UNIVERSITE SIDI MOHAMED BEN ABDELLAH
FACULTÉ DES SCIENCES ET TECHNIQUES FÈS
DÉPARTEMENT D'INFORMATIQUE

PROJET DE FIN D'ETUDES

MASTER SCIENCES ET TECHNIQUES
SYSTÈMES INTELLIGENTS & RÉSEAUX

ETUDE ET RÉALISATION D'UN ANALYSEUR MORPHOLOGIQUE DE LA LANGUE ARABE

LIEU DE STAGE : LABORATOIRE LSIA

RÉALISÉ PAR : ED-DARIOUACHE ADNANE

SOUTENU LE 24 JUIN 2015

ENCADRÉ PAR :
MR BENABBOU ABDERRAHIM

DEVANT LE JURY COMPOSÉ DE :
MRBENABBOU ABDERRAHIM
MRBEN ABBOU RACHID
MME MAJDA AÏCHA
MME MRABTI FATIHA

ANNÉE UNIVERSITAIRE 2014-2015

Résumé

Nous présentons dans ce projet la réalisation d'un analyseur morphologique de la langue arabe. Le développement de cet analyseur nécessite une étude sur la morphologie arabe, les techniques déjà utilisé pour ce but et une vue générale sur les analyseurs existants. Cette étude contient une description sur la langue arabe et la morphologie arabe. Puis, un aperçut sur toutes les approches existantes avec ses avantages et ses inconvénients. Enfin, une description détaillée sur des analyseurs existants.

Notre analyseur a pour but d'offrir plusieurs niveaux de performances et de fiabilité pour répondre aux besoins des différentes applications de Traitement Automatique de la Langue Arabe telles que l'analyse syntaxique, l'analyse sémantique, la traduction automatique , la recherche d'information, la correction orthographique, etc.

Mots Clés : Traitement automatique de la langue naturelle, Langue arabe, La morphologie arabe, Analyse morphologique, Ambiguïté.

Remerciement

Je remercie tout d'abord Dieu tout puissant de m'avoir donné le courage, la force et la patience d'achever ce bon travail.

Je tiens tout d'abord à exprimer ma profonde gratitude et mes sincères remerciements à mon encadrant Monsieur Benabbou Abderrahim, professeur à la faculté des sciences et technique de Fès, pour m'avoir encadré avec un grand intérêt et une grande compétence, pour ses disponibilités, ses soutiens, ses conseils, pour la qualité de son suivi durant toute la période de ce travail, et pour les encouragements qui m'ont permis de mener à bien ce travail.

Je remercie également le doctorant Mourad Brour pour le soutient qui m'a apporté tout au long de la période du travail.

Je tiens également à remercier les membres du jury, qui ont accepté de juger mon travail.

Merci à tout le corps professoral, administratif et technique de la FST FÈS, pour la qualité de l'enseignement qui nous a été dispensé et le séjour agréable durant ces deux année, spécialement ceux de département Informatique pour leur apport en savoir.

Je ne peux qu'être infiniment reconnaissant envers mes parents pour leur soutien indescriptible, leur patience, leur confiance, leurs sacrifice. Je leur dédie avec plaisir ce travail ainsi qu'à toute ma grande famille.

Finalement, je tiens à remercier tous mes amis du Master Systèmes intelligents et Réseaux, et à tous ceux qui ont participé de près ou de loin pour la réalisation de ce travail.

Liste des figures

Liste des tableaux

Sommaire

Introduction générale

Ce projet s'inscrit dans le cadre du traitement automatique de la langue naturel (TALN). Cette discipline repose sur les quatre pôles : la linguistique, l'informatique, les mathématiques et l'intelligence artificielle. Elle a pour objectif la conception et le développement de programmes techniques informatiques capables de traiter de façon automatique les données exprimées dans une langue. Les applications liées au TALN ont fait l'objet d'une attention particulière depuis plusieurs décennies.

Nous présentons dans ce travail une étude de la morphologie de la langue arabe qui ne possède jusqu'à maintenant aucun système d'analyse morphologique qui pourra faire une analyse morphologique valide et complète de la totalité de ses phénomènes. En d'autres termes, on a constaté qu'il y a une augmentation fulgurante de l'utilisation des textes arabes. Actuellement, le traitement et l'utilisation de ces textes restent encore un défi pour les spécialistes dans ce domaine.

Notons que dans ce domaine(TALN), la plupart des applications (traitement automatique, recherche d'information, indexation automatique, traitement automatique de la parole, résumé automatique, vérification et correction orthographique, etc.) nécessitent un module d'analyse morphologique. Le traitement morphologique consiste essentiellement à mettre en œuvre un processus qui cherche à dégager un maximum d'informations permettant de caractériser le mot analysé en vue d'une utilisation ultérieure.

L'analyse morphologique est une étape essentielle dans le traitement automatique de la langue arabe, elle est devenue aujourd'hui incontournable dans le développement des technologies de l'information. Plusieurs recherches sont nées pour offre meilleure performance et plus de fiabilité pour répondre aux besoins des différentes applications de traitement automatique de la langue arabe.

Dans le premier chapitre de ce projet, nous présentons le système de traitement automatique de la langue arabe, nous commençons par définir les systèmes de traitement automatiques des langues naturelles. Puis, nous donnons une description de la langue arabe. Nous terminons ce chapitre par les problèmes de traitement automatique de la langue arabe.

Dans le second chapitre, nous définissons la morphologie arabe. Ensuite, nous discutons les différents types de la morphologie arabe. Puis, nous présentons les propriétés morphologiques ainsi nous citons toutes les éléments de base de la morphologie de la langue arabe et une description schématique sur la morphologie, et finalement une étude sur les techniques d'analyse morphologique.

Dans le troisième chapitre, nous donnons une description sur les analyseurs morphologiques arabes connus.

Le dernier volet de ce projet est consacré à la réalisation d'un système d'analyse morphologique arabe.

Chapitre 1 TALN et la langue arabe

2.1 Introduction

Nos recherches portent sur l'étude de la langue arabe qui, malgré sa position de cinquième langue au monde[1] avec plus de 50000 sites arabes sur le web et plus de 320 millions de locuteurs, n'a encore aucun analyseur capable de traiter de façon robuste la totalité de ses phénomènes morphosyntaxiques. Par ailleurs, nous assistons à un accroissement des contenus textuels en arabe, surtout en ligne. À ce jour, le traitement et l'exploitation de ces ressources documentaires présentent encore un défi pour les chercheurs dans le domaine du traitement automatique des langues naturelles.

L'internet fait maintenant partie de notre vie quotidienne et il contient une quantité phénoménale d'informations qui sont très utiles. Aujourd'hui la langue arabe est parmi les langues les plus utilisées sur le Web.

Bien que ne donnant pas toujours des résultats totalement satisfaisants, certains aboutissent à des résultats intéressants et exploitables par le grand public, comme les logiciels de traduction automatique. Pour avoir un système de traduction d'une langue à une autre, mais également pour l'indexation ou l'extraction d'information, on peut avoir besoin d'un système d'étiquetage robuste et performant. C'est pourquoi beaucoup de recherches ont été menées sur cette tâche. Ces différentes recherches ont donné lieu à des propositions d'approche et à des algorithmes, et ont parfois débouché sur des applications.

D'un point de vue général, pour mettre en œuvre des outils du TALN en arabe, les chercheurs peuvent avoir besoin :

- Des modules de base pour la segmentation en phrases et en mots, l'analyse morphologique, syntaxique… etc.
- Des ressources linguistiques(dictionnaires, corpus, bases de données lexicales…).
- Des ressources et modules de comparaison pour l'évaluation.
- D'utilitaires de traitement de la langue (outils de recherche de texte, outils statistiques sur les textes et corpus annotés, etc.).

Parmi les « modules de base », l'étiquetage morphosyntaxique constitue une étape essentielle pour réaliser la plupart des applications en traitement automatique de la langue, car il permet d'identifier la catégorie grammaticale à laquelle appartiennent les mots du texte. Ainsi, les étiqueteurs constituent un module essentiel dans des applications de grand public telles que la correction grammaticale automatique, la génération automatique des résumés et le repérage d'information.

[1] Source : Ecole d'été en Linguistique par le biais de l'encyclopédie Enra en ligne(consultée le 3 juin 2007).

2.2 Traitement automatique des langues naturelles

Le traitement automatique des langues naturelles (TALN) est un domaine à la frontière de la linguistique et l`informatique, il a pour objectif de développer des logiciels capables de traiter de façon automatique des données linguistiques exprimées dans une langue naturelle donnée et pour une application bien définie. Cet objectif passe nécessairement par l'explicitation des règles de la langue puis les représente dans un formalisme calculable et enfin les implémenter à l`aide des programmes informatiques.

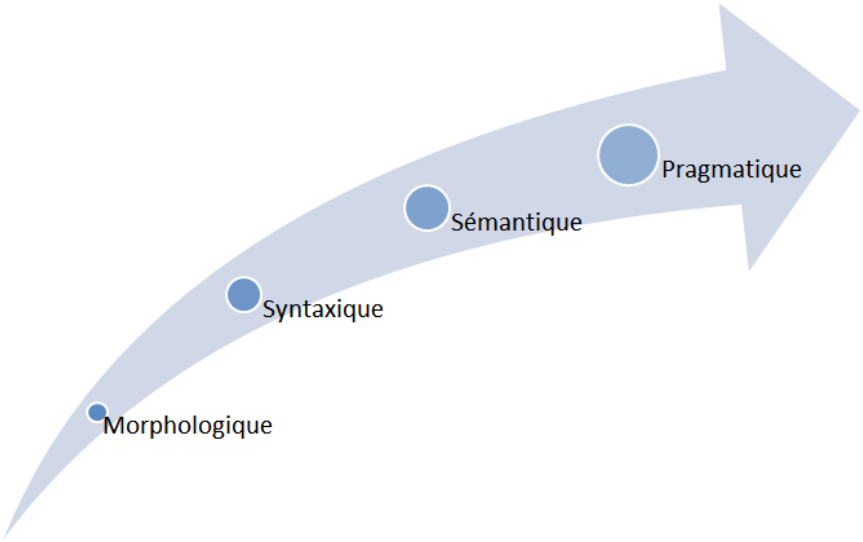

Figure 1: Les différents niveaux d'analyse

L'analyse des langues naturelles est une suite de traitements (morphologiques, syntaxiques, sémantiques…), elle consiste à construire une représentation formelle du texte en entrée, cette représentation doit être facile à manipuler par la machine. À partir des séquences de chaînes de caractères, différents niveaux d'analyses (traitement) peuvent être envisagés. On parle dans la littérature d'analyse morphologique, d'analyse syntaxique, d'analyse sémantique, et d'analyse pragmatique.

Le traitement automatique des langues se heurte à deux principales difficultés(El Amine Abderrahim, 2008):

- L'ambiguïté de la langue : elle concerne les différents types d'ambiguïté propres à chaque niveau d'analyse.
- La complexité des connaissances qui doivent être mises en œuvre à tous les niveaux d'analyse.

Dans ce qui suit, nous allons décrire brièvement les différents niveaux d'analyse d'un texte en langue naturelle.

2.2.1 Analyse morphologique

L'analyse morphologique est indispensable pour tout système de traitement automatique de la langue naturelle, cette analyse permet de regrouper les mots en classes

utilisables par les autres niveaux d'analyse. La définition de ces classes varie en fonction des traitements envisagés. À chaque classe on associe une étiquette appelée catégorie grammaticale ou catégorie lexicale. Il arrive qu'un même mot puisse avoir différentes catégories grammaticales, on dit qu'il y a ambiguïté grammaticale ou une homographie. L'analyse morphologique des langues comme le français ou l'anglais ne pose plus un problème et bons nombre d'analyseurs efficaces sont réalisés. L'analyseur morphologique du français proposé par (Lallich, 1990)en est un bon exemple. Ce dernier va nous servir comme base pour l'élaboration de notre analyseur pour la langue arabe, il comprend trois étapes :

- Préparer le texte en entrée à analyser : l'objectif de cette étape est de simplifier les phases ultérieures de l'analyse par la normalisation des caractères (par exemple le codage du texte à l'aide uniquement du code ASCII standard, la substitution d'une chaîne de caractères par une autre…) et le découpage du texte en formes.
  Ces prétraitements font partie du modèle linguistique, ils sont regroupés dans deux grandes classes : les prétraitements morphographiques (par exemple le traitement des majuscules, le traitement des ponctuations…) et les prétraitements morphosyntaxiques. Ces derniers sont basés sur l'application d'un ensemble de règles pour régulariser la surface du texte tout en amorçant l'analyse. Parmi ces règles on trouve par exemple celles de l'éclatement d'amalgames orthographiques, ou la suppression de formes.
- Chaque forme est traitée isolément par l'analyseur. Une ou plusieurs interprétationspossibles en terme de couple (entrée lexicale, catégorie) sont associées à la forme dans cette étape.
  L'analyse d'une forme dans (Lallich, 1990)revient à trouver tous les découpages base + (flexions) attestés. Les bases sont données par un dictionnaire, par contre les flexions sont données par une liste de flexions qui sont particularisées. Pour réaliser sa tâche, l'analyseur morphologique de (Lallich, 1990)a besoin d'un ensemble de données :
  - la chaîne à analyser (issue du prétraitement)
  - le dictionnaire
  - la liste des modèles des noms, adjectifs et des verbes
  - les régularisations de formes et de base
  - la liste des flexions et leur compatibilité.

À l'heure actuelle, la conception et la réalisation d'un analyseur morphologique pour une langue comme le français ou l'anglais sont très bien maîtrisées et les analyseurs produits sont jugés efficaces, mais malheureusement, pour la langue arabe les choses ne font que commencer et ce domaine pose encore des problèmes.

2.2.2 Analyse syntaxique

C'est une partie de la grammaire qui traite la manière dont les mots peuvent se combiner pour former des propositions et de l'enchaînement des propositions entre elles. Cela consiste à associer, à la chaîne découpée en unités, une représentation des groupements structurels entre ces unités ainsi que des relations fonctionnelles qui unissent les groupes d'unités.

Prenant l'exemple « يأكل عمر التفاحة » et sa représentation morphologique :

U1=يأكل U2=عمر U3=التفاحة

Le résultat de l'analyse syntaxique pourra être par exemple l'arbre suivant :

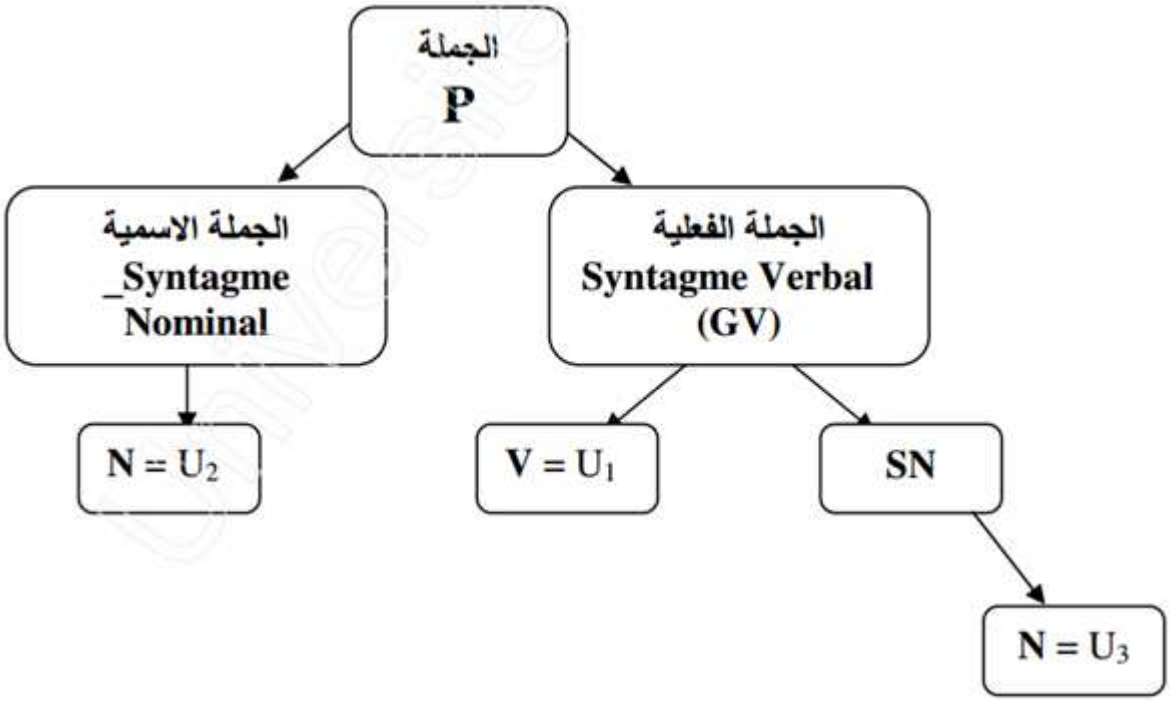

Figure 2:Arbre syntaxique

2.2.3 Analyse sémantique

Le niveau sémantique est encore beaucoup plus complexe à décrire et à formaliser que les niveaux précédemment énoncés. De ce fait, peu d'outils de traitement restentopérationnels ou du moins, concernent des applications très réduites où l'analyse sémantique se limite à un domaine parfaitement étroit ; par contre, il reste beaucoup à apprendre sur la manière de construire en grandeur réelle des analyseurs sémantiques généraux qui couvriraient la totalité de la langue arabe et seraient indépendants d'un domaine d'application particulier.

La phrase est l'unité d'analyse principale que prend en charge le traitement sémantique afin de représenter sa partie significative. Ces phrases, dont l'analyseur sémantique doit décrire le sens, se composent d'un certain nombre de mots identifiés par

l'analyse morphologique, et regroupés en structures par l'analyse syntaxique. Ces mots et ces structures constituent autant d'indices pour le calcul du sens : on pourrait dire que le sens résulte de la double donnée du sens des mots et du sens des relations entre ces mots.

2.2.4 Analyse pragmatique

La pragmatique concerne l'étude de l'environnement d'une phrase, au moment où elle est émise ; elle découle de l'idée qu'une phrase (un énoncé) ne peut prendre tout son sens que si on la (le) replace dans son milieu d'origine ; c'est la prise en compte de toutes les conditions de production d'une phrase, tant il est vrai qu'un acte linguistique effectif ne peut avoir lieu qu'à l'intérieur d'une certaine situation de communication.

Ce niveau d'analyse recouvre tout ce qui est lié à l'implicite dans la communication. C'est donc le niveau qui pose le plus de problèmes à concevoir et par conséquent il est beaucoup plus complexe à établir, ce qui explique qu'il n'existe que peu de réalisation opérationnelle, et qui ne concerne que des applications limites. On est donc encore loin de savoir construire des analyseurs pragmatiques pour le TALN.

2.2.5 Analyse et génération

La plupart des analyseurs et les travaux de la morphologie sont intéressés pour l'analyse et la génération de texte arabe à partir d'un lexique donné, les deux sont utilisés d'une manièrecontrastive.

2.2.5.1 Analyse

Dans l'analyse l'intérêt du système est d'analyser le texte et d'extraire et fournir les caractéristiques de chaque mot[étiquetage].

Exemple :

Mot	préfixe	radical	Suffixe	Schème	racine
الكَاتِبُ	ال	كَاتِبُ	_	فاعِلُ	كَتَبَ
فَارَقَهُم	_	فارَقَ	هُم	فاعَلَ	فرق
Entré	Résultat d'analyse				

Tableau 1: Exemple d'analyse

2.2.5.2 Génération

Contrairement à l'analyse, le but de génération est de construire un texte à partir d'un lexique donné.

Exemple :

Préfixe	radical	Suffixe	Schème	Racine	Mot
ال	كَاتِبُ	_	فاعِلُ	كَتَبَ	الكَاتِبُ
_	فارَقَ	هُم	فاعَلَ	فرق	فَارَقَهُم

Entrée de génération	résultat

Tableau 2: Exemple de génération

On voit souvent les deux systèmes lors de la traduction automatique d'un texte.

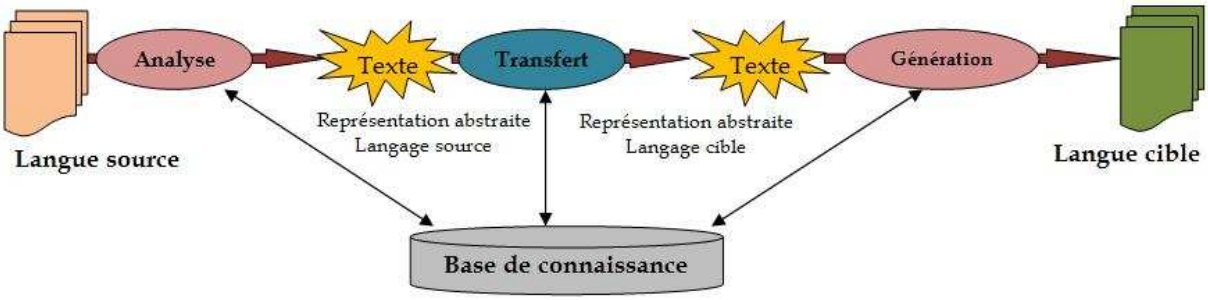

Figure 3: Utilisation d'analyse et génération

2.3 Langue arabe et TALN

Malgré de nombreuses recherches, la langue arabe est considérée comme une langue difficile à maitriser dans le domaine du traitement automatique de la langue à cause de sa richesse morphologique. L'arabe existe et se développe à partir du 7ème siècle grâce à la diffusion du Coran qui est considéré comme la base de cette langue. Avec la diffusion de la langue arabe sur le Web et la disponibilité des moyens de manipulation des textes arabes, les travaux de recherche ont abordé des problématiques variées comme la morphologie, la traduction automatique, l'indexation des documents, etc.

Les langues varient considérablement en complexité morphologique. L'anglais, par exemple, a une morphologie simple par rapport aux langues comme l'arabe et l'hébreu. Les langues européennes ont une morphologie plus complexe que ne le fait d'anglais.

Au cours de ce chapitre, nous présenterons les particularités de la langue arabe ainsi que certaines de ses propriétés morphologiques et syntaxiques.

1.1.1 Particularités de la langue arabe

La langue arabe est une langue orientale, sémitique (comme l'akkadien et l'hébreu)et est une langue très structurée et dérivationnelle où la morphologie joue un rôle très important.Au contraire de nombreuses autres langues, elle s'écrit et se lit de droite à gauche. Une autre originalité concerne l'utilisation facultative des voyelles. Les voyelles sont ajoutées au-dessus ou au-dessous des lettres, sous la forme des signes diacritiques. Elles sont utiles à la lecture et à la compréhension correcte d'un texte, car elles permettent de distinguer des mots ayant la même représentation graphique. Elles sont utiles, notamment, pour effectuer la correcte interprétation grammaticale d'un mot indépendamment de sa position dans la phrase. Le tableau suivant donne un exemple pour les mots s'écrivant «كَتَبَ »sous la forme non-voyelle.

Le mot sans voyelles	Interprétation 1	Interprétation 2	Interprétation 3
كتب	كَتَبَ	كُتِبَ	كُتُبٌ
ذهب	ذَهَبَ	ذُهِبَ	ذَهَبٌ
وقف	وَقَفَ	وَقْفٌ	وَقِفْ

Tableau 3: Les différentes interprétations possibles pour un mot donné

En général les voyelles ne sont utilisées que pour les textes sacrés et didactiques (comme les textes du Coran). En effet, la presse, la littérature et la plupart des textes écrits contemporains ne contiennent pas de voyelles. De plus, même pour des textes non-voyelles, il existe des variations d'usage au niveau des diacritiques. Par exemple, « أ »ou « إ », qui correspondent à des réalisations différentes de la lettre « ا », et qui sont en principe discriminé, sont souvent écrites ا sans les diacritiques. Il est de même pour les lettres « ي »(Y) et « ة »(T) (qui s'écrivent parfois ى(y) et ه(h), ce qui est une grande source (Xu, Fraser, & Weischedel, 2002)l'ambiguïté dans l'interprétation des mots.

Les voyelles jouent donc un rôle analogue à celui des accents en français, par exemple pour le mot « pêche » qui peut être interprété comme pêche, pèche ou péché. Mais, ces ambiguïtés sont démultipliées en arabe, car chaque lettre de chaque mot devrait posséder sa voyelle, ce qui augmente les combinaisons possibles.

