REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE MOHAMED BOUDIAF - M'SILA

FACULTE DES MATHEMATIQUES ET DE L'INFORMATIQUE.

DEPARTEMENT D'INFORMATIQUE

N° :……..

DOMAINE : MATHEMATIQUE ET INFORMATIQUE

FILIERE : INFORMATIQUE

OPTION : SYSTEMES D'INFORMATION AVANCES

Mémoire présenté pour l'obtention du diplôme de Master Académique

Par: BENHALIMA MAISSA

Intitulé

Implémentation d'une méthode hybride (Morphologique & statistique) pour l'analyse des mots arabes

Soutenu devant le jury composé de :

……	Université de M'sila	Président
Mr.MAHDJOUBI Roussafi	Université de M'sila	Rapporteur
……	Université de M'sila	Examinateur

Année universitaire : 2016 /2017

Dedicaces

Je dédie ce modeste travail:

A mes chers parents.

Ames frères et sœurs.

A toute ma famille.

A tous mes fidèles amis.

A mon fiancé.

maissa

Remerciements

Tout d'abord, je remercie ALLAH le tout puissant de m'avoir donné la santé, la volonté et la force pour terminer ce travail.

Je remercie vont à tous ceux qui ont contribué à la réussite de ce travail en particulier a :

Mon encadreur : MAHDJOUBI Roussafi.

Et aux membres du jury d'avoir accepté d'évaluer mon travail.

Table des matières

CHAPITRE 2: LE TRAITEMENT AUTOMATIQUE DE LA LANGUE ARABE

CHAPITRE 3 DESCRIPTION DU SYSTEME REALISE

Listes des figures

Liste des tableaux

INTRODUCTION GENERALE

L’arabe est la langue des populations arabes qui firent leur entrée dans l’histoire depuis 3 millénaires environ et qui occupaient les zones septentrionales de l’Arabie. Cette langue connut un destin extraordinaire et même prodigieux avec l’avènement de l’islam il y a 14 siècles. Elle est actuellement la langue officielle de 22 pays arabes. Elle fait partie de la grande famille des langues sémitique, au même titre que l’akkadien, le phénicien, l’hébreu, etc.

Dans un langage à haute morphologie dérivationnelle telle que l’arabe, la détection des unités lexicales dans un texte électronique devient une tâche assez complexe comme la traduction automatique, l'indexation automatique des documents, la recherche d'information, etc. donc, la langue arabe était considérée parmi les langues les plus difficiles à traiter automatiquement à cause de plusieurs problèmes. Mais le développement de l'informatique et l'avancement des travaux de recherche dans le domaine du traitement automatique de la langue arabe ne cessent d'inventer de plus en plus d'outils bien adaptés qui ont permis de lever certains obstacles. Mais même avec ces progrès informatiques, il existe encore plusieurs problèmes liés au traitement automatique de la langue arabe. Ces obstacles proviennent des caractéristiques de la langue arabe elle-même.

L'analyse d'un mot arabe consiste principalement à déterminer sa racine. Donc notre analyse menée sur les méthodes d'extraction de la racine, contribuent à identifier les problèmes et améliorer les performances du traitement automatique en la langue arabe.

Notre travail se centralise sur une méthode hybride entre l'analyse morphologique et statistique pour analyser les mots arabes. Ce mémoire sera organisé de la manière suivante :

Nous commençons, dans la première partie, par la présentation des caractéristiques de la langue arabe. Dans la deuxième partie, nous nous intéressons à expliquer les méthodes statistiques et morphologiques proposées pour l'extraction des racines arabes. Puis, nous présentons la méthode choisie, suivie de sa réalisation. Finalement, nous terminons ce travail par une conclusion.

La langue arabe est une langue sémitique très populaire en termes de nombre de locuteurs. Elle est parlée par plus de 422 millions de personnes [40] dans le monde et distribué dans la région connue du monde arabe et les autres régions voisines comme la Turquie, le mali et l'Érythrée. L'arabe doit sa formidable expansion à partir du 7^{éme} siècle grâce à la propagation de l'islam et la diffusion du coran [20]. Au cours de ce chapitre, nous présenterons certaines propriétés morphologiques et syntaxiques de la langue arabe.

1. Particularités de la langue arabe

La langue arabe est une langue orientale qui s'écrit de droite à gauche. L'alphabet de la langue arabe compte 28 consonnes (tableau 1.1).

Les consonnes arabes
ا ب ت ث ج ح خ د ذ ر ز س ش ص ض ط ظ ع غ ف ق ك ل م ن ه و ي

Tableau 1. 1: les consonnes de la langue arabe.

La représentation morphologique de l'arabe est assez complexe en raison de la variation morphologique de ses mots et du phénomène d'agglutinement; les lettres changent de formes selon leur position dans le mot (isolée, initiale, médiane et finale). Le tableau 1.2 montre un exemple des différentes formes de la lettre «qaf //ق» dans différentes positions. [14]

Isolée	Initiale	Médiane	Finale
ق	قـ	ـقـ	ـق
طارق	قرآن	القرآن	غسق

Tableau 1. 2. : différentes écritures de la lettre « qaf//ق» selon sa position dans le mot.

Cependant, 6 d'entre elles s'attachent uniquement aux lettres précédentes mais pas aux lettres suivantes. Ces lettres sont les suivantes: {ا د ذ ر ز و}.

De même, il existe 6 voyelles en arabe, 3 longues et 3 courtes, la durée d'une voyelle longue est environ le double de celle d'une voyelle courte. (Tableau 1. 3). [13]

Voyelles longues	Voyelles courtes
اوي	ـِـُـَ

Tableau 1. 3: Les voyelles arabes.

Un mot arabe s'écrit avec des consonnes et des voyelles. Les voyelles courtes sont ajoutées au-dessus ou au-dessous des lettres (ـْ,ـِ ، ـُ ,ـَ). Elles sont nécessaires à la lecture et à la compréhension correcte d'un texte, elles permettent de différencier les mots ayant la même représentation. Le Tableau 1. 4 présente un exemple pour les mots مدرسة et شعر [17]

Mot sans Voyelles	1ère Interprétation		2ème Interprétation		3ème Interprétation	
مدرسة	مَدْرَسَةٌ	Ecole	مُدَرِسَة	enseignante	مُدَرَسَةٌ	enseignée
شعر	شَعَرَ	Il a senti	شِعْر	Poème	شَعْر	Chevelure

Tableau 1. 4: ambiguïté causée par l'absence des voyelles courtes.

2. La structure morphologique d'un mot arabe

En arabe, un mot peut signifier toute une phrase grâce à sa structure composée qui est une agglutination de « morphes » (racines, préfixes, affixes, suffixes, schèmes). La représentation suivante schématise une structure possible d'un mot. Rappelons que la lecture et l'écriture d'un mot se fait de droite vers la gauche. [9]

Post fixe	Suffixe	Corps Schématique	Préfixe	Antéfixe

- Les antéfixes sont des prépositions ou des conjonctions.
- Les préfixes et les suffixes expriment les traits grammaticaux et indiquent les fonctions: cas du nom, mode du verbe et autres catégories d'actualisation (nombre, genre, personne,…)

- Les post fixes sont des pronoms personnels.

Exemple:

« Atatadhakkaronana » أَتَتَذَكَّرُونَنَا

Ce mot exprime la phrase en français: "*Est ce que vous vous souvenez de nous?*" La segmentation de ce mot donne les constituants suivants :

Post fixe	Suffixe	Corps Schématique	Préfixe	**Antéfixe**
نا	**ون**	**تَذَكَّرْ**	**تَ**	**أ**

Tableau 1.5.: La segmentation du mot "أَتَتَذَكَّرُونَنَا" **.**

- Antéfixe: أ conjonction d'interrogation.
- Préfixe: تpréfixe verbal du temps de l'inaccompli.
- Corps schématique: تذكّر dérivé de la racine: ذكّر selon le schèmeفعّل.
- Suffixe: ون suffixe verbal exprimant le pluriel.
- Post fixe: نا pronom suffixe complément du nom.

2. 1. Catégories des mots arabes

L'arabe considère les catégories de mots suivantes **:**

- ✓ Le nom: l'élément désignant un être ou un objet qui exprime un sens indépendamment du temps.
- ✓ Le verbe: entité exprimant un sens dépendant du temps, c'est un élément fondamental auquel se rattachent directement ou indirectement les divers mots qui constituent l'ensemble.
- ✓ Les particules: entités qui servent à situer les événements et les objets par rapport au temps et l'espace, et permettent un enchaînement cohérent du texte.

