoh ». Dans ce cas, pour la détection et la distinction entre les deux dialectales, l'accent sera mis sur le suffixe de la 2ème et la 3ème personne du singulier et non pas sur le radical du verbe.

11.5.4. Détection des dialectales par les particules

Les indices de la personne peuvent constituer un critère fiable pour faire la différence entre les dialectes. En effet, les dialectes du Maghreb sont caractérisés par l'utilisation de la particule « n » à la première personne du singulier, à l'inaccompli, alors que cette particule est presque absente dans les dialectes du Machrek, c'est le cas du dialecte libanais qui emploie à la place du « n » d'autres particules comme le « a », « e », « u ». Pour illustrer ces propos, prenons l'exemple du verbe « kataba » (écrire) qui est conjugué, à la première personne du singulier à l'inaccompli, comme suit dans les deux dialectes :

- ✓ « âna *n*ekteb » (j'écris) : dialectale du Maghreb
- ✓ « ana *e*kteb ou *u*kteb » (j'écris) : dialectale du Machrek

11.5.5. Détection des dialectales par le schème verbal et la forme passive

Le système phonétique des dialectes du Maghreb présentent la caractéristique de succession de deux consonnes au début du mot qui est rare voire inexistante dans son correspondant (système phonétique) dans les dialectes du Machrek. Cette caractéristique influence notablement sur le schème verbal « faʿala» en arabe standard qui se décline en « fʿel » en Algérie et en « faʿal» en Egypte. Voici quelques exemples de l'effet de cette particularité :

Verbe	Arabe Standard	Dialecte Maghrébin	Dialecte du Machrek
Frapper	*Daraba*	*Dreb*	*Darab*
se taire	*Sakata*	*Sket*	*Sakat*
Boire	*Charaba*	*Chreb*	*Charab*

Notons que pour la reconnaissance, la conjugaison de ces verbes au passé, à la 3ème personne du singulier, est en soi un élément intéressant de classification.

Dans le même cadre, la forme passive des verbes constitue aussi un élément de distinction des dialectes entre le Maghreb et le Machrek. En effet, en arabe standard la forme passive est dérivée par apophonie de la forme active avec un simple changement du timbre de

la mélodie vocalique « a →u ». Cependant, dans les dialectes cette forme est obtenue en ajoutant au verbe à l'accompli le préfixe [t] dans le cas du dialecte du Maghreb et les préfixes [it] ou [in] dans le cas du dialecte du Machrek. Par exemple la forme passive du verbe «kataba » (écrire) est « inkatab » ou « itktab » au Machrek, et « tekteb » contre la forme« kutiba » en arabe standard.

11.5.6. Le cas de la consonne « q » dans les dialectales arabes

Prononciation des consonnes : le meilleur exemple est l'utilisation de la consonne occlusive uvulaire sourde « ق » [q] dans certaines régions et l'occlusive palatale sonore « ڤ » [g] dans d'autres régions. La consonne « q » est l'un des sons qui méritent une attention particulière. En fait, selon les dialectales, les régions, les villes, et parfois les localités, ce son qui se prononce « q » en arabe littéral, peut être prononcé : [q,ʾa,k, g ou kh]. Ce son est considéré comme une propriété qui traduit un clivage sociogéographique [D. Lajmi, 2009] entre parler citadin et parler rural et encore parler bédouin. Une distinction de base peut s'avérer utile pour un premier classement. En effet, dans la grande majorité des cas étudiés, on peut esquisser quelques tendances générales concernant la prononciation du son « q » :

- ✓ **q → ʾa** correspond en général aux parlers des citadins au Machrek;
  qalb → ʾalb (coeur); qâla → ʾâl (il a dit)
- ✓ **q → k** correspond en général aux parlers des ruraux :
  qalb → kalb (coeur); qâla → kâl (il a dit)
- ✓ **q → g** correspond en général aux parlers des bédouins :
  qalb → galb; qâla → gâl
- ✓ **q → q** correspond en général aux parlers de groupes qui sont restés plus ou moins fermés et qui continuent à prononcer le « q » à la manière classique.
  qalb → qalb; qâla → qâl

11.5.7. Détection des dialectes par la transcription des lettres

Les systèmes d'écriture latine diffèrent d'une langue à une autre. Par exemple on ne trouve pas en anglais les lettres (é, è, ô, à, ù, â, ê, ç, î...) qui sont utilisées en français (Al-Balawi et al., 2009). Ce fait génère des différences lors de la transcription des mots arabes en écriture latine, car en général les gens du Maghreb sont influencés par la littérature française tandis que les gens du Machrek sont influencés par la littérature anglaise. Citons l'exemple de l'article (ال). Ses règles d'assimilation sont variables d'un dialecte à un autre. Les gens du Maghreb le transcrivent par (**El**) et les gens du Machrek le transcrivent par (**Al**). Le même problème se pose pour certaines lettres comme la lettre (ج) qui est transcrite en (Dj) en Algérie et (J) ou (G) dans une moitié du Maghreb et au Machrek. La lettre (ش) est transcrite en (Ch) au Maghreb et en (Sh) au Machrek.

Tous ces phénomènes peuvent être observés à partir du corps dialectalisé que nous avons constitué tout au long de l'année 2012-2013. Il permet de reconnaître automatiquement les dialectes arabes écrits en caractères latins et de mieux cerner les spécificités morphosyntaxiques et sémantiques de chaque dialecte.

Partie V : Résultats expérimentaux et évaluation

Chapitre 12 Système d'évaluation et d'extraction de connaissances du MSA

Les évolutions rapides des nouvelles technologies sont accompagnées d'un essor important de la quantité d'information disponible sur le Web, et nécessitent le développement d'outils pour analyser et structurer les documents textuels. Ainsi, les documents en arabe sur le Web, à l'instar des autres langues, se multiplient en nombre, en contenus et en volume. Pour traiter cette grande masse de textes, une possibilité est de recourir à des outils de fouille de textes ou d'extraction de connaissances. Pourtant, dans ce domaine en pleine émergence, les outils qui permettent d'analyser les documents arabes se limitent généralement à l'extraction d'entités nommées.

Selon (Fayyad et al.,1996), l'extraction de connaissances à partir des données (ECD) se définit comme « *l'acquisition de connaissances nouvelles, intelligibles et potentiellement utiles à partir de faits cachés au sein de grandes quantités de données* ». En fait, nous cherchons surtout à isoler des traits structuraux *(patterns)* qui soient valides, non triviaux, nouveaux, utilisables et si possible compréhensibles ou explicables.

Extraire des informations ou des connaissances utiles à partir des textes représente de nos jours, un domaine de recherche important et intéressant faisant émerger plusieurs axes d'évolution. Parmi ces axes, la reconnaissance des entités nommées (REN) qui est considérée comme une tâche fondamentale pour l'analyse sémantique. Néanmoins, elle ne représente qu'une étape préalable. Pour aller au-delà de l'extraction des ENs, la détection des relations impliquant ces entités est nécessaire aussi afin qu'un modèle plus structuré de la compréhension du texte soit généré. Ce type de traitement concerne les tâches de découverte des relations utiles entre deux entités d'un texte. Nous signalons à ce stade que l'extraction des connaissances peut avoir de nombreuses applications comme dans les systèmes questions-réponses, le résumé automatique et l'exploration du Web.

Par ailleurs tout système linguistique, et intrinsèquement l'extraction de connaissances, doit être évalué afin de mesurer la pertinence et la qualité de ses résultats. Dans cette perspective, nous avons intégré notre système linguistique dans la plateforme d'extraction de connaissances de *GEOLSemantics*. Cette intégration nous donne la possibilité de faire une évaluation en se basant sur des ressources assez riches et intéressantes, chose qui est primordiale dans l'évaluation de tout système linguistique.

Dans ce chapitre, nous présentons d'abord notre plateforme d'intégration, qui est le système global d'extraction de connaissances de *GEOLSemantics*. Nous donnons ensuite une description générale du système d'extraction de connaissance de l'arabe. Enfin, nous présentons les différentes évaluations, que nous avons effectuées pour mesurer la qualité de nos extractions de connaissances, ainsi que la pertinence du système d'analyse en sa globalité.

12.1. Système d'extraction de connaissances de GEOLSemantics

GEOLSemantics est une jeune entreprise innovante dont l'activité repose principalement sur une grande expertise en traitement linguistique. L'objectif principal de l'entreprise est de développer des outils permettant de traiter un ensemble de documents de manière automatique et efficace afin d'extraire le maximum d'informations pertinentes. En effet, le premier intérêt est de dresser un profil sémantique qui permet d'extraire automatiquement d'un ensemble de textes des connaissances structurées, datées et localisées, qui décrivent des concepts, des relations et des évènements impliquant des entités nommées. L'originalité du travail réalisé à GEOLSemantics repose sur son approche linguistique : la construction des relations et de la structure des connaissances se fait à partir de données purement linguistiques, à l'aide de règles linguistiques. En outre, l'extraction se fait sur des

relations syntaxiques extraites au cours d'une analyse de dépendance. L'analyse syntaxique profonde permet aux linguistes de mettre en place des règles d'extraction qui prennent véritablement en compte le sens porté par la syntaxe d'une phrase.

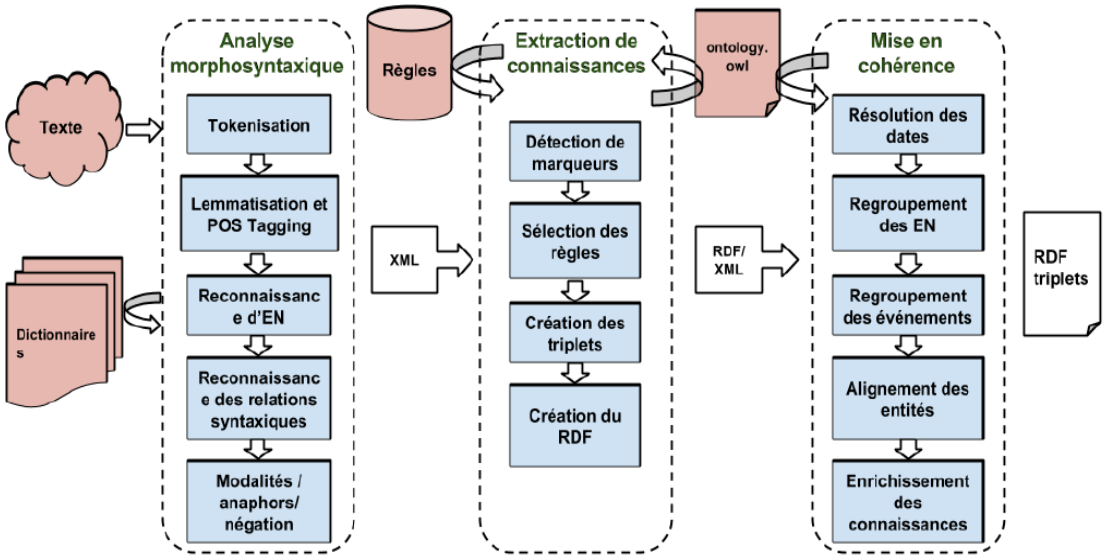


Figure 12. 1. Architecture du système d'extraction de GEOLSemantics

La chaine de traitement, présentée dans la figure 12.1, contient principalement trois modules complémentaires. Les deux premiers modules concernent le traitement automatique des langues. Ces modules représentent une capitalisation des différentes expertises acquises depuis des années dans le TAL. Plus précisément, à partir d'un texte en langage naturel donné en entrée, nous procédons à une analyse linguistique profonde afin d'identifier les relations syntaxiques entre les différentes unités de la phrase. Ensuite, nous effectuons une extraction de connaissances consistant à formaliser ces relations sous forme sémantique. Ces deux modules renvoient en sortie des connaissances formalisée en RDF. Enfin, le dernier module, nommé mise en cohérence, permet d'effectuer des opérations afin de compléter le traitement précédent. Ce module aide à pallier les quelques lacunes dans le résultat RDF dues au traitement intra-phrase des deux analyses précédentes. A l'issue des traitements de ce module, nous obtenons un ensemble de triplets RDF modélisant la connaissance contenue dans le texte. Dans la suite de ce chapitre, nous détaillons les différentes étapes du traitement d'extraction de connaissances pour les textes arabes, effectuées suivant la logique et étapes du système présenté.

12.2. Analyse linguistique profonde

L'analyse linguistique profonde est nécessaire pour assurer une extraction d'informations sûre, pertinente et complète, par exemple en reliant des éléments qui peuvent être éloignés dans la phrase initiale. L'analyse mise au point dans le premier module de notre système est réalisée en plusieurs étapes allant du découpage en mots jusqu'aux relations que ceux-ci entretiennent au sein d'une phrase. Nous donnons ci-après un aperçu de ces étapes, et pour plus de détails nous renvoyons le lecteur au Chapitre 3.

12.2.1. Découpage en mots

La tokenisation permet le découpage du texte en mots, les « tokens », séparés par des ponctuations ou par des espaces. Elle prend aussi en compte les balises, les dates abrégées, etc. Citons l'exemple de la tokenisation en mots de la phrase «باريس مدينة الجن والملائكة. » (Paris la ville des diables et des anges) donnera : باريس | مدينة | الجن | و | الملائكة. C'est une étape qui va permettre d'attribuer ensuite à chaque token des catégories et des propriétés sur lesquelles portera l'analyse profonde.

12.2.2. Analyse morphologique

Il convient de rappeler que la fonction principale de l'analyseur morphologique consiste à retrouver la forme de surface d'un mot stocké dans le lexique à partir de la forme canonique de ce dernier (infinitif du verbe, masculin singulier d'un adjectif, etc...), et d'attribuer à ces unités lexicales simples ou complexes divers types d'informations à partir de deux types d'étiquettes, d'une part l'étiquette syntaxique concernant les catégories grammaticales (nom, verbes, etc.) et d'autre part, l'étiquette morphologique concernant les traits morphologiques (genre, nombre, la voix, le mode, ...etc.). Cette étape est primordiale lors de l'analyse linguistique. Elle se divise à son tour en plusieurs sous-étapes : la consultation du dictionnaire des formes fléchies pour récupérer la normalisation du mot ainsi que ses informations linguistiques (genre, nombre, catégorie grammaticale, etc.). L'une des particularités de la langue arabe est la présence des formes agglutinées (formes avec des proclitiques et des enclitiques). Ces formes ne sont pas présentes dans le dictionnaire des formes fléchies. Pour identifier ces formes et les traiter correctement, nous avons ajouté un segmenteur de clitiques (proclitiques et enclitiques) à l'analyse morphologique.

12.2.3. Repérage des dates

Lors de l'analyse morphologique, un traitement spécifique intervient pour le repérage des dates. Ceci permet ensuite à la désambiguïsation d'être plus efficace, étant donné que les dates ne sont plus constituées d'une suite de catégories, mais sont associées à une catégorie « date ».

Les dates et heures se composent de l'indication normalisée du temps qu'elles représentent. Nous nous sommes basés sur la norme ISO-8601 avec le format AAAAMMJJ où AAAA représente l'année sur 4 chiffres, MM représente le numéro du mois, sur 2 chiffres compris entre 01 et 12 et JJ représente le quantième dans le mois, sur 2 chiffres. Par exemple, «1984 أفريل 01» est normalisé de la manière suivante : «19840401 ». Il arrive qu'une date ne soit renseignée qu'en partie. Les parties inconnues sont alors remplacées par des X Comme nous pouvons le voir sur l'exemple suivant. « أفريل (Avril) 1984 » est normalisé en «198404XX», «غداً : demain» est normalisé en «XXXXXX+1».

12.2.4. Identification des relations syntaxiques

Cette étape permet de représenter la structure syntaxique d'un texte sous forme symbolique et graphique. Elle définit les relations syntaxiques entre les mots. Pour réaliser cette tâche, nous appliquons une analyse linguistique basée sur la grammaire définie par Tesnière. Nous procédons en deux étapes : 1) traiter les syntagmes nominaux, puis 2) présenter les relations sujet-verbe-complément.

12.2.4.1. Les syntagmes nominaux

Une étude linguistique spécifique de la langue arabe nous a permis de définir et d'écrire un certain nombre de règles dans le but d'établir des relations de dépendance (contiguës et non contiguës) entre les mots au sein du syntagme nominal. Ces relations permettent ensuite de reconnaître les mots composés présents dans une phrase.

Citons l'exemple de «أرملـة الشهيد» (la veuve de martyr), nous avons une relation entre deux mots associés par annexion (معرف بالإضافة), qui relie le mot indéfini أرملـة (veuve) et le mot défini الشهيد (martyr) pour donner une relation de type <*NomRelNom*> qui est une relation de type <***DG***> permet de relier des éléments dont la tête est à droite du dépendant.

12.2.4.2. Relations sujet-verbe-complément

Nous avons défini un certain nombre de règles, issues d'une étude expérimentale pour l'identification et le repérage des relations syntaxiques dans une phrase. Notons que certains verbes peuvent avoir un complément d'objet direct, contrairement à d'autres. Ces verbes sont appelés des verbes transitifs. Il faut définir la liste des verbes transitifs et des verbes intransitifs, étant donné que, en arabe, la position des mots ne suffit pas à déduire la fonction syntaxique du mot. Les voyelles courtes l'indiquent, mais elles ne sont généralement pas indiquées dans les textes écrits. C'est pour cela qu'il faut se baser sur la transitivité ou sur la non transitivité du verbe pour déterminer quelles sont les relations qui existent entre un nom et un verbe.

Voici les relations que nous détectons :

- Les relations agent-verbe, qui permettent d'identifier l'agent de l'action (pour répondre à la question : qui a fait l'action?)
- Les relations verbe-complément, qui permettent d'identifier qui a subi l'action, ou encore les circonstanciels qui nous renseignent sur le moyen (comment? Avec quoi?), la date (quand?), le lieu (où ?), ... de l'action.

12.2.4.2.1. Gestion de la forme passive

Afin de rendre l'étape de construction des règles d'extraction des connaissances plus efficace, l'analyse linguistique profonde adopte en interne la même représentation pour une phrase passive et pour son équivalent à la forme active. Cette phase consiste donc à identifier les formes passives et à les transformer en formes actives. Voici quelques structures syntaxiques exprimant un passif (Ziad, 2010) :

- Passif avec un verbe doublement transitif. مُنِح الشاعرُ جائزةً : On a accordé un prix à l'écrivain
- Passif avec un verbe transitif indirect, précédé par une préposition, حُكِمَ عليه بالاعدام : Il a été condamné à mort.
- Emploi de tournures modernes du passif, qui expriment le complément d'agent : مِنْ قِبَل ، مِنْ طَرَفِ ، مِنْ جَانِب ، عَلَى يَد (par).

Dans l'exemple suivant : إعتقلت الفتاة على يد الشرطة (la fille a été arrêtée par la police), pour ne pas confondre entre la personne qui fait l'action et la personne qui la subit, il est important de savoir si la forme est active ou passive. Ici, la forme est active mais emploie une tournure moderne du passif qui exprime le complément d'agent (عَلَى يَد), donc le sujet est la police et le complément est la fille. Nous obtenons donc les relations suivantes :

- relation agent-verbe entre إعْتَقَلَ (arrêter) et شُرْطَةٌ (Police) reliés par le mot عَلَى يَد (par)
- relation verbe-complément entre إعْتَقَلَ (arrêter) et فَتَاة (fille).

```
<relation reltype="SV">
  <head>
    <posBeg>112</posBeg>
    <lemma>إِعْتَقَلَ </lemma>
    <catPos index="no">+verbe</catPos>
<prop
index="no">+vbpassif+acc+3fs</prop>
    <posEnd>118</posEnd>
  </head>
  <dept>
    <posBeg>133</posBeg>
    <lemma>شُرْطَةٌ</lemma>
    <catPos index="no">+nom</catPos>
    <prop index="no">+fs</prop>
    <posEnd>139</posEnd>
  </dept>
  <lingIndication index="no">
    <posBeg>126</posBeg>
    <lemma>عَلَى يَدِ</lemma>
    <catPos
index="no">+prepN</catPos>
    <prop index="no">+passif</prop>
    <posEnd>132</posEnd>
  </lingIndication>
</relation>
```

```
<relation reltype="VC">
  <head>
    <posBeg>112</posBeg>
    <lemma>إِعْتَقَلَ </lemma>
    <catPos
index="no">+verbe</catPos>
    <prop
index="no">+vbpassif+acc+3fs</prop
>
    <posEnd>118</posEnd>
  </head>
  <dept>
    <posBeg>119</posBeg>
    <lemma>فَتَاة</lemma>
    <catPos
index="no">+annppers</catPos>
    <prop index="no">+pers+fs</prop>
    <posEnd>125</posEnd>
  </dept>
</relation>
```

Les principales balises du contenu HTML ci-dessus sont :

- **Head :** unité qui constitue la tête de la relation
- **Dept :** unité qui constitue le dépendant de la relation
- **LingIndication :** balise qui contient des indications sur les unités qui permettent de relier les termes d'une relation, et qui serviront lors de l'extraction sémantique.

12.2.5. Reconnaissance des entités nommées

Cette phase consiste à mettre en œuvre un système de reconnaissance et de typage des entités nommées. Dans notre approche, nous avons opté pour un système à base de règles linguistiques qui exploitent l'étiquetage syntaxique, des marqueurs lexicaux (que nous appelons des déclencheurs) et des dictionnaires de noms propres. La mise en place de règles de reconnaissance d'entités nommées a nécessité une recherche profonde sur certains traits linguistiques propres aux entités nommées en arabe.

