RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE
CENTRE UNIVERSITAIRE SALHI AHMED– NAAMA
INSTITUT DES SCIENCES ET TECHNOLOGIES

Département de Mathématiques et Informatique

Mémoire de fin d'études

Pour l'obtention du diplôme

Mastère en Informatique

Option : **Systèmes d'information**

Thème

Une outil semi automatique pour la génération de Corpus arabe étiqueté

Présenté par :

- **Khelifi Lakhdar**
- **Melkaoui Tayeb**

Soutenu le : 12/07/2021

Devant le Jury composé de :

M. BENDIDA Aissam	***MAA***	***Centre Universitaire Naâma***	***Président***
Dr YAHIAOUI Yasser	***MCB***	***Centre Universitaire Naâma***	***Encadreur***
Dr BOUZIANE A	***MCB***	***Centre Universitaire Naâma***	***Examinateur***

Année universitaire 2020-2021

Remerciements

EN premier lieu, nous tenons à remercier Dr YAHIAOUI Yasser, pour nous avoir proposé ce sujet de recherche. Nous tenons à lui adresser nos sincères et chaleureux remerciements pour la confiance qu'il nous a témoignée dès nous. Ce mémoire n'aurait pu aboutir sans sa présence et son parfait encadrement.

Nos remerciements s'adressent également à tous les enseignants qu'ils ont nous formés durant nos études.

Nos remerciements vont également à l'ensemble des membres du jury et son président pour avoir accepté juger notre travail.

Enfin tous ceux qui ont de prés ou de loin, contribués à la réalisation de ce projet trouveront ici notre reconnaissance.

Remerciements et dédicaces

Je voudrais tout d'abord adresser toute ma gratitude à Monsieur YAHIAOUI YASSER pour son engagement, son aide et ses précieux conseils qu'il a sus me transmettre tout au long de ce projet. Je tiens à le remercier tout particulièrement pour son soutien durant toute cette année. Ma volonté de poursuivre dans ce domaine tient en particulier à son enseignement pour lequel, je souhaite lui témoigner toute ma reconnaissance.

Je souhaite témoigner de la richesse de cette année au travers d'un corps professoral passionné, déterminé et qui a toujours su manifester son soutien. Je remercie toutes ces personnes qui ont contribué au renforcement de mes connaissances et qui m'ont donné les outils indispensables à la poursuite de mes études.

J'adresse mes sincères remerciements à tous les professeurs, intervenants et toutes les personnes qui par leurs paroles, leurs écrits, leurs conseils et leurs critiques m'ont permis de mener à bien mon travail ,et surtout mon ami MELKAOUI TAYEB..

Un grand merci à l'ensemble de ma famille et plus particulièrement à mes parents et ma femme et mes filles pour leurs amour, leurs confiance, leurs conseils ainsi que leurs soutien inconditionnel qui m'a permis de réaliser les études pour lesquelles je me destine et par conséquent ce mémoire.

Khelifi lakhdar

Dédicaces

Je dédie ce modeste travail:

A mes parents.

A mes frères et mes sœurs.

A mon binôme du mémoire Lakhdar.

A tous mes amis.

Tayeb

Résumé

Par sa richesse morphologique et syntaxique, la langue arabe est considéré parmi les langues les plus difficiles a traiter dans le domaine de recherche d'information .cela et du notamment, aux divers difficultés rencontrés dans la dérivation et l'identification des rôles syntaxique des mots, qui n'a pas encore connu une progrès considérable. Notre travail se situent sur cet axe de recherche.

On a développé un outil permettant de générer des flexions et des dérivées des mots d'une façon automatique selon une base de connaissance prédéfinir comme meta-connaissances basé sur une hiérarchie des rôles syntaxiques que peut avoir un mot.. cette partie de corpus est déjà définie et représentée dans des travaux antérieurs selon un formalisme de représentation moins connu appelé l'attribut hiérarchique de subsomption en Anglais « Subsumption Hierarchical Attribute » SHA qui permet de symboliser la connaissance syntaxique sous une forme vectorielle. Et par conséquence faciliter leur manipulation.

Mots clés : analyse morphologique, analyse syntaxique, génération de corpus, flexions, des dérivées des mots, base de connaissance, l'expert, SHA,corpus

Sommaire

Chapitre III: Conception et Architecture du système

Chapitre IV : Implémentation et Résultats

Liste des figures

Liste des Tableaux

Introduction générale

Introduction général

La langue naturel prend sa valeur du fait qu'il représente un moyen de communication permet l'échange des idées quoi qu'elle soient et dans n'importe qu'elle domaine scientifique ,culturelle ou social .par rapport a ca on veut répondre a la question a quoi sert la recherche sur le traitement automatique de la langue .En plus le faits d'avoir de bonne utilisation de la langue ne peut en aucun cas nier l'utilisation pour des mauvaise intention d'où le traitement automatique de la langue naturelle prend encore une dimension sécuritaire.

Dans le temps passe, l'être humain évolue en terme de penser et de développement technologique d'une manière phénoménal. L'homme s'est développé dans tous les domaines, il a constaté que la langue peut entrer dans son champs de développement, et avec l'apparition de l'outil informatique cette alliance entre linguistes et informaticiens a donnée naissance à une nouvelle discipline connue sous l'acronyme TALN dans le but est d'automatisé le langage naturel.

Le traitement automatique de la langue naturelle (TALN) touche plusieurs domaines, telles que, les applications de correction grammaticale, les applications de communication homme/machine, les applications de traduction automatique, etc. L'automatisation de l'une de ces applications nécessite en général une étape d'analyse du texte ou document source, qui se fait à son tour par la décomposition de cette analyse en sous-tâches calquées sur les différents niveaux d'analyse linguistique à savoir l'analyse lexicale, l'analyse morphologique, l'analyse syntaxique, l'analyse sémantique et l'analyse pragmatique.

Pour les langues indo-latines telles que l'Anglais et le Français, certains systèmes existent déjà et sont à une étape assez avancée. Pour la langue Arabe, les systèmes existant souffrent de difficulté inhérente au traitement automatique.

Dans le cadre des travaux de ce mémoire, Notre travail tient sur le traitement de la

langue Arabe écrite, il évoque le problèmes de manque de ressources fiable et efficace comparativement au langue indo-latine .les ressources existants ont été développés sous une forme de traduction des ressources existants pour d'autre langue sans prise en considération des particularité de la langue arabe .

Nous proposons ici une application permet a un expert linguiste de créer une ressource en langage standard a savoir XML et le script CSV .pour être exploitable en particulier notre travail tient a donner place au mécanisme spécial de la langue Arabe tel que la dérivation (الاشتقاق) et le flexion (التفعيلات) et d'autre mécanisme comme (الإقلاب) pour les caractères du EL-ILLA

La classification utilisée est représentée par un formalisme de représentation des connaissances dit SHA pour subsumption hierarchical attribute qui permet de remplacé chaque classe de mots par un vecteur significatif permettant de symboliser la représentation de la classe et ses relations directs et indirects avec les autres classes.

Organisation du document

Le présent mémoire s'organise en (04) chapitres.

- Le premier chapitre intitulé " Traitement automatique de langage Natural " présente la notion du TALN, les différents niveaux du traitement de langage ainsi que d'autres concepts liées a TALN .
- Dans le second chapitre, un rappel sur la langue arabe, à travers sont historique, ces variantes, sa morphologie ainsi que les travaux sur l'automatisation de la langue arabe.
- Le troisième chapitre est consacré à présenter la conception et l'architecture globale du notre système .
- Le quatrième chapitre montre Les résultats obtenus à travers l'environnement de développement, l'interface graphique et les différents tests.

Chapitre I

Traitement automatique de langage Natural

Chapitre I

Traitement Automatique de Langage Natural

1. Introduction

Il y a eu un énorme développement dans le domaine de l'informatique et cela est dû à l'augmentation des ordinateurs personnels standardisés, avec des capacités de stockage et de traitement en progression exponentielle, ainsi que l'apparition du Web qui a marqué l'apogée technologique en informatique. Dans tout ce changement est née « l'ingénierie linguistique ».

La linguistique appelée aussi sciences du langage, est l'étude scientifique des langues naturelles de l'espèce humaine.

Pour distinguer la langue humaine, on parle actuellement des « langues naturelles», donc la langue naturelle est un langage humaine, et contrairement aux « langues artificielles » que sont les langages de programmation informatique tels que: langage C , Pascal ,Delphi ,java ,.. etc.

Le traitement automatique du langage naturel (TALN, ou NLP en anglais) est un domaine multidisciplinaire impliquant la linguistique, de l'informatique et de l'intelligence artificielle, a pour objective de produit des applications, ses programmes et beaucoup de techniques informatiques, au service du langage humain en vue d'appréhender le sens des données en langage naturel.

Dans ce chapitre, on va présente un état des lieux concernant le traitement automatique du langage naturel à travers son objectifs, historique les différents niveaux de traitements, et aussi les domaines d'applications.

2. Définition (TALN)

Le traitement automatique des langues naturelles (TALN) est un domaine à la frontière de la linguistique et l'informatique, il a pour objectif de développer des logiciels capables de traiter de façon automatique des données linguistiques exprimées dans une langue naturelle donnée et pour une application bien définie. Cet objectif passe nécessairement par l'explicitation des règles de la langue puis les représenter dans un formalisme calculable et enfin les implémenter à l'aide des programmes informatiques [1].

3. Objectifs du TALN

L'objectif du traitement automatique des langues est la conception de programmes capables de traiter des données exprimées dans une langue naturelle pour lesquels plusieurs phases d'analyse (morphologique, syntaxique, sémantique et pragmatique) sont nécessaires afin dent extraire des informations.

Avec l'avènement des documents électroniques, des quantités phénoménales d'informations sont générées. Cette montée en volume de textes nécessite la production d'outils informatiques performants dont la tâche est de trouver et d'extraire l'information pertinente sous une forme condensée [2].

4. Histoire du Traitement automatique des langages Naturelles

Historiquement, Le traitement automatique du langage naturel (TALN) est né à la fin des années quarante dans un contexte scientifique imprimé par les premiers travaux sur la traduction mais aussi dans un contexte politique [3], qui peut s'expliquer par la fin de la Seconde Guerre mondiale et le début de guerre froide entre d'une part les États-Unis d'autre part l'Union des républiques socialistes soviétiques (URSS).

Les dates suivantes représentent certaines dates Marquantes dans l'histoire du traitement du langage naturel TALN :

- **1947** : Début des travaux sur la traduction automatique.
- **Entre 1951 et 1954** : Zelling Harris publie ses travaux les plus importants de la linguistique (linguistique distributionnaliste) .
- **1954 :** La mise au point du premier traducteur automatique (très rudimentaire) qui traduit du Russe à l'Anglais.

- **1956** : L'école de Dartmouth (au Etats-Unis) et la naissance de l'Intelligence Artificielle (I.A) sous l'influence de plusieurs figures marquantes de cette époque : J. McCarthy, Marvin Minsky, Allan Newell et Herbert Simon qui discutent sur les possibilités de créer des programmes d'ordinateurs qui se comportent intelligemment et en particulier qui soient capables d'utiliser le langage naturel .

- **1957** : **N. Chomsky** publie ses premiers travaux sur la syntaxe des langues naturelles , et sur les relations entre grammaire formelles et grammaire naturelles .
- **1962** : la première conférence sur la traduction automatique est organisée au MIT (Institut Technologique du Massachussetts) par Y. Bar-Hillel .
- **Entre 1961 et 1966** : beaucoup d'applications ont été mis en place tel que : BASBEL, SIR, STUDENT, ELIZA, …etc. Mettant en oeuvre des mécanismes de traitement simple, à base de mots clés .
- **1966** : L'histoire du TAL fait souvent celle des rendez-vous manqués et des désillusions cruelles Parmi ces faits marquants, on peut citer le rapport de la commission ALPAC (Automatic Language Processing Advisory Committee) en Anglais qui s'interroge sur l'utilité de poursuivre les recherches dans ce domaine.
- **1968**: le premier (vrai) système de traduction automatique commercial nommé SYSTRAN pour objectif d'améliorer la traduction du russe en anglais.
- Dès lors, les crédits sont considérablement réduits et la recherche stagne jusqu'au début des années 70.
- **Depuis 1970**, la plupart des recherches visent surtout la sémantique dans le cadre de la compréhension, mais aussi en parallèle les modèles syntaxiques connaissent en informatique des développements et des raffinements continus, et des algorithmes de plus en plus performants sont proposées pour analyser les grammaires les plus simples.
- **1971**: un système intelligent en mode fermé (SHRDLU c' est un programme qui permet d'interagir avec un robot dans un monde de blocs).
- **1972** : Terry Winogard, réalise le premier logiciel appelé SHRDLU capable de dialoguer en anglais avec un robot.
- **1976** : le système de traduction METEO est un système de traduction automatique conçu spécifiquement pour la traduction des bulletins météorologiques émis quotidiennement par Environnement Canada. Ce système fût développé par John Chandioux.

- **80s**: Approches symboliques. Applications utilisent des connaissances linguistiques et encyclopédiques extensives. Manquent de robustesse.
- **90s** : Premiers corpus, approches statistiques apprentissage automatique. Applications utilisent corpus de grande taille et méthodes statistiques
- **2000s** : Utilisation du World Wide Web comme corpus

5. Domaines d'application du TALN

Concernant les applications du TAL qui sont de nombreuses et variées, en peut les regroupées ces applications en trois grandes tâches, qui correspondent aux aides à la lecture de documents, aux aides à la production de documents, et enfin aux interfaces homme-machines. Ces tâches sont détaillées dans les paragraphes qui suivent : [4]

5.1. Les tâches liées à la gestion de documents ou de bases documentaires

- Traduction automatique (ou l'aide à la traduction automatique).
- La recherche de documents dans des bases documentaires.
- Le routage, classement ou l'indexation automatique de documents électroniques
- Le Résumé automatique.
- Recherche et extraction d'information.
- Plus complexe est la tâche de trouver (ou de produire à la demande) des réponses précises aux questions de l'utilisateur (tâche de "question-réponse").
- L'analyse d'un corpus de documents relatifs à un thème donné (histoire, veille technologique, etc.).

5.2. Les tâches de production ou d'aide à la production de documents

- Correction orthographique ou de syntaxe.
- Intégrée à toute application informatique impliquant la rédaction
- La génération automatique de documents à partir de spécifications formelles.
- La reconnaissance optique de caractères (ROC, OCR pour Optical Character Recognition en Anglais).
- Apprentissage assisté par ordinateur des langues naturelles.

5.3. Les tâches liées à la conception d'interfaces homme-machine

- Agents dialoguant par téléphone.
- Assistants virtuels.

- Reconnaissance de la parole ou commande vocale (Reconnaissance vocale de Windows, Systèmes de navigation routière GPS, Smartphone...).
- Synthèse de la parole (Créer de la parole artificielle à partir d'un texte quelconque).
- Génération et gestion de dialogue.
- l'interrogation en langage naturel de bases de données (traduction langage naturel--SQL) ou de moteurs de recherche sur web.

6. Les Différents niveaux de TALN

Nous introduisons dans cette section les différents niveaux d'analyses (traitement) nécessaires d'une séquence de chaînes de caractères (texte) pour parvenir à une compréhension complète d'un énoncé en langage naturel.

Pour cela, on distingue quatre (04) niveaux de traitement où chaque niveau a une tâche bien précise : d'une analyse morphologique, d'une analyse syntaxique, d'une analyse sémantique et enfin d'une analyse pragmatique.

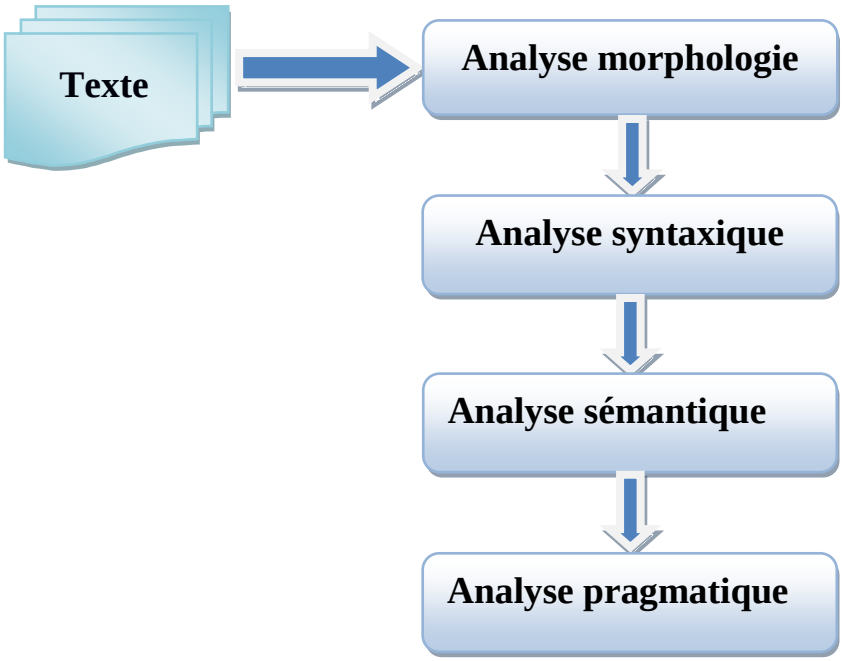

Figure 1 : Les différents niveaux d'analyse d'un texte

6. 1 Analyse morphologie

La morphologie est le domaine de la linguistique qui traite de la structure interne des mots. [5]

Selon l'approche classique, celle de la linguistique structurale, les mots sont formés de morphèmes qui sont les unités linguistiques minimales (c'est-à-dire non décomposables) porteuses de sens. [5]

L'analyse morphologie (morpho-lexicale) prend en charge l'identification des mots du texte (simples, composés, noms propres, abréviations) et leurs traits (genre et nombre). L'analyse morpho-lexicale se décompose en trois étapes :[6]

1- La segmentation, dont le but est de découper le texte en phrases puis en mots distincts.
2- La lemmatisation, qui permet de déterminer les règles et les formes canoniques qui régissent les mots séparés dans l'étape précédente.
3- L'étiquetage, dont l'objectif est d'identifier l'adéquate catégorie morphosyntaxique (verbe, nom…) des mots selon le contexte.

Cette dernière étape est considérablement importante car elle conditionne le processus de l'interprétation du texte. Il se peut qu'elle soit parfois équivoque, car, il y a parfois confusion quant à l'attribution d'une catégorie à un mot. Prenons l'exemple de l'expression « les laissez-passer » ; doit-on l'étiqueter comme un verbe ou comme un nom.