2. 1. 1. Le nom:

Toute unité lexicale référant à un sens indépendant du temps [6], regroupe : les adjectifs féminin et masculin ; les noms déméritent, les noms prolongés ainsi que les noms réduits, les noms communs et les noms propres, les pronoms et leurs types (connectés et séparés), les noms de périphrases, les noms du verbe, les noms de voix [33]. Nous pouvons distinguer deux classes de noms (tableau 1. 6) : la première regroupe les noms conjugables ou semi-conjugables qui peuvent avoir la forme duelle, plurielle, etc. la deuxième classe regroupe les noms non-conjugables qui gardent la même forme quel que soit le contexte. Les noms conjugables sont soit des noms primitifs, qui échappent à toute dérivation comme «عَزَل» (gazelle), soit des noms dérivationnels qui sont formés à partir d'une racine comme « مَكْتَبَة » (bibliothèque) de la racine «كَتَبَ». [14]

Catégorie	Dérivation	Conjugaison	Sous-catégorie	Exemple
Nom	Dérivation nel irrégulier	Non conjugable	Adverbe	أينَ، حَيثُ، قبلَ
			Nom de voix	كخْ، نخْ
			Nom de verbe	هَيْهَات، آه، أف
			Pronom personnel (affixé ou isolé)	هُوَ، أنَا، تُ، تُنَ
			Pronom interrogatif	كَيْفَ، مَتَى، مَا
			Pronom conditionnel	مَنْ، إذا
			Pronom allusif	كَمْ، كأَيَ
		Conjugable	Pronom relatif	الذي، التي
			Nom de nombre	ثلاثة، وَاحد، خَمْسَةُ
			Pronom démonstratif	هَذا، هَذه
			Nom propre	مُحَمَدُ، هِنْدُ، صَحْرَاءُ
			Nom commun	قلمُ، أرْنَبُ، رَجُل
	Dérivation nel régulier	Conjugable	Masdar	كتَابَةَ، القَلْ
			Participe actif	قَالٌ،شَارِبُ
			Participe passif	مَكْتُوبُ، مَضْرُوبُ
			Nom d'une fois	جَلْسَةُ، ضَرْبَةُ
			Nom de manière	خَطْرَةُ، جَلْسَةُ
			Nom de temps	مَغْرِبُ
			Nom de lieu	مَكْتَبُ، مَقْبَرَةُ
			Nom d'instrument	مطْرَقَةُ، مسْمَارُ
			Adjectif	حَسَنُ، جَميلُ، بَطَلُ
			Elatif	أَحْسَنُ، أَفْضَلُ
			Nom diminutif	كُتَيبُ، شُوَيْعِرُ
			Nom de relation	تُونسِيُ، مصْرِيُ
			Intensif	قتَلُ، عَوَاصُ

Tableau 1. 6. : Classement des sous catégories de noms.

2. 1. 2. Le verbe:

Unité lexicale référant à un état ou une action exprimant un sens dépendant du temps comme: عمل (*travailler*),ذهب (partir) [6]. La plupart des mots en arabe dérivent d'un verbe de trois lettres. Chaque verbe est donc la racine d'une famille de mots. Le mot en arabe se déduit de la racine en rajoutant des suffixes et/ou des préfixes.

Nous pouvons classer les verbes arabes selon plusieurs critères [33] :

- Selon le critère de temps, il existe trois types :
  - l'accompli: correspond au passé et se distingue par des suffixes (par exemple pour le pluriel féminin on a **كتبن**(*elles ont écrit*) et pour le pluriel masculin on a **كتبوا**(*ils ont écrit*).
  - Inaccompli présent: les verbes conjugués à ce temps se distinguent par les préfixes. Pour notre exemple, au masculin singulier on obtient **يَفتح** (*il ouvre*) ; et pour le féminin singulier, on obtient **تَفتح**(*elle ouvre*).
  - Inaccompli futur: la conjugaison d'un verbe au futur nécessite d'ajouter l'antéposition au début du verbe conjugué à l'inaccompli. En ajoutant l'antéposition à notre exemple **س,** on obtient **سَيفتح** (*il ouvrira*), qui désigne le futur. On peut également ajouter l'antéposition **سَوف** on obtient **سَوف يَفتح** (*il va ouvrir*). [20]
- Selon leur sens et leur transitivité de sujet au complément aux deux types :
  - Intransitif.
  - transitif.
- Selon leurs modes aux deux types :
  - la voix passive
  - la voix active.
- Selon le nombre des consonnes de la racine, la majorité des verbes, à peu près 85%, sont formés a partir des racines de 3 lettres et le reste entre les racines de 4 et 5 lettres. Ces racines peuvent donner plusieurs mots grâce à des transformations morphologiques selon les schèmes.
- Selon le schème et le nombre de consonnes qui constituent la structure verbale, nous avons, soit:
  - des verbes nus (*Mojarad* مجرد): qui sont formés seulement par les consonnes de leurs racines et des voyelles brèves.

- des verbes augmentés (*Mazid* مزيد): qui sont dérivés des trois consonnes de la racine par modification des voyelles, par redoublement de la deuxième lettre de la racine, par adjonction et même par intercalation d'affixes. Les verbes dérivés se conjuguent avec les mêmes préfixes et suffixes que le verbe nu. Les verbes trilitères peuvent être augmentés au maximum par trois lettres et les verbes quadrilitères par deux lettres. Alors, la longueur maximale d'un verbe arabe est de 6 lettres. [7]

❖ Selon leur conjugaison, il existe :

- le conjugué
- le non conjugué
- l'invariant.

❖ Il existe aussi les verbes d'exclamation ainsi que les verbes panégyrique et les verbes de diatribe.

❖ Selon la nature des consonnes:

- Les verbes sains (صحيح) dont les racines ne contiennent pas des lettres défectueuses qui sont (ا,و , ي)
- Les verbes défectueux (معتل): dont les racines contiennent une ou deux lettres défectueuses qui causent des altérations importantes au cours de la conjugaison. [7]

2. 1. 3. La particule:

Entité invariable exprimant un sens dépendant de compréhension. 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, parmi lesquels on peut citer [33]:

- Particules de préposition: exemple *MaEa,ILA,Fi,Ka,Bi* (بِ,كَ,فِي,إِلى,مَعَ).
- Particules de coordination : exemple *Wa, Fa,Aaw* (أَو,فَ,وَ).
- Particules interrogatives : exemple *Aa,MaA,Hal*(أَ,مَا,هَلْ).
- Particules d'affirmation : exemple *LaA,NaEam,Bala,Ajal*(لا،نَعَمْ،بَلى،أَجَلْ).
- Particules de négation: exemple *Lame,Lane*(لَمْ ،لَنْ).
- Particules distinctive: exemple *Aye* (أَيْ).
- Particules relatives : exemple *MaA*(مَا).
- Particules de future : exemple *Sa,Sawefa,Lane,Aan*(سَ,سَوفَ،لَن،أَنْ).

- Particules conditionnelles : exemple *Ine,law*(إِنْ،لَوْ). [11].

2. 2. Morphologie arabe:

La morphologie est un domaine de la langue qui permet de faire la description des règles régissant la structure du mot et ses changements par l'ajout de particules pour former des dérivés et des formes flexionnelles. En langue arabe, l'analyse morphologique est d'autant plus importante que les mots sont fortement agglutinés, c'est-à-dire qu'ils sont formés dans leur majorité par assemblage d'unités lexicales et grammaticales élémentaires. Le lexique arabe comprend trois catégories de mots: verbes, noms et particules. Les verbes et noms sont le plus souvent dérivés d'une racine à trois consonnes radicales.

2. 3. Les éléments essentiels de la morphologie de la langue arabe:

A. La racine:

Une racine est purement consonantique, elle est formée par une suite de trois ou quatre consonnes (ou même cinq pour les noms) formant la base du mot. Elles sont aux alentours de 10000 racines dont la grande majorité (85%) sont trilatérales. Les restes sont des racines quadrilatérales ou quintilatérales. Une racine définit la signification fondamentale des mots dérivés en utilisant différents diacritiques et affixes avec les lettres de la racine pour créer l'inflexion de la signification. [3]
Par exemple, la racine كتب peut engendrer 15 mots autour du concept (*écriture*): كتاب (*livre*), مكتب (*bureau*), مكتوب (*écrit*) etc. Les lettres de la racine كتب gardent leur position dans les mots engendrés. Des voyelles (ا/ de كتاب) ou des consonnes (م/ de مكتب) peuvent s'y ajouter. C'est le fait le plus caractéristique de la morphologie arabe et plus généralement sémitique.

B. Le schème:

Le schème est un mot composé de trois consonnes ف [f], ع[a], et ل [l], qui sont vocalisées et qui peuvent être augmentées par d'autres lettres (préfixe, suffixe et infixe). Le schème joue un rôle très important dans le processus de génération des formes dérivées à partir d'une racine. Ce processus de génération consiste à remplacer les consonnes de la racine du schème par les consonnes de la racine en question, tout en gardant les mêmes voyelles et les mêmes lettres augmentées et tout en respectant le même ordre des consonnes, autrement dit, le schème peut être considéré comme un

moule sur lequel coule la racine[7]. On peut classer les schèmes en deux catégories: des schèmes verbaux et des schèmes nominaux. Ainsi, à partir d'une racine, on peut générer des noms et des verbes selon la catégorie du schème utilisé.

C. Le lemme:

Le lemme est l'entrée lexicale dans un lexique ou dans un dictionnaire. Il s'agit d'une forme entièrement vocalisée. Chaque mot est rapporté à son lemme qui est sa forme canonique qui dépend toujours de la catégorie grammaticale de ce mot, si c'est un nom il doit être au singulier et si c'est un verbe il doit être à l'accompli de la troisième personne du singulier. . . etc. Un lemme peut être formé par un mot simple ou un mot composé.