En effet, une autre difficulté concernant le système d'écriture, les lettres arabes changent de forme de présentation selon leur position, au début, au milieu ou à la fin du mot, certaines lettres sont écrites différemment selon leurs positions dans le mot ou selon les lettres qui les entourent. De plus, la signification d'un mot n'est décelée que par son contexte, vu qu'un même mot peut avoir plusieurs sens (polysémie).

Mot	لَحَّام	مَلْحَمَة	لَحَمَ
racine	لحم	لحم	لحم
ullistrarion			

Tableau 4: Exemple de racine polysémie "لحم"

Le tableau suivant montre par exemple les variations de la lettre « ع ». Toutes les lettres sont liées entre elles sauf (ذ د ز ر و ا) qui ne se joint pas à gauche.

Lettre	À la fin	Au milieu	Au début
ع	ـع	ـعـ	عـ

Tableau 5: La variation de lettre "ع"

Mais ces variations n'affectent pas le traitement automatique, car elles ne touchent que les glyphes (représentations graphiques définies lors du rendu) et pas les caractères eux-mêmes dans le codage informatique des textes.

2.3.1 L'alphabet arabe

L'alphabet de la langue arabe compte 28 consonnes appelées « الحروف الهجائية » et se compose de deux familles contenant le même nombre de consonnes:

- Familles solaires : contiens 14 consonnes.
- Familles lunaires : contiens 14 consonnes.

Le tableau suivant représente la classification des consonnes arabe :

La famille des lettres solaires	La famille des lettres lunaires
ت ث د ذ ر ز س ش ص ض ط ظ ل ن	أ ب ج ح خ ع غ ف ق آ ه م و ي

Tableau 6: Classifications des Lettres arabes

La différence entre les deux familles a une relation avec la prononciation de la lettre "ل" d'article de définition "ال", où la lettre"ل" ne se prononce pas après les lettres solaires.

L'alphabet arabe compte 7 voyelles qui sont aussi divisées en deux groupes :

- Voyellescourtes : 4 voyelles(َ, ُ,ْ ,)ِ.
- Voyelles longues : 3 voyelles(ا,و)ي.

Par ailleurs, les diacritiques arabescontiennentd'autres types :

- ✓ Les diacritiques simples qui sont au nombre de quatre ◌َ [a], ◌ُ [u], ◌ِ [i] et ◌ْ[o], tous ces diacritiques se prononcent de la même façon que leurs translittérations sauf le dernier qui indique l'absence de tout son.
- ✓ Les diacritiques allongés() المَد : ◌ُو[O] ◌ِي[I] ◌َا[A], conforme aux lettres allongées en ce qu'il arrive au milieu et à la fin du mot, ils ne vient pas dans la première du mot.
- ✓ Les diacritiques doubles sont ◌ً [F], ◌ٌ[N] et ◌ٍ [K] : dites aussi « nunnation », il s'agit de diacritiques casuels, ils produisent, respectivement le même son que les trois premières voyelles simples avec l'ajout du son « n » à la fin. Exemple : ◌ً se prononce « an ».
- ✓ Le diacritique ◌ّ appelé «chadda», qui a pour effet la gémination d'une lettre à laquelle il est associé.
- ✓ Trois graphies supplémentaires :

- hamza (ء), noté, soit sous forme diacritique (plusieurs lettres pouvant lui servir de « support ») soit sur la ligne ;
- La lettre (ى) ;
- La lettre (ة).

2.3.2 Structure d'un mot

Les mots peuvent avoir une structure composée, résultat d'une agglutination de morphèmes lexicaux et grammaticaux. En arabe un mot peut représenter toute une proposition. La représentation suivante schématise une structure possible de mot complexe. Notons bien que la lecturese fait de droite à gauche.

Enclitique	Suffixe	Corps schématique	Préfixe	Proclitique

Tableau 7: Structure du mot arabe

- Les proclitiques sont des prépositions ou des conjonctions.
- Les préfixes et suffixes expriment des traits grammaticaux, tels que les fonctions de noms, le mode du verbe, le nombre, le genre, la personne…
- Les enclitiques sont des pronoms personnels.
- Le corps schématique représente la base de mot « radicale ».

Exemple :

«أَتتفكروننا » ce mot en arabe représente en français la phrase suivante : « Est-ce que vous vous souvenez de nous ? » ,la segmentation correcte de ce mot se fait sous la forme suivante :

Enclitique	Suffixe	Corpsschématiques	préfixe	proclitique
نا	ون	تفكّر	تـ	أ

Tableau 8: La segmentation du mot "أتتفكروننا"

- Proclitique : أconjonction d'interrogation.
- Préfixe : تَـ préfixe verbal exprimant l'aspect inaccompli.
- Corps schématique : تَفَكّر dérivé de la racine (فَ ك ر) selon le schème (تَفعل)
- Suffixe : ونَـ suffixe verbal exprimant le pluriel.
- Enclitique : نَـا pronom suffixe.

Cet exemple montre bien la richesse morphologique de la langue arabe. Pour identifier les différentes formes soudées par ces phénomènes d'agglutination, et envisager un traitement automatique, il va donc falloir mettre en œuvre une phase spécifique de segmentation.

2.3.3 Le principe de composition des mots arabes

Le vocabulaire de la langue arabe est essentiellement construit à partir de la dérivation desracines(khalfallah, 2008). Un mot arabe est construit à partir de sa racine

selon deux étapes,la dérivation et l'inflexion (El Amine Abderrahim, 2008),(Habash & Rambow, 2005).

La première consiste à appliquer un modèle précis sur la racine, ce quiproduit laradicale.Donc uneradicale = une racine + un modèle, cette radicale s'ais construit selon le schéma suivante

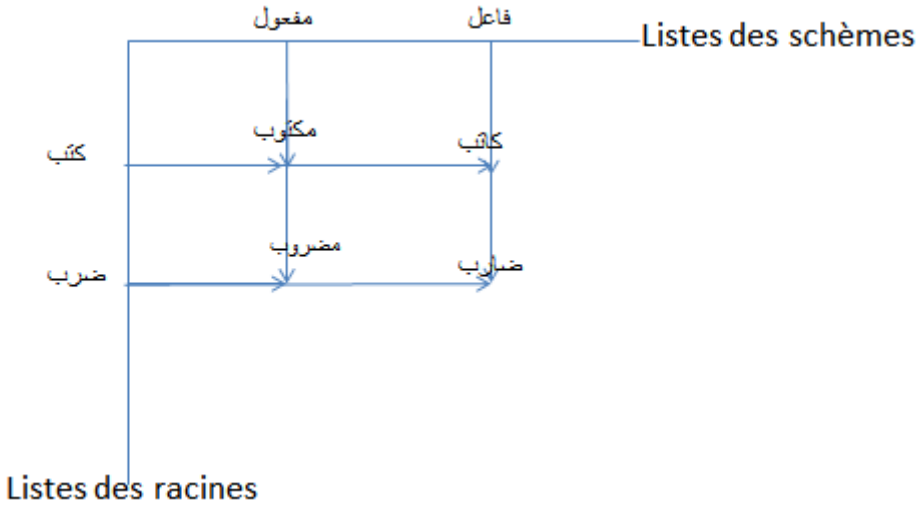

Figure 4: Le principe de composition des mots arabes

En général, le nombre des radicales de la langue arabe est limité (Habash & Rambow, 2005). Si le radical obtenu dans l'étape précédente estutilisable, la deuxième étape consiste à ajouter des affixes (préfixe ou suffixe) au stem ainsiobtenu, selon les caractéristiques du mot demandé (temps, nombre, personne, mode, etc.). La figure suivante présente les étapes de composition du mot arabe.

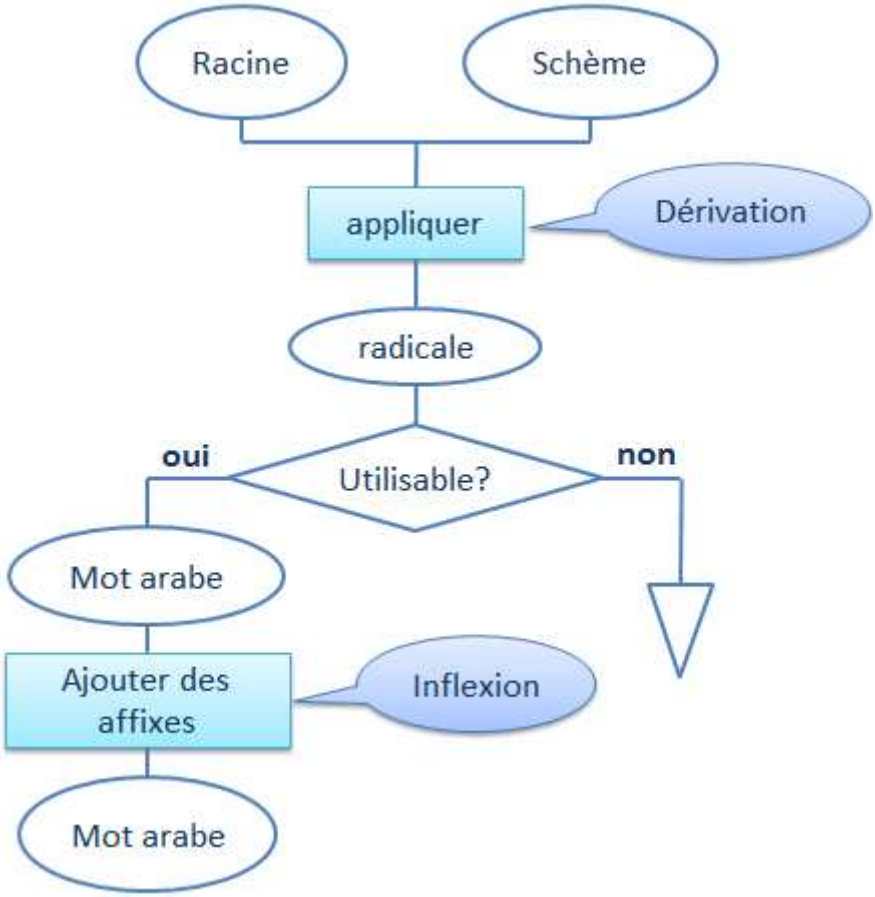

Figure 5: La composition du mot arabe

2.3.4 Les catégories d'un mot

En arabe, la majorité des mots dérivent d'un verbe de troislettres (qui peuvent être tous des consonnes "كتب"ou un mélange entre les consonnes et les voyelles "رمى")qui représente une racine d'un groupe de mots.

La classification des mots de la langue arabe trouvée dans la plupart des références de la linguistique arabe repose essentiellement sur la distinction entre trois catégories de mots : les noms « الأسماء », les verbes « الأفعال » et les particules « الحروف ».

2.3.4.1 Les verbes :

Un verbe est une entité qui exprime un sens dépendant du temps, c'est un élément fondamental auquel se rattachent directement ou indirectement les divers mots qui constituent l'ensemble.Autrement dit, le verbe est un mot qui se conjugue, et qui indique un état ou une action faite ou subie par le sujet.

Les verbes arabes sont classés selon plusieurs critères : selon le nombre et la nature des consonnes de leurs racines est selon leurs schèmes [El-dahdeh, 1999]. En ce qui concerne le nombre des consonnes de la racine, nous pouvons distinguer les verbes trilitères « الأفعال الثلاثية » qui ont trois consonnes et les verbesquadrilatères « الأفعال الرباعية » qui possèdent quatre consonnes.

a. *La nature des consonnes*

Selon la nature des consonnes, nous distinguons entre deux types des verbes

Les verbes sains « الافعال الصحيحة »:qui ne sont pas formés par des lettres défectueuses, ils sont divisés en trois types :

- **Hamzé(مهموز) :** qui contient hamza(ء) parmi les radicales « أكل ».
- **Sourd(مُضَعَّف) :** contient le même consomme dans le deuxième et la troisième consonne « شدَّ »
- **régulière - سالِم :** qui est ni Hamzé ni sourd « كتب ».

Les verbes défectueux[2]« الأفعال المعتلة » : qui contiennent une ou deux lettres défectueuses :

- **Assimilé مثال** : contient la lettre défectueuse au début du verbe « وصل ».
- **Concave أجوف** : contient la lettre défectueuse au milieu du verbe « قال ».
- **Défectueuse ناقص**: Contient la lettre défectueuse à la fin du verbe « رمى».
- **Repliéeséparé لفيف مفروق**: est le verbe qui est assimilé et défectueuxau même temps comme « وفى ».
- **Repliée groupée لفيف مقرون** : est le verbe concave et défectueuxau même temps comme « طوى ».

Ces verbes causent des altérations importantes dans la conjugaison.

b. *Le nombre de consonnes*

Selon le schème et le nombre de consonnes qui constituent la structure verbale, il y a deux types :

Les verbes nus[3] ***« الأفعال المجردة » :*** sont formés seulement par les consonnes de leurs racines et des voyelles courtes, c'est-à-dire toutes les lettres sont originales et on ne peut pas élimer l'une de ses consonnes, il être soit trilatérale ou quadrilatérale.

Verbes trilatéraux	**كتب,لعب,عاش**
Verbes quadrilatérales	دحرج, زلزل

Table 1: Exemple des verbes nus

Les verbes augmentés « الأفعال المزيدة »éventuellement quatre consonnes tellessont le cas du verbe « أخرج »(faire sortir). Les grammairiens de la langue arabe ont confirmé que les lettres qu'ils peuvent s'ajouter sont dix lettres (ا,ه,ي,ن,و,م,ت,ل,ء,س), ces racines peuvent donner naissance à plusieurs schèmes ou modèles à la suite d'une ou plusieurs transformations morphologiques comme :

- Le redoublement d'une consonne« مَدَّدَ ».

[2] Appelées aussi les verbes faible
[3] Appelées dans des autres livres aussi « simples »

- L'allongement d'une voyelle «سَابَقَ ».
- L'adjonction d'un morphème «أوصل ».

Comme les autres langues, la conjugaison des verbes en arabe dépend des facteurs suivants :

- L'aspect: accompli (passé) ou inaccompli (présent).
- Le nombre du sujet : singulier, pluriel ou duel.
- Le genre du sujet : masculin ou féminin.
- La personne : première, deuxième ou troisième.
- La voix : active ou passive.

Comme on l'a dit précédemment, la conjugaison des verbes se fait en rajoutantdes préfixes, des suffixes ou les deux.

La langue arabe possède trois temps :

- **L'accompli :** indique le passé et les verbes conjugués se distinguent par des suffixes. Pour notre exemple, avec le féminin pluriel, on obtient فتحنَ«elles ont ouvert» ;pour le masculin pluriel, on obtient فَتحوا «ouvert ont ils ».
- **L'inaccompli présent :** les verbes conjugués à ce temps se distinguent par les préfixes. Pour notre exemple, au masculin singulier on obtient يَفتح « il ouvre ; et pour le féminin singulier, on obtient تَفتح « ouvre elle ».
- **L'inaccompli futur :** la conjugaison d'un verbe au futur nécessite d'ajouter l'antéposition au début du verbe conjugué à l'inaccompli. En ajoutant l'antéposition" à"س notre exemple on obtient « سَيفتح»(il ouvrira) , qui désigne le futur ; on peut également ajouter l'antéposition سَوف on obtient سَوف يَفتح « il va ouvrir ».

Et il y a un verbe particulier qui s'appelle un verbe d'état « لَيْسَ » qui est un verbe "figé". Il n'existe qu'à un seul temps. Il veut dire "ne pas être" au présent. Son sujet est au nominatif et son attribut est au cas direct. Les verbes suivantsباتَ - أَصْبَحَ - كانَ - peuvent être suivis d'un attribut au cas direct, ou d'un verbe à l'inaccompli indicatif.

2.3.4.2 Les noms :

Comme dans les autres langues, les noms dans la linguistique arabe sont des entités qui nomment une personne, un lieu, une chose ou un concept. Le système morphologique des noms de la langue arabe distingue, d'une part, deux classes de noms selon la possibilité de changement de leur forme graphique en fonction de la valeur grammatical : les noms structurés «الأسماء المبنية »et les noms déclinés « الأسماء المعربة » . D'autre part, il distingue deux autres classes fondamentales de noms selon certaines contraintes qui concernent la morphologie du radical : les noms dérivés « الأسماء المشتقة » et les noms particuliers « الخاصةالأسماء». Dans ce travail nous avons choisi de grouper les noms selon la deuxième classe.

La catégorie des noms dérivés regroupe tous les noms qui sont obtenus en employant les règles de dérivations. Un nom dérivé se caractérise complètement par sa représentation morphologique racine/schème. La seconde catégorie regroupe tous les noms particuliers qui ne respectent aucune règle de dérivation.

Les noms particulières « الأسماء الخاصة » : Ce sont la catégorie des noms qui ne peuvent pas être rattachés à une racine verbale. Cette catégorie des noms constitue un glossaire important de la langue concrète. Notons que cette catégorie peut regrouper les noms suivants :

- ✓ Les noms exclusifs «أسماء الاستثناء»
- ✓ Les noms d'interrogations«أسماء الاستفهام »
- ✓ Les noms de démonstration «أسماء الإشارة »
- ✓ Les noms d'annexion «أسماء الإضافة »
- ✓ Les noms confirmatifs «أسماء التوكيد »
- ✓ Les noms conditionnels «أسماء الشرط»
- ✓ Les noms personnels libres «الضمائرالمنفصلة »
- ✓ Les noms numéraux «أسماء العد »
- ✓ Les nomsconjonctifs «أسماء الموصول»

Les noms dérivés « الأسماء المشتقة » :C'est la catégorie des noms pouvant être dérivés d'une racine verbale. Le nombre et la nature de ces formes varient selon le statut du verbe auquel ils se rattachent. Puisque ce sont des noms, alors ils peuvent recevoir les marques du genre, du nombre, du cas, etc. Cette classe des noms peut contenir :

- ✓ Les noms d'agent «أسماء الفاعل »
- ✓ Les noms de patient «أسماء المفعول »
- ✓ Les noms superlatifs «أسماء المبالغة »
- ✓ Les noms originels «المصدر »
- ✓ Les noms de temps «أسماء الزمان »
- ✓ Les noms de lieu «أسماء المكان »
- ✓ Les noms d'outil «اسم الآلة »
- ✓ Les noms adjectifs «الصفة »

Notons que les linguistes arabes ont assimilé les adjectifs à des noms vus qu'ils y prennent presque tous les descripteurs morphologiques. Ils peuvent être définis ou indéfinis et se caractérisent par les propriétés morphologiques telles que le nombre, le cas, le genre, etc.

2.3.4.3 Les particules

En général, les particules sont les mots outils pour une langue donnée. En arabe, les particules sont classées selon leur fonction dans la phrase, on en distingue plusieurs types (introduction, explication, conséquence). Elles jouent un rôle important dans l'interprétation de la phrase(Kadri & Benyamina, 1992).

Les particules représentent en particulier les mots qui expriment des faits ou des choses par rapport au temps ou au lieu. Par exemple :

- Particules temporelles : منذُ (pendant), قَبل (avant), بَعد (après),…
- Particules spatiales : حَيث (où) les particules peuvent aussi exprimer des pronoms relatifs (la détermination, avec une valeur référentielle) : الَّذي (ce), الَّذين (ceux), الَّتي (cette),…
- Particules d'affirmation : exemple (بلى أجل, نعم)
- Particules de négations : exemple لن,لالم,
- Particules distinctives : exemples أي
- Prépositions: exemplesعن,ل,ك,ب
- Particules conditionnelles : exemple إنَّ ,لو
- Particules futures : exemples لن ,سوف , سَ
- Particules de coordination : exemplesأم,ثمّ,و,ف
- Particules interrogatives : exemples أ ,هل , ما
- Particules relatives : exemples مَا

Le problème, c'est que certaines particules peuvent également porter des préfixes et suffixes, ce qui complique la phase de segmentation pour les identifier.

a. *Les préfixes*

Les préfixes sont représentés par un morphème correspondant à une seule lettre en début de mot, qui indique la personne de la conjugaison des verbes au présent. Les préfixes ne se combinent pas entre eux. Le tableau suivant présente la liste des préfixes verbaux en arabe :

Le préfixe	Signification
أ	Indique la première personne au singulier (je)
نَ	Indique la première personne au pluriel (nous)
تَـ	Indique la deuxième personne féminine, masculine, singulière et duelle
يَـ	Indique la troisième personne masculine au singulier, duel, pluriel, masculin et féminin pluriel.

Figure 6: Liste des préfixes arabe

b. *Les suffixes*

Les suffixes en arabe sont essentiellement utilisés pour des terminaisons des conjugaisons verbales, ainsi que les marques du pluriel et du féminin pour les noms.

c. *Les proclitiques*

Les proclitiques sont en inventaire finis, et se combinent entre eux pour donner les traits syntaxiques (coordonnant, déterminant ...) qui peuvent accompagner le mot arabe. Ils se collent au « mot minimal ›› (forme fléchie) en tant que préfixes.

Dans le cas des verbes, les proclitiques dépendent exclusivement de l`aspect verbal. Dans le cas des noms et les déverbaux, le proclitique dépend du mode et du cas de déclinaison. Dans notre travail, nous n'allons pas retenir l'intégralité de ces clitiques, parce que certains cas sont vraiment rares et d'autres théoriquement possibles, mais jamais n'utilises dans la langue. Dans la liste des proclitiques pris en compte dans notre étude, nous distinguons trois types :

Les proclitiques réservés aux noms et adjectifs :

- ✓ L'article de définition « ال »(le).
- ✓ Les prépositions : «بِـ», «لِ » , «كَ ».

Les proclitiques réservés aux verbes :

- ✓ La particule du subjonctif « نصب» :
- ✓ La particule de la future : " سَ"
- ✓ La particule d'occupation « جزم» : ل

Les proclitiques généraux, utilisés indépendamment de la catégorie des mots auxquelles ils s'attachent :

- ✓ La conjonction de coordination : «فَـ» et « وَ »
- ✓ L'article d'interrogation : « أ»
- ✓ Le marqueur de corroboration « تأكيد» : «ل »

La restriction de la prise en compte à la dizaine de proclitiques cités ci-dessus n`est pas sansdésagrément puisque le même clitique peut jouer plusieurs rôles. En l`occurrence, le proclitique (وَ) utilisé majoritairement comme particule de liaison(conjonction de coordination), peut êtreemployé en tant que particule d`accompagnement ""واو المعيةet, à de moindresusages, en tant que particule de serment "واو القسم" (cas rare).

Les proclitiques peuvent se combiner entre eux et forment ainsi un proclitique composé. Uneanalyse robuste doit permettre l'identification de la catégorie syntaxique de chacun de cescomposants d`une telle composition. Ce découpage est particulièrement utile, voire indispensable,pour les analyses syntaxiques. Les proclitiques sont classés en quatre catégories suivant lapossibilité de leur apparition dans un proclitique composé :

1re position	2e positon	3e position	4e position
L'article d'interrogation"أ"	Les conjonctions de coordination"ف" et "و"	*Les propositionsc: "بِـ" et "لِ" et "كَ" ; *La particule dessurjections « نصب » : "لَ"; *La particule de l'apocope « جزم » : "لِ"	L'article de définition "الْ"

Tableau 9:Agencement possible des proclitiques

La fusion de l'ensemble de proclitiques est régie par deux types de contraintes :

- ✓ **Une relation d'ordre :** chaqueproclitique est incompatible, dans une relation d'ordre strict, avec un proclitique de même position.De même, un proclitique, qui occupe une position d'antériorité par rapport à un autre proclitique sur la classification, n'a aucune chance de le suivre dans la construction d'un lot graphique arabe.
- ✓ **Des règles de compatibilité :** Les proclitiques présentant la relation d'ordre ne sont pas forcément compatibles entre eux pour des raisons d'ordre syntaxique et sémantique. À ce propos, nous signalons que la combinaison de proclitiques appartenant à quatre positions différentes est très peu fréquente. Par exemple, une construction telle que « أَفَبِالبَيْتِ» décomposable en "أ + فَ+بِـ+الـ+بَيْتِ" est rarement utilisée dans les textes courants.

Outre les règles de compatibilité entre proclitiques, d'autres règles s'imposent pour vérifier la compatibilité du proclitique lui-même avec les bases voire même les enclitiques qui s'y rattachent.

d. Les enclitiques

Ils sont en inventaire finis, leurs emplois respectent certaines restrictions.
Contrairement aux proclitiques, nous ne pouvons avoir qu`un seul enclitique à la fois dans un mot graphique.