6. 2 Analyse syntaxique

C'est une partie de la grammaire qui traite la manière dont les mots peuvent se combiner pour former des propositions et de l'enchaînement des propositions entre elles. Cela consiste à associer, à la chaîne découpée en unités, une représentation des groupements structurels entre ces unités ainsi que des relations fonctionnelles qui unissent les groupes d'unités. [7]

Exemple : soit la chaîne de caractères suivant : «Ahmed a mangé des pommes», et sa représentation morphologique: U1= ahmed, U2 = a mangé, U3 = des, U4 = pommes.

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

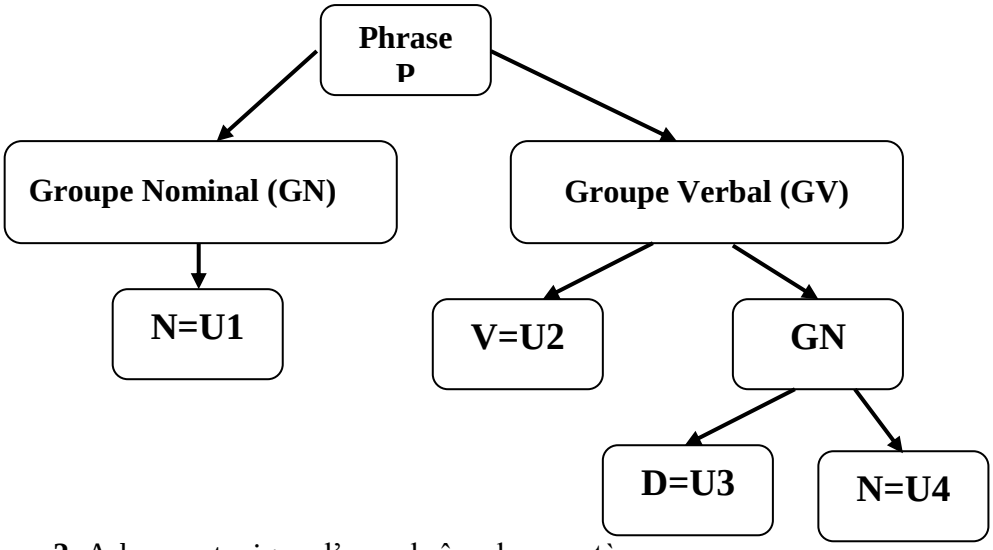

Figure 2: Arbre syntaxique d'une chaîne de caractères

.

P = « Ahmed a mangé des pommes »
GN = Ahmed
GV = a mangé des pommes
GN = des pommes
N = U1= Ahmed
V = a mangé
D = des
N = U2= pommes

6. 3 Analyse sémantique

Le niveau sémantique est encore beaucoup plus complexe à décrire et à formaliser que les niveaux de traitement précédents, par conséquent les réalisations qui sont opérationnelles sont peu nombreuses et elles concernent des applications très limités ou l'analyse sémantique se réduit à un domaine parfaitement circonscrit ; par contre on est encore loin de savoir construire en grandeur réelle des analyseurs sémantiques généraux qui couvriraient la totalité de la langue et seraient indépendants d'un domaine d'application particulier. [8]

Le traitement sémantique prend comme unité d'analyse la phrase et conduit à représenter sa partie significative, les phrases dont l'analyseur sémantique doit décrire le sens, se composent d'un certain nombre de mots identifiés par l'analyse morphologique. Permet de rejeter une phrase qui est correcte syntaxiquement mais qui n'a pas de sens, elle s'appuie sur la notion de catégorie sémantique. [8]

Exemple: « Le courrier lit mon père »

Une phrase correcte syntaxiquement est fausse sémantiquement.

6. 4 Analyse pragmatique

Ce type de traitement permet de lever les ambiguïtés qui ne peuvent pas être éliminées par le traitement sémantique, à cause de certains problèmes ayant un lien avec le contexte dans lequel la phrase est prononcé (donner un sens au mot par rapport au contexte dans lequel il se trouve), c'est-à-dire, il se charge de placer le mot dans le contexte de l'ensemble des connaissances en faisant recours à des informations hors-contexte (géographie, sport, travail,…etc.). [9]

7. Conclusion

Dans ce chapitre nous avons défini le TALN comme étant un domaine à la frontière de la linguistique et l'informatique, dont l'objet est la création de programmes informatiques capables de traiter de façon comparable a celle de l'être humain, et nous avons présenté les différents niveaux de traitement d'une langue naturelle et les domaines d'application existent dans cette discipline.

Dans le chapitre qui suit, on met la lumière sur la langue Arabe en se basons sur les propriétés morphosyntaxique de la langue l'arabe est sa position par rapport au processus d'automatisation. En prend compte les flexions et les dérivations ainsi que les autres mécanismes particulier a la langue Arabe

Chapitre II

Traitement automatique de langage arabe

Chapitre II

Traitement automatique de langage arabe

1. Introduction

Aujourd'hui la langue arabe (classique ou standard) est la langue officielle de 23 pays. Elle fait partie de la famille des langues sémitiques et est utilisée par plus de 377 millions de personnes (selon le Fonds des Nations Unies pour la Population UNFPA pour l'année 2020) en Afrique et en Asie. L'arabe est aussi la langue référentielle pour plus d'un milliard musulmans autour de monde.

L'Assemblée Générale de l'ONU a décidé, en audience plénière n° 2006 le 18 décembre 1973, de l'introduction de la langue arabe au même titre que les autres langues officielles au sein de l'assemblée et ses principales commissions. ce qui la place au 6ème rang dans le monde, derrière l'espagnol et le russe et nettement devant le français et l'allemand. C'est, depuis 1974, la 6ème langue des Nations Unies.[10]

L'alphabet arabe est le deuxième système d'écriture utilisé dans le monde après l'alphabet latin. Où nous trouvons un grand nombre de peuples non-arabes, à transcrire leur langue à l'aide de son alphabet .nous pouvons mentionner:

Les langues *persanes* (le persan et le *balouch* en Iran, le *pashto* et le *dari* en Afghanistan, le *kurde* en Syrie, en Irak et en Iran) et les langues *indo-aryennes* (*l'ourdou* et le *sindhi* au Pakistan, le *kashmiri* en Inde).

Les langues turques, D'autres langues ont connu l'alphabet arabe comme le *swahili* en Afrique de l'Est, le *somali* en Somalie ou encore le *malais* en Indonésie.

Dans ce chapitre, nous commencerons par présenter les caractéristiques de la langue arabe. Ensuite, nous détalions sur la Morphologies arabe Puis, nous présenterons le mécanisme de dérivation des verbes et noms et les Inflexion des verbes et noms. Enfin,

nous donnerons un aperçu sur les problèmes du TALArab, les approches et les ressources linguistique utilisé dans le domaine du traitement automatique de la langue arabe.

2. Caractéristiques de la langue arabe

La langue arabe véhiculaire divisée est divisée en Arabe Classique (AC) et Arabe Standard Moderne (ASM), la première est la langue des textes saints de l'islam : le Coran et le Hadith et du patrimoine culturel, littéraire et scientifique de la civilisation arabo-musulmane.

Cependant l'ASM est la langue officielle du monde arabe actuellement, elle est utilisée dans l'enseignement et dans les médias.

2.1 L'alphabet arabe

L'alphabet de la langue arabe compte 28 consonnes. Cet alphabet contient 25 consonnes et 3 voyelles longues « ا » , «و» et « ي » . L'arabe s'écrit et se lit de droite à gauche les lettres changent de forme de présentation selon leur position (au début, au milieu ou à la fin du mot). Par exemple la lettre ع ,(عـ ، ـعـ ،ـع) et se compose de deux familles contenant le même nombre de consonnes :

- Familles Solaires : contient 14 consonnes.
- Familles Lunaires : contient 14 consonnes

Familles Solaires	ا ب ج ح خ ع غ ف ق آ ه م و ي
Familles Lunaires	ت ث ذ د ر ز س ش ص ض ط ظ ل ن

Tableau 1: Classification des consonnes arabes.

2.2 La diacritisation

Un mot arabe s'écrit avec des consonnes et des voyelles. Les diacritiques (voyelles) sont ajoutées au-dessus ou au-dessous des lettres (ً ٌ ٍ َ ُ ِ ّ ْ) comme le montre le Tableau 2 [11]. Elles sont nécessaires à la lecture et à la compréhension correcte d'un texte, elles permettent de différencier des mots ayant la même représentation.

Voyelle	Nom	Transcription
ـَـ	**فتحة** fathatun	a
ـِـ	**كسرة** kasratun	i
ـُـ	**ضمة** dammatun	ou
ـْـ	**سكون** sukûnun	-

Tableau 2: Les voyelles en langage arabes

Notons que, le concept de "Tanwin" considéré par quelques auteurs comme étant le double de deux voyelles, peut être sous trois formes (ٌ, ً, ٍ) qui sont construits par dédoublement des voyelles. Il est ajouté seulement à la fin des mots indéterminés, par conséquent il n'apparait jamais avec l'article de détermination ال .Le signe du tanwin « ٌ » (à l'accusatif) est suivi toujours par ا *alif*. [11]

2.3 Le mot arabe

La langue arabe recèle 12 302 912 mots sans répétitions alors que la langue anglaise possède 600 000 mots[12].et aussi Le nombre de racines dans la langue arabe est d'environ 6000.

Un mot arabe est un mot qui répond aux deux conditions suivantes : [13]
1- Tous ses caractères sont alphabets arabes nus ou avec diacritiques.
2- Il appartient à l'une des deux catégories suivantes :
a- Les mots d'origine arabe : sont divisés à leur tour en deux sous-catégories :

- ✓ **Mots arabes dérivés :** Ce sont les verbes et les noms qui sont construits selon les règles de dérivation arabe.

- ✓ **Mots arabes fixes :** Ce sont un ensemble de mots moulés par les Arabes, autrefois, et ne respectent pas les règles de dérivation arabe. La plupart de ces mots ne sont ni fixes ni les noms des verbes, la plupart d'entre eux sont des mots fonctionnels (particules) comme les pronoms, les prépositions, conjonctions, mots interrogatifs, et autres. Ils peuvent être considérés comme la meilleure colle qui lie les mots de la phrase arabe ensemble. La catégorie des mots arabes fixes contient un nombre limité.

b- Les mots arabisés les mots arabisés sont des noms empruntés à des langues étrangères (peut-être avec un certain ajustement phonétique pour convenir à la prononciation arabe) et sont devenues communs.

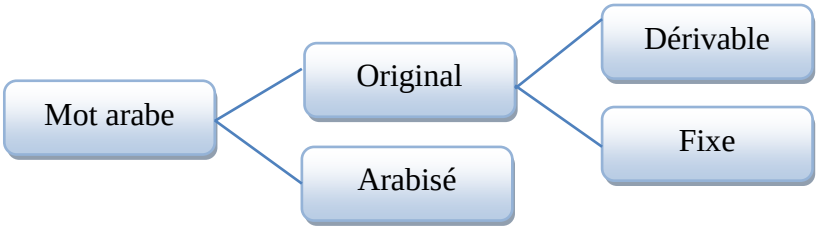

Figure 3 : Classification du mot arabe. [13]

3. Morphologies arabe

Comme nous l'avons dit précédemment. La morphologie est l'étude de la structure interne d'un mot. Elle permet de connaître comment le mot est formé (morphologie dérivationnelle) ainsi que les différentes variantes qu'il peut subir dans la syntaxe (morphologie flexionnelle) [14].

Les grammairiens arabes classiques ont adopté trois catégories pour classifier les mots de la langue Arabe, à savoir les noms, les verbes et les particules.

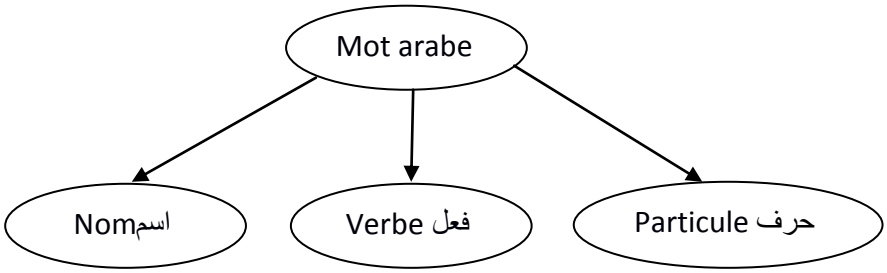

Figure 4: classification des mots de la langue Arabe selon la catégorie grammaticale

a- Le verbe (الفعل)

Entité exprimant un sens dépendant du temps (partir,ذهب), c'est un élément fondamental auquel se rattachent directement ou indirectement les divers mots qui constituent l'ensemble. Chaque verbe est donc l'origine d'une famille de mots. La conjugaison des verbes dépend de plusieurs facteurs: [15]

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

b- Le nom (الاسم)

Toute unité lexicale référant à un sens indépendant du temps, regroupent : Les adjectifs ; féminin et masculin ; les noms démérites, les noms prolongés ainsi que les noms réduits ; les noms communs et les noms propres ; les pronoms et leurs types (connectés et séparés) ; les pronoms relatifs ; les pronoms démonstratifs ; les noms d'interrogations ; les noms déterminés et non déterminés ; les noms de périphrases ; les noms du verbe ; les noms de voix ; les semblables des verbes de noms [15].

c- Les particules (الحروف)

Entité invariable exprimant un sens dépendant de compréhension. La langue arabe contient un nombre limité ne dépasse pas 80 éléments, ils se nommaient en arabe les particules de sens (حروف المعاني), par contre l'alphabet arabe se nommait les particules de construction (حروف المباني) .

Les particules de sens sont de type : unitaire, binaire, tertiaire, quaternaire ou quintette, Elles jouent un rôle important dans l'articulation et l'interprétation de la phrase ainsi la cohérence et l'enchaînement d'un texte.

Les particules sont classées selon leur sémantique et leur fonction dans la phrase. Il existe deux classes selon leur fonction (active, inactive) et 31 classes de particules selon leur sens. [14]

La plupart des verbes et des noms sont les produits d'un mécanisme de dérivation.

3.1 Morphologie dérivationnelle

La majorité du vocabulaire de la langue Arabe est construite selon le formalisme racine-schème (جذر -وزن). En effet, le terme dérivation peut désigner de façon générale le processus de formation des unités lexicales. La dérivation est un processus par lequel on génère des mots à partir des combinaisons d'une racine et des schèmes.

Les Racine peuvent donner naissance à plusieurs schèmes à la suite d'une ou plusieurs transformations morphologiques, comme c'est le cas de la racine "(K+T+B) " كتب à partir de laquelle on peut générer 16 mots représentant 9 catégories grammaticales déférentes.

a-Racine :

Les grammairiens arabes ont utilisé les lettres "ع" "ف" et "ل" comme lettres génériques pour représenter la racine et les schèmes. Pour les mots dérivés, l'ordre de ces lettres est

toujours le même: " ف"est première lettre, عest la seconde et لreprésente le reste des lettres. La majorité des mots arabes sont formés par des radicaux de 3 consonnes tel est le cas du verbe "كتب" (écrire) et éventuellement 4 consonnes tel est le cas du verbe " دحرج "(glisser).

b-Schème

C'est un modèle spécifique représentant d'une façon schématique une structure de la langue ou du comportement verbal des locuteurs. Formellement parlant, un schème "S" est une combinaison des éléments vides (0) et des extensions (si); il prend la forme suivante: [16]

S= 0 s1 0 s2 …

Exemples

soit un mot M= c1 s1 c2 s2 c3, alors S = 0 s1 0 s2 0, avec ci, i=1.3: consonnes de la racine et si, i=1,2: extensions du schème.

La majorité de vocabulaire arabe est ainsi constitué par des dérivées verbales et des dérivées nominales selon des schèmes en nombre limités. La génération d'un mot à partir de la racine et de son schème, s'effectue en remplaçant les vides par les lettres qui forme la racine, selon le schéma suivant :

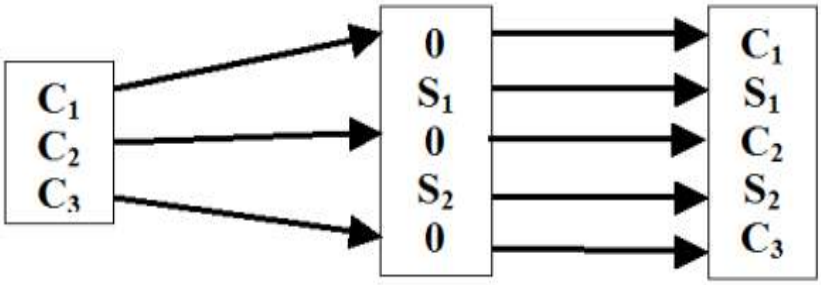

Figure 5: Schéma de dérivation [16]

c- Lemme

Le lemme est le résultat de l'opération dérivation appliquée sur un couple (racine, schème).Il porte le sens principal du mot et représente les entrées des dictionnaires. Pour les verbes, le lemme est sa forme conjuguée à l'accompli à la 3ème personne du singulier du verbe, et pour les noms, le lemme est la forme singulière masculine du nom (quand elle existe).[17]

3.1.1 -Dérivation des verbes

La dérivation des verbes en arabe se fait à l'aide d'un nombre de schèmes limité. Les verbes simples à racine trilitère (الثلاثي المجرد) sont dérivés selon trois schèmes (فَعُلَ, فَعِلَ , فَعَلَ) ,tandis que les verbes simples à racine quadrilatère (الرباعي المجرد) suivent un seul schème (فَعْلَلَ).

Si le verbe ne contient aucune lettre longue on dit qu'il est correct (صحيح *Sahihe*) et se diviser en trois types :

- Le verbe sain (*SAlim* سالم):qui ne contient aucune lettre radicale défectueuse, ni lettre hamza, ni lettre redoublée.
- Le verbe de lettre Alif (*Mahmuz* مهموز) : qui contient une lettre radical hamza comme : Interroger (*SaAla* سأل), Lire (*karaAa* قرأ).
- Le verbe redoublé (MudaEaaf مضعف) : la présence de deux consonnes identiques dans la deuxième et troisième position du radical de verbe nus trilitère et son augmenté comme : passer (*Maraa* مر) ou la première et la troisième lettre identique dans le verbe quadrilatère comme : commotionner (*Zalzala* زلزل)[18]

Sinon le verbe est défectueux et contient une ou deux lettres longues ou bien défectueuses qui causent des altérations importantes au cours de la conjugaison, ce type est distingué en 4 catégories:

- Verbe assimilé (*MivAl* ; مثال) : la première consonne est une longue voyelle, il est nommé comme ça parce qu'il a assimilé le verbe sain dans leur conjugaison au passé. Exemple : promesse *(WaEada-*وعد) [19].
- Verbe creux (*Ajwaf* ;أجوف) : la deuxième consonne est une longue voyelle, il est nommé comme ça parce que leur cavité est vidée d'une lettre saine ; par exemple :kAl(قال, dire) [18] .
- Verbe incomplet (*NaAkis* ناقص):la troisième consonne est une longue voyelle, il est nommé comme ça parce que dans leur conjugaison on supprime cette lettre comme : conquérir (*RazA,*غزا).
- Verbe Ramas (لفيف *Lafif*): il contient deux longues voyelles au même temps, il est divisé en deux selon leur position :
  - Ramas séparé (لفيف مفروق *Lafif Mafruwk*) : la première et la troisième consonne sont des voyelles longues.