- **Exemple :** " الأخ مُعز غرسلاوي " (le frère Moez Garsallaoui) Dans cet exemple, nous avons le titre de civilité "الأخ : frère" suivi d'un prénom et d'un nom propre. Voici la représentation que nous obtenons :

```
<en entype="pers">
 <relation reltype="AnnpNP">
  <head>
   <posBeg>1104</posBeg>
   <lemma>غرسلاوي</lemma>
   <catPos index="no">+np</catPos>
   <posEnd>1111</posEnd>s
  </head>
  <dept>
   <posBeg>1092</posBeg>
   <lemma>أَخٌ</lemma>
   <catPos
index="no">+annppers</catPos>
   <prop
index="no">+pers+ms</prop>
   <posEnd>1097</posEnd>
  </dept>
 </relation>
```

```
<relation reltype="PrenomNP">
  <head>
   <posBeg>1104</posBeg>
   <lemma>غرسلاوي</lemma>
   <catPos index="no">+np</catPos>
   <posEnd>1111</posEnd>
  </head>
  <dept>
   <posBeg>1098</posBeg>
   <lemma>مُعِز</lemma>
   <catPos
index="no">+prenom</catPos>
   <prop index="no">+m</prop>
   <posEnd>1103</posEnd>
  </dept>
 </relation>
</en>
```

12.3. Extraction sémantique

Les inputs de cette étape sont les outputs de l'analyse linguistique (morphosyntaxique) décrite dans les chapitres 2 et 3. Cette analyse fournit les informations suivantes :

- les lemmes des mots ainsi que leur position dans le texte, et leur catégorie grammaticale
- Les relations de dépendance syntaxique entre les mots,
- Les entités nommées typées.

L'extraction sémantique de l'arabe s'insère dans un outil qui effectue aussi de l'extraction en français, anglais et chinois. Nous avons choisi une représentation interlingue en anglais de toutes les informations, quelle que soit la langue du texte d'origine, dans le but de faciliter la lecture des informations extraites par les non arabophones et de faciliter la fusion d'informations provenant de document en plusieurs langues. Nous avons eu recours à deux types d'opérations : l'utilisation de dictionnaires de traduction existants et l'ajout d'un système de translittération pour les entités nommées qui n'existent pas dans les dictionnaires (Saadane et al., 2012).

L'extraction de connaissances permet de mettre en évidence des entités nommées et des relations relatives à un concept particulier et décrit dans l'ontologie du domaine, par exemple : «contact», « arrestation », « attentat », « condamnation », « construction ». Le déroulement de cette étape s'effectue en trois temps : (i) la sélection des concepts potentiellement présents, (ii) la sélection des règles à appliquer, puis (iii) l'application des règles.

12.3.1. Sélection des concepts probables

Une étape primordiale lors de l'extraction des connaissances consiste à repérer les déclencheurs. Ces déclencheurs peuvent être des mots, des expressions ou des relations, et indiquent qu'une relation relative à un concept peut être présente dans le texte et ensuite extraite. Une étude des corpus et des textes, nous a permis de définir et d'enrichir la liste des mots déclencheurs ainsi que leurs synonymes et la construction des ressources linguistiques

nécessaires et pertinentes pour l'identification automatique de l'information présentée dans les textes. Voici une liste non-exhaustive de quelques catégories d'indicateurs des verbes :

Emission :صَرَّحَ, وَجَّهَ, أَرْسَلَ, أَعْلَنَ ...
Transfert : إِنْتَقَلَ,
Contact : إِجْتَمَعَ, تَلَاقَى, نَاقَشَ, إِنْعَقَدَ, مُقَابَلَةٌ, إِجْتِمَاعٌ, حِوَارٌ...
Mort : وَفَاة، مَوْت، مَنِيَّة،

Il convient de rappeler qu'à la sortie de l'analyse morphosyntaxique les « mots » sont sous forme simple, donc les déclencheurs (appelés aussi des marqueurs) sont des formes simples. Par conséquent, les listes ne contiennent que les formes simples des marqueurs, ce qui signifie la forme masculin singulier du verbe à l'accompli et le masculin singulier des noms. Rappelons que l'analyse morphosyntaxique est l'étape qui permet de cerner la problématique des formes fléchies et agglutinées des marqueurs linguistiques.

Ces déclencheurs sont listés sous la forme d'un dictionnaire de données, correspondant à des couples clé/valeur. Ils sont présentés sous forme de deux colonnes :

- La première colonne présente la clé et elle contient les mots, les expressions ou encore les relations qui indiquent la présence du concept dans le texte traité.
- La deuxième colonne présente la valeur et elle définit et précise le concept de l'ontologie (arrestation, transfert, émission, Union,...) associé à l'élément de la première colonne (la clé).

Clé	**valeur**
رَجَعَ	Transfer
إِعْتَقَلَ	Arrest
إِقْتَرَنَ	Union
VC#رِسَالَة#نَقَلَ#	Emission

Comme nous l'avons mentionné et présenté dans l'exemple, les déclencheurs peuvent être des mots (...إِعْتَقَلَ، إِقْتَرَنَ) :(se marier, interpeller…), des expressions ou bien des relations (VC# رِسَالَة # نَقَلَ#) : (VC#transmettre#message#). Il est nécessaire qu'un déclencheur soit présent dans les relations identifiées lors de l'analyse syntaxique afin d'être en mesure d'extraire l'information présente. À partir d'un déclencheur, un concept est proposé ce qui nous permet ensuite de sélectionner les règles à appliquer telles que définies à la section suivante.

12.3.2. Sélection des règles à appliquer

Les déclencheurs nous ont permis d'obtenir la liste des concepts présents dans le texte. Ces concepts nous amènent alors vers une liste de règles qui vont être confrontées aux relations issues de l'analyse syntaxique. Si des relations syntaxiques correspondent aux règles définies, les relations pourront alors être extraites. La définition des règles à appliquer est construite sous forme d'un dictionnaire comprenant aussi deux colonnes :

- La première colonne indique les concepts. Ils sont ensuite comparés aux concepts probables sélectionnés lors de la phase précédente par le biais des déclencheurs
- La deuxième colonne liste les règles spécifiques à un concept pouvant être appliquées.

Concepts	Règles	Traduction
Arrest	SV#إِعْتَقَلَ#<en>#عَلَى يَدِ	SV#arrêter#<en># par
Arrest	SV#إِعْتَقَلَ#<en>#	SV#arrêter#<en>#
Arrest	VC#إِعْتَقَلَ#<en>#	VC#arrêter#<en>#

Pour illustrer notre propos, prenons l'exemple suivant : إعتقلت العروض على يد الشرطة (El Aroud a été arrêtée par la police). Lors de la première étape, le mot «إعتقل» (arrêter) a été repéré comme déclencheur, dégageant alors le concept « Arrest» (arrestation). Ce concept est associé aux règles présentées ci-dessus. Afin de pouvoir être sélectionnées, les règles doivent correspondre à une relation syntaxique présente dans la phrase, il est nécessaire qu'apparaissent dans les relations syntaxiques, une relation sujet-verbe entre « arrêter » et une entité nommée, une relation verbe-complément entre « arrêter » et une entité nommée (ou un annonceur d'un organisme) ou bien une relation verbe-complément entre « arrêter » et un groupe prépositionnel. Or, dans la phrase « El Aroud a été arrêtée par la police », « arrêter » a un complément « El Aroud » de type entité nommée, et « arrêter » a un sujet « police », étant donné que la phrase est au passif. Donc, ce sont les règles SV# إِعْتَقَلَ #<en># عَلَى يَدِ et VC# إِعْتَقَلَ #<en># qui seront sélectionnées.

À ce stade intervient un traitement qui permet l'application des règles sélectionnées, lors de la phase précédente aux relations syntaxiques effectivement présentes, afin d'extraire l'information.

12.3.2.1. L'application des règles sélectionnées

Les règles sélectionnées à l'étape précédente sont transcrites dans la structure de l'automate supportant les traitements linguistiques et sont appliquées aux relations syntaxiques afin d'extraire les connaissances. Cette opération est dirigée par le format de sortie du module d'analyse linguistique indiquant comment la connaissance devra être extraite et quels éléments devront être conservés. La règle d'extraction indique ensuite quels sont les éléments qui doivent être extraits, et quelle est leur sémantique. Si nous reprenons l'exemple « إعتقلت العروض على يد الشرطة : El Aroud a été arrêtée par la police», l'une des règles sélectionnée est la relation verbe-complément entre «arrêter» et une entité nommée. La règle à appliquer sera donc la suivante :

VC# إعتقل*#***<pers>**# → *<gs:Arrest rdf:nodeID="dXXArrest">*
*<wn:undergoer rdf:nodeID="***pers***"/>*
</gs:Arrest>

Cette règle indique que le concept extrait sera «Arrest» (Arrestation), dont l'action est «arrêter» et le patient est l'objet de «arrêter», c'est-à-dire «El Aroud». Le résultat obtenu sera alors de la forme suivante : *<gs:Arrest rdf:nodeID="d17Arrest"><wn:undergoer rdf:nodeID="d1Elaroud"/></gs:Arrest>*

Il convient de souligner que le format de sortie est un graphe RDF.

12.3.2.2. L'extraction des entités nommées et son contrôle

Toutes les entités nommées détectées au niveau de l'analyse linguistique sont extraites en conservant leur type : « Personne », « Lieu », « Organisme », « Mesure », « Date », « Produit » ou encore « Inconnu ». Cependant, il arrive que lors de l'analyse linguistique, le type d'une EN soit erroné. De ce fait, nous proposons lors de cette étape d'effectuer un contrôle sur le type des entités nommées extraites.

Ce contrôle consiste à vérifier l'adéquation entre le type de l'entité nommée issu de l'analyse linguistique, et le type de l'entité nommée proposé par la règle. Les types des entités nommées peuvent être modifiés si la règle d'extraction considère que le type de l'entité nommée est incompatible avec le type de l'entité nommée issu de l'analyse linguistique.

L'exemple suivant illustre ce phénomène d'incompatibilité : «أعلنت باريس أن الجزائر ...: Paris déclare que l'Algérie ...». Au niveau morphosyntaxique, Paris est considérée comme une entité nommée de type «lieu», mais dans le cadre de l'action d'émission d'un message, l'agent ne peut pas être un lieu. En fait, l'émission ne peut être réalisée que par une personne ou une organisation et si à l'origine elle a été considérée comme la capitale de la France, la nouvelle catégorie est une organisation, et nous pouvons en déduire que c'est le gouvernement français.

12.3.3. La création d'entités nommées

Il arrive qu'une règle fasse référence à une entité sans que celle-ci n'existe. C'est le cas lorsque l'entité nommée n'a pas pu être repérée au niveau de l'analyse syntaxique, ou bien lorsqu'il ne s'agit pas d'une entité nommée comme dans l'exemple «أدين : il a été condamné». Pour faire apparaître le patient de l'action, la règle d'extraction va créer l'entité manquante :
VC# أَدَانَ#**<$pers>**# → *<en entype=βers">***<$pers>**</en>

Notons que l'extraction et/ou la création d'entités nommées peuvent introduire des ambiguïtés. Une relation peut demander comme objet ou sujet une entité nommée de type lieu ou organisation. Ces ambiguïtés sont alors générées, pour être résolues plus tard grâce à la mise en cohérence. Cependant, lorsque l'ambiguïté se situe entre les types personne ou organisation, le type « agent », qui est générique, sera indiqué.

12.4. Mise en cohérence

Les connaissances extraites lors du traitement décrit dans la section précédente sont représentées au format RDF faisant référence aux concepts et propriétés issus de l'ontologie intégrée dans le système. Nous disposons alors en entrée de ce traitement d'un graphe de connaissances.

L'étape de mise en cohérence correspond aux opérations de consolidation, résolution d'ambigüités et enrichissement de ce graphe. Cette étape est commune à toutes les langues, et va permettre de rassembler les informations concernant une même entité nommée, ou une même action. Elle permet aussi d'exploiter les métadonnées (date et lieu d'émission par exemple) attachées au document. La construction de notre système d'extraction a nécessité de définir la nature des informations d'intérêt dans le domaine de sécurité. Nous avons choisi, pour cela, de développer une ontologie du domaine qui servira de guide aux différentes étapes d'extraction. La diversité des documents exploités nécessite que l'ontologie soit assez générale tout en définissant les concepts et les propriétés relatifs au domaine de la sécurité. Les traitements effectués sont les suivants :

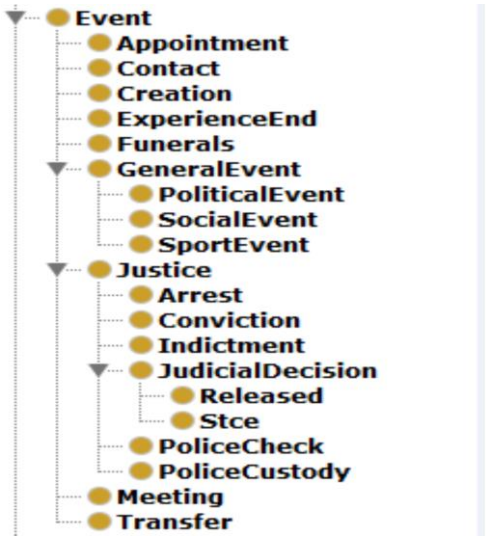


Figure 12. 2. Extrait de l'ontologie GEOLsemantics.

12.4.1. Regroupement des entités nommées

L'un des problèmes des différentes étapes d'extractions réside dans le fait que les graphes obtenus peuvent contenir des duplications inutiles de nœuds. Ce phénomène est particulièrement visible pour les entités nommées que l'on retrouve à plusieurs reprises dans un même document. L'objectif de cette étape consiste à regrouper les différentes occurrences d'une même entité nommée sous un même et unique identifiant. Ce problème est généralement connu sous le nom de 'Record linkage' ou 'Entity resolution' et a été abordé par différentes approches (Elmagarmid et al., 2007).

Dans le contexte d'un graphe RDF, et dans le domaine de l'extraction sémantique, nous adoptons une méthode basée sur un ensemble de règles. Ces règles ont été définies pour identifier les entités nommées dupliquées et permettre leur regroupement. Citons un exemple de ces règles : deux personnes sont identiques dans un même document, si elles ont le même nom et prénom, et qu'il n'y a pas d'autres informations contradictoires, par exemple «junior» et «senior».

12.4.2. Résolution des dates relatives

Parmi les problèmes que l'analyse linguistique ne résout pas il y a les dates relatives. Ces dates ne sont pas toujours exprimées d'une manière explicite dans les textes. Pour résoudre ce phénomène, nous nous appuyons sur les trois aspects suivants :

1. La représentation adoptée par l'ontologie : l'ontologie décrit chaque date comme un intervalle. Elle contient donc les attributs suivant : (1) **dtstart :** date de début, (2) **dtend :** date de fin, (3) **type :** le type de calendrier utilisé, qui correspond à des constantes prédéfinies dans l'ontologie (grégorien, arabe, chinois …), (4) **authorValidation :** donne une indication sur quand a eu lieu l'action, si cette dernière se situe dans le passé ou le futur, grâce notamment aux temps des verbes liés à la date, (5) **day :** le jour de la semaine, lorsqu'il est précisé.
2. la sortie de l'analyse linguistique
3. les métadonnées du document analysé (notamment la date d'édition du document).

La sortie de l'analyse linguistique nous permet d'identifier les occurrences où la date extraite est incertaine. Dans le contexte de la presse écrite, il est fréquent d'extraire des dates relatives à un jour de la semaine ou à une indication dans le temps. Par exemple un événement devant se dérouler « نهاية الأسبوع المقبل : à la fin du week-end prochain » pour un article paru le

lundi 01 avril 2013 (une métadonnée du document). Les métadonnées sont alors exploitées pour définir une date incertaine se situant entre le samedi 06/4/2013 et le dimanche 07/4/2013.

À l'issue de tous ces traitements nous disposons alors d'un ensemble de triplets décrivant les connaissances extraites. Le but de la conceptualisation grâce à l'ontologie est de formaliser les informations tout en gardant leur potentiel sémantique exprimé dans le texte.
La figure 12.3 présente un exemple d'extraction de connaissances obtenu par l'analyse et les différents traitements présentés dans la phrase : « إلْتَقَى الرئيس الجزائري عبد العزيز بوتفليقة بالرئيس السوري بشار الأسد في الجزائر في جانفي 2012 لمناقشة الإرهاب. » 'Le président algérien Abdel Aziz Bouteflika a rencontré le président syrien Bachar El Assad en Algérie, le mois de janvier 2012 pour discuter la question du terrorisme'.

Notre analyse linguistique permet d'extraire les entités nommées ainsi que les relations syntaxiques entre ces entités. En s'appuyant sur les résultats de l'analyse linguistique ainsi que l'analyse sémantique, notre système extrait les connaissances suivantes :

- entité nommée de type ***«Personne»*** : Président algérien Abdel Aziz Bouteflika et Président syrien Bachar El Assad
- entité nommée de type ***«Lieu»*** : Algérie
- concept ***«Contact»*** extrait grâce au verbe «إلْتَقَى» «se rencontrer», avec deux agents : «Président algérien Abdel Aziz Bouteflika » et «Président syrien Bachar El Assad».
- Date de rencontre : 201201XX.

Le concept ***«Contact»***, peut être extrait grâce au verbe «إلْتَقَى» ce contact est établi entre deux ou plusieurs personnes (agents), dans un lieu géolocalisé ou dans une zone inconnu. Ce contact est établi dans une date précise ou une plage temporelle d'incertitude. Il convient de souligner si le jour n'est pas précisé, l'ontologie permet de mettre une plage temporelle du 01-01-2012 jusqu'à 31-01-2012.

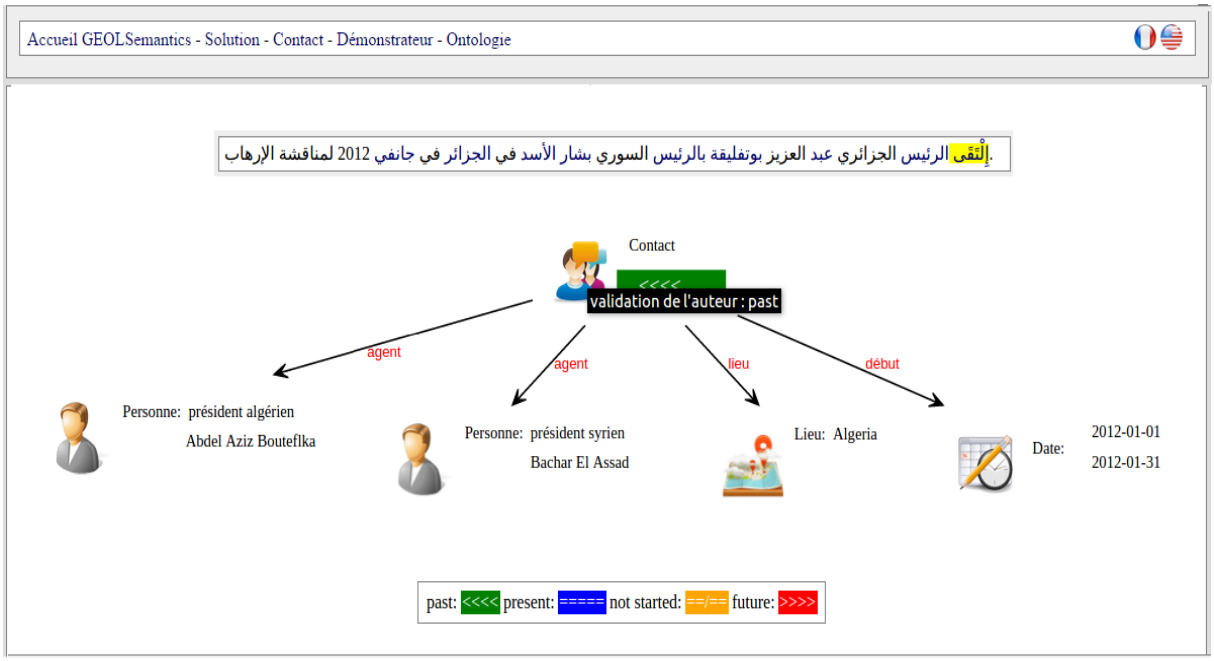


Figure 12. 3. Exemple de graphe RDF

12.5. Évaluation du système arabe

Pour estimer l'efficacité de notre système, nous avons mené deux types d'évaluations : i) une évaluation quantitative concernant la phase de segmentation et l'extraction d'entités nommées et ii) une évaluation qualitative intrinsèque concernant l'extraction de connaissances.

12.5.1. Évaluation de la segmentation

L'évaluation présentée dans cette section concerne la mesure des performances du module de segmentation en mots de phrase. A cet effet, nous avons utilisé deux segmenteurs : le segmenteur GEOLSemantics de notre analyseur et le segmenter de Stanford[23]. Le fait d'utiliser ces deux ségmenteurs est justifié par la nécessité de comparer les performances de notre segmenteur avec un autre outil similaire du marché. Nous renvoyons le lecteur au chapitre 3 (Section 3.2.2.2), pour plus de détails sur les données utilisées pour l'apprentissage des modèles de prédiction d'étiquetage morphosyntaxique et de segmentation.

12.5.1.1. Corpus

Vu que notre analyse morphosyntaxique vise des textes provenant des sites webs d'information, nous avons choisi de construire un corpus qui contient des articles de presse non voyellés d'actualité sur le sport, la politique, la société et l'international. Les sites que nous avons choisi sont : celui de la chaîne *Aljazeera*[24] et celui du journal *El-Khabar*[25]. Nous avons effectué des traitements sur la totalité de corpus mais seulement 500 textes sont choisis pour le calcul du score.

12.5.1.2. Déroulement et résultats

Le corpus est analysé par notre segmenteur, dédié pour la langue arabe, et ainsi que celui de *Stanford*. Nous avons procédé à la mise en œuvre de quelques programmes pour enlever les étiquettes morphosyntaxiques du résultat. Ces opérations de formatage ont pour objectif l'harmonisation des formats des résultats avec ceux produits par l'outil de Standford. Nous avons aussi apporté quelques modifications sur nos règles pour les adapter et pouvoir les comparer avec cet outil. A cet effet, nous citons par exemple que, dans les outils de Stanford, l'article défini fait partie du mot contrairement à notre segmenteur qui le considère comme un proclitique.

Il convient de rappeler par ailleurs, que la phase de segmentation de la langue arabe est compliquée par rapport à celle des autres langues (comme le français ou l'anglais). Cette complication vient de l'absence en arabe des majuscules, des ponctuations régulières, etc. De plus, il existe aussi d'autres facteurs qui rendent la segmentation en arabe plus compliquée, comme le phénomène de l'agglutination. Nous signalons aussi qu'au cours de la phase de segmentation, notre système a besoin de connaître, dans certains cas, la catégorie du mot analysé. Les mots concernés sont ceux positionnés avant et/ou après le signe de ponctuation. Cependant, le système ne dispose pas en général à ce niveau des éléments indispensables pour fournir ce type d'informations. De ce fait, nous sommes contraints d'anticiper l'analyse et de communiquer avec les autres phases de traitement. Les phases concernées sont : l'analyse Morphologique pour lever l'ambiguïté au niveau de la segmentation et l'analyse syntaxique pour vérifier s'il y a une règle grammaticale validant l'une des solutions proposées par l'analyse morphologique (Belguith et al., 2007).