Nous remarquons que les particules gardent toujours leur représentation de base. Pour les autres catégories, le lemme permet de regrouper les mots ayant la même racine, le même schème original et le même sens. Ce regroupement aide à réduire le nombre d'entrées lexicales. [11]

D. Les affixes:

Les affixes sont des lettres qui s'ajoutent au début (les préfixes) ou à la fin des mots arabes (les suffixes). En général, Ils sont utilisés pour accorder aux mots des éléments syntaxiques. Ils marquent l'aspect verbal, le mode, les propriétés transitives, etc. Ils sont aux alentours de 150. [12]

Les préfixes dépendent des mots auxquels ils s'attachant. En effet, la plupart des mots arabes commencent par le préfixe < ال التعريف, al altaåryif, l'article de définition > qui est utilisé en tant que terme déclaratif. Pour cela, il y a trois types de préfixes. Premièrement, les préfixes nominaux qui sont réservés pour les noms et les adjectifs. Deuxièmement, les préfixes verbaux qui sont réservés aux verbes. Et troisièmement, les préfixes généraux qui sont utilisés indépendamment du type des mots.

Il y a deux types de suffixes, les suffixes verbaux et les suffixes nominaux. Les premiers dépendant de la transitivité et de la personne conjuguée. Les suffixes nominaux indiquent la flexion casuelle du nom (nominatif, accusatif, et génitif), le genre (masculin et féminin), le nombre (singulier, duel et pluriel), etc. [3]

E. Le stem:

Un Stem est la dérivation obtenue à partir d'une racine donnée selon un modèle. L'arabe classique à un grand nombre des Stems qui ne sont pas tous utilisables, 2% seulement sont utilisables selon Rashwan; le Stem correspond à un schème si et seulement s'il possède le même nombre de lettres et les mêmes lettres dans les mêmes positions [3]. Une exception est accordée aux consonnes <ف, f>, <ع, å>, <ل, l> qui sont les lettres de la racine de base <فعل, fål, (*faire*)>. Par exemple, on y trouve : <مَكاتِب, mkatb,(*bureaux*)>, il est obtenu à partir de la racine <كتب, ktb, (*il a écrit*)> selon le schème <مفاعل, mfaål>. Les Stems produits ne sont pas tous utilisables.

F. Les mots dérivés:

Selon la grammaire traditionnelle, le lexique arabe comprend trois catégories de mots: verbes, noms, et particules. Hormis les mots propres, les mots des deux premières catégories sont dérivés à partir d'une racine. Ils sont nommés mots réguliers ou mots dérivables. Les mots dérivés sont construits à partir d'un Stem en y ajoutant des affixes. La plupart des mots arabes sont considérés comme des mots dérivés, puisqu'ils sont construits à partir des racines [31].

G. Les mots outils:

Les mots outils sont des entités qui servent à situer des faits ou des objets par rapport au temps ou au lieu. Ils jouent également un rôle clé dans la cohérence et l'enchainement d'un texte. Par exemple, nous avons des particules qui désignent un temps: بعد (*après*),قبل (*avant*), ou un lieu tel que حيث (*où*). Selon leur sémantique et leur fonction dans la phrase, ils peuvent jouer un rôle important dans l'interprétation d'une phrase en exprimant une introduction, explication, conséquence, etc. Les mots outils incluent différentes catégories. Nous citons:

- Les prépositions: par exemple,"في" (*dans*).
- Les conjonctions de coordination: par exemple, "ثُم" (*ensuite*).
- Les adverbes: par exemple, "أَبدًا" (*jamais*).
- Les quantificateurs: par exemple,"كُلّ " (*tout*).

Les mots outils se répartissent en deux sous-groupes : ceux qui sont variables (quantificateurs) et ceux qui sont invariables (adverbes, prépositions, etc.) [38].

H. Les mots isolés:

Les mots isolés sont les mots qui n'ont pas de racines. Les mots sont en général, les noms propres, les noms communs et les particules [3].

Un nom propre désigne toute substance distincte de l'espèce à laquelle elle appartient. Il ne possède en conséquence aucune signification, ni aucune définition. Exemple : Paris, Jules, etc. Par contre, un nom commun est toute substance non distincte de l'espèce à laquelle elle appartient. Il est pourvu d'une signification et d'une définition.

Exemple : < 'بلد', (*pays*)>, <لنسان, (*personne*)>, <حيوان, (*animal*)>, etc.

I. Les diacritique:

Les signes diacritiques sont des signes ajoutés au-dessus ou en dessous des lettres arabes afin de signifier la prononciation du mot, ce rôle phonologique influe aussi sur le sens de ce mot.

Au nombre de trois, ces symboles sont transcrits de la manière suivante :

- La fetha(الفتحة)[a] est symbolisée par un petit trait sur la consonne (بَba)
- La damma(الضمة) [u] est symbolisée par un crochet au-dessus de la consonne (بُbu)
- La kasra(الكسرة) [i] est symbolisée par un petit trait sous la consonne (بِbi)

Un petit rond_symbolisant le soukoun (سكون) et apposé sur une consonne lorsque celle-ci n'est liée a aucune voyelle (بَعْدَ/baåda) [10].

M. La Šhadda :

Šhadda est un signe qui peut être placé au-dessus de toute consonne autre que celle se trouvant à la position initiale du mot. La consonne surmontée de ce signe est analysée comme une séquence de deux consonnes identiques [10]. Exemple: كلم --→/ ka**ll**ama/ (*Il a parlé à*).

N. Tanwin:

Ou bien la désinence (ً an, ٌ un, ٍ in) considérée par quelques auteurs comme étant le double de la même voyelle brève, il est ajouté seulement à la fin des mots indéterminés, par conséquent, il n'apparaît jamais avec l'article de détermination *AL* (ال).

3. Conclusion

Dans ce chapitre, nous avons présenté les caractéristiques de la langue arabe qui sont différentes par rapport à d'autres langues. C'est une langue riche et sensible au contexte, ce qui présente de nombreux défis pour des domaines divers tels que le traitement automatique du langage naturel qui fait face à un certain nombre de problèmes comme le problème de la voyellation, l'agglutination, l'extraction de la racine ou aussi la recherche d'informations.

Par sa richesse morphologique et syntaxique, la langue arabe est considérée parmi les langues les plus difficiles à traiter dans le domaine du traitement automatique du langage. Cela est dû notamment, aux diverses difficultés rencontrées dans sa racinisation. En effet, un mot arabe est construit à partir d'une racine à laquelle on ajoute des affixes selon un modèle précis. De ce fait, pour trouver un mot dans la plupart des dictionnaires arabes, nous devons d'abord extraire sa racine et ensuite trouver cette racine dans le dictionnaire. Pour extraire la racine à partir d'un mot arabe, plusieurs méthodes ont été proposées. Ces méthodes sont principalement basées sur les caractéristiques morphologiques de la langue arabe ou sur des calculs statistiques.

1. Traitement automatique des langues naturelles (TALN)

Le traitement automatique des langues naturelles (TALN) est un domaine à la frontière de la linguistique et de l`informatique. Il a pour objectif de développer des systèmes capables de traiter de façon automatique des données linguistiques exprimées dans une langue naturelle donnée et pour une application bien définie.

L'histoire du TALN commence dans les années 1950[41]. Le champ du traitement automatique du langage couvre de très nombreuses disciplines de recherche qui peuvent mettre en œuvre des compétences aussi diverses que les mathématiques appliquées ou le traitement du signal.

Parmi les applications les plus connues, on peut citer celles en relation avec la production ou la modification de texte:

- La traduction automatique.
- La génération automatique de textes.
- Le résumé automatique de textes.
- La reconnaissance de l'écriture manuscrite.

Les applications en relation avec le traitement du signal:

- La reconnaissance automatique de la parole.
- La synthèse de la parole.
- Le traitement de la parole.

Les applications en relation avec l'extraction d'information:

- La recherche d'information et la fouille de textes.
- La classification et la catégorisation de documents.

2. Langue arabe et TALN

Malgré de nombreuses recherches, la langue arabe est considérée comme une langue difficile à maitriser dans le domaine du traitement automatique de la langue à cause de sa richesse morphologique. 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.

3. Etat de l'art

Au cours des deux dernières décennies, de nombreux chercheurs ont proposé de nouvelles approches pour extraire les racines des mots arabes, Certaines de ces approches utilisent l'analyse morphologique, tandis que d'autres approches s'appuient sur des méthodes statistiques.

3.1. Etat de l'art sur l'analyse morphologique:

Pour extraire la racine d'un mot, Grefenstette et al.[18] Ainsi que Semmar et al.[34] ont présenté une méthode qui procède en deux étapes. Après une première étape de normalisation, la deuxième consiste à éliminer les affixes pour trouver le radical qui est vérifié dans un dictionnaire principal. S'il n'y existe pas, il faut alors appliquer des règles de réécriture qui sont décrites dans Darwish [23]. Le radical réécrit est vérifié à nouveau dans le dictionnaire principal.