- Ramas collé (لفيف مقرون *Lafif makruwn*) : la deuxième et la troisième consonne sont des voyelles longues.

Les transformations morphologiques, subies par ces verbes, produisent des verbes dits augmentés (الأفعال المزيدة) qui suivent 15 schèmes dont 12 pour les verbes à racines trilitères(اِفْعَال , اِفْعَوَّل, اِفْعَوْعَل, اِسْتَفْعَل, تَفَعَّل, تفاعَل, اِفْعَل, اِفْتَعَل, اِنْفَعَل,أفْعَل , فاعَل,فَعَّل) et 3 pour les verbes à racines quadrilatère (اِفْعَنْلَل, اِفْعَلَل,تَفَعْلَل).

Les tableaux 3 et 4 présentent les schèmes verbaux simples et augmentés.

	Les schèmes verbaux	**Exemple**
Trilitère simple (الثلاثي المجرد)	فَعَلَ	كتب (Ecrire)
	فَعِلَ	فهم(Comprendre)
	فَعُلَ	كبر(Grandir)
Quadrilatères Simple (الرباعي المجرد)	فَعْلَلَ	دحرج(Glisser)

Tableau 3 : Schèmes verbaux simples

	Les schèmes verbaux	**Exemple**
Trilitère augmenté (الثلاثي المزيد)	فَعَّل	Présenterقدم
	فاعَل	Participerشارك
	أفْعَل	Donnerاعطى
	اِنْفَعَل	se lancerانطلق
	اِفْتَعَل	Respecterاحترم
	اِفْعَل	Rougirاحمر
	تَفاعَل	Convenirتلائم
	تَفَعَّل	se présenterتقدم
	اِسْتَفْعَل	Extraireاستخرج
	اِفْعَوْعَل	اخضوضر
	اِفْعَوَّل	monter surإعلَوط
	اِفْعَال	Jaunir اصفار
Quadrilatères augmenté (الرباعي المزيد)	تَفَعْلَل	Rouler تدحرج
	اِفْعَلَل	se rassurer اطمأن
	اِفْعَنْلَل	s’entasser احرنجم

Tableau 4: Schèmes verbaux augmentés.

3.1.2 -Dérivation des noms

Les noms dérivés dans la langue Arabe comprennent les types principaux suivants :

a. Participe actif (اسم الفاعل)

Le participe actif est dérivé du verbe (à la voix active) pour désigner l'entité qui a accompli l'acte exprimé parce verbe. Ce nom est utilisé pour indiquer des actes ou des états temporaires, transitoires ou accidentels. Par exemple, le nom "كاتب"(écrivain) est un participe actif dérivé du verbe "كتب"(écrire), à la voix active, pour désigner l'entité qui a effectué l'acte d'écriture. Les participes actifs sont formés en utilisant un ensemble limité de schèmes nominaux [20]

b. Participe adjectif semi-actif (الصفة المشبهة)

Le participe adjectif semi-actif est un nom dérivé qui a le sens du participe actif. Alors que le participe actif indique un acte ou un état temporaire, transitoire ou accidentel, le participe adjectif semi-actif désigne une action continue, un état habituel, ou une qualité permanente.

Le tableau suivant montre des exemples du participe actif semi-actif dérivé du schème verbale فعل : [20]

Verbe	Participe adjectif semi-actif
كرم(être généreux)	كريم(généreux)
حسن(être beau)	حسن(beau)
جبن(être lâche)	جبان(lâche)
شجع(être courageux)	شجاع(courageux)
وقر(être sérieux)	وقور(sérieux)

Tableau 5: Les schèmes du Participe adjectif semi-actif

c. Forme d'exagération (صيغة المبالغة)

La forme d'exagération est un nom dérivé donnant la signification du participe actif, mais avec un sens supplémentaire. Elle dénote, en plus de l'exécution d'un acte, un haut degré de qualité, ou indique un acte qui est exécuté fréquemment ou de manière intensive. Ce nom dérivé est généré à partir d'un certain nombre de schème nominaux qui comprennent les formes suivantes :[20]

Schèmes	Forme d'exagération
فعال	كذّاب *(Menteur)*
فعول	غفور , شكور,*(indulgent)* , *(reconnaissant)*
مفعال	مدرار*(torrentiel)*
فعيل	خبير , *(expert)*
فعل	حذر ,*(prudent)*,فطن , *(perspicace)*

Tableau 6: Forme d'exagération

d. Le participe passif (اسم المفعول)

Le participe passif est un nom dérivé utilisé pour désigner l'entité qui a subit l'action exprimée par le verbe. Par exemple "مكتوب" *(*écrit*)* est un participe passif dérivé du verbe "كتب"*(*écrire*)* ou désigner l'entité quia été affectée par l'action d'écrire. Les participes passifs sont générés à partir des schèmes nominaux qui correspondent aux schèmes verbaux.

e. Nom d'instrument (اسم الالة)

Le nom d'instrument est un nom dérivé qui désigne l'instrument utilisé pour exécuter l'acte exprimé par un verbe. Ce nom est dérivé seulement à partir des verbes transitifs en utilisant un certain nombre de schèmes nominaux tels que ceux donnés dans le tableau ci-dessous.

Schèmes	*Nom d'instrument*
مفعل	مصْعَد *(ascenseur)*
مفعال	مفتاح , *(clés)*
مفعلة	مطرقة*(marteau)*

Tableau 7: Les schèmes du nom d'instrument

f. Nom verbal (المصدر)

Le nom verbal est un nom abstrait qui désigne un acte ou un état indiqué par le verbe correspondant, sans aucune indication de temps, de sujet ou d'objet. Il est semblable au gérondif dans la langue française. Le nom verbal issu des verbes associés aux schèmes trilitères verbaux est obtenu par l'utilisation de près de 44 schèmes de « masdar » . Dans ce contexte, chaque verbe n'est pas nécessairement associé à tous ces 44 schèmes. En effet, la majorité de ces verbes sont associés à un seul schème de « masdar », et très peu d'entre eux sont associés à deux ou à trois des 44 schèmes. Cependant, la liste des schèmes

correspondants à un « masdar » particulier ne peut être définie qu'à partir d'un dictionnaire de la langue classique.

3.2 Morphologie flexionnelle

La flexion en linguistique est une opération de dérivation, qui ne crée pas de nouveaux mots, mais qui permettant d'appliquer des modifications sur un lemme afin de dénoter des traits grammaticaux souhaités. Elle possède deux catégories : la déclinaison pour le système nominal et la conjugaison pour les verbes. Toute langue utilisant cette opération est appelée langue flexionnelle, et l'arabe en est une. En arabe, la flexion se concrétise par l'ajout des suffixes et préfixes un lemmes pour refléter des indices d'aspects, de mode, de temps, de personne, de genre, etc.

3.2.1 Inflexion des verbes

Selon les grammairiens arabes classiques, les verbes de la langue Arabe sont classifiés en trois formes : accompli (الماضي), inaccompli (مضارع) et impératif (أمر). Les verbes à l'accompli indiquent un acte achevé, tandis que les verbes inaccomplis désignent un acte inachevé qui vient de commencer ou en cours. Les verbes accomplis, inaccomplis et impératifs diffèrent dans leur inflexion selon la personne, le nombre et le genre. Cette différence apparaît dans les préfixes et les suffixes. Nous commençons par une présentation du système flexionnelle des verbes à l'accompli.[20]

- **Inflexion à l'accompli**

L'inflexion à l'accompli est réalisée en attachant aux verbes des suffixes qui indiquent la personne, le nombre et le genre. Ces suffixes sont identiques dans les deux voix : active (المبني للمعلوم) et passive (المبني للمجهول) .

- **Inflexion à l'inaccompli**

L'inflexion des verbes à l'inaccompli est obtenue en ajoutant des préfixes et des suffixes qui varient selon la personne, le nombre et le genre. Le paradigme de l'inaccompli inclut 3 modes, l'indicatif (المضارع المرفوع), le subjonctif (المضارع المنصوب) et l'apocopé (المضارع المجزوم).

- **Inflexion à l'impératif**

L'inflexion des verbes arabes à l'impératif se fait en ajoutant des suffixes aux verbes à la deuxième personne. Les tableaux suivants ressentent respectivement les suffixes à l'impératif ainsi qu'un exemple de conjugaison.[21]

Singulier tu		**Duel vous**		**Pluriel vous**	
Mas	**Fem**	**Mas**	**Fem**	**Mas**	**Fem**
	ي	ا	ا	او	ن

Tableau 8: Les suffixes de l'impératif

Singulier tu		**Duel vous**		**Pluriel vous**	
Mas	**Fem**	**Mas**	**Fem**	**Mas**	**Fem**
اُدْرُسْ	اُدْرُسِي	اُدْرُسَا	اُدْرُسَا	اُدْرُسُوا	اُدْرُسْنَ

Tableau 9: Le verbe درس conjugué à l'impératif

3.2.2 Inflexion des noms

L'inflexion des noms en langue Arabe comporte trois cas (حالة إعرابية) : le nominatif (المرفوع), l'accusatif (المنصوب) et le génitif (المجرور) suivant les trois nombres : singulier (مفرد), duel (مثنى), et pluriel (جمع). Les noms en arabe sont majoritairement déclinables (معرب) c'est-à-dire qu'ils se mettent à l'un de ces trois cas suivant leur fonction dans la phrase. Il diffère selon la nature de la forme (simple, diptote, etc.) et le nombre de celle-ci (singulier, duel ou pluriel).[21]

3.2.2.1 Inflexion du singulier (المفرد)

Les mots déclinables aux trois cas : C'est le cas le plus fréquent, il prend la voyelle (ضمة) comme une marque du nominatif; la voyelle (فتحة) à l'accusatif et la voyelle (كسر) au génitif. Quand le nom est indéfini, le tanwîn apparaît marqué respectivement par les trois signes diacritiques : «ً», «ٌ», et «ٍ ». A l'accusatif indéfini, excepté le cas des noms qui se terminent par «ة »ou par «ى », le caractère «ا » est ajouté à la fin du mot pour renforcer le *tanwîn*: par exemple, à l'accusatif indéfini, le nom كتاب *(*livre) produit كتابا *(*livre à l'accusatif indéfini) et le nom جزيرة *(*île) produit جزيرة *(*île à l'accusatif indéfini). [21]

3.2.2.2 Inflexion du duel (المثنى)

L'inflexion du duel en langue Arabe est marquée par la voyelle longue « ا » dans le cas du nominatif et par la voyelle longue « ي » à l'accusatif et le génitif. Dans le cas du duel nom indéfini ou défini par l'article, la consonne «ن » est ajoutée aux marques de déclinaison. Par exemple, le duel du nom رجل (homme) prend la forme "رجلان " (deux hommes, au nominatif) et "رجلين " (deux hommes, à l'accusatif et au génitif). [22]

3.2.2.3 Inflexion du pluriel (الجمع)

Il existe deux grandes catégories de pluriel en arabe : [22]

a- Les pluriels réguliers (الجمع السالم) Ces pluriels sont formés par l'ajout d'un suffixe au singulier sans changement de la structure du mot. Nous distinguons :

- Le **pluriel régulier masculin (جمع المذكر السالم)** au nominatif, ce pluriel prend le suffixe « ن و » et dans le cas de l'accusatif et le génitif le suffixe « ين » est ajouté. Notons que dans le cas de la définition par annexion (التعريف بالإضافة) la consonne « ن » est supprimée du suffixe. Par exemple, le singulier " مدرس"(enseignant) devient "مدرسون"(des enseignants) au nominatif et مدرسين à l'accusatif et au génitif.

- **Le pluriel régulier féminin (جمع المؤنث السالم)** le suffixe ajouté dans le cas de ce pluriel est « ات » auquel s'ajoute la voyelle «ٌ» au nominatif et la voyelle«ٍ» à l'accusatif et génitif. Par exemple, le mot " شجرة"(arbre) devient شجراتٌ (arbres) au nominatif et شجراتِ à l'accusatif et au génitif.

b- Les pluriels brisés (جمع التكسير) ces pluriels doivent ce nom aux modifications et infixations qu'ils causent par rapport à la forme du singulier, à la différence des pluriels réguliers (masculin et féminin). Les formes du pluriel brisé sont très nombreuses et généralement imprévisibles. Par exemple : le nom مفتاح (clé) se transforme pour donner les deux formes plurielles مفاتيح et مفاتح (clés). [22]

4. L'automatisation de traitement de la langue arabe

La caractéristique d'une dérivation morphologique abondante qui avec son abondance, est semi-régulière, et cette régularité rend la langue arabe éligible au traitement automatique, et dans ce cas le traitement morphologique automatique représente la composante de base de l'automatisation du Lexique arabe et le développement de systèmes

automatiques pour l'analyse grammatical automatique (نظم آلية للإعراب الآلي) et la vocalisation automatique (التشكيل التلقائي) . [23]

La particularité du morphologique arabe étant de dépendre des racines et non de l'ordre alphabétique des mots Malgré le petit noyau du morphologique arabe (moins de 10 mille racines), le vocabulaire est énorme grâce à la propriété de dérivation morphologique.

4. 1 Les approches de Traitement automatique de langage

D'un point de vue général, pour mettre en œuvre des outils du TAL (Traitement automatique de la langue) en arabe, les chercheurs peuvent avoir besoin : [24]

- Des modules de base pour la segmentation en phrases et en mots, l'analyse morphologique, syntaxique voire sémantique.
- Des ressources langagières (dictionnaires, corpus, bases de données lexicales,…) ;
- Des ressources et modules de comparaisons pour l'évaluation ;

D'utilitaires de traitement de la langue (outils de recherche de texte, outils statistiques sur les textes et corpus annotés, etc.) ;

Parmi les « modules de base », l'étiquetage morphosyntaxique constitue une étape essentielle pour réaliser la plupart des applications en traitement automatique de la langue car il permet d'identifier la catégorie grammaticale à laquelle appartiennent les mots du texte. Ainsi, les étiqueteurs constituent un module essentiel dans des applications de grand public telles que la correction grammaticale automatique, la génération automatique des résumés et le repérage d'information. Ils sont aussi très utiles dans le traitement de la parole pour réaliser des systèmes de synthèse ou de reconnaissance vocale. En général, l'étiquetage morphosyntaxique est une étape préalable dont il est difficile de faire l'économie dans la plupart des applications du TAL. [24]

4 .2 Propriétés des corpus

Comme nous l'avons dit précédemment Pour développer les outils de traitement informatique du langage, les chercheurs peuvent avoir besoin à des ressources langagières tel que corpus, alors c'est quoi corpus ? Et quelle sont leur propriétés ?

Un corpus est une collection de divers matériaux rassemblés selon un ensemble de critères afin qu'il soit représentatif et balancé. Parmi les paramètres qu'il faut prendre en

considération pour qu'un corpus soit balancé, nous citons à titre d'exemple, le genre, le domaine, la longueur, le temps et l'époque qu'il représente ainsi que le registre du langage. [25]

Les propriétés du corpus varient également selon les applications pour lesquelles il sera utilisé. Si le corpus est assemblé pour un objectif précis, il n'est pas nécessaire d'avoir des exemples de divers genres. Le plus important dans le domaine de collecte des corpus est d'inclure les spécificités de la langue que le chercheur essaie de maîtriser.

Un des premiers corpus représentatifs de l'anglais américain écrit est le Brown Corpus. Il contient environ un million de mots provenant de textes variés. Il a été assemblé en 1961.[25][26]

Concernant la langue arabe, le premier corpus annoté est celui réalisé par Khoja [27]. Il contient 50 000 mots, annotés avec les étiquettes suivantes : nom défini ou indéfini, verbe, particule, point de ponctuation ou numéro. Dans ce corpus, seuls 1 700 mots ont été annotés avec un jeu d'étiquettes détaillé utilisant le genre, le nombre et autres informations morphosyntaxiques.

Parmi les corpus de la langue arabe, annotés avec des données morphologiques et syntaxiques, on trouve : le Penn Arabic Treebank et le Prague Arabic Dependency Treebank .

5. Travaux sur l'automatisation de la langue arabe

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

En terme de travaux de référence concernant le traitement automatique de la langue arabe, il est impératif de citer l'étude pionnière de David Cohen « Vers un traitement automatique de l'arabe » qui date de 1960 concernaient notamment le lexique et la morphologie.

Dès le milieu des années 1970, les travaux de chercheurs tels que Yahya Hlal, puis ceux de Fathi Debili dans les années 1980, ont montré la possibilité d'un traitement automatique de la langue arabe. Dans les années 1990, on peut également citer aussi, toujours les travaux de Joseph Dichy notamment dans le cadre du projet européen DIINARMBC (Dictionnaire informatisé de l'arabe, multilingue et basé sur corpus) .

Depuis plusieurs travaux ont vu le jour touchants différents thématiques : l'analyse morphologique [28], l'analyse morphosyntaxique [29], la lemmatisation automatique (Khoja, Isri, JIdr, ….etc.), Résumeurs automatique, la correction orthographique, ainsi que d'autres projets.

6. Problèmes du traitement automatique de la langue arabe

L'arabe, comme toutes les langues naturelles, est caractérisée par un ensemble de phénomènes créant des difficultés et des problèmes qu'il faut prendre en considération lors d'un traitement automatique. En plus des phénomènes classiques, comme l'ambiguïté, la coordination ou l'anaphore, nous trouvons aussi dans le cas de l'arabe d'autres phénomènes propres aux langues sémitiques tel que l'absence de voyelles, l'agglutination et l'ordre des mots dans une phrase. Dans la présente section, nous présentons les phénomènes que nous considérons les plus importants pour l'arabe.[30]

6.1 L'absence de voyelle – voyellation

Nous trouvons plusieurs définitions pour décrire le phénomène de la voyellation qui est concrétisée par l'absence des voyelles courtes, appelées aussi les diacritiques, dans les textes en arabe. Cette absence génère plusieurs cas d'ambiguïté compliquant ainsi le traitement automatique. Ces ambiguïtés lexicales sont du essentiellement au fait que chaque consonne peut prendre l'une des sept voyelles de l'arabe, ce qui crée des combinaisons de mots dont le nombre diffère d'un mot non voyellé à un autre en fonction de l'existence de la combinaison obtenue dans le vocabulaire ou pas. Selon, l'absence de diacritiques en arabe entraîne une complexité de calcul d'un ordre de grandeur plus grand que la manipulation de ses homologues langues latines.[30]

Exemple :

Figure 6: Ambiguïté causée par le manque de diacritiques pour le كتب

6.2 Agglutination

L'arabe montre une forte tendance à l'agglutination : l'ensemble des morphèmes collés les unes aux autres et constituant une unité lexicale véhiculent plusieurs informations morphosyntaxiques. Ces unités lexicales sont souvent traduisibles par l'équivalent d'une phrase en français. La structure d'une unité lexicale arabe est donc décomposable en cinq éléments : proclitique, préfixe, base, suffixe et enclitique.