[23] http://nlp.stanford.edu/projects/arabic.shtml
[24] http://www.aljazeera.net/portal
[25] http://www.elkhabar.com/

12.5.1.3. Protocole expérimental

Afin d'évaluer la performance des systèmes de segmentation, nous utilisons un ensemble de métriques. Dans notre expérimentation nous avons utilisé les métriques des taux de précision (P) et de rappel (R) ainsi que la F-Mesure. Selon (Mesfar, 2008) ces métriques sont définies comme suit :

- **Le rappel (R) :** est une évaluation de la couverture du système. Il mesure la quantité de réponses pertinentes d'un système par rapport au nombre de réponses idéales selon la formule suivante :

$$R = \frac{total\ des\ mots\ correctes\ du\ test}{\text{total des mots de test}}$$

- **La précision (P)** : est une évaluation du bruit du système. Elle mesure la proportion de réponses pertinentes du système parmi l'ensemble des réponses qu'il a fourni en suivant la formule :

$$P = \frac{total\ des\ mots\ correctes\ du\ test}{\text{total des mots de la référence}}$$

- **La F-Mesure (F) :** est une métrique qui permet de combiner en une seule valeur les mesures de précision et de rappel de manière à pénaliser les trop grandes inégalités entre ces deux mesures. Elle favorise les systèmes dont les deux valeurs sont homogènes. La formule de cette métrique est donnée comme suit :

$$F = \frac{2 \times P \times R}{P+R}$$

L'évaluation de notre système de segmentation a montré un taux de précision de 97,53% et un taux de rappel de 97,87 % contre un taux de précision de 96, 61 % et un taux de rappel 94,73 % pour celui de Stanford. Nous remarquons que le système Stanford n'a pas pu segmenter certaines phrases à un tel point qu'il ne les affiche même pas. Il est à signaler aussi qu'une grande partie des mots non segmentés correctement est engendrée par le fait qu'il existe des mots propres qui n'existe pas dans le dictionnaire ou des mots inconnus (hors vocabulaire).

Par rapport à la couverture lexicale de nos ressources, elle est mesurée par le taux des mots inconnus détectés lors de l'analyse automatique du corpus. Nous notons que notre analyse de corpus a abouti à une couverture lexicale de 91,92% contre 8,08 % des mots hors vocabulaires ou inconnus. Il est à noter que la liste des mots inconnus comporte principalement :

- Les erreurs d'orthographe, parmi ces erreurs, nous citons l'omission de l'espace entre les mots, la répétition des lettres, …
- Les noms propres (noms de personnes, d'organisation, de lieu) et spécialement les entités nommées transcrites, des mots étrangers ou empruntés, etc.

Pour ce qui est de l'étiquetage morphosyntaxique (POS), sur les textes analysés nous avons eu un taux d'erreur de 4,6%.

12.5.2. Évaluation de l'extraction des entités nommées

12.5.2.1. Description du Corpus arabe

Les expériences menées dans cette phase concernant la détection des entités nommées qui ont été réalisées sur le corpus ANER[26] (Benajiba et al., 2007), qui est composé de 150 000 occurrences de mots. Ce corpus distingue les types d'entités nommées suivants : lieux (LOC) avec 40 % des ENs observées, personnes (PERS) avec 32 % des ENs observées, organisations (ORG) avec 18% et un autre type 'divers' (MISC) qui regroupe tous les autres types avec une observation de 10 %). Nous notons que seuls les trois premiers types sont utilisés pour notre évaluation. La répartition en ENs est présentée dans le tableau suivant :

	LOC	**ORG**	**PERS**	**MISC**
Entités nommées	4431	2026	3602	1117
Entités nommées distinctes	1004	657	1446	437

Tableau 12. 1. La répartition des entités nommées dans le corpus ANER

Il convient de signaler que nous avons apporté quelques modifications sur les frontières des entités nommées proposées dans le corpus ANER. L'exemple suivant montre un extrait d'une phrase en format BIO. L'étiquette B-X (Begin) indique le premier mot d'une EN de type X. L'étiquette I-X (Inside) indique qu'un mot fait partie d'une EN mais qui n'est pas le premier mot. L'étiquette O (Outside) est utilisée pour les mots qui ne sont pas des ENs.

وزير	الخارجية	السوري	وليد	المعلم
O	O	O	B-PERS	I-PERS
Président	Affaires étrangères	syrien	Walid	Maalam

Tableau 12. 2. Exemple d'un extrait de phrase en format BIO

Dans notre traitement, nous n'avons pas les mêmes frontières pour les ENs. Au moment où l'outil ANER ne considère que le prénom et le nom propre pour l'EN, nous considérons pour l'EN dans notre analyseur en plus de ces éléments les annonceurs de personnes, organisations, etc. Le tableau suivant donne la sortie de notre traitement en prenant les mêmes mots :

وزير	الخارجية	السوري	وليد	المعلم
I-PERS	I-PERS	I-PERS	B-PERS	I-PERS
Président	Affaires étrangères	syrien	Walid	Maalam

Tableau 12. 3. Exemple d'un extrait avec notre analyse

Nous notons que l'annonceur exprimé par '*le président des affaires étrangères syrien*' fait aussi partie de notre entité nommée de type personne, le prénom *Walid* et le premier élément de notre entité mais il ne présente pas la tête de notre relation syntaxique. Le nom propre المعلم *Maalam* présente le deuxième élément de l'entité nommée mais il s'agit de la tête autrement dit l'élément le plus important de la relation syntaxique.

12.5.2.2. Protocole expérimental

Les expériences sont réalisées à partir de données segmentées et analysées avec les outils que nous avons développés. L'extraction des entités nommées est à base de règles. Les

[26] http://users.dsic.upv.es/~ybenajiba/downloads.html

scores sont calculés en utilisant l'outil d'évaluation développé pour la tâche de repérage des ENs proposée dans le cadre de CoNLL 2002[27]. Cet outil calcule les métriques d'évaluation présentées dans la Section 12.5.1.3. Il convient de rappeler que selon les lignes directrices d'évaluation CoNLL, une entité nommée est considérée comme correctement détectée si seulement si la classification est correcte et ainsi que tous les mots constitutifs de l'entité nommée sont reconnus. Les résultats obtenus sont présentés dans le tableau suivant :

Catégories	Précision	Rappel	F-mesure
Personne	75,40%	55,58%	64,31%
Location	89,11%	81,49%	85,12
Organisation	63,62%	43,33%	52,25

Tableau 12. 4. Performances du système de reconnaissance des entités nommées

Les résultats montrent que la performance du système varie selon le type d'entité nommée à identifier. Pour la détection des entités nommées de type '*lieu*', le tableau montre que nous avons obtenu le meilleur F-mesure (85,12%). Ceci peut être expliqué par la bonne couverture lexicale des lieux ainsi que les règles définies pour l'identification de ce type d'entités nommées. Pour les entités de type '*personne*' et '*organisation*', notre système présente encore quelques faiblesses. Dans ce registre, il convient de rappeler que les autres systèmes présentent aussi des difficultés similaires pour le traitement de ces deux catégories (Zaghouani, 2012).

12.5.2.3. Discussion des erreurs détectées

L'analyse des erreurs de détection pour les ENs de type '*Organisation'* a montré qu'ils sont dus parfois à des facteurs tels que l'incohérence de la transcription des noms d'organisation étrangère à l'arabe. Ce phénomène peut être vu aussi lors de la transcription des ENs de type '*personne'* ou '*lieu'*. Les résultats ont révélé aussi de nombreux cas de qualification erronée en raison de l'ambiguïté de certains mots arabes. Cette ambiguïté est engendrée en partie par le problème de la non-vocalisation des textes qui augmente son ambiguïté, affectant ainsi les systèmes de reconnaissance des entités nommées. Un autre type d'erreur est aussi remonté par les résultats et concerne les éléments des ENs partiellement reconnus, en particulier ceux des noms d'organisation et de personne.

Par ailleurs, l'absence de normes rigoureuses pour la saisie au clavier arabe a conduit à des incohérences dans l'orthographe de certains mots et donc a influencé nos résultats. Cette absence augmente le nombre de possibilités orthographiques pour une EN données qu'elles ne peuvent pas être toutes renseignées dans nos ressources. Les possibilités non renseignées créent par conséquent des ambiguïtés supplémentaires. De la même manière, certaines erreurs sont également dues à la présence de variantes orthographiques des entités traduites ou transcrites qui ne figuraient pas dans nos dictionnaires. Par exemple, un nom de lieu comme *Angleterre* pourrait être écrit de trois manières différentes : انجلترا، انكلترا، انقلترا.

12.5.3. Évaluation de l'extraction de connaissances

L'absence ou l'indisponibilité des outils et des travaux de référence, traitant ce domaine d'extraction des connaissances pour le traitement de l'arabe, a été un vrai obstacle pour mesurer la performance de notre système. Ceci nous a handicapés dans la mesure où nous n'avons pas le moyen de comparer notre approche avec d'autres propositions. C'est la raison pour laquelle, nous avons lancé des phases de tests afin d'améliorer et de compléter

[27] http://bredt.uib.no/download/conlleval.txt

l'extraction d'informations. Dans cette phase nous avons tenté de répondre à des questions comme la suivante : sur quel corpus peut-on tester notre module ? Pour y répondre nous avons opté pour les corpus suivants :

– Corpus de textes sur Malika El-Aroud. Ce corpus est assez général et regroupe une grande partie des concepts présents dans notre ontologie.

– Ensemble de corpus propres à chaque concept, composés d'articles journalistiques. Ces corpus ne sont pas généraux mais peuvent permettre d'étudier et d'améliorer en profondeur un type de concept. Les corpus spécifiques sont les suivants : « arrestation », « attentat », « condamnation », « construction », « décès », « divorce », « émission », « mariage », « paiement », « rencontre » et « transfert ».

Afin de recouvrir un maximum de cas tout en améliorant la reconnaissance et l'extraction d'information, l'utilisation conjointe de ces deux types de corpus paraît être la meilleure solution. Voici un exemple d'un extrait de phrases issu du corpus Malika El-Aroud :

مليكة العروض من أصول مغربية الملقبة باسم أم عبيدة و أميرة الجهاد تبلغ من العمر إثنين وخمسين سنة .
. . . . أم عبيدة أو أميرة الجهاد أرملة الشهيد عبد الستار دهمان التونسي منفذ العملية الاستشهادية
. . . . وفي مرحلة لاحقة اقترنت العروض بالأخ معز غرسلاوي من صول تونسية وانتقلت معه لسويسرا
. . . وإعتقلت الأخت أم عبيدة من طرف الشرطة البلجيكية في 11 ديسمبر 2008 في بلجيكا

Notre analyse linguistique permet d'extraire les entités nommées ainsi que les relations syntaxiques entre ces entités. En s'appuyant sur les résultats de l'analyse linguistique ainsi que l'analyse sémantique, notre système extrait les connaissances suivantes :

– entité nommée de type ***«Personne»*** : Malika Elaroud, Moez Garsallaoui et Abd Elsattar Dahmane ;

– entité nommée de type ***«Lieu»*** : Switzerland, Belgium ;

– entité nommée de type ***«Organisation»*** : Police ;

– concept ***«Union»*** extrait grâce au verbe «إقترنت» «se marier», avec deux bénéficiaires : «Umm Obeyda » et « Moez Garsalloui».

– concept ***«Transfert»*** est extrait grâce au verbe «انتقل» d'une personne Malika Elaroud à un lieu « Switzerland »;

– concept ***«Relation Familiale»*** : Malika Elaroud veuve d'Abd Elsattar Dahmane ;

– concept ***«Arrestation»*** : d'une personne *Malika Elaroud* par une organisation *police* ;

– Date d'arrestation : 2008-12-11 ;

– Lieu d'arrestation *Belgium*.

Dans les résultats obtenus, nous remarquons que pour le concept « ***transfert*** », nous avons obtenu seulement un transfert pour une seule personne, à savoir Malika El-Aroud. La deuxième personne, en occurrence « Moez Garsallaoui », a été exprimé par l'anaphore, et c'est pour cette raison que cette personne n'a pas été extraite dans le concept *transfert*. Cette omission peut être expliquée par l'absence de traitement des anaphores à ce stade de notre analyse.

Voici la représentation des connaissances extraites, dans notre outil de visualisation :

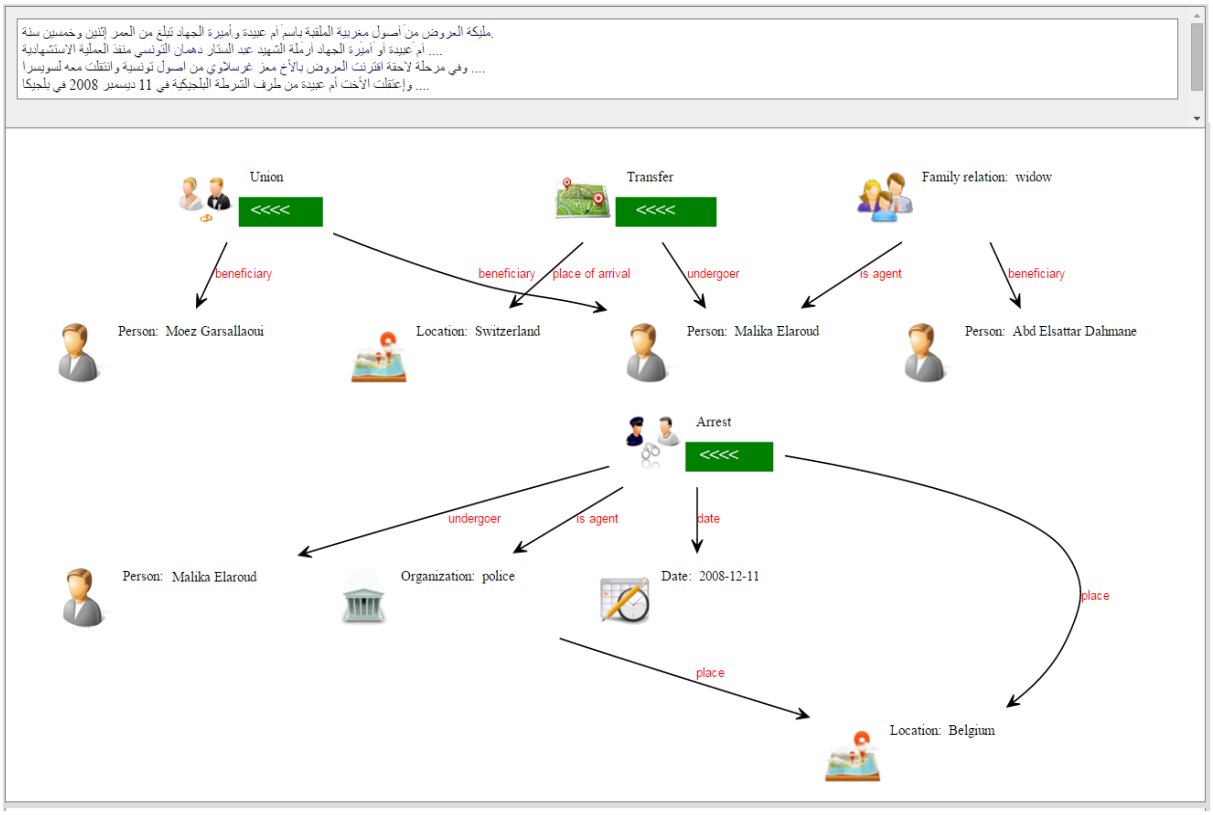


Figure 12. 4. Connaissances extraites du texte de Malika El-Aroud.

Nous notons que les résultats extraits sont présentés par le graphe de la figure 12.4. Les éléments textuels de ce graphe peuvent être transcris en arabe, français ou en anglais, selon le paramétrage de la langue lors de la génération de ce graphe. Cela nous permet par conséquent d'avoir une visualisation multilingue.

12.6. Conclusion

Nous avons décrit dans ce chapitre notre système d'extraction des connaissances dans des textes arabes, basé d'une part sur une analyse linguistique profonde, et d'autre part sur une extraction sémantique utilisant une ontologie du domaine. L'évaluation effectuée sur ces premiers travaux nous a permis de déceler globalement la qualité de notre analyse linguistique et l'extraction mais aussi de donner naissances à d'autres problématiques à étudier.

Chapitre 13 Système d'évaluation de la translittération des noms propres

Les dictionnaires bilingues jouent un rôle important dans les applications de Traitement Automatique des Langues (TAL) telles que la Recherche d'Information Interlingue (RII) et la Traduction Automatique (TA). La construction manuelle de ces dictionnaires est lente et coûteuse. C'est la raison pour laquelle depuis quelques années de nombreux travaux ont fait appel aux techniques d'alignement pour automatiser le processus de construction de lexiques bilingues. Ces travaux ont montré que l'alignement d'unitermes et de syntagmes nominaux à partir de corpus parallèles est une tâche relativement bien maîtrisée pour les langues à écriture latine. En revanche, l'appariement de textes parallèles n'utilisant pas la même écriture demeure une opération complexe. Ce qui a conduit plusieurs chercheurs à exploiter la transcription ou la translittération de certains mots des textes parallèles comme « points d'ancrage » pour améliorer la mise en correspondance bilingue.

Dans la perspective d'évaluer à une large échelle l'impact de l'utilisation de la translittération de noms propres sur la qualité d'un lexique bilingue français-arabe produit par l'outil d'alignement de mots intégrant notre outil de translittération, nous présentons, d'une part un outil d'alignement de mots simples et composés à partir de corpus de textes parallèles français-arabe, et d'autre part, les résultats d'évaluation de ce lexique bilingue selon deux approches : une évaluation de la qualité d'alignement à l'aide d'un alignement de référence construit manuellement et une évaluation de l'impact de cet alignement sur la qualité de traduction du système de traduction automatique statistique Moses. Les résultats obtenus montrent que la translittération améliore aussi bien la qualité de l'alignement que celle de la traduction. Ces résultats consolident ceux obtenus dans nos travaux antérieurs (Saâdane et Semmar, 2012).

La suite du chapitre est organisée comme suit : nous présentons dans la section (13.1) un état de l'art sur l'alignement de mots à partir de corpus parallèles. Nous montrons dans la section (13.2) comment cette translittération est utilisée pour améliorer les résultats d'un outil d'alignement de mots simples et composés. Nous consacrons la section (13 .3) aux expériences menées ainsi qu'à la présentation des résultats obtenus en précisant les taux d'amélioration de la qualité de l'alignement et de la traduction. La section (13.4) conclut notre étude et présente nos travaux futurs.

13.1. État de l'art sur l'alignement de mot

L'alignement de mots ou l'extraction de lexiques bilingues à partir de corpus de textes parallèles peut se décomposer conceptuellement en deux aspects: il s'agit de repérer les mots du texte source et du texte cible, puis de les mettre en correspondance.
Il existe principalement trois approches pour l'alignement de mots à partir de corpus de textes parallèles alignés phrase à phrase:

- Les approches à dominante statistique qui s'appuient sur les modèles IBM (Brown et al., 1993). L'outil d'alignement GIZA++ (Och et Ney, 2000) implémente notamment ce type d'approche. Cet outil implémente divers modèles de traduction (IBM 1, 2, 3, 4, 5 et HMM). Ces modèles utilisent l'algorithme EM (Dempster et al., 1977) pour l'apprentissage à partir de corpus bilingues. L'alignement des mots est réalisé à l'aide d'un algorithme de recherche de type Viterbi. GIZA++ est un outil efficace pour aligner les mots simples, mais il est moins performant, d'une part, lorsque les langues source et cible ont des morphologies et des structures syntaxiques différentes, et d'autre part, pour aligner les expressions multimots (Allauzen et Wisniewski, 2009) (Abdulhay, 2012).

- Les approches linguistiques qui utilisent généralement des dictionnaires bilingues déjà disponibles mais aussi les résultats de l'analyse morpho-syntaxique des phrases source et cible (Debili et Zribi, 1996) (Bisson, 2001). Les méthodes proposées par (Debili et Zribi, 1996) ainsi que (Bisson, 2001) utilisent des ressources linguistiques externes (lexiques, règles, etc.) pour apparier les mots des textes parallèles alignés au niveau de la phrase. Ces méthodes font l'hypothèse que pour que des phrases soient en correspondance de traduction, il faut que les mots qui les composent soient également en correspondance. Elles n'utilisent qu'une information interne, c'est-à-dire que toute l'information nécessaire (et en particulier les correspondances lexicales) est dérivée des textes à aligner eux-mêmes (ancrage lexical).

- Une combinaison des méthodes statistiques avec différentes sources d'information linguistique (Daille et al., 1994) (Gaussier et Langé, 1995) (Smadja et al., 1996) (Blank, 2000) (Barbu, 2004) (Ozdowska, 2004) (Ozdowska et Claveau, 2006) (Semmar et al., 2010). La méthode proposée par (Gaussier et Langé, 1995) est fondée sur des modèles statistiques pour établir les associations entre mots anglais et mots français, et ce en exploitant la propriété de dépendance entre les mots et leurs traductions respectives. La prise en compte des positions des mots dans les phrases permet de constituer un modèle de distorsion qui aide à la construction des associations. Ensuite, les structures morpho-syntaxiques représentant les séquences admissibles d'étiquettes grammaticales et de mots ont été recensées. Les correspondances et non-correspondances entre les structures anglaises et françaises sont utilisées pour élaborer les modèles statistiques permettant de retrouver les équivalences entre termes anglais et termes français. Quant à l'approche développée par (Ozdowska, 2004), elle consiste d'abord à apparier les mots à un niveau global grâce au calcul des fréquences de cooccurrence dans des phrases alignées. Ensuite, ces mots constituent les couples amorces qui servent de point de départ à la propagation des liens d'appariement à l'aide des différentes relations de dépendance identifiées par un analyseur syntaxique dans chacune des deux langues.