Dans le cas d'une famille verbale, le lien entre celle-ci et l`enclitique dépend de sa propriété detransitivité. Nous distinguons principalement les enclitiques compatibles avec les verbes transitifsaux humains uniquement et les enclitiques compatibles avec les verbes transitifs aux humains etaux non humains. Les verbes intransitifs et ceux conjugués au passif ne peuvent prendre aucunenclitique.

Contrairement aux enclitiques à la première personne tels que "ني" (moi/mon) ou "نا" (nous/notre) et ceux a la deuxième personne tels que "كَ" (toi/ton) ou "كُم" (vous/votre [masculin pluriel]) dont la forme reste invariable indépendamment des propriétés de la forme à laquelle ils se rattachent, les enclitiques à la troisième personne sont variables. En effet, ils prennent différentes vocalisations suivant les règles suivantes :

- Dans le cas des noms, seul le mode déterminé par annexion est susceptible de prendre des enclitiques. Selon la flexion casuelle de la base nominale, cette dernière prend l'un ou l'autre enclitique :
  - o Si le nom est fléchi au nominatif ou accusatif, il nécessite l'utilisation des enclitiques suivants : هُ[PRON+3+m+s], هُما[PRON+3+m|f+d], هُم [PRON+3 + m+Pp] etهُنَّ[PRON+3+f+p].
  - o Si le nom est fléchi au génitif, il nécessite l'utilisation des enclitiques suivants: هِ[PR0N+3+m+s],هِما[PRON+3+m|f+Dd], هِمْ[PRON+3+m+p], هِنَّ[PRON+3+f+p].

Dans le même contexte, nous signalons que les mots qui se terminent par un hamza, une « ى » ou une « ي »nécessitent des transformations morphologiques avant leurs suffixations. Par exemple, la forme "مقهى » (café, salon de thé), qui se termine par la lettre «ى», nécessite une transformation de celle-ci en "ا" avant sasuffixation pour produire la forme agglutinée "مَقهاهُ"(son café, son salon de thé).

- Quant aux verbes, l'enclitique peut varier en fonction de l'aspect et du pronom. La répartitionde l'utilisation de ces enclitiques selon les aspects est la suivante :
    - Si le verbe est conjugué à la voix active, inaccomplie, à la 2eme personne, féminin, singulier,il prend les pronoms qui portent la marque du nominatif tel que : هُ,هُما,هُنَّ
    - Si le verbe est conjugué à la voix active, inaccomplie, à la 2e personne, masculin ou féminin duel, ou 3eme personne, masculin ou féminin duel, il prend les pronoms qui portent la marque du génitif tel que : هِنَّ,هِمْ,هِما,هِ
    - Si le verbe est conjugué à l'accompli actif, accompli passif, inaccompli subjonctif actif, inaccompli apocopeactif ou impératif, à la 2e personne, féminin singulier, il prend lespronoms portant la marque du génitif : هِنَّ,هِمْ,هِما,هِ
    - Si le verbe est conjugué à l'accompli actif, accompli passif, Inaccompli Subjonctif actif, inaccompli apocopeactive, impérative ou future, 2e personne, masculin ou féminin duel, ou3ème personne, masculin ou féminin duel, il prend les pronoms portant la marque dunominatiftel que : هُنَّ, هُم, هُما,هُ
    - Le cas échéant, si le verbe est conjugué à la voix passive, il ne prend jamais d'enclitique.

e. Les prébases

Les prébases sont obtenues par combinaison entre le(s) proclitique(s) et le préfixe. Lagénération des prébases se fait d'une manière automatique. Le tableau suivant représenteun exemple de groupe de prébase :

Prébase	Préfixe	Proclitique
أَتَـ	تَـ	أَ
سَتَـ	تَـ	سَـ
أَفَتَـ	تَـ	أَفَـ
أَسَتَـ	تَـ	أَسَـ
وَسَتَـ	تَـ	وَسَـ
فَسَتَـ	تَـ	فَسَـ
أَفَسَتَـ	تَـ	أَفَسَـ

Tableau 10:Exemple de groupe de prébase

f. *Les post-bases*

En arabe, les post-bases sont obtenues par combinaison entre le suffixe et le(s) enclitique(s). Les compatibilités dépendent des pronoms décrits par chacune des particules:

- Les suffixes de la première personne se combinent très souvent avec les enclitiques de la deuxième et la troisième personne.
- Les suffixes de la deuxième personne se combinent très souvent avec les enclitiques de la première et la troisième personne.
- Les suffixes de la troisième personne se combinent très souvent avec les enclitiques de la première, la deuxième personne et la troisième personne.

Enfin, il existe en arabe des suffixes qui jouent le rôle de caractère terminal du mot. En effet ces types des suffixes ne se combinent avec aucun enclitique(s). Le tableau suivant présente un exemple de groupe de post-bases :

Post-base	Enclitique	Suffixe
وك	ك	و
وهم	هم	و
وننا	نا	ونَ
ونني	ني	ونَ
أنكم	كَم	أنَ
أنهم	هم	أن
تموه	ه	تمو

Tableau 11: Exemple de poste-base

Ces deux derniers éléments représentent l'ensemble des particules qui fait partie de début et la fin du mot, le schéma suivantprésentent l'emplacement de chaque composant du mot arabe.

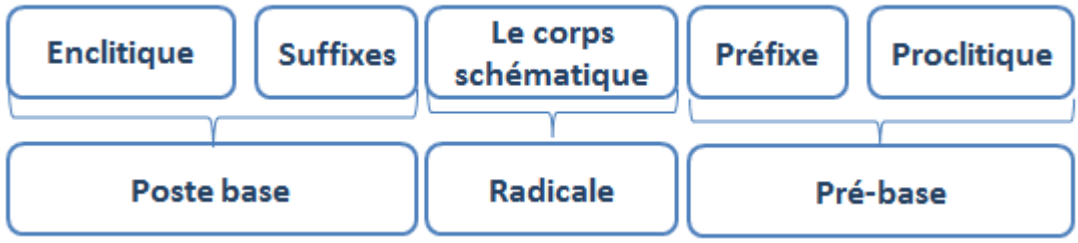

Figure 7: schéma présentatif de la structure générale d'un mot arabe.

2.4 Conclusion

Après avoir situé notre étude dans la chaîne du traitement automatique d'une langue (à savoir le niveau morphologique), nous avons dégagé les principales étapes et recommandations pour élaborer avec succès un système pour le TALN. Les attentes les plus

prégnantes d'un tel système étant : la grandeur réelle, la performance et la robustesse. Dans le chapitre suivant, nous commençons par l'explicitation des connaissances externes (connaissances liées à la langue arabe). Dans cette perspective nous présentons un modèle pour le mot graphique arabe et nous discutons les problèmes du traitement automatique de ce dernier.

Le traitement automatique de langue arabe souffre de plusieurs problèmes parmi eux:

- la voyellation multiple
- la structure complexe du mot graphique arabe (phénomènes d'agglutinations qui caractérisent la langue ; un mot graphique arabe peut correspondre à une phrase française).
- L'ordre des mots est relativement libre dans une phrase arabe (Verbe + Sujet + Complément ; verbe + Complément + Sujet ; complément + Verbe + Sujet).
- Traitement des cas particuliers : traitement de la 'hamza'', traitement de la 'Shedda', traitement de l'altération de la forme du mot.
- Traitement des racines analogues ne donnant pas lieu à des dérivations analogues.
- Traitement de la racine d'un mot issu d'une racine anormale.
- Traitement des mots homographes (une même chaîne de caractère qui suivant le contexte recouvre deux notions différentes) exemple : le mot verbe impératif ou préposition.

Chapitre 2 : Morphologie de la langue arabe

3.1 Introduction

La langue arabe est la langue principale de tous les pays arabes, elle est parmi les langues les plus anciennes connues dans le monde entier qui a constitué un facteur très puissant dans le développement de l'humanité. Dans ce travail, l'arabe standard sera notre objectif : c'est la langue de communication officielle ; en outre c'est la langue essentiellement écrite dans la littérature et dans la presse, parlée ordinairement à la radio et utilisée dans les discours officiels, les cours et les conférences universitaires dans tous les domaines scientifiques, administratifs, techniques, etc.

Ce chapitre sera consacré à l'étude de la morphologie de l'arabe standard, rappelons que la morphologie est la branche de la linguistique qui consiste en l'étude de la structure interne des mots. Un mot peut être décomposé en unités morphologiques. C'est-à-dire en unités de sens appelées des morphèmes. Dans ce chapitre tel qu'elle est présentée par les grammairiens arabes.

Nous commençons par données une définition de la morphologie arabe. Puis, les types de la morphologie arabe. Ensuite, nous présentons ses éléments essentiels de la morphologie arabe

3.2 La morphologie

La morphologie est un domaine de la langue naturelle qui permet la description des règles régissant la structure interne des mots (appelé unité lexicale), chez un grammairien la morphologie est l'étude des mots (flexions et dérivation), en d'autres termes, la morphologie est l'étude des mots considérés isolement sous le double aspect de la nature et les variations qu'ils peuvent subir. Donc, le principal objectif d'une analyse morphologique est de reconnaitre ces unités et d'attribuer à chacune divers types d'informations telles que la catégorie grammaticale (verbe, adjectif…) et les traits morphologiques (genre, la voix, le mode…,etc.).

À chaque classe on associe une étiquette appelée catégorie grammaticale ou catégorie lexicale. Il arrive qu'un même mot puisse avoir différentes catégories grammaticales, on dit qu'il y a ambiguïté grammaticale ou une homographie.

L'analyse morphologique est indispensable pour tout système de traitement automatique de la langue naturelle, cette analyse permet de regrouper les mots en classes utilisables par les autres niveaux d'analyse.

La différence principale de la langue arabe et les autres langues est que la langue arabe est dérivationnelle contrairement aux autres langues qui on flexionnelle, la figure suivante présente le système de dérivation arabe :

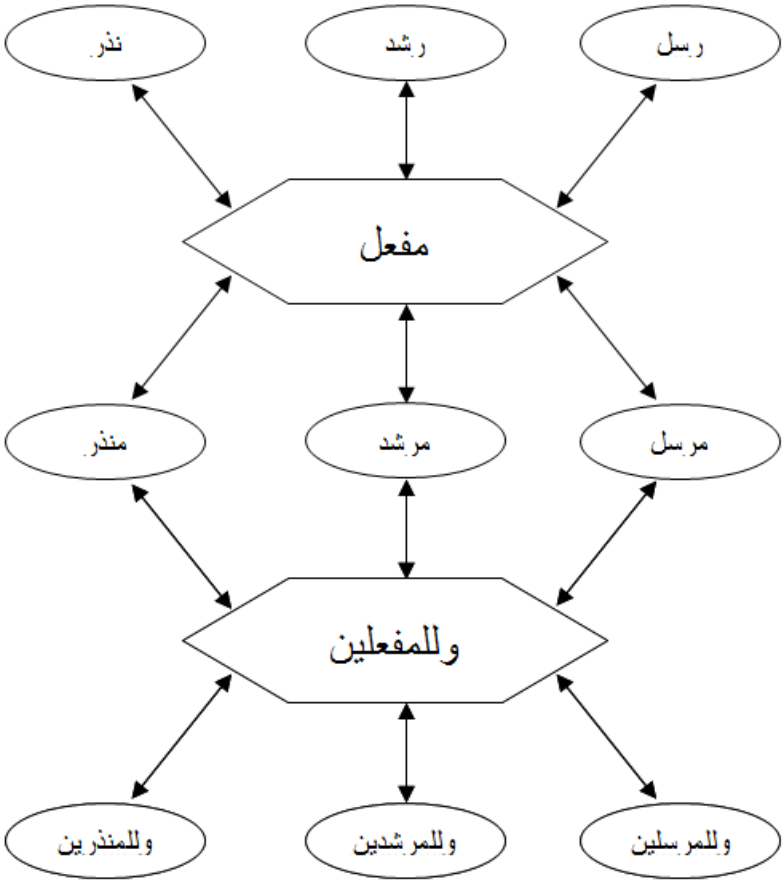

Figure 8: Le système de dérivation arabe et flexionnel

3.3 Morphologie flexionnelle

La langue arabe est aussi une langue flexionnelle, ce type de morphologie est employé, principalement, pour la déclinaison des noms et la conjugaison des verbes, des indices d'aspect, de temps, de mode, Personne, de nombre, de cas, de mode, etc., qui sont présentés généralement sous forme de préfixes et suffixes (Mesfar, 2008).

- Le mode des verbes : par exemple, pour le verbe " ذهَبَ " (aller), les formes à l'accompli sont repérables à l'aide de leurs suffixes tel que "ذهَبَتُ" (je suis allé) ou de leurs préfixations telles que "أذهَبُ" (je vais) ;
- La fonction des noms à l'aide des suffixations tels que "رَجلانِ" (deux hommes au nominatif) ou "رجلين" (deux hommes à l'accusatif ou génitif).

3.3.1 Flexion des verbes

La conjugaison des verbes décrit la variation de leurs formes en fonction des circonstances.Généralement, la conjugaison regroupe un certain nombre de valeurs dont :

- **La valeur aspectuelle:** L'aspect est un trait grammatical associé, le plus souvent, au verbepour indiquer la façon dont le procès ou l'état exprimé par le verbe est envisagé du point de vue de son développement (commencement, déroulement, achèvement, évolution globale, etc.), indépendamment du moment où l'on parle ;

- **La valeur modale :** Le mode dénote la manière dont l'action exprimée par le verbe est conçue et présentée. L'action peut être mise en doute, affirmée comme réelle ou éventuelle. Ils se combinent à la sémantique des verbes et par là créent les aspects ;
- **La valeur temporelle :** Le temps est un trait grammatical permettant de situer un fait (qui peut être un état ou une action) dans l'axe du temps de l'énonciation par rapport à trois repères : passé, présent et le futur. Les indications temporelles sont souvent accompagnées d'indications aspectuelles qui lui sont plus ou moins liées.

Ces trois principales valeurs sont étroitement liées (CHAIRET & Mohamed, 1996); elles permettent de décrire deuxformes fondamentales du verbe :

- ✓ **L'accompli " الماضي ":** il indique que le déroulement de l'action exprimée par le verbe est achevé, ce qui implique le passé. Il se caractérise par une suffixation des marques de la personne, du genre, du nombre et du mode à la racine verbale. Par exemple, pour le pluriel féminin du verbe "كَتَبَ " (écrire), nous ajoutons le suffixe "نَ " pour avoir la forme « كَتَبْنَ » (elles ont écrit) et pour le pluriel masculin, nous ajoutons le suffixe "وا " (elles ont écrit) et pour le pluriel masculin, nous ajoutons le suffixe " " pour avoir la forme " كَتَبوا " (ils ont écrit) ;
- ✓ **L'inaccompli « المضارع » :**il signale un déroulement inachevé, ce qui peut impliquer le présent. Il se caractérise par une préfixation de ses éléments ainsi qu'une ou plusieurs infixations sous forme de duplication de lettres ou de substitution de voyelles. Par exemple, pour le verbe "مَدَّ" (tendre), nous pouvons obtenir "أَمُدُّ" (je tends) ou « يُمْدِدْنَ». (elles tendent).L'inaccompli inclut deux types de flexions modales :
  - o L'inaccompli indicatif de mode réel où le locuteur énonce le caractère réel (relise, devant être réalisé, en cours de réalisation, etc.) de l'action ou l'état exprimé par le verbe ;
  - o L'inaccompli subjonctif et apocope de mode potentiel où le locuteur se contente d'énoncer la nature possible ou virtuelle de l'action ou l'état exprimé par le verbe.

Dans la littérature, on convient d'ajouter un paradigme supplémentaire qui est :

- ✓ **L'impératif** : il exprime l'ordre, le commandement, la défense ou l'exhortation et dont les éléments n'existent qu'à la 2ème personne au singulier, féminin duel et pluriel ;

Les formes ainsi obtenues peuvent combiner des valeurs aspectuelles, modales et temporelles bien que, dans l'usage moderne, l'aspect temporel semble être plus saillant. À l'exception des verbes sains dont la conjugaison est régulière et suit des règles flexionnelles bien définies, tous les autres types de verbes nécessitent un traitement particulier selon le type de défectuosité.

3.3.2 Flexion des noms

En arabe, la déclinaison des noms comporte trois cas : "مَرْفوع"(nominatif), "مَنْصوب" (accusatif) et " مَجْرورٌ" (génitif). À l'exception de certains cas particuliers, les noms sont

"مُعْرَبةٌ"(déclinables) et se met à l'un de ces trois cas suivants leur fonction dans la phrase. Sur le plan de la graphie, le cas ne correspond qu'à un élément graphique adjoint à la fin des formes nominales.

Le système nominal de l'arabe admet différents systèmes de déclinaison suivant la nature de la forme (simple, diptotes, etc.) et le nombre de celle-ci (singulier, duel ou pluriel). Nous pouvons distinguer :

- ✓ **Déclinaison du nom au singulier :**
  - o Déclinaison de base à trois cas : C'est le cas le plus fréquent, il prend la voyelle "ضَمَّةٌ" (ُ) comme une marque du nominatif, la "فَتْحَةٌ" (َ) à l'accusatif et la "كَسْرَةٌ" (ِ) au génitif. Quand le nom est indéfini, la nunnation(التنوين) apparaît marquée respectivement par les trois signes diacritiques ٍ ٌ ً" " .A l'accusatif indéfini, excepté le cas des noms qui se terminent par " ة " ou par " اء " , la lettre« ا »vient renforcer la nunnation(ً)par exemple, à l'accusatif indéfini, le nom "كِتاب" (un livre) produit "كِتابًا " (un livre à l'accusatif indéfini) et le nom "جَزيرة" (île) produit "جَزيرةً " (île à l'accusatif indéfini).
  - o Déclinaison des diptotes : Les diptotes sont les noms qui, indéfinis grammaticalement, n'acceptent pas de nunnation et prennent la même marque à l'accusatif et le génitif, soit la"فَتْحَةٌ"(َ). Par contre, quand ils sont définis, ils suivent la déclinaison de base à trois cas. C'est le cas des noms féminins qui se terminent par " اء "tel que "صحراء" (désert), les adjectifs masculins de couleur ayant pour schème "أفْعل"tel que "أحْمر" (rouge) et ceux qui sont féminins de schème "فَعْلاء" tel que "بَيْضاءٌ" (blanche)
  - o Déclinaison des cinq noms : Ce sont des noms bilitères qui prolongent leur voyelle finale quand ils sont définis par un complément, les cinq noms sont :
    - les 3 noms : "أبو"(père), "أخو"(frère) et "حمو" (beau-père) ;
    - une variante de "فم" (bouche) : "فو", "فا" et "في" ;
    - le nom "ذو" (possesseur).
  - o Déclinaison de déverbaux de racines défectueuses : Certains participes actifs et noms verbaux des verbes à racine défectueuse tels que le participe actif "ماضٍ" (passé) et le nom verbal "تَخَلٍ" (abandon) ne prennent la marque du cas qu'à l'accusatif: le "ي" (dernière lettre de la racine) est remplacé par la nunnation(ٍ) aux nominatif et génitif indéfini. Quant aux participes passifs qui se terminent par "ى" ou "ا" tel que "مُعْطى" (donné), ils perdent leur flexion casuelle. Une nunnation différencie le nom indéfini du nom défini. À ce niveau, il nous importe de signaler que l'usage de cette règle de déclinaison est abandonné. En effet, dans les textes courants la forme du nom verbal "قاضٍ" (avocat) est, généralement, altérée en "قاضي" (avocat) par adjonction du glide "ي" à la fin de la forme initiale.
  - o Déclinaison du nom au duel : Il existe en arabe "المثنّى" (le duel) pour désigner deux choses ou deux personnes. Il prend la place entre le singulier (pour

désigner une chose ou une personne) et le pluriel (à partir de trois choses ou trois personnes).

Il s'agit d'une déclinaison avec deux alternatives où la marque du nominatif est le "ا", et celle de l'accusatif et le génitif est le "ي". Pour former le duel d'un nom indéfini ou défini par l'article, nous lui suffixons : "ان" au nominatif et "يْن" à l'accusatif et le génitif. Par exemple, la forme duelle du nom "سَيَّارةٌ" (une voiture) prend la forme "سَيَّارتَانِ" (deux voitures, au nominatif) ou "سَيَّارَتَيْنِ" (deux voitures, accusatif et génitif)

Dans certains cas, notamment pour les mots dont la racine est défectueuse ou qui se terminent par un "ى" (a) , un (و)ou un hamza (ء), la terminaison du nom se transforme devant le suffixe du duel. En l'occurrence, la forme "مقهى" (un café) a pour forme duelle "مَقْهَيَانِ" (deux cafés)

- ✓ **Déclinaison du nom au pluriel** :

  Il existe deux grandes catégories de pluriel en arabe :

  - o <u>Les pluriels externes ou réguliers</u> : Les pluriels externes sont formés par l'ajout d'un suffixe au singulier sans changement de la structure du mot. Nous distinguons :
    - Le pluriel externe masculin : Pour le pluriel masculin nous rajoutons les deux lettres "ين" ou "ون" dépendamment de la position du mot dans la phrase (sujet ou complément d'objet), exemple : "مسلم"(musulman) devient "مسلمون" (musulmans, au nominatif) ou "مسلمين" (musulmans, accusatif ou génitif) ;
    - Le pluriel externe féminin : De la même manière, nous rajoutons pour le pluriel féminin le morphème " ات " (àt), exemple "سيَّارَة" (une voiture) devient "سيَّارَات" (des voitures).
  - o <u>Les pluriels internes ou brisés</u>: Les pluriels internes sont désignés par pluriels brisés à cause des modifications et infixations qu'ils nécessitent par rapport à la forme du singulier, à la différence de ce qui se passe avec lespluriels réguliers (masculin et féminin). Les formes du pluriel brisé sont nombreuses et généralement imprévisibles; elles suivent une diversité de règles complexes et dépendent du nom ; par exemple : le nom "كاتب" (un écrivain) se transforme pour donner les deux formes plurielles "كُتَّابٌ" (écrivains) ou "كَتَبَةٌ" (écrivains). Notons aussi que les grammairiens arabes ont formulé des distinctions entre pluriels de petit nombre et pluriels collectifs ; par exemple : le nom "شَهْرٌ " (mois) admet deux formes plurielles : " أَشْهُرٌ " (moins de 12 mois) et " شُهُورٌ " (au-delà) ;

    Seuls les pluriels externes suivent des déclinaisons propres. Les pluriels internes se rattachent aux déclinaisons du singulier (déclinaisons de base à trois cas et diptotes).

3.3.3 Flexion des mots outils

Lorsqu'il s'agit de la flexion des particules, nous en distinguons deux catégories :

- **Les mots outils non déclinables ou invariables** : leurs formes sont constantes et n'acceptent aucune déclinaison ; par exemple : "عَلى" (sur), "مُنذ" (depuis), etc.
- **Les mots outils déclinables ou variables** : ils suivent le système de déclinaison à trois cas selon leurs fonctions dans la phrase. Par exemple, le quantificateur "كلّ" (tout) peut accepter les trois voyelles casuelles filiales pour désigner le nominatif, accusatif ou génitif selon sa fonction dans la phrase.

3.4 Morphologie dérivationnelle

La morphologie dérivationnelle est la branche de la morphologie qui s'intéresse à la construction de nouvelles primitives morphologiques à partir de celles existantes selon des règles de dérivation adéquates. Tout verbe a dans son sillage des formes dérivées qui lui sont associées et avec les lesquelles il entretient des relations morphologiques, syntaxiques et sémantiques. Le nombre et la nature de ces formes varient selon le statut du verbe.