La base est une combinaison de lettres radicales (le plus souvent trois) et d'un schème. La base – avec préfixe et suffixe - forme le noyau lexical, éventuellement entouré d'extensions.

Comme le montre l'exemple suivant : ولِيَضربها Les éléments clitiques sont séparés par le symbole "+" :

Wa + li + ya +Dribu + haA
(COORD) + (CONJONCTION) + (V) SUBJONCTIF + (PRO)
et pour frappent elle
" et pour la frapper"

Cet exemple révèle la complexité morphologique de l'arabe. Il s'agit du verbe يضرب employé au présent du subjonctif, 3ème personne du masculin pluriel, la base verbale est /ضرب/et la racine /ضرب/. Le pronom sujet n'est pas réalisé. En position proclitique, on utilise la conjonction de coordination "wa" وet la conjonction "li" ل .En position enclitique, on utilise le pronom complément d'objet 3ème personne du féminin singulier "haA" ها"elle".[21]

6.3 Ambiguïté lexicale et syntaxique

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

Par ailleurs, le problème ne réside pas dans l'analyse d'un langage ambigu en soi; mais c'est plutôt au niveau de son traitement de façon robuste et réaliste. En effet, après une première phase de segmentation du texte en unités lexicales, il est convenu de chercher dans le lexique les interprétations correspondant à chacune d'entre elles. A chaque interprétation, nous associons une catégorie syntaxique reconnue par la grammaire.

L'un des aspects de la langue arabe qui cause cette ambiguïté, c'est le fait que beaucoup de mots en arabe sont homographiques : une même forme orthographique peut avoir des prononciations différentes. Cette homographie peut être accentuée lorsqu'elle est associée à d'autres phénomènes (absence de voyellation, morphologie flexionnelle et agglutinante, etc) ce qui donne des taux d'ambigüité assez élevés.[30]

6.4 Irrégularité de l'ordre des mots dans la phrase

La construction des phrases en arabe est flexible, dans le sens où l'ordre des mots dans une phrase donnée est relativement libre. Généralement, un mot placé au début de la phrase est un terme sur lequel nous voulons attirer l'attention, s'en suit le terme le plus long ou le plus riche en sens ou en sonorité. Cette flexibilité provoque des ambiguïtés syntaxiques artificielles due à la prise en compte de toutes les règles de combinaison possibles des composants d'une phrase. Pour illustrer cette propriété prenons les phrases suivantes :[30]

- Verbe + sujet + complément :

(- L'Algérie s'est qualifiée pour la coupe du monde)تأهلت الجزائر إلى كاس العالم

- Sujet + verbe + complément :

(- C'est l'Algérie qui s'est qualifiée en coupe du monde) الجزائر تأهلت إلى كاس العالم

- Complément + verbe + sujet :

(- C'est pour la coupe du monde que l'Algérie s'est qualifiée)إلى كاس العالم تأهلت الجزائر

7. Conclusion

Dans ce chapitre, on a présenté la langue arabe et ses propriétés morphologiques, ensuite, nous avons discuté de la possibilité d'un traitement automatique de la langue arabe puits on a présenté quelques travaux et produits de traitement automatique de la langue arabe. et finalement, nous avons mis le point sur les différents problèmes de traitement automatique de la langue arabe.

Dans le chapitre suivant, on va détaillons sur l'architecteur le système TALArab proposé et également sur les choix de conception qu'on opté pour l'utilisation de UML comme langage de description.

Chapitre III

Conception et Architecture du système

Chapitre III

Conception et Architecture du système

1. Introduction

L'importance des outils de traitement automatique de la langue arabe a considérablement augmenté dans la dernière décennie en raison de développement énorme du contenu numérique arabe sur internet.

Ce fait augmente l'importance de créer les outils de traitement automatique qui peuvent traiter ce contenu.

L'analyse morphosyntaxique est une étape importante dans le traitement automatique de l'arabe

L'arabe est une langue dont la morphologie est riche comparée à d'autres langues, et elle est langue particulière, basé sur la morphologie dérivative et flexionnelle on plus de la richesse lexicologique et le champ sémantique

La phase de la conception reste une étape essentielle pour n'importe quel système en informatique. Dans le chapitre actuel, on propose une conception pour notre système et ses différents modules de traitement, et Nous rappelons dans la suite le fonctionnement de processus la dérivation, la flexion pour la génération des mots de la langue arabe.

2. Classification du mot arabe

Le mot en langue arabe peut être vu comme une occurrence d'un nœud d'un graphe de dépendances représentant une organisation hiérarchique des classes existant pour les rôles syntaxiques comme suit: le mot est appelé en arabe «Kalima», ce modèle comprend les trois nœuds qui sont «فعل» (verbe), «اسم» (nom) et «حرف» (propositions, conjonctions, etc.). Et tous ces termes ont une dépendance avec d'autres pour créer la hiérarchie des concepts.[31]

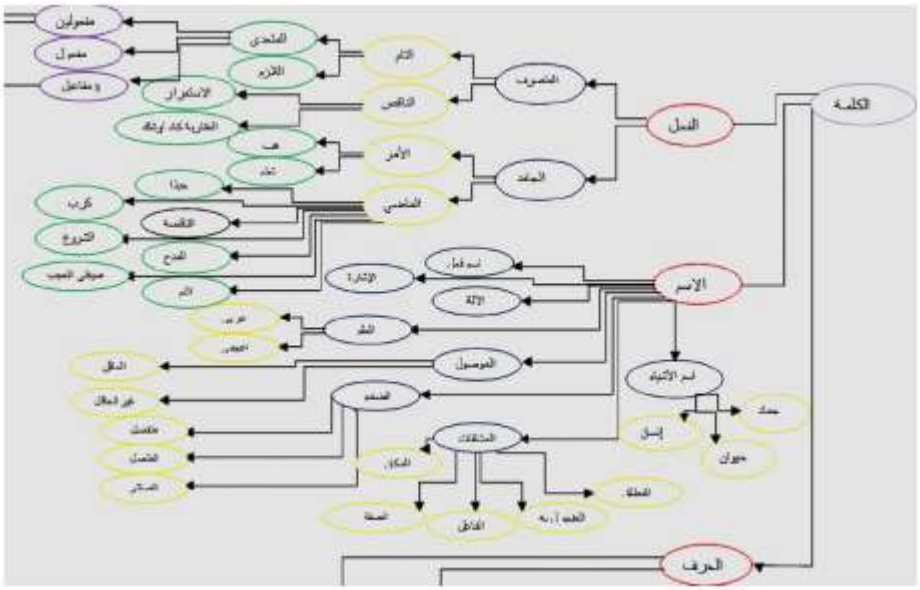

Figure 7: Classification du mot arabe sous forme d'hiérarchie de concept.[31]

La relation représentée par les arêtes est «**IS A**» qui est interprétée comme une subsomption dans la terminologie des DL ou un sous-ensemble dans l'interprétation des DL en théorie des ensembles en considérant les concepts comme un ensemble des individus.[31]

3. Le formalisme utilisé (SHA)

SHA (Subsumption hierarchical attribute) : est un formalisme qui permet de rendre la structure de représentation une symbolisation manipulable et significative en même temps. Apres une représentation sur un graphe de dépendance basé sur la relation de subsomption.[32]

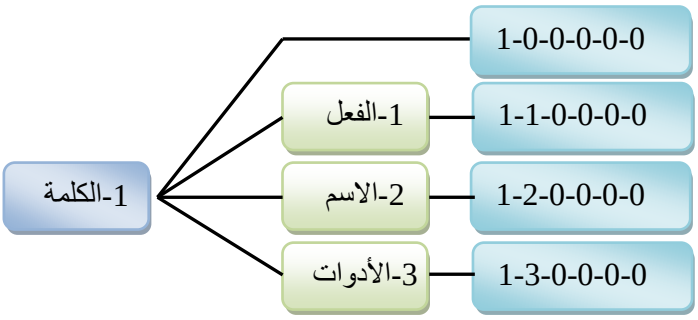

Figure 8: Classification du mot arabe avec sa vecteur SHA

Le verbe est en relation avec ces héritiers comme le montre la figure suivante :

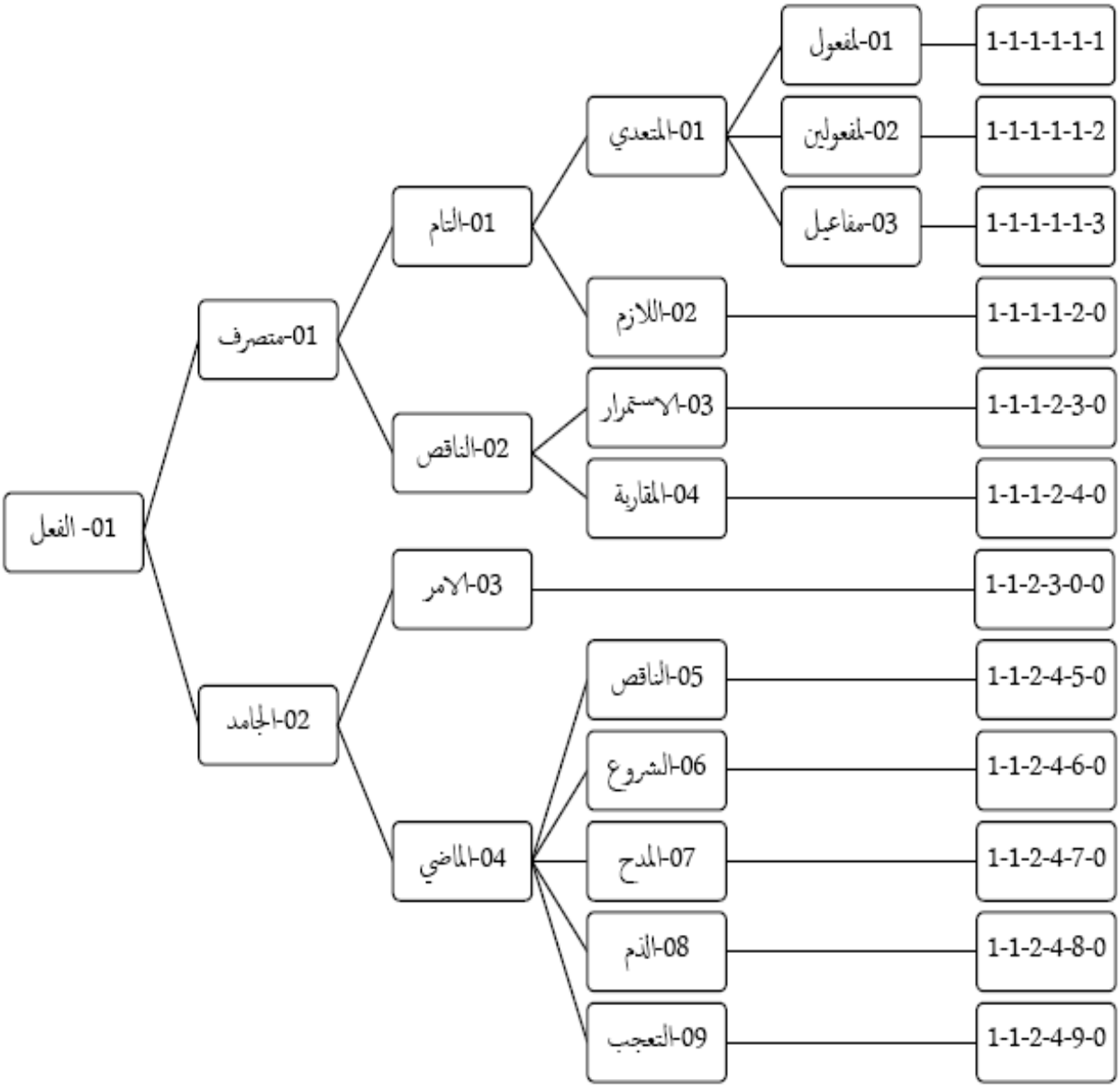

Figure 9 : Classification du verbe arabe.

De la même façon en peut décrire le nom comme dans la figure qui suit :

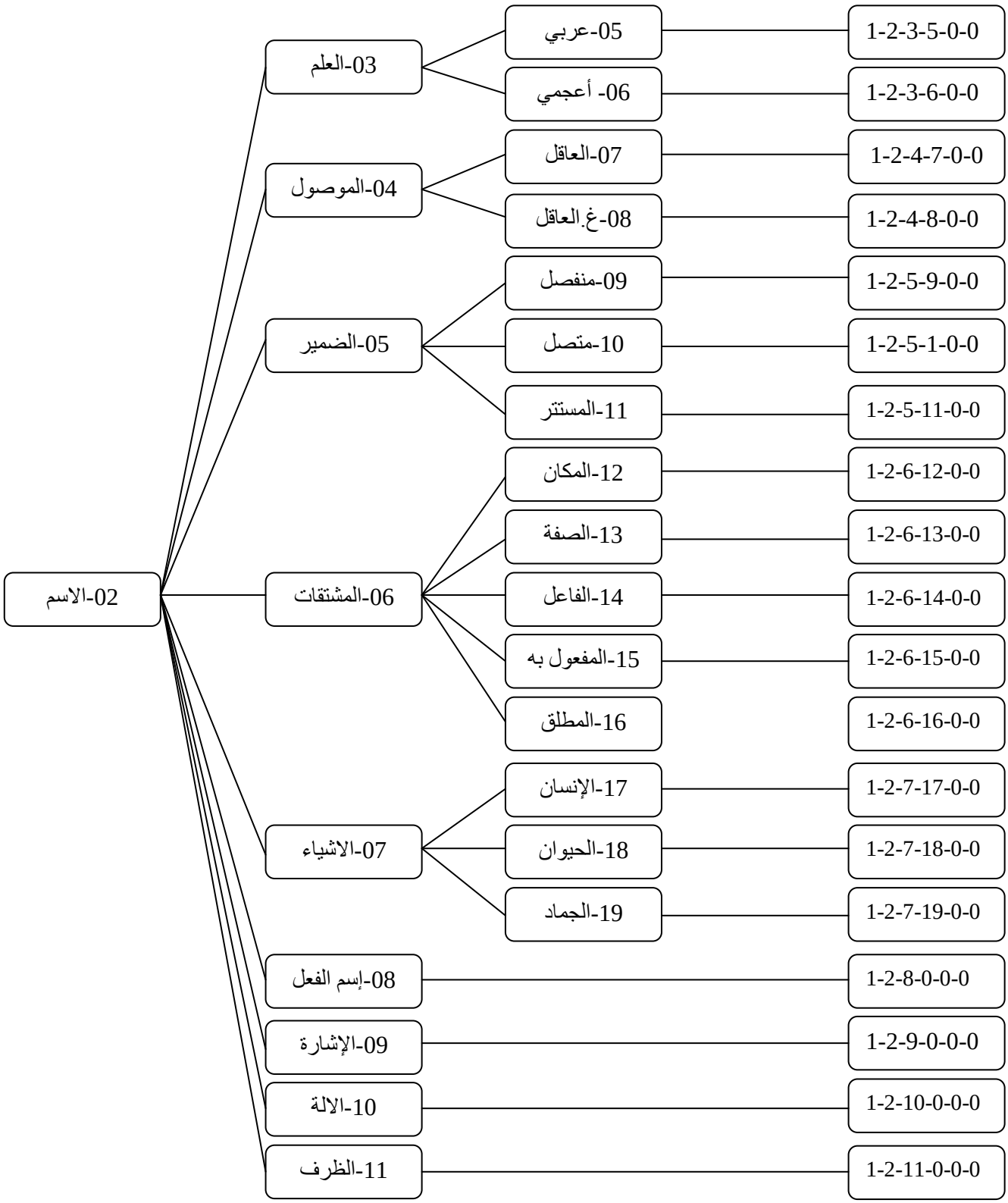

Figure 10 : Classification du nom arabe.

De la même façon en peut décrire la particule comme dans la figure qui suit :

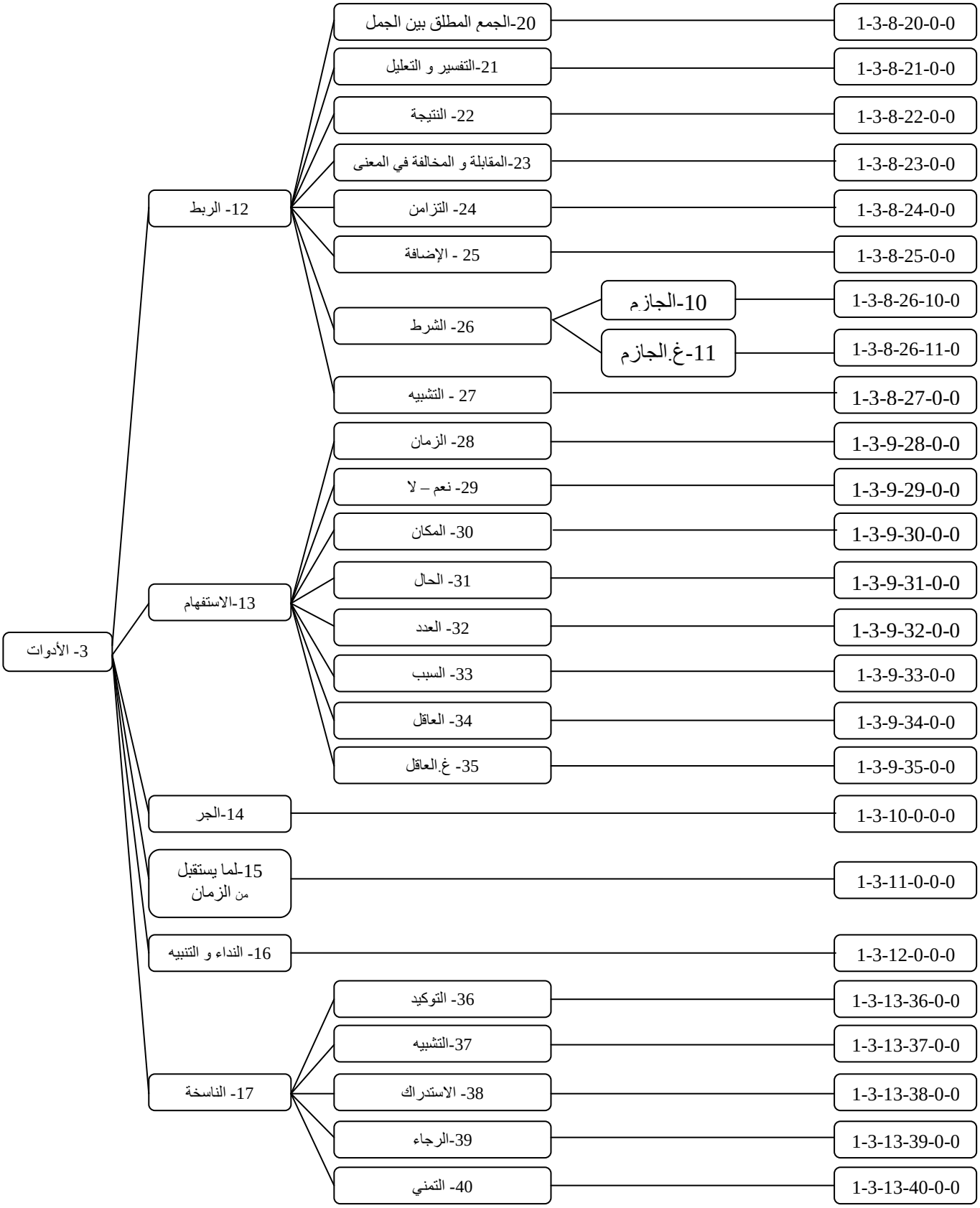

Figure 11 : Classification de la particule arabe.