Contrairement à l'alignement de mots simples qui est désormais une tâche bien maîtrisée plus particulièrement pour les langues à écriture latine, l'alignement d'expressions multimots continue à susciter de nombreux travaux de recherche (Ozdowska et Claveau, 2006) (MacCartney et al., 2008) (Lefever et al., 2009) (Bouamor et al., 2012). La plupart de ces travaux commencent tout d'abord par identifier les expressions multimots dans chaque partie du corpus parallèle, ensuite, utilisent différentes approches d'alignement pour les apparier. Les approches pour l'extraction monolingue d'expressions multimots peuvent être: (1) symboliques en reposant sur des patrons morpho-syntaxtiques (Okita et al., 2010), (2) statistiques en utilisant des mesures d'association pour classer les expressions multimots candidates (Vintar et Fisier, 2008), et (3) hybrides combinant (1) et (2) (Daille, 2001) (Seretan et Wehrli, 2007). Pour identifier les correspondances entre expressions multimots dans différentes langues, plusieurs travaux font appel à des outils d'alignement de mots simples pour guider l'alignement d'expressions multimots. D'autres se basent sur des algorithmes d'apprentissage statistique. Une hypothèse largement suivie pour acquérir des expressions multimots bilingues est qu'une expression multimots dans une langue source garde la même structure syntaxique que son équivalente dans une langue cible donnée (Seretan et Wehrli, 2007) (Tufis et Ion, 2007). Or, cette hypothèse n'est pas toujours vérifiée puisque certaines expressions multimots ne se traduisent pas forcément par des expressions ayant la même structure syntaxique. Par exemple, l'expression « gestion des ressources en eau » peut se traduire par l'expression « إدارة الموارد المائية», mais cette dernière n'a pas la même structure

syntaxique que l'expression source. De même, certaines expressions ne se traduisent pas systématiquement par une expression de même longueur. C'est le cas du mot simple « informatique » qui se traduit par l'expression multimots « علم الحاسوب» (Science de l'ordinateur).

Pour les langues n'utilisant pas l'écriture latine, de nombreux travaux ont été réalisés pour aligner automatiquement les translittérations à partir de corpus de textes multilingues en vue de l'enrichissement de lexiques bilingues. Citons notamment les travaux de (Al-Onaizan et Knight, 2002) et (Sherif et Kondrak, 2007) sur l'alignement arabe-anglais, (Tao et al., 2006) sur l'utilisation de la translittération pour l'extraction d'entités nommées à partir de corpus comparables ainsi que (Shao et Ng, 2004) qui utilisent l'information apportée par les translittérations sur la base de leur prononciation. Ils combinent l'information apportée par le contexte des traductions avec l'information apportée par les translittérations entre l'anglais et le chinois. L'intérêt de ce travail réside dans le fait qu'il permet l'alignement de mots très spécifiques mais rares.

Nous décrivons, dans la section suivante, notre démarche pour extraire un lexique bilingue de mots simples et de mots composés à partir d'un corpus parallèle français-arabe aligné au niveau de la phrase.

13.2. Approche proposée pour l'alignement de mots à partir de corpus de textes parallèles français-arabe

La démarche que nous proposons pour la construction de lexiques bilingues à partir de corpus de textes parallèles, est composée des trois étapes suivantes:

- alignement de mots simples,
- alignement de mots composés se traduisant mot à mot,
- alignement d'expressions multimots.

Notre approche pour l'alignement de mots est basée, d'une part, sur un modèle linguistique utilisant un dictionnaire bilingue, les caractéristiques des cognats, les catégories grammaticales, les relations de dépendance syntaxique et les règles de reformulation pour l'alignement de mots simples et composés, et d'autre part, sur un modèle hybride combinant patrons morpho-syntaxiques et méthodes statistiques pour l'alignement d'expressions multimots. Les entrées de l'outil d'alignement, implémentant notre approche, sont les sorties normalisées d'une analyse morpho-syntaxique effectuée à l'aide de la plate-forme d'analyse linguistique LIMA (Besançon et al., 2010) sur le corpus de textes parallèles. Cette plate-forme fournit pour chaque couple de phrases source et cible :

- la liste des lemmes et des formes fléchies des mots ainsi que leur position dans la phrase,
- les catégories grammaticales des mots,
- les relations de dépendance syntaxique entre les mots et les mots composés.

Le processus de normalisation consiste à supprimer les mots vides de la liste des lemmes des mots retournés par la plate-forme LIMA. Les mots vides sont identifiés à partir de leur catégorie grammaticale (prépositions, articles, ponctuations et certains adverbes). Nous considérons les mots restants comme des mots significatifs (pleins).
Nous décrivons ci-dessous uniquement les principaux modules composant l'aligneur de mots simples et nous nous focalisations sur l'étape qui concerne l'alignement de mots utilisant la détection de cognats et d'entités nommées dans les phrases source et cible. C'est cette étape qui utilise la translittération des noms propres de l'arabe vers l'écriture latine. Les modules

d'alignement de mots composés et d'expressions multimots sont décrits respectivement dans (Semmar et al., 2010) et (Bouamor et al., 2012). L'alignement de mots simples se déroule selon les trois étapes suivantes:

- alignement utilisant le dictionnaire bilingue préexistant,
- alignement utilisant la détection de cognats et d'entités nommées dans les phrases source et cible,
- alignement utilisant les catégories grammaticales des mots des phrases source et cible.

L'alignement en utilisant uniquement le dictionnaire bilingue préexistant consiste, d'une part, à extraire les traductions des lemmes significatifs des phrases de la langue source en interrogeant le dictionnaire bilingue, et d'autre part, à rechercher la traduction dans la phrase cible et en comparant sa position avec celle du lemme à aligner. Si les positions des deux lemmes source et cible sont dans une même fenêtre de taille *n* respectivement dans les phrases source et cible, alors ils seront considérés traduction l'un de l'autre. Nous avons fixé expérimentalement la valeur de *n* à 6. Ainsi, le mot de la phrase source Mot_{source} est considéré comme traduction du mot de la phrase cible Mot_{cible} si les conditions [1] et [2] sont vérifiées :

$$Position\ (Mot_{source}) - 3 <= Position\ (Mot_{cible}) \quad [1]$$

$$Position\ (Mot_{cible}) <= Position\ (Mot_{source}) + 3 \quad [2]$$

Par exemple, dans la phrase source « Le général Garner a laissé entendre que l'occupation de l'Irak ne serait pas éternelle. » et sa traduction en langue cible « اشار الجنرال غارنر الى ان احتلال العراق لن يدوم الى الابد.», le dictionnaire bilingue français-arabe utilisé par l'outil d'alignement propose pour le lemme « occupation» les traductions وَقْفَة , مَشْغَلَة , اِسْتِمْلَاك , اِحْتِلَال , شُغْل mais seule la traduction « إِحْتِلَال» est retenue puisque ce lemme est présent dans la phrase cible et sa position (6) vérifie les conditions [1] et [2]. Nous avons constaté que malgré la couverture importante du dictionnaire français-arabe (124568 entrées), plusieurs mots de la phrase source n'ont pas de traductions présentes dans la phrase cible. Citons par exemple le mot « général » qui possède plusieurs traductions dans le dictionnaire bilingue (عَامّ, عِمَاد, قَائِدُ, جَيْشٍ, شَامِل) mais aucune n'est présente dans la phrase cible. Le cas du lemme « éternel » est particulier : aucune des traductions fournies par le dictionnaire bilingue ne correspond au lemme « أَبَد» de la phrase cible, mais parmi les traductions du lemme « éternité » dans ce dictionnaire on trouve le lemme « أَبَد». Ainsi, le dictionnaire bilingue n'a pas permis l'alignement du lemme « éternel » avec le mot « أَبَد» car l'analyseur morpho-syntaxique a produit pour le mot « éternelle » le lemme « éternel » et non pas le lemme « éternité ».

Par ailleurs, nous avons constaté aussi que beaucoup de noms arabes ne sont pas reconnus comme entités nommées par la plate-forme LIMA. Cela vient du fait que cette plateforme utilise des listes ainsi que des règles de déclencheurs pour reconnaître des entités telles que les noms de personnes, d'organisations, de lieux… mais ces listes sont limitées et plus particulièrement pour les langues peu dotées comme l'arabe. C'est pour cette raison que nous avons ajouté une étape supplémentaire à notre outil d'alignement de mots simples. Cette étape est utilisée pour permettre l'appariement des cognats présents dans les phrases source et cible. En linguistique, les cognats sont des paires de mots de langues différentes qui partagent des propriétés phonologiques, orthographiques et sémantiques. Nous pouvons étendre cette définition aux noms propres et aux expressions numériques puisqu'ils varient en général légèrement d'une langue à une autre. Plusieurs travaux ont montré que la détection et la mise en correspondance des cognats dans les textes source et cible permettent d'améliorer les résultats d'alignement au niveau des phrases (Simard et al., 1993) mais aussi des mots

(Bisson, 2001); (Al-Onaizan et Knight, 2002) ; (Kondrak, 2005). Récemment, (Frunza et Inkpen, 2009) ont évalué une méthode qui utilise 13 mesures de similarité orthographique pour identifier les cognats et les « faux amis ». Nous considérons dans une première étape comme cognats les mots dont les quatre premiers caractères sont identiques. Cette étape est simple à implémenter lorsque les phrases source et cible sont écrites avec le même script ou dans deux scripts proches. Dans notre étude, l'alignement de mots est réalisé à partir de corpus de textes parallèles français-arabe. Or ces deux langues sont écrites avec deux scripts différents. Pour détecter les cognats présents dans ces textes, nous avons utilisé le système de translittération décrit précédemment pour transformer les noms propres écrits en arabe vers l'écriture latine. Cette première étape a permis de détecter que les noms propres « Garner » et « Irak » et leur translittération respective en écriture latine «garnir» (du nom propre « غارنر») et « irak » (du nom propre «العراق») sont des cognats. En revanche, cette étape ne permet pas d'aligner des couples de mots comme « Algérie » et « aljezeyr » (translittération du nom propre « الجزائر»). Pour ce faire, nous avons utilisé la distance Jaro–Winkler (Winkler, 1990), une mesure de similarité basée sur le nombre de lettres en commun entre le mot de la langue source *ms* et le mot de la langue cible *mc*.

$$DJ(ms,mc) = \begin{cases} 0 & si\ m = 0 \\ \frac{1}{3}\left(\frac{m}{|ms|} + \frac{m}{|mc|} + \frac{m-t}{m}\right) & sinon \end{cases}$$

Où:

- *m* est le nombre de caractères correspondants. Deux caractères identiques des mots ms et mc sont considérés comme correspondants si leur éloignement (la différence entre leurs positions dans leurs chaînes respectives) ne dépasse pas :

$$\left(\frac{max(|ms|,|mc|)}{2}\right) - 1$$

- *t* est le nombre de transpositions. Ce nombre est obtenu en comparant le ième caractère correspondant du mot ms avec le ième caractère correspondant du mot mc. Le nombre de fois où ces caractères sont différents, divisé par deux, donne le nombre de transpositions.
- $|ms|$, $|mc|$ correspondent aux longueurs en nombre de caractères des mots *ms* et *mc*.

La mesure de similarité Jaro–Winkler est une variante de la distance Jaro *DJ* (Jaro, 1989).

$$DJW(ms,mc) = DJ(ms,mc) + (lp(1 - DJ(ms,mc)))$$

Où:

- *l* est la longueur du préfixe commun.
- *p* est un coefficient qui permet de favoriser les chaînes avec un préfixe commun.

Pour fixer les valeurs de *l* et *p* ainsi que le seuil pour lequel deux mots sont considérés comme cognats, nous avons utilisé un échantillon de 100 noms propres arabes translittérés en écriture latine. Dans cet échantillon, un nombre propre écrit en arabe peut avoir en moyenne 37 translittérations en écriture latine mais il existe des noms propres qui peuvent dépasser les 1000 translittérations comme c'est le cas du mot « الجزائر» (Algérie) qui en a 1120. Nous avons constaté que les valeurs de *l* et *p* qui permettent d'accepter le plus grand nombre de translittérations pour un nom propre sont respectivement 2 et 0,1 pour un seuil de cognats égal à 0,9. Ces paramètres fixés empiriquement permettent certes d'identifier comme cognats le

mot «Algérie» et la translittération «aljezeyr» mais génèrent aussi des erreurs puisque cet aligneur considère par exemple que les mots «mohamed» et la translittération «mahmoud» du nom propre arabe «محمود» sont des cognats. Pour réduire ce type d'erreurs, nous vérifions les conditions [1] et [2] relatives aux positions des mots respectivement dans les phrases source et cible.

Certes, la détection de cognats améliore significativement les résultats de l'alignement mais ça concerne uniquement les corpus de textes ayant une forte présence de noms propres. Pour détecter de nouvelles correspondances, nous prenons en compte les paires de mots des langues source et cible qui ont les mêmes catégories grammaticales et dont les positions vérifient les conditions [1] et [2] décrites précédemment. Cette étape est particulièrement performante pour identifier les traductions des mots entourés par des mots déjà traduits. Par exemple, le lemme « général » a la catégorie grammaticale « Nom Commun », il est précédé par le lemme « Garner » déjà aligné au lemme « غارنر» en utilisant la détection de cognats. De même, le lemme « جِنِرَال» a la catégorie grammaticale « Nom Commun » et il précède le lemme « غارنر». De plus, les positions des lemmes « général » et « جِنِرَال» vérifient les conditions [1] et [2]. Par conséquent, le lemme « جِنِرَال» sera considéré comme un alignement du lemme « général ». Le même procédé est utilisé pour aligner le lemme « laisser » avec le lemme «أَشَارَ».

Le tableau ci-dessous (Table 2) présente le résultat de l'alignement de mots simples et de mots composés se traduisant mot à mot de la phrase source « Le général Garner a laissé entendre que l'occupation de l'Irak ne serait pas éternelle. » et de sa traduction en langue cible « اشار الجنرال غارنر الى ان احتلال العراق لن يدوم الى الابد ».

Lemmes des mots de la phrase en langue source	**Lemmes des mots de la phrase en langue cible**	**Etape d'alignement utilisée**
Général	جِنِرَال	Appariement de catégories grammaticales
Garner	غارنر	Appariement de cognats
Laisser	أَشَارَ	Appariement de catégories grammaticales
Occupation	إِحْتِلاَل	Dictionnaire bilingue
Irak	العِرَاق	Appariement de cognats
général_garner	جِنِرَال_غارنر	Mise en correspondance de mots composés
occupation_Irak	العِرَاقإِحْتِلاَل_	Mise en correspondance de mots composés

Tableau 13. 1. Résultat de l'alignement de mots simples et composés.

Ce tableau montre, d'une part, que les lemmes « entendre », « être » et « éternel » de la phrase source n'ont pas été alignés, et d'autre part, que l'alignement du lemme « laisser » n'est pas correct. En vérifiant dans le dictionnaire bilingue, nous avons trouvé plusieurs

traductions pour ces lemmes, mais ils n'ont pas été alignés car ces traductions ne sont pas présentes dans la phrase cible. Cet exemple montre bien l'intérêt des alignements n:m (dans notre exemple il s'agit d'un alignement 2:1 pour le lemme « laisser entendre » qui aurait dû être aligné avec le lemme «أَشَارَ») même s'ils ne sont pas aussi fréquents que les alignements 1:1. Notons que le lexique bilingue construit à l'issue du processus d'alignement de mots contient les alignements corrects et incorrects, mais, les lemmes qui n'ont pas été alignés ne seront pas pris en compte. Les symboles « _ » séparant les lemmes des mots composés seront remplacés par des espaces.

13.3. Résultats expérimentaux et discussion

Pour illustrer l'apport de la translittération sur la qualité du lexique bilingue produit par l'alignement de mots simples et composés, nous avons évalué les résultats de l'alignement selon deux approches différentes :

- une évaluation manuelle comparant les résultats de notre aligneur de mots par rapport à un alignement de référence,
- une évaluation automatique en intégrant les résultats de notre aligneur de mots dans le corpus d'apprentissage du modèle de traduction du système de traduction statistique libre Moses[28] (Koehn et al., 2007).

L'évaluation manuelle de l'aligneur de mots a été réalisée sur une partie composée de 1000 phrases du corpus MD (Monde Diplomatique) français-arabe de la campagne ARCADE II (Véronis et al., 2008). Cet alignement de référence au niveau des mots simples et composés a été construit manuellement à l'aide de l'outil Yawat (Germann, 2008). Pour les métriques d'évaluation, nous avons utilisé celles du protocole défini lors de la conférence HLT/NAACL 2003 (Mihalcea et Pedersen, 2003). La table 3 résume nos résultats en termes de précision et de rappel selon que l'aligneur de mots utilise ou non l'appariement de cognats avec la translittération de noms propres arabes. Ces résultats montrent que l'utilisation de la translittération arabe permet d'augmenter aussi bien la précision que le rappel et confirment les résultats que nous avons obtenus précédemment sur un petit corpus de 283 phrases (Saâdane et Semmar, 2012) ainsi que ceux de (Kondrak et al., 2003) qui ont pu réduire de 10% le taux d'erreurs de l'alignement de mots en utilisant l'appariement de cognats.

Le lexique bilingue extrait à partir des 1000 paires de phrases en utilisant notre outil d'alignement de mots contient 16291 entrées dont 2023 noms propres. L'analyse de ce lexique montre qu'il contient un nombre important de doublons plus particulièrement pour les noms propres mais aussi quelques traductions de mots polysémiques. En outre, environ 53% des mots alignés se trouvaient dans le dictionnaire bilingue et 12% ont été alignés à l'aide du module d'appariement de cognats qui utilise la translittération.

Alignement de mots	**Précision**	**Rappel**	**F-Mesure**
sans l'appariement de cognats (sans translittération)	0,82	0,86	0,83
avec l'appariement de cognats (avec translittération)	0,87	0,88	0,87

Tableau 13. 2. Résultats de l'évaluation de l'alignement de mots.

[28] http://www.statmt.org/moses.

L'évaluation automatique de notre aligneur de mots a été réalisée en utilisant le corpus OPUS[29] (Tiedemann, 2009) pour la paire de langues français-arabe. Ce corpus regroupe 74067 paires de phrases parallèles extraites des résolutions des Nations Unies. Ces résolutions citent certains noms de dirigeants, et beaucoup de noms de pays et d'organisations. Nous avons divisé ce corpus en trois parties : 70067 paires de phrases pour l'apprentissage du modèle de traduction, 3500 paires de phrases pour la construction du lexique bilingue en utilisant notre aligneur de mots et 500 paires de phrases pour l'évaluation du système de traduction Moses. Pour estimer le modèle de traduction du système de référence, nous avons construit un corpus d'apprentissage contenant 70067 paires de phrases auquel nous avons ajouté les 3500 paires de phrases utilisées pour l'alignement de mots.

Pour étudier l'impact du lexique bilingue produit par l'outil d'alignement de mots intégrant la translittération sur le modèle de traduction du système Moses, nous avons ajouté ce lexique bilingue construit à partir des 3500 paires de phrases au corpus d'apprentissage. Le modèle de traduction utilisé est appris sur les lemmes des mots composant le corpus parallèle d'apprentissage et les lemmes des mots produits par notre aligneur. Nous avons aussi entraîné un modèle de langue (tri-grammes) sur la totalité du corpus OPUS en langue arabe (74067 phrases) en utilisant la boite à outils IRSTLM[30]. Deux types de corpus de test ont été utilisés pour mener nos expérimentations : *Tout-Corpus-Test* et *Noms-propres-Corpus-Test*. Le premier corpus de test *Tout-Corpus-Test* est constitué de 500 paires de phrases parallèles extraites aléatoirement du corpus OPUS. Pour mesurer l'apport réel du lexique bilingue des noms propres translittérés, nous avons constitué un corpus de test noté *Noms-propres-Corpus-Test* où nous ne conservons que les phrases du corpus *Tout-Corpus-Test* contenant au moins un nom propre. Ce corpus contient 173 paires de phrases parallèles. La qualité de traduction du système de référence (celui qui n'intègre pas les translittérations) ainsi que celui intégrant les translittérations est évaluée sur les deux corpus de test sur la base de la métrique BLEU (Papineni et al., 2002). Nous avons préféré utiliser la métrique BLEU car elle est la plus appropriée pour évaluer les systèmes de traduction statistique à base de séquences (n-grammes) tels que Moses. Nous avons considéré qu'à chaque phrase source correspond une seule phrase de référence en langue cible. Les résultats de traduction obtenus pour les deux configurations sont regroupés dans la table 13.3.

Corpus d'apprentissage	**Tout-Corpus-Test**	**Noms-propres-Corpus-Test**
sans les résultats de l'appariement de cognats (sans translittération)	15,79	17,67
avec les résultats de l'appariement de cognats (avec translittération)	16,49	19.52

Tableau 13. 3. Résultats de traduction selon le score BLEU.

Tout d'abord, nous constatons que le score BLEU obtenu est satisfaisant compte tenu de la taille du corpus d'apprentissage et du modèle de traduction utilisé et qui a été estimé sur des lemmes plutôt que sur des formes de surface (Sadat et Habash, 2006). Ce score varie en fonction du type du jeu de test. Le corpus de test *Noms-propres-Corpus-Test* qui ne considère

[29] http://opus.lingfil.uu.se
[30] http://hlt.fbk.eu/en/irstlm.

que les phrases contenant des noms propres du lexique bilingue rapporte des scores BLEU plus élevés que le corpus de test *Tout-Corpus-Test* dans les deux configurations (corpus d'apprentissage sans l'ajout de translittération ou avec translittération). Les résultats obtenus montrent que l'intégration dans le corpus d'apprentissage du modèle de traduction des alignements obtenus par le module d'appariement de cognats utilisant la translittération a permis d'obtenir un gain de +0,70 points BLEU pour le corpus de test *Tout-Corpus-Test* et un gain de +1,85 pour le corpus de test *Noms-propres-Corpus-Test*. Ces résultats confirment ceux de (Huang et al., 2003) qui ont obtenu une F-Mesure de 81% pour l'alignement d'entités nommées à partir d'un corpus parallèle chinois-anglais et un gain de +0,06 en score NIST pour la traduction.

