ROYAUME DU MAROC
Université Mohammed V - RABAT -
جامعة محمد الخامس – الرباط –
Faculté des sciences
كلية العلوم
CENTRE D'ETUDES DOCTORALES - SCIENCES ET TECHNOLOGIES

N° d'ordre : **3500**

THESE

En vue de l'obtention du : ***DOCTORAT***

Structure de Recherche : Laboratoire de Recherche en Informatique et Télécommunications (LRIT)

Discipline : Sciences de l'ingénieur

Spécialité : Informatique et télécommunication

Présentée et soutenue le 10-07-2021 par :

Said IAZZI

L'analyse morphologique des mots arabes à base des schèmes de surface et de graphes.

JURY

Abdellah YOUSFI	PES	Faculté des Sciences Juridiques, Economiques et Sociales, Souissi, Université Mohammed V – Rabat.	Président
Said OUATIK ELALAOUI	PES	Ecole Nationale des Sciences Appliquées, Université Ibn Tofail, Kenitra	Rapporteur / Examinateur
My Ahmed FAQIHI	PH	École Nationale Supérieure d'Informatique et d'Analyse des Systèmes, Université Mohammed V - Rabat	Rapporteur / Examinateur
Si Lhoussain AOURAGH	PH	Faculté des Sciences Juridiques, Economiques et Sociales, Salé, Université Mohammed V – Rabat	Rapporteur / Examinateur
Mostafa BELLAFKIH	PES	Institut National des Postes et Télécommunications, Rabat.	Examinateur
Khalid MINAOUI	PES	Faculté des Sciences, Université Mohammed V – Rabat.	Directeur de thèse

Année Universitaire : 2020 – 2021

Résumé

L'analyse morphologique est parmi les outils indispensables dans le traitement automatique des langues. Dans cette thèse, nous avons amélioré deux approches pour faire l'analyse morphologique des mots arabes.

Dans la première approche, nous avons utilisé les schèmes de surface des mots dérivables arabes et le degré de similarité entre le mot et le schème. Cette approche vise à traiter les mots dérivés arabes ; elle est basée principalement sur la construction de la base des données des schèmes de surface des mots.

Cette approche a été testée sur un corpus de 4000 mots arabes (1400 verbes et 2600 noms dérivés), les résultats obtenus sont très intéressants et montrent l'utilité et l'importance de cette démarche.

Dans la deuxième approche, nous avons utilisé les graphes pour faire l'analyse morphologique des mots arabes. Le système utilise un graphe et un dictionnaire très restreints de stems de Buckwalter. Ensuite, on cherche les solutions dans ce graphe en utilisant l'algorithme de Viterbi. Notre approche a été testée sur un corpus de 4000 mots arabes. Les résultats obtenus sont encourageants, impressionnants et montrent l'importance de cette seconde approche.

Mots-clefs : *TALN, Nom dérivé Arabe, Schème de Surface des mots, Analyse Morphologique, Degré de Similitude, Graphe; Algorithme de Viterbi, Tables de compatibilité, Automates à états finis .*

Abstract

Morphological analysis is among the essential in natural language processing tools. In this thesis, we have developed two approaches to the morphological analysis of Arabic words.

First, we presented a morphological analysis system for the Arabic language. Which is based on the surface patterns of Arabic words.

Our work in this approach aims at dealing with Arabic derived nouns. It is based mainly on the building of a database for the surface patterns of the latter (In order to deal with Arabic derived nouns, the article is also based on a previous study by (Yousfi, 2010) for the analysis of Arabic verbs).

Our approach was tested on a corpus of 4000 Arabic words (1400 verbs and 2600 derived nouns), the obtained results are very interesting and show the utility and importance of this approach.

In a second step, we propose a new morphological analysis system for Arabic words. This system combines both the Buckwalter approach as well as a graph-based morphological analytical approach. This system is based on very restricted dictionaries and looks for solutions in a global network by using the Viterbi Algorithm. Our approach was tested on a corpus of 4000 Arabic words. The results achieved are very impressive and show how important our new approach can be.

Key Words: *ANLP, Arabic Derived Nouns, Surface Pattern of Words, Morphological Analysis, Degree of Similarity, Compatibility Table, Graph, the Viterbi Algorithm, Morphological Analysis.*

Dédicaces

Mes dédicaces ne sont que l'expression de mes profondes gratitudes, de mes salutations chaleureuses et de ma sincère reconnaissance à tous ceux qui comblent ma vie et lui confèrent son goût et sa saveur.

Je dédie cette thèse à :

Mes parents :

Ma mère, qui a œuvré pour ma réussite, de par son amour, son soutien, tous les sacrifices consentis et ses précieux conseils, pour toute son assistance et sa présence dans ma vie, reçois à travers ce travail aussi modeste soit-il, l'expression de mes sentiments et de mon éternelle gratitude.

Mon père, qui peut être fier et trouver ici le résultat de longues années de sacrifices et de privations pour m'aider à avancer dans la vie. Puisse Dieu faire en sorte que ce travail porte son fruit ; Merci pour les valeurs nobles, l'éducation et le soutient permanent venu de toi.

Ma femme, pour sa patience, ses conseils, son soutien pendant ces dernières années de thèse, pour ses encouragements pendant toute cette période et sa présence à mes côtés dans les moments de joie et de peine.

Mes enfants adorés Alaa et Safae et ma petit Hajar pour l'espoir qu'ils gravent de jour en jour dans mon cœur.

Mes frères et sœurs qui n'ont cessé d'être pour moi des exemples de persévérance, de courage et de générosité.

Enfin je dédie ce travail à tous mes amis et collègues qui n'ont cessé d'être des sources de soutien et d'amitié.

À tous les membres de ma famille sans aucune exception. À tous ceux qui me sont chers.

Remerciements

Au terme de cette recherche, je ne pourrai m'empêcher de témoigner ma reconnaissance aux personnes qui, d'une manière ou d'une autre, ont rendu sa réalisation possible.

Le travail présenté dans cette thèse a été réalisé au sein du **Laboratoire de Recherche en Informatique et Télécommunications (LRIT)** à la Faculté des Sciences de Rabat (FSR) de l'Université Mohammed V-Rabat, sous la direction du professeur **Khalid MINAOUI** de la Faculté des Sciences de Rabat.

Je tiens tout d'abord à exprimer ma profonde gratitude et mes remerciements à mon directeur de thèse **Mr. Khalid MINAOUI** Professeur d'Enseignement Supérieur à la faculté des sciences, université Mohammed V-Rabat. Je tiens à le remercier pour la confiance qu'il m'a toujours témoignée, en m'accordant à la fois une large autonomie et un soutien permanent.

Mes remerciements s'adressent particulièrement à **Mr. Abdellah YOUSFI** Professeur de l''Enseignement Supérieur à la Faculté des Sciences Juridiques, Economiques et Sociales, Souissi, Université Mohammed V-Rabat., pour l'honneur qu'il me fait pour présider le jury de cette thèse.

Je remercie **Mr. Said OUATIK ELALAOUI** Professeur d'Enseignement Supérieur de l'Ecole Nationale des Sciences Appliquées, Kenitra, d'avoir accepté de rapporter et examiné cette thèse.

Je remercie **Mr. My Ahmed FAQIHI** Professeur Habilité à l'École Nationale Supérieure d'Informatique et d'Analyse des Systèmes, Université Mohammed V-Rabat, d'avoir accepté de rapporter et examiné cette thèse.

Je remercie **Mr. Si Lhoussain AOURAGH** Professeur Habilité à la Faculté des Sciences Juridiques, Economiques et Sociales, salé, Université Mohammed V- Rabat, d'avoir accepté de rapporter et examiné cette thèse.

Je remercie **Mr. Mostafa BELLAFKIH** Professeur d'Enseignement Supérieur à l'Institut National des Postes et Télécommunications, Rabat, pour avoir examiné ce travail.

Je remercie le professeur **Mohammed OUADOU**, directeur du laboratoire **LRIT** pour nous avoir permis de réaliser ce travail au sein du laboratoire.

A tous ceux qui ont participé de près ou de loin à la réalisation de ce travail.

Mes gratitudes vont aussi à tous mes enseignantes pendant tout mon cursus d'études pour leurs disponibilités et pour le savoir qu'ils m'ont fidèlement transmis.

Résumé

L'analyse morphologique est parmi les outils indispensables dans le traitement automatique des langues. Dans cette thèse, nous avons amélioré deux approches pour faire l'analyse morphologique des mots arabes.

Dans la première approche, nous avons utilisé les schèmes de surface des mots dérivables arabes et le degré de similarité entre le mot et le schème. Cette approche vise à traiter les mots dérivés arabes ; elle est basée principalement sur la construction de la base des données des schèmes de surface de ces derniers.

Cette approche a été testée sur un corpus de 4000 mots arabes (1400 verbes et 2600 noms dérivés), les résultats obtenus sont très intéressants et montrent l'utilité et l'importance de cette démarche.

Dans la deuxième approche, nous avons utilisé les graphes pour faire l'analyse morphologique des mots arabes. Le système utilise un graphe et un dictionnaire très restreint de stems de Buckwalter. Ensuite, on cherche les solutions dans ce graphe en utilisant l'algorithme de Viterbi. Notre approche a été testée sur un corpus de 4000 mots arabes. Les résultats obtenus sont encourageants, impressionnants et montrent l'importance de cette seconde approche.

Mots clés : TALN, Nom dérivé Arabe, Schème de Surface des mots, Analyse Morphologique, Degré de Similarité, Graphe, Algorithme de Viterbi, Tables de Compatibilité ; Automates à état fini.

ملخص

يعتبر التحليل الصرفي من بين المراحل الأساسية في معالجة اللغات الطبيعية. في هذه الأطروحة، اقترحنا مقاربتين للتحليل الصرفي للكلمات العربية المشتقة من الجذور.

في البداية، قدمنا نظام جديد في مجال التحليل الصرفي للغة العربية. ويستند هذا النظام على الأوزان السطحية لكلمات اللغة العربية المشتقة من الجذور، ودرجة التشابه بين الكلمة والأوزان. ويهدف عملنا الى معالجة الأسماء العربية المشتقة؛ ويستند في المقام الأول على بناء قاعدة معطيات خاصة بالأوزان السطحية.

وفي المقاربة الثانية، اقترحنا نظاما جديدا للتحليل الصرفي للكلمات العربية. هذا النظام يجمع بين نهج **بك والتر Buckwalter** والتحليل الصرفي من خلال نهج المبيان. ويستند هذا النظام على قواميس جد صغيرة، ويسعى الى إيجاد حلول في الشبكة العامة باستخدام خوارزمية **Viterbi**. وقد تم تجريب مقاربتنا باعتماد مثن يخضع حوالي 4000 من الكلمات العربية. وكانت النتائج جد مهمة أظهرت أهمية المقاربة الجديدة.

تم تقييم المقاربتين بالاعتماد على متن لحوالي 4000 كلمة. كل كلمة تضم الجدر، السوابق، اللواحق، ...

النتائج المحصل عليها كانت مشجعة وبينت أهمية المقاربتين وتشجعنا على المواصلة في البحث من أجل تحسين النتائج وتعميم المقاربتين على جميع أنواع الكلمات العربية وبالخصوص الغير المشتقة منها.

كلمات مفتاحية : المعالجة الآلية للغة العربية، الأسماء العربية المشتقة، الأوزان السطحية، المحللات الصرفية، درجة التشابه، الرسوم، قائمة توافق السوابق واللواحق، أوتومات الأوضاع النهائية.

Abstract

Morphological analysis is among the essential in natural language processing tools. In this thesis, we have developed two approaches to the morphological analysis of Arabic words.

First, we presented a morphological analysis system for the Arabic language. Which is based on the surface patterns of Arabic words.

Our work in this approach aims at dealing with Arabic derived nouns. It is based mainly on the building of a database for the surface patterns of the latter (In order to deal with Arabic derived nouns, the article is also based on a previous study by (Yousfi, 2010) for the analysis of Arabic verbs).

Our approach was tested on a corpus of 4000 Arabic words (1400 verbs and 2600 derived nouns), the obtained results are very interesting and show the utility and importance of this approach.

In a second step, we propose a new morphological analysis system for Arabic words. This system combines both the Buckwalter approach as well as a graph-based morphological analytical approach. This system is based on very restricted dictionaries and looks for solutions in a global network by using the Viterbi Algorithm. Our approach was tested on a corpus of 4000 Arabic words. The results achieved are very impressive and show how important our new approach can be.

Keywords: ANLP, Arabic Derived Nouns, Surface Pattern of Words, Morphological Analysis, Degree of Similarity, Compatibility Table, Graph, the Viterbi Algorithm, Morphological Analysis.

Table des Matières

Liste des tableaux

Liste des figures

Liste des abréviations

Abréviations	Définition du terme
ALECSO	Arab League Educational, Cultural and Scientific Organization
TALN	Traitement Automatique des Langues Naturelles
TAL	Traitement Automatique de la Langue
IBM	International Business Machines
FST	Finite-State Transducer
MAGEAD	Morphological Analyzer and Generator for the Arabic Dialects
MSA	Modern Standard Arabic
IR	Recherche d'Information

Introduction générale

Le traitement automatique des langues naturelles (TALN) est un domaine d'actualité qui sert à améliorer l'interaction entre l'homme et la machine. Il a pour objectif de développer des applications capables de traiter de façon automatique des données linguistiques exprimées dans une langue naturelle donnée. Nous avons remarqué ces dernières années, que plusieurs chercheurs se sont orientés vers ce domaine.

Le Traitement Automatique des Langues Naturelles donne aux ordinateurs la capacité de comprendre la façon dont les êtres humains apprennent la langue et de l'utilisation. Les techniques du Traitement Automatique des Langues Naturelles analysent les entrées linguistiques (un mot, une phrase, un texte, un dialogue) selon les règles (règles de dérivation, règles flexionnelles, règles grammaticales, etc.) et les ressources (comme le lexique, le corpus, le dictionnaire) de la langue ciblée.

Le traitement automatique de la langue a besoin de plusieurs efforts aux niveaux développement de l'ensemble des méthodes et des techniques informatiques avancées, pour de nombreuses tâches de traitement automatique de la langue, une base de données lexicale complète et riche est essentielle, même une simple liste de mots peut souvent constituer une source d'information inestimable. L'un des problèmes les plus difficiles avec les lexiques est celui des mots hors vocabulaire, surtout pour les langues qui ont une morphologie plus riche comme l'arabe. Pour évaluer nos systèmes d'analyse morphologique, nous avons besoin d'un corpus adapté pour les analyseurs morphologiques arabes qui permet de faciliter les tâches de traitement et qui donne des résultats efficaces.

Morphologiquement, la langue arabe est une langue complexe et riche. Des dizaines, voire des centaines de mots peuvent être formés en utilisant une racine, des schèmes et des affixes. La langue arabe présente des caractéristiques et des spécificités qui la rendent plus ambigüe que d'autres langues naturelles, sa morphologie, sa syntaxe ainsi que sa sémantique sont en corrélation et se complètent. Pour la langue arabe, il y a un entrelacement fort entre ces différents niveaux.

L'analyse morphologique des mots arabes est une condition préalable pour de nombreuses applications du Traitement Automatique des Langues Naturelles, telle que : l'analyse syntaxique, la recherche d'information, la traduction automatique, etc. L'importance de cette procédure a

attiré de nombreux chercheurs de différentes disciplines pour travailler sur l'analyse morphologique (les linguistes, les chercheurs dans le domaine de l'intelligence artificielle, les chercheurs en informatique, etc).

L'analyse morphologique est une étape indispensable dans nombreuses applications de TALN, telles que : l'analyse syntaxique, l'analyse sémantique, la recherche de l'information, la correction orthographique, la reconnaissance de l'écriture manuscrite, la traduction automatique, le résumé automatique et l'apprentissage automatique de la langue arabe (Al-Sughaiyer et Al-Kharashi 2004).

L'analyseur morphologique des mots, a pour objectif de traiter et d'analyser la structure interne des mots et de déterminer les différents morphèmes composant ces mots, tel que les préfixes, les suffixes, les infixes, les lemmes et les racines, ainsi que les caractéristiques morphologiques (genre, nombre, personne, cas, humeur, voix, etc) (Kiraz; 2001).

Plusieurs travaux, dans cet axe, ont été élaborés ces dernières années, et qui se basent généralement sur l'une des approches suivante : l'approche à deux niveaux, l'approche par concaténation, l'approche à base des dictionnaires Buckwalter (Buckwalter, 2004), l'approche à base de schèmes des mots Beesly (Beesley, 1996) et Sabri et yousfi (Sabri et yousfi, 2006) , l'approche à base d'automate à états finis (Al-Sughaiyer and Al-Kharashi 2004), (Dichy & Fargaly, 2003) et (Iazzi, S, Yousfi, A, SITA 2020), l'approche à base de règles linguistiques Buckwalter (Buckwalter, 2004) et Al-Fedagi et Al-Anzi (Al-Fedagi et Al-Anzi, 1989) , Approche statistiques [Darwish, (2002)] ou des approches qui combinent entre ces différentes approches (Approche hybrides))(Buckwalter, 2002).

Dans cette thèse, nous avons choisi de travailler sur ce sujet, qui consiste à faire d'abord une étude sur les approches existantes, définir leurs faiblesses, et de proposer à la fin d'autres approches qui améliorent les méthodes existantes.

Le travail est consacré à l'amélioration des performances de l'analyse morphologique des mots dérivés en langue arabe, moyennant l'approche linguistique à base racine-schème et à base de graphe. Il vise d'une part l'amélioration de la technique de racinisation (stemming) permettant d'extraire la racine d'un mot analysé en utilisant des schèmes de surface dont l'objectif est de

garder les lettres défectueux à leurs places dans les schèmes. D'autre part nous proposons l'utilisation d'une liste des schèmes de surface qui représentent les différentes classes des schèmes des verbes. Un autre travail qui propose une nouvelle approche qui combine la technologie des automates finis et l'utilisation des ressources linguistiques utilisés dans l'analyseur de Buckwalter (dictionnaire et les tables de compatibilité entre les préfixes, suffixe et infixes).

Notre thèse est constituée de quatre chapitres :

Le premier chapitre définit les spécificités de la langue arabe, les catégories et les structures des mots arabes, une représentation des mots et leurs structures, les catégories des mots arabes, sa morphologie et ses différentes difficultés, les types des schèmes de la langue arabe et les différents niveaux de traitement des langues naturelles.

Une présentation détaillée de l'analyse morphologique sera l'objectif du deuxième chapitre. Dans ce dernier, on a donné une définition de l'analyse morphologique pour une langue flexionnelle, une présentation des différentes approches de l'analyse morphologique et nous avons terminé avec un état de l'art sur les différentes approches d'analyse morphologique de la langue arabe.

Notre première contribution d'analyse morphologique est présentée dans le chapitre trois, dans lequel nous avons amélioré l'approche de l'analyse morphologique à base de schèmes de surface (Yousfi, 2010). Dans notre contribution, nous avons d'abord présenté les schèmes de surfaces et les phases de construction de la base de schèmes de surfaces liés aux noms dérivés. Ensuite, nous avons détaillé le concept théorique de cette approche. Puis, nous avons terminé avec le test et le corpus d'évaluation de notre analyseur morphologique en termes de précision et rappel.

Dans le dernier chapitre, nous avons présenté notre deuxième approche qui s'appuie sur les graphes et tables de compatibilité entre affixes pour faire l'analyse morphologique des mots arabes. Dans notre contribution, nous avons détaillé le concept théorique de cette approche, nous avons d'abord présenté l'utilisation des automates à états finis pour faire l'analyse morphologique de la langue arabe et une présentation de notre analyseur IBN-GINNY et le réseau global des mots arabes dans notre analyseur, ensuite, une analyse morphologique à base du réseau global et les tables de compatibilité entre affixes et les racines. Puis, nous avons cité quelques

inconvénients des approches de l'analyse morphologique, nous avons terminé par une évaluation de notre approche en termes de précision et rappel en se basant sur un corpus d'évaluation constitué de 4000 mots de la langue arabe construit par des experts en langue arabe.

Chapitre I : Caractéristiques et spécificités de la langue arabe

I- Introduction :

Actuellement, l'arabe fait face à de nombreux défis à cause de plusieurs raisons telles que la croissance du nombre de sites Web en arabe, les médias arabes, et des sociétés dans le monde entier qui utilisent la langue arabe. Pour ces raisons, beaucoup de recherches dans le domaine ont été développées pour satisfaire cette demande croissante. La morphologie arabe est l'un des besoins essentiels dans ce domaine, c'est pourquoi beaucoup d'analyseurs morphologiques sont disponibles maintenant. Certains d'entre eux ont un but commercial, d'autres sont disponibles pour la recherche et l'évaluation.

L'analyse morphologique est une étape importante dans le traitement de la langue arabe. En raison de son utilisation dans la plupart des applications du traitement automatique du langage naturel (TALN), elle constitue une branche de l'intelligence artificielle qui a pour objectif d'inventer des théories, de découvrir des techniques et de construire des logiciels qui peuvent comprendre, analyser et générer les composantes des langues humaines.

Ce processus a pour but général d'élaborer des interfaces informatiques qui donnent la possibilité aux utilisateurs d'interagir avec ces interfaces.

II- Spécificités de la langue arabe

La langue arabe fait partie des langues sémitiques les plus parlées au monde. Elle occupe la cinquième rang mondial. Elle est l'une des dix langues les plus utilisées sur Internet. Pour une

population globale de 422 millions de locuteurs natifs répartis dans plus de 57 pays dans le monde, plus de 300 millions de personnes utilisent la langue arabe comme première ou deuxième langue. En outre, elle est la langue officielle dans 22 pays. Et on pense que c'est la langue la plus riche du monde avec environ 12 millions et 300 000 mots. Cependant, en l'absence d'une théorie solide sur l'origine de l'arabe, les linguistes considèrent que l'arabe coranique est un point de départ et de référence [1].

Contrairement aux langues d'origine latine, le système d'écriture arabe est orienté de droite à gauche, et la plupart des lettres dans les mots arabes sont liées. Vingt-deux parmi les vingt-huit lettres peuvent être réunis sur les deux côtés et prennent des formes différentes en fonction de leur emplacement dans le mot (ـعـ ، ـع، عـ، ع).

Actuellement, la langue arabe peut être classée en trois types: l'arabe classique, l'arabe standard moderne et l'arabe dialectal (Fleisch, Henri., (1964)).

III- Catégories et structures des mots arabes

Un mot est une séquence de lettres se terminant par un espace. Il y a des cas où des mots sont attachés par des particules (" والداخل، فالداخل , للداخل", 'wadaʾẖilon, fadaʾẖil, lldaʾẖil'). L'orthographe et la morphologie de la langue arabe donnent lieu à un large vocabulaire des mots composés.

Egalement, la variabilité et la richesse supplémentaire des mots découlent de la productivité dérivationnelle et flexionnelle de la langue arabe. Un mot donné peut être trouvé sous différentes formes qui devraient éventuellement être confondues pendant la recherche d'information. Citons l'exemple des mots ("يدخلون ، يتداخلون ، داخلون" 'daʾẖilown, ytdaʾẖalown, yadẖolown').

L'arabe a deux genres, masculin et féminin. Elle dispose également de trois personnes : le singulier, le duel et le pluriel. Il dispose également dans des situations du discours de celui qui parle : locuteur ("المتكلم",'al-mutakalim'), l'interlocuteur ("al-muẖaʾṭab", 'المخاطب') et l'absente ("al-ġaʾaýib", 'الغائب')[(2)].

[1] https://notesread.com/characteristics-of-arabic-language/

[(2)] Dans ce manuscrit, nous utiliserons, pour plus de commodité, les caractères latins pour représenter les caractères arabes sur la base de la translittération API (Annexe A).

À un niveau plus profond, la plupart des mots arabes sont construits à partir des racines, qui sont généralement composées de trois consonnes, mais il y a également des racines composées de quatre consonnes "دحرج" 'daḥraĝa' (Dichy, 1999). Les mots sont dérivés à partir de ces racines suivant un processus fixe en y ajoutant des préfixes, des suffixes et des infixes.

D'autre part, les lettres racines[3] "ف، ع، ل" sont la base de tous les radicaux arabes. Le schème est obtenu en comparant l'ensemble des additions effectuées sur les lettres racines qui forment le radical du mot en cours de traitement.