4. Le Architecture globale du système

Notre système consiste à la découpé des textes arabe en mots (tokens), puis a partir de expert de la langue arabe en produisons des ressources lexicales arboré a la forme des racines reliés avec tous leurs dérivés, afin de construire une base de connaissances lexicales.

La liste des mots et leur dérivation doit être définis et étiqueter par leurs vecteurs appropriés, obtenu lors de l'étape de l'analyse lexicale, dont chaque un de ces vecteurs représente une catégorie de mot arabe,

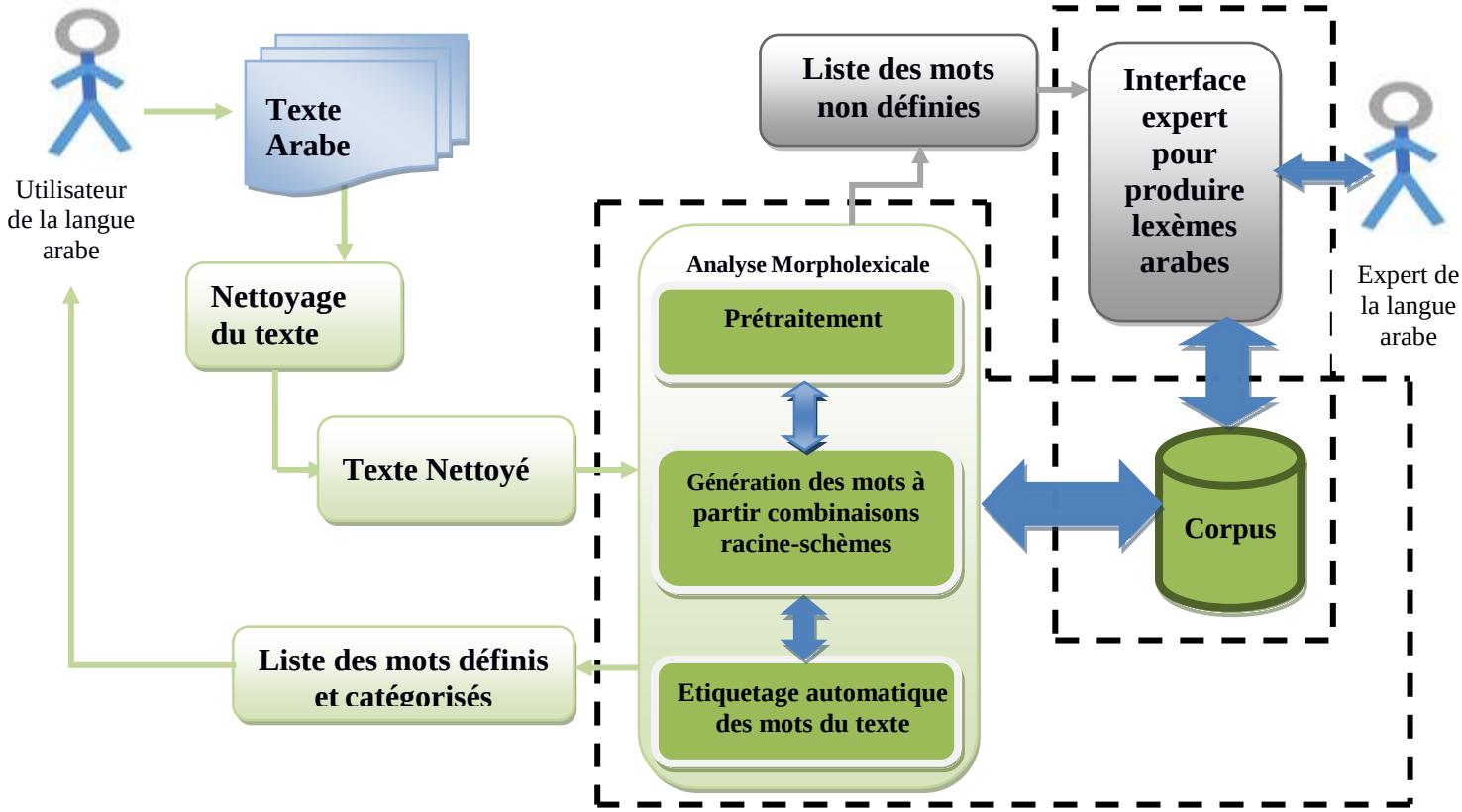

Figure 12: Architecture globale du système morphosyntaxique.

4.1 Le démarche du système

Le texte concerné par l'analyse de ce système est concédé comme flux de caractères, on a proposé l'exécution suivante :

a. Nettoyage du texte

Consiste a supprimé tous les caractères ce qui diffèrent de la lettre arabe

b. Analyse lexicale

Elle admet deux processus :

- ✓ **Prétraitement**

Est une phase de préparation du texte pour la découpassions en des mots , et de vérification de l'existence des mots de texte dans le corpus, dans le cas où il y a lieu à des nouveaux mots, on donne la possibilité a un expert de la langue arabe d'enrichir le corpus à l'aide d'un interface expert dont les mots peuvent être identifier et catégoriser.

- ✓ **Etiquetage**

Consiste a symbolisé les mots de texte par leur vecteur approprié selon le corpus, afin de donner la possibilité au système de sélectionné le rôle syntaxique de chaque mot du texte.

c. Analyse Morphologie

- ✓ **génération des mots à partir combinaisons racine-schèmes**

C'est la dérivation de la racine des mots avec l'utilisation des schèmes des verbes ou mots connus. Donc la dérivation est un processus par lequel on génère des mots à partir des combinaisons d'une racine et des schèmes.

L'avantage de cette phase est offre la proposition automatique de toutes les possibilités de dérivation d'un racine, grâce à combinaisons de ce dernier avec les schèmes, a fin de riches notre corpus avec nouveaux des mots de façon automatique. et dans ce cas le traitement morphologique automatique représente la composante de base de l'automatisation du Lexique arabe et le développement de systèmes automatiques pour l'analyse.

5. Conception UML

Dans cette partie on va présenter une conception de notre système en utilisant UML (Unified Modeling Language) que l'on peut traduire par " langage de modélisation unifié " qui est un formalisme de modélisation objet. Ce langage de modélisation est né de la fusion de plusieurs méthodes existant auparavant, et est devenu désormais la référence en terme de modélisation objet.

5.1 Diagramme de classe

Le diagramme de classe est l'un des diagrammes structurels d'UML. Utilisé pour présenter les classes et les interfaces des systèmes ainsi que les relations entre elles, ce diagramme fait partie de la partie statique d'UML .

L'intérêt de concevoir ce diagramme est d'avoir plus d'éclaircissement à la solution du problème posé,

La figure ci-dessous illustre notre le diagramme de classe:

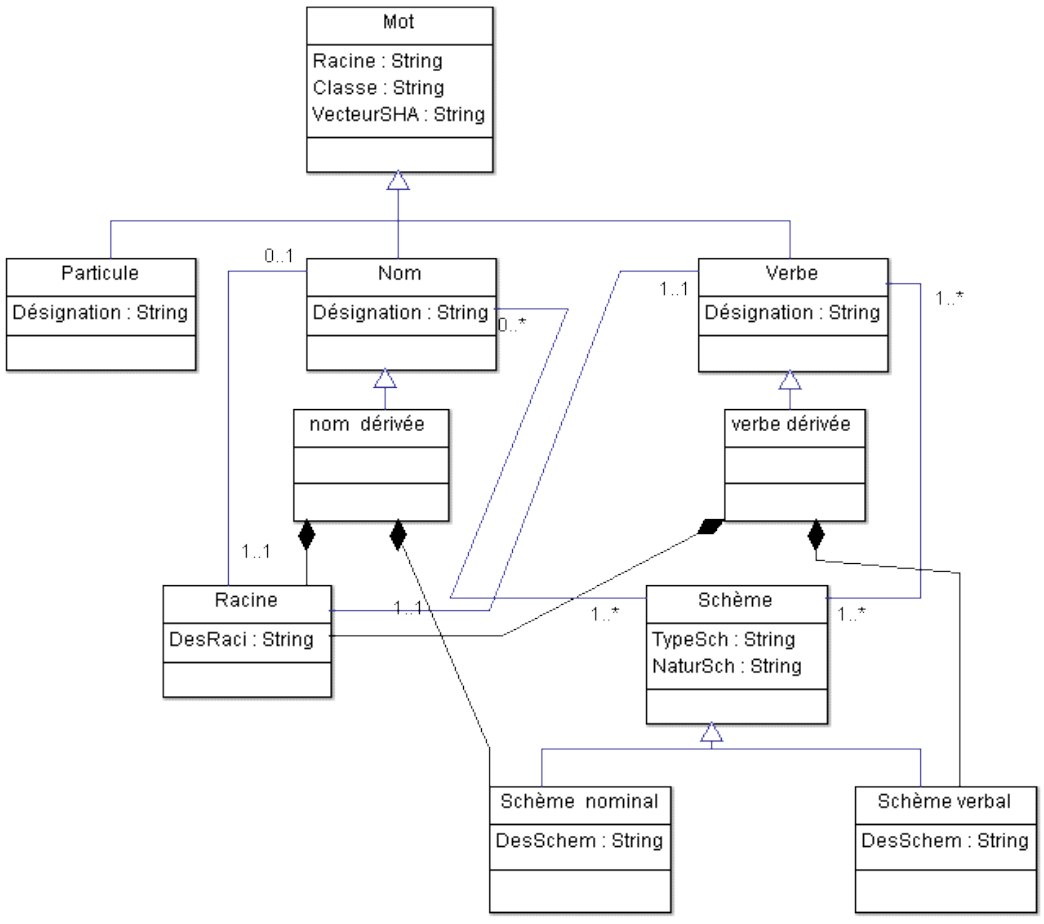

Figure 13: Diagramme de classe du modèle de base

5.2. Diagramme de séquence

Le diagramme de séquence fait parties des diagrammes comportementaux (dynamique) et plus précisément des diagrammes d'interactions, Il permet de représenter des échanges entre les différents objets et acteurs du système en fonction du temps, généralement diagramme de séquence ailleurs pour illustrer un cas d'utilisation.[33]

La figure ci-dessous illustre notre le diagramme de séquence :

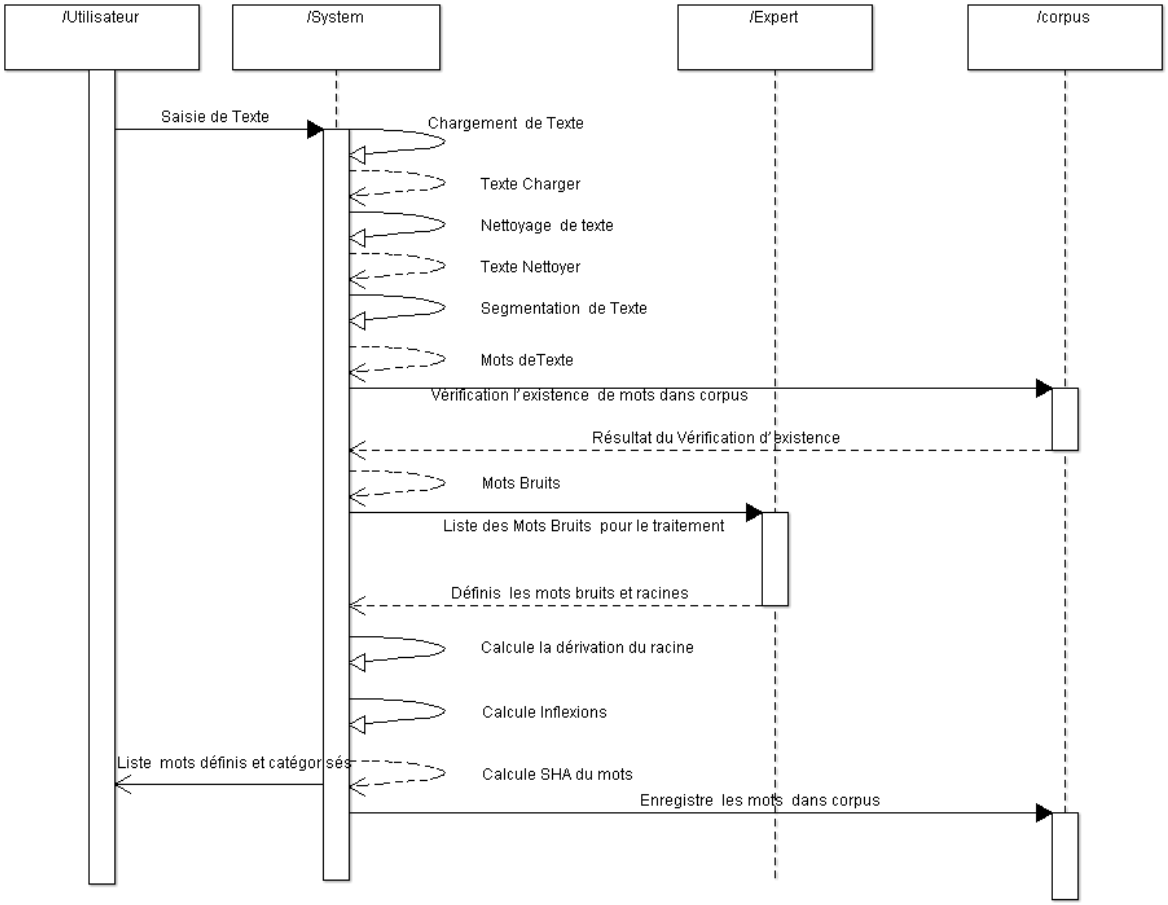

Figure 14: Diagramme de séquence.

5.3. Diagramme d'activité

Les diagrammes d'activités permettent de déterminer des traitements a priori séquentiels. Ils sont bien adaptés à la spécification détaillée des traitements en phase de réalisation. On peut également utiliser pour décrire des enchaînements d'actions, en particulier pour la description détaillée en cas d'utilisation.[34]

Un diagramme d'activité est composant de éléments suivant : activité (exécution d'actions atomiques) et Transition, Nœud (initial, final, Nœud de décision, nœud d'union,..), Couloir d'activité (les acteurs)

Le diagramme suivant met en évidence les responsabilités partagées entre les acteurs (utilisateur, system, expert) impliqués dans le processus, c'est le diagramme d'activité de définition les mots du texte et l'enregistrement sur le corpus.

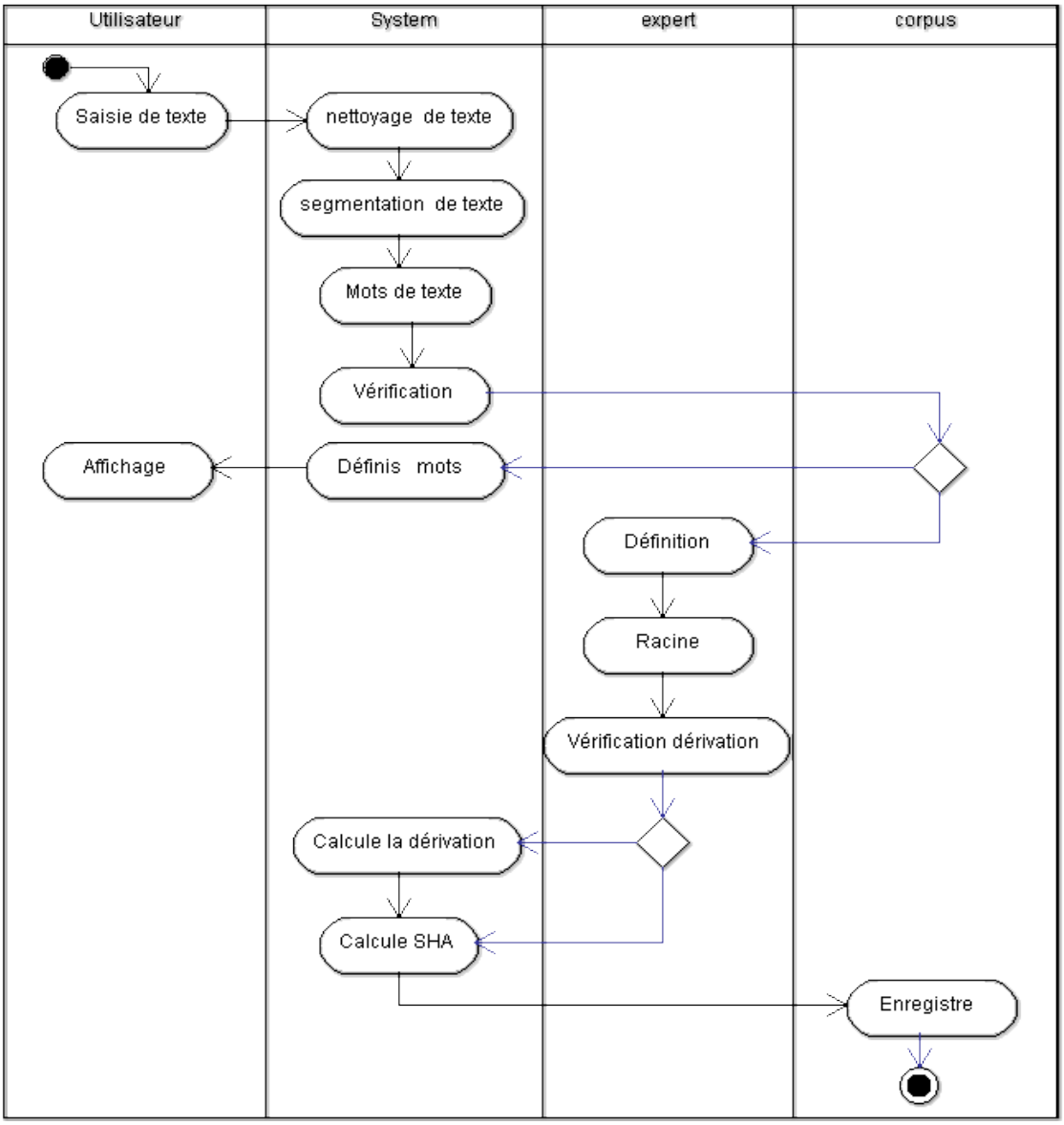

Figure 15: Diagramme d'activité de définition les mots.