- **Le nom verbal (اسم مشتق)**: c'est le type des noms qui sont dérivés à partir de la même racine que le verbe associé avec un contenu sémantique pareil. Généralement, tous les verbes arabes possèdent un nom verbal associé ou plus dans certains cas. En ce qui concerne les verbes augmentés « الأفعال المزيدة », ils possèdent un seul nom verbal. Par contre les verbes nus « الأفعال المجردة » peuvent avoir jusqu'à cinq noms verbaux (Mesfar, 2008). Par exemple le verbe « كَتَبَ » (écrire) admet quatre noms dérivés différents « كِتابٌ » (un livre) , « مَكتَبَةٌ » (une bibliothèque) , «مَكتوبٌ »(écrit), « مَكتَبٌ» (un bureau).
- **Le nom de l'agent**[4]**«**اسم الفاعل**»** : ce type de noms sont généralement associés aux verbes transitifs « متعدي » ou intransitifs « لازم » en montrant l'agent qui effectue l'action. Par exemple, les verbes à racine simple tels que «خرج » (sortir) suit le schème « فاعل » pour produire le nom de l'agent « خارج » (celui qui sort).
- **Le dû patient**[5]**اسم المفعول :** c'est le type de noms dérivés qui sont, généralement, associés aux verbes transitifs « متعدي » en indiquant le patient qui a subi l'action. Par exemple, le verbe à racine simple « سَرَقَ »(voler), il subit le schème « مَفعولٌ » pour produire le nom du patient « مَسروقٌ »(volé) .

Ces trois déverbaux sont ceux qui existent pour le plus grand nombre de verbes. Leurs formations obéissent, pour un type donné de verbe, à des règles extrêmement générales. Habituellement, nous assimilons le participe actif au participe présent français et le participe passif au participe passé. Cette assimilation n'occulte pas les propriétés spécifiques que ces déverbaux en arabe. Pour tous les verbes, simples et augmentés, les participes se forment sur des schèmes stables ; ils ont, donc, un comportement morphologique d'une grande régularité. En tant que noms, les participes peuvent recevoir toutes les marques morphologiques de cette classe : genre, déclinaison, nombre et détermination.

[4] Appelé aussi le participe actif
[5] Appelé aussi le participe passif

Aux formes dérivées ci-dessus s'ajoutent d'autres formes dont le rang d'utilisation est moins important.

- Le nom de lieu « اسم المكان ».
- Le nom de temps « اسم الزمان ».
- Le nom d'instrument« اسم الآلة ».
- Le nom de fois« اسم المرَّة ».
- Le nom de manière« اسم المكان ».

Contrairement aux verbes, les noms primitifs échappent au système dérivationnel. Cependant, certaines règles peuvent être mises en place pour décrire les gentilés et ethnonymes qui sont des noms ou adjectifs par lesquels nous désignons des habitants d'un lieu, une nationalité, une identité nationale, etc.

3.5 Les propriétésmorphologiques

3.5.1 Les propriétés morphologiques verbales

En général, un verbe peut avoir les propriétés morphologiques suivantes :

a. L'aspect

En langue arabe, on peut différencier entre trois types d'aspects des verbes :

- ✓ **L'accompli (الماضي) :** appelé aussi « le passé », désigne une action achevée.
- ✓ **L'inaccompli (المضارع) :** appelé aussi « le présent »,désigne une action encours de se produire, sans être achevée.
- ✓ **L'impératif (الأمر) :** désigne l'ordre. Il peut être conjugué seulement avec les deuxièmes personnes.

b. Le mode

On distingue trois modes :

- ✓ **Le nominatif (المرفوع) :** il se caractérise par(الضمة - ُ) à la fin.
- ✓ **L'accusatif (المنصوب):**il se caractérise par(الفتحة -َ) à la fin.
- ✓ **L'apocope (المجزوم) :**il se caractérise par l'absence de la marque (السكون -ْ) à la fin.

c. La voix

La langue arabe a deux voix :

- ✓ **L'actif** (المعلوم).
- ✓ **Le passif** (المجهول) .

d. La personne

Comme les autres langues, on distingue trois personnes :

- ✓ **Première personne :** أنا , نحن
- ✓ **Deuxième personne :** أَنتَ, أنتِ, أنتُما, أنتُم, أَنتُنَّ

- ✓ **Troisième personne :** هُوَ, هِيَ, هُما, هُمْ, هُنَّ

e. Le genre

Dans la langue arabe, il existe deux genres :

- ✓ **Masculin** : pour examiner tous les verbes masculins.
- ✓ **Féminin** : pour examiner tous les verbes féminins.

f. Le nombre

Un verbe arabe peut avoir trois nombres :

- ✓ **Le singulier.**
- ✓ **Le duel.**
- ✓ **Le pluriel.**

3.5.2 Les propriétés morphologiques nominales

En général, un nom arabe peut avoir les propriétés suivantes :

g. Le genre du nom

Dans la langue arabe, il existe deux genres :

- ✓ Masculin.
- ✓ Féminin.

h. Le nombre du nom

Un nom arabe peut avoir trois types des nombres :

- ✓ Singulier.
- ✓ Dual.
- ✓ Pluriel : le pluriel externe ou sain est un pluriel à suffixe de masculin ou féminin.

Il y'a un cas particulier de pluriel qui s'appelle le pluriel brisé[6] « جمع تكسير », qu'in s'ajoute, d'une part, pour enrichir cette langue, et d'autre part, pour compliquer son traitement automatique. Ce pluriel est souvent rencontré dans la langue arabe, et il suit des règles imprévisibles et complexes.

i. La déclinaison du nom

On distingue trois cas de déclinaison d'un nom en arabe :

- ✓ Nominatif « مرفوع »
- ✓ Accusatif « منصوب »
- ✓ Génitif « مجرور »

j. La détermination d'un nom « التَّعْريف »

On distingue deux cas :

[6] Appelé broken plural en anglais

- ✓ **Déterminé :** il est signalé par une terminaison vocalique sans «Tanwin ».
- ✓ **Indéterminé :** il est signalé par une terminaison «Tanwin ».

Le nom peutêtre déterminé par :

a. Le vocatif (النداء) : يا مُرَادُ
b. L'annexion d'un complément de nom (بالإضافة) : باب الكلية
c. L'article السَّيارة :ال

3.6 Les éléments essentiels de la morphologie arabe

La langue arabe a une morphologie riche et différente, par rapport aux langues occidentales. L'analyse morphologique d'un mot arabe, consiste principalement à déterminer la structure générale de ce mot, s'il existe, et les autres éléments utilisés pour construire ce mot (les affixes, les modèles). Les éléments essentiels de la morphologie de la langue arabe sont (ZREIK & HAJJAR, 2010)(Buckwalter, 2004).

3.6.1 Les racines« الجذر »

Les racines sont à l'origine de la plupart de mots arabes. Elles sont des verbes formés de trois à cinq lettres consonnes (khalfallah, 2008). Elles sont aux alentours de 10000 racines dont la grande majorité (85%) est trilatérale. Les restes sont des racines quadrilatérales. Une racine définit la signification fondamentale des mots dérivés en utilisant différents diacritiques et affixes avec les lettres de la racine pour créer l'inflexion de la signification. Par exemple, la racine « كتب »(il a écrit) a la signification de base «écrire». Plusieurs mots sont dérivés à partir de cette racine, en la conjuguant sous plusieurs formes (présent, imparfait, futur simple, passe simple, impérative, etc.). Il y a aussi des formes supplémentaires telles que les noms verbaux.

	La racine « كتب » (écrire)			
Verbes	كتب	Il a écrit	تكتب	Il écrit
	كتبا	Nous avons écrit	يكتبون	Ils écrivent
	كتبت	Elle a écrit	تكتب	Tu écris
	تكتبون	Vous écrivez	نكتب	Nous écrivons
Noms	كاتب	Écrivain	كتابة	Ecriture
	كتاب	Livre	مكتوب	Ecrit
	مكتب	bureau	اكتتاب	Enregistrement

Figure 9: quelque dérivation du verbe "كتب"

3.6.2 Les schèmes[7] « الأوزان »

Un schème représente une forme ou modèle général composé par une séquence de caractères. Quelque caractère estconstant et d'autres sont variables. Les caractères variables sont destinés à être substitués par d'autres d'une racine pour générer radicale.

Les schèmes servent à produire la plupart des mots arabes à partir d'une racine ou inversement à extraire la racine d'un mot. Ils sont aux alentours de 900 schèmes. Quel que soit le mot, il est donc issu d'une racine et inséré dans un schème. En fait le schème est une sorte de moule. De plus, ils permettent de déterminer la racine d'un mot arabe(khalfallah, 2008). En général, les racines trilatérales sont représentées par le modèle <فَعَلَ, faire>, les racines quadrilatérales sont représentées par le modèle <فَعْلَلَ>. Par contre, on peut trouver le schème<عل> qui représente les mots qui ont perdu l'une des leurs lettres. Par exemple, le mot <زن, pèses> a pour racine <وزن>(khalfallah, 2008).

La racine	Le schème	Résultat
كتب	فِعَالٌ	كُتَّابٌ
غلب	مَفْعولٌ	مَغْلوبٌ
لعب	فاعِلٌ	لاعِبٌ
روى	فاعٍ	راوٍ

Table 2:Exemple de construction des mots à partir d'un schème

Pour construire un mot à partir d'une racine il suffit de modifier les lettres 'ف','ع'et 'ل' successivement par les lettres composant de la racine.

3.6.3 Les affixes « الزّوائد »

Les affixes sont des lettres qui s'ajoutent au début (les préfixes) ou à la fin des mots arabes(les suffixes). En général, ils sont utilisés pour accorder aux mots des éléments syntaxiques.Ils marquent l'aspect verbal, le mode, les propriétés transitives, etc. Ils sont aux alentours de150 (Mesfar, 2008).

3.6.3.1 Les préfixes « السوابق »

Les préfixes dépendent des mots auxquels ils s'attachent. En effet, la plupart des mots arabescommencent par le préfixe « التعريف » (l'article de définition) qui est utilisé en tantque terme déclaratif. Pour cela, il y a trois types des préfixes. Premièrement, les préfixesnominaux qui sont réservés pour les noms et les adjectifs. Deuxièmement, les préfixesverbaux qui sont réservés aux verbes. Et troisièmement, les préfixes généraux qui sont utilisésindépendamment de type des mots. La table suivent présente des exemples de chaque type despréfixes (Sanan, 2008).
Les préfixes peuvent s'enchaîner dans un mot pour former des préfixes composés qui peuventatteindre jusqu'à quatre lettres <وبال, et avec' le', 'la' ou 'les'>, <وال , et 'le', 'la' ou 'les'>, <بال ,avec >, <كال , comme ' le', 'la' ou 'les'>, etc.).

[7] Dites aussi modèles ou patrons ou gabarit

Type	Les préfixes		
	Non en français	Signification	Préfixe
Nominaux	L'article de définition	Le	ال
	Les propositions	Avec	بـ
		Pour	لـ
		Comme	كـ
	…	…	…
Verbaux	La particule de la future	Sera	سـ
	Les particules du subjonctif	Pour	لـ
	…	…	…
Généraux	Les conjonctions de coordination	Et	فـ
		Et	و
	..	…	…

Figure 10: Exemple des préfixes

Dans ce cas, certains préfixes nepeuvent prendre que la première position, l'article d'interrogation <أ> par exemple, d'autrespeuvent prendre n'importe quelle position, l'article de définition <ال التعريف, l'article de définition >.

3.6.3.2 Les suffixes « اللواحق »

Il y a deux types des suffixes, les suffixes verbaux et les suffixes nominaux. Les premiersdépendent de la transitivité et de la personne conjuguée. Les suffixes nominauxindiquent la flexion casuelle du nom (nominatif, accusatif, et génitif), le genre (masculin etféminin), le nombre (singulier, duel et pluriel), etc.

Type	Nombre	Suffixes	
		Signification	**suffixe**
Première personne	**Singulier**	Moi/Mon	ني
	Duel/ pluriel	Nous/Notre	نا
Deuxième personne	**Singulier**	Toi/Ton	ك
	Duel	Votre/Vous	كما
	pluriel	Votre/Vous	كم
		Votre/Vous	كن
Troixième personne	**Singulier**	Lui/ Son	ه
	Duel	Eux/leur	هما
	pluriel	Eux/leur	هم
		Eux/leur	هن

Tableau 12: Un exemple des suffixes divisés selon leurs types.

3.6.4 Les radicales « الجذوع »

Un radical est la dérivation obtenue à partir d'une racine donnée selon un modèle. L'arabe classique à un grand nombre des Stems qui ne sont pas tous utilisables, 2% seulement sont utilisables. Le Stem correspond à un modèle si et seulement s'il possède le même nombre de lettres et les mêmes lettres dans les mêmes positions. Une exception est accordée aux consonnes <ف>, <ع>, <ل> qui sont les lettres de la racine de base <فعل, faire>.

Racine	schème	radicale	Utilisable ?
كتب	فَعَلَ	كَتَبَ	Oui
درس	فَاعَلَ	دَارِسٌ	Oui
أكل	مَفْعُولٌ	مَأكولٌ	Oui
لعب	فَعْلاَّءٌ	لَعْبَاءُ	Non

Tableau 13: Exemple de génération des radicales

Par exemple, on y trouve : <مكاتب, bureaux>, il est obtenu à partir de la racine <كتب,il a écrit> selon le modèle <مفاعل>. Les radicaux produits ne sont pas tous utilisables(Wikipedia).

On peut arriver auxradicales à partir de la racine on appliquant le un schème ou à partir d'un mot agglutiné en supprimant les affixes.

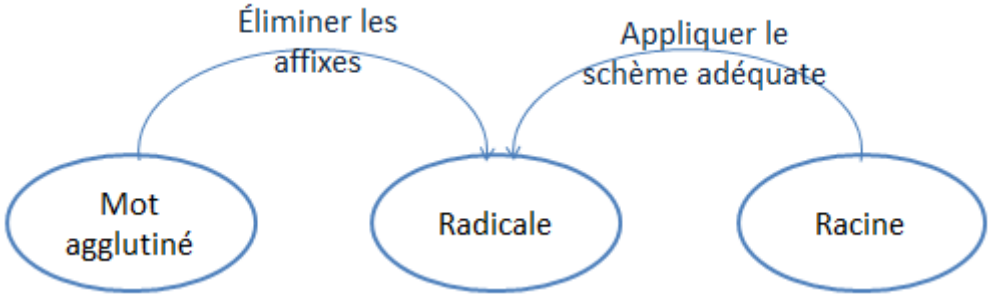

Figure 11: Le système de dérivation arabe

3.6.5 Les mots dérivés « الأسماء المعربة »

Les mots dérivés sont construits à partir d'un radical en y ajoutant des affixes. La plupart desmots arabes sont considérés comme des mots dérivés, puisqu'ils sont construits à partir desracines en utilisant les schèmes. Ainsi, les mots qui dérivent d'une même racine ont des significations similaires. Eneffet, certains mots dérivés peuvent avoir la signification d'une phrase entière, comme c'est lecas du nom <أتطلبون , Est-ce que vous demandez?>.

Mot dérivé	Préfixe	radical	Suffixe	racine
أتطلبون	أ	تطلب	ون	طلب

Tableau 14: Exemple de formation du mot "أتطلبون"

Il y a deux catégories des mots dérivés : les verbes et les noms [Baloul et al., 2002].Dans la plupart des cas, un verbe représente une racine. La conjugaison du verbe consiste à ajouter des affixes. Ces affixes dépendent du nombreux paramètre :

- du temps (présent, passé, futur, etc.),
- du nombre(singulier, duel et pluriel),
- mode (actif et passif),
- du genre (masculin, féminin),
- lapersonne (première, deuxième et troisième),

- …etc. (Sanan, 2008).

Prenons l'exemple du mot« يلعبن » (elles jouent) qui dérive de la racine « لعب » (il a joué) par l'ajout du préfixe<ي> et du suffixe <ن>. Dans ce cas, le temps est présent, le nombre est pluriel, lemode est actif, le genre est féminin et la personne est troisième.
Dans le cas des noms, la dérivation est utilisée pour indiquer le genre, le nombre, etc. Parexemple, le féminin singulier nécessite d'ajouter le suffixe <ة> comme « مدرسة » (école) mais le féminin pluriel nécessite d'ajoute le suffixe <ات> comme « مكتبات » (des librairies). En général, le pluriel se fait en ajoutant quelques suffixes comme<ات>,<ون> ,<ين, >, etc.). Par contre, il y a des mots qui ont des règles de composition plus complexe, comme le cas despluriels irréguliers, comme le mot « أشجار » (arbres) qui estle pluriel du mot « شجرة » (arbre)(Sanan, 2008).

3.6.6 Les mots isolés « الأسماء الجامدة »

Les mots isolés sont les mots qui n'ont pas des racines. Les mots sont en général, les nomspropres, les noms communs et les particules (Sanan, 2008).
Un nom propre désigne toute substance distincte de l'espèce à laquelle elle appartient. Il nepossède en conséquence aucune signification ni aucune définition. Exemple : « فاس », « محمد », etc.

Par contre, un nom commun est toute substance non distincte del'espèce à laquelle elle appartient. Il est pourvu d'une signification et d'une définition.

Exemple :

<بلد, pays>, <إنسان, une personne>, <حيوان,un animal>, etc.
La particule est un mot court qui ne représente qu'une expression grammaticale, comme les conjonctions de coordination et de subordination, des prépositions et desadverbes.
Il y a plusieurs types des particules :

- **Les particules du temps:** <حينما,lorsque>, <عندما, tant que>, <بعد , après>,etc.
- **Les particules du lieu:**<فوق , au-dessus>, <عند, chez>, <حيثما, où>, etc.
- **Les particules denégation**:<بلا, sans>, <دون, sans>, <ليس n'est pas>, etc.).

3.6.7 Les signes diacritiques « التشكيل »

Les signes diacritiques sont des signes ajoutés au-dessus ou en dessous des lettresarabes afin de spécifier la prononciation du mot. Ce rôle phonologique influe aussi sur le sensde ce mot. En effet, deux mots peuvent être écrits de la même manière, mais différenciés parl'ajout des signesdiacritiques différents.

Les signes diacritiques ne sont pas utilisés largement dans les documents arabes. Ils ne sontutilisés que pour spécifier une signification donnée d'une écriture qui peut être confuse. Pourcela, la plupart des études dans le traitement automatique de la langue arabe ignorent lessignes diacritiques et les suppriment durant une phase préalable qu'on appelle la normalisation. La normalisation consiste aussi à remplacer quelques lettres par d'autres selon des règles prédéfinies. Ce processus peut créer des ambigüitésdans certains cas. En effet, plusieurs mots, ayant des sens différents, peuvent avoir la mêmeforme normalisée (Khreisat, 2006).

3.7 Conclusion

Dans ce chapitre, nous avons explicité les différentes connaissances liées au traitement automatique à la langue arabe. Comme nous l'avons déjà laissé entendre précédemment, le

mot graphique arabe représente l'entité de base sur lequel nous allons construire notre système pour le TALN. Dans ce contexte, nous avons essayé de dégager les problèmes qui peuvent entraver la construction d'un système de TALN arabe.

Chapitre 3 : Analyseur morphologique de la langue arabe

4.1 Introduction

L'analyse morphologie est considérée comme une étape primordiale dans le traitement automatique de langues naturelles. Notons que les étapes d'analyse syntaxique et l'analyse sémantique nécessitent obligatoirement la réalisation d'un analyseur morphologique robuste. En d'autres termes, la plupart des applications de TALN, telles que par exemple la recherche d'information, traduction automatique, la correction orthographique, etc.nécessite une étape principale qui est l'analyse morphologique. De ceci, on conclut que l'opération d'analyse morphologique est considérée le cœur du traitement automatique de la langue.

Dans ce travail, nous intéressons à l'analyse morphologique arabe. Notons que les recherches dans le domaine de traitement automatique de la langue arabe sont continuées pour proposer un système d'analyse morphologique arabe valide et robuste. David Cohenest l'un des premiers chercheurs dans ce domaine qui a proposé un système d'analyse automatique arabe, dès les années 60.

Actuellement, plusieurs travaux dans le domaine ont vu le jour pour le développement de systèmes d'analyse morphologiques arabes. Mais, cette dernière reste une langue sémitique présentant de nombreux problèmes morphosyntaxiques ce qui complique le processus d'analyse.

Nous commençons ce chapitre par présenter le processus d'analyse morphologique. Puis, nous avons introduit une étude sur les approches d'analyse morphologique arabe. Ensuite nous avons concentré sur l'ambiguïté suivie par les approches proposés de désambiguïsation. Enfin nous avons donnée une description sur les analyseurs existants.

4.2 L'analyse morphologie

L'analyse morphologique a pour objectif de donner à chaque unité lexicale, en entrée, les traits morphologiques. De plus, l'analyseur morphologique doit permet l'analyse des différents types de mots arabes, à savoir, les noms, les verbes, les particules, etc. Pour réaliser cela le système doit baser les étapes à savoir, la segmentation du texte en mots, le prétraitement morphologique, l'analyse affixale, l'analyse morphologique et le post-traitement(Belguith & Chaâben, 2003).En effet, pour faire cela toutesla plupart des Segmentations du texte en phrases et en mots

4.2.1 Segmentation

Cette partie peut être appelée aussi analyse lexicale, il a pour but de segmenter le texte en mots est faite en deux étapes : une segmentation du texte en phrases, en premier lieu, et une segmentation des phrases en mots, en second lieu.

Il y en a beaucoup des systèmes qui font cette tâche parmi eux on trouve STAr(Belguith & Chaâben, 2003), chaque système de segmentation de textes basé sur l'exploration contextuelle des signes de ponctuation, des mots connecteurs jouant le rôle de séparateurs de phrases ainsi que celles de certaines particules, telles que les conjonctions de coordination. La segmentation de la phrase en mots se base sur la détection des espaces, des signes de ponctuation et de certains caractères spéciaux.

4.2.2 Prétraitement morphologique

Cette étape consiste à supprimer les proclitiques et les enclitiques pouvant être agglutinés au mot en se basant sur les listes de proclitiques et d'enclitiques. Le mot restant sera filtré pour tester s'il s'agit d'une particule, d'un nom propre, d'un nombre ou d'une date. Si ce n'est pas le cas, il va subir l'analyse affixale dans le but est de déterminer toutes ses caractéristiques morphologiques possibles.

4.2.3 Stemming[8]

Stemming est une technique qui a pour but d'extraire la racine ou la radicale d'un mot, d'où les raciner sont des systèmes qui font l'extraction de racines des mots entrants et retourne comme résultat les préfixes, les suffixes, le schème utilisé, ainsi la racine et la radicale, ce type d'analyse «simpliste », traite de façon identique affixes flexionnels et dérivationnels.

Les algorithmes de racinisation en arabe les plus connus sont ceux de (Larkey, Ballesteros, & Connell) et (khoja, 2001). Ci-dessous une description succincte de ces raciner. Plusieurs chercheurs qui ont utilisé le Stemming (ELHADJ, AL-SUGHAYEIR, & AL-ANSARI, 2009)(Darwish, 2002)(ELHADJ, AL-SUGHAYEIR, & AL-ANSARI, 2009)…, de nombreux algorithmes de cette technique sont présentés au cours des quinze dernières années. Mais le Stemming arabe présenté par Khoja est considéré par un certain nombre de chercheurs comme un Stemming standard (LARKEY, BALLESTEROS, & CONNELL, 2002)pour Modern Standard Arabe (MSA) et aussi pour les dialectes arabes, une des étapes principales dans la Stemming est la normalisation qui vise à transformer une copie du document original dans un format standard plus facilement manipulable. Cette étape est considérée nécessaire à cause des variations qui peuvent existerlors de l'écriture d'une même unité lexicale. Donnant comme exemple :

- Suppression des caractères spéciaux ;
- Remplacer la lettre(ة) par (ه) ;
- Remplacement les lettres إ,أ, آ, ئpar ا ;
- Remplacement de la lettre finale ىparي .

En effet, plusieurs mots, ayant des sens différents, peuvent avoir la même forme normalisée. Nous distinguons deux types de Stemming :

[8] Appelées aussi racinisation ou désufixation

- **Light Stemming :** s'intéresse seulement par le Stem.

Figure 12:Exemple de light Stemming

- **Heavy Stemming :** son objectif est d 'extraire le Stem et la racine.