Une approche similaire qui est basée sur la normalisation et l'élimination des affixes, mais sans utiliser de dictionnaires, a été proposée initialement par Aljlayl et Frieder[29] et ensuite modifiée par d'autres. Chen et Gey [1] ont identifié d'autres ensembles d'affixes et de règles.

Kanaan et al. [19] ont présenté un nouvel algorithme pour extraire des racines arabes quadrilatérales. Deux matrices temporaires sont utilisées, une pour sauvegarder les lettres du suffixe et une autre pour sauvegarder les racines.

Kadri et Nie [39] ont introduit deux nouvelles catégories d'affixes : les antéfixes qui sont situés avant les préfixes et les postfixes qui sont situées après les suffixes. Les deux catégories sont supprimées avant les préfixes et les suffixes.

Hawas [15] présente une nouvelle méthode d' extraction des racine pour les mots arabes qui essaie d'assigner une racine unique pour chaque mot arabe sans avoir une liste de racines, une liste de stems de mots ou la liste des préfixes et suffixes du mot. L'algorithme proposé prédit les positions des lettres qui peuvent former la racine une à une, en utilisant des règles fondées sur les relations entre les lettres et leurs placements dans le mot.

3.2. Etat de l'art sur l'analyse statistique :

Ahmed et Nürnberger [16] ont présenté le modèle N-gram qui peut être utilisé pour calculer la ressemblance entre deux chaînes de caractères en comptant le nombre des N-grammes semblables qu'ils partagent. Le coefficient de ressemblance est donné par l'équation (1):

$$\delta_{\mathrm{n}}(a,b) = \frac{|\alpha \cap \beta|}{|\alpha \cup \beta|} \quad (1)$$

Où α et β sont les ensembles de N-gram.

Mustafa [30] a testé la performance des techniques basées sur l'analyse statistique dans la recherche des racines arabes. Les techniques ont été testées sur un corpus textuel qui contient environ 25 mille mots (avec une dimension totale d'environ 160KB) et un ensemble de 100 mots de requête textuelle. Les résultats de l'approche hybride ont montré une performance meilleure par rapport à l'approche contiguë.

Sanan et al. [32] ont proposé une approche statistique qui est basée sur le mot et les lettres. Quatre types d'expression ont été explorés : mot, racine lexicale, racine et n-gram, parfois séparés et parfois en associés. Ils ont trouvé que les n-grams basés sur les stems sont meilleurs que ceux basés sur les mots.

Khreisat [26] a proposé une autre approche statistique pour classer des documents arabes. La technique emploie une mesure de la dissemblance appelée «Distance Manhattan» et une mesure de ressemblance appelée «Dice measure». Un corpus de documents arabes a été collecté des journaux arabes en ligne. 40% du corpus a été utilisé pour l'apprentissage et le reste pour la classification. Une phase de normalisation a été utilisée.

4. Processus d'extraction de la racine d'un mot arabe

Une étape préalable à toute analyse consiste à appliquer des traitements communs qui seront détaillés dans la section qui suit. L'étape suivante de ce processus consiste à vérifier si le mot à traiter est un mot outil (Figure2.1). En général, cette vérification est faite dans une base de données qui contient l'ensemble des mots isolés, [5]. La dernière étape consiste à appliquer une méthode d'extraction de la racine. Cette méthode peut être basée sur l'analyse morphologique ou sur des calculs statistiques de la langue arabe.

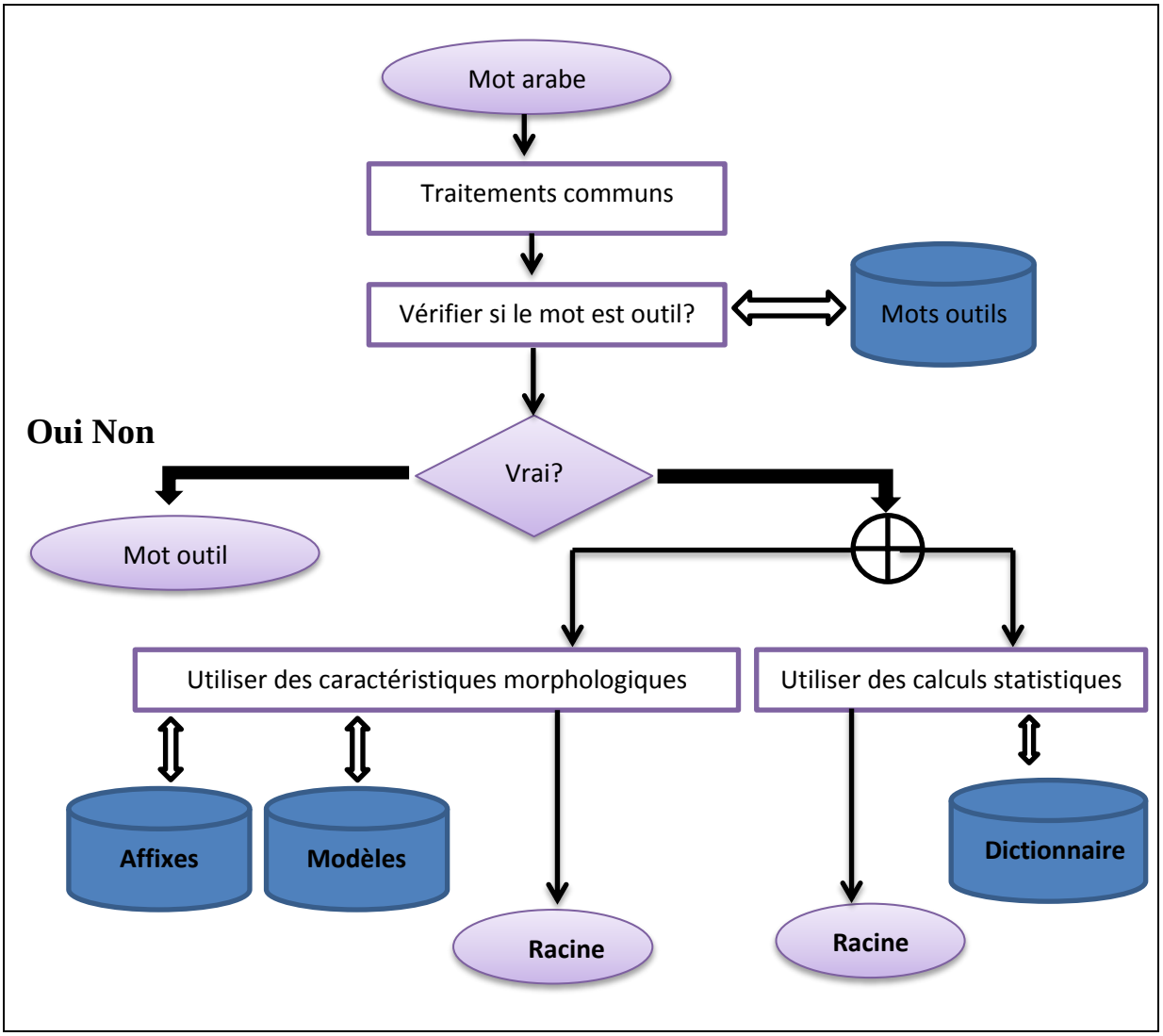


Figure 2.1: Processus d'extraction de la racine d'un mot arabe.

4.1. Traitements communs

La première phase du traitement automatique d'un texte arabe est la spécification du codage utilisé dans ce texte. Vient ensuite, la phase de normalisation qui est commune à la plupart de ces méthodes. Par contre, la phase de transcription est utilisée par peu de méthodes.

4.1.1. Normalisation

La phase de normalisation vise à transformer une copie du document original dans un format standard plus facilement manipulable. Cette étape est considérée nécessaire à cause des variations qui peuvent exister lors de l'écriture d'une même unité lexicale. [39]

Le document est normalisé comme suit:

- Suppression des caractères spéciaux et les chiffres.
- Remplacement de إ , آ et أ par ا.
- Remplacement de la lettre finale ي par ى.
- Remplacement de la lettre finale ة par ه.

En effet, plusieurs mots, ayant des sens différents, peuvent avoir la même forme normalisée.