La langue arabe est parmi les langues qui présentent une structure morphologique très systématique, mais complexe, basée dans des cas sur l'approche schème. Par conséquent, l'enquête concernant ces techniques s'avère indispensable.

Dans les références de la grammaire arabe traditionnelle, les linguistes classent les mots arabes en trois catégories principales: noms, verbes et particules [Sibawayh alkitab]. Khoja (Khoja, 1999) a ajouté les catégories de signes de ponctuation [4] et les mots non arabes comme par exemple: les devises, les chiffres et les mots d'autres langues (non arabes) (Dahdah 1987; Dahdah 1993; Wright 1996; Al-Ghalayyni, 2005; Ryding, 2005).

III-1 Représentation des mots et des structures

En ce qui concerne l'utilisation des signes diacritiques, les textes arabes peuvent être classés en trois types: les textes voyellés, non voyellés et semi-voyellés:

- Le premier type représente les textes arabes entièrement voyellés, dans lequel chaque consonne est suivie d'un signe diacritique qui peut être placé en dessous (ِ، ٍ) ou au dessus (ً، َ ، ُ، ٌ). Ce genre de textes se trouve dans les manuels scolaires et dans le Coran pour assurer une lecture correcte.
- Le deuxième type se compose de textes sans signes diacritiques, la majorité des textes disponibles dans les livres, les journaux et les textes sur Internet sont non voyellés.

[3] Les lettres composant la racine d'un mot.

[4] Point, point d'interrogation, point d'exclamation, virgule, point-virgule, deux-points, points de suspension

- Dans certains cas on peut trouver des mots partiellement voyellés dans des textes non voyellés, où certains signes diacritiques, généralement un ou deux, sont ajoutés afin d'éliminer toute ambiguïté de sens de ces mots, les textes partiellement voyellés sont en nombre limité.

III-2 Catégories des mots arabes

Les mots arabes sont classés en deux catégories :

Les mots dérivés : ils sont générés à partir d'entités de base appelées racines, selon un ensemble de règles de dérivation ou selon des schèmes morphologiques (tableau n°1).

Pronom personnel	Dérivation des verbes	Conjugaison des verbes	Verbes
مفرد-مؤنث	ضَرَبْتِ	ماضي	ضَرَبَ
مفرد-مذكر	مُسْتَاءٌ	اسم-الفاعل	اِسْتَاءَ
مفرد- مذكر	يُعَلَّمُ	مضارع-معلوم	عَلَّمَ
مفرد- مذكر	تَعِظَ	مضارع-معلوم	وَعَظَ
مثنى-مذكر	وَاقِيَانِ	اسم-الفاعل	وَقَى
جمع-مذكر	مُنْصَاعُونَ	اسم-الفاعل	اِنْصَاعَ

Tableau n° 1 : Un exemple des mots dérivés

Les mots non dérivés: ce sont les mots qui ne respectent pas les règles de dérivation standard, par exemples les mots fonctionnels, les mots empruntés d'autres langues étrangères (زكرياء ، مهندس ،التلفزة ، صالون،...).

III-2.1 Le verbe arabe

C'est un mot qui désigne une action qui pourrait être combinée avec certaines particules. Le verbe pourrait être au passé, au présent ou à l'impératif. Un futur du verbe existe, mais c'est un dérivé du temps présent qui se réalise en attachant un préfixe au présent du verbe ("سأدخل", 'saádḫolo'). Les particules peuvent être ajoutées en tant que préfixes et / ou suffixes indiquant le nombre, le genre et la personne du sujet. Trois modes sont possibles pour les verbes: indicatif, subjonctif et conditionnel.

Un verbe peut avoir une forme morphologique saine ("صحيح", 'ṣaḥiḥ') comme le verbe ("دخل", 'daḫala', ou défectueuse ("معتل", 'moʿtal') comme ("دعى", 'daʿąą '). Le verbe peut être transitif ("متعدي", 'motaʿadiy') ou intransitif ("لازم", 'laʾzim').

Il y a un autre type de verbes [5] qui ne nécessitent ni sujet ni complément, mais ils ont besoin de topique ("مبتدأ ", 'mobtadáʾ) et d'attribut ("خبر ",'ḫabar ') : "كان محمد مريضا".

a) Verbes sains et défectueux

En s'appuyant sur la structure morphologique des mots, le verbe arabe peut être réparti en deux classes : sain ("صحيح", ' ṣaḥiḥ'), comme exemples : "دَخَلَ، كَتَبَ ،سَمِعَ، صَبَرَ " et défectueux ("معتل", moʿtal"), exemples : "نَمَا، هَدى ".

- ✓ **Verbes sains ("الصَحيحُ", ''lṣaḥiḥ') :** se sont les verbes qui ne contiennent aucune lettre défective et lors de la génération morphologique, les lettres composants ces verbes ne subissent aucune transformation. comme دخل ⟶ مدخول، داخل، يدخلون .

  - o **Verbes hamzé ("الفعل المهموز", ''lmahmowz'):** la hamza n'est pas considérée comme une lettre de cet alphabet, Les règles d'écritures de la hamza et de son support éventuel dépendent de la nature de la hamza, de sa place dans le mot, où "ء" est l'une de ces consonnes (سأل, أكل, بدأ).

  - o **Verbe sourd /doublés("المضعف", ''lmoḍaʿaf'):** ce sont des verbes qui se terminent par deux consonnes identiques sans voyelle entre les deux. Ceci fait référence à n'importe quel verbe se terminant par un shadda (ّ) sur la dernière lettre (consonnes doublées sans une voyelle entre les deux) " مَرَرَ-مَرَّ, عَدَدَ-عَدَّ, مَدَدَ-مَدَّ, دقّ-دقق" .

- ✓ **Verbes faibles ("المعتلة", ''lmoʿtalä')** : ce sont des verbes dont les racines contiennent un ou plusieurs lettres faibles[6] (ى ،ا ، و، ي), exemple "...دعى، سعى، وقف". ces verbes se trouvent sous trois types :

  - o **Verbes creux ("الأجوف", ''lágwaf') :** la deuxième lettre dans la racine est le "alif" "ا" comme " باع –يبيع" et "عاد -يعود".

[5] le verbe "كان", "أصبح",…

[6] Les lettres faibles sont : alif "ا", waw "و" , ya: "ي".

- **Verbes assimilés ("المثال", ''lmiṯa'l') :** ils commencent par (و ou ي) ou (généralement و); dans les imparfaits et dans d'autres situations, le waw "و" disparaît souvent "وقف – يقف".

- **Verbes défectueux ("الناقص", ''lna'qiṣ') :** ce sont des verbes dont la troisième lettre radical est un ى ou un ي comme (سعى، دعى).

b) Verbes transitifs et intransitifs (transitivité des verbes)

De même si on fait référence à la syntaxe, le verbe arabe peut avoir deux types:

- ✓ **Verbes transitifs ("الفعل المتعدي", ''lmotaʿdiy') :** ce sont des verbes qui prennent un complément. Le verbe transitif est retracé au sujet et dirigé vers le complément, il est le verbe qui se rattache au complément et l'affecte. Comme l'exemple " فَتَحَ طَارِقٌ الأَنْدَلُسَ " (Tariq conquis al-Andalus).
- ✓ **Verbes intransitifs ("اللازم", ''ll'zim')** : sont les verbes qui n'ont pas des compléments, le verbe intransitif nécessite uniquement un sujet "فاعل", Par exemple "جَلَسَ" (il était assis), ainsi, si l'on dit "جَلَسَ زَيدٌ", cela serait une phrase complète, de même on ne peut pas dire "جَلَسَهُ".

c) Les facteurs intervenants dans le verbe

Les verbes arabes sont une unité essentielle dans la structure morphologique de la langue. Ils subissent plusieurs transformations selon les catégories grammaticales avec lesquelles ils sont liés. Ces catégories grammaticales varient selon les cas. Tout d'abord, le verbe doit s'accorder avec la personne (le sujet), notamment le nombre (المفرد, **'lmofrad**) singulier, duel (المثنى, **'lmoṯnĄ**) et pluriel (الجمع, **'lĝalʿ),** et le genre; masculin (المذكر) et féminin (المؤنث). Ensuite, le temps intervient aussi dans ces catégories; parfait (الماضي) et imparfait (المضارع). Puis, le mode du verbe se veut un élément important aussi; indicatif (المرفوع), subjonctif (المنصوب) et jussive (المجزوم). Enfin, la voix intervient aussi dans les changements qui affectent le verbe, actif (المعلوم) et passif (المجهول) .

d) Forme passive et active

La forme passive (المبني للمجهول) est très importante et intéressante. Il s'agit de changer la forme et le sens des phrases dans une certaine mesure. Dans les constructions passives, en français par exemple le complément de la phrase active devient un sujet grammatical. En arabe " كتب صديقي الكتاب" est une phrase active qui commence par un sujet. Son équivalent passif "كُتب الكتاب". Les formes morphologiques passives et actives de la plupart des verbes en arabe sont étroitement liées. La principale différence entre les formes actives et passives des verbes arabes sont les voyelles courtes "كَتَبَ، كُتِبَ".

e) Classification des verbes au niveau du temps :

Les verbes peuvent être classés en fonction de la forme de leur morphologie en trois groupes:

- **Verbe parfait ("ماضي", 'ma’ḍy') :** il indique une action dans le passé "عَلَّمَ-عَلَّمْتَ". Le parfait est principalement utilisé dans des contextes où l'action du verbe s'est produite au passé ("حضرت حفلة موسيقية يوم أمس") . Le verbe parfait peut être sous les deux modes : actif et passif.

- **Verbe imparfait ("المضارع", ''lmoḍa’ri‘') :** il indique une action qui se produit au moment de l'énonciation **(عَلَّمَ-أُعَلِّمُ).** L'imparfait est utilisé pour évoquer des actions en train de se dérouler ("يلعبون / ندرس / يعمل").

  Le verbe imparfait peut être sous les deux modes : mode actif et passif. De même, il y a d'autres types des verbes imparfaits comme : المضارع المعلوم، المضارع المجزوم، المضارع

- **L'imparfait du subjonctif (المضارع المنصوب)** : Le subjonctif indique l'attitude de celui qui parle envers l'événement et il est marqué (précédé) par un ensemble de particules: لن - حتى - لكي - كي - ل – أن. Comme "لندرس / كي ندرس / لكي ندرس / حتى ندرس /..."

- **L'imparfait apocopé (المضارع المجزوم) :** Le mode «jussive» est généralement marqué par la particule négative "لم" et la particule conditionnel "إن". Exemple "لم يذهب"

  Le verbe apparaît également dans le jussive à l'impératif. L'impératif est identique à la forme du verbe imparfait sans le préfixe ta-"ت". Exemple " **أُدرُسْ! اجلسي !** "

  Toutefois, quand le négatif est à l'impératif, le verbe apparaît sous la forme imparfaite avec le préfixe ta "ت", et toujours dans le mode jussive. Comme " **!** لاتدرسن **!** لاتجلسي"

- **Voix passi**f (**المبني للمجهول**), **dans ce mode, on trouve deux sous modes :**

  - **Verbe passif parfait:** La forme du verbe passif du parfait est formée en mettant un signe diacritique "ُ" ضمة sur la première consonne du mot et un signe diacritique "ِ" كسرة de l'avant dernière. قال قيل / شاهد شوهِد / درّس دُرِّس/كتب كُتِب / أكل أُكِل.
  - **Verbe passif inaccompli** : La forme du verbe passif de l'imparfait est formée en mettant un signe diacritique ضمة sur la première consonne du mot et un signe diacritique فتحة avant la dernière lettre (يَقُول-يُقال ، يَكْتُبُ - يُكْتَبُ).

- **Verbe impératif (فعل الأمر)** : indique une action à effectuer dans le futur, ou une demande (commandation) pour faire une action "عَلَّمَ-عَلِّمَا".

III-2.2 Nom arabe

C'est un mot qui désigne une personne, une chose ou une idée. Il peut être déterminé "البَيتُ" ou indéterminé "بَيتٌ ". Il peut être catégorisé par la personne, le cas (nominatif, accusatif et génitif), le nombre (singulier, duel, pluriel), le sexe (masculin, féminin) et par les cas grammaticaux "الرفع", "الجر" et "النصب". Le nom peut être soit un nom propre, soit un nom non propre qui peut être à son tour un nom commun ou un nom construit. Il peut être également un nom relatif (اسم موصول), un pronom personnel ou un nom démonstratif (اسم إشارة).

Un nom en arabe peut être également un substantif, un adjectif, un adjectif numéral, un pronom ou un nom propre (Khoja, 1999). Les pronoms peuvent être démonstratifs, relatifs, personnels, interrogatifs ou indéterminés.

Les noms à l'infinitif sont des substantifs abstraits, qui expriment l'action, la passion ou l'état indiqué par le verbe correspondant, sans aucune référence à l'objet ou le temps. Il s'agit notamment des dérivations de la racine, les noms formés à partir des formes dérivées du verbe, sont les noms qui expriment l'accomplissement d'une action, les noms du genre, les noms de lieu et de temps et les noms des instruments.

Il existe deux types de nom en arabe: noms dérivés et noms non dérivés. Les noms dérivés proviennent de la racine à laquelle ils sont sémantiquement liés. Il y a près de 400 formes

morphologiques de substantifs dérivés et non dérivés (El Sadany et haschich, 1989)(M.al-Ğalāyīnī, 2000).

a) Noms dérivés arabes :

Les noms dérivés sont les noms qui peuvent être dérivés d'une racine verbale. Le nombre et la nature de ces formes varient selon le statut du verbe auquel ils se rattachent. Parmi les noms dérivés, nous citons (Sawalha, 2011; Mesfar, 2006):

- **Le participe actif (اسم الفاعل) :** est un nom associé à tout verbe d'action, et qui désigne l'agent du verbe, c'est-à-dire celui qui fait l'action. (ضَارِبٌ-ضرب ، فاعل)

- **Le participe passif (اسم المفعول) :** est un nom associé à tout verbe d'action transitif. Il désigne le patient qui subit l'action ou le résultat de cette action. Un nom dérivé qui indique un sens abstrait qui décrit quelque chose ou quelqu'un affecté par une action.

  (مفعول، مسموع- سمع)

- **Nom verbal (المصدر) :** est un nom abstrait formé sur la même racine que le verbe auquel il est associé et il exprime le même contenu sémantique que le verbe. Un nom qui indique un cas ou une action qui n'est pas liée au temps. Un verbe peut avoir plus qu'un nom verbal (وَدَّ ،وُدًّا – وِدَادا- وِدادةً – مَوَدَّةً). Les schèmes de ce type de noms sont respectivement (فعلا – فعالا – فعالة مفعلة").

- **La qualité similaire (الصفة المشبهة) :** Les noms de la qualité similaire indiquent la présence absolue de la qualité de celui qui a fait l'action, Un nom dérivé qui indique un sens de la fermeté à savoir l'existence absolue de la qualité dans son possesseur.

  (رؤف-رؤوف ، كَرُمَ- كَرِيم ، شَجُعَ- شُجَاع)

- **Le nom préférence (اسم التفضيل) :** Un nom dérivé utilisé pour les personnes ou les choses comparatives et superlatives lorsque l'on compare. Il indique la qualité commune de deux noms dont l'un exprime un degré supérieur (أخوَفُ - أفعل ، فضل- أَفْضَل , صغُر- أصغر , كبر - أطول-طال , أغلى- غلا, أكْبَر).

- **Les noms de lieux et de temps (اسم الزمان والمكان) :** Ils indiquent l'endroit ou le temps de l'action, comme "مَأْخَذٌ , ملعب" et son schème "مفعل" est le même pour les noms des lieux suivants "مقام– قام , مَـقـرّ- قـرَّ " c'est le cas aussi pour les noms des temps (وسم , موعِد- وعد – موسِم).

- **Le nom d'instrument (اسم الآلة) :** Un nom dérivé qui indique un outil utilisé pour certains travaux. Il indique le moyen par lequel l'action a été réalisée, comme "ملعقة" et son schème "مفعلة" et le même pour (الغسالة– غَسَلَ , المطرقة– طَرَقَ , المنظار– نَظَرَ).

- **Al-masdar al-mimi (المصدر الميمي) :** Un nom qui indique un cas ou une action qui n'est pas liée au temps. Il présente un schème auquel on ajoute la lettre "mim" "م" au début du mot (موعد- وعد ,شَرِبَ-مَشْرَبٌ-أخرج-مُخرَج).

- **Nom propre (اسم العلم) :** Le nom des personnes ou de pays ou des villes. comme "أحمد , المغرب".

- **Diminutifs (اسم التصغير) :** Un nom déclinable qui a le son yaa "ي" après sa deuxième lettre de la racine. Il indique la rareté, le mépris ou l'affection (عُصفُور , بُنَيَّة-بِنْت , جُبَيْل-جَبَل-عُصَيْفِيْر).

- **Forme d'exagération (صيغة المبالغة) :** Il indique l'exagération de la qualité du nom qualifié et se produit comme un nom dérivé de la signification de base du participe présent. (علم – عليم , مبذال– بذل , عوَّام- عام , بشير- بشَّر).

- **Masdar El-marra (مصدر المرة) :** Un nom qui décrit une action qui a eu lieu qu'une seule fois. Il est formé en ajoutant la terminaison féminine (ة) pour le nom verbal. (رحمة دعوة...).

- **Nom de l'Etat (مصدر الهيئة) :** Un nom qui décrit une action. Il indique la manière (l'état, le caractère, et la représentation) de l'action exprimée par le verbe. (ابتدأ , انتِفاضةً- انتفض – ابتداءَة).

III 2-3 Les particules arabes :

Les particules sont classées en deux grandes catégories. La première est constituée de particules non significatives (حروف المباني) "huruf al-Mabani" comme (أ، ب، ت، ه، و، ي،...) ce sont les 28 lettres de la langue arabe. La deuxième est constituée de particules significatives (حروف المعاني) "ḥuruwf âl-ma'ani" comme (في، هل، بل، لا، نعم، بلى، إي، أجل، ألا، هلاّ، لولا، لوما،).

Les particules arabes sont des mots qui n'appartiennent ni au nom ni au verbe, mais elles ont la particularité d'affecter la forme du nom ou du verbe comme les particules régissant (حروف عاملة) (...لا، نعم، بلى، إي، أجل) qui (أن، إن، لكن، كأن، ليت،.., et les particules non régissant (حروف غير عاملة) n'affectent pas la forme des noms ou des verbes (Dahdah, 1987; Dahdah, 1993). de même il y a certaines particules qui coordonnent deux ou plusieurs phrases (Les particules de conjonctions) (حروف العطف), comme (و، ف، ثم، أو، لكن،...).

IV- Les différents niveaux de traitement des langues naturelles

Les langues naturelles s'appuient sur des règles phonologiques, morphologiques, syntaxiques et sémantiques pragmatiques pour former des mots ou des phrases. Le niveau de difficulté et de complexité dépend de la nature de chaque langue.

IV -1 Niveau Phonologiques

La phonologie s'intéresse à tous les aspects sonores de la langue. Parmi les aspects particulier de la phonologie, les phonèmes, c'est-à-dire les unités qui permettent de distinguer les mots les uns des autres. On peut distinguer dans l'analyse phonologique plusieurs niveaux, qui sont des niveaux de traitement de l'information est le niveau de stockage à long terme de l'information. De nombreuses études ont ainsi été consacrées aux relations entre l'apprentissage de la lecture et les habiletés phonologiques, c'est-à-dire la capacité à opérer une analyse phonologique du langage oral(reference;;;;;;) .

IV -2 Niveau Morphologique

La morphologie est la branche de la linguistique, elle étudie les modèles de formation des mots à travers les langues, et les tentatives de formuler des règles qui modélisent la connaissance des locuteurs de ces langues, c'est aussi l'identification, l'analyse et la description de la structure des morphèmes d'une langue donnée.

IV -3 Niveau Syntaxique

En linguistique, la syntaxe est l'étude des règles et des processus par lesquels les phrases sont construites dans une langue donnée (Chomsky, 1957).

En plus, le terme "syntaxe" est utilisé pour se référer directement aux règles et aux principes qui régissent la structure des phrases d'une langue. La recherche moderne dans la syntaxe tente de décrire les langues en termes de ces règles. De nombreux linguistes ont tenté de trouver des modèles généraux qui s'appliquent à toutes les langues naturelles.

IV -4 Niveau Sémantique

La sémantique est un mot grec qui signifie l'étude du sens. Elle s'intéresse à la relation entre signifiants, comme les mots, les phrases, les signes et les symboles. Elle a pour objectif la compréhension de l'expression humaine à travers le langage.

V- La morphologie des langues naturelles

La morphologie est un élément essentiel dans le traitement du langage naturel. Comme la morphologie en arabe est très dérivationnelle, donc l'analyse morphologique peut être facilement automatisée.

La langue arabe a un vocabulaire trés riche et sa représentation morphologique est assez complexe en raison des variations morphologiques et le phénomène d'agglutination (Kadri et Benyamina, 1992). L'arabe est une langue très flexionnelle avec 85% des mots dérivés de racines trilatérales. De plus, il y a environ 10.000 racines indépendantes (Al-Fedaghi et Al-Anzi, 1989). Les racines arabes sont entourées d'un nombre de préfixes, suffixes, ou des deux. En outre, le préfixe et le suffixe pourraient être associés à tout type de mots arabes comme le nom, le verbe, l'adjectif, etc. Le préfixe et le suffixe sont considérés comme des indicateurs de la catégorie grammaticale du mot comme la personne, le nombre et le genre. Les langues flexionnelles sont classées morphologiquement en deux grandes classes:

Les langues flexionnelles sont celles dont les mots se forment d'une racine à laquelle sont accolés des morphèmes supplémentaires. On peut distinguer ces derniers facilement de la racine, ils peuvent fusionner avec la racine ou entre eux.

Dans les langues flexionnelles, on accorde moins d'importance à l'ordre des mots que dans les langues analytiques. Par voie de conséquence, la morphologie devient plus importante que la syntaxe dans les langues flexionnelles.

Les langues flexionnelles se divisent en trois sous-types, selon que les morphèmes sont clairement différenciables ou non. Ces trois sous types sont :

1. **Les langues agglutinantes:** Les langues agglutinantes sont celles où les morphèmes sont identifiables phonétiquement les uns des autres. Une langue agglutinante a tendance à avoir un plus grand nombre de morphèmes par mot, et à être extrêmement régulière. Parmi les langues agglutinantes, nous pouvons citer le finno-ougriennes, le turc, le japonais, etc.

   Exemple:

   - En Finnois, le mot "taloissani" (dans ma maison), décomposable en talo-i-ssa-ni («maison»-pluriel-inessif-possesseur singulier 1ère personne).
   - En Turc, le mot "evlerimde" (dans ma maison), décomposable en ev-ler-im-de («maison»-pluriel-possesseur singulier 1ère personne-inessif).

2. **Les langues fusionnelles (synthétiques):** Dans ces langues, telles que les langues indo-européennes (excepté l'arménien), les flexions se font aussi désinences (terminaisons) où les différents constituants de la flexion ne sont généralement pas distincts. Il devient parfois difficile de distinguer les morphèmes de la racine, ou les morphèmes les uns des autres. Plusieurs morphèmes peuvent former un seul affixe, et les affixes peuvent, à leur tour, interagir et fusionner entre eux.

3. **Langues à flexion interne:** Dans cette catégorie de langues, les consonnes indiquent le sens et les voyelles marquant la flexion du mot, ce type se rencontre surtout dans les langues sémitiques (arabe, hébreu...).

   Ce type des langues connait des modifications morphologiques qui ne se manifestent pas par addition d'éléments à une racine, mais par la transmutation de la racine elle-même. C'est le cas des langues sémitiques, où, en général, la structure de la racine est conservée, et les

morphèmes grammaticaux sont exprimés par des changements vocaliques (كِتَاب" - كُتُبٌ, foot - feet, man - men).