Pour évaluer la significativité statistique des résultats obtenus, nous utilisons la méthode par ré-échantillonnage par amorce décrite par (Koehn, 2004). Cette méthode estime la probabilité (p-valeur) qu'une différence mesurée entre les scores BLEU surgit par hasard et ce par la création à plusieurs reprises (10 fois) d'échantillons uniformes avec remise à partir des corpus de tests. Nous exploitons cette méthode pour comparer les deux configurations (corpus d'apprentissage sans l'ajout de translittération ou avec translittération) selon le corpus de test utilisé. Sur un intervalle de confiance (IC) de 95%, les résultats varient de non significatifs (quant $p > 0.05$) à hautement significatifs. Les p-valeurs obtenues sur les corpus de test *Tout-Corpus-Test* et *Noms-propres-Corpus-Test* sont respectivement de 0,02 et 0,01. Par conséquent, les améliorations apportées par l'utilisation de la translittération sont significatives dans les deux configurations de test.

13.4. Conclusion et travaux futurs

Nous avons décrit dans ce chapitre, d'une part, un système de translittération des noms propres de l'écriture arabe vers l'écriture latine, et d'autre part, un outil d'alignement de mots simples et composés à partir de corpus de textes parallèles français-arabe. Nous nous sommes particulièrement intéressés à l'étude de l'impact de l'utilisation de la translittération sur la qualité du lexique bilingue produit par l'outil d'alignement de mots. Pour réaliser cette étude, nous avons évalué l'outil d'alignement de mots intégrant la translittération en utilisant deux approches : une évaluation de la qualité d'alignement à l'aide d'un alignement de référence construit manuellement et une évaluation de l'impact de cet alignement sur la qualité de traduction du système de traduction automatique statistique Moses. Les résultats obtenus montrent que la translittération améliore aussi bien la qualité de l'alignement de mots que celle de la traduction.

Dans nos expérimentations sur l'outil d'alignement de mots, le modèle de traduction a été estimé sur des lemmes plutôt que sur des formes de surface qui généralement diminuent la qualité de traduction plus particulièrement pour une langue morphologiquement riche comme l'arabe. De même, les traductions du lexique bilingue produit par l'outil d'alignement de mots ne sont pas pondérées, ce qui nous prive d'intégrer ce lexique directement dans la table de traduction. Nos travaux futurs sur l'alignement de mots s'orientent, d'une part, vers l'utilisation d'un modèle de génération pour produire les formes de surface adéquates à partir des résultats de traduction présentés en lemmes dans cette étude, et d'autre part, vers une amélioration des résultats de notre outil d'alignement en lui intégrant l'appariement d'expressions multimots et en pondérant les traductions du lexique bilingue qu'il produit.

Par ailleurs, nos expérimentations sur le système de translittération nous ont montré que les corpus étudiés contenaient aussi des noms propres latins et que la précision de l'alignement de mots est très élevée lorsque des noms propres arabes sont présents dans les

phrases source et cible. Nos travaux futurs en translittération s'orientent vers une prise en compte plus large des noms propres latins et une translittération géolocalisée qui permet d'avoir des indications sur l'origine et/ou le profil de celui qui les utilise (francophone ou anglophone, du Maghreb ou du Macherek, du nord ou du sud...).

Conclusion générale

Dans ce dernier chapitre nous donnons un résumé succinct de nos contributions, ainsi qu'une description des perspectives et des évolutions du travail développé dans la présente thèse que nous envisageons.

Contributions

Dans notre travail nous avons mis l'accent sur l'étude de l'arabe standard et ses variantes dialectales. Cette étude vise à identifier les écarts entre l'arabe standard et ses dialectes, ainsi que la mise en évidence du caractère discriminant de certains de ces écarts pour une population localisée géographiquement. Ces travaux sont faits dans la perspective d'avoir des outils de traitement automatique de la langue, ce qui nous a contraints à effectuer entre autre une collection de données d'étude par des moyens automatisés ainsi que la conception et la réalisation d'un analyseur automatique.

En effet, nous avons développé un système linguistique de la langue arabe et de ses dialectes. Plus précisément, le système élaboré contient deux modules : analyseur morphosyntaxique et un détecteur des entités nommées.

- *Analyseur morphosyntaxique :* cet analyseur effectue une analyse en trois étapes : la tokenisation, l'analyse morphologique et l'analyse syntaxique. L'analyse morphologique permet de réaliser la segmentation des formes agglutinées, la désambiguïsation pour l'attribution des catégories grammaticales et la transformation morphologique. Quant à l'analyse syntaxique, cette dernière permet d'identifier les relations syntaxiques dans les syntagmes nominaux et verbaux.

- *Détecteur d'entités des entités nommées :* d'après les évaluations de (Souhir, 2013) et celles de (Habash, 2008), entre 25 % et 40 % des mots hors vocabulaires sont des entités nommées. Nous avons donc complété notre analyse linguistique profonde par le développement d'un système de détection et de typage des entités nommées en arabe basé sur l'écriture de règle d'identification dans des transducteurs à états finis. Nous avons livré dans cette partie une étude complète du traitement des entités nommées de l'arabe, depuis la détection de ces entités jusqu'à leur traduction ou leur translittération. Il convient de souligner que l'intégration du module de détection des entités nommées réduit d'une manière significative le taux de mots hors vocabulaire.

Ces développements nous ont permis d'observer que la plus part des mots en arabes ont une structure complexe traduite par l'ajout des clitiques aux formes fléchies. Nous avons remarqué que souvent, cette agglutination s'accompagne d'une altération de la forme initiale du mot qui entraîne l'application d'un ensemble de contraintes lexicales.

Par ailleurs, nous avons effectué une collecte de données afin de construire des lexiques et des corpus de ressources linguistiques pour l'arabe standard et les dialectes. Ces constructions sont une réponse au manque de ressources dont souffre le TAL notamment pour les dialectes. Nous avons effectués dans cette partie les réalisations suivantes :

- *Construction des lexiques dialectaux* : nous avons opté pour une construction des ressources textuelles des dialectes basée sur les ressources de l'arabe standard disponibles. Nous avons à cet effet étudié les écarts (lexicaux et morphologiques) ce qui nous a permis d'identifier les ressources MSA qui peuvent être réutilisées pour les dialectes et de déterminer celles qui sont propres aux dialectes. En se basant sur cette étude, les lexiques obtenus ont été construit selon les deux méthodes suivantes :
    - ***Exploitation des ressources du MSA :*** à partir des ressources MSA, des lemmes dialectaux sont déduits en transformant et convertissant des lemmes du MSA en leurs correspondants dans les quatre dialectes étudiés : algérien, tunisien, égyptien et marocain.
    - ***Transcription des mots écrits en caractères latins vers l'écriture arabe :*** cette deuxième technique s'appuie sur la translittération en écriture arabe des lemmes dialectaux transcrits en écriture latine contenus dans différents dictionnaires.

- *Développement d'une convention d'écriture nommée CODA* : en réponse de l'absence de convention de transcription des dialectes, nous avons développé la convention d'écriture CODA pour le dialecte algérien. Cette convention proposée est complémentaire pour les autres conventions développées selon la même approche pour les dialectes tunisien et égyptien.

- *Translittération des noms arabes en écriture latine et inversement* : nous avons travaillé sur une nouvelle approche de traitement des noms propres hors-vocabulaire en réponse à l'incompatibilité des stratégies conventionnelles de traitement de ce type de noms en raison de la différence de l'alphabet des langues source et cible. Nous avons développé un outil de translittération automatique des noms arabes voyellés et non voyellés vers les différentes transcriptions possibles en écriture latine (essentiellement vers le français et l'anglais) et inversement. Cet outil implémente notre méthode de transcription fondée sur les automates d'états finis pondérés de type transducteur.

- *Constitution des corpus dialectaux* : nous avons développé des corpus dialectaux rédigés à la fois en écritures arabe et latine. La construction de ces corpus a été faite à travers l'exploitation des commentaires des utilisateurs des contenus web comme les vidéos, les journaux et les échanges dans les réseaux sociaux. Les corpus élaborés concernent les dialectes algérien, tunisien, marocain et égyptien. Nous soulignons à ce niveau que la construction de ces corpus nous a permis d'identifier des pratiques langagières de l'arabe standard, fortement teintées de dialecte local ou mixées avec une langue étrangère comme le français ou l'anglais ou encore directement transcrits en lettres latines.

- *Elaboration d'une approche d'annotation des corpus et identification des dialectes* : nous avons effectué une annotation, appelée aussi *labélisation*, des corpus développés à l'aide d'abord d'un traitement automatique du MSA et des dialectes via des analyseurs linguistiques développés par *GEOLSemantics*. Ensuite ces annotations ont été corrigées et validées manuellement. Ces mêmes opérations ont été aussi appliquées sur les corpus rédigés en écriture latine en utilisant des analyseurs linguistiques du français et de l'anglais. En plus de la labélisation, nous avons élaboré des méthodes pour l'identification de l'origine dialectale ainsi que la séparation des segments dans

différentes langues au sein d'un même texte. Ces méthodes s'appuient sur des techniques linguistiques et ont montré des résultats probants.

- *Développement d'une interface d'annotation* : nous avons développé une interface permettant de visualiser les résultats d'analyse linguistique, et de permettre d'annoter manuellement des mots hors vocabulaire dans la perspective d'enrichir les dictionnaires initiaux. Cette interface permet aussi d'afficher l'origine dialectale d'un texte ou d'un corpus annotés.

Enfin, nous avons effectué un ensemble d'expérimentations afin d'évaluer l'ensemble des modules réalisés et des ressources élaborées. Une partie de ses expérimentations concernent l'analyseur linguistique du MSA. A cet effet, nous avons mené deux types d'évaluation afin de mesurer l'efficacité de notre système : une évaluation quantitative et une évaluation qualitative. Nous rappelons à ce stade que notre analyseur a été intégré dans une plateforme d'extraction de connaissances de GEOLSemantics. En ce qui concerne l'évaluation quantitative, elle concerne principalement i) l'outil de segmentation via la comparaison de ses résultats avec ceux du segmenteur de Stanford, et ii) le module d'extraction des entités nommées. Quant à l'évaluation qualitative, elle concerne l'estimation de la performance de nos règles d'extraction de connaissances dans le domaine de la sécurité. Ces expérimentations ont montré que notre outil de segmentation possède des performances meilleures que celle de Stanford notamment au niveau de la rapidité de segmentation avec un taux de traitement de 100% des entrées lexicales. Pour le module d'extraction des entités nommées, nous signalons que les évaluations nous ont conduits à observer que la performance du système varie selon le type d'entité nommée à identifier. De ce fait, pour certaines entités nous avons obtenu des performances intéressantes, par exemple nous avons un F-mesure de 85,12% pour les lieus, contrairement à d'autres entités où les performances sont moyennes.

Nous avons aussi effectué des évaluations afin d'illustrer l'apport de l'utilisation de la translittération sur la qualité du lexique bilingue. Nous nous sommes intéressés au lexique français-arabe produit par l'alignement de mots simples et composés intégrant la translittération, où nous avons évalué les résultats de l'alignement selon deux approches différentes :

- Une évaluation manuelle comparant les résultats de notre aligneur de mots par rapport à un alignement de référence,
- Une évaluation automatique en intégrant les résultats de notre aligneur de mots dans le corpus d'apprentissage du modèle de traduction du système de traduction statistique libre Moses.

A travers ces évaluations, nous avons montré comment cette translittération est utilisée pour améliorer les résultats d'un outil d'alignement de mots simples et composés. Les résultats obtenus précisent les taux d'amélioration de la qualité de l'alignement et de la traduction intégrant la translittération.

Perspectives

Les travaux effectués dans cette thèse ainsi que les résultats obtenus nous ont permis d'entreprendre de nouvelles pistes et perspectives de recherche dans le domaine de TAL et du traitement des dialectes arabes. Ces nouvelles ouvertures peuvent être résumées dans les points suivants :

- *Extraction des entités nommées :* à ce niveau nous envisageons de travailler sur l'amélioration du critère de performance *rappel*. De plus, nous travaillerons aussi sur l'aspect technique de l'approche proposée afin de résoudre d'avantage de problèmes

liées surtout à l'ambiguïté caractérisant la langue arabe et ses dialectes. Nous envisageons aussi enrichir les règles d'extraction afin de prendre en considération les contextes particuliers non considérés dans la version actuelle. Ces règles devront intégrer entre autre les spécificités régionales de la langue arabe ainsi que l'utilisation d'autres méthodes statistiques ou hybrides pour l'analyse des ENs.

- *Analyse syntaxique :* les améliorations concernant l'analyse syntaxique sont orientées vers l'ajout de la recherche des antécédents des anaphores présentes dans les textes. En effet, si les pronoms, utilisés fréquemment dans les textes pour éviter les répétitions, ne sont pas liés à l'entité à laquelle ils font référence, nous risquons de perdre beaucoup des informations présentes dans les textes. Il faut noter que les limites des systèmes linguistiques et statistiques actuels nous orientent vers une future combinaison de ces approches pour une meilleure extraction.

- *Extraction des connaissances :* l'objectif est d'améliorer les performances du système actuel en travaillons d'avantage sur certaines problématiques TAL auxquelles nous nous sommes confrontés, comme la gestion des modalités. Cette évolution passe par la création des corpus annotés pour une évaluation à grande échelle, tout en étendant notre système à d'autres domaines comme l'économie, le sport, etc. De plus, au niveau de l'évaluation nous projetons d'enrichir le corpus de tests afin de proposer une campagne d'évaluation. Ainsi, nous encouragerons le développement des systèmes d'extraction des connaissances en arabe. Ceci nous permettra de mettre en place une évaluation quantitative sur un jeu de données conséquent.

- *La translittération :* nous envisageons d'orienter nos recherches vers la translittération géolocalisée pour répondre à la question de savoir comment les différentes translittérations peuvent fournir des indications sur l'origine et/ou sur le profil de celui qui les utilise (francophone ou anglophone, du Maghreb ou du Macherek, du nord ou du sud...). Cette orientation répond à la fois à une demande industrielle urgente et à une problématique de recherche intéressante. La prochaine étape consiste à étendre le système de translittération aux noms d'origine non arabe.

- *Création des corpus parallèles Arabizi-Arabe* : nous envisageons décrire et développer le processus de création d'une ressource écrite en latin afin de créer des corpus parallèles : arabe dialectal translittéré en latin vers l'arabe dialectal translittéré en caractères arabes. Le langage utilisé dans les médias sociaux expose plusieurs particularités comme la non formalisation de son vocabulaire, les déviations intentionnelles de l'orthographe standard telles que la répétition des lettres pour l'accentuation, les erreurs de frappe et les abréviations non standard ; et les contenus non linguistiques sont écrits, tels que les rires, les sons, et les émoticônes. Nous pensons que ce corpus serait utile pour les travaux du TALN sur les dialectes arabes et les genres informels. Il est aussi intéressant d'explorer l'usage de ce composant dans des applications spécifiques telles que la traduction automatique de l'Arabizi en anglais ou en français, et l'analyse des émotions dans les médias sociaux.

- *Création des lexiques et corpus dialectaux :* à ce stade les améliorations que nous visons concernent l'enrichissement des bases lexicales afin de compléter les dictionnaires proposés, contenant des mots propres à chaque dialecte, et d'aider à l'amélioration de l'identification de l'origine dialectale des locuteurs ou internautes. Il

sera aussi intéressant d'étudier d'autres dialectes du Moyen Orient en utilisant et adaptant les outils de flexion que nous avons proposés pour ces types des dialectes. En ce qui concerne les corpus, il est aussi nécessaire d'étendre sa couverture à d'autres dialectes comme ceux du moyen orient. Cette extension devra introduire de nouvelles spécificités des autres dialectes, non considérés actuellement, et de prendre en charge les forts contrastes entre les différents dialectes comme c'est le cas entre le syrien et le libanais.

- *Les variations syntaxiques dialectales :* afin de pouvoir étudier les variations de la syntaxe, il sera également nécessaire d'introduire, dans un dictionnaire supplémentaire, les termes n'appartenant pas à l'arabe moderne et qui ont une certaine fréquence d'usage dans le corpus. Cela permettra de constituer des corpus étiquetés par catégorie grammaticale, ouvrant ainsi la voie à une analyse de la syntaxe par des moyens automatiques. Cette étude essayera de répondre aux questions suivantes :
    - Comment traiter l'écart existant avec l'arabe standard moderne, en particulier lorsqu'il ne relève pas seulement du lexique mais d'une distorsion de la syntaxe ?
    - Comment expliquer et traiter les distorsions de la syntaxe ?
    - Est-ce seulement dû au style de simplification introduit par Internet et par l'usage de langages SMS ou bien s'agit d'une influence plus profonde du dialecte local ?

- *Extraction des traits les plus discriminants :* l'étude vise à permettre la réalisation d'analyses automatiques complètes des textes intégrant ces diverses variétés d'arabe. En effet, une fois les divers écarts par rapport à l'arabe standard identifiés et normalisés, on mettra en place des méthodologies statistiques pour faire ressortir les traits les plus discriminants. Grâce à des outils de catégorisation de type SVM (Support Vector Machine), on pourra réaliser, sur notre corpus, un apprentissage des origines géographiques des textes qui permettra de déterminer cette origine sur un nouveau texte présenté et dont la machine ignore l'origine.

- *Détection de l'arabe écrit en latin (Arabizi)* : pour améliorer la classification d'un mot en tant qu'Arabizi, anglais ou français qui se fait suivant le contexte, nous envisageons d'employer d'autres techniques pour l'étiquetage des séquences en utilisant par exemple les CRF (champs markoviens conditionnels). Cet étiquetage permettra de détecter l'Arabizi dans son contexte. Le CRF sera formé en utilisant les fonctions au niveau des mots et des séquences.

- *Reconnaissance des dialectes :* des efforts supplémentaires sont aussi à investir afin d'extraire les traits dialectaux. Comme la source de nos corpus sont issus de journaux et de contenus web, il est logique de s'attendre à ce que les différences dans les n-grammes soient dues uniquement à la couverture d'actualités locales de chaque journal et non aux différences inhérentes dans les dialectes. Une solution possible à ce problème consiste à chercher des commentaires dans plusieurs sites dans le même pays et de comparer la précision obtenue. Cependant, il est financièrement impossible d'annoter toutes les données tirées des forums en ligne, néanmoins, il peut être possible d'améliorer les performances grâce à l'utilisation de techniques semi-supervisées.

Bibliographie

Abdulhay Authoul. (2012). *Constitution d'une ressource sémantique arabe à partir d'un corpus multilingue aligné*. Thèse de Doctorat de l'Université Stendhal – Grenoble III.

Abbas-Mekki Wijdan. (1998). *Définition et description des unités linguistiques intervenant dans l'indexation automatique des textes en arabe*. Doctoral dissertation, Thèse en Science de l'Information et de la Communication. Lyon: Université Lumière Lyon 2.

Abbès Ramzi et Dichy Joseph. (2008). Extraction automatique de fréquences lexicales en arabe et analyse d'un corpus journalistique avec le logiciel AraConc et la base de connaissances DIINAR.1. *Serge Heiden & Bénédicte Pincemain, Proceedings of JADT*, 12-14.

Abbès Ramzi. (2002). AraFreq: un outil pour le calcul de fréquences de mots arabes, in A. Braham (ed.), *Proceedings of the International Symposium on The Processing of Arabic* (Avril 18-20, 2002), Université de la Manouba, Tunis.

Abbès Ramzi. (2004). La conception et la réalisation d'un concordancier électronique pour l'arabe. Thèse de doctorat en sciences de l'information, Lyon, ENSSIB/INSA.

Abdallah Sherief, Shaalan Khaled, et Shoaib Muhammad. (2012). Integrating Rule-based System with Classification for Arabic Named Entity Recognition. In *Computational Linguistics and Intelligent Text Processing* (pp. 311-322). Springer Berlin Heidelberg.

Abdel-Malek Zaki N. (1972). *The Closed-List Classes of Colloquial Egyptian Arabic*. The Hague: Mouton.

Abdellatif Karim. (2010). Dictionnaire « le Karmous » du Tunisien. Version 1.

AbdelRahman Samir, Elarnaoty Mohamed, Magdy Marwa et Fahmy Aly. (2010). Integrated Machine Learning Techniques for Arabic Named Entity Recognition. *International Journal of Computer Science Issues (IJCSI)*, Vol. 7, Issue 4, No 3, pages 27-36.

Abdul-Hamid Ahmed, et Darwish Kareem. (2010, July). Simplified feature set for Arabic named entity recognition. In *Proceedings of the 2010 Named Entities Workshop* (pp. 110-115). Association for Computational Linguistics.

Abduljaleel Nasreen, Larkey Leah S. (2003). Statistical transliteration for English-Arabic Cross Language Information Retrieval. In *Proceedings of the Twelfth ACM International Conference on Information and Knowledge Management* (pp. 139-146), New Orleans, Louisiana, 2003.

Aboul-Fetouh Hilmi Mohamed. (1969). *A Morphological Study of Colloquial Egyptian Arabic*. (Vol. 33). Walter de Gruyter GmbH & Co. KG.

Abu Al-Chay Najim. (1988), Un Système expert pour l'analyse et la production des verbes

arabes dans une perspective d'Enseignement Assisté par Ordinateur. Thèse de doctorat.

Al-Ghamdi Mansour. (2005). Alghorithms for Romanizing Arabic names. In *Journal of King Saud University-Computer and Information Sciences*, 17, 105-128.

Al-Ghulayaini Mustafa. (2010). جامع الدروس العربية *jAmç Aldrws Alçrbyh, Part II.* IslamKotob.

Al-Onaizan Yaser, Knight Kevin. (2002). Translating named entities using monolingual and bilingual resources. In *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL'02), pages 400–408*. Philadelphia. Association for Computational Linguistics.

Al-Toma Salih J. (1969). *The Problem of Diglossia in Arabic: A Comparative Study of Classical and Iraqi Arabic* (Vol. 21). Cambridge, Mass: Harvard University Press.

Allauzen Alexandre, Wisniewski Guillaume. (2009). Modèles discriminants pour l'alignement mot à mot. TAL Volume 50 – n° 3/2009, pages 173 – 203.

Alsalman Abdulmalik, Alghamdi Mansour, Alhuqayl Khalid, Alsubai Salih . (2007). Romanization System for Arabic Names. In *Proceedings of The First International Symposium on Computer and Arabic Language (ISCAL – 07)*, Riyadh, pp. 214-227.