6. Conclusion

Dans ce chapitre, on a présenté l'architecture du système d'interfaçage pour traitement automatique de texte arabe et on a terminé avec une conception UML qui facilite la compréhension des idées et des choix de concept qu'on a fait.

Dans le prochain chapitre on présentera l'implémentation, l'expérimentation, et l'évaluation de système

Chapitre IV

Implémentation et Résultats

1. Introduction

Après avoir expliqué dans les chapitres précédents l'aspect théorique concernant le traitement automatique de la langue arabe, il serait intéressant de franchir l'aspect pratique pour la réalisation de un module de programmation en python (version 3.4) capable de donner une interface de création de corpus d'une semi-automatique avec suffisamment d'efficience.

.

On va entamer dans ce chapitre la partie réalisation et qui a pour objectif de réaliser une interface efficace et fiable. Pour ce faire, on va commencer tout d'abord par préciser l'environnement matériel et logiciel de ce travail. Ensuite, on va présenter l'interface graphique de notre projet

2. Enivrement de développement

2.1 Langage de programmation

Le langage de programmation que nous avons adopté pour implémenter notre application est le python est un langage développé en 1989 par Guido van Rossun très utilisé dans la programmation WEB (accès aux bases de données, programmation objet), comme langage de scripts (manipulation de fichiers, administration de systèmes, configuration de machines), le calcul scientifique (librairies mathématiques)[35]

2.1.1 Pourquoi choisir PYTHON ?

- ✓ Python est multiplateforme et open source
- ✓ Parce qu'il est facile à apprendre.
- ✓ C'est un langage complet et puissant dans de nombreux domaines.
- ✓ Python permet de créer des fonctions avec moins de lignes de code
- ✓ Python dispose de l'un des gestionnaires de paquets les plus matures : PyPI

2.1.2 Les Bibliothèques utilisée dans l'application :

Dans Python, une bibliothèque est un ensemble logiciel de modules ajoutant des possibilités étendues à Python : calcul numérique, graphisme, programmation internet ou réseau, formatage de texte, génération de documents, …

Il en existe un très grand nombre, et c'est d'ailleurs une des grandes forces de Python.

Dans ce point, nous allons montrer les modules (intégrés et tiers) qui ont été utilisés pour créer notre application :

a-*Tkinter*

Tkinter est un module intégré pour le développement de GUI (interface utilisateur graphique) applications.

b-*messagebox*

Accès aux boîtes de dialogue Tk standard

c-*tkinter.filedialog*

Boîtes de dialogue standard permettant à l'utilisateur de spécifier un fichier à ouvrir ou à enregistrer, et pour l'affichage d'une boîte de dialogue d'ouverture de fichier nous avons utilisé *askopenfilename, et* pour enregistrement nous avons utilisé *asksaveasfilename.*

d-*tkinter.constants*

Tkinter fournit un *tkinter.constantsmodule* qui contient toutes les constantes (aka. Top, Left, ...).

e-*Numpy*

NumPy est un projet open source visant à permettre le calcul numérique avec Python. Elle a été créée en 2005 [36], Cela nous permet de travailler sur des tableaux multidimensionnels.

f-*CSV*

Nous avons utilisé le *CSV* (valeurs séparées par des virgules) fichiers pour stocker les données tabulaires. Python est livré avec un module appelé CSV pour gérer les fichiers classe.csv, MND.csv, sheme.csv...etc.

g- *xml.etree.ElementTree*

C'est une bibliothèque intégrée qui a des fonctions pour lire et manipuler des XML.

2.2 Environnement matériel

Pour la réalisation de ce projet, On a utilisé un ordinateur de type HP ayant les caractéristiques suivantes :

- Un microprocesseur Intel(R) CORE i2.
- Un espace de disque dur de 465 Go.
- Une RAM de 4,00 Go

3. Interface graphique d'Application

Nous allons présenter dans ce point, la description de l'interface graphique de notre application, de ce fait, on va essayer de décrire chaque outil et chaque composant de l'interface toute en mentionnant sa fonctionnalité.

L'interface graphique principale de notre application est Construite dans la figure suivante :

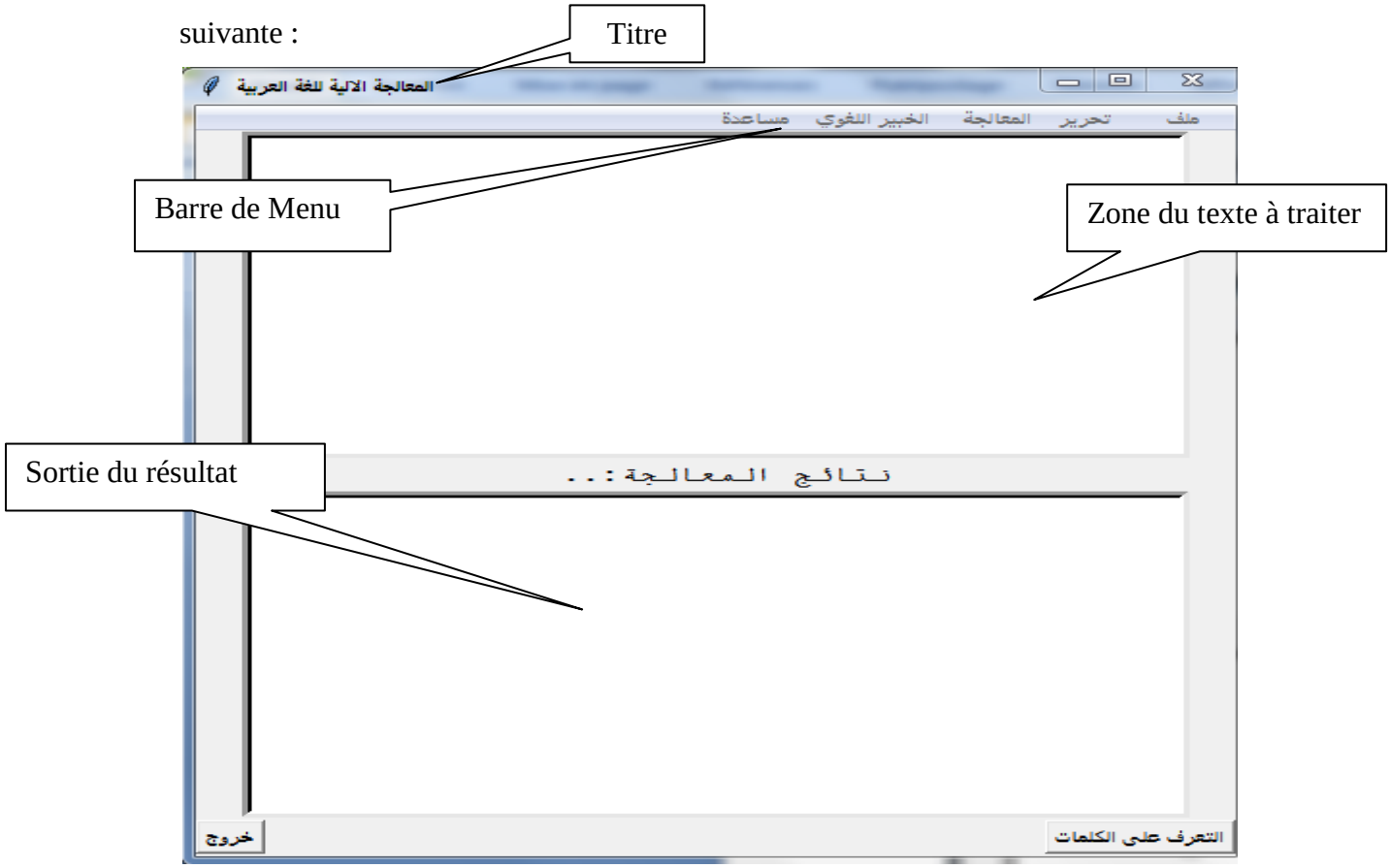

Figure 16: Interface principale d'Application

3.1. Interface d'utilisateur

Cette fenêtre offre à l'utilisateur la possibilité d'éditer le texte pour appliquer le différent traitement connus. Et tous les outils d'éditeur de texte sont également pris en compte (sous menu de ملف et تحرير), voir le Figure 18.

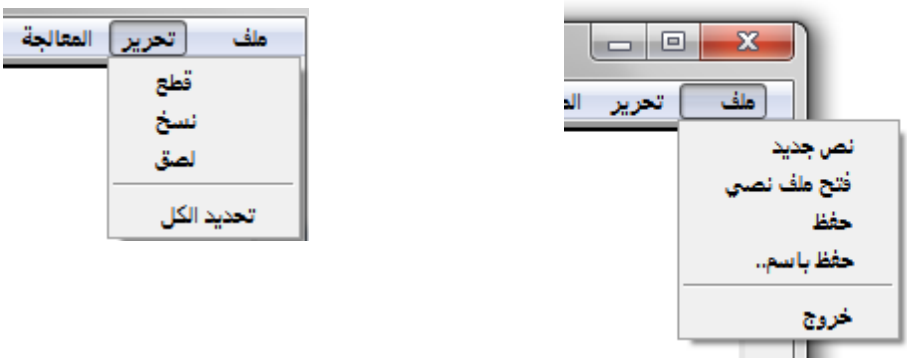

Figure 17: Les outils d'éditeur de texte

Nous avons aussi la préparation du texte (prétraitement) pour la Tokenisation

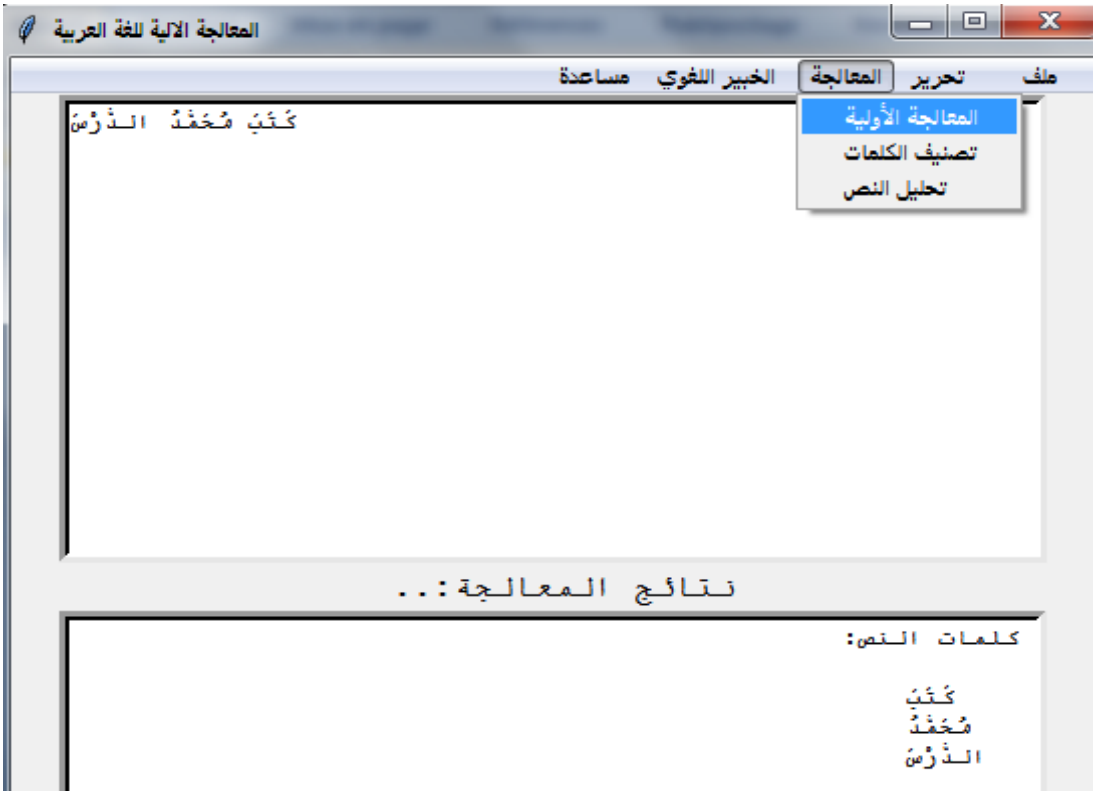

Figure 18: Le prétraitement du texte.

On peut aussi classifier les mots du texte selon le jeu du rôle syntaxique de chaque mot du texte par rapport à notre metabase terminologique (voir chapitre III)

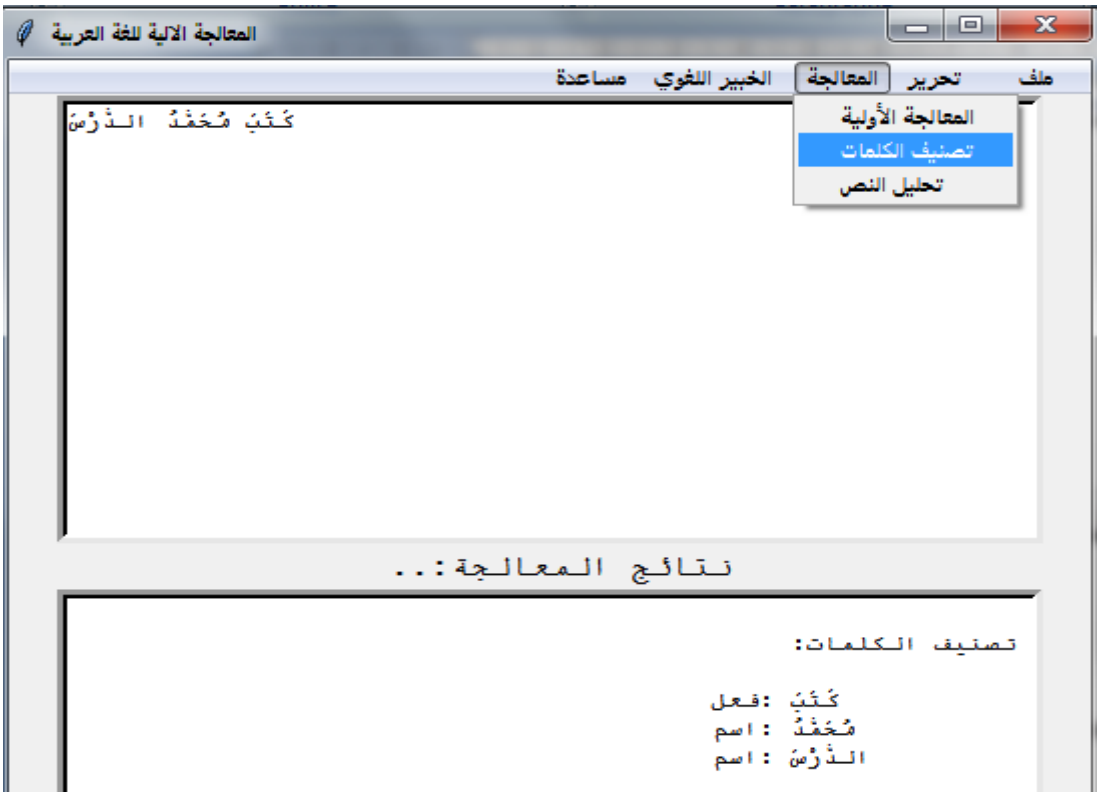

Figure 19: Classification des mots du texte.

Même aussi la possibilité d'analyse le texte, c.-à-d la vérification de l'existence des mots du texte dans le corpus et l'affichage de classe de ces mots.

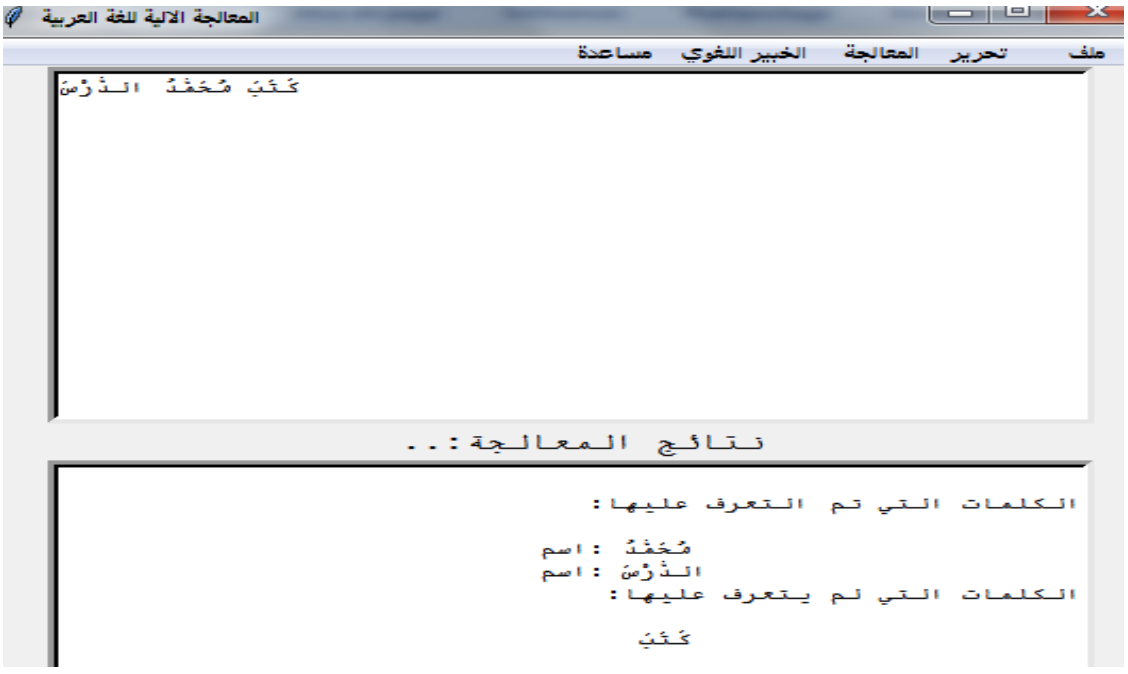

Figure 20: L'analyse du texte.