VI- La morphologie de la langue arabe

La morphologie est un langage spécifique est composante, qui contient les règles morphologiques de la langue. Il est un niveau intermédiaire sur la branche forme phonologique qui est responsable de l'exploitation sur les structures syntaxiques afin de les rendre conformes à ces règles avant qu'elles ne soient réalisées phonologiques. Afin d'étudier la morphologie arabe d'une façon complète et efficace, nous exposerons les points suivants:

VI-1 Spécificités de la morphologie arabe

a. **La conjugaison ou génération :**

- **la conjugaison des verbes** : l'étude de la façon dont les verbes sont dérivés de la racine et comment ils se conjuguent dans les différents temps selon le genre, le pluriel, la voix et d'autres aspects associés à ces verbes.
- **noms dérivés** : ils sont issus à partir d'une phase de génération qui se fait à partir des racines.

b. **Le paradigme verbal** : prend la forme verbale de la racine et toutes ses conjugaisons en plus de l'ajout de certaines lettres supplémentaires afin d'améliorer son sens et ajouter certaines connotations.

VI-2 Difficultés de la morphologie arabe

La langue arabe présente des caractéristiques morphologiques qui la rendent plus ambigüe que d'autres langues naturelles.

La représentation morphologique arabe est assez complexe à cause du phénomène d'agglutination et que les lettres changent de forme en fonction de leur position dans le mot (début, milieu, fin et séparées).

Parmi les difficultés présentées dans la langue arabe, nous citons par exemple :

VI-2.1 Gémination [7] (الإدغـام):

La gémination est le phénomène de renforcement de l'articulation consonantique qui en prolonge la durée et en augmente l'intensité. Ce phénomène est parfois appelé redoublement, bien qu'il n'y ait pas véritablement répétition de la consonne. En arabe standard, la gémination est parfois indiquée par le signe diacritique spécial appelé chaddah, elle est du à la fusion entre deux consonnes de même nature de telle façon à constituer une seule lettre (مدد - مدّ). La gémination se trouve en deux lieux :

- Le premier cas est le plus récurent dans l'étude morphologique, il se trouve sous forme de deux lettres (consonnes) semblables qui se suivent dans le mot. (مدد - مدّ).

- Le deuxième cas entre plutôt dans les préoccupations des sciences du coran et consiste en la gémination de deux lettres (consonnes) différentes, mais de même nature et qui se suivent dans le mot par la transformation d'une des deux consonnes de façon à la rendre identique avec la seconde consonne, puis le fusionnement des deux consonnes semblables (انمحى-امَّحى).

VI-2.2 Affaiblissement (الإعلال)

Ce sont les modifications qui affectent les voyelles dans un mot arabe. Elles s'effectuent soit par le fait de marquer ces voyelles par le signe diacritique suku:n "سكون" "يجري l'origine et "يجريْ" , soit par la modification de la nature de la lettre "يقولُ l'origine et يَقْوُلُ ", ou par la suppression de cette lettre (l'élision) "الإعلال بالحذف" "al-i'lal bi-alhadf " يَعِدُ l'origine et يَوْعِدُ ", ou par la transformation complète "عاد l'origine et عَوَدَ".

4. L'affaiblissement par transport : il s'agit de transporter le signe diacritique qui vient après la lettre faible en mouvement vers la consonne sans mouvement qui la précède. En voici un exemple : le verbe dit (قال) devient (يَقْوُلُ). Ce type d'affaiblissement n'intervient que si nous avons les voyelles waw "و" et ya: "ي" car elles peuvent admettre un mouvement, contrairement à la voyelle alif "ا".

[7] : Pour plus de détails voir l'annexe.

5. L’affaiblissement par la suppression : il s’agit ici de supprimer la lettre faible ; ainsi le verbe s’arrêter (وقف) devient au présent يقف.

6. L’affaiblissement par la transformation : il s’agit de transformer les lettres و et ي en ا. C’est le cas qu’on trouve dans les verbes dire et vendre (قال و باع). La racine de ces verbes est comme suit : (قول، بيع).

VI-2.3 Substitution (الإبدال)

C'est le fait de remplacer une lettre par une autre pour faciliter la prononciation. Elle consiste en réalité dans la substitution des lettres saines (ahrufun sahihatun أحرف صحيحة) entre elles; ou bien par leur substitution par des lettres affaiblies (ahrufu al- 'ilati). La substitution ne s'effectue qu’à l'intérieur d'un certain nombre connu et limité de lettres que les grammairiens délimitent dans le nombre de neuf, regroupées dans" هدأت موطيا". Il existe plusieurs types de substitution, mais on se contentera de citer les cas les plus connus :

7. **La substitution de la lettre "و " et la lettre "ي"** par la lettre "ت" : ce changement affecte les verbes dont le schème est (افتعل) et tous les dérivés de ces verbes. cette transformation a lieu si la première lettre de la racine est faible. En voici un exemple concret : Les mots " اوتقد " et " اوتصف " ("وقد " et "وصف") deviennent "اتّصف" et "اتّقد".

8. **La substitution de la lettre "تاء" par la lettre "د"**: ce changement s’effectue si la première lettre de la racine est l’une des consonnes ز , د et ذ. Alors dans ce cas , la lettre d’ajout "ت" est substituée par "د" (ادْتَحر- ادّحر، ادْتَثر- ادّثر،).

9. **La substitution de la lettre "تاء" par la lettre "طاء " :** il s’agit d’une substitution qui ne peut avoir lieu que si nous avons une consonne emphatique au début de la racine (ص، ض، ط، ظ), mais il faut qu'elle soit suivie dans le schème par la lettre d’ajout "ت" (اطتلع - اطّلع).

10. **La substitution de la lettre** "النون الساكنة" par la lettre "الميم" (انمحى - امّحى).

VI-3 Morphèmes arabes :

Pour les morphèmes arabes, on trouve: les préfixes, les racines, les radicaux, les suffixes et les infixes:

1. **Racine :** la racine est l'unité lexicale primaire d'un mot, ou d'une famille de mots (racine est appelée aussi mot de base). Elle porte sur les aspects les plus importants du contenu sémantique et ne peut être réduite en petits constituants. La langue arabe contient plus de 10000 racines, et elles sont constituées souvent de trois consonnes, rarement de quatre consonnes (ق و ل، ع ل م).

2. **Le préfixe :** est un morphème ajouté au début du radical pour générer un autre mot. Les préfixes ajoutent aussi de sens au mot, tels que: l'accent, la transitivité, etc. Pour la langue arabe, il existe plusieurs préfixes; par exemple (.... ف، ال ، فسي) (voir l'annexe B).

3. **Le suffixe :** c'est un morphème attaché à la fin du radical pour générer un autre mot. Les exemples courants sont les terminaisons de cas, qui indiquent le cas grammatical des substantifs ou des adjectifs, et les terminaisons qui forment la conjugaison des verbes. En particulier dans l'étude des langues sémitiques, un suffixe est appelé afformatif, car il peut modifier la forme des mots auxquels il est fixé. (...ون، هم، ها، كما)(voir annexe B).

4. **L'infixe :** il est inséré à l'intérieur d'un radical pour former un autre mot (ي، و، ا) exemple (مدخول).

5. **Le radical :** En linguistique, Le terme est utilisé avec des significations légèrement différentes. Le radical est la partie d'un mot à laquelle s'attachent les affixes flexionnels. (يقول-قول، فكاتب-كاتب), appelé aussi stem dans le cas de Buckwalter.

6. **Lemme** : dans la morphologie et la lexicographie, un lemme est la forme canonique, forme de dictionnaire, ou forme de citation d'un ensemble de mots. Par exemple le verbe "Ecrire - كتب " produit plusieurs lemmes "Livre- كتاب", "Ecrivain-كاتب", "Ecrit-مكتوب", sont les formes du même lexème. Le lexème, dans ce contexte, fait référence à l'ensemble de toutes les formes qui ont le même sens, et le lemme se réfère à la forme particulière qui est choisie par convention pour représenter le lexème. Le processus de détermination du lemme pour un mot donné est appelé lemmatisation.

VII- Les schèmes[8]

VII-1 Schèmes classiques

La majorité des mots arabes sont des mots dérivés. Tous les types de mots (verbes, noms, adjectifs et adverbes) sont générés à partir des racines basées sur un nombre limité de schèmes. Le schème est utilisé pour générer un mot sous la même forme que ce schème est à partir d'une racine donnée "د خ ل","مفعول-مدخول". La plus grande partie des mots dérivés arabes est générée à partir des racines de trois lettres et de quatre lettres.

Les racines les plus utilisées dans la langue arabe moderne sont de l'ordre de 5000 (Beesley, 1996), (Alesco, 2007). Environ 3500 schèmes et schèmes étendus sont utilisés pour générer les formes des mots à partir de ces racines, Le nombre des schèmes est significativement réduit dans le cas des textes non voyellés, et moins de 50 % des schèmes sont couramment utilisés dans les textes modernes (Awajan.A, 2011).

Les schèmes sont composés de trois lettres origines "ف", "ع" et "ل", où "ف" correspond à la première lettre, "ع" à la deuxième lettre et "ل" pour la troisième "فعل" et la quatrième lettre "فعلل", plus les affixes pour les schèmes étendus (سيفعلون) (Al-Rajhi,1979; Al-Hamlawy,1957).

Exemples : يفعلون → يقولون

الفاعلون → الداخلون

VII-2 Schèmes de surface

Dans notre étude, nous avons construit les schèmes de surface (Sabri et Yousfi, 2006) à partir des mots dérivés. On remplace, dans chaque mot dérivé, les lettres de la racine par les lettres "ف", "ع" et "ل", où ف correspond à la première lettre de la racine, ع à la deuxième lettre de la racine et ل à la troisième lettre de la racine, par contre si l'un des caractères est égale à (ا , و , ي), on le garde à son place.

(8) Schèmes : أوزان

Schèmes de surface	Schème classique	Verbe	Mot
يفولون	يفعلون	جال	يجولون
الفائلون	الفاعلون	قال	القائلون
فاعِيَةٌ	فاعلةٌ	رعى	رَاعِيَةٌ
مَأعول	مفعول	أَخَذَ	مَأخُوذٌ

Tableau n° 2 : les schèmes de surface a partir des mots

Ce type de schèmes est utilisé pour modéliser les transformations morphologiques de la racine dans un mot dérivé.

VIII- Le traitement automatique de la langue arabe

Au cours des dernières années, le traitement automatique de la langue arabe a pris une importance croissance, et plusieurs systèmes ont été développés pour une large gamme d'applications. Ces applications ont dû faire face à plusieurs problèmes complexes pertinents à la nature et la structure de la langue arabe (Farghaly; 2009). Le manque de ressources disponibles et leurs limites ont motivé de nombreux chercheurs à suivre l'approche à base de règles et de s'appuyer sur des règles linguistiques pour le développement des outils, des systèmes et des ressources linguistiques.

Des techniques permettant le traitement automatique ou quasi-automatique de l'information textuelle ou vocale à l'aide des outils linguistiques, statistiques, physiques, informatiques. Faciliter les interactions Homme-Machine, Les outils du TALN, sont utilisés dans l'analyse des contenus textuels ou vocaux, Data Mining, Big Data, etc.

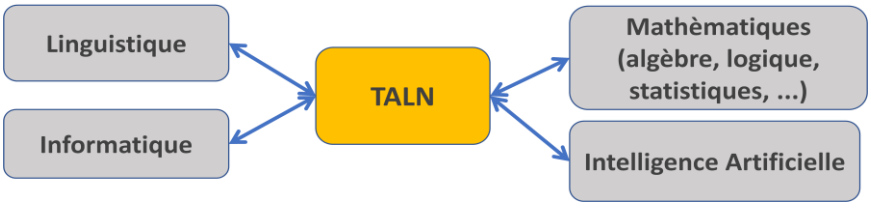

Figure n° 1 : Outils et techniques de Traitement Automatique du Langage Naturel.

L'arabe est une langue morphologiquement riche et complexe, chose qui présente des défis importants pour son traitement automatique. Des différents chercheurs ont travaillé sur plusieurs axes du traitement automatique de la langue (TAL) arabe, parmi ces axes nous citons :

- L'analyse syntaxique étudie la structure syntaxique des phrases, en prenant en considération les positions respectives des mots (Ouersighni, 2001).
- L'analyse sémantique : L'analyse sémantique, qui vise à lever les ambiguïtés subsistant après l'analyse syntaxique, par référence aux relations sémantiques qui lient les concepts entre eux. Il permet de déterminer le sens des mots. Un même mot peut signifier plusieurs choses. Par exemple le mot "الكرة" peut signifier "كرة القدم", comme il peut prendre le sens de la terre "الكرة الأرضية". Il ne se base pas sur l'écriture d'un mot mais sur le sens de celui-ci. Cela s'oppose à l'analyse lexicale qui se base sur un lexique et à l'analyse grammaticale qui se base sur la grammaire.
- L'analyse pragmatique : qui replace la phrase dans le contexte du domaine général de la connaissance afin de lever les ambiguïtés qui ne peuvent être levées par l'analyse sémantique (Hilal, 1985).
- L'analyse morphologique est un procédé de calcul qui analyse les mots naturels en considérant leurs structures internes. La structure interne d'un mot peut inclure le radical, la racine, les affixes, et les schèmes (T.Buckwalter, 2004; KR.Beesly, 1998; k.Darwish, 2002).
- La recherche d'information est l'activité de l'obtention de ressources d'informations pertinentes à un besoin d'information à partir d'une collection des ressources d'informations (Frakes, William B., 1992).
- La traduction automatique : est une traduction effectuée par ordinateur, sans intervention humaine, en se basant sur des règles grammaticales, des règles linguistiques et des dictionnaires de mots courants. Ce processus permet de traduire un texte d'une langue naturelle (par exemple, l'anglais) vers une autre (par exemple, l'arabe).
- Le développement des corpus, ces derniers sont presque indispensables dans n'importe quelle application de traitement automatique des langues (voir annexe III**(9)**).

[9] Corpus arabe

Chapitre II : Analyse morphologique de la langue arabe

I- Introduction

L'analyse morphologique est une tâche centrale dans le traitement du langage qui peut prendre un mot comme entrée et détecter les différentes entités morphologiques du mot et en fournir une représentation morphologique. L'analyse morphologique a été et reste le centre des chercheurs dans le traitement automatisé de la langue arabe.

Les études sur la morphologique arabe au niveau de l'ordinateur ont reçu une grande attention de la part des ingénieurs informaticiens et des linguistes depuis le début des années quatre-vingt. Un grand nombre d'analyseurs morphologiques sont conçus pour être utilisés dans diverses applications. L'attention est due à la richesse et la complexité des morphologiques Arabes, l'importance apparaît également pour les analyseurs morphologiques dans les applications principales pour faciliter et apporter des solutions dans les domaines de la traduction automatique, la recherche d'informations et récupérer des informations.

Le premier analyseur morphologique automatique était proposé par David Cohen en 1961 (D.Cohen, 1961) d'après (Mesfar, 2006). C'était l'un des premiers chercheurs dans le domaine du traitement automatique de la langue arabe. En 1983, une analyse morphologique pour la langue

Finlandaise à deux niveaux était proposée par Koskenniemi (Koskenniemi, 1983). D'autres chercheurs ont adapté cette approche pour la langue arabe. Depuis 1996, le centre de recherche de Xerox a amélioré ce système en utilisant un algorithme de combinaison automatique entre les racines et les schèmes des radicaux.

L'importance des outils de traitement automatique de la langue arabe a considérablement augmenté dans la dernière décennie, en raison de l'énorme croissance du contenu numérique arabe dans le monde (en particulier sur le Web), d'où l'importance de créer des outils de traitement de ces contenus, et de les faire interagir avec les utilisateurs dans des meilleures conditions. La morphologie automatique est un domaine de recherche qui est devenu très actif pendant les 25 dernières années. La morphologie computationnelle est principalement concernée par l'analyse et la génération morphologique des mots.

Pour la langue, arabe plusieurs analyseurs ont été élaborés, nous citons par exemple celui de Beesly (Beesley, 1996), de Buckwalter (Buckwalter, 2004), de Hegazi et Elsharkawi (Hegazi et Elsharkawi, 1986) et celui de Al-Fedagi et Al-Anzi (Al-Fedagi et Al-Anzi, 1989)….. Nous allons détailler ces différents analyseurs ultérieurement.

II- Analyse morphologique pour une langue flexionnelle:

Soit W un mot d'un vocabulaire V d'une langue donnée, on note par $R = \{R_1, R_2, \dots, R_n\}$ l'ensemble des racines ou radicaux associées à cette langue $S=\{S_1,S_2,\dots..,\ S_p\}$ l'ensembles des suffixes et $P = \{P_1, P_2, \dots, P_k\}$ l'ensemble des préfixes, associés à cette langue.

L'analyse morphologique du mot W consiste à extraire l'ensemble des solutions:

$$E = \{(P_i, R_j, S_k) \in P \times R \times S \ tel\ que\ W = P_i f(R_j) S_k\} \qquad (1)$$

Avec $f(R_j)$ est une fonction de transformation.

Exemple : فقلت ⟶ $E\{($ف،قال،ت$)$، $($ف،قل،ت$)\}$

Dans la première solution (ف, قال, ت): $w =$ ف $f\ ($قال$)$ت et $f($قال$) =$ قال

Dans la deuxième (ف, قل, ت) : $w =$ ف $f\ ($قل$)$ت et $f($قـ$) =$ قل

III- Approches utilisées dans les systèmes d'analyse morphologique

Il existe différentes manières de construire des analyseurs morphologiques, il y a eu des recherches visant à programmer des règles de grammaire, morphologiques, grammaticales et orthographiques afin de construire des systèmes d'analyseur morphologique. Le développement exige l'étude des propriétés des données des mots en soulevant essentiellement des problèmes concernent l'analyse morphologique et la présentation des mots arabes.

Les méthodes utilisées dans la construction des analyseurs morphologiques sont assez variées. En effet, certains chercheurs ont développé des méthodes basées sur la recherche des signes diacritiques au niveau du caractère, d'autres ont exploité ces méthodes pour identifier les signes diacritiques au niveau du mot, un groupe de chercheurs a développé des méthodes hybrides couplant les approches pour améliorer ces méthodes. Darwish (2002a) a suggéré de classer les approches dans l'approche symbolique, l'approche statistique et l'approche hybride.

Les chercheurs dans le domaine d'analyseur morphologique utilisent plusieurs méthodes pour analyser un mot :

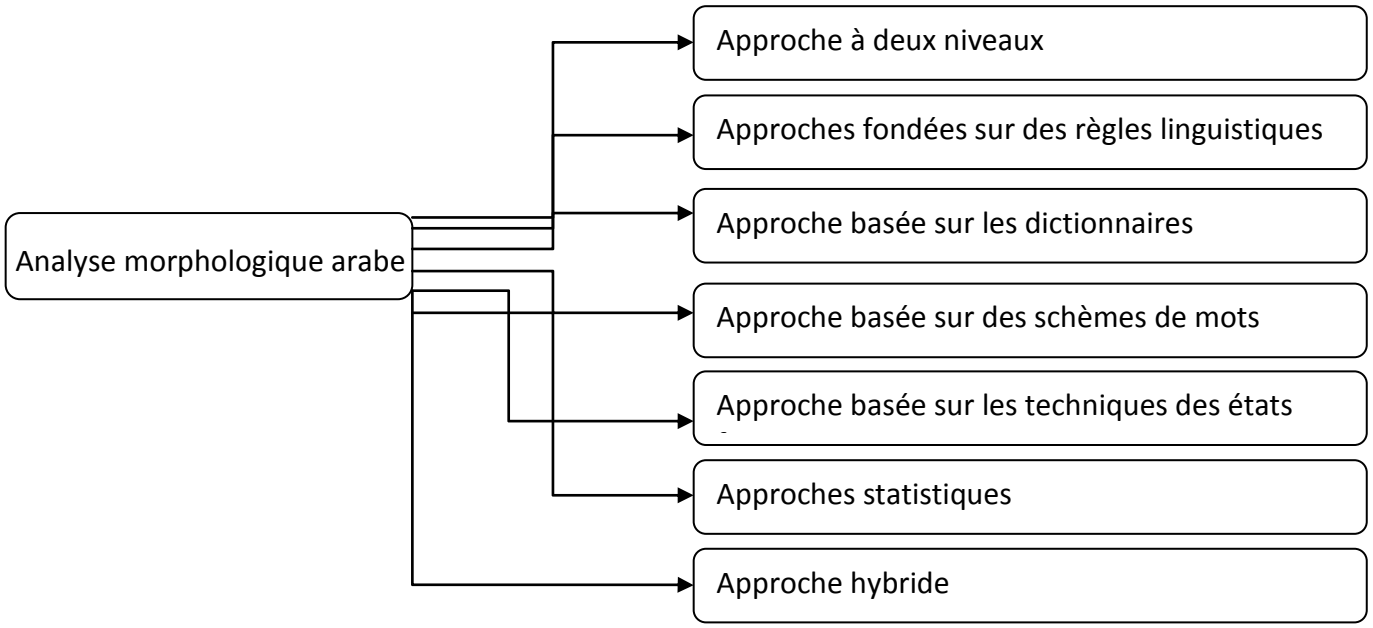

Figure n° 2: Approches utilisées dans les analyseurs morphologiques arabes

III-1 L'approche à deux niveaux

Elle comporte le niveau lexical qui représente les morphèmes et les niveaux de surface qui représentent les formes de surface. Elle a été définie pour la 1ère fois par (Koskenniemi,1983) pour la langue finlandaise. L'approche à deux niveaux considère le mot comme une composition de deux couches: la racine et le schème.

III-2 Approche symbolique

Cette approche est basée sur la segmentation du mot en préfixes, radicale(stem), infixes et suffixes dans le but d'extraire la racine du mot arabe, après la suppression de tous les préfixes, les infixes et les suffixes attachés. Plusieurs analyseurs morphologiques ont été élaborés et se sont basés sur cette approche (Darwish, 2002; Buckwalter, 2002 ; Hegazi et ElSharkawi, 1986 ; Koskenniemi, 1983 ; Beesly, 1998 ; El-Sadany et Hashish, 1989 ; Khoja et Garside, 1999 ; Soudi, 2002).

III-3 Approches fondées sur des règles linguistiques

Pour les analyseurs morphologiques arabes plusieurs chercheurs utilisent des approches basées sur des règles linguistiques. Ils utilisent une base de connaissances de règles écrites par des linguistes pour attribuer des solutions aux différents attributs morphologiques des mots arabes. L'approche basée sur des règles linguistiques qui utilisent les algorithmes qui se basent purement sur les connaissances morphologiques de la langue. Il nécessite des règles pour couvrir toutes les formes morphologiques. Ces règles sont souvent classées en catégories grammaticales, structurelles et logiques. Consiste à utiliser des critères et des propriétés linguistiques sous forme de règles exprimant le fonctionnement du langage naturel utilisé.

L'approche linguistique nécessite un grand nombre de listes et de tableaux. Pour élaborer un ensemble de règles permettant de trouver la décomposition appropriée, cette approche se base sur une analyse morphologique approfondie de la langue arabe (Buckwalter, 2002) (Al-Fedaghi et Al-Anzi, 1989).

L'approche linguistique subjective simule le processus utilisé par un expert linguistique. Elle consiste à supprimer des affixes par comparaison avec des listes prédéfinies et à transformer ce

qui reste, le radical en racine, après une éventuelle altération par l'ajout, la suppression ou la modification de certaines de ses lettres.

III-4 Approche basée sur les dictionnaires

Les ressources utilisées pour un analyseur morphologique sont un dictionnaire de mots racines qui a été créé manuellement à l'aide de différentes ressources. La principale ressource pour la base de données du dictionnaire et également pour les données de formation et de test de cet analyseur morphologique. De plus, un dictionnaire morphologique est utilisé à la fois pour un analyseur morphologique et un générateur morphologique, en fonction du sens dans lequel il est lu par le système (Buckwalter, 2002).

III-5 Approche basée sur des schèmes de mots

L'utilisation de schèmes morphologiques, selon ce qui est implique des affixes morphologiques sous toutes ses formes, il existe des schèmes pour les verbes, des schèmes pour les noms, des schèmes pour les adjectifs, etc. Il existe des schèmes communs entre ces types.

Et pour faire ce type d'analyse, il est nécessaire de déterminer les schèmes morphologiques, et en comptant les morphes qui les entrent, et en comptant les morphèmes qui y sont incluses et la liste entre eux et les affixes qu'ils partagent avec eux au début ou à la fin des mots. (Beesley, 1998), (Iazzi et Yousfi, 2013) , (Yousfi, 2010).