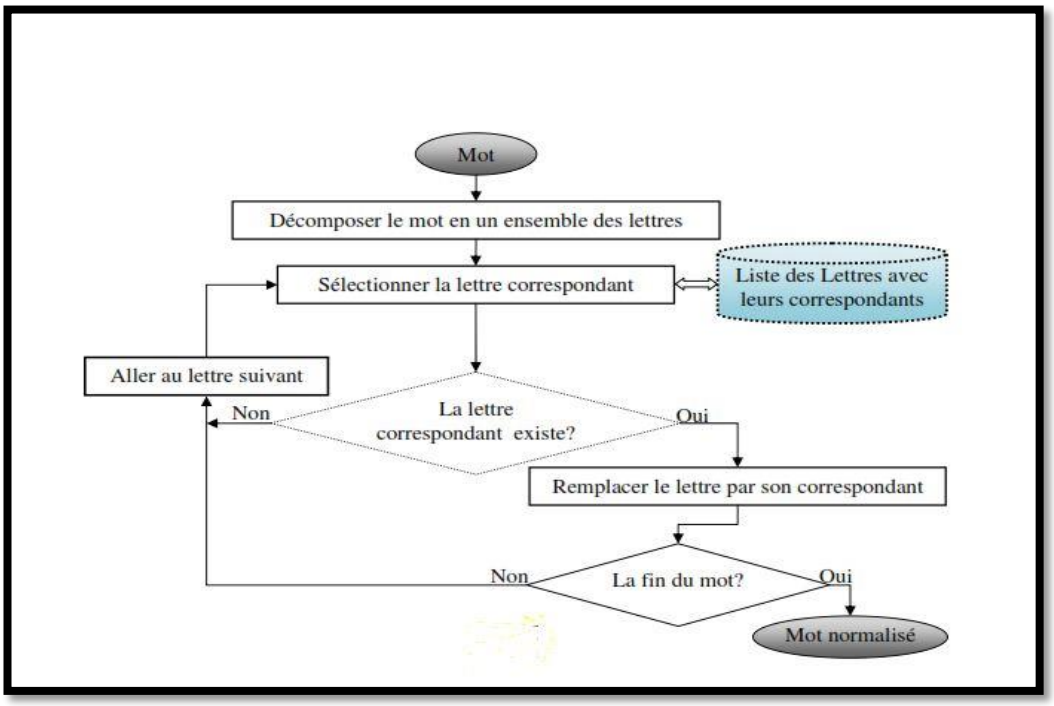


Figure 2.2.: Processus de normalisation. [3]

4.1.2. Élimination des mots outils

Est une phase qui vise à supprimer les particules, les propositions, les conjonctions, les adverbes et les quantificateurs à partir d'une base de donnes, qui contient ces mots outils.

4.2. Les méthodes d'extraction de la racine d'un mot arabe

Dans cette partie, nous présentons quelques méthodes d'extraction de la racine d'un mot arabe. Ces méthodes sont citées et réparties en trois grandes classes : les méthodes basées sur l'analyse statistique, les méthodes basées sur l'analyse morphologique et les méthodes hybrides (basées sur l'analyse morphologique et statistique).

La figure 2.3 présente une classification générale des méthodes d'extraction de la racine d'un mot arabe à trois niveaux.

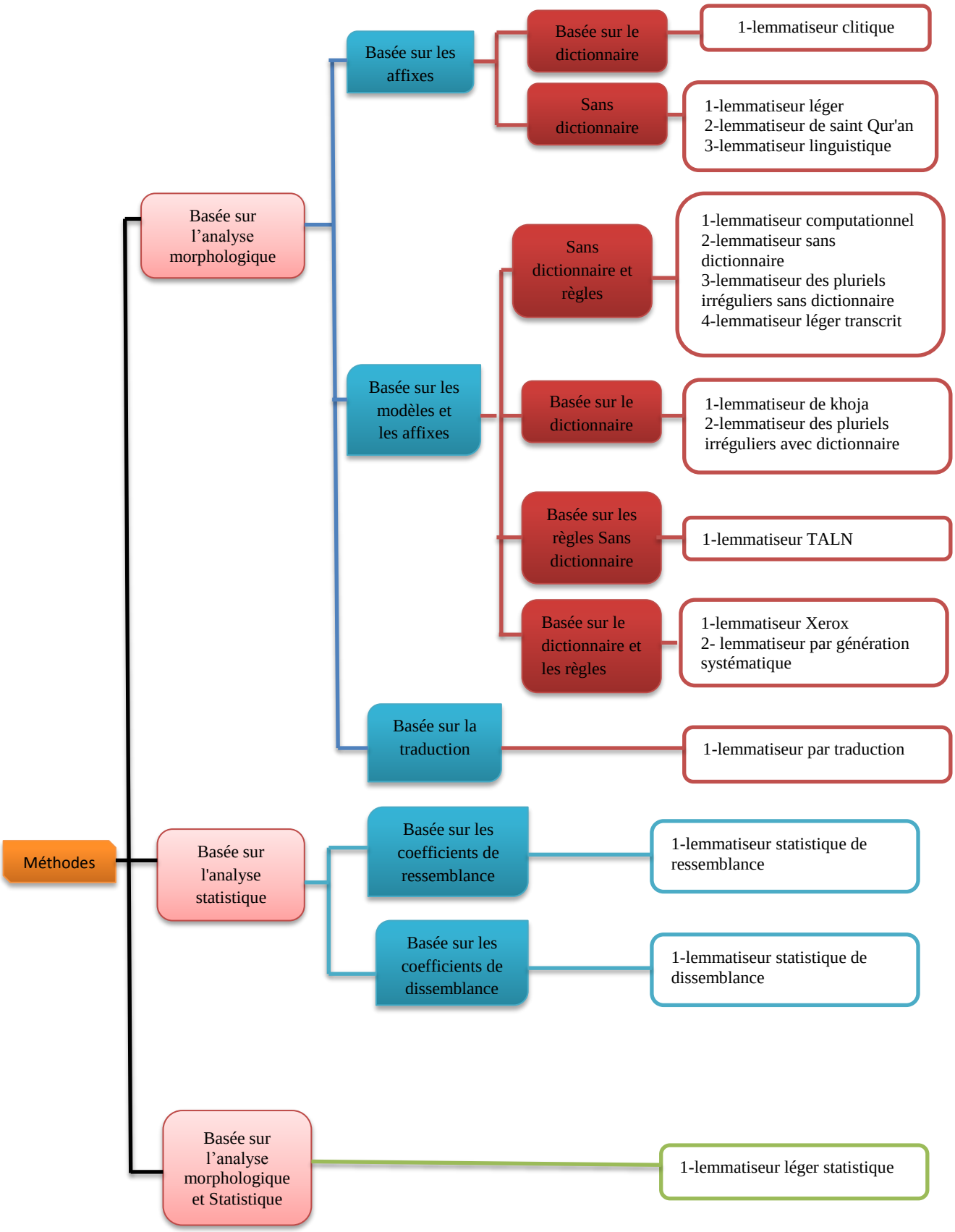


Figure 2.3.: Schéma générale de la classification des méthodes d'extraction de la racine en arabe.

4.2.1. Classe de l'analyse morphologique

L'analyse morphologique sert à séparer et identifier des morphèmes semblables aux mots préfixés, comme les conjonctions, les prépositions préfixées, l'article défini, les suffixes de pronom possessif. La phase d'analyse morphologique détermine un schème possible. Les préfixes et suffixes sont trouvés en enlevant progressivement des préfixes et des suffixes candidats et en essayant de faire correspondre toutes les racines produites par un schème afin de retrouver la racine. Les méthodes basées sur l'analyse morphologique sont reparties en trois catégories: suppression des affixes, reconnaissance des modèles et traitement des affixes ou traduction. [23]

4.2.1.1. Basée sur les affixes

Dans cette catégorie, les méthodes repèrent les différents types d'affixes du mot à traiter et ensuite les suppriment afin d'extraire la racine probable. Deux approches sont mises en place, la première utilise un dictionnaire pour valider la racine probable. Et la deuxième considère la racine extraite comme finale.

a. Basée sur les affixes avec dictionnaire

Dans cette approche, on trouve la méthode : lemmatiseur clitique. Pour extraire la racine d'un mot avec cette méthode, il faut appliquer une première étape de normalisation et une deuxième étape qui consiste à trouver le radical, par l'élimination des clitiques, en utilisant les dictionnaires de proclitiques et d'enclitiques. Ensuite, ce radical est vérifié dans le dictionnaire principal. S'il n'y existe pas, il faut alors appliquer des règles de réécriture. Ces règles sont décrites dans Darwish [23]. Le radical réécrit est vérifié à nouveau dans le dictionnaire principal (Figure 2.4). Cette méthode utilise les ressources linguistiques suivantes : un dictionnaire principal (5,4 millions d'entrées) contenant pour chaque mot ses caractéristiques possibles (genre, nombre, etc.), un dictionnaire de proclitiques (77 entrées) et un dictionnaire d'enclitiques (65 entrées).