3.2. Interface expert

Nous avons mis à la disposition de l'expert un ensemble d'interfaces, Ceci afin que nous puissions bénéficier de l'expérience de l'expert ainsi que d'enrichir notre application avec les connaissances linguistiques de la langue arabe.

Pour notre application l'expert a la possibilité de :

- La définition des mots.
- La gestion des concepts (ajoute un nouveaux concept, modifier concept)
- La gestion des schèmes (ajouter, modification, suppression)
- La gestion des Inflexions des verbes (ajouter, modification, suppression)
- Vérification de l'existence des mots du texte dans le corpus, dans le cas ou un mot n'est pas déjà définie nous devons passer a l'interface expert pour faire entrer les mots.
- Consultations à la liste des mots du corpus.

Figure 21: Les taches pour L'expert

3.2.1 Ajoute un mot

Dans notre application l'expert a la possibilité d'ajouter et de définir un mot (le mot et son racine s'il existe et sa classe) la fenêtre d'ajoute un mot est comme suit :

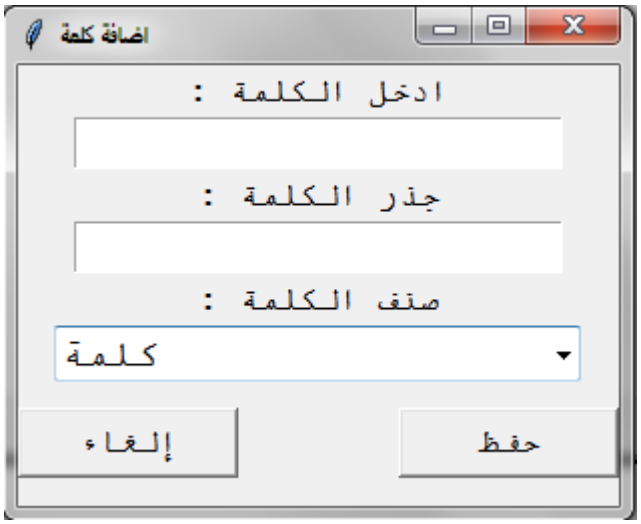

Figure 22: L'ajoute d'un mot

3.2.2 Gestion des concepts

Dans la présente application l'expert à la possibilité d'ajouter un nouveau concept et de modifier un concept existe déjà, les fenêtres de gestion des concepts sont montrées comme suit :

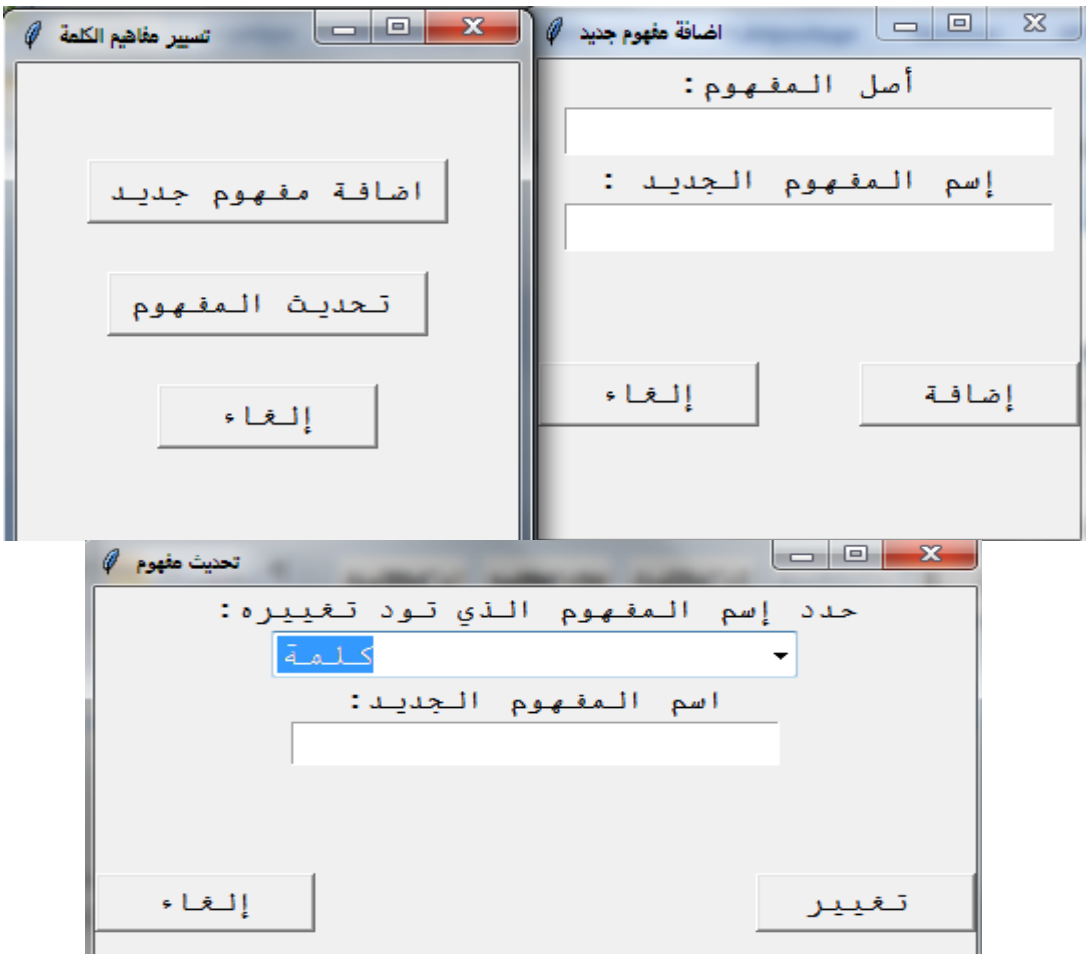

Figure 23: Interfaces de gestion des concepts.

3.2.3 Gestion des schèmes:

Pour la gestion des schèmes l'expert à la possibilité d'ajouter un nouveau schème comme il peut modifier ou supprimer un qui existe déjà, les fenêtres de gestion des concepts sont montrées comme suit :

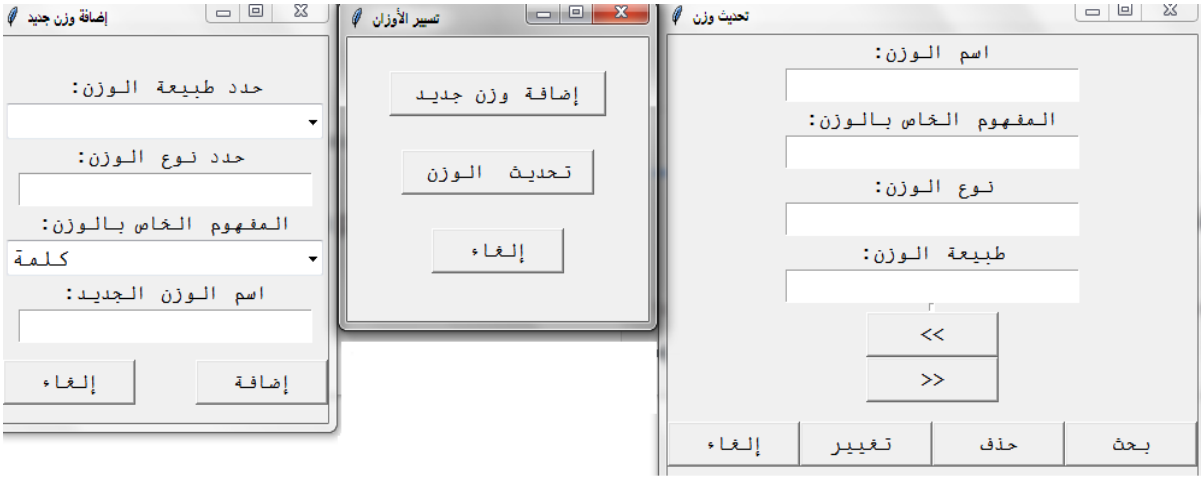

Figure 24: Interfaces de gestion des schèmes.

3.2.4 Gestion des Inflexions des verbes

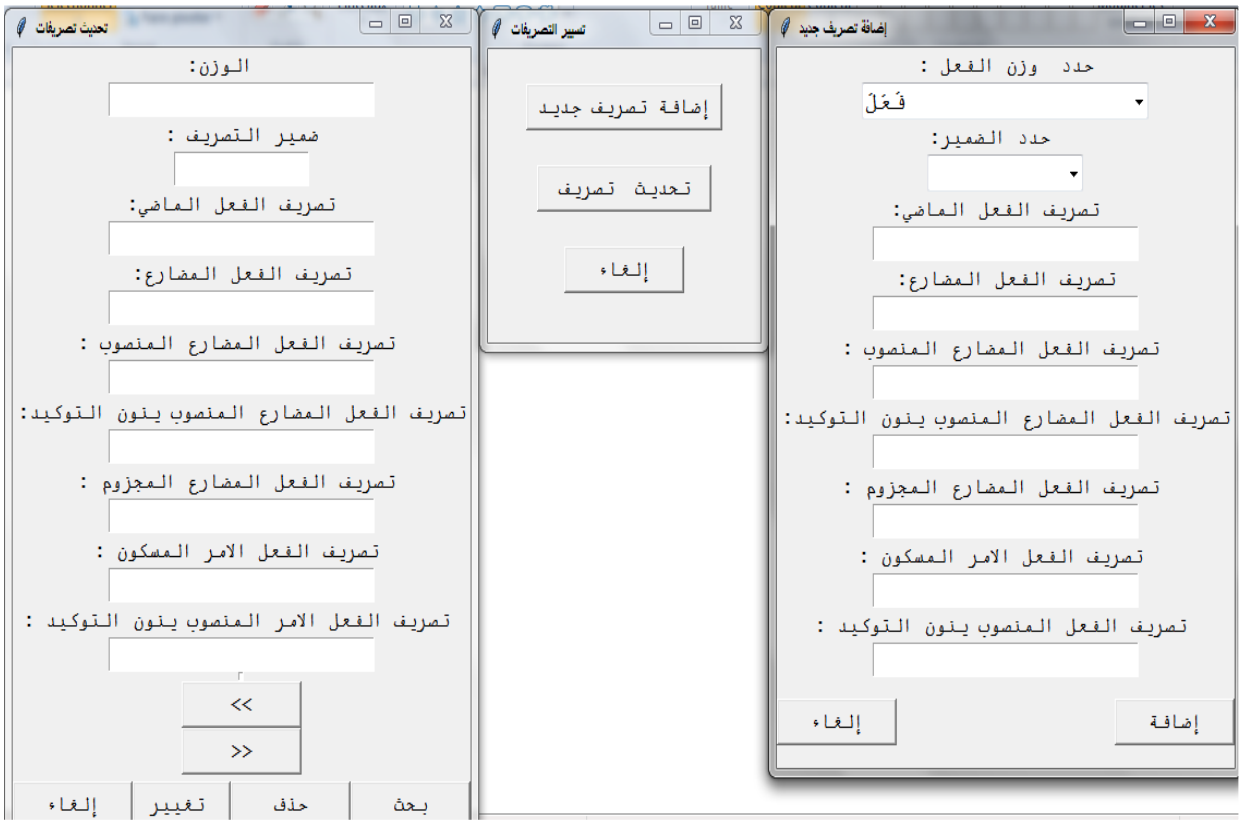

Figure 25: Interfaces de gestion des Inflexions des verbes.

3.2.5 Définition des mots du texte

Ce présent système doit vérifier l'existence de mots de texte dans le corpus, dans le cas ou le mot n'est pas déjà défini, on doit passer a l'interface expert pour définir ce mot. L'interface experte pour définir des mots est montrée comme suit :

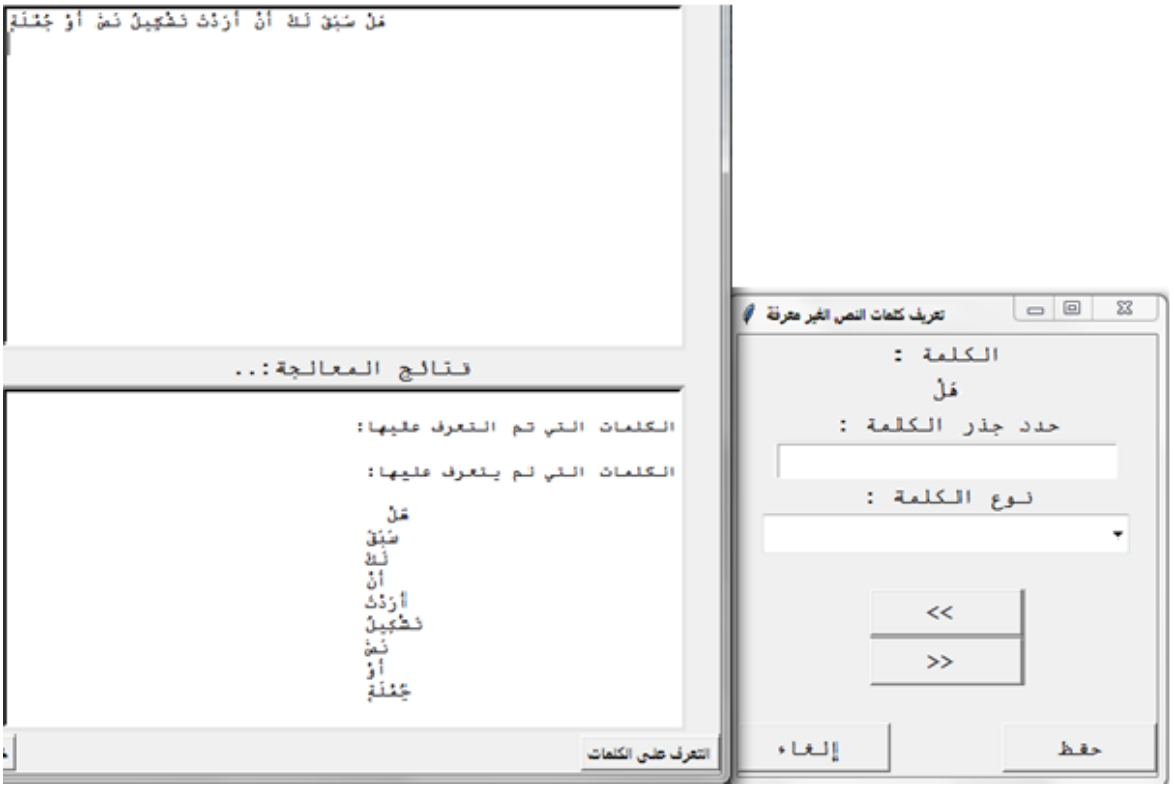

Figure 26: Interfaces de définition des mots du texte.

3.2.6 Consultations à la liste des mots du corpus

Figure 27: La liste des mots du corpus.

4. Résultat

A l'issue de la finalisation de cette application de traitement automatique de la langue arabe, les résultats obtenus indiquent certains objectifs atteints. Ces derniers figurent dans l'insertion des textes arabes qui seront par la suit prêt pour beaucoup de type traitement automatique.

Puis, elle accède à la détermination de l'étiquetage consistant à accorder à chaque mot un vecteur en vue de déterminer sa position en texte, elle lance le processus de génération des étiquettes mots.

4.1 Prétraitement d'un texte

Texte arabe	هَلْ سَبَقَ لَكَ أَنْ أَرَدْتَ تَشْكِيلُ نَصَّ أَوْ جُمْلَةٍ texte arabe
Filtrage	هَلْ سَبَقَ لَكَ أَنْ أَرَدْتَ تَشْكِيلُ نَصَّ أَوْ جُمْلَةٍ
Prétraitement (les tokens)	هَلْ/ سَبَقَ/ لَكَ /أَنْ /أَرَدْتَ/ تَشْكِيلُ/ نَصَّ/ أَوْ/ جُمْلَةٍ
Mots non reconnus par le corpus	هَلْ سَبَقَ لَكَ أَنْ أَرَدْتَ تَشْكِيلُ نَصَّ أَوْ جُمْلَةٍ

Tableau 10 : Prétraitement d'un texte, contient des mots non reconnus par le corpus.

4.2 Étiquetage

Concernant l'étiquetage qui Consiste a symboliser les mots de texte par leur vecteur SHA, c'est un vecteur qui représentant la connaissance syntaxique pour chaque mots, et qui est calculé de façon automatique.

Texte arabe	كَتَبَ مُحَمَّدُ الدَّرْسَ
Prétraitement (les tokens)	كَتَبَ /مُحَمَّدُ/ الدَّرْسَ
Etiquetages	[1, 2, 0, 0, 0, 0],[1, 2, 0, 0, 0, 0] ,[1, 1, 0, 0, 0, 0]

Tableau 11 : Prétraitement et étiquetage d'un petit texte**.**

4.3 Processus de génération des mots

A pour objective d'automatisation les taches de l'expert linguiste au maximum, c.-à-d mettre à jours le corpus par des nouveaux mots, on a utilisé des processus de génération des mots qui se fait selon les demarches suivants : la dérivation (la combinaison d'une racine et d'un schème), et la flexion.

Donc pour le processus de dérivation en utilisons table *sheme.csv* et combiner avec racine qui son définis par l'expert linguiste.

mots générés (corpus)				**Schème**	**Racine**
vecteur	**classes**	**racine**	**mot**		
[1, 1, 0, 0, 0, 0]	فعل	دخل	دَخَلَ	فَعَلَ	د خ ل
[1, 1, 0, 0, 0, 0]	فعل	دخل	اسْتَدْخَلَ	اِسْتَفْعَل	
[1, 1, 0, 0, 0, 0]	فعل	دخل	تَدَخَّلَ	تَفَعَّل	
[1, 1, 0, 0, 0, 0]	فعل	دخل	تَدَاخَلَ	تَفاعَل	

Tableau 12: Exemples de mots générés par dérivations.

Et aussi la même chose pour le processus de flexion en utilisons table *tasrif.csv* et combiner avec racine qui son aussi définis par l'expert.

mots générés (corpus)				flexion	Schème	Racine
vecteur	classes	racine	mot			
[1, 1, 0, 0, 0, 0]	فعل	دخل	دَخَلْتَ	فَعَلْتَ	فَعَلَ	د خ ل
[1, 1, 0, 0, 0, 0]	فعل	دخل	تَدْخُلُ	تَفْعَلُ		
[1, 1, 0, 0, 0, 0]	فعل	دخل	أُدْخُلْ	اِفْعَلْ		

Tableau 13: Exemples de mots générés par flexion.