III-6 Approche basée sur les techniques des états finis

L'approche des états finis est dominante depuis les années 1980. L'approche des états finis pour l'analyse morphologique a été initialement étudiée chez Xerox et la première application pratique était due à Koskenniemi (Koskenniemi 1983) ; cela a été utilisé pour développer des analyseurs morphologiques à large couverture pour plusieurs langues.

Dans ce type d'approche, les règles morphotactiques et orthographiques sont programmées dans un transducteur à états finis (FST), elles nécessitent trop de traitements manuels pour énoncer des règles dans un FST et ne pas analyser des mots qui n'apparaissent pas dans les dictionnaires arabes (KR Beesley et L. Karttunen, 2003) (Al-Sughaiyer et Al-Kharashi 2004), (Dichy & Fargaly, 2003). D'autres analyseurs utilisent des graphes pour faire l'analyse morphologique des mots arabes (Iazzi et Yousfi, 2013) (Iazzi, S, Yousfi, A, SITA, 2020).

III-7 Approches statistiques

Cette approche utilise la probabilité de succession de certains morphèmes pour faire l'analyse morphologique d'un mot donné. Les données statistiques obtenues par un corpus permettant d'acquérir des connaissances sur la morphologie de la langue, il apprend des préfixes, des suffixes et des schèmes à partir d'un corpus ou d'une liste de mots dans la langue cible sans aucune intervention humaine. Cette approche utilise une liste de préfixes, de suffixes et de schèmes pour transformer le stem à une racine. Les combinaisons possibles préfixe-schème-suffixe sont construites pour un mot afin d'obtenir les racines possibles. Cela prend plus de temps, car il faut effectuer des tâches de prétraitement et demande des corpus de grande taille (Darwish, 2002).

C'est les possibilités et les probabilités que peut avoir un préfixe, un suffixe et un radical à paraitre ensemble dans une base de données des mots (Goldsmith et John, 2001). Cette approche utilise une liste de préfixes, une liste de suffixes, et des schèmes pour extraire un radical d'un mot. Une telle approche permet d'obtenir une large couverture morphologique de la langue arabe.

Les inconvénients de cette approche est la nécessité d'inclure des règles linguistiques manuelle, et ceci exige beaucoup de temps et une bonne connaissance du système orthographique et morphosyntaxique de langue arabe.

L'avantage de cette méthode est que la ressource est complète et prend en compte tous les faits linguistiques, c'est la possibilité d'obtenir une couverture morphologique plus large de la langue arabe.

III-8 Approche hybride

Les méthodes hybrides sont des algorithmes qui combinent plusieurs approches déjà mentionnées. Elle utilise une liste de préfixes, une liste des suffixes et des règles de transformation à partir d'un radical à une racine. Les combinaisons (préfixe – schème- suffixe) possibles sont construites pour un mot pour extraire les racines possibles. Bien que ces approches atteignent une couverture morphologique plus large, la plupart des travaux récents utilisent cette approche puisqu'elle donne les meilleurs résultats. Par exemple dans le cas de l'analyseur Buckwalter, ce dernier combine entre les règles linguistiques (quand il introduit la notion de compatibilité entre préfixe, suffixe et radical) et les dictionnaires (quand il utilise le dictionnaire

des radicaux arabes) (Buckwalter, 2002). D'autres travaux utilisent ce type d'approche, nous citons par exemple (Alkhalil, 2017) (Azmi, Aqil, Reham S Almajed, 2015).

IV- Présentation de quelques analyseurs morphologiques arabes:

De nombreux travaux ont été menés au cours de ces dernières années et de nouvelles théories sont nées. Plusieurs analyseurs morphologiques pour la langue arabe ont été élaborés. Certains d'entre eux sont disponibles pour la recherche et l'évaluation, tandis que les autres sont des applications exclusivement commerciales. L'étude de ces analyseurs nous permet de dévoiler leur méthode de travail et l'approche utilisée.

Parmi les travaux connus dans la littérature, nous citons :

(Rafea et al., 1984) : ces chercheurs ont produit une analyse morphologique utilisée dans la compréhension automatique des mots arabes flexionnels. L'analyseur morphologique est appuyé par un algorithme pour supprimer les affixes et clitiques joints au mot et de vérifier si le résultat final est acceptable. Le processus se répète si nécessaire. Les radicaux sont stockés dans leurs différentes formes dans un lexique, qui comprennent des entrées qui ne peuvent pas être acceptées seules et n'ont pas de sens sans ajouts. Même avec la taille d'un lexique qui contient ces éléments supplémentaires, cette approche est beaucoup mieux que de stocker les mots flexionnels.

(Hilal, 1985) : le système de Hilal a basé son analyse sur les trois classes : une classe des plus grands préfixes, une classe du plus grand suffixe et une troisième classe utilisant le nombre de lettres restantes dans le mot). Hilal classe le mot arabe en outils et mots ordinaires. Les mots ordinaires suivent des règles grammaticales. Cependant, les outils ne suivent pas de telles règles. La méthode d'extraction des racines trilittérales d'un mot ordinaire suit les étapes générales indiquées ci-dessous :

I- le préfixe et le suffixe les plus longs possibles sont éliminés en comparant les caractères de début et de fin avec les préfixes et suffixes connus.

2- en fonction de la longueur de la partie restante, différentes règles sont examinées :

- Éliminer les lettres supplémentaires

- Modifier une lettre supplémentaire pour former une racine trilitère.
- Changer une lettre à sa valeur originale.

Beaucoup de tables de consultation sont utilisées pour accomplir la tâche, y compris les tableaux des schèmes, des racines, des préfixes, des suffixes et des racines non normalisées (filial, 1985; Al Fedaghi et Al Anzi, 1989).

- **(Hegazi.N, ElSharkawi.A., 1986) :** Ils ont mis en place, en 1986, un système morphologique des mots arabes. Le système reconnaît la racine d'un mot, le schème morphologique et sa catégorie morphologique. Ces chercheurs ont élaboré un système d'analyse morphologique basé sur un algorithme permettant d'extraire les positions des caractères constituant la racine. Cet outil permet dans une première phase d'extraire le plus long préfixe possible du mot à analyser. Ceci permet de déduire, ensuite, que la racine du mot est constituée par le premier caractère du mot restant. Le système utilise les règles de dérivation morphologique et les schèmes de chaque mot. Le système traite d'environ 400 schèmes, qui couvrent un peu tous les schèmes dans la langue arabe. Un équilibre syllabique est fait pour chaque schème morphologique.
- **(Al-Fedaghi et Al-Anzi, 1989) :** Ils ont présenté un algorithme d'extraction de la racine et du schème d'un mot arabe donné. Il utilise des méthodes mathématiques qui génèrent des formes de schèmes de mots arabes. Cet algorithme est basé sur la détection de la position des trois lettres de la racine trilitère et quadrilitère pour vérifier son existence dans la liste des racines déjà connues (Fedaghi, 89). Il examine, toutes les combinaisons des trois lettres dans le mot donné et produit sa représentation racine-schème. Deux fichiers sont utilisés dans cet algorithme en entrée: le fichier des racines trilatérales et le fichier des schèmes. L'algorithme examine les schèmes qui ont la même longueur que le mot d'entrée. Le concept principal de cet algorithme est d'examiner les tendances afin d'identifier les positions des lettres de la racine dans un radical donné. Ces lettres sont ensuite testées pour déterminer s'ils constituent une racine arabe ou non. L'algorithme a été testé sur divers textes, et le pourcentage de réduction des mots (racines extraites) varie entre 50 à 80%.
- (**Thalouth et Al Dannan, 1990) :** cet analyseur utilise les catégories des schèmes correspondant au mot et met en œuvre les différentes règles morphologiques. En outre, les préfixes et les suffixes ont été définis de telle manière que leur interférence avec des

schèmes a été minimisée. L'algorithme utilisé dans ce système découpe le mot en suffixes et préfixes les plus fréquents ; puis le mot correspondant à un ensemble de schèmes fréquents de manière à obtenir l'origine trilitère du mot. L'algorithme divise les mots en trois catégories, chacune avec son propre traitement.

- **(EI-Sadany et haschisch, 1989) :** Ils ont développé un système morphologique arabe également conçu pour effectuer l'analyse et la génération, capable de faire face à des mots voyellés, semi voyellés et des mots arabes non voyellés. Ce système a été développé au Centre Scientifique IBM du Caire.
- **(Beesley, 1989) :** il a développé un analyseur morphologique arabe en utilisant les générations de Xerox[(10)], basées sur des modèles de langages implémentés par des automates à états finis de koskenniem. Cette analyse, basée sur la morphologie à deux niveaux, est capable d'identifier tous les mots arabes avec une analyse morphologique valide. Cet analyseur morphologique est conçu comme un outil pédagogique et peut également constituer une étape de traitement dans un système plus vaste de traitement automatique de la langue.
- **(Saliba et Al-Dannan, 1989) :** Ils ont élaboré un système d'analyse et de génération morphologique pour le centre scientifique IBM au Kuwait. Cet analyseur permet de trouver toutes les analyses possibles d'un mot (Saliba 89). Dans le processus d'analyse, le plus long préfixe et suffixe valides sont supprimés du mot, et la partie restante du mot qui est appelée radical, est utilisée pour identifier un mot arabe valide. Si le radical est accepté comme un mot de la langue arabe (nom ou verbe), alors d'autres processus d'analyse seront effectués.
- **(Xerox, 1998):** Il a développé un système d'analyse et de génération morphologique, ce système est bien connu dans la littérature, disponible pour l'évaluation et bien documenté. Cet analyseur est construit en utilisant la technologie des états finis (FST) (Beesley, 1996 et 2000). Il adopte l'approche des schèmes-racines. En outre, il comprend 4930 racines et 400 schèmes, générant efficacement 90000 radicaux. Les avantages de cet analyseur sont d'une part sa large couverture et d'autre part, il est basé sur des règles et il fournit également un glossaire anglais pour chaque mot.

[(10)] Il existe une version de démonstration à http://www.xrce.xerox.com/research/mltt/arabic

- **(Khoja et Garside, 1999) :** L'analyseur morphologique de Khoja est basé sur quatre listes de préfixes, de suffixes, de racines et de schèmes. Quand il supprime le plus long préfixe et suffixe, il compare ensuite le mot restant en consultant la liste des schèmes et ce, pour extraire la racine. Puis, il vérifie dans une liste des schèmes pour déterminer si le reste peut être une racine connue avec un schème connu à l'aide d'un dictionnaire des racines arabes. L'analyseur de Khoja supprime les signes diacritiques, les signes de ponctuation, les mots vides et les nombres, il remplacer les lettres " ا,و,ي" par "و" et toutes les occurrences de hamza" ئ,ء,ؤ " par "أ". L'algorithme permet d'obtenir un taux allant jusqu'à 96% de précision et traite correctement la plupart des mots arabes dérivés.
- **(Attia, 2006) :** Lorsqu'un mot peut être dérivé de plusieurs racines différentes, la détection de la racine est encore plus difficile, et particulièrement en absence des voyelles. L'analyseur morphologique de Attia s'appuie sur la technologie d'états finis. Il est adapté à la fois pour l'analyse et la génération. il a utilisé un corpus d'articles de 4,5 millions de mots et prend le radical comme la forme de base. Il contient 9741 lemmes.
  L'avantage de ce système est le traitement des expressions de plusieurs mots. Le système peut gérer efficacement les noms composés de personnes, les lieux et les organisations. Un inconvénient du système, cependant, est sa couverture limitée. Il ne gère pas les textes diacritiques et vise une application particulière (analyse syntaxique).
- **(Berri, Zidoum et Atif; 2001):** Ils ont développé un autre analyseur morphologique qui se compose de trois éléments principaux: une base de connaissances qui contient les règles morphologiques régulières et irrégulières de la grammaire arabe, un ensemble de listes de mots contenant les exceptions gérées par des règles irrégulières et un algorithme de compatibilité qui correspond aux règles.
- **(Kiraz, 1995 2001) :** Il présente un système de correction orthographique qui s'intègre parfaitement à l'analyse morphologique en utilisant un formalisme multi-bande. les modules d'analyse morphologiques doivent rendre compte de la phonologie et de la syntaxe. Ces dernières augmentent la complexité de l'élaboration des systèmes d'analyse morphologique pour le texte arabe (Kiraz, 2001).

- **(Buckwalter, 2002)** : Cet analyseur est considéré comme l'un des plus référenciés dans la littérature. Il est bien documenté et sa source se trouve sur internet[(11)] et disponible pour évaluation. Buckwalter a développé un système d'analyse morphologique de la langue arabe qui consiste à déterminer toutes les segmentations possibles d'un mot, puis à chercher les résultats dans des listes des stems, des suffixes et des préfixes.

  L'approche de Buckwalter utilise trois fichiers de lexique: dictionnaire des préfixes, dictionnaire des suffixes et dictionnaire des stems:

  - Le dictionnaire des préfixes : contient tous les préfixes de la langue arabe et leurs enchaînements. Exemples : "و، ف، ب، ك، أ، ي، س، ل، ت، ن، ال". Le préfixe peut être de 0 à 4 caractères.
  - Le dictionnaire des suffixes : contient tous les suffixes de la langue arabe et leurs enchaînements. Exemples : "ا، ت، ه، أ، ك، ن، ف، ي، و،ة". Le suffixe peut être de 0 à 6 caractères.
  - Le dictionnaire des stems : contient tous les stems arabes. Le stem au sens de Buckwalter est défini comme le reste du mot après le retrait des préfixes et des suffixes.

  Exemple : après l'analyse de mot "ففلت" on trouve les partie suivantes "ف|قل|ت" avec "ف" et le préfixe, "ت" le suffixe de mot et "قل" est un stem associé à la racine 'ق و ل'.

  - Les tables de compatibilité : pour l'analyseur AraMorph version 2002, Buckwalter a utilisé une base de données composée de trois fichiers du lexique arabe-anglais: les préfixes (299 entrées), les suffixes (618 entrées) et les stems (82158 entrées représentant 38.600 lemmes). Les lexiques sont complétés par trois tables de compatibilité morphologique permettant d'ordonner les combinaisons préfixe- stems (1.648 entrées), des combinaisons stems-suffixe (1285 entrées) et des combinaisons préfixe-suffixe (598 entrées).

Exemple : l'analyse morphologique de mot 'فيسكتهم', il donne deux solutions :

- la1^{er} 'سك' après l'extraction de préfixe 'في' et suffixe 'تهم'.
- la 2ème 'سكت' après l'extraction de préfixe 'في' et le suffixe 'هم'.

[(11)] http://www.qamus.org/

Après l'application de la table de compatibilité, la première solution sera éliminée car le préfixe 'في' ne peut pas être regroupée avec le suffixe "تهم". Donc il garde seulement la 2ème solution 'سكت'.

- **(K.Darwish, 2002) :** Darwish 2002 propose une méthode pour extraire automatiquement des règles de lemmatisation en suppression des préfixes et suffixes à partir d'une liste des mots et racines arabe appariés, Ce système s'intéresse à la recherche des racines possible d'un mot. il développé l'analyseur morphologique «Sebawai» pour la langue arabe. Cet analyseur utilise une liste des préfixes, des radicaux, des racines, des suffixes et une liste des schèmes. Il applique dans la première étape une racinisation légère, et dans la deuxième, il cherche le schème approprié pour enlever les infixes et extraire les racines. Il réduit le radical à la forme de la racine sur la base d'une liste crée manuellement des paires de mot-racine.

  L'approche exploite des tables de consultation pour une liste de préfixes et de suffixes en essayant de faire correspondre toutes les racines produites par un schème et les modèles orthographiques arabes afin de retrouver la racine des mots. Son approche donne une précision de 92,7%.

  L'avantage de l'analyseur «Sebawai» est sa rapidité de la production des racines. Or, l'inconvénient de cet analyseur est qu'il ne peut pas donner des informations sur le mot analysé, ce qui le rend très limité pour la plupart des applications du traitement automatique des langues.

- **(Al-Shalabi et al., 2003) :** Ils ont développé un algorithme d'extraction des racines trilitères des mots arabes, Il dépend de calculs mathématiques des valeurs numériques attribuées aux lettres du mot puis on multiplie ces valeurs numériques par la position des lettres dans le mot. Des valeurs numériques plus élevées sont affectées aux lettres au début et à la fin du mot. Ensuite, l'algorithme sélectionne les lettres avec une valeur numérique plus faible que les lettres racine. Ils ont classé les lettres arabes en deux groupes: le premier groupe contient les lettres qui ne figurent dans aucun affixe. Ensuite, ils ont attribué une valeur numérique à ce groupe. Le second contient des lettres qui apparaissent dans les affixes, et ils ont attribué des différentes valeurs numériques à ces lettres.

- **(Sakhr; 2004)** : Ce logiciel utilise l'analyse morphologique pour identifier toutes les formes possibles du radical d'un mot: la racine, le schème morphologique et les affixes. Il fonctionne dans un mode inverse pour générer les différentes formes morphologiques (radical, racine, schèmes morphologique, partie du discours et/ou affixes) à partir d'un mot. L'analyseur morphologique utilisé par Sakhr couvre une grande partie des mots arabes classiques et modernes. Malheureusement, il n'existe pas une version ouverte pour son utilisation.
- **(Nizar Habash and Owen Rambow, 2006):** Ces auteurs ont développé un analyseur et génération morphologique de dialectes arabes, cet analyseur morphologique est ouvert pour la recherche. C'est un système qui forme des modèles basés sur la morphologie (M. Altantawy et al, 2010). Pour développer MAGEAD[(12)], ils utilisent une représentation morphologique pour tous les morphèmes, et définissent explicitement les règles morphologiques et orthographiques pour dériver les allomorphes (racines). Le lexique est développé en prolongeant celui de Elixir-FM. L'avantage de cet analyseur est qu'il traite des mots de la morphologie des dialectes. Mais, malheureusement, cet analyseur a besoin d'un lexique complet des dialectes, pour rendre l'évaluation plus intéressante et convaincante et pour vérifier ces allégations.
- **(Sabri et yousfi, 2006) :** Ils ont élaboré un Système d'analyse morphologique en utilisant un dictionnaire des schèmes de surface du mot. Cette approche est nécessaire dans la phase initiale d'un classement de tous schèmes de surface par le nombre de caractères (chaque classe d'un schème de surface contient un nombre d'alphabet arabe). La deuxième phase est la construction de la base de données des schèmes de surface. Ce système est capable d'analyser tous les verbes arabes.
- **(Soudi, 2001, 2007) :** C'est un analyseur morphologique arabe qui repose sur la morphologie des lexèmes où la racine du mot est l'information cruciale qui doit être extraite de mot. Soudi (2007), il constitue une version développée du premier, puisqu'il prend en considération le radical des mots, la grammaire et les spécifications lexicales.(Soudi, Cavalli-Sforza et Jamari 2001; Soudi, Bosch et Neumann, 2007).

[(12)] Morphological Analyzer and Generator for the Arabic Dialects

- **(Elixir-FM, 2007):** Cet analyseur morphologique arabe est en ligne. Il traite les mots de l'arabe moderne. Il est développé par (Otakar Smrz, 2009). Il est bien documenté et disponible pour l'évaluation. Elixir-FM s'inspire de la méthodologie de la morphologie fonctionnelle (Forsberg et Ranta, 2004) et invoque initialement le lexique de Buckwalter re-transformé. Il contient deux composantes principales: une bibliothèque de programmation multi-usages et un lexique linguistique morphologique. L'avantage de cet analyseur est qu'il donne à l'utilisateur quatre différents modes de fonctionnement (Resolve, infléchir, Derive et Lookup) pour analyser un mot ou un texte en arabe. Mais le système a une portée limitée car il n'analyse que les mots dans la langue arabe moderne.
- **(Hamada, Salwa, 2009; 2010):** Il a défini l'analyse morphologique de texte arabe comme une série de processus. L'analyse morphologique pour le texte arabe consiste à extraire la racine du mot à analyser tous les dérivés possibles de cette racine.
- **(AlKhalil, 2010) :** Alkhalil est un analyseur morphologique pour le texte arabe standard. Cet analyseur permet d'analyser trois types de texte: non voyellé, partiellement voyellé et texte entièrement voyellé MSA[(13)]. Il est basé sur la modélisation d'un très grand nombre de règles morphologiques arabes, sur l'intégration des ressources linguistiques qui sont utiles à l'analyse tels que : la base de données des racines voyellées, les schèmes morphophonémiques associés aux racines, les listes enclitiques et proclitiques. Les résultats de l'analyse des mots arabes sont présentés dans un tableau qui montre le radical totalement voyellé; sa catégorie grammaticale, les caractéristiques morphosyntaxiques des mots en langage naturel; ses racines possibles associées aux schèmes correspondants; ses proclitiques enclitiques (Boudlal et al. 2010). Les listes des schèmes de noms et des schèmes de verbes obtenues en utilisant le système d'ALECSO Sarf (Système Morphologie Arabe) (ALECSO 2008b) et le corpus NEMLAR (Attia et al., 2005). Ces listes sont de taille très large, 28.000 schèmes totalement voyellés. Alkhalil contient environ 7000 racines obtenues à partir de Sarf où chaque racine est liée aux schèmes de dérivation utilisés pour obtenir les mots de cette racine (Mazroui et al 2009;.. Boudlal et al 2011).

[(13)] Modren Standard Arabic

- **(Sadik Bessou et Mohamed Touahria; 2011) :** Ils ont développé un système d'analyse de génération morphologique pour la traduction automatique, depuis et vers l'arabe. Il présente une traduction automatique qui nécessite un traitement linguistique, puis présente une méthode d'analyse et de génération sur la base de caractéristiques linguistiques des mots arabes. Il traite les schèmes et extrait des informations morphologiques. Cette information est très utile dans la génération des dérivés du mot et pour le transfert structurel.

Chapitre III : Analyse morphologique à base de schèmes de surface

I- Introduction:

En raison de sa complexité, la morphologie de la langue arabe est devenue une partie intégrante de nombreux systèmes d'analyse morphologique arabe. Dans le domaine de traitement automatique de la langue arabe et pour l'analyse d'un mot, l'analyseur morphologique à base de schème détermine un ou plusieurs schèmes possibles afin de retrouver la racine. Ces schèmes sont nominaux ou verbaux, et pour détecter la racine d'un mot, il faut connaître les schèmes solutions. Différents travaux utilisent l'approche racine-schèmes pour trouver la racine (ou les racines possibles) d'un mot analysé, des algorithmes de combinaison automatique entre racines et schèmes. Parmi ces travaux celle de (Darwish, 2003) et (Khoja, 1999), (Al-Fedaghi et Al-Anzi, 1989), (Hegazi.N, ElSharkawi.A., 1986), (Xerox, 1998), (Sabri et yousfi, 2006), (AlKhalil, 2010), (Sadik Bessou et Mohamed Touahria; 2011):

Ce chapitre sera consacré à présenter l'approche racine-schème et montrer l'opportunité de l'utilisation des schèmes de surface par rapport aux schèmes classiques, les schèmes de surface

apportent à notre analyseur de grandes possibilités pour trouver les racines possibles d'un mot.

Notre travail vise dans un premier lieu à traiter les mots dérivés arabes, en ce basant principalement sur la construction de la base des données des schèmes de surface des mots traités. Ensuite, on adoptera un travail antérieur de (Yousfi, 2010) (Sabri et Yousfi, 2006) pour l'analyse des verbes arabes afin de traiter tous les mots dérivés arabes (les noms dérivés plus les verbes).

Malgré les travaux réalisés dans le domaine de l'analyse morphologique, on remarque que ces derniers présentent toujours des limites, nous citons :

- Le volume des dictionnaires utilisés qui sont de taille très grande.
- le taux de couverture de ces dictionnaires est très restreint et il ne couvre pas la totalité des mots dérivés arabes.
- Ces dictionnaires des mots contiennent une sorte de redondance représentée par des mots ayant les mêmes règles morphologiques "دخل–داخل ، جمع–جامع".
- De même, les systèmes d'analyse existants utilisent plusieurs règles au moment de l'analyse morphologique (les règles de transformations morphologique pour les racines non régulières).

Pour remédier à ces inconvénients, nous avons développé un analyseur morphologique indépendant du dictionnaire des mots et de l'utilisation des règles au moment de l'analyse morphologique, et qui utilise uniquement les schèmes de surface des mots arabes.

II- Construction de la base des schèmes de surface des mots dérivés

Les seules catégories des mots arabes qui sont régies par des schèmes sont les mots dérivés (les verbes et les noms dérivés).