Figure 13: Exemple de heavy Stemming

4.2.4 Analyse affixale

Cette étape a pour objectif de reconnaître les éléments de base qui entrent dans la constitution d'un mot à savoir, la racine (R) ou la forme canonique et les affixes (préfixe (P)et suffixe(S)),cette opération est effectuée en plusieurs étapes ,d'une étape à l'autre un mécanisme de filtrage permet d'éliminer les décompositions parasites reconnues. On distingue les principales étapes suivantes :

- identification des couples (P, S),
- identification des couples affixales candidates,
- filtrage lexical, contrôle des associations (R) et (P, S),

4.2.5 Analyse morphologique

Cette étape consiste à déterminer, à partir de la forme (radicale, préfixes,suffixes) obtenue pour chaque mot, tousles cas possibles où chaque cas avec ces avec ses caractéristiques morphosyntaxiques possibles (c.-à-d. partie de discours, genre, nombre, temps, personne, etc.). La détection des caractéristiques morphosyntaxiques se fait en trois phases(hadrich, 1999).

a. Consiste à identifier la catégorie principale du mot (partie de discours), à savoir verbe, nom, adjectif, etc.
b. Permets de déterminer pour chaque catégorie, identifiée au niveau de la phase 1, la liste de ses caractéristiques morphologiques.

c. Un filtrage des listes de caractéristiques.

Le schéma suivant représente le processus d'analyse morphologique pour un mot donné:

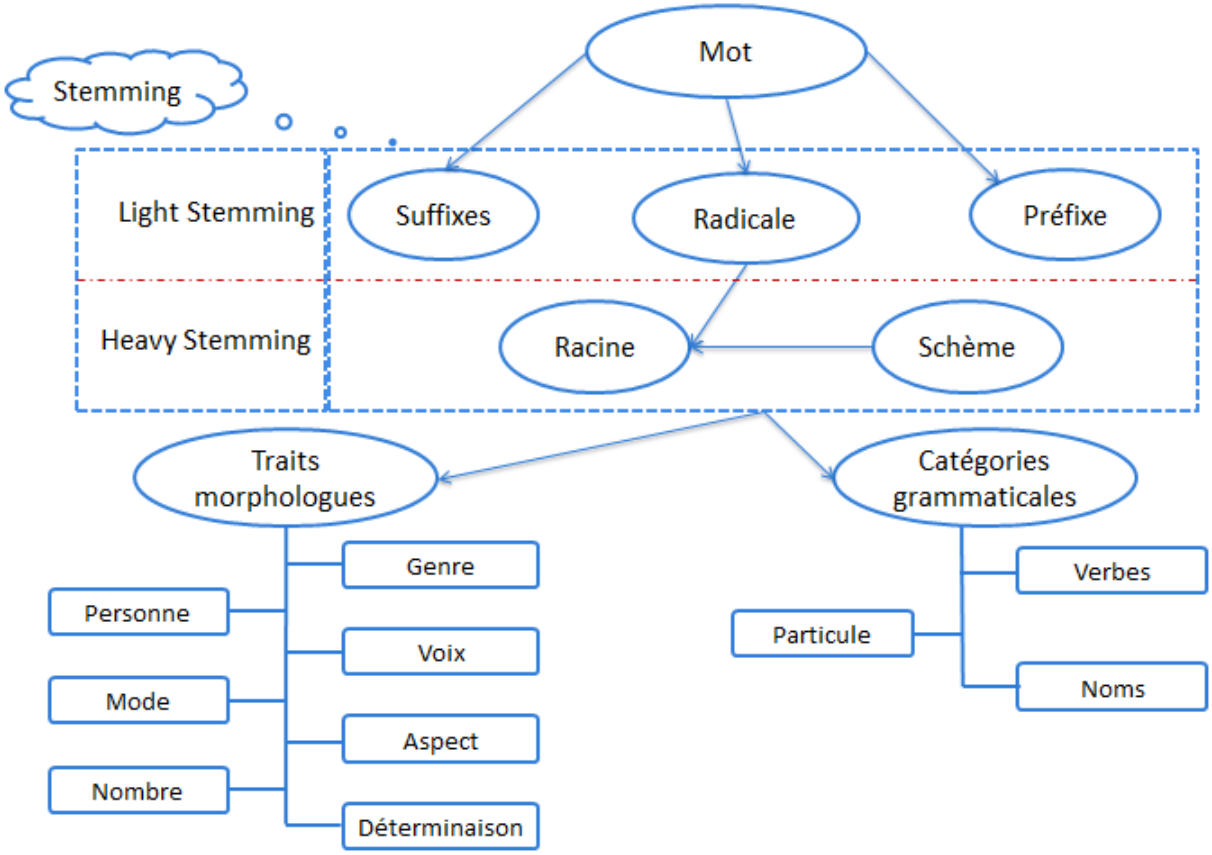

Figure 14: Analyse morphologie d'un mot

4.2.6 Désambiguïsation

La désambiguïsation est le processus qui permet d'enlever l'ambiguïté, c'est une étape qui est cruciale dans le processus d`étiquetagemorphosyntaxique,à ce niveau du traitement si un motest mal étiqueté, les règles de la grammaires`appliqueront mal ou pas du tout.

للطالب	الدرس	الأستاذ	فهم
لَلطَّالِب	الدَّرْسَ	الأسْتَاذَ	فَهَّمَ
لَلطَّالِبَ	الدَّرْسِ	الأسْتَاذَ	فَهُمْ
لَلطَّالِبُ	الدَّرْسُ	الأسْتَاذِ	فَهِمَ
لِلطَّالِبِ			فَهْمٌ
			فَهَمَّ
لِلطَّالِبِ	الدَّرْسَ	الأسْتَاذَ	فَهَّمَ

Tableau 15: Exemple de désambiguïsation

Cependant la phase de désambiguïsation n'est pas toujours nécessaire ouobligatoire au bon déroulement du processusd'étiquetage.

Il faut dire que le module de désambiguïsation rentre en jeu dans un seul cas de figure, celui où l'unité lexicale (mot) reçoit plus d`uneétiquette (plus d'une information morphosyntaxique),ce qui va générer une situation de confusion ouambiguïté.

4.3 Étude sur les techniques de l'analyse morphologie

Le but principal de l'analyse morphologique est d'associer à chaque unité lexicale ces catégories grammaticales et ces traits morphologiques, pour se faire il doit être basé sur un algorithme capable de segmenter chaque unité lexicale en trois parties (préfixe, radical, suffixe), ces dernières sont capables de lier chaque élément parmi elleavec des informations qu'elle soit catégorie grammaticale ou trait morphologique, cette opération est connue par le Stemming.

Le Stemming en détailest une technique qui a pour but d'extraire la racine ou la radicale d'un mot, d'où les raciner sont des systèmes qui font l'extraction de racines des mots entrants et retourne comme résultat les préfixes, les suffixes, le schème utilisé, ainsi la racine et/ou la radicale, ce type d'analyse «simpliste », traite de façon identique les affixes flexionnels et dérivationnels.

Plusieurs recherches sont élaborées pour l'étude des techniques de la morphologie arabe, où chacun a été mis en œuvre selon une méthodologie spécifique, dans les sections suivantes, nous allons représenter un aperçu de ces algorithmes de Stemming utilisés pour les systèmes arabes.

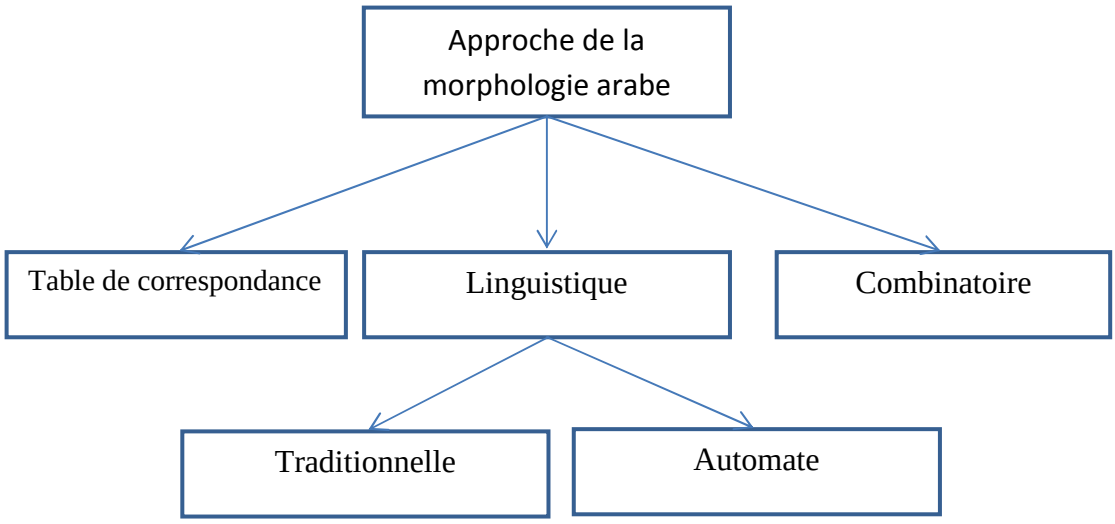

Figure 15: Les différentes techniques de la morphologie arabe.

4.3.1 La table de correspondance

L'approche de table de correspondance dépend principalement de très grands tableaux classés par ordre alphabétique, stockant des mots arabes qu'ils ont trouvés dans les

textes naturels avec leurs parties morphologiques correspondant. Ou chaque mot accompagné par sa radicale, sa racine et ses affixes.

Les mots peuvent inclure des mots fonctionnels, les mots étrangers, et les noms propres, où chaque mot utilise une entrée unique dans le tableau. Les entrées multiples peuvent exister pour les mots orthographiés de même pour refléter la possibilité de multiples analyses (Al-Fedaghi, 1989).

Suffixes	préfixes	racine	Radicale	Le mot
	ال	ح ل ل	تحليل	**التحليل**
				…
		ح ل ل	تحليل	**تحليل**
				…
ه		ح ل ل	تحليل	**تحليله**
				…
	وال	ح ل ل	تحليل	**والتحليل**
				…
ات		ح ل ل	تحليل	**وبتحليلات**
				…

Tableau 16: exemple de table de correspondance

Avantage

- Permets l'analyse des mots qui n'ont pas d'origine arabe.
- La précision
- Plusieurs entrés pour les mots qui ont ambiguë

Inconvénient

- Nécessite beaucoup de travail pour collecter tous les mots arabes.
- Nécessite un grand temps de recherche d'un mot.

4.3.2 Les approches combinatoires

Les approches combinatoires comparent les mots à tester contre les listes préparées par les racines, les schèmes, les particules, et les affixes.

La comparaison est basée sur un algorithme combinatoire qui teste toutes les combinaisons de trois ou quatre lettres d'un mot donné afin d'en extraire la racine. En général, de telles approches sont simples, mais prennent beaucoup de temps pour traiter et nécessitent de très grandes listes.

(Al-Fedaghi, 1989)A amélioré un algorithme proposé auparavant par Al-Fedaghi et Al-Sadoun. L'amélioration est obtenue par la résolution de certains problèmes phonologiques et orthographiques importants, comme l'assimilation, la mutation, vocalisation, et la germination. L'algorithme modifié utilise des listes de racines et de schèmes trilatérales avec toutes les combinaisons d'affixes. Il commence en comparant le mot d'entrée contre le schème de la liste pour extraire la racine. L'algorithme se dirige ensuite dans quatre modes de couvrir tous les cas orthographiques et phonologiques possibles.

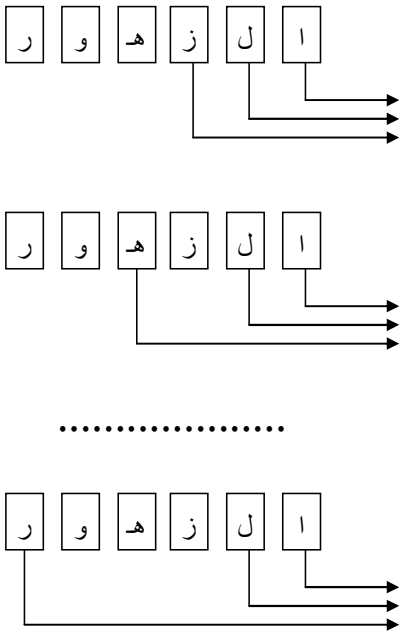

Figure 16: l'approche combinatoire pour le mot'الزهور'

Ils ont remarqué que quelques règles peuvent être ajoutées à l'algorithme pour construire un algorithme non homogène, mais qu'il est caractérisé par les deux avantages.

Avantage

- Facile et claire.
- Donne toutes les possibilités

Inconvénient

- Le taux d'erreur élevé
- Restriction par les racines trilittérales
- Trop long O(n3), d'où n est la longueur du mot

4.3.3 Les approches linguistiques

Les approches linguistiques nécessitent une analyse approfondie du système morphologique arabe.Ils simulent le comportement d'un linguiste lors de l'analyse d'un mot arabe donné.

Dans cette approche, les mots testés sont comparés aux listes des affixes pour extraire les radicales et les comparer ensuite avec les listes de schème et de racines afin d'extraire les racines.

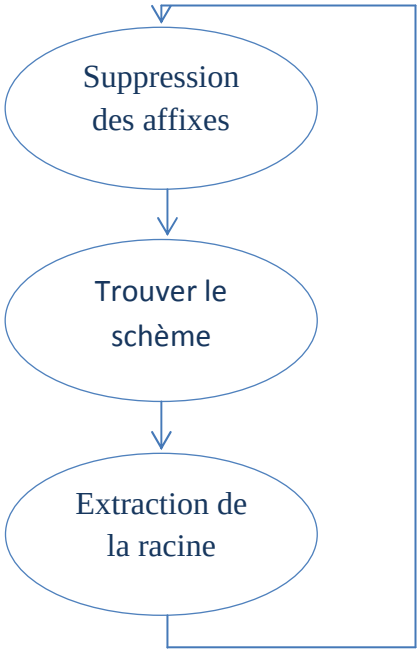

Figure 17: l'approche linguistique

En général, ces approches linguistiques sont plus précises, mais exigent trop de listes qui doivent être préparées et contrôlées linguistiquement. Ces listes encourent une surcharge de temps pour y accéder. En outre, le mécanisme pour enlever les affixes est presque un processus d'essais et d'erreurs, dont les résultats ne sont pas garantis pour être précis. Les approches linguistiques peuvent être distinguées en deux approches.

4.3.3.1 L'approche traditionnelle

Les systèmes basés sur cette approche devinent le type de mot d'entrée comme verbe, nom ou particule. Un peu d'information peut être utilisé comme un indice que le type de mot, comme une marque de notation. L'analyse dumot commence par supprimer les affixes.

Après l'élimination partielle ou totale des affixes, le reste sera comparé avec des schèmes qui sont de même longueur. Les entrées dans la liste de schèmes sont triées en fonction de la longueur du schème et la vocalisation est constituée de schèmes, des racines, et des propriétés morphologiques. De nombreux processus de l'essai sont effectués pour assurer une analyse correcte y compris la compatibilité entre le préfixe et les schèmes, la disponibilité de la racine dans la liste, et les pronoms contrôle et le cas de la syntaxe des verbes.

Al-Afandi a donné un cadre de travail, pour une technique linguistique basée sur des matrices de compatibilité, et puisqu'il n'y a pas des règles claires qui peut lier les racines

avec les schèmes correspondants, la matrice illustrée dans la figure suivante peut être utilisée pour l'analyse morphologie d'où '1' signifie que le schème est compatible avec la racine contrairement au '0' qui signifie la non-compatibilité entre la racine et le schème.

	Schème1	Schème2		. . . .			Schème (n-1)	Schème (n)
Racine1	1	0	1	1	1	0	1	1
Racine2	1	1	1	0	0	1	0	1
. .	0	1	1	1	0	1	1	1
.	1	1	0	1	0	1	0	0
	1	0	0	1	0	0	0	0
Racine n	0	0	1	0	0	1	0	0

Tableau 17: La structure de la matrice de compatibité

Si les matrices de compatibilité pour toutes les racines sont disponibles, vous pouvez ensuite utiliser l'algorithme linguistique. Lorsque les racines sont trouvées, il faut être sûr d'eux dans des tableaux pour confirmer la compatibilité entre les racines et les schèmes proposés.

Avantage

- Les solutions précises
- Le taux d'erreur bas

Inconvénient

- La précision des résultats non garantie
- La nécessité de plusieurs listes et sa vérification linguistique.
- Le temps d'accès à ces listes et la matrice.

4.3.3.2 Les automates morphologiquesarabes

La théorie des automates a été exploitée dans plusieurs domaines et en particulier dans le traitement automatique des langues naturelles et par conséquent, les automates ont été utilisés dans le traitement automatique de la phonétique et de la morphologie. En outre, des applications telles que la reconnaissance vocale et de traitement de la langue utilisent les expressions régulières, les langages rationnels et les automates et les transducteurs d'état finis.

En ce qui concerne la morphologie, Koskenniemi était le premier qui a exploité les automates dans le traitement automatique de la morphologie, plusieurs travaux sur l'analyse morphologique arabe où différentes approches sont mises en œuvre pour satisfaire ce domaine de recherche (Al-kharashi & Al-Sughaiyer).

Généralement, un automate morphologique arabe est un « cinq-uplet » représenté de la manière suivante $< Q, \sum, q0, F, \tau >$ où :

- Q est un ensemble fini d'états représentant ainsi les états de l'automate morphologique arabe.
- ∑ Est un alphabet fini de symboles, pour un automate morphologique arabe, il est constitué de l'alphabet de la langue arabe.
- q0 est l'état initial de l'automate morphologique.
- F est un sous-ensemble de l'ensemble Q, il représente l'ensemble des états accepteurs de l'automate morphologique.
- La fonction τ qui représente la transition de l'automate morphologique arabe.

Par conséquent, l'implémentation de cet automate nécessite l'utilisation directe de la base des données morphologiques (XMODEL,DINAR,…)(Gridach, 2010). Elle nécessite l'extraction des règles morphologiques de la base de données et le développement d'un ensemble d'automates morphologiques arabes pour chaque règle.

La figure suivante présente l'automate qui présente les noms d'agent « فَاعِلٌ », ou λ est la transition vide et 'ℓ' représente les lettres de la racine.

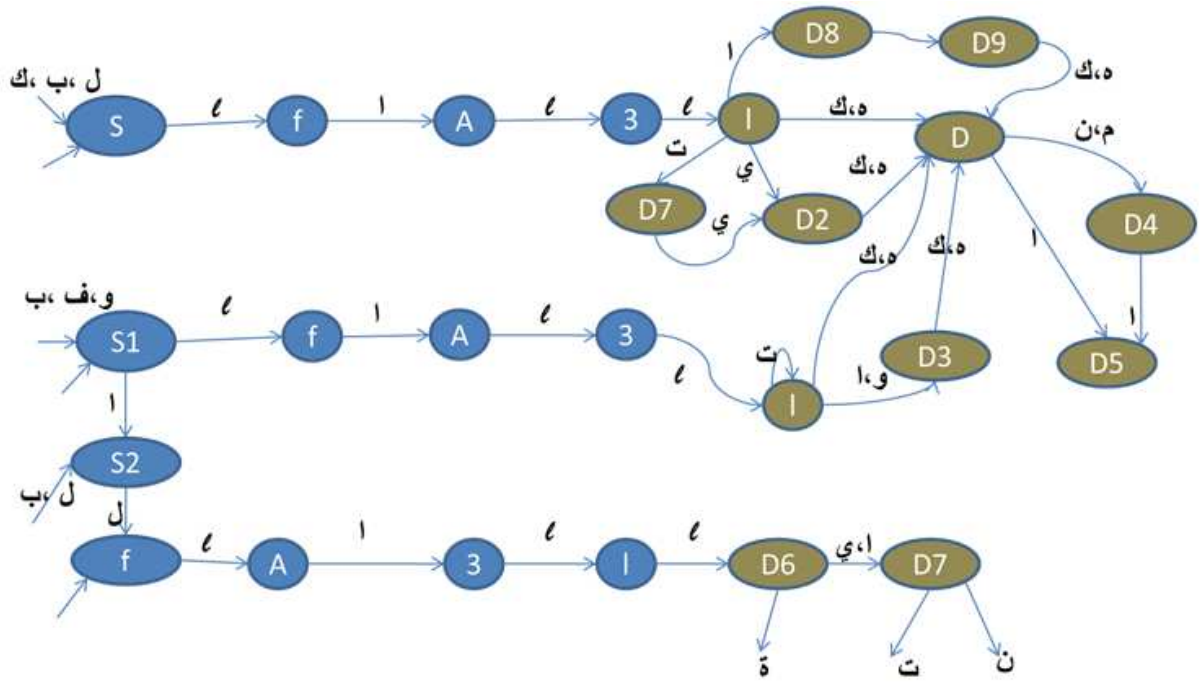

Figure 18: l'automate présentant les noms d'agent « فاعل »

La réalisation de chaque règle peut imposer quelques opérations sur les automates à savoir l'opération de concaténation et l'opération de fusion ou l'union.

Avantage

- Le taux d'erreurs faible
- Plus rapide

Inconvénient

Restriction sur les verbes et les noms dérivés

4.4 L'ambiguïté

L'ambiguïté morphologique se manifeste lorsque l'analyse associe, à une unité lexicale, plusieurs informations non conformes au contexte du mot, autrement dit quand l'analyse fournit plusieurs valeurs pour certains attributs morphologiques (hajji, 2000). Un mot est considéré ambigu si l'analyseur morphologique fournit plus qu'une seulesolution pour ses attributs morphologiques. Par ailleurs, une approche pour la désambiguïsation morphologique arabe est nécessaire pour faire face à l'ambiguïté des mots non voyellés. La désambiguïsation consiste, donc, à attribuer la valeur exacte d'un attribut morphologique parmi celles proposées par l'analyseur.

D'abord, les mots peuvent être ambigus aux niveaux lexical ou grammatical. Par exemple le mot «ذهب » est ambigu lexicalement. Il peut désigner « l'or » en français ou encore le verbe « aller ». «كتّب» quant à lui, est ambigu grammaticalement. Il peut appartenir à plusieurs catégories grammaticales différentes : verbe ou nom. Le sens de ce mot sera très différent selon sa catégorie : nom = « كُتُبٌ », verbe = « écrit ». Il existe aussi des ambiguïtés qui relèvent du niveau syntaxique. Une même phrase peut avoir plusieurs sens possibles en fonction de ses interprétations syntaxiques.

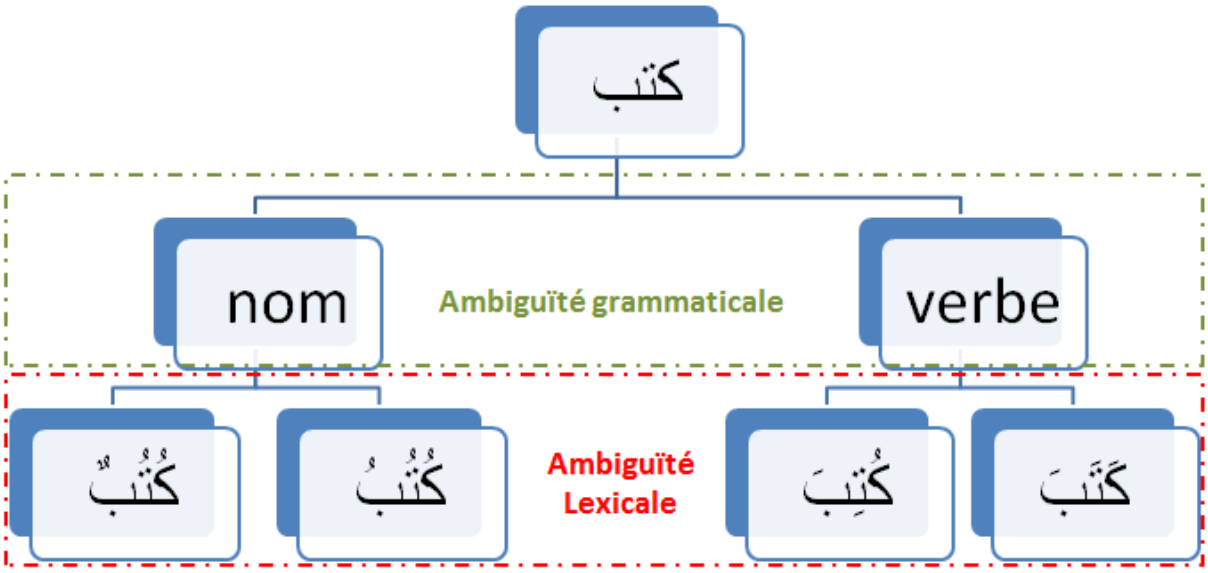

Figure 19:Les types d'ambiguïté

Ce phénomène d'ambiguïté est un problème omniprésent dans toutes les langues naturelles. Le code linguistique explique l'existence de plus d'une fonction. Les contextes linguistiques de la langue arabe sont de nature lucide.