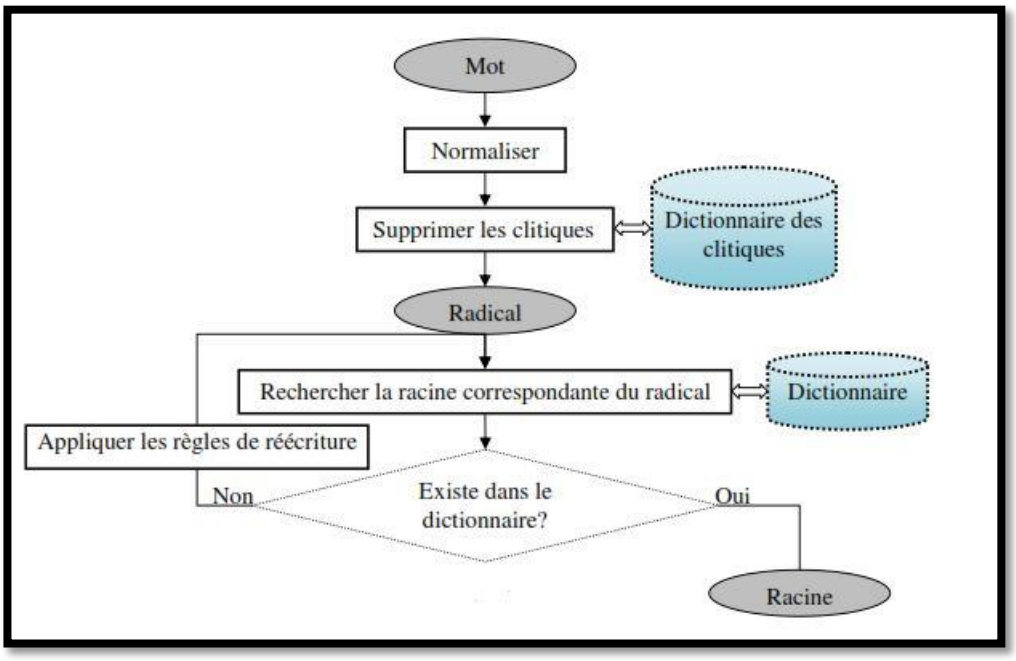


Figure 2.4.: Lemmatiseur clitique. [3]

b. Basée sur les affixes sans dictionnaire

Dans cette approche plusieurs méthodes sont proposées, ces méthodes sont : lemmatiseur léger, lemmatiseur du saint Qur'an et lemmatiseur linguistique.

b.1. Lemmatiseur léger:

Cette méthode est proposée par Aljlayl et Frieder [29] . Elle est principalement basée sur la normalisation et l'élimination des affixes pour les mots de longueur supérieure ou égale à trois lettres. Chen et Gey [1] ont identifié d'autres ensembles d'affixes. Pour supprimer les affixes à partir des ensembles déjà définis, chaque algorithme propose ses propres règles (Figure 2.5).

Par exemple, ils appliquent les règles suivantes: Si le mot est formé d'au moins cinq lettres, on enlève les préfixes de trois lettres suivants: <لال, lal>, <سال, sal>, <مال, mal>, <كال, kal>, <ولل,wull>, <اال, aal>, <فال, fal>, <بال, bal> et <ال, al>. Si le mot est formé d'au moins quatre lettres, on enlève les préfixes de deux lettres suivants: <فا, fa>, <كا, ka>, <ول, wul>,<وي, wuyi>, <وس,wus>, <سي, syi>, <لا, la>, <وب, wub>, <وت, wut>, < وم ,wum, >, <لل, ll>, <با, ba>, <وا, wua>, <ال,al>. Si le mot est formé d'au moins quatre lettres et commence par <و, wu> on élimine<و, wu> . Si le mot est formé d'au moins quatre lettres et commence par <ب, b> ou <ل, l> on élimine <ب, b> ou <ل, l>.

Al Ameed et al. [21] ainsi que Larkey et al. [28] ont développé un algorithme qui est basé sur la suppression de «و» au début du mot, des articles définies (<لل, ll>, <فال, fal>, <كال, kal>,<بال, bal>, <وال, wual>, <ال, al>) et des suffixes (<ي, yi >, <ة, ah>, <ه , h>, <ييه, yiyiah>, <يه,yih>, <ين, yin>, <ون, wun>, <ات, at>, <ان, an>, <ها, ha>).

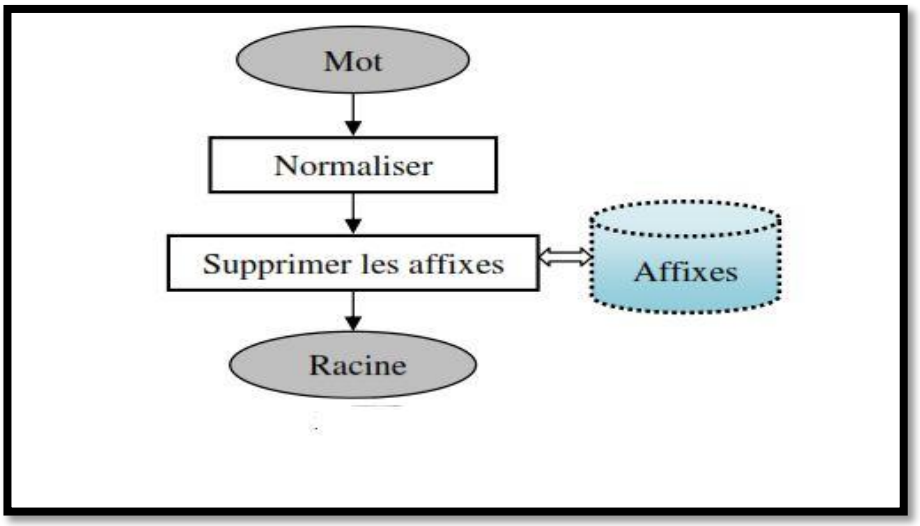


Figure 2.5: Lemmatiseur léger.[3]

b.2. Lemmatiseur du saint Qur'an:

N. Thabet [35] a proposé une nouvelle approche basée sur la technique du Lemmatiseur léger. Cette approche a été appliquée sur une version transcrite du Saint Coran en anglais (Figure 2.6). Dans cette méthode, les préfixes sont d'abord supprimés, ensuite, les suffixes qui sont répartis en six groupes selon leurs longueurs, sont supprimés pour produire la racine transcrite. Cette racine est ensuite de-transcrite pour obtenir la racine arabe.

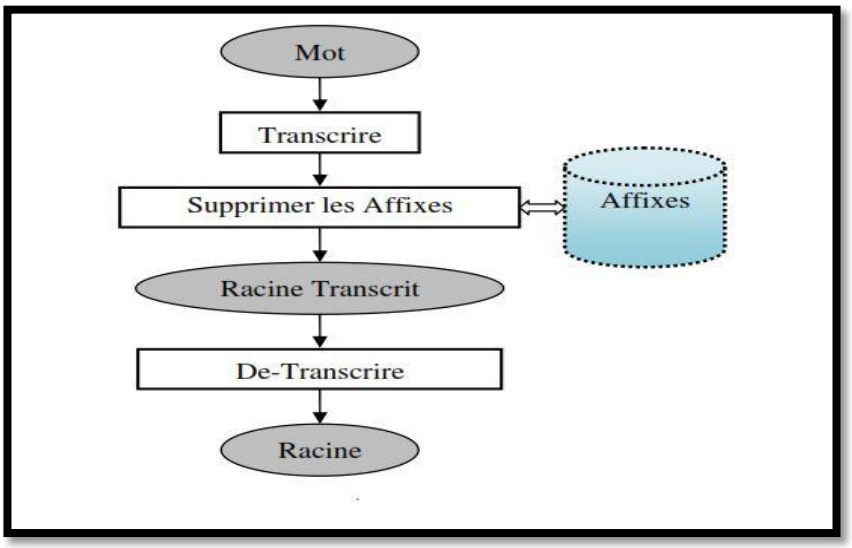


Figure 2.6: Lemmatiseur du Saint Qur'an [3]

b.3. Lemmatiseur linguistique:

Kadri et Nie [39] ont proposé une nouvelle méthode qui est basés sur la technique Lemmatiseur léger (Figure 2.7). Dans cette méthode, deux nouvelles catégories d'affixes sont introduites : les antéfixes qui sont situés avant les préfixes et les postfixes qui sont situés après les suffixes. Les deux catégories sont supprimées avant les préfixes et les suffixes.

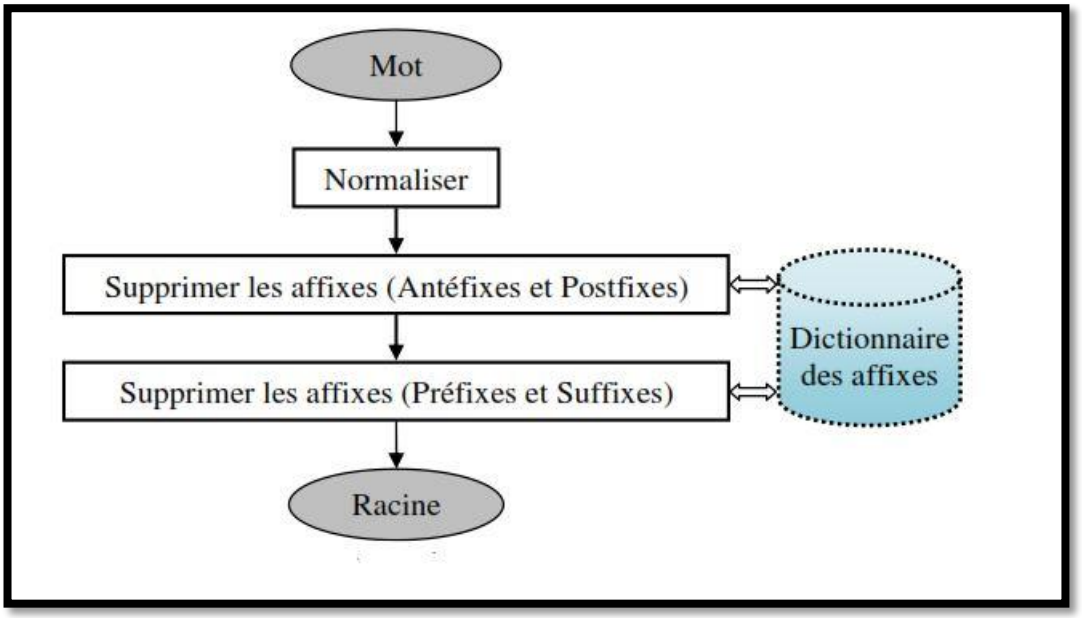


Figure 2.7: Lemmatiseur linguistique.[3]