II-1 Noms dérivés arabes

Les noms dérivés sont les noms qui peuvent être dérivés à partir d'une racine verbale. Le nombre et la nature de ces formes varient selon le statut du verbe auquel ils se rattachent. Parmi les noms dérivés qui existent, nous citons : le participe actif, le participe passif, etc.

Nature et nombre du pronom	Type de noms	Dérivation de la racine	Racine
مثنى-مذكر	اسم-الفاعل	ضاربان	ض ر ب
جمع-مذكر	اسم-الفاعل	قائلون	ق و ل
مثنى-مذكر	اسم-المفعول	موليان	ولي
جمع-مؤنث	اسم-المفعول	مضروبات	ضرب
مفرد-مذكر	التصغير	وويق	وقى
مفرد-مذكر	اسم-الألة	مرمى	رمى
مفرد-مذكر	المصدر-الميمي	مرعى	رعى
مفرد-مذكر	اسم-التفضيل	أخوف	خاف

Tableau n° 3 : Un exemple des mots dérivés en fonctions de leurs racines et leurs pronoms

II-2 Les verbes :

Les verbes arabes sont des mots qui se conjuguent aux différents temps, et qui indiquent un état ou une action faite ou subie par le sujet: فعل ماض معلوم، فعل ماض , (singulier, duel ou pluriel), la voix (active et passive) et le mode (parfait, imparfait, subjonctif, apocopé et impératif).

II-3 Schème de surface

Le schème d'un mot permet de détecter les lettres radicales constituant sa racine ainsi que les suffixes et les préfixes liés à ces lettres. Le schème de 'مُكْرِمُونَ' est 'مُفْعِلُونَ', les lettres " ف،ع،ل" remplacent les lettres de la racine de "ك ر م", et le schème de "صَارَعَ " est " فَاعَلَ " (Sam.A et Youssef , 1999 ; Bahrak, 869-930; Hanafi, 1914 ; Zanjani, 1343).

L'utilisation des schèmes de surface facilite l'identification des verbes défectueux et exactement la nature et l'emplacement des caractères défectueux dans le racine. Ce type de schèmes classiques ne présente pas les variations morphologiques du mot, comme par exemple le mot قَائِلٌ du verbe قَالَ, c'est pourquoi nous avons utilisé le schème de surface (Yousfi, 2006, 2010). Ce type de schème permet de présenter les variations morphologiques, par exemple pour le mot "قائل" , on a une variation de "ا" à "أ" c'est pourquoi le schème de surface est "فائل" au lieu de "فاعل" "أ" a remplacé le "ع".

La méthode de construction de ce type de schème est la suivante: si on suppose que le mot dont on cherche son schème est $w = l_1 l_2 ... l_n$ (l_i Caractère du mot) et R sa racine.

Le schème de surface de w est $m = f_1 f_2 \ldots . . f_n$ avec :

Et le schème de surface de la racine $R = g_1 g_2 \ldots \ldots g_k$ ($g_i$est un caractère de R) est

$$P = f'_1 f'_2 \ldots f'_k$$

avec :

$$\begin{cases} f'_i = l'une\ des\ trois\ lettre\ \text{ل،ع،ف}\ si\ g_i\ est\ une\ lettre\ constante \\ \qquad\qquad au\ moment\ de\ la\ transformation\ de\ R \\ f'_i = g_i \qquad\qquad sinon \end{cases}$$

Exemple:

Le schème de surface de la racine "رَعَى " est "فَعَى ", et "فَاعٍ " est le schème de surface de "راعٍ". Le schème de surface de "آجِرٌ " est "آفِعٌ " et de "آجِرَاتٌ " est "آفِعَاتٌ".

II-4 Construction de la base des schèmes de surface

Pour la construction de la base des schèmes de surface des mots dérivés arabes, nous avons traité 127 racines qui représentent presque toutes les classes possibles[14] pour générer les mots dérivés arabes adoptées par Youssef Dishet (Youssef , 1999). Des linguistes ont généré tous les mots dérivés arabes à partir de ces 127 racines. Ensuite, ils les ont conjugués aux différentes personnes (masculin singulier, masculin duel, masculin pluriel, féminin singulier, féminin duel, féminin pluriel), et à partir de ces mots, ils ont dégagé les schèmes de surface de chaque mot dérivé.

A la fin, nous avons obtenu plus de 6216 schèmes de surface qui représentent presque tous les mots dérivés arabes.

[14] Chaque classe de racines se conjuguent de la même façon.

Racine	Genre et nombre	Type de noms	schèmes de surface des mots dérivés
فاء	مثنى – مؤنث	اسم-الفاعل	فائيتان
فعل	جمع – مؤنث	اسم-الفاعل	فاعل
فاء	مفرد-مذكر	اسم-المفعول	مفيء
فاء	مثنى-مذكر	اسم-المفعول	مفيئان
فاء	جمع –مذكر	اسم-المفعول	مفيؤون
فاء	مفرد – مؤنث	اسم-المفعول	مفيئة
فاء	مثنى – مؤنث	اسم-المفعول	مفيئتان
فاء	جمع – مؤنث	اسم-المفعول	مفيئات
فاء	مفرد-مؤنث	المصدر	مفيئة
فاء	مفرد-مذكر	المصدر	فيئا

Tableau n° 4 : Un exemple des schèmes de surface en fonction de leurs racines et leurs pronoms

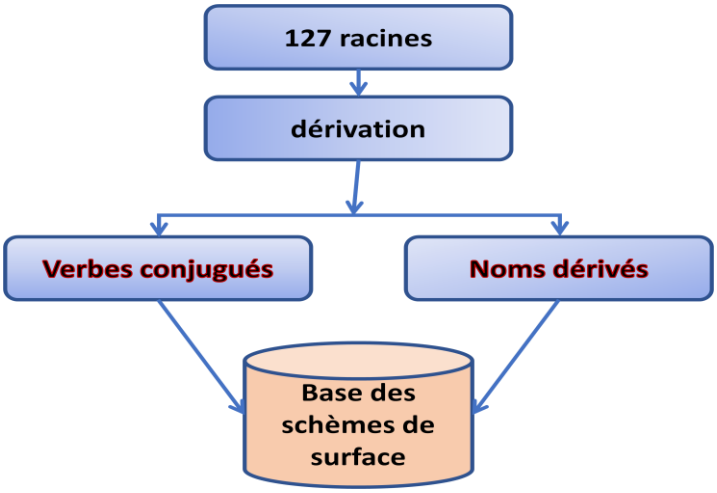

Figure n° 3 : Schéma de la construction de la base des schèmes de surface

III- L'approche utilisée dans notre analyseur morphologique

Dans l'approche déjà utilisée par Sabri et Yousfi en 2006, on a remarqué que pour la construction de la base des schèmes de surface des verbes, il fallait ajouter une étape de concaténation de tous les suffixes et les préfixes possibles avec les schèmes de surfaces des verbes conjugués. Ceci rend la taille de la base des données des schèmes assez grande.

Dans notre cas, nous avons ignoré cette étape et nous avons intégré dans le système une phase de

segmentation du mot en suffixe et préfixe avant de trouver le schème de surface de ce mot. Ensuite, nous avons cherché les schèmes du mot restant dans l'ensemble des schèmes de surface ayant la même longueur.

Prenons comme exemple le mot "فواقيانهم" après l'extraction du préfixe "ف" et du suffixe "هم" on trouve " واقيان", donc le schème de surface est "واعيان".

De même pour ce travail, nous avons pu formuler la fonction qui mesure la similarité entre le mot à analyser et les schèmes de surface. Cette fonction a été formulée comme suite :

$$f(m; w) = \sum_{i=1}^{N} 1_{[m_i ; w_i]}$$

avec : $1_{[m_i,w_i]} = \begin{cases} 1 \;\; si\; m_i = w_i \;\; ou\; (m_i = ف, ou\; ع, ou\; ل) \\ f(m,w) = 0\; sinon\; et\; on\; sort\; de\; l'algorithme \end{cases}$

m_i: $i^{ème}$ $Caractère\ de\ schème\ m$ et w_i: $i^{ème}$ $Caractère\ de\ schème\ w$

La fonction f dégage un ensemble de solutions de schèmes de surface, que nous notons par S :

$$S = \{m \in P_{L(w)} \;\; / \;\; f(m,w) \neq 0\}$$

$P_{L(w)}$: L'ensemble de tous les schèmes de surface de longueur$L(w)$.

$L(w)$: La longueur du mot$w$.

Exemple : pour le mot "واقيان", la fonction qui mesure la similarité entre le mot à analyser et les schèmes de surface, prend les valeurs suivantes :

f ('واقيان ' ; 'وافيان ') = 6

f ('واقيان ' ; 'فاعيان ') = 6

f ('واقيان ' ; 'متفاعى ') = 0

Ensuite, pour chaque schème de surface m du mot w nous cherchons ces racines R . Pour trouver

les racines du mot à analyser W, nous cherchons, dans un premier temps, les positions des caractères"ف", "ع", "ل" dans les schèmes de surface du mot W et nous dégageons les caractères associés à ces positions dans le mot W, ces caractères sont remplacés, ensuite, dans les schèmes de surface de la racine dans les mêmes positions.

Par exemple, pour le mot قائلون on trouve les schèmes de surface suivants :

- "فاعلون" avec le schème de surface فعل pour sa racine.
- "فائلون" avec le schème de surface فال pour sa racine.

Après l'application de notre méthode, nous trouvons les deux solutions suivantes :

قائِلون ⟵ فاعِلونَ　　فعل ⟵ قَئَلَ

قائِلُون ⟵ فائِلون　　فَالَ ⟵ قال

Comme le radical قَئَلَ n'existe pas dans la liste des racines, on garde donc seulement la deuxième solution قَالَ.

VI- La mise en œuvre :

VI.1 Architecture de notre analyseur:

Pour tester notre approche, nous avons d'abord construit 6216 schèmes de surface des mots dérivés arabes. Cette étape a été réalisée par des linguistes qui ont utilisé un ensemble de références arabes (Mustapha, 1999 ; Bahrak, 869-930; Hanafi et al., 1914 ; Zanjani, 1343, Youssef, 1999).

Pour la mise en œuvre de notre approche, nous avons développé un programme en java constitué des parties suivantes (voir la figure 4).

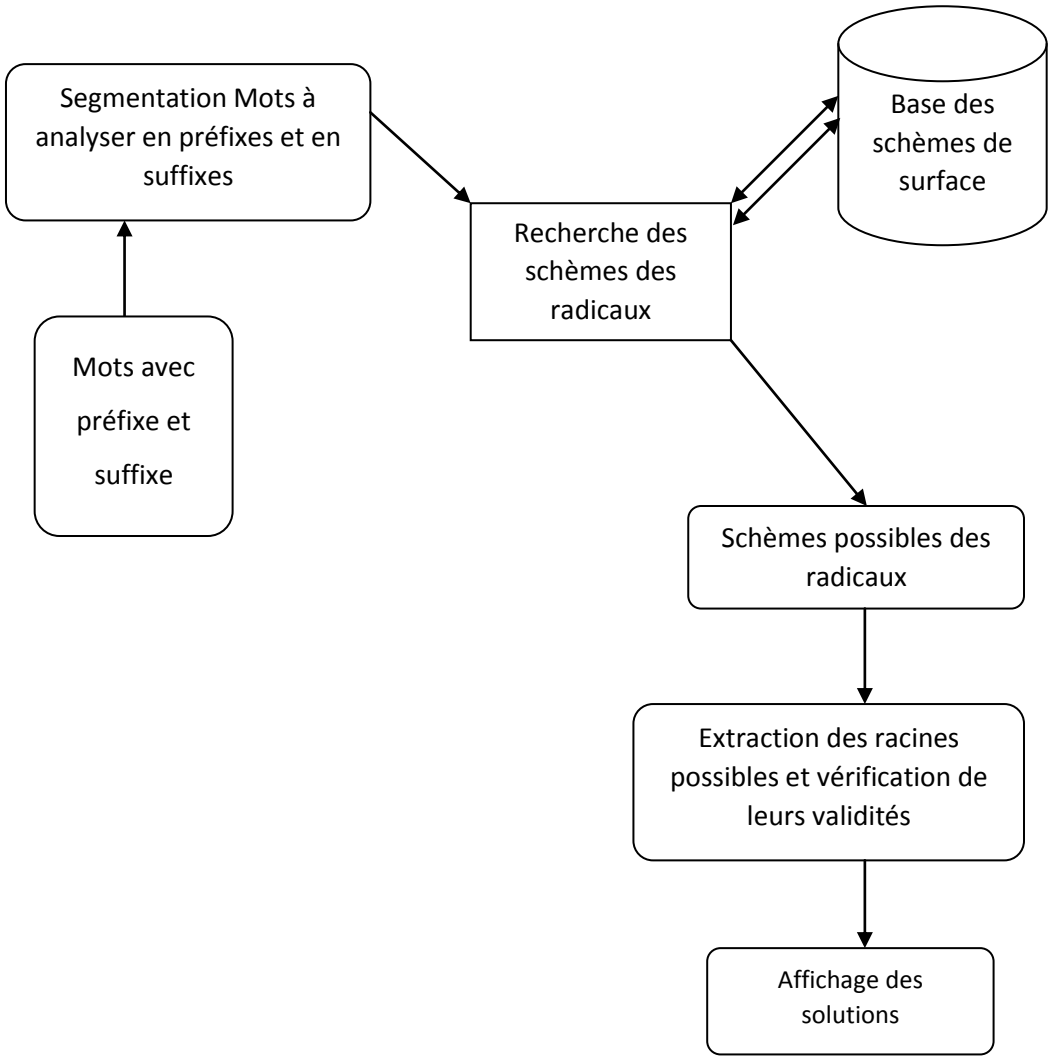

Figure n°4 : Les étapes de notre analyseur morphologique des mots dérivés arabe.

- Partie 1 : la segmentation du mot à analyser en suffixes, préfixes et radical.
- Partie 2 : la recherche des schèmes de surface pour les radicaux retournés par la partie 1.
- Partie 3 : la recherche des racines à partir de tous les schèmes de surfaces retenus par la partie 2.
- Partie 4 : la vérification de la validité de ces racines en cherchant s'ils existent dans la base des racines ou non.

VI.2 Corpus d'évaluation de notre analyseur morphologique

Pour évaluer notre analyseur morphologique, nous avons préparé manuellement un corpus de référence en collaborations avec un groupe de linguistiques de la langue arabe. Ce corpus est constitué de 10000 entrées, chaque entrée contient dans la première colonne le mot dérivé, ensuite le radical, la racine, le type de mot, genre et nombre des pronoms affixe, le préfixe et le suffixe du mot (voir la tableau n° 5).

Mots dérivé	Mot après sup des affixes (Radical)	Radical	Type de mot	Nature et Nombre du Pronom	Proclitique+ Préfixe	Suffixe+ enclitique
بذكرا	ذكرا	ذكر	الماضي-المعلوم	هما	ب	.
مذكورهم	مذكور	ذكر	اسم-المفعول	مفرد-مذكر	.	هم
كمذكرهم	مذكر	ذكر	اسم-الزمان	مفرد-مذكر	ك	هم
فمذكركم	مذكر	ذكر	اسم-الفاعل	مفرد-مذكر	ف	كم
ولمزيدهم	مزيد	زاد	اسم-المفعول	مفرد-مذكر	ول	هم
فمزيدهن	مزيد	زاد	اسم-المفعول	مفرد-مذكر	ف	هن
بمتفقدكم	متفقد	تفقد	اسم-الفاعل	مفرد-مذكر	ب	كم
كواقية	واقية	وقى	اسم-الفاعل	مفرد-مؤنت	ك	.
فموقى	موقى	وقى	اسم-المكان	مفرد-مذكر	ف	.
الأوقى	أوقى	وقى	المضارع-المنصوب-المجهول	أنا	ال	.
كالزائد	زائد	زاد	اسم-الفاعل	مفرد-مذكر	كال	.
فتفقدا	تفقدا	تفقد	الماضي-المجهول	هما	ف	.
متفقدكما	متفقد	تفقد	اسم-المفعول	مفرد-مذكر	.	كما
فوال	وال	ولى	اسم-الفاعل	مفرد-مذكر	ف	.
كالولاة	ولاة	ولى	اسم-الفاعل	جمع-مذكر	كال	.

Tableau n°5 : Un extrait du corpus de référence.

VI.3 Test et résultats:

Pour évaluer notre analyseur, nous avons fait notre test sur un corpus de 10000 mots dérivés (verbes et noms dérivés) représentants presque toutes les catégories des racines (127 classes de racines) déjà cité. Ensuite, nous avons comparé les résultats de notre analyseur avec ceux de Buckwalter et d'Al-Khalil[15], en utilisant le rappel et la précision.

(15) : Les résultants retournée par les deux analyseurs de Buckwalter et Al-Kalil, sont retournées par le système d'évaluation développés par Bouzabaa et Jaafar sur le site sibawayh : www.sibawyh.net

Nature	Type Mot	Schèmes mots(w)	Schèmes radical	Radical	Mots (w)	Mots
مفرد-مذكر	الأمر	افع	فعى	فأى	افئ	وافئكما
مثنى-مؤنت	اسم الفاعل	وافيتان	وفى	وقى	واقيتان	كواقيتانهم
مثنى-مؤنت	اسم الفاعل	فاعيتان	فعى	وقى	واقيتان	كواقيتانهم
جمع-مؤنت	اسم المفعول	مفوءات	فاء	ساء	مسوءات	بمسوءاتكم
جمع-مؤنت	اسم-المفعول	مفوعات	فاع	ساء	مسوءات	بمسوءاتكم
جمع-مؤنت	اسم-المفعول	منفاعات	انفاع	انصاع	منصاعات	بمنصاعاتهم
مفرد-مذكر	اسم-الفاعل	منفاع	انفاع	انصاع	منصاع	ولمنصاعهم

Tableau n° 6 : Exemple de quelques analyses morphologiques des mots retournées par notre système

Les résultats obtenus en termes de précision et rappel (Annexe I), sont mentionnés dans le tableau suivant :

Nom d'analyseur	Rappel	Précision	Temps d'exécution/mot
Analyseur Buckwalter	3.21%	26.93%	2.11 ms
Analyseur Al-Khalil	30.2%	40.22%	31.7 ms
Analyseur à base de schèmes	94,20%	95,97%	24 ms

Tableau n° 7 : Précision et rappel de notre analyseur, de Buckwalter, et d'Al-khalil

Le corpus de référence est constitué seulement des mots dérivés (les verbes et les noms dérivés). Notre analyseur retourne un rappel de 94.20% contre 3.21% et 30.2% pour les analyseurs de Buckwalter et d'Al-khalil[16]. Ceci montre que le nombre de solutions oubliées pour les deux derniers analyseurs est plus important par rapport à notre analyseur. Ceci est dû à ce que notre système utilise les schèmes de surface qui sont, d'une part, de nombre très réduit et, d'autre part qu'il regroupe et modélise presque toutes les variations morphologiques de tous les mots dérivés.

La même remarque se fait pour la précision qui est très élevée par rapport à la précision des deux autres analyseurs. Ceci montre que notre analyseur retourne plus d'analyses justes 95.97% par rapport aux deux autres analyseurs de Buckwalter 26.93%, et 40.22% et d'Al-Khalil.

(16)

http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=6847312&url=http%3A%2F%2Fieeexplore.ieee.org%2Fiel7%2F6839546%2F6846554%2F06847312.pdf%3Farnumber%3D6847312

Conclusion :

Notre contribution dans ce chapitre a été le traitement de tous les mots dérivés arabes. Ensuite, nous avons formulé la fonction qui mesure la similarité entre les mots et les schèmes de surface. De même nous avons pu réduire la taille de la base des schèmes en éliminant la phase d'ajout des préfixes et des suffixes aux schèmes de surface.

Comme conclusion, on peut dire que notre analyseur, en se basant sur le corpus de référence déjà cité, est plus performant (Précision, Rappel) pour l'analyse des mots dérivés, et ceci grâce d'une part à ce que notre système reconnait presque tous les mots dérivés arabes, et d'autre part il retourne presque toutes les analyses justes.

Chapitre IV : Analyse morphologique à base de graphes et tables de Compatibilité entre affixes.

I- Introduction:

Dans le traitement automatique des langues, plusieurs méthodes sont utilisées dans la construction des analyseurs morphologiques. Parmi ces méthodes, nous citons celle à base d'automates à état finis. L'ensemble des mots arabes est organisé sous la forme d'un automate à états finis où chaque arc correspond à une lettre, et chaque mot est représenté par un graphe, et l'ensemble des chemins possibles pour tous les mots, représente le réseau global.

Chaque analyse est présentée par un chemin, et chaque chemin possible est défini par un état initial et un état final. Entre ces deux derniers, on trouve les états des préfixes, les états des racines, les états des infixes et les états des suffixes. Parmi les chercheurs qui ont adopté cette méthode, nous citons (Wehrli, 1997); (Bouillon, 1998).

II- Analyseur à base d'automate à états finis :

Pour l'analyse morphologique, les automates à états finis sont utilisés pour représenter les relations morphologiques dans une langue donnée (Wehrli, 1997). Un automate à états finis a pour objectif de définir les différentes combinaisons possibles entre les alphabets d'un lexique donné.

L'approche à états finis pour l'analyse morphologique automatique a été étudiée chez Xerox. La première application pratique était établie par Koskenniemi (Koskenniemi 1983), et elle a été utilisée après, pour développer des analyseurs morphologiques pour plusieurs langues dont l'anglais, le français, l'Allemand, l'arabe, l'espagnol et le japonais (Beesley et Karttunen , 2003).

L'invention de l'approche de la morphologie à deux niveaux en 1983, par Koskenniemi, a été une percée majeure en linguistique computationnelle. Cela a conduit à l'utilisation et à la généralisation des automates à états finis.

Un automate à états finis ou bien machine à états finis est défini par les états et leurs transitions entre les états. Les automates à états finis et leurs propriétés sont bien connus comme des objets mathématiques. Cependant, leur utilisation pour le traitement du langage naturel a réellement été explorée depuis les années 80. Les automates à états finis utilisent au cours de l'analyse morphologique des mots, la notion des préfixes, des racines, des suffixes et des infixes de la langue arabe.

Automates à états finis / réseau à états finis: un réseau composé d'Etats, y compris un état de départ et un ou plusieurs états finals. Les transitions entre états ne sont possibles que si l'entrée requise est reconnue. Un chemin est une séquence de transitions sur des arcs à un état particulier.

Les automates à états finis sont définis par les éléments suivants :

- L'ensemble d'états finis: $Q = \{q_I, q_1, q_2, \dots, q_n, q_F\}$
- Une fonction de transition notée δ, et qui a pour paramètres : un état et un symbole qui renvoie un état: $\delta : Q \times \sum \rightarrow Q$
- Un ou plusieurs états finaux $q_F \epsilon Q$

Plusieurs auteurs ont utilisé les automates à états finis pour faire l'analyse morphologique arabe tels que : (Audebert et Jaccarini, 1988), (Beesly, 1998), (Beesly, 1996), (Gaubert, 1996), (Jaccarini, 1997).

III- Inconvénients des approches pour l'analyse morphologique

Parmi les inconvénients des approches déjà citées, nous citons :

- L'approche à base d'automate à état fini, utilise tous les mots du lexique pour former le réseau global (Wehrli, 1997). Elle ne prend pas en compte les règles de compatibilité entre les morphèmes (infixe, préfixe, suffixe, radical).
- Pour l'approche à base des dictionnaires, il y a la difficulté de la construction des différents dictionnaires des radicaux. Cette construction nécessite la connaissance des différentes transformations morphologiques de chaque racine. de plus, ce dictionnaire est assez large. car pour chaque racine on trouve au minimum cinq radicaux dans ce dictionnaire, par exemple pour la racine "كتب", dans le dictionnaire des radicaux de Buckwalter on trouve : "كتب" "كاتب" "كتوب" "كتيب" "كتاب".

Pour remédier à ces inconvénients, nous proposons une nouvelle approche qui utilise les graphes et un dictionnaire restreint des radicaux pour faire l'analyse morphologique d'un mot. Nous avons appelé cet analyseur : l'analyseur morphologique IBN-GINNY17. Ce système combine l'approche de Buckwalter et les techniques utilisées dans les graphes pour faire l'analyse morphologique d'un mot donné. De même, nous avons adapté l'algorithme de Viterbi, pour chercher toutes les analyses possibles d'un mot, dans un réseau global de graphes formé seulement à partir des affixes de la langue arabe. L'approche a été testée sur le même corpus déjà utilisé dans le chapitre précédent (4000 mots dérivés).