Arbaoui Nora. (2010). *Les dix formes de l'arabe classique à l'interface Phonologie/Syntaxe—Pour une déconstruction du gabarit* (Doctoral dissertation, Doctoral Dissertation, Université Paris-Diderot).

Arezki Abdenour. (2008). *Le rôle et la place du français dans le système éducatif algérien*. Revue du Réseau des Observatoires du Français Contemporain en Afrique, (23), 21-31.

Attia Mohamed. (2000). A large-scale computational processor of the Arabic morphology. *A Master's Thesis, Cairo University,(Egypt)*.

Attia Mohammed. (2006). An ambiguity-controlled morphological analyzer for modern standard arabic modelling finite state networks. In *Challenges of Arabic for NLP/MT Conference, The British Computer Society, London, UK* (Vol. 200610, No. 1.72).

Attia Mohammed. (2008). Handling Arabic morphological and syntactic ambiguities within the LFG framework with a view to machine translation. Thèse de doctorat. University of Manchester.

Awni Nahid. (1999). Egyptian colloquial/ آنستونا : العامية المصرية. Volume 1, Cairo.

Baccouche Taïeb. (1994). *L'emprunt en arabe moderne*. Académie tunisienne des sciences, des lettres, et des arts, Beït al-Hikma.

Baccouche Taïeb. (1998). La langue arabe dans le monde arabe. *L'Information Grammaticale*, *2*(1), 49-54.

Badawi, Al-Saeed Muhammad. (1973). *Mustawayaatu al-'arabiyya a'-mu'aasira fi misr*. [

Levels of Contemporary Arabic in Egypt]. Cairo, Egypt: Dar al-ma'aarif.

Badawi, Al-Saeed Muhammad. (1985). Educated Spoken Arabic: A Problem in Teaching Arabic as a Foreign Language. *Scientific and Humanistic Dimensions of Language: Festschrift for Robert Lado*. ed. by Kurt R. Jankowsky. Amsterdam: Benjamins. 15-22.

Bahou Younès, Belguith Hadrich Lamia, Aloulou Chafik, et Ben Hamadou Abdelmajid. (2006). Adaptation et implémentation des grammaires HPSG pour l'analyse de textes arabes non voyellés. *Actes du 15e congrès francophone AFRIF-AFIA Reconnaissance des Formes et Intelligence Artificielle RFIA'2006*, *25*, 26-27.

Barbu Ana-Maria . (2004). Simple linguistic methods for improving a word alignment algorithm. In *Proceedings of 7th International Conference on the Statistical Analysis of Textual Data* (pp. 88-98).

Barkat Melissa. (2000). Détermination d'indices acoustiques robustes pour l'identification automatique des parlers arabes. *De la caractérisation…… à l'identification des langues.* Thèse de Doctorat Université Lumière Lyon 2.

Barkat-Defradas Melissa, Hamdi Rim, Pellegrino François. (2004). De la caractérisation linguistique à l'identification automatique des dialectes arabes. *In Proceedings of MIDL 2004*, 51-56.

Beesley Kenneth. R, Karttunen Lauri. (2003). Finite State Morphology. Stanford, CA: CSLI Publications (distributed by the University of Chicago Press), 2003. xviii + 505pp. and CD-ROM. ISBN hardbound 1-57586-433-9, paperbound 1-57586-434-7.

Benajiba Yassine, Diab Mona, et Rosso Paolo. (2008a). Arabic Named Entity Recognition: An SVM-Based Approach. In *Proceedings of 2008 Arab International Conference on Information Technology (ACIT)* (pp. 16-18).

Benajiba Yassine, Diab Mona, et Rosso Paolo. (2008b). Arabic named entity recognition using optimized feature sets. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 284-293). Association for Computational Linguistics.

Benajiba Yassine, Diab Mona, et Rosso Paolo. (2009). Arabic named entity recognition: A feature-driven study. *Audio, Speech, and Language Processing, IEEE Transactions on*, *17*(5), 926-934.

Benajiba Yassine, Diab Mona, et Rosso Paolo. (2009b). Using Language Independent and Language Specific Features to Enhance Arabic Named Entity Recognition. In *The International. Arab Journal of Information Technolgy*, 6(5), 463-471.

Benajiba Yassine, et Rosso Paolo. (2007, December). ANERsys 2.0: Conquering the NER Task for the Arabic Language by Combining the Maximum Entropy with POS-tag Information. In *Proceedings of Workshop on Natural Language-Independent Engineering, 3rd Indian International Conference on Artificial Intelligence* : *IICAI* (pp. 1814-1823).

Benajiba Yassine, et Rosso Paolo. (2008, May). Arabic Named Entity Recognition using Conditional Random Fields. In *Proc. of Workshop on HLT & NLP within the Arabic World, LREC* (Vol. 8, pp. 143-153).

Benajiba Yassine, Rosso Paolo, et Benedíruiz, José Miguel. (2007). Anersys: An arabic named entity recognition system based on maximum entropy. In *Computational Linguistics and Intelligent Text Processing* (pp. 143-153). Springer Berlin Heidelberg.

Benajiba Yassine. (2009). *Arabic named entity recognition* (Doctoral dissertation, Universidad Politécnica de Valencia Valencia, Spain).

Benrabah Mohamed. (1999). *Langue et pouvoir en Algérie: Histoire d'un traumatisme linguistique*. Seguier Editions.

Besançon Romaric, De Chalendar Gaël, Ferret Olivier, Gara Faiza, Laïb Meriama, Mesnard Olivier, & Semmar Nasredine. (2010). LIMA: A Multilingual Framework for Linguistic Analysis and Linguistic Resources Development and Evaluation. In *Proceedings of LREC 2010*.

Biadsy, Fadi, Julia Hirschberg, and Nizar Habash. (2009). Spoken Arabic dialect identification using phonotactic modeling. In *Proceedings of the EACL Workshop on* Computational Approaches to Semitic *Languages*, pages 53–61, Athens.

Bisson Arga F. (2001). Méthodes et outils pour l'appariement de textes bilingues. Thèse de Doctorat de l'Université Paris VII.

Blachère Régis & Gaudefroy-Demombynes Maurice. (1966). *Grammaire de l'arabe classique: morphologie et syntaxe*. Maisonneuve & Larose.

Blank Ingeborg. (2000). Terminology extraction from parallel technical texts. In *Véronis J. (Ed.). Parallel Text Processing* (pp. 237-252). Springer Netherlands.

Bouamor Dhouha, Semmar Nasredine, Zweigenbaum Pierre. (2012). Identifying bilingual Multi-Word Expressions for Statistical Machine. In *Proceedings of the Eighth international conference on Language Resources and Evaluation (LREC)*, p. 674-679, Istanbul, Turkey.

Bouamor Houda, Habash, Nizar, & Oflazer Kemal . (2014). A Multidialectal Parallel Corpus of Arabic. In *Proceedings of the Language Resources and Evaluation Conference, LREC* (pp. 1240-1245).

Boujelbane Rahma, Ellouze Mariem, Béchet Frédéric, & Belguith Lamia. (2015). « De l'arabe standard vers l'arabe dialectal: projection de corpus et ressources linguistiques en vue du traitement automatique de l'oral dans les médias tunisiens ». *Traitement automatique des langues (TAL)*.

Boujnab Abderrahmane. (2011). Le Moroccan Arabic textbook. In *Peace Corps Morocco*.

Boukadida Nahed. (2008). *Connaissances phonologiques et morphologiques dérivationnelles et apprentissage de la lecture en arabe (Etude longitudinale)*(Doctoral dissertation, Université Rennes 2; Université de Tunis).

Brame Michael. (1970). *Arabic Phonology: Implications for Phonological Theory and Historical Semitic*. Doctoral dissertation, Massachusetts Institute of Technology, Cambridge,

Mass.

Broselow Ellen. (1976). The Phonology of Egyptian Arabic. Doctoral dissertation. University of Massachusetts, Amherst.

Buckwalter, Tim (2004). Buckwalter Arabic Morphological Analyzer Version 2.0. Catalog No. LDC2004L02. . University of Pennsylvania: Linguistic Data Consortium. Isbn: 1-58563-257-0.

Calvet Louis-Jean. (1987). *La guerre des langues et les politiques linguistiques*. Hachette littératures, Paris.

Cantineau Jean. (1960). *Cours de phonétique arabe:(édition originale réimprimée) suivi de Notions générales de phonétique et de phonologie*. C. Klinicksieck.

Cantineau, Jean. (1960). *Études de Linguistique arabe* (Vol. 2). Paris, Librairie C. Klincksieck.

Carter Michael. (1996). Signs of change in Egyptian Arabic. *A. Elgibali, Understanding Arabic: Essays in contemporary Arabic Linguistics in Honor of El-Said Badawi*, 137-143.

Chachou Ibtissem (2013). Langues de la publicité et publicisation des langues dans la presse algérienne d'expression arabophone. *Maghreb et sciences sociales*, (331), 179-199.
Chalabi Achraf. (2000). MT-based transparent Arabization of the internet TARJIM. COM. In *Envisioning Machine Translation in the Information Future (pp. 189-191)*. Springer Berlin Heidelberg.

Chanod, Jean-Pierre, et Tapanainen Pasi. (1996, August). A non-deterministic tokeniser for finite-state parsing. In *Proceedings of the Workshop on Extended finite state models of language (ECAI'96)*.

Cheng Xu, Dale Cameron, & Liu Jiangchuan.(2007). Understanding the characteristics of internet short video sharing: YouTube as a case study. *arXiv preprint arXiv:0707.3670*.

Chiang David, Mona Diab, Nizar Habash, Owen Rambow, and Safiullah Shareef. 2006. Parsing Arabic dialects. In *Proceedings of EACL*, pages 369–376, Trento.

Chung, Wingyan. (2008). Web searching in a multilingual world. *Communications of the ACM*, 51(5), pp. 32-40.

Cohen David. (1970). Essai d'une analyse automatique de l'arabe. *Études de linguistique sémitique et arabe*, 49-78.

Cohen David. (1973). Pour un atlas linguistique et sociolinguistique de l'arabe. In *Actes du 1er Congrès d'étude des cultures méditerranéennes d'influence arabo-berbère*, pp. 63-69, Alger, Algérie.

Cohen Marcel Samuel Raphaël. (1912). *Le parler arabe des Juifs d'Alger* .(Vol. 4). Champion :Paris.

Cotterell Ryan, Renduchintala Adithya, Saphra Naomi, & Callison-Burch Chris. (2014, May). An algerian arabic-french code-switched corpus. In *Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools Workshop Programme* (p. 34).

Crystal David. (1985). *A Dictionary of Linguistics and Phonetics*. Oxford: Basil Blackwell.

Daille Béatrice, Gaussier Éric, Lange Jean-Marc. (1994). Towards automatic extraction of monolingual and bilingual terminology. In *Proceedings of the 15th International Conference on Computational Linguistics-Volume 1 (COLING'94)*, pp. 515-521. Association for Computational Linguistics. Kyoto, Japan.

Daille Béatrice. (2001). Extraction de collocations à partir de textes. In Actes de TALN 2001 (Traitement automatique des langues naturelles), Tours.

Dalila Morsly. (1986). Multilingualism in Algeria. *The Fergusonian Impact: In Honor of Charles A. Ferguson on the Occasion of His*, 65.

Darwish Kareem. (2014). Arabizi Detection and Conversion to Arabic . In *Arabic Natural Language Processing Workshop, EMNLP*, Doha, Qatar.

Darwish. Kareem. (2002, July). Building a shallow Arabic morphological analyzer in one day. In *Proceedings of the ACL-02 workshop on Computational approaches to semitic languages* (pp. 1-8). Association for Computational Linguistics.

Debili Fathi et Achour Hadhemi. (1998, August). Voyellation automatique de l'arabe. In : *Proceedings of the Workshop on Computational Approaches to Semitic Languages* (pp. 42-49). Association for Computational Linguistics, Monréal, Canada.

Debili Fathi, & Souissi Emna. (1998, August). Étiquetage grammatical de l'arabe voyellé ou non. In *Proceedings of the Workshop on Computational Approaches to Semitic Languages* (pp. 16-25). Association for Computational Linguistics.

Debili Fathi, Achour Hadhémi, et Souissi Emna. (2002). La langue arabe et l'ordinateur de l'étiquetage gramatical à la voyellation automatique. *Correspondances: bulletin de l'IRMC*, (71), 10-26.

Debili Fathi, Zribi Adnane. (1996). Les dépendances syntaxiques au service de l'appariement des mots. *Actes du 10ème Congrès Reconnaissance des Formes et Intelligence Artificielle (RFIA'96)*.

Debili Fathi. (2001). Traitement automatique de l'arabe voyellé ou non. *Correspondances*, (46).

Dempster Arthur P., Laird Nan M., Rubin Donald B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. In *Journal of the Royal Statistical Society. Series B (Methodological),* Vol. 39, No. 1, p. 1-38.

Diab Mona, Nizar Habash, Owen Rambow, Mohamed Altantawy, and Yassine Benajiba. 2010. COLABA: Arabic dialect annotation and processing. In Proceedings of the LREC Workshop on Semitic Language Processing, pages 66–74.

Diab Mona. (2009). Second generation AMIRA tools for Arabic processing: Fast and robust tokenization, POS tagging, and base phrase chunking. In *2nd International Conference on Arabic Language Resources and Tools*.

Dichy Joseph (1984/89). « Vers un modèle d'analyse automatique du mot graphique non-vocalisé en arabe », in Dichy et Hassoun, éds., 1989, p. 92-158.

Dichy Joseph et Zmantar Yasser. (2009). L'analyse automatique des mots-outils en arabe, in Ghenima, Malek, Ouksel, Aris et Sidhom, Sahbi (eds.), *Systèmes d'Information et Intelligence Economique, 2ème Conférence Internationale (SIIE 2009)*, organisée par l'université de Nancy, France et luniversit de la Manouba, école supérieure de commerce électronique (ESCE), Tunis, Tunis, Hammamet, 12-14 fvrier 2009, IHE éditions, p. 586-597.

Dichy Joseph. (1997). Pour une lexicomatique de l'arabe: l'unité lexicale simple et l'inventaire fini des spécificateurs du domaine du mot. *Meta: Journal des traducteurs Meta:/Translators' Journal*, 42(2), pp. 291-306.

Dichy, Joseph (1990). L'écriture dans la représentation de la langue : la lettre et le mot en arabe. Thèse d'État (en linguistique), Université Lumière-Lyon 2.

Djili Abdelaziz. (2011). L'arabe une langue des défis. URL: http://www.youtube.com/watch?v=LbMQ4HVOyo4.

Eisele, John C. (1990). Time Reference, Tense and Formal Aspect in Cairene Arabic. *Perspectives on Arabic Linguistics I.* 173-214. ed. by Mushira Eid.

El Isbihani Anas, Khadivi Shahram, Bender Oliver, et Ney Hermann. (2006). Morpho-syntactic Arabic preprocessing for Arabic-to-English statistical machine translation. In *Proceedings of the Workshop on Statistical Machine Translation* (pp. 15-22). Association for Computational Linguistics.

El Kassas Dina, et Kahane Sylvain. (2004). Modélisation de l'ordre des mots en arabe standard. *Atelier sur le traitement automatique de la langue arabe, JEP-TALN-RECITAL*, 6.

El Kassas Dina. (2005). *Une étude contrastive de l'arabe et du français dans une perspective de génération multilingue* (Doctoral dissertation, Paris 7).

EL-Bèze Marc, Mérialdo Bernard, Rozeron Bénédicte, A.M. Derouault. (1994). Accentuation automatique de textes par des méthodes probabilistes. *TSI. Technique et Science informatiques*, 13(6), 797-815.

Elfardy Heba, & Diab Mona. (2012b). AIDA: Automatic identification and glossing of dialectal Arabic. In *Proceedings of the 16th EAMT Conference (Project Papers)* (pp. 83-83).

Elfardy Heba, & Diab Mona. (2012c). Simplified guidelines for the creation of large scale dialectal Arabic annotations. In *LREC* , pages 371–378.

Elfardy Heba, & Mona Diab. (2013). Sentence level dialect identification in Arabic. In *ACL (2)*. p. 456-461.

Elsebai Ali, Meziane Farid, et Belkredim, Fatma Zohra. (2009). A rule based persons names Arabic extraction system. *Communications of the IBIMA*, *11*(6), 53-59.

Embarki Mohamed. (2008). Les dialectes arabes modernes: état et nouvelles perspectives pour la classification géo-sociologique. *Arabica*, *55*(5), 583-604.

Eskander Ramy, Al-Badrashiny Mohamed, Habash Nizar, & Rambow Owen. (2014). Foreign words and the automatic processing of Arabic social media text written in Roman script. *EMNLP 2014*, 1.

Ezzahid Samia. (1996). Méthodologie d'élaboration d'une base de données lexicale de l'arabe (vocabulaire général) d'après la théorie Sens-Texte d'Igor Mel'cuk. Thèse de doctorat, Université Lyon 2.

Farghaly Ali, & Shaalan Khaled. (2009). Arabic Natural Language Processing: Challenges and Solutions. *ACM Transactions on Asian Language Information Processing (TALIP)*, 8(4), 14.

Farghaly Ali. (2010). The Arabic Language, Arabic Linguistics and Arabic Computational Linguistics. Arabic Computational Linguistics, 43-81.

Ferguson Charles. (1959a). Diglossia. *Word*, 325–340.

Ferguson Charles. (1959b). The Arabic Koine. *Language,* 616–630.

Fleish Henri. (1961). *Traité de philologie arabe. 1. Préliminaires, phonétique, morphologie nominale.* Dar el-Machreq Éds..

Fluhr Christian, Schmit Dominique, Elkateb Faïza, Ortet Philippe & Gurtner Karine. (1997). Multilingual database and crosslingual interrogation in a real internet application. In *AAAI Symposium on Cross-Language Text and Speech Retrieval. American Association for Artificial Intelligence*.

Fluhr Christian, Schmit Dominique, Ortet Philippe, Elkateb Faïza, Gurtner Karine & Radwan Khaled. (1998). Distributed cross-lingual information retrieval. In *Cross-Language Information Retrieval* (pp. 41-50). Springer US.

Fourour Nordine. (2002). Nemesis, un système de reconnaissance incrémentielle des entités nommées pour le français. In *Actes de la 9ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'02)* (pp. 265-274).

Franjié Lynne .(2003). *Étude sémantique et traductologique de verbes arabes dans les dictionnaires bilingues: le Larousse (arabe-français) et le H. Wehr (arabe-anglais).* Doctoral dissertation, Lyon 2.

Frunza Oana, & Inkpen Diana. (2010). Identification and disambiguation of cognates, false

friends, and partial cognates using machine learning techniques. *International Journal of Linguistics*, *1*(1), E2.

Gadalla Hassan. A. (2000). Comparative Morphology of Standard and Egyptian Arabic (Vol. 5). Lincom Europa.

Gahbiche-Braham Souhir, Bonneau-Maynard Hélene, Lavergne Thomas, et Yvon François. (2012). Joint Segmentation and POS Tagging for Arabic Using a CRF-based Classifier. In *LREC* (pp. 2107-2113).

Gahbiche-Braham Souhir, Bonneau-Maynard Hélène, Yvon François. (2013). Traitement automatique des entités nommées en arabe : détection et traduction. In *TAL Volume 54 – n 2/2013*, p. 101-132

Gaussier Eric, et Langé J. M. (1995). Modèles statistiques pour l'extraction de lexiques bilingues. *TAL. Traitement automatique des langues*, Volume 36(1-2). ATALA, p. 133-155.

Germann Ulrich . (2008). Yawat: yet another word alignment tool. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session* (pp. 20-23). Association for Computational Linguistics.

Ghenima Malek. (1998). Analyse morpho-syntaxique en vue de la voyellation assistée par ordinateur des textes écrits en arabe. Thèse de doctorat, ENSSIB.

Grainger Jonathan, Dichy Joseph, EL-HALFAOUI Mohamed et Bamhamed Mohamed. (2003). Approche expérimentale de la reconnaissance du mot écrit en arabe. *Faits de langue*, (22), 77-86.

Guella Noureddine. (2010). La suppléance linguistique en arabe dialectal: reflet d'une dynamique conversationnelle. *Arabica: Journal of Arabic and Islamic Studies*, 477-490.

Guella Noureddine. (2011). Emprunts lexicaux dans des dialectes arabes algériens. *Synergies Monde arabe*, *8*, 81-88.

Guidère Mathieu. (2003). *Kalimât : le vocabulaire arabe*, Paris, Ellipses.

Guidère Mathieu. (2004). Le traitement de la parole et la détection des dialectes arabes. *Langues stratégiques et défense nationale, Publications du CREC*, Saint-Cyr, pages 53–75.

Guidère Mathieu. (2005). *La traduction arabe: méthodes et applications: de la traduction à la traductique*, Paris, Ellipses.

Guidère Mathieu. (2006). Al-Qaeda's Noms de Guerre : How Should We Decode Terrorists' Names?. *In Defense Concepts, CADS Press, Vol. 1, Edition 3, Fall 2006, 6-16.* http://www.c4ads.org/files/defense_concepts_I.3.pdf

Habash Nizar, Diab Mona, & Rambow Owen. (2012). *Conventional Orthography for Dialectal Arabic*. In: *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Istanbul.

Habash Nizar, Owen Rambow, Mona Diab, and Reem Kanjawi-Faraj. 2008. Guidelines for annotation of Arabic dialectness. In *Proceedings of the LREC Workshop on HLT & NLP within the Arabic World*, pages 49–53, Marrakech.

Habash Nizar, Rambow Owen and Roth Ryan. (2009, April). 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* (pp. 102-109).

Habash Nizar. (2010). Introduction to Arabic Natural Language Processing. *Synthesis Lectures on Human Language Technologies*, *3*(1), 1-187.

Hamdi Ahmed, Boujelbane Rahma, Habash Nizar, & Nasr, Alexis. (2013, June). Un système de traduction de verbes entre arabe standard et arabe dialectal par analyse morphologique profonde. In *Actes de la conférence conjointe JEP-TALN-RECITAL*, pp. 396-406.

Hastie Trevor, Tibshirani Robert, Friedman Jerome. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, *27*(2), Springer.