4.2.1.2. Basée sur les modèles les affixes :

Dans cette catégorie, Il y a quatre approches. La première consiste à traiter les affixes et rechercher les modèles correspondants. La deuxième consiste à utiliser un dictionnaire pour valider la racine extraite de la même façon que l'approche précédente. La troisième consiste à appliquer des règles spécifiques et prédéfinies sur la racine extraite de la même façon que la première. La quatrième consiste à générer dans une phase préalable, un dictionnaire global qui contient la plupart des mots arabes et de leurs racines. Ensuite, pour extraire la racine d'un mot donné, une simple recherche est effectuée dans le dictionnaire ainsi construit.

a. Basée sur les modèles les affixes sans dictionnaire**:**

Dans cette approche, plusieurs méthodes sont proposées:

a.1. Lemmatiseur computationnel**:**

Cette méthode traite seulement les racines trilatérales et procède en deux étapes principales. La première consiste à enlever le préfixe le plus long. En effet, les trois lettres de la racine doivent se trouver quelque part dans les quatre ou cinq premières lettres qui en restent. La deuxième étape permet d'extraire la racine en faisant une comparaison entre tous les trigrammes possibles qui sont formés à partir des cinq premières lettres et les modèles enregistrés dans une base de données (Figure 2.8). [3]

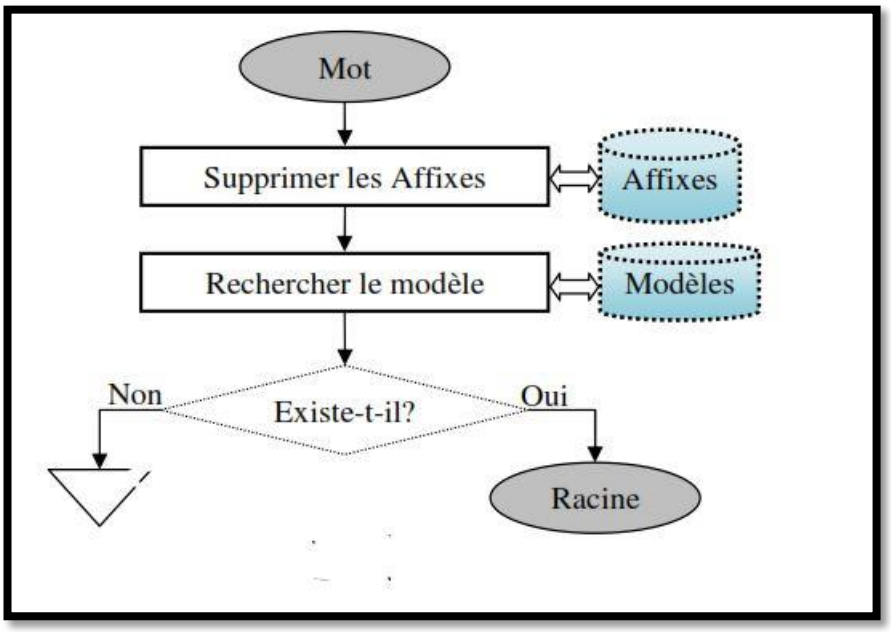


Figure 2.8 : Lemmatiseur computationnel. [3]

a.2. Lemmatiseur sans dictionnaire:

Taghva et al. [25] ont présenté cet algorithme pour extraire des racines quadrilatérales. Plusieurs ensembles ont été définis, comme il est présenté dans le tableau (2.1) :

Ensemble	Signification
D	les diacritiques.
P1	les préfixes de longueur un.
P2	les préfixes de longueur deux.
P3	les préfixes de longueur trois.
S1	les suffixes de longueur un.
S2	les suffixes de longueur deux.
S3	les suffixes de longueur trois.
PR4	les modèles de longueur quatre.
PR53	les modèles de longueur cinq dont la racine est de longueur trois.
PR54	les modèles de longueur cinq dont la racine est de longueur quatre.
PR63	les modèles de longueur six dont la racine est de longueur trois.
PR64	les modèles de longueur six dont la racine est de longueur quatre.

Tableau 2.1: table des abréviations.

Cet algorithme contient les étapes suivantes La première étape est la normalisation. Ensuite, l'extraction de la racine se fait selon la longueur du mot. Si la longueur est 4, on extrait le stem pertinent et on le retourne en enlevant les préfixes et les suffixes de longueur 1 (S1, P1). Si la Longueur est 5, on extrait le stem de longueur 3 (modèles PR53), si aucun de ces modèles ne correspond, alors on enlève les préfixes et les suffixes pour avoir un stem de longueur 3. Si le mot reste de longueur 5, il faut utiliser le modèle PR54 pour déterminer s'il contient un stem de longueur 4.

a.3. Lemmatiseur des pluriels irréguliers sans dictionnaire:

Pour extraire les racines à partir des pluriels irréguliers, Cette méthode procède en deux étapes. La première utilise la méthode Lemmatiseur léger [29], pour produire le Stem. La deuxième consiste à rechercher le modèle correspondant dans un ensemble de 39 modèles, de pluriels irréguliers, répertoriés par les auteurs à partir des livres de grammaire (Figure 2.10).

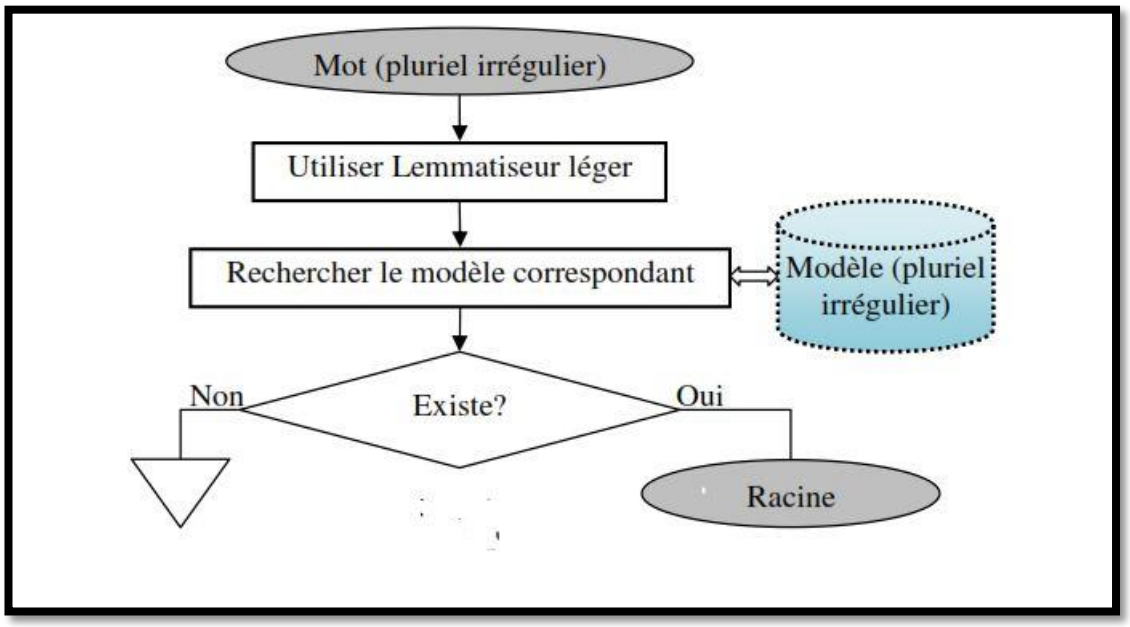


Figure 2.9: Lemmatiseur des pluriels irréguliers sans dictionnaire. [3]

a.4. Lemmatiseur léger transcrit**:**

Cette méthode propose en sortie une liste des racines possibles. Elle a une phase préalable d'initialisation qui consiste à construire trois listes, une pour les modèles, une pour les préfixes et une pour les suffixes. Ces listes sont construites à partir d'un ensemble de couples (racine, mot dérivé). Un taux d'apparition est calculé pour chaque élément de chacune de ces listes. Pendant la phase de l'extraction de la racine, cette méthode procède en cinq étapes. Les deux premières sont la transcription et la normalisation. La troisième consiste à sélectionner les 0 à 3 premières lettres consécutives et à les supprimer s'ils existent dans la liste de préfixes déjà définie dans la phase préalable. La quatrième consiste à sélectionner les 0 à 3 dernières lettres consécutives et à les supprimer s'ils existent dans la liste des suffixes. La cinquième consiste à extraire la racine selon la longueur du stem. S'il s'agit d'un stem de trois lettres alors il est considéré comme racine, sinon on effectue une recherche du modèle correspondant dans la liste de modèles, pour déduire la racine. Les trois dernières étapes sont répétées pour chaque préfixe et suffixe. Les racines possibles sont ajoutées à une liste qui présente la sortie de cette méthode. [3]

b. Basée sur les modèles les affixes avec le dictionnaire:

Dans cette approche plusieurs méthodes sont proposées, ces méthodes sont :

b.1. Lemmatiseur de khoja **:**

Cette méthode est proposée par Khoja et Garside [37] qui consiste à enlever les affixes après une première étape de normalisation. Ensuite, le résultat est comparé avec une liste de modèles. Si une correspondance est trouvée, les lettres représentant la racine dans le modèle sont extraites. Ensuite, la racine ainsi extraite est validée dans un dictionnaire (Figure 2.12).