IV- Présentation de l'analyseur IBN-GINNY

L'analyseur IBN-GINNY [16] est un analyseur morphologique à base de graphe, où chaque mot dans la langue arabe est représenté par un chemin dans ce graphe. Pour faire l'analyse d'un mot, l'analyseur IBN-GINNY passe par les deux étapes suivantes :

[17] Nous avons nommé notre analyseur en l'honneur du grand linguistique Arabe IBN-GINNY.

- Il construit d'abord un réseau global.
- Il cherche les solutions possibles de l'analyse de ce mot.

IV -1 construction de réseau global des mots arabes

Pour diminuer la taille du dictionnaire des radicaux (stems) dans l'analyseur de (Buckwalter, 2002), on va utiliser les graphes pour regrouper tous les préfixes, les suffixes et les infixes de la langue arabe. Ensuite la recherche de l'analyse morphologique d'un mot se fait dans ce graphe.

Dans ce cas, chaque racine va être présentée par un seul radical, comme par exemple "وعد" est présenté seulement par ce verbe lui même :

وعد ⟶ وعد

Pour mettre en œuvre notre idée, nous nous sommes basés sur les graphes où chaque mot est modélisé par un chemin d'états. Les lettres racines de ce dernier sont présentées par un état qui boucle sur lui-même, et les affixes sont présentés par les caractères formant ces affixes.

Exemple : Les mots 'فداخلها' , 'فجامعها'…, sont présentés par le seul chemin suivant :

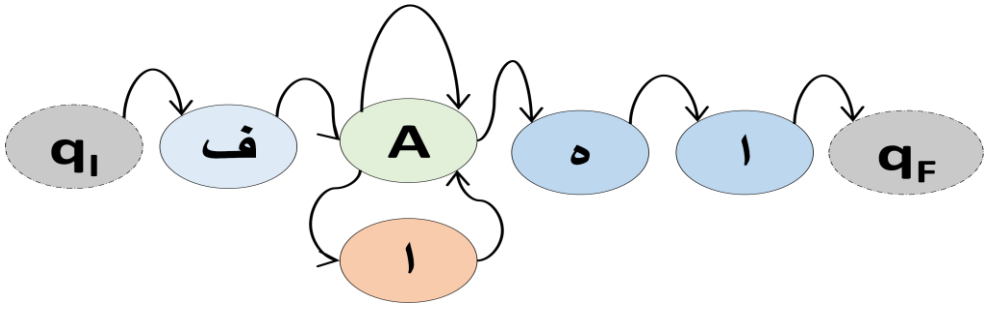

Figure n° 5 : graphe du mot 'فجامعها' et 'فداخلها'.

Un chemin du mot "فجامعها", est constitué des éléments suivants:

- q_I l'état début.
- q_F l'état final.
- "ف" un caractère du préfixe du mot "فجامعها ".

- "ا" infixe du mot "فجامعها ".
- "ا","ه" deux caractères du suffixe "ها" du mot "فجامعها ".
- A = "ج", "م", "ع" représentent les lettres qui composent la racine du mot جمع, dans ce cas, on boucle trois fois sur l'état A.

IV-1 Réseau global de notre analyseur:

En se basant sur tous les préfixes, les infixes et les suffixes de la langue Arabe, on construit un réseau global d'états avec un seul état d'entrée q_I et un seul état de sortie q_F.

Notre réseau global est défini entièrement par :

- L'ensemble de tous les états, il est constitué de tous les caractères composant les affixes (suffixes, préfixes et infixes), de l'état A (représentant les lettres radicales d'un mot donné), l'état initial qI et l'état final qF :

Q={qI, qF, A,"ف","و","ي","ل",…,"ه","م","ت",…}

- L'ensemble de toutes les transitions possibles reliant les caractères des suffixes, des préfixes et des infixes avec les états A, qI et qF.

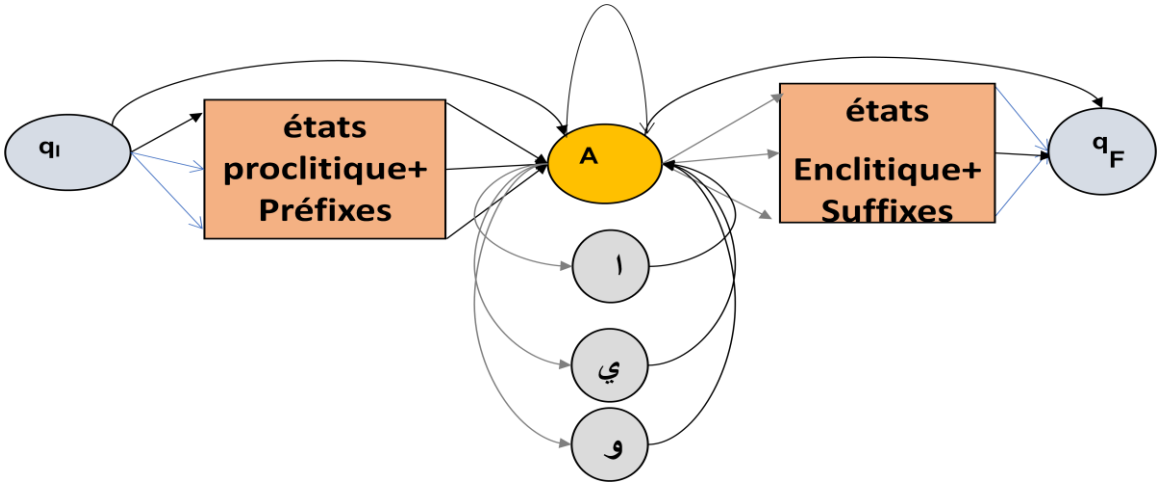

Figure n° 6 : Schéma du réseau global dans l'analyseur IBN-GINNY.

IV-2 Analyse morphologique à base du réseau global:

Pour analyser un mot donné w, on cherche dans le réseau global les différents chemins possibles associés à w.

L'ensemble de ces chemins est donné par :

$$S = \left\{ \xi \ \epsilon B / P(w / \xi) \neq 0 \right\} \quad \textbf{(1)}$$

B: l'ensemble de tous les chemins possibles dans le réseau global.

ξ : Un chemin possible de même longueur que *w*, dans le réseau global, et qui peut présenter le mot *w*.

Les solutions sont tous les chemins qui permettent d'émettre le mot *w* avec une probabilité non nulle. Pour faciliter et diminuer le calcul dans la formule (1), on utilise l'algorithme de Viterbi. Cet algorithme est donné par la formule suivante:

$$\delta_t(c_j) = \underset{c_i \to c_j}{\mathrm{NL}} (\delta_{t-1}(c_i) \cdot a_{ij} \cdot 1_{c_j}(w_t))$$

NL(x) est la fonction qui donne les valeurs non nulle de x.

On cherche les états c_i qui donnent des valeurs non nulles de $\delta_{t-1}(c_i) \cdot a_{ij} \cdot 1_{c_j}(w_t)$.

$\delta_T(q_F)$: est la probabilité non nul d'émission du mot *w* à partir d'un chemin donné. Par un calcul récursif, on récupère tous les chemins possibles qui donnent les valeurs non nulles de la fonction NL (T la longueur du mot *w*).

Avec :

c_j: Le j$^{\text{ème}}$ état ($c_j \in Q$).

a_{ij}: La probabilité de transition de l'état c_i vers l'étatc_j.

$$a_{ij} = \begin{cases} 1 & si\ la\ transition\ est\ possible \\ 0 & sinon \end{cases}$$

w_t: t[ème] caractère du mot w.

$$1_{c_j}(w_t) = \begin{cases} 1\ si\ c_j = w_i\ ou\ c_j = A \\ 0 \qquad\qquad sinon \end{cases}$$

Dans ce cas : $1_{c_j}(w_t) = 1$

Initialisation :

$$\delta_0(c_i) = \begin{cases} 1\ si\ c_i = q_I \\ 0 \quad sinon \end{cases}$$

Exemple : Soit le mot w= "فنقول" la recherche des chemins associées à ce mot se fait selon les étapes suivantes:

$$\delta_0(q_i) = \begin{cases} 1\ si\ q_i = q_I \\ 0 \quad sinon \end{cases}$$

$$\delta_1(\text{ف}) = \underset{q_i \to \text{ف}}{NL}(\delta_0(q_i).\, a_{q_i\text{ف}}.\, 1_{\text{ف}}(w_1)) = 1$$

En effet : $\delta_0(q_I) = 1$ et $\delta_0(q_i) = 0 \quad q_i \neq q_I$

w_1="ف" : le premier caractère du mot "فنقول" $a_{q_I\text{ف}} = 1$, tous les états sont liés à l'état initial.

$1_{\text{ف}}(w_1) = 1$ car $w_1 =$ "ف" en suite $\delta_1(A) = \left(\delta_0(q_I).\, a_{q_{IA}}.\, 1_A(w_1)\right) = 1 \quad et\ 1_A(w_1) = 1$

On calcule de la même façon $\delta_t(c_i)$ pour t=2,….,5 et pour tous les états.

A la fin, on trouve les chemins suivants:

Suffixe	préfixe	chemins proposés
0	0	q_I A A A A A q_F
0	ف	q_I ف A A A A q_F
0	فن	q_I ف ن A A A q_F
0	فن	q_I ف ن A و A q_F
0	ف	q_I ف A A و A q_F

Tableau n° 8 : Chemins possibles associés au mot "فنقول".

L'analyseur IBN-GINNY est un système efficace et très puissant au niveau du temps d'exécution et du nombre des analyses morphologique retournées par ce dernier (il retourne toutes les

analyses possibles d'un mot), mais malheureusement il y a un inconvénient qui est due à la non prise en compte des règles de compatibilité entre les infixes et les racines, les suffixes et les préfixes. Si on prend par exemple le mot "يقولون", l'analyse morphologique par l'analyseur IBN-GINNY, retoure les solutions suivante :

- Préfixe "ي", racine "قال", suffixe "ون"
- Préfixe "ي", racine "قل", infixe "و", suffixe "ون"

Or la solution "قل" (قَلَّ) est fausse, cette erreur est due à la non prise en compte des règles de compatibilité entre l'infixe "و" et le préfixes "ي".

Pour remédier à ce problème, nous proposons dans ce chapitre de faire intégrer ces règles de compatibilité entre infixes-racines, suffixes, préfixes dans l'analyseur IBN-GINNY.

Pour les autres compatibilité entre préfixes-racines, et préfixes-suffixes, racines-infixes on utilise celle de Buckwalter[5].

V-Construction des tables de compatibilité entre affixes et les racines

La décomposition des mots en forme préfixes, suffixes et infixes est très importante dans l'analyse morphologique arabe.

Nous avons cherché, d'abord, tous les préfixes, suffixes et infixes de la langue arabe. Ensuite, nous avons travaillé sur un classement. Pour chaque préfixe on cherche les suffixes et les positions des infixes correspondants, et pour chaque infixe, on cherche la liste des suffixes correspondants.

Préfixes	infixes	suffixes
ا	AاAتA	ات
ت	AيAA	ي
ت	AAاA	ي
ت	AAيA	ية

Tableau n° 9 : Table de compatibilité entre préfixes-infixes-suffixes

Préfixes	Infixes
م	AAاA
م	AوAA
م	AيAاA
ا	AاAتA
ت	AAاA
مُسْتَـ	AاAA

Tableau n° 10 : Table de compatibilité entre préfixes-infixes

infixes	suffixes
AAA	ان
AائAA	ـة
AA|A	ية
AيAA	ـة
AA|A	ون

Tableau n° 11 : Table de compatibilité entre infixes-suffixes

Les aaa représentent les caractères de la racine d'un mot arabe.

Radicales	racine
قول، قل، قيل، قال ..	قال
صل، يصل، أصل،...	وصل

Tableau n° 12 : Table de compatibilité entre radicales-racines

VI- Tests et Résultats

Pour pouvoir comparer entre les résultats obtenus par notre ancien analyseur IBN-GINNY, nous avons gardé le même corpus test déjà utilisé pour évaluer cet analyseur. Le corpus est constitué de 4000 mots qui présentent les différentes variations morphologiques de la langue arabe.

Pour évaluer notre nouvelle approche, nous avons intégré les tables de compatibilité entre préfixe-infixe, infixe-suffixe et infixe-racine dans l'analyseur IBN-GINNY.

Pour la construction du dictionnaire des radicaux, on établit la différence entre deux types de racines: racines saines et racines défectives (معتلة), la présentation des racines saines dans notre dictionnaire se fait par les racines elles-mêmes, mais pour les verbes défectifs on garde les radicaux de Buckwalter [5].

La racine "قال" est présentée par les radicaux dans notre dictionnaire: (قيل، قول،قال). Mais pour la racine saine"دخل", elle est présentée seulement par "دخل".

Avec cette présentation, on a diminué la taille du dictionnaire des radicaux de 62.5% par rapport à celui de Buckwalter [5].

VI-1 Mise en oeuvre informatique de notre approche:

Pour mettre notre approche en œuvre, nous avons développé un programme en Java constitué des trois grandes classes :

- Classe **ConstReseau**: elle permet de construire le réseau global à partir des listes des préfixes, des suffixes et des infixes.

- Classe **AnalyMorpholo**: elle utilise l'algorithme de Viterbi pour chercher tous les chemins possibles associés à un mot donné dans le réseau global déjà construit par la classe précédente.

- Classe **VerfCompatib**: elle vérifie la validité des solutions proposées par la classe **AnalyMorpholo**, en vérifiant l'existence de la racine dans le dictionnaire des radicaux et la compatibilité entre les suffixes et les préfixes.

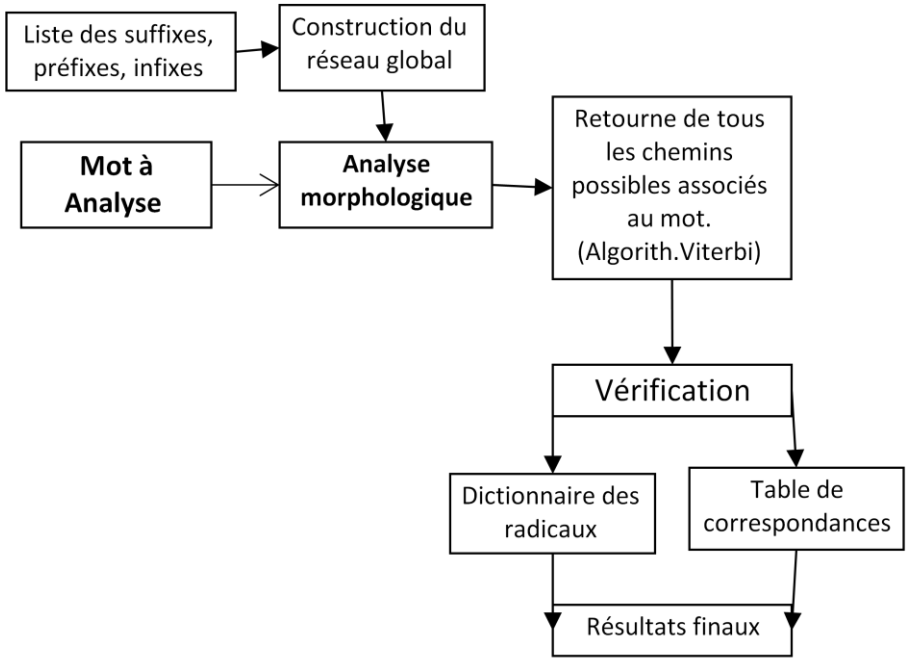

Figure n° 7 : Schéma du processus de l'analyseur morphologique IBN-GINNI.

Pour l'analyse du mot "فيسكتهم", notre système suit les étapes suivantes:

- La recherche des différents chemins dans notre réseau globale :

- L'extraction des racines associées à ces chemins en cherchant les positions des états "A" dans le mot "فيسكتهم" (voir la table 9).

- La vérification de l'existence des racines dans le dictionnaire des radicaux, cette étape garde seulement les solutions suivantes:

    - La solution "سك، في ، تهم", racine "سك" , le préfixe "في", le suffixes et "تهم"

    - La solution "سكت، في، هم", racine "سكت" , le préfixe "في", le suffixes et "هم"

Suffixe	Préfixe	Chemins possibles	Racines Proposées	N°
تهم	Φ	AAAAت ه م	فيسك	1
هم	Φ	AAAAAه م	فيسكت	2
Φ	Φ	AAAAAAA	فيسكتهم	3
تهم	ف	فـAAA ت ه م	يسك	4
هم	ف	فـAAAA ه م	يسكت	5
0	ف	فـAAAAAA	يسكتهم	6
تهم	في	ف يAA ت ه م	سك	7
هم	في	ف يAAA ه م	سكت	8
Φ	في	ف يAAAAA	سكتهم	9

Tableau n°13 : Chemins possibles du mot "فيسكتهم " dans le réseau global avec leurs racines proposées.

- La vérification de la compatibilité entre les suffixes et les préfixes des solutions restants, il garde seulement la solution "سكت، في، هم".

Car dans la première solution, le préfixe "في" n'est pas compatible avec le suffixe "تهم".

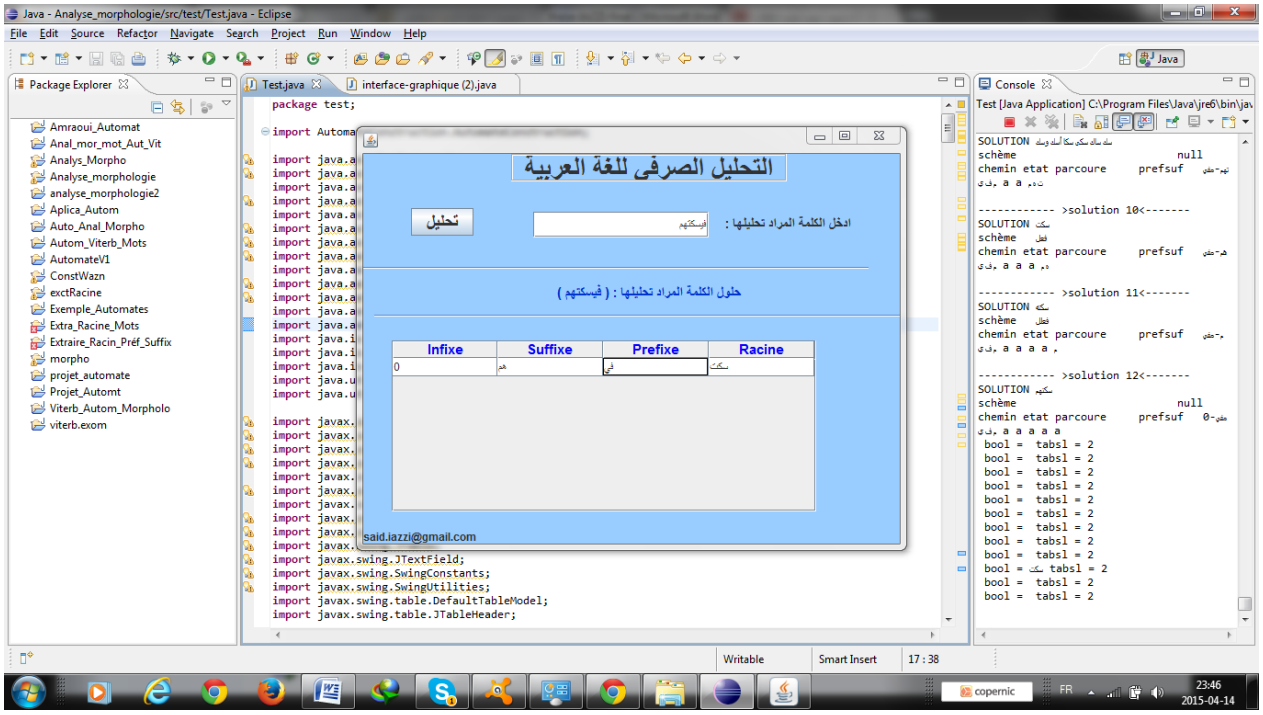

Figure n° 8 : Résultat de l'analyse le mot 'فيسكتهم'

VI-2 Mise en oeuvre de notre approche:

Pour évaluer notre nouvelle approche, nous avons d'abord créé les différents dictionnaires des suffixes, des préfixes et des radicaux. Ensuite, à partir de la liste des préfixes, des suffixes et des infixes, nous avons généré un réseau global d'états comme il a été signalé précédemment sans utilisation des dictionnaires lexicaux (ceci est le grand avantage de notre analyseur par rapport à l'analyseur de Buckwalter et l'analyseur à base d'automate à état fini) .

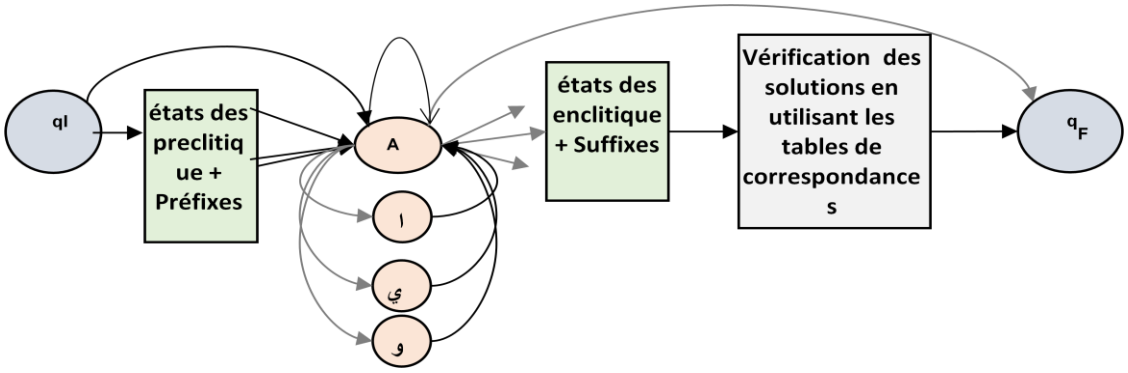

Figure n° 9 : Schéma du réseau global dans l'analyseur IBN-GINNY.

Pour la construction du dictionnaire des radicaux, on a fait la différence entre deux types de racines: racines saines et racines défectives (معتلة), la présentation des racines saines dans notre dictionnaire se fait par les racines elles-mêmes. Mais pour les verbes défectifs, on garde les radicaux de (Buckwalter, 2002).

La racine "قال" est présentée par les 5 radicaux suivants: قال، قل، قول، قيل، قائل. Mais pour la racine saine par exemple "كتب" elle est présentée seulement par "كتب".

Avec cette présentation, on a diminué la taille du dictionnaire des radicaux de 62.5% par rapport à celui de (Buckwalter, 2002).

De même on a construit une table d'compatibilité entre préfixe-suffixe et ceci pour éliminer les solutions impossibles.

Préfixes	**Suffixes**
ي	ت
ي	تما
ي	نا
ن	تم

Tableau n° 14 : Extrait de la table du non compatibilité Preff-Suff.

Conclusion :

Pour augmenter la précision et le rappel de la dernière approche, nous avons introduit pour la première fois la notion de compatibilité entre infixe-racine, infixe-préfixe et infixe-suffixe[18], et nous les avons introduit dans notre approche. Ceci a diminué d'une façon considérable la taille du dictionnaire des radicaux, et il a augmenté la précision de notre système.

Les résultats obtenus à l'aide de l'intégration de ces nouvelles tables de compatibilité sont très importants et montrent l'importance et l'utilité de cette dernière.

[18] Buckwaler a utilisé les tables de compatibilités entre préfixés, racine, et suffixe.

Chapitre V : Comparaison entre les analyseurs morphologique : à base de schèmes de surface et à base de graphe

I. Introduction :

L'analyse morphologique des mots, est une étape très importante dans le domaine du traitement automatique des langues. Plusieurs travaux, dans cet axe, ont été élaborés ces dernières années, et qui se basent généralement sur l'une des approches suivante : l'approche à deux niveaux, l'approche par concaténation, l'approche à base d'automate à état fini, l'approche à base de règles, ou des approches qui combinent entre ces différentes approches.