Heba Elfardy and Mona Diab. 2012a. Token level identication of linguistic code switching. In : *COLING (Posters)*. 2012. p. 287-296.

Hinds Martin, & As-Said Muhammad Badawi. (1986). *A Dictionary of Egyptian Arabic: Arabic-English*. Libr. du Liban.

Huang Fei, Vogel Stephan, & Waibel Alex. (2003). Automatic extraction of named entity translingual equivalence based on multi-feature cost minimization. In *Proceedings of the* 41st Annual Conference of the Association for Computational *Linguistics (ACL'03), workshop on Multilingual and mixed-language named entity recognition-Volume 15* (pp. 9-16). Association for Computational Linguistics. Sapporo, Japan.

Huang Fei, Vogel Stephan, & Waibel Alex. (2004). Improving Named Entity Translation Combining Phonetic and Semantic Similarities. In *HLT-NAACL* (Vol. 2004, pp. 281-288).

Hulden Mans. (2009, April). Foma: a finite-state compiler and library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session* (pp. 29-32). Association for Computational Linguistics.

Hussein Mohamed Hamza. (1973). Adjectives in Egyptian Colloquial Arabic and in English. A Contrastive Study. M.A. Thesis. AlAzhar University.

Ibrahim Zeinab. (2010). Cases of written code-switching in Egyptian opposition newspapers. *Arabic and the Media. Linguistic Analyses and Applications. Leiden-Boston, E.-J. Brill*, 23-46.

Jiang Long, Zhou Ming, Chien, Lee-Feng, & Niu, C. (2007, January). Named Entity Translation with Web Mining and Transliteration. In *Proceedings of the 20th International Joint Conference on Artif cial Intelligence 'IJCAI'* (Vol. 7, pp. 1629-1634).

Kahane Sylvain. (2001). Grammaires de dépendance formelles et théorie Sens-Texte. *TALN 2001 (Traitement automatique des langues naturelles)*.

Khadraoui Errime. (2010). *Pour une étude lexicale des pratiques langagières des internautes: le cas des forums de discussions.* Doctoral dissertation, Université El Hadj Lakhdar de Batna.

Khemakem Aïda . (2006). ArabicLDB: une base lexicale normalisée pour la langue arabe. *Mémoire de mastère, SINT, FSEGS, Sfax*.

Khoja Shereen, Garside Roger et Knowles Gerry. (2001). An Arabic Tagset for the Morphosyntactic Tagging of Arabic (Doctoral dissertation).

Kihm Alain. (2003). Les pluriels de l'arabe: systèmes et conséquences pour l'architecture de la grammaire. *Recherches linguistique de Vincennes 32, p.* 109-156.

Knight Kevin, & Graehl Jonathan. (1997). Machine transliteration. *Computational linguistics*, 24(4), p. 599-612.

Koehn Philipp, Hoang Hieu, Birch Alexandra, Callison-Burch Chris, Federico Marcello, Bertoldi, Nicola, Cowan Brooke, Shen Wade, Moran Christine, Zens Richard, Dyer Chris, Bojar Ondrej, Constantin Alexandra, & Herbst Evan. (2007, June). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions* (pp. 177-180). Association for Computational Linguistics.

Koehn Philipp. (2004). Statistical Significance Tests for Machine Translation Evaluation. In *EMNLP* (pp. 388-395).

Kondrak Grzegorz. (2005). Cognates and word alignment in bitexts. In *Proceedings of the tenth machine translation summit (mt summit x)*, 305-312.

Kondrak Grzegorz, Marcu Daniel, & Knight Kevin. (2003). Cognates can improve statistical translation models. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003--short papers-Volume 2* (pp. 46-48). Association for Computational Linguistics.

Kouloughi Djamel Eddine. (1991). *Lexique fondamental de l'arabe standard moderne*. Editions L'Harmattan, Paris.

Lajmi Dhouha. (2009). Spécificités du dialecte Sfaxien. *Synergies Tunisie*, vol. 1, p. 135-142.

Langacker Ronald. W. (1977). Syntactic Reanalysis. *Mechanisms of Syntactic Change*. Vol. 58.

Laporte Eric. (2000). Mots et niveau lexical. *Ingénierie des langues*, 25-49.

Lavergne Thomas, Cappe Olivier, et Yvon François. (2010, July). Practical very large scale CRFs. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 504-513). Association for Computational Linguistics.

Lee Young-suk (2004). Morphological analysis for statistical machine translation. In *Proceedings of HLT-NAACL 2004: Short Papers* (pp. 57-60). Association for Computational Linguistics.

Lefever E., Macken L., & Hoste V. (2009). Language-independent bilingual terminology extraction from a multilingual parallel corpus. In *Proceedings of the 12th Conference of the Conference of the European Chapter of the Association for Computational Linguistics*, (pp. 496-504). Association for Computational Linguistics.

Lei Yun and John H. L. Hansen. 2011. Dialect classification via text-independent training and testing for Arabic, Spanish, and Chinese. *IEEE Transactions on Audio, Speech,and Language Processing*, 19(1):85–96.

Lewis M. Paul, Simons Gary F, et Fennig Charles. D. (2009). *Ethnologue: Languages of the world* (Vol. 9). Dallas, TX: SIL international.

Lui Marco, Lau, Jey Han, & Baldwin Timothy. (2014). Automatic detection and language identification of multilingual documents. *Transactions of the Association for Computational Linguistics*, *2*, 27-40.

MacCartney Bill, Galley Michel, & Manning Christopher D. (2008). A Phrase-Based Alignment Model for Natural Language Inference. In P*roceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 802-811). Association for Computational Linguistics.

Mahadin Radwan Salim. (1982). The Morphophonemics of the Standard Arabic Tri-Consonantal Verbs. Ph.D. Dissertation. University of Pennsylvania.

Malik Sayed Hamza Ayantunde. (1976). *A Contrastive Study of the Verbal Patterns in Standard Arabic and Spoken Egyptian Arabic*. Ph.D. Dissertation. University of Ibadan, Nigeria.

Maloney John, & Niv Michael. (1998). TAGARAB: a fast, accurate Arabic name recognizer using high-precision morphological analysis. In *Proceedings of the Workshop on Computational Approaches to Semitic Languages* (pp. 8-15). Association for Computational Linguistics.

Mansour Saab (2010, December). MorphTagger: HMM-based Arabic segmentation for statistical machine translation. In *IWSLT* (pp. 321-327).

Mansour Saib, Sima'an Khalil et Winter Yoad. (2007, June). Smoothing a lexicon-based POS tagger for Arabic and Hebrew. In *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources* (pp. 97-103). Association

for Computational Linguistics.

Marçais William, & Jelloûli Farès. (1933). *Trois textes arabes* d'*El-Hamma* de Gabès. Impr. Nationale.
Marçais William. (1902). *Le dialecte arabe parlé à Tlemcen: grammaire, textes et glossaire* (Vol. 26). E. Leroux.

Marçais William. (1908). *Le Dialecte arabe de Ulâd Brahîm de Saïda (département d'Oran)*. H. Champion.

Marçais William. (1930). La diglossie arabe dans l'enseignement public. Alger, In *Revue pédagogique, tome CIV, n° 12*, p. 401-409.

Marçais, Philippe. (1977). *Esquisse grammaticale de l'arabe maghrébin* (Vol. 3). Librairie d'Amérique et d'Orient.

Masmoudi Abir, Ellouze Khmekhem Mariem , Estève Yannick , Hadrich Belguith Lamia, & Habash Nizar. (2014, May). A corpus and a phonetic dictionary for Tunisian Arabic speech recognition. In *of the Language Resources and Evaluation Conference, Iceland*.

McCarthy John & Prince Alan. (1990a). Prosodic Morphology and Templatic Morphology. In *Perspectives on Arabic Linguistics II : papers from the second annual symposium on Arabic linguistics (pp. 1-54).*

McDonald David. (1996). Internal and external evidence in the identification and semantic categorization of proper names. *Corpus processing for lexical acquisition*, 21-39.

McGuirk Russell, H. (1986). *Colloquial Arabic of Egypt*. Psychology Press, London: Routledge.

Mejri Salah, Said Mosbah, & Sfar Inès. (2009). Pluringuisme et diglossie en Tunisie. *Synergies Tunisie n*, *1*, 53-74.

Mesfar Slim. (2007). Named Entity Recognition for Arabic Using Syntactic Grammars. In *Natural Language Processing and Information Systems* (pp. 305-316). Springer Berlin Heidelberg.

Mesfar Slim. (2008). Analyse morpho-syntaxique automatique et reconnaissance des entités nommées en arabe standard. *University of Franche-Comté, Ecole doctorale langages, espaces, temps, socits*.

Meunier Mariette. (2009). La constitution d'un corpus, parcours initiatique en linguistique.

Mihalcea Rada, & Pedersen Ted. (2003). An evaluation exercise for word alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts: data driven machine translation and beyond-Volume 3* (pp. 1-10). Association for Computational Linguistics.

Miller Catherine. (2008). Quelles voix pour quelles villes arabes?. *Les boites Noires de Louis*

Jean Calvet, 371-397.

Miller Catherine. (2012). Langues et Médias dans le monde arabe/arabophone. Entre idéologie et marché, convergences dans la glocalisation, *in* A. Lachkar (dir.), *Langues et médias en méditerranée*, Paris, L'Harmattan, Langue et parole, 157-171.

Mitchell, Terence Frederick. (1956). *An Introduction to Egyptian Colloquial Arabic*. London: Oxford University Press.

Mitchell, Terence Frederick. (1990). *Pronouncing Arabic I*. Oxford: Clarendon Press.
Mohamed Emad, Mohit Behrang , & Oflazer Kemal. (2012, July). Transforming standard Arabic to colloquial Arabic. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2 (pp. 176-180). Association for Computational Linguistics.

Moore John. (1979). Arabic Derivational Morphology and Lexical Theory. *Papers from the Regional Meeting, Chicago Linguistic Society* 15: 228-43.

Och Franz. Josef, & Ney Hermann. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, *29*(1), 19-51.

Okita Tsuyoshi, Guerra Maldonado, Alfredo Graham Yvette, Way Andy. (2010). Multi-word expression sensitive word alignment. In *Proceedings of the 4th International Workshop on Cross Lingual Information Access at COLING,* pages 26–34, Beijing.

Orphanos Giorgos, Kalles Dimitris, Papagelis Thanasis, et Christodoulakis Dimitris. (1999). Decision Trees and NLP: A Case Study in POS Tagging. In *Proceedings of annual conference on artificial intelligence (ACAI).*

Oudah Mai, & Shaalan Khaled. (2012). A Pipeline Arabic Named Entity Recognition Using a Hybrid Approach. In *COLING* (pp. 2159-2176).

Ouerhani Bechir. (2009). Interférence entre le dialectal et le littéral en Tunisie: Le cas de la morphologie verbale. *Synergies Tunisie n*, *1*, 75-84.

Ouersighni Riadh. (2002). La conception et la réalisation d'un système d'analyse morpho-syntaxique robuste pour l'arabe: utilisation pour la détection et le diagnostic des fautes d'accord. 2002. Thèse de doctorat. Lyon 2.

Ozdowska Sylwia. (2004). Appariement bilingue de mots par propagation syntaxique à partir de corpus français/anglais alignés. *In Proceedings of the 11ème conférence TALN-RECITAL.*

Ozdowska Sylwia, Claveau Vincent. (2006). Inférence de règles de propagation syntaxique pour l'alignement de mots. *TAL (Traitement Automatique des Langues)*, 47(1), 167-186.

Papineni Kishore, Roukos Salim, Ward Todd, & Zhu Wei-Jing. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 311-318). Association for Computational Linguistics.

Petasis Georgios, Vichot Frantz, Wolinski Francis, Paliouras Georgios, Karkaletsis Vangelis, et Spyropoulos Constantine. D. (2001). Using Machine Learning to Maintain Rule-based Named-Entity Recognition and Classification Systems. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics* (pp. 426-433). Association for Computational Linguistics.

Queffélec Ambroise, Derradji Yacine, Debov Valéry, Smaali-Dekdouk Dalila, & Cherrad-Benchefra Yasmina. (2002). *Le français en Algérie : lexique et dynamique des langues*, Ed. Duclot, AUF.

Radwan Mohamed Ramzy. (1975). *A Semantico-Syntactic Study of the Verbal Piece in Colloquial Egyptian Arabic*. Doctoral dissertation, School of Oriental and African Studies, University of London.

Robertson Alice Marian. (1970). *Classical Arabic and Colloquial Cairene: An Historical Linguistic Analysis*. Ph.D. Dissertation. University of Utah at Salt Lake City.

Roman André. (1999). La création lexicale en arabe (ressources et limites de la nomination dans une langue humaine naturelle). *Études arabes*.

Roman André. (2007). La création lexicale en arabe : étude diachronique et synchronique des sons et des formes de la langue arabe. Lyon, *Presses universitaires de Lyon*, *2001*, éd. revue et augmentée 2007.

Saâdane Houda et Semmar Nasredine. (2012). Transcription des noms arabes en écriture latine. In *Revue RIST| Vol*, *20*(2), 57.

Saâdane Houda, Semmar Nasredine. (2012). Utilisation de la translittération arabe pour l'amélioration de l'alignement de mots à partir de corpus parallèles français-arabe. In *Actes de la conférence conjointe JEP-TALN-RECITAL*, volume 2: TALN, Grenoble, pages 127–140.

Saâdane Houda, & Habash Nizar (2015, July). A Conventional Orthography for Algerian Arabic. In *Proceedings of the Second Workshop on Arabic Natural Language Processing ANLP,* (pp. 69–79), Beijing, China.

Saâdane Houda, Christian Flhur, & Mathieu Guidère. (2013). La reconnaissance automatique des dialectes arabes à l'écrit. In *Colloque international Traduction et Champs Connexes : quelle place pour la langue arabe aujourd'hui?*. Alger, 18-20 décembre 2013.

Saâdane Houda, Rossi Aurélie, Fluhr Christian, & Guidère Mathieu. (2012). Transcription of Arabic names into Latin. In *Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), 2012 6th International Conference on* (pp. 857-866). IEEE.

Saâdane Houda, Semmar Nasredine. (2012). Utilisation de la translittération arabe pour l'amélioration de l'alignement de mots à partir de corpus parallèles français-arabe. In *Actes de la conférence conjointe JEP-TALN-RECITAL*, volume 2: TALN , pages 127–140.

Saâdane Houda, & Semmar Nasredine. (2013). Transcription des noms arabes en écriture latine. In *Revue RIST| Vol*, 20(2), 57.

Saâdane Houda. (2011). Dialectologie arabe et transcription automatique des noms. In *Colloque RJCP – GIPSA-lab,* université de Grenoble; 25 mai 2011.

Sadat Fatiha, & Habash Nizar. (2006). Combination of Arabic preprocessing schemes for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pp. 1-8.*

Salama Ahmed, Bouamor Houda, Mohit Behrang & Oflazer Kemal . YouDACC: the Youtube Dialectal Arabic Commentary Corpus. In *Proceedings of the Language Resources and Evaluation Conference, LREC* (pp. 1246-1251)., At Reykjavik.

Saleem Abuleil 2004. Extracting Names from Arabic Text for Question-Answering Systems. In *Proceedings of Coupling approaches, coupling media and coupling languages for information retrieval (RIAO 2004),* Avignon, France. pp. 638- 647.

Salib Maurice Boulos. (1981). *Spoken Arabic of Cairo*. Cairo: American University in Cairo Press.

Schmidt Richard. (1975). Sociolinguistic Variation in Spoken Egyptian Arabic: A Reexamination of the Concept of Diglossia. Ph.D. Dissertation Brown University.

Schramm Gene M. (1962). An Outline of Classical Arabic Verb Structure. *Language* : 360-75.

Semmar Nasredine, Laïb Meriama. (2010). Using a Hybrid Word Alignment Approach for Automatic Construction and Updating of Arabic to French Lexicons. In *Proceedings of the LREC 2010: Workshop on Language Resources and Human Technologies for Semitic Languages, Malta,* 2010.

Semmar Nasredine, Servan Christophe, de Chalendar Gaël, Le Ny Benoît, & Bouzaglou, Jean-Jacques. (2010). A Hybrid Word Alignment Approach to Improve Translation Lexicons with Compound Words and Idiomatic Expressions. In *Proceedings of the 32nd Translating and the Computer conference-ASLIB*, London (England).

Seretan Violeta, & Wehrli Eric. (2007). Collocation translation based on sentence alignment and parsing. In *Actes de la 14e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2007),* pp. 401-410.

Sfar Inès. (2005). Morphologie des noms de professions: incorporation et paraphrase. In *Journée scientifique de formation et d'animation régionale* (pp. 156-163). AUF.

Sfar Inès. (2006). Fonctions syntagmatiques et incorporation dérivationnelle (affixale et par schèmes) des noms de professions. *Les noms de professions : Approches Linguistiques, Contrastives et Appliquées*, 79-94, Université Autonome de Barcelone.

Shaalan Khaled, & Raza Hafsa. (2007, June). Person name entity recognition for Arabic. In *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources* (pp. 17-24). Association for Computational Linguistics.

Shaalan Khaled, & Raza Hafsa. (2008). Arabic named entity recognition from diverse text types. In Advances in Natural Language Processing (pp. 440-451). Springer Berlin Heidelberg.

Shaalan Khaled, et Raza Hafsa. (2009). NERA: Named entity recognition for Arabic. *Journal of the American Society for Information Science and Technology*, *60*(8), 1652-1663.

Shao Lie, & Ng Hwee Tou . (2004). Mining new word translations from comparable corpora. In *Proceedings of the 20th international conference on Computational Linguistics (COLING'04),* p. 618. Association for Computational Linguistics.

Sherif, T., & Kondrak, G. (2007, June). Bootstrapping a stochastic transducer for Arabic-English transliteration extraction. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007),* Vol. 45, No. 1, p. 864-871. Prague.

Simard Michel, Foster George. F, & Isabelle Pierre. (1993). Using cognates to align sentences in bilingual corpora. In *Proceedings of the 1993 conference of the Centre for Advanced Studies on Collaborative research: distributed computing-Volume 2* (pp. 1071-1082). IBM Press.

Smadja Frank, McKeown Kathleen, Hatzivassiloglou Vasileios. (1996). Translation Collocations for Bilingual Lexicons: A Statistical Approach. *Computational Linguistics*, 22 (1), 1-38.
Smith, Percy (1917). A Plea for Literature in Vernacular Arabic. *The Muslim World* 7(4), 333-342.

Stalls Bonnie, Knight Kevin. (1998). Translating names and technical terms in Arabic text. In *Proceedings of the COLING/ACL Workshop on Computational Approaches to Semitic Languages*, (pp. 34-41). Association for Computational Linguistics. Montreal, Québec.

Stolcke Andreas. (2002). SRILM-an extensible language modeling toolkit. In *INTERSPEECH.*

Taleb Ibrahimi Khawla. (1997). *Les Algériens et leur (s) langue (s): éléments pour une approche sociolinguistique de la société algérienne*. Éds. ElHikma.

Taleb Ibrahimi Khawla. (2006). L'Algérie: coexistence et concurrence des langues. *L'Année du Maghreb*, (I), 207-218.

Tao Tao, Yoon Su-Youn, Fister Andrew, Sproat Richard, & Zhai ChengXiang. (2006). Unsupervised named entity transliteration using temporal and phonetic correlation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'06),* pp. 250-257. Sydney. Association for Computational Linguistics.

Tesnière, L. (1959). *Éléments de syntaxe structurale*. Librairie C. Klincksieck.

Thackston Wheeler McIntosh. (1984). *An Introduction to Koranic Arabic*. Cambridge, Mass: Department of Near Eastern languages and Civilizations, Harvard University.

Tiedemann Jörg. (2009). News from OPUS-A collection of multilingual parallel corpora with tools and interfaces. In *Recent advances in natural language processing* (Vol. 5, pp. 237-248).

Traboulsi Hayssam. (2006). *Named Entity Recognition: A local grammar-based approach* (Doctoral dissertation, University of Surrey).

Traboulsi Hayssam. (2009, October). Arabic named entity extraction: A local grammar-based approach. In *Proceedings of the International Multiconference on Computer Science and Information Technology IMCSIT* (pp. 139-143).

Trask Robert Lawrence. (1996). *A Dictionary of Phonetics and Phonology*. New York: Routledge.

Travis, D. Ann (1979) *Inflectional Affixation in Transformational Grammar: Evidence from the Arabic Paradigm* (No. 192). Indiana University Linguistics Club.

Tufis Dan Ioan, & Ion Radu. (2007). Parallel corpora, alignment technologies and further prospects in multilingual resources and technology infrastructure. In *Proceedings of the 4th International Conference on Speech and Dialogue Systems* (pp. 183-195).

Vapnik Vladimir. (2013). *The nature of statistical learning theory*. Springer Science & Business Media.

Véronis Jean, Hamon Olivier, Ayache Christelle, Belmouhoub Rachid, Kraif Olivier, Laurent Dominique, Nguyen Thi Minh Huyen, Semmar Nasredine, Stuck François, & Zaghouani Wajdi. (2008). Arcade II Action de recherche concertée sur l'alignement de documents et son évaluation. Chapitre 2, Editions Hermès, 2008.

Versteegh Kees. (1997). The Arabic Language. New York: Columbia University Press.

Versteegh Kees. (2011). Les dialectes arabes. *Dictionnaire des langues*, 336-346. Paris: Presses Universitaires de France.

Vintar Spela, Fisier Darja. (2008). Harvesting multi-word expressions from parallel corpora. In *Proceedings of LREC, Marrakech, Morocco*.

Wise Hilary. (1975). *A Transformational Grammar of Spoken Egyptian Arabic*. Blackwell.

Wright William. (1967). *A Grammar of the Arabic Language*. 3rd ed. Cambridge: Cambridge University Press.

Zaafrani Riadh .(1997). Morphological analysis for an Arabic Computer-aided learning system. In *Proceedings of DIALOGUE*, *97*, 10-15.

Zaghouani Wajdi. (2009). Le repérage automatique des entités nommées dans la langue arabe: vers la création d'un système à base de règles. Master's thesis, University of Montreal.