L'objective d'utilisation de ces processus est de développer un analyseur morphologique performant, et d'autre part de génération des nouveaux mots dans le corpus de façons automatique, Il résulte une diminution de l'intervention manuelle de l'expert linguiste.

Remarque : Dans la langage arabe les règles de flexion ne pas toujours stable, par exemple pour les verbes faibles (الأفعال المعتلة) la flexion de ces verbes ce fait selon le type de verbe (Verbe assimilé, Verbe creux, Verbe incomplet, Verbe Ramas), donc on doit choisir manuellement quel type de conjugaison utilisé sachons que le nombre de cas que on a recensé dans cette étude n'excéder pas les 25 trois chacun de trois premiers et 8 pour les deux derniers (ramas collé, ramas séparé).

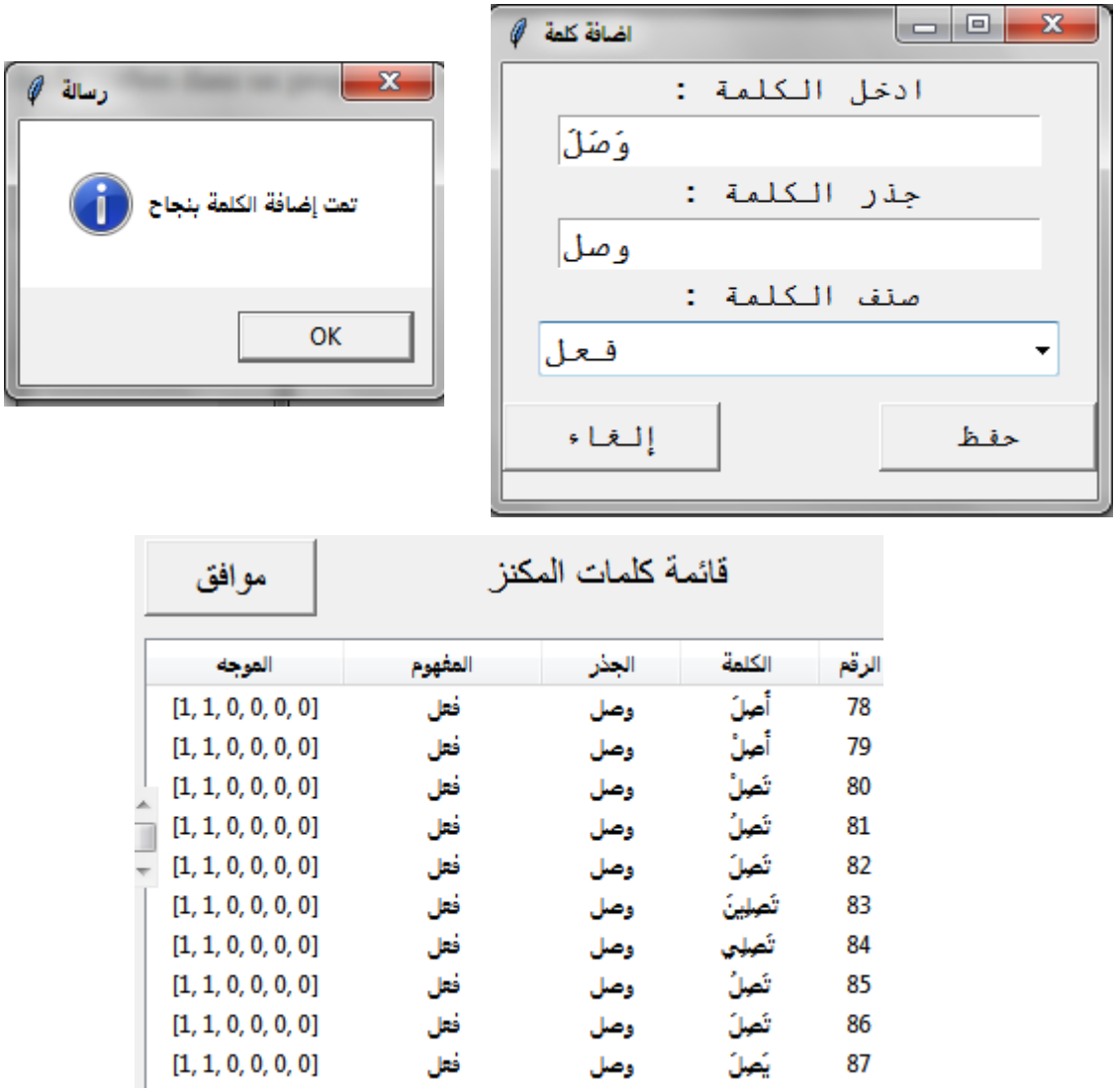

Figure 28: La flexion de verbe faible وصل (معتل المثال الواوي)

4.4 Propriétés du corpus

Comme on a dit précédemment Pour développer les outils de traitement informatique du langage, on a besoin des ressources langagières tel que corpus.

Et pour l'objectif de création de cette ressource qui est utile aux applications futures,

On a créé ce corpus décrit on utilise de descripteur standard open source, et connus à savoir le XML et le format CSV

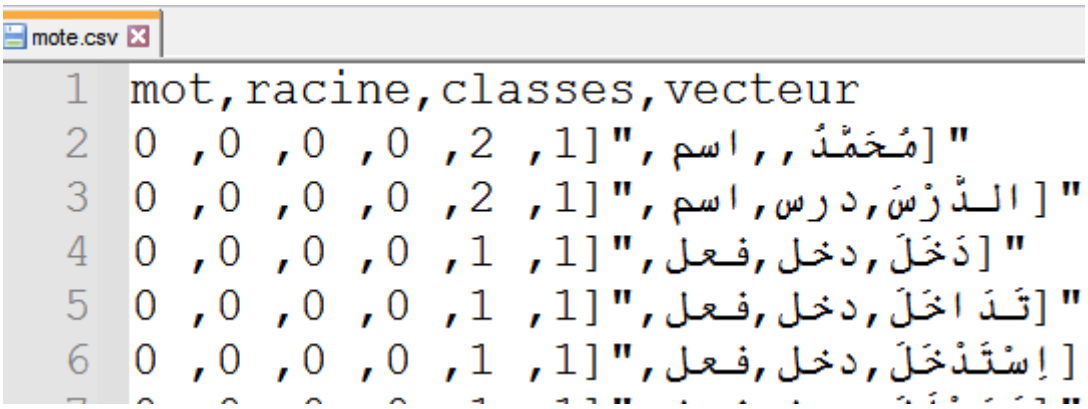

Figure 29: La corpus avec extension csv (mote.csv) .

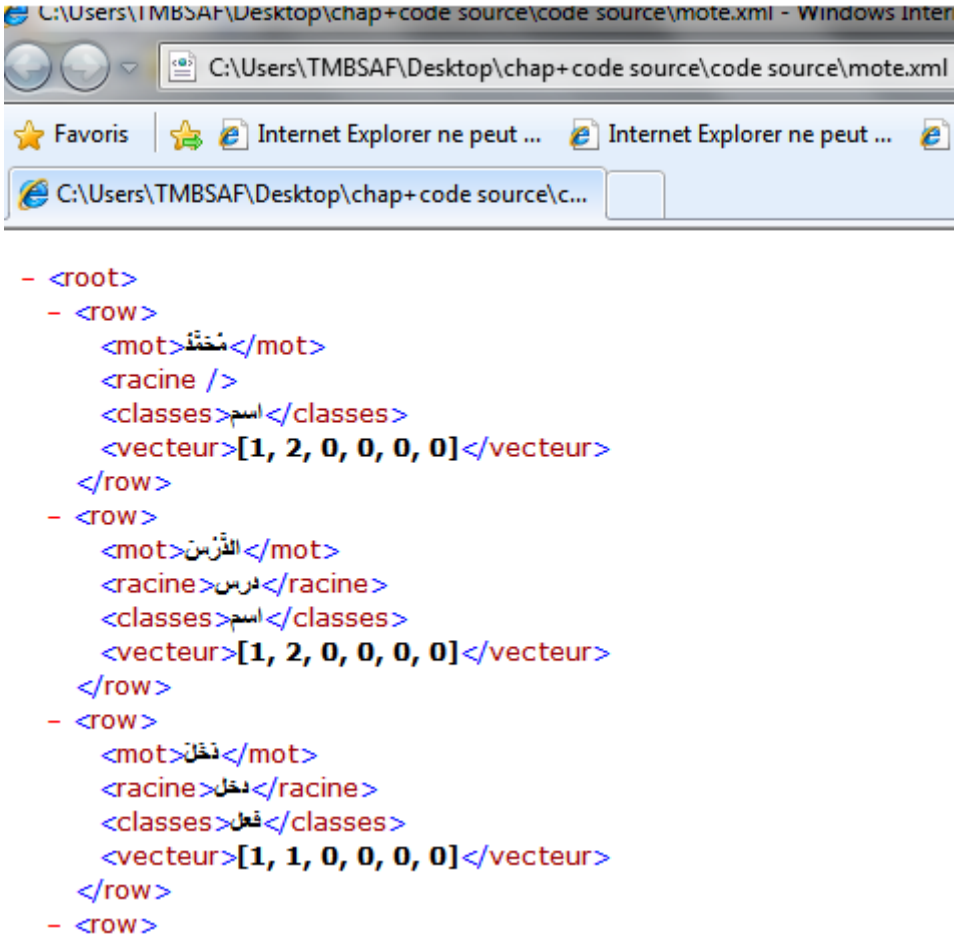

Figure 30: La corpus avec extension XML (mote.xml) .

5. Conclusion

Dans ce chapitre, nous avons présenté l'implémentation de notre application du traitement automatique de la langue arabe ainsi que le langage de programmation choisi. en plus les caractéristiques techniques de l'environnement qui a servi à développer cet outil. Parmi les points de ce chapitre figure la description de l'interface graphique de l'application proposée.

Par conséquent, nous avons essayé de décrire chaque outil et chaque composant de l'interface comme nous l'avons détaillé en mentionnant leurs fonctions.

Conclusion général

Conclusion général

A l'issue de ce mémoire, nous avons présenté une architecture de système à base de connaissance pour la langue arabe fondée sur l'acquisition d'un nombre important de résultats (dérivées) à partir d'une entrée élémentaire (cardinal) pour le TALArabe.

Le traitement est basé sur l'identification des rôles syntaxique des mots représentent en utilisant un formalisme proposé sous l'appellation Subsumption Hierarchical Attribute SHA qui permet de symboliser la connaissance syntaxique sous une forme vectorielle manipulable.

Le système résultant de notre étude propose un ensemble d'interfaces permettant à l'expert linguiste de mettre à jours le corpus au cas où en tombe sur des mots non définis, et aussi afin que on puisse bénéficier de l'expérience de l'expert ainsi que d'enrichir notre corpus avec les connaissances du domaine.

Pour l'objective d'automatisation du Lexique arabe et le développement de systèmes automatiques pour l'analyse grammatical automatique en na utilise des processus de génération des flexions et dérivations des mots.

Le problème majeur auquel nous nous sommes confronté est celui de l'absences des études l'linguistiques arabes orientées TAL, le cas qui nous a exigé d'aller fouiller dans les livres des grammairiens afin de dénicher les règles syntaxiques engendrant des phrases arabes syntaxiquement correctes .

Afin de résoudre ce problème à l'avenir, nous suggérons qu'il y aura projets de fin d'études conjointes entre linguistes et informaticiens afin que chacun puisse bénéficier de l'expertise de l'autre.

En fin nous espérons que notre travail soit une véritable contribution au domaine de recherche sur le traitement automatique de la langue Arabe.

Références bibliographiques

Références bibliographiques

[1] **Mohammed El Amine ABDERRAHIM**, «Reconnaissance des unités linguistiques signifiantes», thèse de doctorat, Université Abou Bekr BELKAID TLEMCEN ,le 08 Juillet 2008.

[2**] Fouad Soufiane Douzidia**, « Résumé automatique de texte arabe» , magistère d'informatique , Université de Montréal , Septembre, 2004.

[3**] https://perso.limsi.fr/anne/coursM2R/intro** consulté le .06/06/2021 a 10 :00

[4] **DAHOU Abdelghani**, «Acquisition de Connaissances à partir d'un texte Arabe non vocalisé (JEEM BOX)», Master en Informatique, Université d'Adrar,2014 .

[5] **Delphine BERNHARD**, « Apprentissage de connaissances morphologiques pour l'acquisition automatique de ressources lexicales », doctorat., l'Université Joseph Fourier – Grenoble I, 8 Décembre 2006.

[6] **Titt Karima**. « Traitement Automatique du langage, Ou Ingénierie linguistique ». Magister, Université d'Oran, 2010 -2011.

[7] **BENAISSA Bedr-Eddine,** «Construction semi-automatique d'ontologies à partir de textes arabes», magister, Université Abou Bakr Belkaid Tlemcen ,2012

[8] **DOUMI Noureddine,** Cour TALN 2016/2017 Master2 Département de M.I. Centre Universitaire de NAAMA

[9] **Zoulikha BENBLAL, Fatima BELOUAFI,** «Integration d'un lemmatiseur arabe dans le cadre d'un système de recherche d'information», Master en Informatique. Université Ahmed Draia – Adrar,2014-2015

[10] **https://www.un.org/ar/observances/arabiclanguageday** consulté le 14/03/2021 a 18:35

[11] **Saidi Abderrahmane ,** «Conception d'une interface en langage naturel (arabe) pour la conception des systèmes d'information », Magister en Informatique. Université des Sciences et de la Technologie d'Oran –USTO ,2010-2011

[12] **https://www.elwatan.com**/edition/contributions/la-langue-arabe-a-lere-de-la-revolution-numerique-10-01-2021 consulté le 19/04/2021 a 14:25

[13] **BENDEBICHE Rafiq, BOUDJENANE Saleh,** «Etiquetage et segmentation des textes arabes basée sur la classification syntaxique des mots», Master en Informatique. Centre Universitaire de NAAMA ,2017-2018

[14] **Ryding KC**. «A reference grammar of modern standard Arabic», Cambridge university; 2005 Aug 25.

[15] **Mustapha Al-Ghalayini**. « جامع الدروس العربية », , livre édité "المكتبة العصرية صيدا" en 2007 en Bierut, Lebanon. Page De 34 a 36

[16] **Mohsen Maraoui & Georges Antoniadis, Mounir Zrigui,** «Un système de génération automatique de dictionnaires étiquetés de l'arabe», In proceding CITALA 2007, Tunisie, january 2007

[17] **Mohamed Boudchiche,** «Toolkit AlKhalil pour l'analyse et la désambiguïsation morphologique des textes arabes», thèse de Doctorat en informatique, Faculté des Sciences Oujda, Université Mohamed Premier ,2020.

[18**] Ahmed Al-hamlawi**, " شذا العرف في فن الصرف" , livre "دار الفكر العربي"publié en 17 décembre 2005.

[19**] Slim MESFAR**, " Analyse Morpho-Syntaxique Automatique et Reconnaissance des entités nommées En Arabe Standard", thèse de Doctorat présenté en 24 Novembre 2008, Université De Franche-Comté.

[20] **Mohamed Ould Abdallahi Ould Bebah** ,«Contribution à l'analyse morpho-syntaxique

de la langue Arabe et application à la voyellation automatique» ,thèse de Doctorat en informatique, Faculté des Sciences Oujda, Université Mohamed Premier , Octobre 2013.

[21] **Siham Boulaknadel**,« Traitement Automatique des Langues et Recherche d'Information en langue arabe dans un domaine de spécialité : Apport des connaissances morphologiques et syntaxiques pour l'indexation» ,thèse de Doctorat, Université de Nantes, Le 18 Octobre 2008.

[22] **GASMI Mounira ,** «Utilisation des ontologies pourl'indexation automatique des sites Web en Arabe» ,Mémoire de MAGISTER Spécialité : Informatique ,UNIVERSITE KASDI MERBAH OUARGLA ,mai 2009

[23**] Nabil Ali**, « العرب و عصر المعلومات ,« سلسلة كتب ثقافية للمجلس الوطني للثقافة والفنون والأداب بالكويت livre publié en Avril 1994.de 330 a 337

[24] **GHOUL DHAOU, «**Outils génériques pour l'étiquetage morphosyntaxique de la langue arabe : segmentation et corpus d'entraînement », Mémoire de master, Université Stendhal - Grenoble 3 -,2010-2011.

[25] **Mohamed Outahajala , Lahbib Zenkouar , Paolo Rosso** , «CONSTRUCTION D'UN GRAND CORPUS ANNOTÉ POUR LA LANGUE AMAZIGHE**»,** Dans Études et Documents Berbères 2014/1 (N° 33), pages 77 à 94

[26] **KURČERA H. and FRANCISW**. «Computational Analysis of Present-Day American English**»**. Brown University Press, Providence, RI, 1967

[27] **KHOJA S., GARSIDE R., and KNOWLES**. «A Tagset For The Morphosyntactic Tagging Of Arabic». In Proceedings of Corpus Linguistics. Lancaster, UK, pp. 341-353. 2001

[28] **Tim BUCKWALTER**, « Buckwalter Arabic Morphological Analyzer Version 1.0. ». le numéro de catalogue est LDC2002L49. Rapport interne ISBN 1-58563-257-0 en 2002.

[29] **Khalil El-Basri**, « برنامج الخليل الصرفي في دليل الاستعمال»,2010.

[30] **Houda Saadane**, «Le traitement automatique de l'arabe dialectalisé : aspects méthodologiques et algorithmique», thèse de Doctorat, UNIVERSITÉ GRENOBLE, décembre 2015.

[31] **Yasser YAHIAOUI, Ahmed LEHIRECH,** « A META DESCRIPTION LOGICS KNOWLEDGE BASE FOR ARABIC LANGUAGE PROCESSING». In proceding ICDIPC 2013 ,Dubai, january 2013

[32] **Yasser YAHIAOUI, Ahmed LEHIRECH, Djelloul Bouchiha** , «Proposed Representation Approach Based on Description Logics Formalism», IJCSA , N° 10.5815 ,le mai 2016

[33] **http://remy-manu.no-ip.biz/UML/Cours/coursUML5** : Langage de modélisation objet unifié Cours n° 5 : Diagramme de séquences, Consulté le 20/05/2021 a 21 :21

[34]**https://fr.wikiversity.org/wiki/Mod%C3%A9lisation_UML/Le_diagramme_d%27activit%C3%A9,** Consulté le 25/05/2021 a 10 :21

[35] **Loic Gouarin**, «Les base du langage python». laboratoire de mathématiques d'Orsay, 6 décembre 2010

[36] **http:// numpy.org/** Consulté le 13/06/2021 a 17 :00