En dépit de cela, ils nécessitent la présence d'un critère témoin de la fonction du code linguistique dans son contexte. La détermination de ces critères et leur compréhension dépend des capacités de l'individu des points de vue linguistique et du volume de ses

connaissances. Plus ces capacités augmentent, plus l'espace de l'ambiguïté diminue aux abords de la compréhension de la langue, et celui de la clarté s'amplifie pour atteindre son point culminant chez le spécialiste. Lorsque l'individu arrive à choisir parmi les diverses solutions celui qui correspond au contexte, on appelle ceci "désambiguïsation".

La levée de l'ambiguïté ou la désambiguïsation morphosyntaxique a pour objectif la réduction du nombre d'interprétations issues de l'analyse morphologique à l'aide du contexte immédiat de point de vue linguistiques et du volume de ses connaissances. Plus ces capacités augmentent, plus l'espace de l'ambiguïté diminue aux abords de la compréhension de la langue, et celui de la clarté s'amplifie pour atteindre son point culminant chez le spécialiste. Lorsque l'individu arrive à choisir parmi les diverses solutions celui qui correspond au contexte, on appelle ceci "désambiguïsation".

4.4.1 Ambiguïtés dérivationnelles et flexionnelles

La flexion est la variation de la forme des mots en fonction de facteurs grammaticaux telle que la conjugaison pour les verbes (exemple : le mot " يتأثّرون " (ils s'influencent) est le résultat de la concaténation du préfixe " ي" indiquant le présent et du suffixe " ون " indiquant le masculin pluriel du verbe " تأثّر "). Le problème en analyse morphologique de l'arabe se rapporte surtout au niveau de la dérivation qui est un phénomène plus complexe que la flexion. En effet, la dérivation est la formation de nouveaux mots à partir de mots existants. Dans le cas de la langue arabe, la plupart des mots sont dérivés à partir de racines trilitères ou quadrilatères. Le mot arabe n'est pas le résultat d'une simple concaténation de morphèmes comme c'est le cas pour l'anglais (exemple : unfailingly=un+fail+ing+ ly), mais c'est à partir d'une racine, d'une combinaison de voyelles, de préfixes, d'infixes, de suffixes et d'un schème morphologique qu'on obtient un mot (exemple : à partir de la racine " أثر " (choisir/citer à) on peut dériver plusieurs verbes tels que " تأثّر " (s'influencer) et plusieurs noms tel que " متأثّر "(ému)).

4.4.2 Ambiguïtés dues à l'agglutination

Contrairement aux langues latines, en arabe, les articles, les prépositions, les pronoms, etc. collent aux adjectifs, noms, verbes et particules auxquels ils se rapportent. Comparé au français, un mot arabe peut parfois correspondre à une phrase française (exemple : le mot arabe "أتتذكَّروننا " correspond en français à la phrase «Est-ce que vous vous souvenez de nous»). Cette caractéristique engendre une ambiguïté morphologique au coursde l'analyse. En effet, il n'est pas toujours facile de distinguer un proclitique ou enclitiqued'un caractère original du mot. Par exemple, le caractère " و " dans le mot " وصل " (il est arrivé)est un caractère original alors que dans le mot " وفتح " (et il a ouvert), il s'agit plutôt d'uneproclitique.

4.4.3 Ambiguïtés dues à la non-voyellation

La morphologie arabe est assez régulière lorsque les mots sont présentés sous leurs formes voyellées. Cependant, la majorité des documents arabes sont non voyelles sauf pour

le Coranet pour certains ouvrages scolaires pour débutants et donc c'est pour cette raison que nousnous sommes intéressés à l'arabe non voyellé.

En fait, les mots non-voyelles engendrentbeaucoup de cas ambigus au cours de l'analyse (exemple : le mot non voyellé " فصل " pris horscontexte peut être un verbe au passé conjugué à la troisième personne du singulier " فَصَلَ"(il alicencié), ou un nom masculin singulier " فَصْلٌ " (chapitre/ saison), ou encore une concaténation dela conjonction de coordination "فَـ " (puis) avec le verbe "صل " : impératif du verbe lié conjugué à la deuxième personne du singulier masculin.

4.5 Les approches dedésambiguïsation :

Les approches de désambiguïsation morphologique a fait l'objet de plusieurs travaux de recherches, qui sont classées principalement en deux, et chaque approche englobe une ou plusieurs techniques :

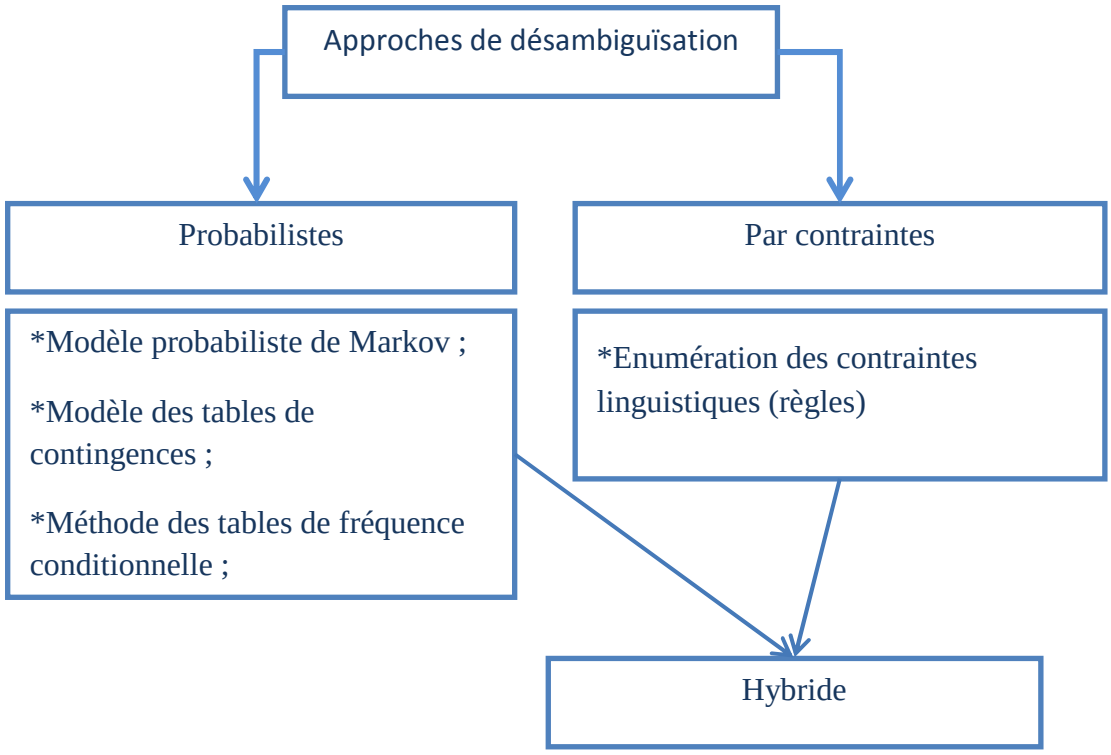

Figure 20: Diffèrent technique de désambiguïsation

4.5.1 Approche par Contrainte

L'approche par contrainte (ou base de règles ou linguistiques ou symbolique),elle a été utilisée avec succès dans le développement de nombreux systèmes de traitement du langage naturel. Les systèmes qui utilisent des transformations à base de règles sont basés sur un noyau de connaissances linguistiques solides. Les connaissances linguistiques acquises pour un système naturel de traitement du langage peuvent être réutilisées pour construire

les connaissances requises pour une tâche similaire dans un autre système. Elle consiste essentiellement à créer les règles manuellement.

Au lieu de trouver ces règles à l'aide de calculs complexes et de corpus, on utilise les connaissances linguistiques et la logique pour les établir, c'estl'approche la plus ancienne qui a été mise en place pour remédier le problème de l'ambiguïté morphologique. Ce type des approchesutilise une base de connaissances des règles écrites par des linguistes permettant d'attribuer des étiquettes aux différentes catégories morphologiques(Daoud, 2009).

Nous parlons, principalement, des heuristiques, des règles contextuelles et des règles non contextuelles, elles se basent principalement sur l'intervention d'un linguiste(ou grammairien) afin d'établir une liste de règles permettant de lever l'ambiguïté. Ces règles sont généralement classées en catégories de type : grammatical, structural, sémantique, logique,…etc.

Donc les systèmes et les outils qui utilisent l'approcheà base de règles sont basés sur un noyau de connaissances linguistiques solides. Parmi les caractéristiques de cetteapproche sont :

- Il a un sens strict du bien formé à l'esprit,
- Il impose des contraintes linguistiques pour satisfaire une bonne formation,
- Il permet l'utilisation de procédés heuristiques (comme un verbe ne peut pas être précédé d'une préposition),
- Elle s'appuie sur les règles construites à la main qui doivent être acquises auprès de spécialistes de la langue plutôt qu'automatiquement formés à partir des données.
- Il est facile à intégrer les connaissances de domaine dans les connaissances linguistiques qui fournit des résultats très précis.
- Les connaissances linguistiques acquises pour le système de traitement d'un langage naturel peuvent être réutilisé pour construire les connaissances requises pour une tâche similaire dans un autre système

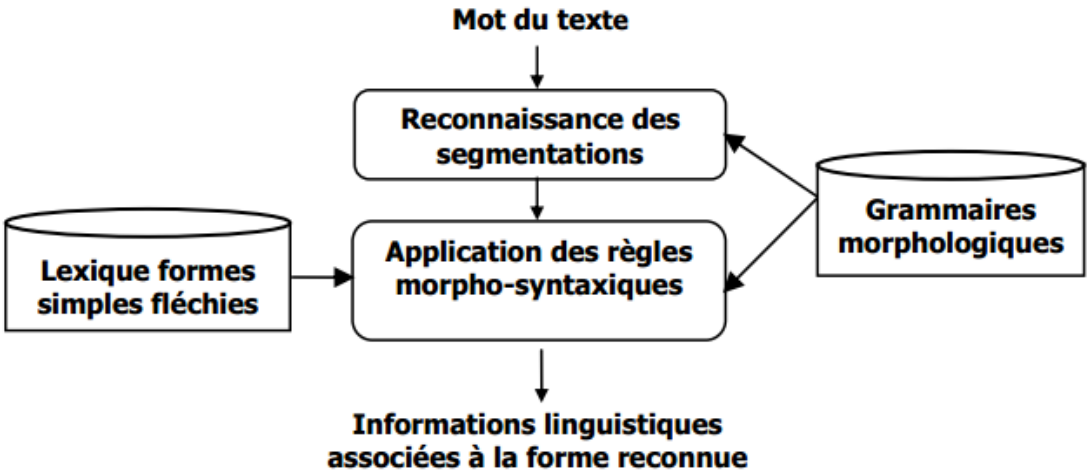

Figure 21: schéma présentatif des approches linguistique

Il y en a beaucoup des méthodes qui utilisent cette approche, parmi eux on cite :

Les arbres de décision

Les arbres de décision sont conçus pour exposer des bases de règles. Un arbre de décision est un modèle prédictif utilisé pour représenter les règles de classification avec une structure en arbre qui partitionne de façon récursive l'ensemble de données d'apprentissage. Chaque nœud interne d'un arbre de décision représente un test sur une valeur d'un attribut de classification, et chaque branche représente un résultat de test. Une prédiction est faite quand un nœud feuille est atteint. Cette approche est étendue pour extraire et calculer des mesures statistiques utilisées pour l'étiquetage grammatical.

4.5.2 Approche statistique

Les approches statistiques forment des modèles d'apprentissage à partir des corpus annotés. Elles incorporent des méthodes de classification telles que les modèles de Markov, SVM, etc. pour calculer des taux de probabilité de chaque valeur résultante d'une catégorie grammaticale d'un mot.

Un modèle peut être utilisé pour classer automatiquement les autres textes en se référant aux taux déjà calculés.(Diab, 2004)Développe un classifieur morphologique utilisant SVM. Ils entrainent et testent le classifieur sur un Treebank arabe de 4000 phrases d'apprentissage et 100 phrases de test.(Habash & Rambow, 2005)Utilisent SVM en se basant sur des informations fournies à partir d'un analyseur morphologique.Il y en a des autres qui combinent les probabilités calculées sur des ensembles d'apprentissagesarabes et hébreux pour classer les catégories grammaticales des mots des textes arabes. Ils utilisant les mêmes paramètres de test de (Diab, 2004).

Quelques travaux de recherches comprennent les modèles de Markov cachés (HMM). (ELHADJ, AL-SUGHAYEIR, & AL-ANSARI, 2009)Présente un système d'étiquetage grammatical qui combine l'analyse morphologique et le modèle de Markov. L'étiqueteur se base sur la structure de la phrase arabe. Dans un premier lieu, le texte est entièrement analysé morphologiquement pour réduire le nombre devaleurs possibles de POS. Dans un second lieu, le modèle statistique (HMM), fondé sur la structure de la phrase arabe, est utilisé pour attribuer à chaque mot la valeur exacte de sa catégorie grammaticale. (ELHADJ, AL-SUGHAYEIR, & AL-ANSARI, 2009)A utilisé leur propre corpus annoté qui est composé de vieux livres arabes. Le total des mots, dans ce corpus, est environ 21 000 mots.

4.5.3 Approche hybride

Une approche hybride combine les règles linguistiques avec les informations statistiques afin de résoudre l'ambiguïté morphologique dont on propose une approche qui analyse les affixes grammaticaux et flexionnels et les règles grammaticales en se basant sur l'approche MBL (**Memory basedlearning)**. Elle est appliquée pour classer une collection de textes coraniques et éducatifs. (ZRIBI, TORJMEN, & AHMED, 2006)Combinent l'approche à base de règles avec un étiqueteur trigramme HMM (COLLINs, 2002). L'apprentissage du

classifieurs trigramme a été fait sur des textes comportant 6000 mots. Des règles heuristiques ont été appliquées pour sélectionner parmi les résultats proposés.

(khoja, 2001)Amit en œuvre une approche hybride qui utilise l'algorithme de Viterbi. Elle calcule deux probabilités sur un corpus annoté composé de 50000 mots:

- une probabilité lexicale, qui est la probabilité qu'un mot ait une certaine valeur d'un attribut morphologique spécifie.
- Une probabilité contextuelle, qui est la probabilité d'une étiquette à suivre une autre.

Une liste de règles grammaticales est préparée à partir de ces statistiques dans le but plus de 90% de précision.

Les outils de désambiguïsation linguistiques sont plus rapides et plus efficaces et fiables que les outils statistiques (HOCEINI, 2011)l'approche linguistique qui n'a besoin que de l'intervention manuelle d'un linguiste, définit un ensemble de règles spécifiques à un domaine particulier. Alors que, les statistiques calculées pour l'apprentissage sont appliquées à n'importe quel domaine.

Néanmoins, les deux approches statistiques et hybrides nécessitent une phase d'apprentissage dans le but est d'apprendre les paramètres requis pour la désambiguïsation. Par conséquent, l'approche hybride est considérée comme la plus efficace et cohérente en termes d'analyse, car elle combine les deux approches et tire profit de leurs avantages.La plupart des désambiguïser morphologiques Arabes ne traitent que la catégorie grammaticale (POS).

4.5.4 Approche basée à l'aide multicritère à la décision.

Le principe de la méthode consiste à réduire, d'emblée, le nombre de scénarios de désambiguïsation en écartant les scénarios dominés (c.-à-d. scénarios ne possédant aucune meilleure évaluation selon tous les critères utilisés) et à classer les scénarios efficaces (c.-à-d. ceux qui ne sont pas dominés) afin de faire émerger le meilleur scénario de désambiguïsation(HOCEINI A. , 2012)celui qui jouit globalement des scores les plus performants selon les différents critères utilisés.

Les étapes de la méthode proposée sont les suivantes :

- **Étape 1 :** La mise en place d'un ensemble qui contient toutes les actions ou solutions (dans notre cas il s'agit des étiquettes ambiguës) possibles
  Soit « E » cet ensembleE= {e1, e2,…, en} où, e1 : est considéré comme étant une étiquette candidate, qui génère systématiquement une information morphosyntaxique ;
- **Étape 2** : Construction d'une famille cohérente de critères F = {f1, f2, …, fp} ;

- **Étape 3 :** Définir une fonction d'évaluation : générer une fonction d'évaluation pour chaque critère. Le résultat est un tableau d'évaluation appelé matrice d'évaluation.
- **Étape 4 :** la pondération et l'agrégation des critères
  - La pondération : consiste à déterminer le poids de chaque critère selon son importance, la méthode de pondération des critères va permettre une discrimination entre les critères en désignant un poids pour chaque critère, ce qui va générer un vecteur de pondération « α ». Pour pondérer les différents critères, nous adoptons la méthode de **l'Entropie**;
  - L'agrégation : Le but est de réduire le nombre d'étiquettes, et de les classer selon leurs scores globaux. Le choix d'une méthode d'agrégation va permettre de normaliser le tableau d'évaluation et facilite une bonne lecture de ce tableau (onobtiendra un tableau « N » ou une matrice « N » normalisé). Afin d'agréger lesdifférentes évaluations d'un scénario calculées selon les critères retenus, nous proposons la méthode **TOPSIS14**.
- **Étape 5 :** Choisir l'étiquette ayant le plus grand score

4.6 Les travaux de domaine

(Gridach, 2012)A fait une étude sur les analyseurs existants dans sa thèse doctorale, les analyseurs morphologiques arabes différents d'un analyseur à un autre selon l'objectif de chacun d'eux et ils possèdent différent objectif : certains sont développés pour la recherche et l'évaluation, par contre les autres analyseurs ont un but commercial.

4.6.1 Analyseur morphologique de Khoja

(khoja, GARSUDE, & KNOWLES, 2001)Ont développés un algorithme linguistique qui analyse les mots arabes voyellés et non-voyellés. L'algorithme utilise une liste de mots de fonction pour les détecter et les filtrer. Cet algorithme utilise aussi le « backtracking » comme un remède pour enlever les erreurs causées par la suppression des affixes. Des modèles simples sont utilisés pour extraire les racines en utilisant une liste préparée des racines pour vérifier la validité des racines extraits. Cet algorithme a pour objectif aussi de trouver une solution pour les cas spéciaux tels que les mots qui contiennent des lettres défectueuses, la gémination ou encore « hamza ». À l'affichage, ce système morphologique donne quelques statistiques très utiles. Ce travail possède des avantages aussi bien que des inconvénients qu'on va les résumer dans le paragraphe suivant.

Ce système possède un avantage majeur concernant le taux de précision et la vitesse d'exécution. L'algorithme de Khoja montre une grande supériorité par rapport aux autres travaux concernant la détection de la racine (Al-Jlayl, 2002).

En plus de ces avantages, ce système possède des inconvénients. Il est incapable de traiter les noms propres. L'inconvénient majeur de ce système reste l'incapacité de traiter les noms étrangers ou encore les analyser correctement. En outre, l'algorithme produit des racines au lieu de fournir les radicaux ou les modèles.

4.6.2 Analyseur morphologique de Buckwalter(BAMA)

L'analyseur morphologique de Buckwalter, connu sous le nom de BAMA (BuckwalterArabicMorphological Analyzer), est l'un des analyseurs morphologiques arabes dans la littérature et il est considéré comme la ressource lexicale la plus respectée dans ce domaine de recherche (hajji, 2000).

D'abord développé en Perl, ce système d'analyse morphologique a été repris avec Java et il adopte un système de translittération du mot arabe. Ce système de translittération a été développé par Buckwalter lui-même. Il est conçu comme une base de données principale de formes des mots qui interagissent avec d'autres bases de données concaténées. Chaque forme de mot est entrée séparément. Il prend comme forme de base le radical (stem), ensuite il fournit des informations sur la racine. Il fournit aussi une traduction anglaise.

Les ressources linguistiques du système sont représentées en adoptant une approche procédurale. Par conséquent, les règles morphologiques sont spécifiées dans le lexique ce qui les rend moins complexes. En ce qui concerne sa base de données lexicale, elle contient trois sous base : la base de données des préfixes qui contient 299 entrées, la base de données des suffixes qui contient 618 entrées et la base de données des lexèmes qui contient 82158 entrées représentant ainsi38600 lemmes (Buckwalter, 2002).

Basé sur la notion de lexèmes, l'analyseur morphologique de Buckwalter utilise une approche d'analyse assez simple voyant que toutes les décisions d'analyse sont codées dans le lexique lui-même et les matrices de comptabilité (Habash & Rambow, 2005). Malgré tout cela, ce (Buckwalter, 2004)système d'analyse souffre d'un certain nombre de problèmes d'analyse(Attia, 2005):

- Cet analyseur morphologique arabe ne se base pas sur la notion de règle. Après chaque entrée, toutes les formes qui appartiennent à cette entrée dans les différents niveaux flexionnels sont listées. Donc, il ne permet pas de capturer les généralités ce qui augmente le coût de maintenance.
- L'analyseur morphologique de Buckwalter n'est pas adapté pour faire la génération. Cela veut dire que c'est impossible de fournir à cet analyseur un ensemble de chaines de caractères pour produire des formes de surface.
- Cet analyseur morphologique manque de couverture concernant les proclitiques qui préfixent les verbes et les noms. Il est conçu de cette manière pour réduire le problème de l'ambigüité, mais il est tombé dans le problème de la couverture qui reste très limité. Les exemples ci-dessous ne sont pas trouvés par ce système morphologique.
  - أَأُحاوِلُ(Est-ce que j'essaie ?)
  - أَمُرَادُ(Est-ce Mourad ?)
- L'insuffisance de couverture de certaines formes à l'impératif: parmi les 9198 Verbes, seuls 22 verbes contiennent des formes à l'impératif. Cette insuffisance de couverture

des formes à l'impératif restreint l'analyse morphologique de ce système. Par exemple, système morphologique ne fournit pas les formes à l'impératif des deux verbes suivants :

- سَامِحْ (pardonne)
- أُتْرُكْ (laisse)

L'insuffisance de couverture de certaines formes à la voix passive. Seulement 15% des verbes contenus dans la base de données lexicale peuvent être analysées. Les exemples ci-dessous ne sont pas analysés par l'analyseur morphologique de Buckwalter.

- يُسَامِحُ (est pardonné)
- يَتْرُكُ (est laissé)

· L'analyseur morphologique de Buckwalter possède un autre inconvénient concernant le préfixe ل/ la/ qui peut préfixer les noms. Cela représente une ambigüité avec la préposition qui possède la même forme et par conséquent augmente le niveau d'ambigüité.

لأَحْزَابْ.

- Manque de spécification de certaines règles morphologiques pour orthographique. L'auteur a confirmé que l'utilisation de certaines règles de formes mal orthographiées par le fait qu'elles sont communes dans les données analysées (Attia, 2005). Les deux exemples ci-dessous concrétisent ce problème d'analyse :
- فاشل(inopérant), l'analyse de ce mot donne aussi le résultat suivant : فأشل (alors je paralyse)
- واقف (être debout),de la même manière, l'analyse de ce mot donne le résultat suivant: وأقف(et je me tiens debout)

Analyseur morphologique de Xerox

L'analyseur morphologique de Xeror développé par Kenneth Beesley est basé sur la technologie à états finis (Finite State Technology) (Dicky, 2003).il adopte une approche d'analyse morphologique basée sur la racine et le schème. IL utilise une base de données contenant 4930 racines et 400 schèmes, d'où la production de 90000 radicaux. Basé sur la notion de règles, l'analyseur morphologique de Xerox permet de faire une large couverture des mots analysés et fournit une traduction en anglais de tous les mots. Malgré les avantages affichés, ce système d'analyse morphologique souffre de nombreux inconvénients dont on cite:

- L'utilisation de règles morphologiques produit une surgénération dans la dérivation des mots. L'analyseur morphologique de Xerox fournit des analyses correctes pour des mots qui ne sont pas des entrées dans les dictionnaires arabes.
- Manque de spécification concernant la classification de chaque mot analysé ce qui le rend incapable d'être utilisé dans l'analyse syntaxique.