Zaghouani Wajdi. (2012). RENAR: A rule-based Arabic named entity recognition system. *ACM Transactions on Asian Language Information Processing (TALIP)*, *11*(1), 2.

Zaidan Omar F., & Callison-Burch Chris. (2011a). The Arabic online commentary dataset: an annotated dataset of informal Arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2* , pages 37–41. Association for Computational Linguistics.

Zaidan Omar F., & Callison-Burch Chris. (2014). Arabic dialect identification. *Computational Linguistics, 40(1)*, pages 171-202.

Zbib, Rabih, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris Callison-Burch. 2012. Machine translation of Arabic dialects. In the *2012* Conference of the North American Chapter of *the Association for Computational Linguistics*, pages 49–59, Montreal.

Zmantar Yasser, Dichy Joseph. (2009). L'analyse automatique des mots-outils en arabe. In *Systèmes d'Information et Intelligence Economique, 2ème Conférence Internationale* (pp. 586-597). IHE éditions.

Zoll, Cheryl Cydney. (1996). Parsing Below the Segment in a Constraint Based Framework. Doctoral dissertation. University of California at Berkeley.

Zribi Ines, Boujelbane Rahma, Masmoudi Abir, Ellouze Mariem, Belguith Lamia, & Habash Nizar. (2014). A Conventional Orthography for Tunisian Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland*.

TABLE DES MATIERES

INTRODUCTION GENERALE---------- 3

PROBLEMATIQUE DU SUJET : ---------- 5
PISTES D'INVESTIGATION EMPIRIQUES :---------- 5
CONTEXTE DE MES TRAVAUX DE RECHERCHE ---------- 6
LE PROJET SAIMSI (SUIVI ADAPTATIF INTERLINGUE ET MULTISOURCE DES INFORMATIONS) ---------- 6
LE PROJET ORELO (ORIGINE DES REDACTEURS ET DES LOCUTEURS)---------- 7
ORGANISATION DE LA THESE---------- 8

PARTIE I : DESCRIPTION DE L'ARABE STANDARD ET DIALECTAL---------- 13

CHAPITRE 1 LA LINGUISTIQUE DE LA LANGUE ARABE---------- 14

INTRODUCTION ----------15
1.1. PRESENTATION DE LA LANGUE ARABE ----------15
1.2. SYSTEME D'ECRITURE DE L'ARABE ----------16
1.3. LEXIQUE ET GRAMMAIRE ----------16
1.3.1. Nom----------*16*
1.3.2. Verbe ----------*17*
1.3.3. Pronoms ----------*18*
1.3.4. Les mots outils ----------*19*
1.4. MORPHOLOGIE FLEXIONNELLE:----------19
1.4.1. Flexions des verbes (conjugaison) ----------*20*
1.4.1.1. Genre du verbe : ---------- 20
1.4.1.2. Paradigme de conjugaison du verbe ---------- 20
1.4.1.3. Trait grammatical---------- 22
1.4.2. Flexions des noms ----------*22*
1.4.2.1. Les déclinaisons au singulier ---------- 22
1.4.2.2. Les déclinaisons au duel ---------- 23
1.4.2.3. Les déclinaisons au pluriel---------- 24
1.4.2.3.1. Le pluriel externe ou régulier :---------- 24
1.4.2.3.2. Le pluriel externe masculin : الجمع المذكر السالم ---------- 24
1.4.2.3.3. Le pluriel externe féminin : الجمع المؤنث السالم---------- 24
1.4.2.3.4. Le pluriel interne ou brisé (جمع التكسير) ---------- 24
1.5. LES PROBLEMES D'ANALYSE DU TRAITEMENT AUTOMATIQUE DE LA LANGUE ARABE ----------24
1.5.1. L'absence de voyelle – voyellation – ----------*25*
1.5.2. Agglutination----------*26*
1.5.2.1. Les proclitiques---------- 27
1.5.2.2. Les enclitiques---------- 28
1.5.3. Ambiguïté lexicale et syntaxique----------*29*
1.5.4. Irrégularité de l'ordre des mots dans la phrase ----------*30*

CHAPITRE 2 INTRODUCTION AUX DIALECTES ARABES ---------- 32

INTRODUCTION ----------33
2.1. LES VARIANTES DE LANGUES ARABES----------34
2.1.1 L'arabe Classique ----------*35*
2.1.2 L'arabe standard (MSA)----------*35*
2.1.3 L'arabe dialectal ----------*36*

2.2. LES VARIETES DIALECTALES DE LA LANGUE ARABE ---- 38
2.2.1 Les dialectes de la péninsule arabique (Golf) ---- *38*
2.2.2 Les dialectes mésopotamiens (Irakien) ---- *38*
2.2.3 Les dialectes levantins ---- *39*
2.2.4 Les dialectes égyptiens ---- *39*
2.2.5 Les dialectes maghrébins ---- *39*
2.3. LA SITUATION LINGUISTIQUE DE LA LANGUE ARABE ---- 40
2.4. APERÇU HISTORIQUE DE L'ARABE ALGERIEN (AA) ---- 42
2.5. COMPARAISON ENTRE L'ARABE ALGERIEN, EGYPTIEN, TUNISIEN ET LE MSA ---- 44
2.5.1 Variations phonologiques ---- *44*
2.5.2 Variations Morphologiques ---- *47*
2.5.3 Variations Orthographiques ---- *49*
2.5.4 Variations lexicales ---- *50*
2.5.4.1. *La dérivation* ---- 50
2.5.4.2. *L'emprunt* ---- 50
2.5.5 La variation syntaxique ---- *51*

PARTIE II : ANALYSE LINGUISTIQUE DE LA LANGUE ARABE ---- 54

CHAPITRE 3 ANALYSE MORPHOSYNTAXIQUE ---- 55

INTRODUCTION ---- 56
3.1. ETAT DE L'ART SUR LE TRAITEMENT AUTOMATIQUE DE L'ARABE ---- 56
3.2. SYSTEME D'ANALYSE LINGUISTIQUE PROPOSE ---- 60
3.2.1. Segmentation locale (Tokenisation) ---- *60*
3.2.2. Analyse morphologique ---- *62*
3.2.2.1. Segmentation des formes agglutinées ---- 63
3.2.2.2. La désambiguïsation ---- 64
3.2.2.3. Transformation morphologique (Règles réécriture) ---- 66
3.2.3. L'analyse syntaxique ---- *69*
3.2.4. Les relations syntaxiques ---- *70*
3.2.4.1. Les relations syntaxiques gouvernées par le nom ---- 71
3.2.4.1.1. La modification ---- 71
3.2.4.1.2. Le complément de nom ---- 73
3.2.4.1.3. Le complément d'objet indirect ---- 73
3.2.4.1.4. L'apposition ---- 74
3.2.4.1.5. La corroboration (al-tawabi' – al-tawkîd) ---- 75
3.2.4.1.6. La quantification numérale ---- 77
3.2.4.1.7. La conjonction de coordination ---- 78
3.2.4.2. Les relations syntaxiques gouvernées par un adjectif ---- 78
3.2.4.2.1. Les relations syntaxiques de surface contrôlées par la valence de l'adjectif ---- 79
3.2.4.2.2. La relation complément de l'adjectif ---- 80
3.2.4.2.3. La relation modificative ---- 80
3.2.4.2.4. La relation comparative ---- 81
3.2.4.2.5. La relation superlative ---- 81
3.2.4.2.6. La relation conjonction de coordination ---- 82
3.2.4.3. Les relations syntaxiques gouvernées par les mots outils ---- 82
3.2.4.3.1. L'interjection d'appel ---- 83
3.2.4.3.2. La préposition ---- 83
3.2.4.3.3. La conjonction de subordination ---- 84
3.2.4.3.4. La conjonction de coordination ---- 84
3.2.4.3.5. L'exception ---- 85
3.2.4.4. Les relations syntaxiques gouvernée par le verbe ---- 86
3.2.4.4.1. Relation Sujet {(V) -sujet→(N)} ---- 87

3.2.4.4.2. Le complément d'objet direct ---- 88
3.2.4.4.3. Le complément d'objet indirect ---- 92
3.2.4.4.4. L'agent prépositionnel ---- 92
3.2.4.4.5. L'attribut ---- 93

CHAPITRE 4 IDENTIFICATION ET TYPAGE DES ENTITES NOMMEES ---- 95

INTRODUCTION ---- 96
4.1. TYPOLOGIE DES ENTITES NOMMEES ---- 96
4.2. APPLICATIONS ---- 97
4.3. PARTICULARITE DE LA LANGUE ARABE LIEE A LA DETECTION DES ENTITES NOMMEES ---- 99
4.4. ÉTAT DE L'ART SUR LES SYSTEMES DE RECONNAISSANCE DES ENTITES NOMMEES EN ARABE ---- 100
4.4.1. Identification des entités nommées ---- 100
4.4.1.1. Les preuves internes ---- 101
4.4.1.2. Les preuves externes ---- 101
4.4.2. Systèmes de reconnaissance des entités nommées ---- 101
4.4.2.1. Les systèmes à base de règle ---- 102
4.4.2.2. Les systèmes Statistiques ---- 103
4.4.2.3. Les systèmes Hybrides ---- 105
4.5. APPROCHE PROPOSEE POUR LA RECONNAISSANCE DES ENTITES NOMMEES ---- 106
4.6. RECONNAISSANCE DES NOMS PROPRES – ENAMEX ---- 109
4.6.1. Les noms de personnes ---- 109
4.6.1.1. Structure des noms de personnes : ---- 110
4.6.1.2. Identification des noms de personnes ---- 110
4.6.2. Identification des lieux ---- 115
4.6.3. Les noms d'organisation ---- 115
4.6.3.1. Structure des noms d'organisation ---- 115
4.6.3.2. Identification des noms d'organisation ---- 116
4.7. RECONNAISSANCE DES EXPRESSIONS NUMERIQUES – NUMEX ---- 117
4.7.1. Identification des déterminants numériques ---- 117
4.7.2. Identification des expressions numériques ---- 118

PARTIE III : TRAITEMENT DES DIALECTES ARABES ---- 120

CHAPITRE 5 ANALYSE PHONOLOGIQUE ---- 121

INTRODUCTION ---- 122
5.1. SYSTEME CONSONANTIQUE ---- 122
5.1.1. Prononciation de la consonne qaf ---- 122
5.1.2. Prononciation de la consonne jim ---- 124
5.1.3. Prononciation des spirantes interdentales ---- 124
5.1.4. Traitement du hamza ---- 126
5.1.5. Autres cas de prononciation particulière : ---- 127
5.2. SYSTEME VOCALIQUE ---- 128
5.3. LES ALTERNANCES PHONOLOGIQUES (VARIATIONS OU DEGRADATIONS PHONOLOGIQUES) ---- 131
5.3.1. Assimilation (الإِدْغَامْ) ---- 132
5.3.2. Métathèses ---- 135
5.3.3. Epenthèse ---- 136
5.3.4. Elision ---- 138
5.3.5. Raccourcissement ---- 139

CHAPITRE 6 ANALYSE MORPHOLOGIQUE VERBALE ---- 142

INTRODUCTION ---- 143

6.1. VERBE, STEMS ET CLASSES ---- 143
6.1.1. Verbes tri-radicaux ---- *143*
6.1.1.1. Verbes sonores ---- 144
6.1.1.2. Verbes géminés (sourds) ---- 152
6.1.1.3. Verbes glottalisées (hamzé) ---- 156
6.1.1.4. Les verbes faibles ---- 159
6.1.1.4.1. Les verbes ayant une glide initiale (appelés aussi verbes assimilé) ---- 159
6.1.1.4.2. Les verbes concaves (creux) ---- 162
6.1.1.4.3. Les verbes défectueux ---- 169
6.1.1.4.4. Les verbes doublement faibles ---- 174
6.1.2. Les Verbes quadrilitères ---- *175*
6.2. L'ASPECT ET LE MODE DE LA FLEXION: ---- 178
6.3. LA VOIX DE LA FLEXION: ---- 188

CHAPITRE 7 ANALYSE MORPHOLOGIQUE NOMINALE ---- 191

INTRODUCTION ---- 192
7.1. LES NOMS SIMPLES ---- 192
7.1.1. Les racines bilitères ---- *193*
7.1.2. Les racines trilitères ---- *194*
7.1.3. Les racines quadrilitères ---- *201*
7.2. LES NOMS DEVERBAUX ---- 201
7.2.1. Les noms verbaux (المصدر) ---- *202*
7.2.2. Les substantifs d'exagération (صيغ المبالغة) ---- *207*
7.2.3. Les noms de lieu et de temps (أسماء المكان والزمان) ---- *207*
7.2.4. Les noms d'instruments (أسماء الآلة) ---- *210*
7.3. NOMS DEFINIS VS NOMS INDEFINIS ---- 211
7.4. LE CAS DE FLEXION ---- 211
7.5. LE GENRE DANS LA FLEXION ---- 214
7.6. LE NOMBRE DANS LA FLEXION ---- 216
7.6.1. Le singulier ---- *216*
7.6.2. Le duel ---- *216*
7.6.3. Le pluriel ---- *218*
7.6.3.1. Le pluriel masculin sain ---- 218
7.6.3.2. Le pluriel féminin sain ---- 219
7.6.3.3. Le pluriel brisé ---- 220

CHAPITRE 8 ANALYSE MORPHOLOGIQUE ADJECTIVALE ---- 224

INTRODUCTION ---- 225
8.1. LES FORMES DES RACINES ADJECTIVALES ---- 225
8.2. ADJECTIFS DEFINIS VS INDEFINIS ---- 229
8.3. LES CAS DE FLEXION ---- 230
8.4. LE GENRE DE LA FLEXION ---- 231
8.5. LE NOMBRE DE LA FLEXION ---- 232
8.5.1. Le Singulier ---- *232*
8.5.2. Le Duel ---- *232*
8.5.3. Le pluriel ---- *233*
8.5.3.1. Le Pluriel Masculin Sain ---- 233
8.5.3.2. Le Pluriel Féminin Sain ---- 234
8.5.3.3. Le Pluriel Brisé ---- 235
8.6. LE DEGRE DE LA FLEXION: ---- 239

8.6.1. Le Degré Positif ---- *239*
8.6.2. Le Degré Comparatif ---- *239*
8.6.3. Le degré Superlatif: ---- *241*
8.7. LES ADJECTIFS RELATIONNELS (NISBAS) ---- 243

PARTIE IV : CONTEXTE ET MATERIEL / GENERATION AUTOMATIQUES DES RESSOURCES 246

CHAPITRE 9 CREATION DES LEXIQUES---------- 247

INTRODUCTION ---- 248
9.1. CONSTRUCTION DES DICTIONNAIRES DIALECTAUX ---- 248
9.1.1. Préliminaires ---- *249*
9.1.2. Présentation de la méthode ---- *250*
9.1.3. La convention CODA ---- *252*
9.2. CONSTRUCTION DES DICTIONNAIRES VERBAUX ---- 254
9.2.1. Construction des lemmes ---- *254*
9.2.2. Construction des schèmes ---- *254*
9.2.2.1. Classification selon le modèle du verbe ---- 255
9.2.2.2. Classification selon la voyelle de la deuxième consonne ---- 256
9.2.2.3. Classification selon la marque de l'inaccompli ---- 258
9.2.3. Construction des racines ---- *259*
9.3. CONSTRUCTION DES NOMS ET DES MOTS OUTILS ---- 260
9.4. LE CAS DE L'EMPRUNT ---- 261
9.5. GENERATION AUTOMATIQUE DES FORMES FLECHEES ---- 261
9.5.1. Les règles morpho-phonémiques et orthographiques ---- *262*
9.5.1.1. Les règles morpho-phonémiques ---- 263
9.5.1.2. Les règles orthographiques ---- 263
9.5.2. Présentation des règles utilisées ---- *264*

CHAPITRE 10 TRANSLITTERATION DES NOMS PROPRES ARABES ---------- 266

INTRODUCTION ---- 267
10.1. LES DIFFERENTS ASPECTS DU SUJET ---- 267
10.1.1. Aspect linguistique ---- *267*
10.1.2. Aspect cognitif ---- *268*
10.1.3. Aspect dialectologique ---- *269*
10.2. ÉTAT DE L'ART SUR LA TRANSLITTERATION ET LA TRANSCRIPTION ---- 270
10.3. TRANSLITTERATION DE NOMS ARABES EN ECRITURE LATINE ---- 271
10.3.1. Normes de translittération pour l'arabe ---- *272*
10.3.2. Correspondances proposées pour la translittération des lettres arabes vers le latin ---- *272*
10.3.3. Approche proposée pour la translittération de noms propres arabes en écriture latine ---- *275*
10.4. TRANSCRIPTION DES NOMS ARABES EN ECRITURE LATINE VERS L'ARABE ---- 278
10.4.1. Les problèmes de la transcription du latin vers l'arabe ---- *278*
10.4.2. Le fonctionnement du module de translittération du latin vers l'arabe ---- *279*
10.4.2.1. Stratégie 1 : Désambiguïsation par la racine consonantique ---- 280
10.4.2.2. Stratégie 2 : Désambiguïsation par le contexte phonologique ---- 281
10.5. VALIDATION DES RESULTATS ---- 281

CHAPITRE 11 CONSTITUTION DES CORPUS ---------- 285

INTRODUCTION ---- 286
11.1. ÉTAT DE L'ART SUR LA CONSTITUTION DES CORPUS ---- 286
11.2. CONSTITUTION DES CORPUS ---- 292

11.2.1. Obtention des liens ---- *294*
11.2.1.1. La presse ---- 295
11.2.1.2. Réseaux sociaux (Facebook) ---- 296
11.2.1.3. Youtube ---- 297
11.2.1.4. Forums de discussion ---- 299
11.2.2. Téléchargement des pages HTML ---- *299*
11.2.2.1. Présentation du logiciel HTTrack ---- 300
11.2.3. Extraction des données ---- *301*
11.3. ANNOTATION DES CORPUS ET IDENTIFICATION DES DIALECTES ---- 303
11.3.1.1. Les difficultés de l'identification des dialectes ---- 304
11.3.1.2. Applications de l'identification des dialectes ---- 304
11.3.1.3. Annotation des textes arabes ---- 304
11.3.1.4. Annotation des textes Arabizi ---- 305
11.3.1.5. Approche d'annotation au niveau du mot ---- 306
11.3.1.5.1. Prétraitement ---- 306
11.3.1.5.2. Analyseurs morphologiques ---- 307
A. Marquages des mots arabes ---- 307
B. Marquage des mots étrangers ---- 307
C. Fonctionnement de l'analyseur morphologique ---- 308
11.3.1.5.3. Combinaison : ---- 310
11.3.1.6. Approche d'annotation au niveau des textes ---- 311
11.3.1.7. Description des balises et des attributs ---- 313
11.4. INTERFACE D'ANNOTATION ---- 313
11.5. EXTRACTION DES TRAITS DE RECONNAISSANCE AUTOMATIQUE DES DIALECTES ARABES ---- 319
11.5.1. Détection des dialectales par le biais des pronoms personnels isoles ---- *319*
11.5.2. Détection des dialectes par les pronoms et les adverbes interrogatifs ---- *320*
11.5.3. Détection des dialectales par les pronoms personnels suffixes ---- *320*
11.5.4. Détection des dialectales par les particules ---- *320*
11.5.5. Détection des dialectales par le schème verbal et la forme passive ---- *320*
11.5.6. Le cas de la consonne « q » dans les dialectales arabes ---- *321*
11.5.7. Détection des dialectes par la transcription des lettres ---- *321*

PARTIE V : RESULTATS EXPERIMENTAUX ET EVALUATION ---- 322

CHAPITRE 12 SYSTEME D'EVALUATION ET D'EXTRACTION DE CONNAISSANCES DU MSA 323

12.1. SYSTEME D'EXTRACTION DE CONNAISSANCES DE GEOLSEMANTICS ---- 324
12.2. ANALYSE LINGUISTIQUE PROFONDE ---- 325
12.2.1. Découpage en mots ---- *326*
12.2.2. Analyse morphologique ---- *326*
12.2.3. Repérage des dates ---- *326*
12.2.4. Identification des relations syntaxiques ---- *326*
12.2.4.1. Les syntagmes nominaux ---- 326
12.2.4.2. Relations sujet-verbe-complément ---- 327
12.2.4.2.1. Gestion de la forme passive ---- 327
12.2.5. Reconnaissance des entités nommées ---- *328*
12.3. EXTRACTION SEMANTIQUE ---- 329
12.3.1. Sélection des concepts probables ---- *329*
12.3.2. Sélection des règles à appliquer ---- *330*
12.3.2.1. L'application des règles sélectionnées ---- 331
12.3.2.2. L'extraction des entités nommées et son contrôle ---- 331
12.3.3. La création d'entités nommées ---- *332*
12.4. MISE EN COHERENCE ---- 332

12.4.1. Regroupement des entités nommées ---- *333*
12.4.2. Résolution des dates relatives ---- *333*
12.5. ÉVALUATION DU SYSTEME ARABE ---- 335
12.5.1. Évaluation de la segmentation ---- *335*
12.5.1.1. Corpus ---- 335
12.5.1.2. Déroulement et résultats ---- 335
12.5.1.3. Protocole expérimental ---- 336
12.5.2. Évaluation de l'extraction des entités nommées ---- *337*
12.5.2.1. Description du Corpus arabe ---- 337
12.5.2.2. Protocole expérimental ---- 337
12.5.2.3. Discussion des erreurs détectées ---- 338
12.5.3. Évaluation de l'extraction de connaissances ---- *338*
12.6. CONCLUSION ---- 340

CHAPITRE 13 SYSTEME D'EVALUATION DE LA TRANSLITTERATION DES NOMS PROPRES 341

13.1. ÉTAT DE L'ART SUR L'ALIGNEMENT DE MOT ---- 342
13.2. APPROCHE PROPOSEE POUR L'ALIGNEMENT DE MOTS A PARTIR DE CORPUS DE TEXTES PARALLELES FRANÇAIS-ARABE ---- 344
13.3. RESULTATS EXPERIMENTAUX ET DISCUSSION ---- 348
13.4. CONCLUSION ET TRAVAUX FUTURS ---- 350

CONCLUSION GENERALE ---- 352

CONTRIBUTIONS ---- 352
PERSPECTIVES ---- 354

BIBLIOGRAPHIE ---- 358