Dans ce chapitre, nous proposons une comparaison entre nos deux analyseurs morphologiques, que nous avons développés ces dernières années. Le premier est à base de schèmes de surface des mots arabes, le deuxième est un analyseur qui combine entre l'approche de Buckwalter et l'approche d'analyse morphologique à base de graphe.

La comparaison est faite sur un corpus de 10000 mots arabes qui généralisent tous les cas des mots dérivés arabes. Les résultats obtenus montrent l'intérêt et les avantages de chaque analyseur.

II. Caractéristiques des deux analyseurs

Notre apport personnel se veut le développement de deux approches de l'analyse morphologique de la langue arabe. La première s'appuie sur les schèmes de surface de tous les mots dérivés arabes, et la deuxième approche, elle fait appel aux graphes pour faire l'analyse morphologique d'un mot donné, nous proposons une comparaison entre ces deux approches. Avant de présenter la comparaison en terme de précision et de rappel, nous présentons les avantages et les inconvénients de chaque approche.

Parmi les avantages de ces deux approches, nous citons en guise d'exemple :

- Les deux approches ont le mérite de réduire la taille des dictionnaires utilisé dans les autres analyseurs : la première approche utilise la base de schèmes de surface en plus de la base des racines pour modéliser toutes les variations morphologiques des mots dérivés. Quant à la deuxième approche, elle s'articule seulement sur la base des racines.
- Dans la deuxième approche nous avons généré un réseau global d'états sans utilisation des dictionnaires lexicaux (ceci est le grand avantage de notre analyseur par rapport à l'analyseur de Buckwalter et l'analyseur à base d'automate à état fini) .
- L'approche à base de graphe donne toutes les analyses morphologiques d'un mot donné, mais l'approche à base de schèmes, donne seulement les analyses associées aux schèmes de surface existants dans la base des schèmes.
- La deuxième approche n'utilise aucune base de connaissance linguistique pour faire l'analyse morphologique d'un mot.
- le taux de couverture de ces deux approches pour les mots dérivés arabes, est très élevé par rapport aux autres analyseurs.

Malheureusement, comme tout ce que fait l'homme, les deux approches présentent quelques inconvénients :

- L'approche à base de schème traite seulement les mots dérivés, tandis que l'approche à base de graphe peut être utilisée même pour les mots non dérivés.
- La deuxième approche elle ne prend pas en compte toute les règles de compatibilité entre les affixes (infixes, préfixes et suffixes) et les racines. Par voie de conséquence, cette approche retourne des erreurs à cause de cette négligence.

III. Evaluation de l'approche à base de schèmes

Pour l'analyseur à base de schèmes de surface, nous avons utilisé :

- Un lexique de 6216 schèmes de surface. Ce lexique contient toutes les classes morphologiques des mots dérivés arabes.
- Un dictionnaire des racines, il contient 1200 racines.
- Un dictionnaire des stems, il contient 6000 stems.

L'évaluation est effectuée sur deux corpus, le premier contient 10063 mots dérivés, et il contient presque tous les catégories des mots dérivés, et le deuxième contient 4600 mots dérivés et il est utilisé pour comparer entre les deux approches. Le tableau suivant, représente le taux d'erreur, de l'analyse à base de schèmes de surface, pour chaque catégorie des mots dérivés.

Type du mot dérivé	Nombre de mots	Taux d'erreur
فعل	2964	7,02%
اسم-الفاعل	1603	9,11%
اسم-المفعول	1661	3,91%
اسم-الآلة	348	2,59%
اسم-التفضيل	250	4,00%
اسم-الزمان	700	0,86%
اسم-المكان	715	1,26%
التصغير	73	8,22%
المرة	210	0,95%
المصدر	315	2,86%
المصدر-الصناعي	109	1,83%
المصدر-الميمي	641	0,31%
الهيئة	191	1,05%
صيغة-المبالغة	237	5,49%
الصفة-المشبهة	46	0,00%
Total	10063	4,86%

Tableau n° 15 : Taux d'erreur de l'analyse à base de schèmes de surface pour chaque catégorie des mots dérivés.

En général, on remarque que le taux d'erreur d'analyse, pour chaque type des mots dérivés, est proche de 4,86%. Le taux le plus faible est obtenu pour le type "الصفة-المشبهة" avec un taux d'erreur de 0%, tandis que les taux les plus élevés sont celui de "اسم-الفاعل" et de "التصغير" avec des pourcentage de 9,11% et de 8,22%. Le taux moyen d'erreur pour tous ces types, il est de l'ordre de 4,86%.

La plupart de ces erreurs proviennent essentiellement de la non prise en compte des règles de compatibilité entres les affixes et les racines.

IV. Comparaison entre les deux approches :

Pour comparer entre la performance de chaque approche, nous avons utilisé les corpus suivant :

- L'ensemble de test, il contient 4000 mots dérivés arabes, et ils représentent presque tous les cas des mots dérivés arabes. Il est utilisé pour évaluer les deux analyseurs.

Pour la comparaison entre ces deux analyseurs, nous avons utilisé les indicateurs suivants : la précision, le rappel et le temps d'exécution. Les résultats trouvés sont représenté dans le tableau suivant :

Analyseurs	Précision	Rappel	Temps d'exécution
A base de schèmes	97.08%	94,20%	24 ms
A base de Graphe	95.97%	98.58%	14 ms

Tableau n° 16 : Précision, le Rappel et le Temps d'exécution pour les deux analyseurs

L'analyseur à base de schèmes de surface donne une précision de 97%, le 3% restant provient des erreurs d'analyse de cet analyseur. Ces erreurs proviennent essentiellement de la non prise en considération les règles de correspondance entre les préfixes, les suffixes et les radicaux. Pour le rappel, notre système a pu retourner 94,20% des analyses juste, parmi toutes les analyses possibles, le 5,8% restant du rappel, provient de l'insuffisance des lexiques des schèmes de surface et des racines qui ne recouvrent pas à 100% les 4000 mots.

L'analyseur à base de graphe donne une précision de 95,97%, et le 4,03% restant représente le nombre des analyses fausses parmi les analyses retournées. Le rappel, il est de l'ordre de 98,85% , c'est-à-dire que cet analyseur représente une insuffisance de 0,15% dans le lexique des radicaux utilisés pour analyser les 4000 mots.

Pour la comparaison entre ces deux analyseurs, l'analyseur à base de schèmes de surface retourne plus de solutions valides par rapport à celui à base de graphe. Tandis que l'analyseur à base de graphe retourne plus d'analyses en un temps presque le moitié par rapport à celui à base de schèmes de surface.

V. Comparaison des résultats de notre analyseur avec d'autres analyseurs

En intégrant les tables de compatibilité infixes-préfixes, infixes-suffixes et infixes-racines pour l'évaluation de notre deuxième approche, nous avons utilisé le même corpus de référence de 10000 entrées et le même ensemble des mots dérivés composés de 4000 mots dérivés. Ensuite, nous avons comparé le rappel et la précision de notre deuxième approche avec ceux de Buckwalter et Al-Khalil et par rapport à notre première approche. Les résultats obtenus sont affichés sur le tableau ci-dessous.

Analyseurs	Précision	Rappel	Temps d'exécution
Analyseur Buckwalter	26.93%	3.21%	2.11 ms
Analyseur Al-Khalil	40.22%	30.2%	31.7 ms
A base de schèmes	97.08%	94,20%	24 ms
A base de Graphe	95.97%	98.58%	14 ms

Tableau n° 17 : Résultats après la mesure de la précision et le rappel de notre analyseur et de Buckwalter et d'Al-khalil

En comparant notre deuxième approche avec l'analyseur de Buckwalter et d'Al-khalil, on remarque que notre analyseur retourne un rappel et une précision qui sont très élevés par rapport à ces deux analyseurs: 98,58% contre 30.2% pour Al-khalil et 3,21% pour Buckwalter pour le rappel, et 95,97% pour la précision contre 40.22% pour Al-Khalili et 26,93% pour Buckwalter.

Tandis que le rappel de la deuxième approche est supérieur à celle à base de schèmes de surface ceci est expliqué par le fait que le nombre des analyses oubliées par la première approche est plus élevé par rapport à la deuxième approche, et ceci est un bon avantage pour l'approche à base de graphe. Pour la précision, l'analyseur à base de schèmes donne plus d'analyses valides que l'analyseur à base de graphe avec un taux de 97% contre 96%. Cela est dû à ce que la deuxième approche donne plus de solutions qui ne sont pas valides, à cause du problème de la non prise en considération de la compatibilité entre les infixes et les autres morphèmes du mot (racine, suffixe, préfixe) par exemple pour le mot "يقول "notre analyseur donne les "قل " et "قال".

De même on remarque d'une façon très claire que le système à base de graphe est très rapide au niveau du temps d'analyse, il se classe le deuxième après l'analyseur du Buckwalter.

Afin de prendre en considération la compatibilité entre les infixes et les autres morphèmes dans un mot donné, nous avons élaboré et intégré des tables de compatibilité entre les infixes et les autres morphèmes[19] dans IBN-GINNI . Les résultats obtenus sont présentés dans le tableau suivant[20].

Analyseurs	Précision	Rappel	Temps d'exécution
Analyseur Buckwalter	26.93%	3.21%	2.11 ms
Analyseur Al-Khalil	40.22%	30.2%	31.7 ms
A base de schèmes	97.08%	94,20%	24 ms
A base de Graphe	95.97%	98.58%	14 ms
A base de graphe avec table de compatibilité entre affixes	97.71%	98.9%	16 ms

Tableau n° 18 : Résultats après la mesure de la précision et le rappel de notre analyseur et de Buckwalter et d'Al-khalil en utilisant la compatibilité entre les affixes.

Après l'introduction de la table de compatibilité on a réduit le taux d'erreur par une unité de 1%. Cette table qu'on a élaborée est de taille restreinte et le chantier reste ouvert pour mener une amélioration à ce niveau là.

Conclusion :

Nous avons proposé une comparaison entre les deux analyseurs morphologiques à base de schèmes de surface et à base de graphe. Cette comparaison a permis de détecter les points forts et les points faibles de chaque analyseur. Par la suite, il sera très intéressant de combiner entre ces

[19] On note ici que c'est la première fois qu'on parle de la correspondance entre les infixes et les autres morphèmes, Buckwalter a utilisé compatibilité entre les préfixe-suffixe, préfixes-radical, radical-suffixes.

[20] Les résultats de ce travail sont acceptés pour publication dans le Journal of Theoretical and Applied Information Technology.

deux approches, dans un seul analyseur pour augmenter la précision et le rappel en même temps, en gardant un temps d'exécution très proche du celui du deuxième analyseur.

CONCLUSION GENERALE:

Dans cette thèse, nous avons traité le problème de l'analyse morphologique pour la langue arabe. Il s'agit d'un travail complexe et difficile. Mais grâce aux directives et aux conseils de l'encadrant, j'ai pu mener ce travail jusqu'au bout. En effet, nous avons commencé dans un premier temps, par une approche générale du traitement automatique de la langue arabe, en montrant tout ce que cette langue a de spécifique et de particulier. Cela veut dire que la langue arabe présente des caractéristiques et des spécificités qui la rendent plus ambiguë que d'autres langues maternelles. Sa morphologie, sa syntaxe ainsi que sa sémantique sont en corrélation et se complètent. Pour cette langue, il y a un grand enchevêtrement entre ses différents niveaux.

Dans un deuxième temps, nous nous sommes attardés sur l'analyse morphologique de la langue arabe en focalisant mon attention sur les différents niveaux d'analyse. Cela présuppose de ma part une présentation de quelques analyseurs morphologiques arabes et l'évaluation de leurs résultats. Avant cela, il fallait passer par un aperçu historique qui éclaire davantage l'évolution de l'analyse morphologique de la langue arabe.

Notre apport personnel se veut le développement de deux approches de l'analyse morphologique de la langue arabe. La première s'appuie sur les schèmes de surface. En d'autres termes, j'ai traité tous les mots dérivés arabes. Ensuite, j'ai formulé la fonction qui mesure la similarité entre les mots et les schèmes de surface. La deuxième approche, quant à elle, fait appel aux graphes pour faire l'analyse morphologique d'un mot. Cette méthode combine entre l'approche à base de dictionnaires et l'approche à base d'automates à états finis. Ces deux approches ont démontré une grande fiabilité et leur efficacité est devenue on ne peut plus claire. Parmi les avantages de ces deux approches, nous citons en guise d'exemple :

- Les deux approches ont le mérite de réduire la taille des dictionnaires utilisé dans les autres analyseurs : la première approche utilise la base de schème de surface en plus de la base des racines pour modéliser toutes les variations morphologiques des mots dérivés. Quant à la deuxième approche, elle s'articule seulement sur la base des racines.
- Le taux des analyses valides est très élevé pour les mots dérivés par rapport aux autres analyseurs.
- La deuxième approche n'utilise aucune base de connaissance linguistique pour faire l'analyse morphologique d'un mot.

Malheureusement, comme tout ce que fait l'homme, mon travail présente quelques inconvénients. En fait, les deux approches que je viens de présenter rencontrent certaines limites que je formule comme suit :

- Les deux approches traitent seulement les mots dérivés
- La deuxième approche elle ne prend pas en compte toute la base de la compatibilité entre les infixes et les racines, les suffixes et les préfixes. Par voie de conséquence, cette approche retourne plus d'erreurs à cause de cette négligence.

Ces imperfections stimulent mon intérêt pour de nouveaux travaux à travers lesquels j'essayerai de généraliser ces deux approches sur tous les mots de la langue arabe (mots non dérivés). Pour le deuxième inconvénient, je tâcherai de minimiser le plus possible le nombre d'erreurs retourné, et ce en prenant en considération la fluctuation entre les infixes et les racines. Cela se présente déjà comme une tâche colossale. Laquelle tache mérite d'être traiter afin d'améliorer le traitement automatique de la langue arabe.

Annexes

Liste des annexes :

Annexe I : Précision et rappel des analyseurs morphologique

Annexe II : Corpus arabe

Annexe III: Liste des analyseurs morphologiques arabes.

Annexe IV : Préfixes et suffixes et leurs compatibilités

Annexe V : Les analyseurs morphologiques arabes

Annexe I : Précision et rappel pour les analyseurs morphologiques

En particulier, on les rencontre en permanence en traitement automatique de la langue, aussi bien en analyse. Dans le domaine de l'analyse morphologique, il est parfois difficile de donner des indicateurs de qualité et de comparer les performances. Pour caractériser la qualité d'un analyseur morphologique, on utilise souvent les critères : la précision et le rappel. L'évaluation consiste à mesurer la différence entre un résultat attendu et un résultat obtenu.

Ces mots sont fréquemment utilisés dans des domaines faisant appel à des techniques proches de la statistique.

Pour un système la Précision et le Rappel sont définis l'aide des paramètres suivant :

	Pertinent (vrai)	Non pertinent (faux)	Total
Retrouvé	a	b	a+b
Non retrouvé	c	d	c+d
	a+c	b+d	a+c+b+d

- **la Précision** = (Nbr des résultats retrouvés et vrais)*100/(Nbr bons résultats retrouvés+Nbr des résultats retrouvés et faux)

$$\text{Précision} = \frac{a}{a+b}$$

- **le Rappel** = (Nbr bons résultats retrouvés)*100/(Nbr bons résultats retrouvés +Nbr résultats non retrouvés et vrais)

$$\text{Rappel} = \frac{a}{a+c}$$

Les résultats sont évalués en termes de précision et rappel par un linguiste.

Annexe II : Liste des approches des analyseurs morphologiques arabes

année	Nom analyseur	source		Groupe	Information électroniques (Email, site,..)
2012	Apport de la diacritisation dans l'analyse morphosyntaxique de l'arabe	Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 3: RECITAL, pages 247–254, Grenoble, 4 au 8 juin 2012.c 2012 ATALA & AFCP	Ahmed Hamdi	Aix Marseille Université, LIF-CNRS, Marseille	ahmed.hamdi@lif.univ-mrs.fr
2011	Morphological Analysis and Generation for Machine Translation from and to Arabic.	International Journal of Computer Applications 18(2):14-18, March 2011. Published by Foundation of Computer Science.	Sadik Bessou Mohamed Touahria	Ferhat Abbas University Sétif- Algeria	Bessou.s@gmail.com
2011	Developing a New Approach for Arabic Morphological Analysis and Generation		Mourad Gridach Noureddine Chenfour	Mathematics and Computer Science Department, Sidi Mohamed Ben Abdellah University – Faculty of Sciences, Fez, Morocco.	mourad_i4@yahoo.fr chenfour@yahoo.fr
2010	The morphological analysis of Arabic verbs by using the surface patterns.	IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, No 11, May 2010	A. Yousfi	The institute of studies and research for the arabization University Mohamed V Souissi, Rabat,Morocco	yousfi abdellah <yousfi240ma@yahoo.fr>;
2010	Fine-Grain Morphological Analyzer and Part-of-Speech Tagger for Arabic Text. in Language Resource and Evaluation	Conference LREC 2010 2010. Valleta, Malta: European Language Resources Association (ELRA).	Majdi Sawalha and Eric Atwell	School of Computing, University of Leeds, Leeds, LS2 9JT, UK	sawalha.majdi@gmail.com/sawalha@comp.leeds.ac.uk/eric@comp.leeds.ac.uk
2010	A robust morphological analyzer for Arabic texts.	in JADT 2010: 10th International Conference on Statistical Analysis of Textual Data. 2010: SAPIENZA, Italy.	Nouha Chaâben Kammoun, Lamia Hadrich Belguith Abdelmajid Ben Hamadou	MIRACL Laboratory FSEGS – B.P. – 1088 –3018 Sfax – Tunisia ISIMS – B.P. – 242 –3021 Sakiet-Ezzit Sfax – Tunisia	
2010	Morphological Analysis of Ill-formed Arabic Verbs in Intelligent Language Tutoring Framework		Khaled Shaalan Marwa Magdy Aly Fahmy	The British University in Dubai, PO Box 345015 Dubai, UAE Faculty of Computers & Information, Cairo University, 5 Ahmed Zewel St., Giza 12613 Egypt	khaled.shaalan@buid.ac.ae, m.magdy@fci cu.edu.eg, a.fahmy@fci cu.edu.eg
2010	Alkhalil Morpho Sys: A Morphosyntactic analysis system for Arabic texts.	IJCSI International Journal of Computer Science Issues.	a. boudlal, a.lakhouaja, a. mazroui, a. meziane m. ould abdallahi ould bebah and m. shoul	Faculty of Letters and Human Sciences, University Mohamed I, Oujda, Morocco	rahimboudlal@hotmail.com, mshoul@hotmail.com abdel.lakh@gmail.com, azze.mazroui@gmail.com, abdelouafi_meziane@yahoo.fr, medbebaha@yahoo.fr
2010	Morphological Analysis and	The International Journal of ACM Jordan	Amjad M Daoud	Computer Science Department Isra	

	Diacritical Arabic Text Compression.	(ISSN 2078-7952), Vol. 1, No. 1, March 2010		University Amman, Jordan	
2010	Morphological analyzers evaluation",	The Morphological analyzers experts meeting, Syria,2009.	Salwa Hamada	Electronics Research Institute, Egypt Cairo Electronics Research Institute Dokki- Tahrir st.	hesalwa@yahoo.com
2009	Morphological analyzers, The Morphological analyzers experts meeting ,Syria,2009.		Salwa Hamada	Egypt Cairo Electronics Research Institute Dokki- Tahrir st.	hesalwa@yahoo.com
2008	Un analyseur morphologique pour l'arabe voyellé ou non.	SIIE'2008 1ère Conférence Internationale, Systèmes d'Information et Intelligence Economique, SIIE 2008 Hammamet – Tunisie, 14-16 Février 2008, Proceedings tome II, IHE éditions, ISBN 9978-9973-868-20-6, pp. 324--339 (2008)	Mohammed El Amine abderrahim	Université Abou Bekr Belkaid Tlemcen, Algérie. Faculté des sciences de l'ingénieur Département d'informatique BP 230 chetouane	medamineabd@yahoo.fr
2007	A Morphological Analyser for Arabic Language.	ICTA'07, April 12-14, Hammamet, Tunisia	Mourad mars Georges antoniadis Mounir zrigui	LIDILEM laboratory University of Stendhal, Grenoble3, France UTIC laboratory Faculty of Sciences of Monastir, tunisie.	Mourad.Mars@e.u-grenoble3.fr Georges.Antoniadis@ugrenoble3.fr Mounir.zrigui@fsm.rnu.tn
2006	MAGEAD: A Morphological Analyzer and Generator for the Arabic Dialects	A Morphological Analyzer and Generator for the Arabic Dialects. In Proceedings of ACL, Sydney, Australia, 2006.	Habash,Nizar and Owen Rambow	In Proceedings of ACL, Sydney, Australia, 2006.	habash@ccls.columbia.edu habash@umiacs.umd.edu
2006	An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modelling Finite State Networks.	The Challenge of Arabic for NLP/MT Conference, The British Computer Society, London, 2006.	Mohammed Attia	School of Languages, Linguistics and Cultures. School of Informatics, The University of Manchester	mattia@computing.dcu.ie
2006	Analyse et désambiguïsation morphologiques des textes arabes non voyellés.	Actes de la 13ème édition de la conférence sur le Traitement Automatique des Langues Naturelles (TALN 2006), pp. 493-501. Belgique.	L.B.Hadrich, N.Chaâben,	ANLP Research Group- MIRACL Laboratory, University of Sfax, Tunisia.	l.belguith@fsegs.rnu.tn nouha.chaaben@laposte.net
2006	S. Mesfar, Analyse lexicale et morphologique de l'arabe standard utilisant la plateforme linguistique NooJ,	Actes de la 13ème Conférence sur le Traitement Automatique des Langues Naturelles RECITAL/ TALN 2006, Presses universitaires de Louvain, Louvain-la-Neuve, Belgique, Avril 2006	Slim mesfar	LASELDI, Franche-Comté University, France	mesfarslim@yahoo.fr
2004	Arabic morphological analysis techniques: A comprehensive survey.	Journal of the American Society for Information Science and Technology, 2004. 55(3): p. 189-213.	Al-Sughaiyer, I.A. and I.A. Al-Kharashi	Computer and electronics research institute king abdulaziz city for science and technology P.O.Box 6086, Riyadh 11442, Saudi arabia	{Kharashi, Imad}@kacst.edu.sa
2003	New approach for extracting Arabic	Paper presented at the International Arab	Al-Shalabi	Department of Computer Science and	

	roots.	Conference on Information Technology (ACIT'2003), Alexandria, Egypt.		Applied Mathematics Illinois Institute of Technology 10 West 31st Street Chicago, IL 60616	http://hdl.handle.net/2086/1209 alshriy@minna.cns.iit.edu,
2002	Arabic Morphological Analyzer Version 2.0. Linguistic Data Consortium	Linguistic Data Consrtium, catalog. Number LDC2002L49 and ISBN 1-58563-257-0.	T.Buckwalter	Linguistic Data Consortium, University of Pennsylvania, United States	http://ufal.mff.cuni.cz/padt/PADT_1.0/docs /index.html http://www.qamus.org/wordlist.htm timbuck2@ldc.upenn.edu
2002	Morphological analysis for automatic Arabic text classification	in Proceedings of the ACLWorkshop on Computational Approaches to Semitic Languages, Philadelphia, PA, 2002.	k.Darwish	Electrical and Computer Engineering Dept. University of Maryland, College Park College Park, MD 20742	kareem@darwish.org / kareem@glue.umd.edu kdarwish@qf.org.qa,
2002	La conception et la réalisation d'un système d'analyse morphosyntaxique robuste pour l'arabe : utilisation pour la détection et le diagnostic des fautes d'accord.	Doct. dissert., ENSSIB/Université Lyon 2.	Riadh Ouersighni	ENSSIB 17/21 Bld du 11 novembre 1918, 69623 Villeurbanne Cedex, France	ouersighni@enssib.fr
2001	Web-based Arabic Morphological analyser.	In Gelbukh, A (Ed.): CICLing 2001, LNGS 2004, pp. 216-225. Springer-Verlag Berlin Heidelberg.	Jawad Berri, Hamza Zidoum, & Atif, Y	King Saud University · Department of Information Systems	{j.berri, hamza.zidoun, yacine.atif}uaeu.ac.ae
2001	Arabic Computational Morphology.	Knowledge-based and Empirical Methods. Dordrecht, The Netherlands: Springer	A., A.V.D.Bosch G.Neumann	Ecole Nationale de I'Industrie Minérale, Rabat, Morocco. Tilburg University, The Netherlands. Deutsches Forschungszentrum fur Kunstliche Intelligenz, Saarbrucken, Germany.	asoudi@gmail.com, Mobile: (+212) 6 61 05 51 09, Fax: (+212) 37 77 10 55 Neumann@dfki.de Antal.vdnBosch@uvt.nl
1996	Arabic Finite-State Morphological Analysis and Generation	Proceedings of the 16th conference on Computational linguistics, Vol 1. Copenhagen,Denmark: Association for Computationa Linguistics, pp 89-94	k.Beesley	Xerox Research Centre Europe Grenoble Laboratory 6, chemin de Maupertuis 38240 meylan France	Beesley@xrce.xerox.com / Ken.Beesley@xrce.xerox.com ken.beesley@xrce.xerox.com
1989	A new algorithm to generate Arabic root pattern forms.	In Proceedings of the 11th national computer conference. King Fahd University of Petroleum & Minerals, Dhahran, Saudi Arabia, pp. 391–400, 1989.	Al-Fedaghi, S. S. and Al-Anzi, F. S.	Computer Engineering Department P.O. Box 5969, Safat 13060, Kuwait	sabah@alfedaghi.com, sabah.alfedaghi@ku.edu.kw fawaz.alanzi@ku.edu.kw, alanzif@eng.kuniv.edu.kw
1989	Automatic Morphological Analysis of Arabic: A Study of Content Word Analysis.	Proceedings of the First Kuwait Computer Conference, Kuwait, March, 3-5.	Saliba.B and Al-Dannan.A		
1986	Natural Arabic Language Processing.	Proceedings of the 9th National Computer Conference and Exhibition, Riyadh, Saudi Arabia, 1-17.	Hegazi, N., and ElSharkawi		
1985	Morphological analysis of Arabic speech.	In: Computer processing of the Arabic language.	Hilal, yahiah		
1984	Ahmed Rafea	Nineteenth Annual Conference on	Ahmed Rafea	Computer Science Dept.,	rafea@aucegypt.edu

		Statistics, Computer Science and Operations Research		American University in Cairo 113, Sharia Kasr El-Aini, P.O. Box 2511, 11511, Cairo, Egypt	
1961/ 1970	Essai d'une analyse automatique de l'arabe.	Dans: David Cohen. Etudes de linguistique sémitique et arabe. Paris:Mouton, p. 49-78, (1970).	David Cohen	Paris:Mouton, p. 49-78, (1970).	