- Chaque racine dérivée est autorisée de combiner avec un ensemble sélectionné de formes pour produire juste les mots qui peuvent être trouvés dans les dictionnaires standards arabes. Les mots arabes qui sont morphologiquement possibles et qui ne sont pas stockés dans ces dictionnaires ne sont pas considérés.
  - Les inconvénients présentés par cet analyseur produit un autre problème d'analyse qui est l'ambigüité.
  - Le système morphologique Xerox n'a pas de mécanisme précis pour éliminer l'ambigüité.
  - Le système morphologique est développé par des non-arabophones. Alors, la couverture du système final est moins excellente surtout lorsque le système est testé avec une littérature orientale ou encore un ancien texte arabe.

4.6.3 Analyseur morphologique ElixiarFMd'OtakarSmrž

AlixirFM(smrž, 2007) est un analyseur morphologique développé par Otakar Smrž. c'est une implémentation open source de haut niveau de la morphologie fonctionnelle (functionalmorphology) arabe réutilisable par les diverses applications de TALN. II permet d'analyser les textes de l'Arabe standard moderne. Il utilise une base de données morphologique inspirée de celle de l'analyseur morphologique de (Buckwalter, 2002). Cependant, ce système (bielický, 2009)a eu une considérable correction computationnelle et lexicographique ainsi qu'une considérable raffinement et extension. ElixirFM est développé dans le cadre du projet de la « DependancyTreebank » de l'arabe à Prague en utilisant le langage Haskell, par contre les interfaces sont développées avec le langage Perl.

L'Implémentation d'interface d'ElixirrFM (voir. 2.1) peut servir comme un exemple de base de données lexicale. Cette interface fonctionne en quatre modes différents :

a) Le mode de résolution fournit l'analyse lexicale et l'analyse morphologique du texte inséré. L'utilisateur a le choix entre trois options de notation: Unicode, Buckwalter ou encore arabe Tex
b) Le mode flexionnel offre des paradigmes complets ou partiels des formes fléchies pour un lexème donné.
c) Le mode dérivé fournit des informations dérivées (verbes, nom verbal, voix active et passive) pour une classe particulière de lexèmes.
d) Le mode lookup permet de chercher le lexème désiré ainsi que toutes les racines dérivées existantes et possibles. Ce mode permet de faire la traduction en anglais du lexème fourni.

Analyseur morphologique MAGEAD de NizarHabash

L'analyseur morphologique MAGEAD développé par NizarHabash est l'un des analyseurs morphologiques arabes les plus connus dans la littérature ces derniers temps, et il est destiné pour la recherche et l'évaluation. C'est l'un des analyseurs morphologiques se basant sur les systèmes de la morphologie fonctionnelle,la base de données lexicale de

MAGEAD est construite en étendant celle de l'analyseur morphologique ElixirFM d'Otakar Smtz .

Parmi les avantages de cet analyseur morphologique, c'est qu'il permet d'analyser les mots de la morphologie des dialectes qui est considérée comme un nouveau travail dans le domaine de l'analyse morphologique arabe. Il permet aussi, en plus de l'analyse morphologique, de faire la génération. Mais ; malheureusement il a besoin d'une base de données lexicale complète représentant les dialectes arabes pour faire de l'évaluation un processus intéressant et convaincant.

4.6.4 Analyseur morphologiqueSebawai de Darwish

Sebawai est un analyseur morphologique arabe développé par Darwish en 2003 en un seul jour et exactement comme a été indiqué par le développeur en 12 heures et avec 200 lignes de code en utilisant les langages de programmation Perl. Le système permet de trouver les racines des mots avec un taux de réussite de 84%. En plus, cet analyseur est capable de dériver des racines de 40000 mots par minute (Darwish, 2002).

Ce système morphologique utilise l'approche hybride. Cette approche est basée sur la notion de règles en conjonction avec les statistiques. Elle emploie une liste de préfixes, une liste de suffixes et des schèmes pour transformer un radical en une racine. Les combinaisons préfixe-suffixes-modèle possible sont construites pour un mot afin de dériver toutes les racines possibles. La dérivation manuelle des règles reste une opération difficile et finalement nécessite une bonne connaissance des règles orthographiques et morpho tactiques de la langue arabe.

Ce système utilise deux modules principaux :

- ✓ Le premier module utilise une liste de couple mot racine arabe pour :
  - o dériver une liste de préfixes et suffixes ;
  - o pour construire des modèles des radicaux ;
  - o pour calculer la probabilité d'apparition d'un préfixe, un suffixe ou encore un schème.
- ✓ Le second module accepte des mots arabes en entrée, essaie de construire toutes les combinaisons préfixe-modèle-suffixe possibles et affiche une liste classée de toutes les racines possibles.

Cependant, il est restreint dans les aspects suivants :

- ✓ Basé sur la notion de la racine, ce système est incapable d'analyser les mots arabes provenant des autres langues étrangères.
- ✓ Dans certains cas de la langue arabe, on peut trouver des racines qui contiennent entièrement une seule lettre longue, mais ces racines originales contiennent trois lettres. Par exemple, le mot qui veut dire « protéger » à l'impératif. Ce système est incapable d'analyser cette catégorie de mots.

- ✓ Il est incapable d'analyser des mots qui constituent des phrases complètes .par exemple le mot agglutiné« أَسَتَتَذَكَّرُونَنَا » qui représente une phrase complète en arabe ne peut pas être analysé par ce système.
- ✓ Cet analyseur est incapable de déchiffrer les combinaisons préfixe-suffixe qui sont légales. Bien que déchiffrer des combinaisons légales reste théoriquement faisable avec les statistiques, ce processus nécessite un très grand nombre des exemples pour s'assurer que le système ne rejette pas les combinaisons légales.

4.6.5 Analyseur morphologique d'Hilal

L'analyseur morphologique d'Hilal est basé sur trois classes : une classe des préfixes, une classe des suffixes et une troisième classe utilisant le nombre de lettres qui reste dans le mot. D'après cette approche, les mots arabes sont classés en mots outils et mots ordinaires. Les mots ordinaires suivent les règles grammaticales. Cependant, les mots outils ne suivent pas ces règles.

La méthode d'extraction des racines trilitères à partir d'un mot ordinaire suit les étapes générales citées ci-dessous :

a) Le préfixe et le post fixe le plus long possible sontéliminés en comparant les caractères les plus dominants et trainant avec les préfixes et les suffixes les plus connus.
b) Selon la longueur de la partie restante, les différentes règles sont examinées :
   - Éliminer les lettres étrangères
   - Modifier une lettre supplémentaire afin de former la racine trilitère
   - Changer une lettre à sa valeur originale

De nombreuses consultations des tables sont utilisées pour accomplir cette tâche, y compris les tables des modèles, des racines, des préfixes, des suffixes ainsi que les racines non standardisées (Al-Fedaghi, 1989).

4.6.6 Analyseur morphologique de Hegazi and El-Sharkawi

(Hegazi & El-sharkawi, 1986)A proposé un algorithme de la hiérarchie morphologique assisté par ordinateur (computer-AidedMorphologicalHierarchy (CAMH)) pour dériver la racine d'un mot arabe voyellé, son modèle morphologique et les attributs morphologiques. D'après (Hegazi & El-sharkawi, 1986), les approches que les linguistes arabes ont utilisées pour s'occuper de la morphologie arabe sont les approches que les linguistes arabes ont utilisées pour s'occuper de la morphologie arabe sont les suivantes :

- La première approche utilise les règles morphologiques qui nécessitent des techniques de mémoire et qui sontdépendantes de la connaissance humaine accumulée.
- La seconde approche utilise des règles phonétiques qui peuvent produire des résultats incorrects. D'après Hegazi et El-Sharkawi, les deux approches peuvent produire de meilleurs résultats. Ils ont utilisé leur algorithme pour construire un

analyseur lexical automatique qui nécessite une liste de racines et affixes arabe, des morphologiques ainsi que des mots de fonction.

L'organigramme de cet analyseur morphologique n'a été fourni avec aucune explication. Il est basé sur les règles morphologiques et il prend en considération les mots étrangers, les erreurs d'orthographe et il nécessite que les mots soient voyelles.

Parmi les avantages principaux de ce système, c'est qu'il permet de fournir un bon outil pour la traduction automatique. Par contre, cet analyseur n'est pas bien expliqué et il n'a fourni aucune évaluation ainsi qu'aucunrésultatexpérimental.

4.6.7 Analyse morphologique de Thalouth et Al Dannan

(Dannan, 1987)a proposé un algorithme basé sur des règles linguistiques solides utilisant un système étendu de pattern-matching. Pour concrétiser cette approche, ils ont créé une liste de préfixes, de suffixes, de racines solides de mots, de modèles, de mots étrangers et des fonctions de mots en utilisant des études statistiques. Ces listes sont accompagnées de large nombre de flags tels que la comptabilité entre les racines et les modèles, la comptabilité entre les modèles et les suffixes et les préfixes ainsi que la comptabilité entre les préfixes et les suffixes.

Le but était de réaliser un analyseur et générateur pour les mots arabes non-voyelle. L'algorithme proposé permet de modifier les modèles prédéfinis crées par les linguistes arabes. Il utilise les équilibres morphologiques computationnels qui surmontent quelques problèmes tels que la mutation et la vocalisation. Dans cet algorithme, les mots arabes sont placés dans deux catégories principales qui sont respectivement les mots de fonction et de contenu. Les mots de fonction consistent principalement d'environ 200 prépositions et pronoms. Avec les suffixes, le nombre passera à environ 600 mots. Les mots de contenu sont tous les mots arabes y compris les mots étrangers (Al-kharashi & Al-Sughaiyer).

L'avantage majeur de ce travail se concrétise principalement dans l'approche d'analyse morphologique du texte arabe non-voyelle(Al-Kharashi, 1991) et la construction claire et isolée des listes (Al-Uthman, 1990).par contre, cet analyseur possède des inconvénients qui se concrétisent dans l'inefficacité et l'utilisation de nombreuses listes (Al-Uthman, 1990) (El-Affendi, 1998).la nécessité d'un processus long pour faire back-tracking afin de supprimer les affixes (El-Affendi, 1998)et enfin l'absence d'une évaluation expérimentale ainsi que l'absence du taux de succès.

4.6.8 Analyseur morphologique d'Al-Fedaghi et Al-Anzi

Al-Fedaghi et Al-Anzi ont développé un algorithme (Al-Fedaghi, 1989)pour examiner chaque combinaison des trois lettres du mot fourni et produit sa présentation racine-modèle. L'algorithme utilise deux fichiers en entrée, le fichier des racines trilitères et le fichier des modèles. L'algorithme examine ces modèles qui possèdent la même longueur que le mot fournit en entrée. Le concept principal de cet algorithme reste d'examiner les modèles pour identifier les positions des lettres de la racine. Ces lettres sont ensuite testées

pour décider s'ils forment une racine arabe ou non. L'algorithme a été testé dans quatre modes qui sont les suivants :

- **Mode 1 :** le mot d'entrée contient sa racine trilitère complète.
- **Mode 2** : la troisième lettre de la racine du mot d'entrée est perdue.
- **Mode 3 :** une lettre parmi les lettres de la racine du mot d'entrée est manquante.
- **Mode 4 :** deux lettres de la racine du mot d'entrée sont manquantes.

Ils comparent chaque mode en termes de sa vitesse et sa capacité de réduire le mot en sa racine. Ils ont trouvé que le quatrième mode prend plus d'une heure pour réduire le mot en sa racine, par contre les autres modes prennent entre une et cinq minutes pour faire ce traitement. En ce qui concerne la réduction d'un mot en sa racine, le premier mode atteint 63%, le deuxième mode atteint 70 % et le troisième mode atteint 79 %.

4.6.9 Analyseur morphologique Multi-Mode

Analyseur morphologique Multi-Mode (Multi-ModeMorphological Processor (MMMP)) a été développé en 1986 par la société Electrique d'Al-Alamaih en Caire (Al-AlamaihElectronicCompany). Il était l'un des analyseurs les plus connus dans la littérature. Cela est dû au fait que cet analyseur morphologique arabe est utilisé pour la recherche d'informations ce qui n'est pas le cas pour d'autres analyseurs et il est capable d'analyser n'importe quel mot arabe quel que soit son niveau de voyellation dans ses primitives dérivationnelles, flexionnelles, affixes (Nabil, 1992). Les algorithmes de l'analyseur morphologique multimode sont conçus de telle sorte qu'ils vont être adaptés aux différentes applications de traitements automatiques du langage naturel tel que l'analyse du texte, la recherche d'information et la traduction automatique.

L'année suivante (1987) a connu une incorporation de cet analyseur dans la AFTDB (Arabic Full Text Data Base) et il est implémenté pour la recherche du texte du Coran. L'AFTDB en 1989 pour développer la base de données d'Al-Hadith. Depuis, le groupe d'Al-Alamiah a utilisé l'AFTDB comme un moteur de base de données pour la production de plusieurs textes arabes.

Les développeurs de l'analyseur MMMP ont conçu ce système pour qu'il soit capable d'émettre des hypothèses d'un certain nombre de traitements pour un mot donné, en analysant le mot et ensuite en le resynthétisant. Si l'analyse morphologique du mot a donné un résultat positif, ensuite le traitement est considéré correct. Sinon, l'analyse devrait répéter le même processus, mais avec un nouveau traitement.

Brièvement, la fonction principale de l'analyseur morphologique MMMP est de fournir toutes les formes valides voyelles d'un mot en entrée. Pour chaque voyellation, il fournit une décomposition morphologique du mot (affixes, radicaux). Plus loin, le radical est analysé en termes de sa racine et ces formes morphologiques. Le système MMMP est divisé en quatre parties qui sont les suivantes :

- Processus Morpho-Syntaxique
- Processus dérivationnel
- Processus de Parsing
- Processus de voyellation

4.6.10 Analyseur morphologique Morpho3 d'Attia

Ce système morphologique a été développé dans le laboratoire RDI(Attia, 2005). Morpho2 et Morpho3 développés par Attia sont considérés comme améliorations de Morpho1 développé par Nagy Fathy Mohammad(fathi, 1995). Dans sa forme finale, l'analyseur morphologique Morpho3 possède les avantages suivants :

- Chaque racine dérivée est autorisée de combiner avec n'importe quelle autre forme tant que cette combinaison est autorisée. Ceci permet à l'analyseur morphologique Morpho3 de traiter tous les mots arabes possibles et d'éliminer la nécessité d'être lié à un vocabulaire fixe.
- Le modèle morphologique de l'analyseur Morpho3 est un modèle homogène. Il traite de la même façon les mots dérivés réguliers, les mots irréguliers et les mots fixes.
- D'après l'auteur, ce système morphologique couvre tous les phénomènes de la morphologie arabe. Il traite les textes arabes modernes aussi bien que les anciens textes arabes. Il peut traiter aussi les textes arabes en relation avec le business, la littérature ainsi que les textes scientifiques.

4.6.11 Analyseur morphologique G-LexAr

Le système G-LexAr est un programme d'analyse morpho-grammaticale de l'arabe destiné pour l'arabe classique ainsi que l'arabe standard moderne (Modern Standard Arabic), pouvant traiter des textes d'entrée voyelles ou non. Il produit en sortie des analyses où les mots peuvent être indépendamment segmentés, voyelles, lemmatisés ou étiquetées. Il est fondé sur la mise en œuvre d'un grand nombre de dictionnaires et règles qui privilégient la rapidité des traitements à l'espace mémoire (Mekki, 2011).

D'après les auteures, l'analyseur morphologique G-LexAr opère en trois étapes :

- La segmentation du texte d'analyse en unités morphologique est considérée comme la première étape de ce système d'analyse. Le résultat de segmentation consiste à dégager les formes simples et agglutinées de l'arabe (hyper-forme). Ensuite, les chaines de caractères qui ne relèvent pas de l'analyse morphologique de l'arabe sont filtrées dans l'étape suivante.
- L'analyse de ces hyper-formes indépendamment de leur contexte est réalisée en seconde étape. À chaque hyper-forme est attribué, sous forme d'un arbre, l'ensemble de ses segmentations, voyellations, lemmatisations et étiquettes grammaticales possibles. Lerésultant de cette deuxième étape est une succession de mots accompagnés de leurs arborescences lexicales.

- La dernière étape de ce système est l'étiquetage grammatical en élaguant ces arborescences lexicales. Il consiste à garder que les branches dont les feuilles ont des étiquettes. Ces étiquettes permettent d'examiner un certain nombre de règles portant sur la légalité de la succession de ces étiquettes. Un dictionnaire précompilé est mis en œuvre afin de rendre le traitement encore plus rapide (Mekki, 2011).

En conclusion, l'analyseur morphologique G-LexAr est destiné spécialement pour faire la traduction automatique ce qui représente l'une de ses insuffisances, c'est-à-dire cet analyseur est incapable d'être utilisé pour d'autres applications de TALN. Cet analyseur morphologique affiche une faiblesse concernant l'indexation des mots d'emprunt ce qui représente un autre inconvénient

4.6.12 Analyseurs morphologique de DAVID COHEN :

Dès les années 60, le premier essai d'analyse automatique arabe a vu le jour par David Cohen dans la revue de l'Association pour le Traitement automatique des Langues naturelles(ATALA) et intitulé « Essai d'une analyse de l'arabe » . Il a proposé un analyseur automatique arabe qui a été repris en LISP par J.MACCARTHY puis en COMIT par Y.YNGVEV. Dans cet essai, DAVID COHEN propose un schéma général des mots graphiques appelés « mots maximaux ». L'analyse d'un mot maximal permet de dégager les proclitiques, les préfixes, la base, les suffixes et les enclitiques. Notons que les préfixes, la base et les suffixes forment ce que David Cohen appelle « un mot minimal ».

Par la suite, David Cohen a mené une étude sur la dérivation des noms en Arabes où il a enregistré que la formule de croisement de racines et schéma était insuffisante pour une bonne partie du lexique arabe d'où la nécessité d'autres constituants comme les suffixes pour la construction des mots. Malgré ces défaillances de l'analyse, cette première expérience est souvent considérée comme un premier essai d'analyse automatique arabe (Mesfar, 2008)

4.6.13 Autres analyseurs morphologiques arabes

Plusieurs autres travaux d'élaboration d'analyseurs morphologiques et morphosyntaxiques ont vu le jour, mais sans jamais être développés pour la communauté scientifique pour l'évaluation. Nous pouvons encore citer les travaux d'analyse morphologique arabe suivants:

- AMSAAR (Analyseur MorphosyntaxiqueAssisté de l'Arabe) est un système d'analyse morphosyntaxique conçu et développé au CNRS par une équipe dérigée par Débile. Ce système d'analyse permet de faire le contrôle et la correction des mots arabes analysés en faisant une intervention à chaque étape d'analyse morphosyntaxique(Mesfar, 2008).
- Le système d'analyse morphologique de Khalid Shaalan est considéré parmi les anciens travaux dans le domaine de l'analyse automatique de l'arabe. Il repose sur

les principes de l'intelligence artificielle où il a développé un ensemble de règles en SICStus Prolog (Shaalan, 1989).

- En appliquant la méthode d'ÉricBrill sur la langue arabe, Freeman a développé un étiqueteur pour l'établissement d'un système d'apprentissage de la langue en utilisant 146 étiquettes pour étiqueter des lexèmes (Mars, 2008).
- L'analyseur morphologique de l'équipe LIDILEM : cet analyseur morphologique développé par le laboratoire LIDILEM à l'université Stendhal de Grenoble sert pour la création automatique et semi-automatique d'activité pédagogique pour l'apprentissage de l'arabe [Mars and Belgacem, 2006]. Cet analyseur est développé en utilisant une base de données linguistique sous forme de dictionnaires.

4.7 Conclusion

Dans ce chapitre, nous avons présenté les étapes d'analyse morphologique. Puis, nous avons présenté une étude sur les techniques de la morphologie. Ensuitenous avons introduit la théorie d'ambiguïté suivie par une étude sur les approches de désambigüité. Nous terminons ce chapitre par donner une description surles différents analyseurs morphologiques arabes les plus connus dans les littératures. Généralement, on peut distinguer entre deux catégories d'analyseurs morphologiques : la première concerne les analyseurs morphologiques développés pour la recherche et l'évaluation et ils sont disponibles et téléchargeables sur internet, elle regroupe l'analyseur morphologique de Buckwalter connu sous le nom de BAMA, le système de Xerox d'analyse morphologique et de génération développé par Beesley, l'analyseur morphologique MAGEAD développé par NizarHabash de l'université américaine de Columbia et le système morphologique Elixir FM développé par Otakar Smrz. La deuxième catégorie concerne les analyseurs développés pour des laboratoires de recherche privés ou encore pour des raisons commerciales.

Chapitre 4 : Développement d'un analyseur morphologique

5.1 Introduction

La plupart des applications qui mettent en jeu le texte ou encore des applications de traitement automatique du langage naturel notamment l'analyse syntaxique, l'analyse sémantique, la traduction automatique, la recherche de l'information, la correction orthographique réalisent de performances remarquables en intégrant de meilleurs et de robustes systèmes d'analyse morphologique.

Un analyseur morphologique de n'importe quelle langue naturelle est défini comme étant un programme qui permet de connaitre un mot sous les diverses formes qu'il peut prendre dans une phrase quelconque pour chaque forme analysée, l'analyseur doit déterminer les valeurs morphologiques du mot analysé telles que la catégorie du mot(nom, verbe, etc.), le genre, le nombre, la voix, les informations sur les clitiques et les traits morphosyntaxiques.

Il convient de signaler que la morphologie arabe est très riche par rapport aux autres langues, en particulier les langues indo-européennes. Elle repose à la fois sur la morphologie flexionnelle et dérivationnelle. Cette richesse rend le processus d'analyse difficile à traiter et à accomplir. L'analyse morphologique arabe est une étape essentielle dans le traitement automatique de l'arabe, elle est devenue aujourd'hui incontournable dans le développement des technologies de l'information. Cette étape n'est pas une tâche facile parce que l'arabe présente des particularités qui rendent le processus d'analyse difficile par rapport aux autres langues.

Actuellement, plusieurs travaux sur l'analyse morphologique arabe ont vu le jour (Al-kharashi & Al-Sughaiyer).La plupart des analyseurs ne traitent pas toutes les fonctionnalités du mot analysé et certains d'entre eux sont destinés pour certaines applications. Notons que les analyseurs existants ont des buts différents, certains d'entre eux ont un objectif commercial, les autres sont développés pour la recherche et l'évaluation (Attia, 2005).

Nous commençons ce chapitre par présenter les défis et les objectifs de l'analyse morphologique arabe. Nous présentons le processus d'analyse dans la section suivante en faisant une description détaillée des différentes étapes de notre système d'analyse.

5.2 Défis et objectifs de l'analyse morphologique arabe

La langue arabe est une langue arabe sémitique différente par rapport aux autres langues(les langues Indo-Européennes). Cette différence se manifeste dans plusieurs niveaux : morphologique, syntaxique et sémantique.

Autrement dit la langue arabe est une langue fortement flexionnelles et dérivationnelle et son vocabulaire peut être étendu facilement en utilisant un Framework latent en ce qui concerne l'utilisation créative des racines et des schèmes. Notons que 85% des mots arabes sont dérivés à partir des racines trilitères et il existe à peu près 10000 racines indépendantes (Al-Fedaghi, 1989).

Par conséquent, les mots arabes sont développés en appliquant des schèmes aux racines. La racine est une séquence ordonnée de trois ou quatre caractères valides de l'alphabet et rarement cinq caractères. Un schème est une séquence ordonnée de caractères dont certains sont constants et d'autres sont variables. Les caractères variables sont substitués par les caractères de la racine pour générer un mot nommé radical. Signalons que le schème et la racine ne sont pas des mots arabes valides, par contre, le radical est un mot valide.

Un autre aspect qui caractérise la langue arabe est la possibilité d'emprunter des mots étrangers et de les ajouter à son système de morphologie dérivationnelle comme «أكسدة».

L'analyse morphologique arabe fait face à de nombreux défis. Le premier concerne la qualité d'analyse et les informations fournies. Le second, son utilisation dans les systèmes de TALN tels que les systèmes de traduction automatique. Koehn et hoang ont montré que les modèles de traduction contenant l'information morphologique aboutissent à une meilleure performance. Par conséquent, l'analyse morphologique devient plus importante quand on traduit des langues morphologiquement très riches telles que la langue arabe (koehn & hoang, 2007). Le troisième concerne son importance comme étape principale avant l'analyse syntaxique ou encore l'analyse sémantique.

Récemment, les travaux de recherches et de développement des systèmes de traitement automatique ont montré que l'analyse morphologique de n'importe quel mot consiste à déterminer de nombreuses valeurs telles que la catégorie du mot, le genre, la personne, le nombre, la voix, des informations, concernant les clitiques, etc. (Habash & Rambow, 2005).

L'analyse morphologique arabe est considérée comme l'une des tâches les plus difficiles par rapport aux autres langues. Les causes de cette difficulté sont la longueur et la complexité des phrases, la structure des mots est très complexe et morphologiquement ambiguë et cela est dû à l'utilisation fréquente d'un très grand nombre d'affixes, ainsi que le phénomène d'agglutination.

5.3 Processus d'analyse morphologique

L'analyse morphologique d'un texte arabe en utilisant notre système comprend cinq étapes principales. En entrée, le système d'analyse accepte un texte arabe. La première étape consiste à lire le fichier qui contient le teste arabe. Puis, un programme de segmentation du texte arabe en mots. La troisième étape s'impose pour faire un chargement de toutes les listes (affixes, schèmes, racines, mots spéciaux). L'étape suivante est considérée une étape primordiale dans l'analyse morphologique, on parle alors du Stemming. Ensuite, l'étape de validation des vecteurs résultants de l'étape précédente, enfin de générer les résultats. La figure suivante l'architecture de notre système.

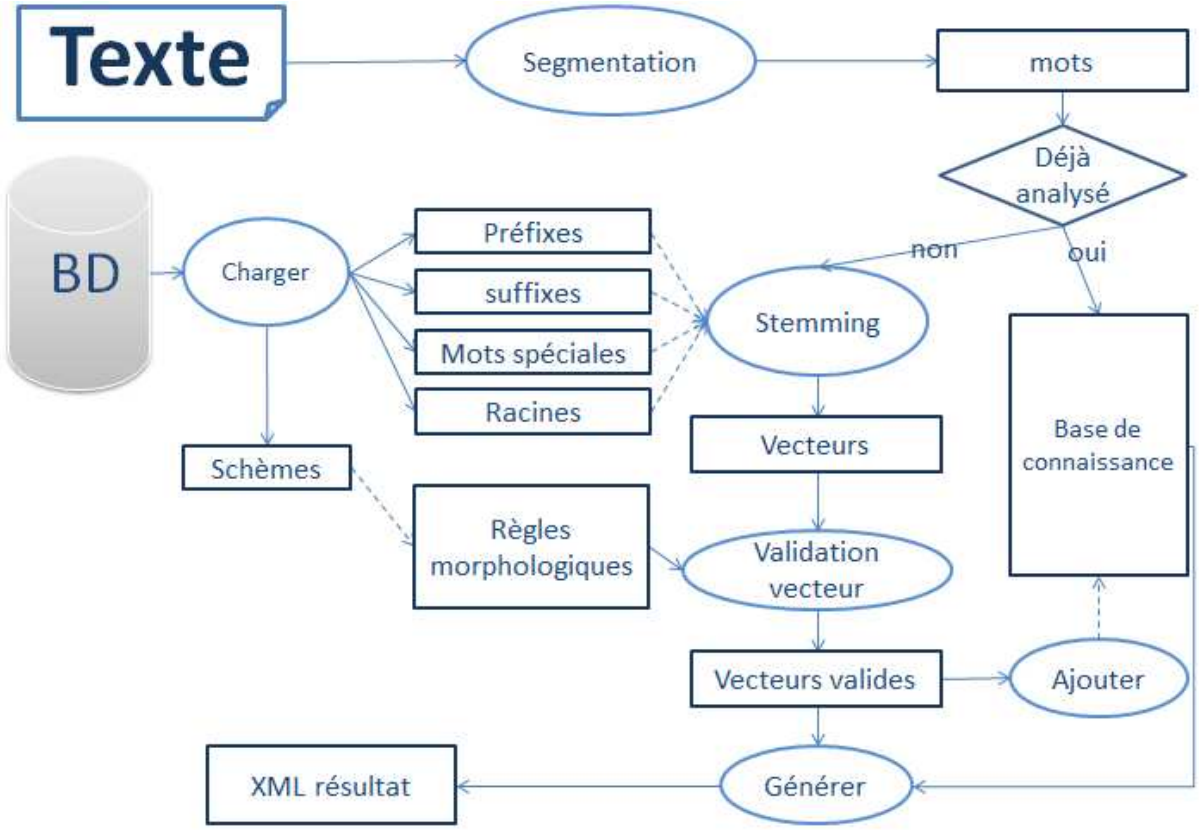

Figure 22: schéma complet de notre analyseur morphologique

5.3.1 Lecture du fichier

La première étape de l'analyseur morphologique consiste à lire le fichier contenant les phrases arabes ou les mots arabes. Signalons que le texte arabe fournit en entrée est un texte arabe voyellé. Pour lire ce fichier, nous avons utilisé les classes du langage java pour a lecture d'un fichier texte.

5.3.2 Segmentation du texte arabe

Pour faire l'analyse morphologique, l'étape de la segmentation du texte reste une étape importante dans le traitement automatique des langues naturelles. Son objectif principal est de segmenter le texte en unités définies et repérées auparavant. Le processus de la segmentation des textes en traitement automatique des langues naturelles consiste à

délimiter les segments de ses éléments de base qui sont les caractères, en éléments de différents niveaux structurels : paragraphe, phrase, syntagme, mot graphique, mot-forme, morphème.

Il convient d'indiquer que l'analyse automatique de la langue arabe est un processus difficile à accomplir et cela aux différents niveaux d'analyse. Ces difficultés viennent des particularités de la langue arabe et de sa morphologie. Citons l'exemple de l'agglutination qui reste parmi les défis de l'analyse automatique de l'arabe.

Selon le but de l'analyse à effectuer qu'elle soit morphologique ou syntaxique, on peut généralement distinguer entre trois grands types d'application de la segmentation (Zoubeir, 2008) :

- ✓ **tokenization:** elle est appelée aussi segmentation lexicale, et elle consiste à segmenter le texte en mot ou unité lexicaux(tokens).
- ✓ **La segmentation morphologique :** consiste au contraire de la segmentation lexicale, à segmenter le texte en unités distinctes appelées les morphèmes.
- ✓ **La segmentation syntaxique :** a pour objectif la segmentation des différents constituants du texte en unités indépendantes, supérieures aux mots, comme les propositions.

Dans notre cas(l'analyse morphologique), le premier type de segmentation sera notre objectif. La segmentation lexicale ou itération consisteà segmenter un texte en mots-formes ou unités lexicales. C'est une opération consistant à structurer le texte en passant d'un ensemble continu de caractères à une suite discrète d'items lexicaux. Par conséquent, notre programme de segmentation permet de fournir des unités lexicales (tokens) qui seront analysées dans les prochaines étapes de l'analyse.

5.3.3 Chargement de ressources

Avant de commencer la procédure de Stemming on doit charger les ressources nécessaires qui sont :

Préfixes

Tous les préfixes sont rassemblés dans un fichier XML ou chaque préfixe est défini comme un nœud de qui a les attribue suivant:

- ✓ **Unvoweledform :**représente le préfixe non voyellé.
- ✓ **Voweledform :**représente le préfixe voyellé.
- ✓ **Desc :**description du préfixe
- ✓ **Classe :**des informations sur le préfixe comme s'il est verbal ou nominal.

La figure suivante présente un extrait sur la ressource des préfixes.

```
<prefixe unvoweledform="و" voweledform="وَ" desc="حرف العطف" classe="C1">
</prefixe>
<prefixe unvoweledform="ف" voweledform="فَ" desc="حرف العطف أو الاستئناف" classe="C1">
</prefixe>
<prefixe unvoweledform="أ" voweledform="أَ" desc="همزة الاستفهام" classe="C2">
</prefixe>
<prefixe unvoweledform="أو" voweledform="أوَ" desc="همزة الاستفهام" classe="C2">
</prefixe>
<prefixe unvoweledform="أف" voweledform="أفَ" desc="همزة الاستفهام" classe="C2">
</prefixe>
<prefixe unvoweledform="ل" voweledform="لَ" desc="حرف الابتداء" classe="C3">
</prefixe>
<prefixe unvoweledform="ول" voweledform="ولَ" desc="حرف العطف+حرف الابتداء" classe="C3">
</prefixe>
<prefixe unvoweledform="فل" voweledform="فلَ" desc="حرف العطف+حرف الابتداء" classe="C3">
</prefixe>
<prefixe unvoweledform="ال" voweledform="ال" desc="التعريف" classe="N1">
</prefixe>
```

Figure 23: extrait de ressource des préfixes

Suffixes

Tous les suffixes sont rassemblés dans un fichier XML ou chaque suffixe est défini comme un nœud de qui a les attribue suivant:

- ✓ **Unvoweledform :**représente le suffixe non voyellé.
- ✓ **Voweledform :**représente le suffixe voyellé.
- ✓ **Desc :**description du suffixe
- ✓ **Classe :**des informations sur le préfixe comme s'il est verbal ou nominal.

La figure suivante présente un extrait sur la ressource des suffixes.

```
<suffixe unvoweledform="ك" voweledform="كَ" desc="ضمير المخاطب" classe="C2">
</suffixe>
<suffixe unvoweledform="ك" voweledform="كِ" desc="ضمير المخاطب" classe="C2">
</suffixe>
<suffixe unvoweledform="كما" voweledform="امك" desc="ضمير المخاطبين" classe="C2">
</suffixe>
<suffixe unvoweledform="كم" voweledform="كمْ" desc="ضمير المخاطبين" classe="C2">
</suffixe>
<suffixe unvoweledform="كن" voweledform="كنَّ" desc="ضمير المخاطبات" classe="C2">
</suffixe>
<suffixe unvoweledform="ه" voweledform="ه" desc="ضمير الغائب" classe="C3">
</suffixe>
<suffixe unvoweledform="ها" voweledform="اه" desc="ضمير الغائبة" classe="C3">
</suffixe>
<suffixe unvoweledform="هما" voweledform="اهم" desc="ضمير الغائبين" classe="C3">
</suffixe>
<suffixe unvoweledform="هم" voweledform="همْ" desc="ضمير الغائبين" classe="C3">
</suffixe>
```

Figure 24: extrait du ficher des suffixes

Mots d'outil

Ce fichier rassemble tous les mots outils comme les verbes d'affirmation « إن » avec ses attributs :

- ✓ Unvoweledform : représente le mot outil non voyellé
- ✓ Voweledform : représente le mot outil voyellé
- ✓ Type : représente la description de ce mot
- ✓ Prefixeclass : représente la classe des préfixes qui peut être liée avec ce mot
- ✓ Suffixeclass : représente la classe des suffixes qui peut être liée avec ce mot

Prenons comme exemple le mot « على » est enregistré dans ce fichier par le balise suivante :

<toolwordunvoweledform="على"voweledform="عَلَى"type="حرف جر" prefixeclass="C1C2C3C4" suffixeclass="C1C2C3N1" ">

D'où le mot est « على » , la forme voyellé est « عَلَى» , son type est « حرف جر» , et il peut être accompagné par les préfixes de la classeC1, C2,C3 et C4 et par les suffixes de la classe C1,C2,C3 et N1.

Propernouns

Ce fichier contient presque 8400 noms propres « اسماء علم»avec leur type

```
<propernoun unvoweledform="سنبل"  type="اسم علم"/>
<propernoun unvoweledform="سندس"  type="اسم علم"/>
<propernoun unvoweledform="سها"  type="اسم علم"/>
<propernoun unvoweledform="سهاد" type="اسم علم"/>
<propernoun unvoweledform="سهى"  type="اسم علم"/>
<propernoun unvoweledform="سهبة" type="اسم علم"/>
<propernoun unvoweledform="مقديشي"  type="نسبة إلى اسم علم"/>
<propernoun unvoweledform="مقديشية"  type="نسبة إلى اسم علم"/>
<propernoun unvoweledform="ملاوي"  type="نسبة إلى اسم علم"/>
<propernoun unvoweledform="ملاوية" type="نسبة إلى اسم علم"/>
```

Figure 25: Extrait du fichier des noms propres

Racines

Ce fichier contient presque 4000 racines arabes, ordonnées selon les lettres alphabétiques, ce fichier a pour objectif de valider le résultat de Stemming.

Schèmes

Ce fichier contient tous les schèmes arabes, avec la règle permettant d'extraire la racine afin de le valider par la liste des racines.

```
<pattern value="أفاعل" rules="245" />
<pattern value="أفعاء" rules="233 23و"/>
<pattern value="أفعال" rules="235" />
<pattern value="أفعلة" rules="234" />
<pattern value="أفعلي" rules="234" />
<pattern value="أفعول" rules="235" />
<pattern value="أفعيا" rules="233" />
<pattern value="أفعية" rules="233" />
<pattern value="أفعيو" rules="233" />
<pattern value="أفعيي" rules="233" />
<pattern value="أفيعل" rules="245" />
<pattern value="إفالة" rules="2و4 2ي4" />
<pattern value="إفعاء" rules="23 و23ي" />
<pattern value="إفعال" rules="235" />
<pattern value="إفعلة" rules="234"/>
```

Figure 26:Extrait du fichier schème

5.3.4 Stemming

Cette partie est unepartie primordiale dans l'analyse morphologique, puisqu'il a un rôle de distinguer entre le radical et les affixes. Dans cette partie nous comparons les listes d'affixes pour qu'on puisse arriver à tenir le triplet (préfixe, radical, suffixe).

Le résultat obtenu par cette étape peut être sous plusieurs triplets à cause de l'agglutination. Prenons comme exemple le mot « وقف » donne les résultats suivante :

	suffixes	Stem	Préfixes
أتتذكروننني	---	أتتذكروننني	---
	---	تتذكروننني	أ
	ي	أتتذكرونن	---
	ني	أتتذكرون	---
	ني	تتذكرون	أ
	ي	تتذكرونن	أ

Tableau 18 : Exemple de stemming du mot "أتتذكروننني "

5.3.5 Validation des segments

Cette étape a pour but de filtrer les vecteurs résultants de l'étape précédente, parmi ces résultats il y a des résultats qui sont fausses, pour faire cela on doit appliquer des règles morphologiques.

Parmi ces règles, il y a la correspondance entre les affixes nominales et verbales, selon la classe d'affixe, par exemple le mot « الفواكه » .

	vecteur	suffixes	Radical	préfixes
الفواكه	1	---	الفواكه	---
	2	---	فواكه	ال
	3	ه	الفواك	---
	4	كه	الفوا	---
	5	كه	فوا	ال
	6	ه	فواك	ال

Tableau 19: Exemple de Stemming du mot"الفواكه "

Pour le cas de ce mot, la validation élimine les vecteurs « 3,4,5,6 » à causes des règle suivantes :

- Les suffixes « كه,ه » sont des suffixes verbaux, il sont impossible d'être avec les préfixes nominaux.
- Aucun schème adéquat avec le radical.
- Le radical n'existe pas ni dans la liste des racines, ni dans les noms propres, ni dans les mots outils.
- …etc

5.3.6 Génération des résultats

La dernière étape de notre système est l'affichage des résultats (Vecteurs) associée à chaque mot, notre résultat est affiché dans un fichier XML, pour qu'on puisse intégrer dans le système de traduction de texte en langue arabe vers la langue des signes arabes.

```
<texte>
    <phrase value="الجملة">
        <mot value="الكلمة">
            <exception description="نوع الكلمة">
            </exception>
            <candidat>
                <prefixe value="" description=""></prefixe>
                <stem value="" scheme="" racine=""></stem>
                <suffixe value="" description=""></suffixe>
            </candidat>
        </mot>
    </phrase>
</texte>
```

Figure 27: La structure de fichier de résultat

La figure précédente présente le fichier XML de résultat.

5.3.7 Mise à jour de la base de connaissance

Dans cette étape notre système va insérer les vecteurs trouvés à la base de connaissance pour qu'on puisse les récupérer une fois que nous voulons analyser le même mot une autre fois.

```
<mot value="الكلمة">
    <exception description="نوع الكلمة">
    </exception>
    <candidat>
        <prefixe value="" description=""></prefixe>
        <stem value="" scheme="" racine=""></stem>
        <suffixe value="" description=""></suffixe>
    </candidat>
</mot>
```

Figure 28: structure d'entré de la base de connaissance

Après un certain temps nous ne serons pas besoin de passer par le processus d'analyse du mot en étude.

5.3.8 Conclusion

Nous avons présenté dans ce chapitre une description détaillé de notre système d'analyse morphologique arabe basé sur l'approche linguistique, afin d'analyser les textes arabes et qui permet d'extraire les traits morphologiques et la catégorie grammatical d'un mot.

Ce travail peut être utilisé dans la plupart des futurs travaux de traitement automatique de la langue arabe (analyse syntaxique, analyse sémantique, recherche d'information).

Nous avons développé ce système avec les mêmes standards des technologies (Java, XML, SAX,…), par conséquent, il peut être exécuté dans les différentes plateformes (Windows, Linux, Unix,…).

Conclusion générale

Aujourd'hui le traitement automatique de la langue naturel est devenu un domaine technologique en pleine effervescence dû à l'émergence de plusieurs entreprises telles que IBS, Microsoft, Apple,...etc. Donnant un nombre incalculables des applications ou encore linguiciels tels que les traducteurs automatique, les correcteurs d'orthographe automatique, les systèmes de résumé automatique, …etc. C'est un domaine scientifique qui attire un grand nombre de chercheurs en vue des progrès qui restent encore à accomplir.

Comme toutes les langues, l'arabe est convoité par le TALN. Cette convoitise n'est pas hasardeuse, mais due à deux facteurs principaux : l'un démographique, l'arabe parlée par plus de 453 million de personne à travers la planète répertoriée sur 23 pays et l'autre est linguistique, lié à sa structuration et son organisation qui est le résultat de plusieurs années de recherches de nos ancêtres linguistes et grammairiens (traitement morphologique, classifications des mots, règles de la grammaire, etc.).

Notresystème permet de donner des résultats raisonnable et efficace pour certains cas. Nous visons dans la suite du travail d'ajouter un certain nombre d'amélioration au niveau de la précision des résultats, l'ajout des autres règles morphologiques, la correction des résultats au niveau de la base de connaissance générée d'une manière automatique, d'implémenter une approche capable d'enlever l'ambigüité afin de produire finalement un résultat convenable.

Bibliographie

Al-Fedaghi, A. (1989). A new algorithm to generate root-pattern forms. *In Proceedings of the 11th National Computer Conference*.

Al-Jlayl, M. a. (2002). On Arabic Search : Improving the retrival effectivenesse via light stemming approche. *In Proceeding of the 11th ACM international Conference on Information and Knowledge Managment*.

Al-Kharashi. (1991). A microcomputer-based Arabic informationsretrival system comparing words stems, and roots as index terms. . *doctoral dissertation, illinois institute of technlogy* .

Al-kharashi, & Al-Sughaiyer. (s.d.). Pattern-based arabic stemmer. *In Procedding of 2nd saudi conference and exhibition*.

Al-Uthman. (1990). A morphological analyser for arabic . *Master's thesis KFUPM* .

Attia, M. (2005). Developing a Reboust arabic Morphological Transducer Using Finte State Technology. *8th Annual research colloquium . Manchester, UK.*

Belguith, & Chaâben. (2003). Analyse et désambiguïsation morphologiques de textes arabes non voyellés. *In Actes des Troisièmes journées scientifiques des jeunes chercheurs en Génie Électrique et Informatique (GEI'2003)* .

bielický, S. a. (2009). ElixirFM High-Level implementation of Functional Arabic Morphology . *http//sourceforge.net/projects/elixir-fm*.

Buckwalter. (2002). *Arabic morphological analyser Version 0.1.*

Buckwalter. (2004). *Issues in arabic orthography and morphology analysis.*Genève, suisse.

CHAIRET, & Mohamed. (1996). *Linguistique contrastive et traduction. N° spécial : Fonctionnement du système verbal en arabe et en français.*Paris.

COLLINs. (2002). Discriminative training methods for hidden Markov models: theory and experiments with perceptron algorithme.

Dannan, T. e. (1987). A comprehensive Arabic morphological analyser /generator. *IBM kuwait scientific center*.

Daoud. (2009). Synchronized Morphological and Syntactic Disambiguation for Arabic.

Darwish. (2002). Building a Shallow Arabic Morphological Analyzer in One Day. *In Actes du workshop Computational approaches to Semitic languages*.

Diab. (2004). Automatic Tagging of Arabic Text: From Raw Text to Base Phrase.

DICHY. (1992). *L'écriture dans la représentation de la langue: la lettre et le mot en arabe.*

Dicky, F. (2003). Roots & Patterns vs. Stems plus Grammer-lexis Spesification. *9ème MT Conference Workchop on Machine Traslation*.

El Amine Abderrahim, M. (2008). *Reconnaissance des unitées linguistiques signifiantes.*TLEMCEN: Université Abou Bekr BELKAID TLEMCEN.

El-Affendi. (1998). Performing Arabic morphological search on the internet . *King Saud university* .

ELHADJ, AL-SUGHAYEIR, & AL-ANSARI. (2009). Arabic Part-Of-Speech Tagging using the Sentenc.

fathi, N. M. (1995). An integrated Morphological and syntactic technology:contrubution of morphosemantics.

Gridach. (2010). Developing a New Approach for Arabic Morphological Analysis and Generation.

Gridach. (2012). *réalisation d'un analyseur Morphologique Arabe Basé sur XML et les Automates Morphologiques.*

Habash, & Rambow. (2005). Arabic tokenization, part-of-speech tagging and morphological.

hadrich, B. (1999). *Traitement des erreurs d'accord de l'arabe basé sur une analyse.* tunis: Faculté des sciences.

hajji. (2000). Morphological Tagging.

Hegazi, & El-sharkawi. (1986). Natural arabic language processing .

HOCEINI. (2011). Towards a New Approach for Disambiguation in NLP by Multiple Criterian Decision-Aid.

HOCEINI, A. (2012). Méthodologie Multicritère de Désambiguïsation Morphosyntaxique de la Langue Arabe.

Kadri, & Benyamina. (1992). *A syntax semantic analyzer for Arabic language.*

khalfallah. (2008). *عنَا صِر أ ً ولِيَّة من النَّحو العربِي.* Lorraine: Département d'Arabe - Université de Lorraine.

khoja. (2001). APT: Arabic part-of-speech tagger. *In Proceedings of The Student Workshop at the second meeting of the North American Chapter of the Association for Computational Linguistics*.

khoja, GARSUDE, & KNOWLES. (2001). A tagset for the morphosyntactic tagging of arabic.

Khreisat. (2006). Arabic Text Classification Using N-gram Frequency Statistics.

koehn, & hoang. (2007). Factored translation models.

Lallich, a. (1990). *analyse du francais achévement et l'imlentation de l'analyseur morphosyntaxique.*Grenoble: université des sciences sociale.

LARKEY, BALLESTEROS, & CONNELL. (2002). Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis.

Larkey, Ballesteros, & Connell. (s.d.). Towards Improving Khoja Rule-Based Arabic Stemmer.

Larky. (2002). Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis.

Mars. (2008). Nouvelles ressources et nouvelles pratiques pédagogiques avec les outilsTAL. *journal information sciences for decision for decision Marketing* .

Mekki. (2011). évaluation de G-lexir pour la traduction automatique statique. *in proceeding of traitement automatique du langage natural* , Montpellier.

Mesfar. (2008). *Analyse morphosyntaxique autommatique et reconnaissance des entités signifiants .*

Nabil, A. (1992). Parsing and automatic diacritization of written arabic . *In Proceedings of the 13th National Computer Conference* .

Sanan. (2008). *Arabic documents classification.* toulouz.

Shaalan. (1989). Arabic morphological analysis and the lexicon.

smrž. (2007). Functional Arabic Morphology : Formal System and Implementation . *PhD thesis Charles University in Prague*.

Wikipedia. (s.d.). *en.wikipedia.org/wiki/Arabic_grammar*.

Xu, Fraser, & Weischedel. (2002). *Empirical Studies in Strategies for Arabic Retrieval.*

Zoubeir, M. (2008). AraSeg: un segmenteur semi-automatiquedes textes arabes.

ZREIK, & HAJJAR. (2010). *EXTRACTION ET GESTION DE L'INFORMATION A PARTIR DES DOCUMENTS ARABES.*SAINT DENIS: UNIVERSITE PARIS VIII .

ZRIBI, TORJMEN, & AHMED, B. (2006). An Efficient Multi-agent System Combining POS-Taggers for arabic.