Annexe III : Liste des préfixes et suffixes arabe et avec leurs compatibilité

a. Table : les préfixes de la langue arabe

Nombre de lettres	Proclitique + Préfixes
1	"ك" , "ب" , "و" , "ف"،"ل"،"ت"،"أ"،"ن"،"ي"،"ا"
2	","بم","لم","كمـ"،"فم"،"لن"،"لت"،"لي"،"با","فا"،"كا"،"سأ","سن"،"وس","ست"،"سي"،"يا","لل"،"ال" "،"لأ"،"فت"،"في"،"فن"،"فأ","وا","وت","وي"،"ون"،"وأ"،"أت"،"أي"،"أن"،"أأ","فل","نت","مت "بت" ,"يت" ,"فك" ,"فب"،"وم"،"ول","وك"،"وب"،"أك"،"أب"،"أس","أل","
3	،"ولي"،"ولا"،"وسأ"،"وسن","وسي","وست","لال","سال","اال","مال","ولل"،"كال"،"فال"،"بال"،"وال" "أسن"،"أسأ"،"أفت"،"أفي","أفن","أفأ","أوي","أوت"،"أون"،"أوأ"،"ألت"،"ألي"،"ألن"،"ألأ","ولن"،"ولت" ,"فست" , "فسي" , "فسن"،"فسأ"،"فلت"،"فلي"،"فلن"،"فلأ" , "ولت" , "ولي","ولن","ولأ"،"أست"،"أسي"، "فلل"،"أول"،"أفب" , "أوب"
4	,"أفلأ" , "أولت" , "أولي"،"أولن"،"أولأ"،"أفست"،"أفسي"،"أفسن"،"أفسأ","وكال"،"وبال","بكال"،"فبال" "أفلت"،"أفلي" , "أفلن"

b. Table : les suffixes de la langue arabe

Nombre de lettres	Suffixes + Enclitique
1	"و" ,"ت"، "ي" ،"ا"، "ة"، "ك"، "ن"، "ه "
2	,"ته"، "تن" ،"تي"، "تم"، "وا"، "نا"، "كن " ,"كم" ,"ها"، "هن" ،"هم"، "ات"، "ون"، "ين"، "ان " ,"ات"، "نه" ،"نك"، "اه"، "اك"، "ية"، "يه " ,"وه" ,"ته"، "نا" ،"ني"، "ما"، "يا"، "تا"، "تك " ,"يه" "وك"
3	،"وكن"، "وكم"، "وها"، "ونا"، "تكن " ,"تكم" ,"هما"، "تهن" ،"تهم"، "وهن"، "وني"، "وهم"، "تها " "يني"، "ينه"، "ننا " ,"ونه" ,"نكن"، "تنا" ،"نكم"، "نهم"، "نها"، "تاي"، "تاك " ,"اتك" ,"اتي"، "تني" ،"يهم"، ,"كم" ,"يها"، "ناك" ،"ناه"، ,"اكم" ,"اكن"، "اني" ،"انا"، ,"ونك" ,"اته"، "اهم" ،"اها"، "تاه"، "تما"، "تمو" ،"تان" ,"تان" ,"تين"، "انه" ،"تنه" ,"ينا" ,"يني"، "يهن"
4	"تاهن"، "تاهم"، "تاها " ,"وننا" ,"ونني"، "ونكن" ،"ونكم"، "ونهن"، "ونها"، "ونهم"، "يهما " ,"تهما " ,"نهما"، "اتنا" ،"اتكن"، "اتكم"، "اتهن"، "اتهم"، "اتها " ,"وهما" ,"تانا"، "تاني" ،"تاكن"، "ناكم"، "تنهم"، "تنها"، "تماه"، "تاكن " ,"تاكم" ,"ناهم"، "ناها" ،"اهما"، "يننا"، "ينني"، "ينهن"، "ينهم " ,"ينها" "تموه"،"انكن"، "انكم"، "انهن " ,"انهم" ,"اتها"، "تتني" ،"تنهن"،
5	"تماهم"، "تماها " ,"تمونا" ,"تموني"، "تموهن" ،"تموهم"، "ناهما"، "تموها"، "ونهما"، "تاهما " "تينهن" ,"تينهم"، "تينها"، "انهما " ,"تنهما "، "تماني"، "تماهن"،
6	"تماهما"، "تموهما"، "تينهما "

Annexe A :

Tableau de translittération de l'alphabet arabe API

lettres	Transcription international (API)	Lettres	Transcription international (API)
ا	’		’
ب	b		b
ت	t		t
ث	ṯ		ṯ
ج	ĝ		ĝ
ح	ḥ		ḥ
خ	ḫ		ḫ
د	d		d
ذ	ḏ		ḏ
ر	r		r
ز	Z		Z
س	S		S
ش	Š		Š
ص	ṣ		ṣ
ض	ḍ		ḍ
	ṭ		ṭ
	ẓ		ẓ
	ġ		ġ
	f		f
	q		q
	k		k
	l		l
	m		m
	n		n
	h		h
	w		w
	y		y
	a		a
	u		u
	i		i
	ã / an		ã / an
	ũ / un		ũ / un
	ĩ / in		ĩ / in
	ä (at en annexion)		ä (at en annexion)
	ą		ą
	å		å
	á		á
	ù		ù
	í		í
	ẃ		ẃ
	ý		ý

Références :

Al Fedaghi.S and Al-Anzi, 1989 : A new application to generate Arabic Root-Pattern Forms, Procedings of the 11th National Computer Conference and Exhibition, March, Dahran, Saudia Arabia, 391-400.

Al-Hamlawy A. 1957. Shaza Al-Orf in the art of morphology. (by) Dar Al-Kiaan, Riadh, KSA, (Arabic book).

ALESCO, "Arabic Language Derivation and morphological System," Published by the Arab League Educational, Cultural and Scientific Organization, http://www.reefnet.gov.sy/ed4-2.htm, Last Visited 2007.

Al-Ghalayyni. 2005. جامع الدروس العربية "Jami' Al-Duroos Al-Arabia". Saida - Lebanon: Al-Maktaba Al-Asriyiah "المكتبة العصرية".

Al-Rajhi A. 1979. The application of morphology. (by) Dar Al-Nahdha Al-Arabia Beirut, (Arabic book).

Al-Sughaiyer, I. A. and Al-Kharashi, I. A. 2004. Arabic morphological analysis techniques: A comprehensive survey. Journal of the American Society for Information Science and Technology 55(3): 189-213.

Al-Shalabi et al., 2003 : New approach for extracting Arabic roots. Paper presented at the International Arab Conference on Information Technology (ACIT'2003), Alexandria, Egypt.

M.Al- Ğalāyīnī, 2000; جامع الدروس العربية, المكتبة العصرية بيروت, صيدا لبنان, 2000.

Alexia Blanchard. Analyse morphologique des réponses d'apprenants en environnement d'Apprentissage Assisté par Ordinateur. Université Stendhal-Grenoble III,UFR des Sciences du Langage.

Attia, M. 2006 : An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modelling Finite State Networks. The Challenge of Arabic for

NLP/MT Conference, the British Computer Society, London.

Audebert C, Jaccarini A. 1988. De la reconnaissance des mots outils et des tokens. Annales islamologiques 24, Institut francais d'archeologie orientale du Caire.

Awajan.A, 2011: "Multilayer Model for Arabic Text Compression", The International Arab Journal of Information Technology, Vol. 8, No. 2, April 2011

Bahrak. (869- 930هـ) ، جمال الدين محمد بن عمر بن مبارك الحميري الحضرمي .بحرق، فتح الأقفال وحل الإشكال بشرح لامية الافعال.

Beesly.KR 1998. Arabic Morphology Using Only Finate-State Operations, Proceedings of the Workshop on Computational Approaches to Semetic languages. Montreal, Quebec, pp 50-57.

Beesley KR 1996. Arabic Finite-State Morphological Analysis and Generation. Proceedings of the 16th conference on Computational linguistics, Vol1. Copenhagen,Denmark: Association for Computational Linguistics, pp 89-94.

Beesley, Kenneth, Tim Buckwalter, and Stuart Newton, "Two-Level Finite-State Analysis of Arabic Morphology." Proceedings of the Seminar on Bilingual Computing in Arabic and English, Cambridge, England, 1989.

Beesley, Karttunen; Finite-State Non-Concatenative Morphotactics 2000.

Beesley, 2001, Finite-State Morphological Analysis and Generation of Arabic at Xerox Research: Status and Plans - Beesley – 2001.

Beesley, Karttunen ; Finite-State Morphology: Xerox Tools And Techniques - (Show Context) 2003.

Berri,j, Zidoum,H, & Atif, Y: Web-based Arabic Morphological analyser. In Gelbukh, A (Ed.): CICLing 2001, LNGS 2004, pp. 216-225. Springer-Verlag Berlin Heidelberg.

Boudlal, A., Lakhouaja, A., Mazroui, A., Meziane, A., Bebah, M. O. A. O. and

M.Shoul. 2010. Alkhalil Morpho Sys: A Morphosyntactic analysis system for Arabic texts. IJCSI International Journal of Computer Science Issues.

Buckwalter.T 2002. Buckwalter Arabic Morphological Analyzer. Version 1.0. Linguistic Data Consrtium, catalog. Number LDC2002L49 and ISBN 1-58563-257-0.

Buckwalter, T. 2004. Buckwalter Arabic Morphological Analyzer Version 2.0. Linguistic Data Consortium, catalog number LDC2004L02 and ISBN 1-58563-324-0.

Chomsky, Noam (1957). Syntactic Structures. The Hague: Mouton and Co.

Cohen.D, 1961/1970: Essai d'une analyse automatique de l'arabe. Dans: David Cohen. Etudes de linguistique sémitique et arabe. Paris:Mouton, p. 49-78, (1970).

Dahdah, A. 1987. A Dictionary of Arabic Grammer in Charts and Tables " معجم قواعد اللغة العربية العربية –في جداول ولوحات". Beirut, Lebanon: Librairie du Liban publisher.

Dahdah, A. 1993. A dictionary of Arabic Grammatical nomenclature Arabic – English "معجم لغة النحو العربي عربي-انكليزي". Beirut, Lebanon: Librairie du Liban publishers.

Darwish K. (2002). Building a Shallow Morphological Analyzer in One Day. Proceedingsof the workshop on Computational Approaches to Semitic Languages in the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02). Philadelphia, PA, -USA.

Dichy. J, S. Ammar, Les Verbes Arabes, Bescherelle, Paris, 1999.

Dukes, K. and Habash, N. 2010. Morphological Annotation of Quranic Arabic. Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta,19-21 May 2010.: European Language Resources Association (ELRA).

El-Sadany.T.A and Hashish.M.A 1989. An Arabic Morphological System. IBM Systems Journal. Vol.28, No.4, 600-612.

Farghaly. A, K. Shaalan. Arabic Natural Language Processing: Challenges and Solutions, ACM Transactions on Asian Language Information Processing (TALIP),

the Association for Computing Machinery (ACM). TALIP Vol 8, Issue 4, December 2009.

Fleisch, Henri., (1964), « Arabe Classique et arabe dialectal ». Travaux et Jours 12, pp.23- 64 (repr. in Henri Fleisch., Etudes d'arabe dialectal, Beirut: Imprimerie Catholique, 1974, pp. 3-43).

Gaubert C. Analyse morphologique d'un texte par ordinateur – Résultats et évaluation. AnIsl 29 (1996), IFAO, p. 283-311

Goldsmith and John.A (2001). Unsupervised learning of the morphology of a natural language. Computational Linguistics, 27(2), 153-198.

Habash, Nizar and Owen Rambow. MAGEAD: A Morphological Analyzer and Generator for the Arabic Dialects. In Proceedings of ACL, Sydney, Australia, 2006.

Habert et al., 1997 Les linguistiques de corpus. U Linguistique. Paris: Armand Colin/Masson.

Hamada, S. 2009b. "المحللات الصرفية للغة العربية" proposal for evaluating morphological analyzers for Arabic text. Proceedings of the workshop of morphological analyzer experts for Arabic language, organized by Arab League Educational, Cultural and Scientific Organization ALECSO, King Abdul-Aziz City of Science and Technology (KACST) and Arabic Language Academy., Damascus, Syria. 26-28 April 2009.

Hamada, S. 2010. Evaluation of the Arabic Morphological Analyzers Proceedings of The Sixth International Computing science Conference ICCA, Hammamet, Tunisia.

Hanafi. (1914). حنفي بك ناصف-محمد بك دياب-مصطفي طموم-محمود افندي عمر-سلطان بك محمد, طبعه مصر سنه 1914 اديان. علوم الدين.قواعد اللغه العربيه لتلاميذ المدارس الثانويه

Hegazi.N and ElSharkawi.A 1986 : Natural Arabic Language Processing, Proceedings of the 9th National Computer Conference and Exhibition, Riyadh, Saudi Arabia, 1-17.

Hilal, Yahiah, 1985 : Morphological analysis of Arabic speech. In: Computer

processing of the Arabic language.

S.Iazzi, A.Yousfi, M.Bellafkih 2021: Comparison Of Two Approaches: Morphological Analysis Based On Graph And The One Based On Surface Patterns. International Journal of Computer Science and Applications, 2021, 18(1), pp. 102–115.

S.Iazzi, A.Yousfi, M.Bellafkih, D.Aboutajdine 2018 : Arabic Morphological Analysis Based On Graphs And Correspondence tables Between Affixes And Root. ISIVC 2018: 318-322

S.Iazzi, A.Yousfi, M.Bellafkih. 2020: Comparison between the morphological analyzers based on graph and the one based on surface patterns. SITA 2020: 26:1-26:4

Iazzi, S, Yousfi, A, Bellafkih, M. Analyseur morphologique des mots arabe en utilisant le dérivé et le schème de surface. 5ème Edition de la conférence internationale sur les Technologies d'Information et de Communication pour l'Amazighe (TICAM 2012) 26 - 27 novembre 2012, Institut Royal de la Culture Amazighe (IRCAM), Morocco.

Iazzi, S, Yousfi, A, Bellafkih, M, Aboutajdine, D. Morphological Analyzer of Arabic Words Using the Surface Pattern. IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 2, No 1, March 2013, ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784.

Iazzi, S, Yousfi, A, Bellafkih, M, Aboutajdine, D. Analyse morphologique à base de graphe. 8TH international conference on intelligent systems: theories and applications, 08-09 may 2013, Rabat, Morocco. SITA'13.

Iazzi, S, Yousfi, A, Bellafkih, M, Aboutajdine, D. Graph-based morphological analysis. Journal of Computer Science and Engineering Volume 19, Issue 2 June 2013.

Jaccarini A., (1997). Grammaires modulaires de l'arabe. These de doctorat. Universite de Paris-Sorbonne.

Kadri, Y., Benyamina, A..: 1992, Un système d'analyse syntaxico-sémantique du langage arabe non voyellé", Mémoire d'ingénieur, Université d'Oran.

Khoja.S and Garside.R (1999). Stemming Arabic text. Computer Science Departement, Lancaster University, Lancaster, UK.

Kiraz, G.A., Computational Nonlinear Morphology with Emphasis on Sematic Languages. Studies in Natural Language Processing, ed. B. Boguraev and S. Bird. 2001, Cambridge: Cambridge Univeristy Press.

Koskenniemi and Kimmo (1983). Two Level Morpology. A General Computational Model for Word-form Recognition and Production. Publication No. 11, Dep. of General Linguistics, University of Helsinki, Helsinki.

Mesfar.S, (2006), Analyse lexicale et morphologique de l'arabe standard utilisant la plateforme linguistique NooJ, Actes de la 13ème Conférence sur le Traitement Automatique des Langues Naturelles RECITAL/ TALN 2006, Presses universitaires de Louvain, Louvain-la-Neuve, Belgique, Avril 2006

Mustapha.G (1999). جامع الدروس العربية .مصطفى الشيخ .الغلاييني. المكتبة ، العصرية

Ouersighni, R. (2001) A major offshoot of the DIINAR-MBC project: AraParse, a morpho-syntactic analyzer of unvowelled Arabic texts, In ACL 39th Annual Meeting. Workshop on Arabic Language Processing: Status and Prospect, Toulouse, pp. 66-72.

Ryding, K. C. 2005. A Reference Grammar of Modern Standard Arabic. Cambridge University Press.

Rafea.A and Aisha Rafea, 1884 : 'Understanding an Arabic word in a text automatically', Nineteenth Annual Conference on Statistics, Computer Science and Operations Research, Cairo, Egypt, 1984, pp. 52–77.

Sabri, S., Yousfi, A., Bouyakhf, E. (2006b) : Système d'analyse morphologique des noms Arabes. MCSEAI'06 December 07-09, 2006, Agadir.

Saliba.B and Al-Dannan.A (1989). Automatic Morphological Analysis of Arabic: A Study of Content Word Analysis. Proceedings of the First Kuwait Computer Conference, Kuwait, March, 3-5.

Sam.A et Youssef D (1999). بيشرال الأفعال العربية مجموعة ديشي، يوسف عمار، سام .باريس، هاتييه

Selim. S, M. Gheith and M. Mashhour, 'A general morphological analyzer', 20th Annual Conference on Statistics, Computer Science and Operations Research, Cairo, Egypt, 1985, pp. 21–42.

Smrz, O. 2009 : ElixirFM Functional Arabic Morphology: Case Studies. Proceedings of the workshop of morphological analyzer experts for Arabic language, organized by Arab League Educational, Cultural and Scientific Organization (ALECSO), King Abdul-Aziz City of Science and Technology (KACST) and Arabic Language Academy., Damascus, Syria.26-28 April 2009.

Soudi, A., Violetta Cavalli-Sforza (2001). A Computational Lexeme-Based Treatment of Arabic Morphology. In Proceedings of the Association for Computational Linguistics, Arabic Processing Workshop, Toulouse, July 2001, France.

Soudi A., Van den Bosch A. and Neumann G (2007). Introductory chapter to the book Arabic Computational Morphology: knowledge-based and Empirical Methods, Editors: Soudi A., Van den Bosch A. and Neumann G. In Kluwer/Springer's series on Text, Speech, and Language Technology (series editors Nancy Ide and Jean Veronis).

Thalouth, B. Al-Dannan, A., 1990 : "A comprehensive Arabic Morphological Analyzer. Trujillo A., Translation Engines: Techniques for Machine Translation, Springer Verlag, 1999.

Wehrli, E. 1997. L'analyse syntaxique des langues naturelles : problèmes et méthodes, Paris,Masson.

Wright, W. 1996. A Grammar of the Arabic Language, Translated from the German of Caspari, and Editted with Numerous Additions and Corrections. Beirut: Librairie

du Liban.

William B. Frakes and Ricardo A. Baeza-Yates, editors. 1992. Information Retrieval: Data Structures ~ Algorithms. Prentice-Hall.

Yousfi.A (2010). The morphological analysis of Arabic verbs by using the surface patterns. IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, No 11, May 2010.

Yousfi.A Sabri.S and Bouyakhf.E (2006). Système d'analyse morphologique des noms Arabes. JETALA (Journées d'Etudes sur le Traitement Automatique de la Langue Arabe), 2006, Rabat 5-7 juin, 2006.

Yousfi.A, Iazzi.S : "نحو محلل صرفي عربي يعتمد على أوزان الكلمة". 7th International Computing Conference in Arabic (ICCA'11). Riyadh, Saudi Arabia (May 31- June 2, 2011).

Zanjani. (1343 هـ) الزنجاني . متن البناء و متن التصريف العربي. مطبعة مصطفى البابي الحلبي القاهرة, 1343 هـ

Travaux réalisés :

1 - Journaux :

- Iazzi, S, Yousfi, A, Bellafkih, Comparison Of Two Approaches: Morphological Analysis Based On Graph And The One Based On Surface Patterns, International Journal of Computer Science and Applications, 2021, 18(1), pp. 102–115.
- Iazzi, S, Yousfi, A, Bellafkih, M, Aboutajdine, D, Morphological analysis by surface patterns and by graph, International Journal of Engineering and Technology (UAE), Vol 7, No 3.4 (2018) Special Issue 4.
- Iazzi, S, Yousfi, A, Bellafkih, M, Aboutajdine, D. Graph-based morphological analysis, Journal of Computer Science and Engineering Volume 19, Issue 2 June 2013.
- Iazzi, S, Yousfi, A, Bellafkih, M, Aboutajdine, D. Morphological Analyzer of Arabic Words Using the Surface Pattern. IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 2, No 1, March 2013, ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784.

2 - Conférences:

- Iazzi, S, Yousfi, A, Bellafkih, Comparison between the morphological analyzers based on graph and the one based on surface patterns. ACM International Conference Proceeding Series, 2020, pp. 151–156
- Iazzi, S, Yousfi, A, Bellafkih, M, Aboutajdine, D, Arabic Morphological Analysis Based on Graphs and Correspondence tables between Affixes and Root. 9th International Symposium on Signal, Image, Video and Communications, ISIVC 2018, pp. 318–322, 8709237
- Iazzi, S, Yousfi, A, Bellafkih, M, Aboutajdine, D. Analyse morphologique à base de graphe. 8TH international conference on intelligent systems: theories and applications, 08-09 may 2013, Rabat, Morocco. SITA'13.
- Iazzi, S, Yousfi, A, Bellafkih, M. Analyseur morphologique des mots arabe en

utilisant le dérivé et le schème de surface. 5ème Edition de la conférence internationale sur les Technologies d'Information et de Communication pour l'Amazighe (TICAM 2012) 26-27 novembre 2012, Institut Royal de la Culture Amazighe (IRCAM), Morocco.

- Yousfi.A, Iazzi.S : "نحو محلل صرفي عربي يعتمد على أوزان الكلمة". 7th International Computing Conference in Arabic (ICCA'11). Riyadh, Saudi Arabia (May 31- June 2, 2011).