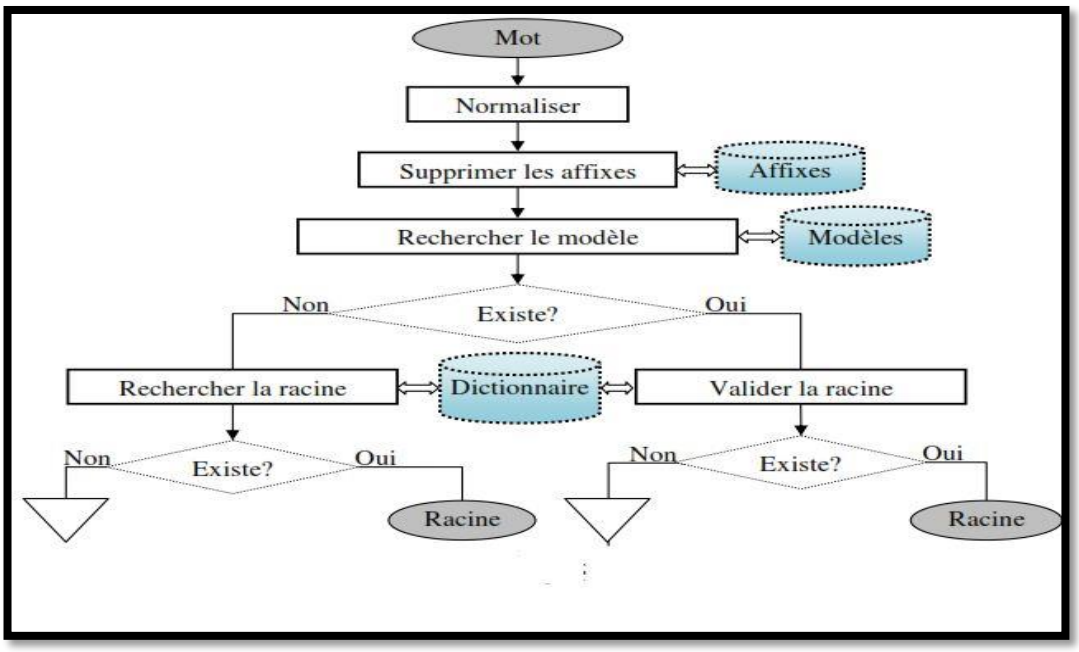


Figure 2.10: Lemmatiseur de Khoja. [3]

b.2. Lemmatiseur des pluriels irréguliers avec dictionnaire**:**

Cette méthode sert à trouver la racine d'un mot qui est un pluriel irrégulier. Le principe de cette méthode est simple : rechercher simplement la racine dans une base de données (Figure 2.13). Le dictionnaire a été construit manuellement et validé par un linguiste à partir de 127.000 types de Stems, pour récupérer tous les types des modèles de pluriels irréguliers. Une liste d'environ 3600 Stems de pluriels irréguliers a été extraite et triée par ordre alphabétique et en fonction de chaque modèle du pluriel irrégulier. [2]

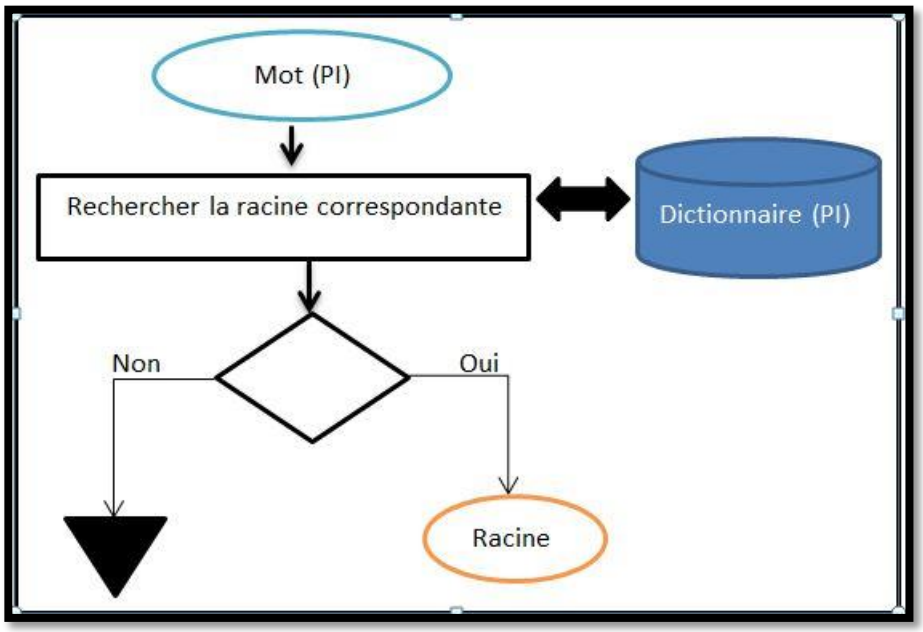


Figure 2.11: lemmatiseur des pluriels irréguliers avec dictionnaire. [3]

c. Basée sur les modèles les affixes avec des règles et sans dictionnaire **:**

Dans cette approche, la méthode lemmatiseur TALN est proposée. Cette méthode procède en trois étapes principales (Figure 2.14). La première est la normalisation. La deuxième consiste à extraire le stem en supprimant les lettres qui représentent les caractéristiques flexionnelles (temps, nombre, personne, etc.). La dernière étape consiste à rechercher le modèle qui a été utilisé pour construire le lexème obtenu dans l'étape précédente. La racine est ensuite déduite à partir du lexème et du modèle, s'il a été trouvé. [3]

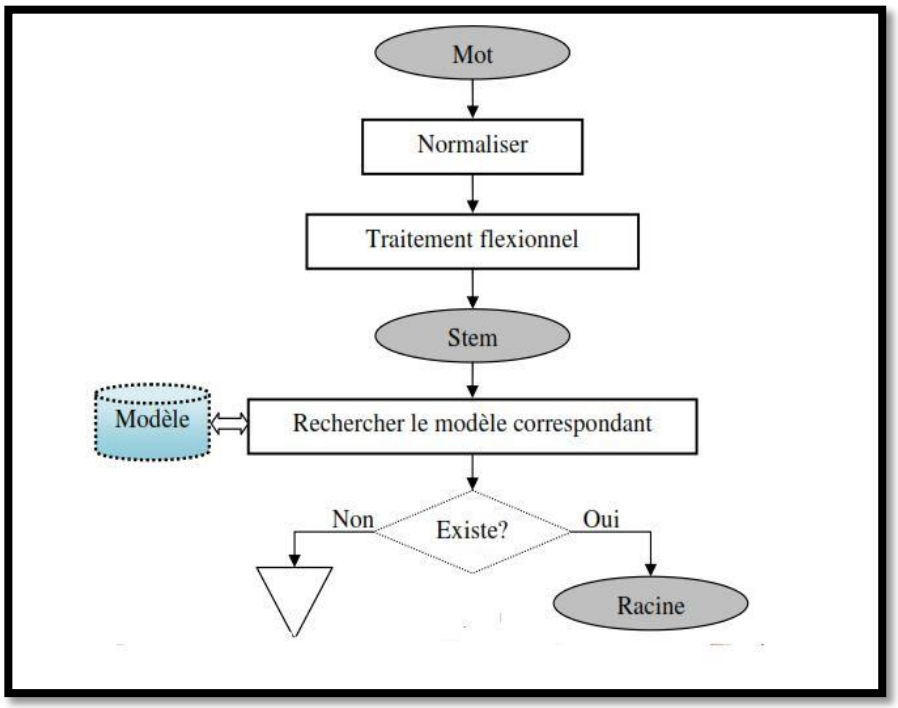


Figure2.12: Lemmatiseur TALN. [3]

d. Basée sur les modèles les affixes avec des règles et dictionnaire**:**

Dans cette approche plusieurs méthodes sont proposées, ces méthodes sont :

d.1. Lemmatiseur Xerox :

Cette méthode procède en deux phases (Figure 2.15, Figure 2.16). Dans une phase préalable, on construit un dictionnaire global qui contient la plupart des mots arabes et leurs racines. Ensuite, pour extraire la racine d'un mot donné, une simple recherche est effectuée dans le dictionnaire ainsi construit.

La génération du dictionnaire est faite en utilisant des racines, des modèles, des règles et des affixes. Le centre Européen de recherche Xerox a proposé un analyseur morphologique pour l'arabe standard moderne et qui est basé sur des dictionnaires. Le système a été largement modifié en utilisant la technologie d'états finis de Xerox. [24]

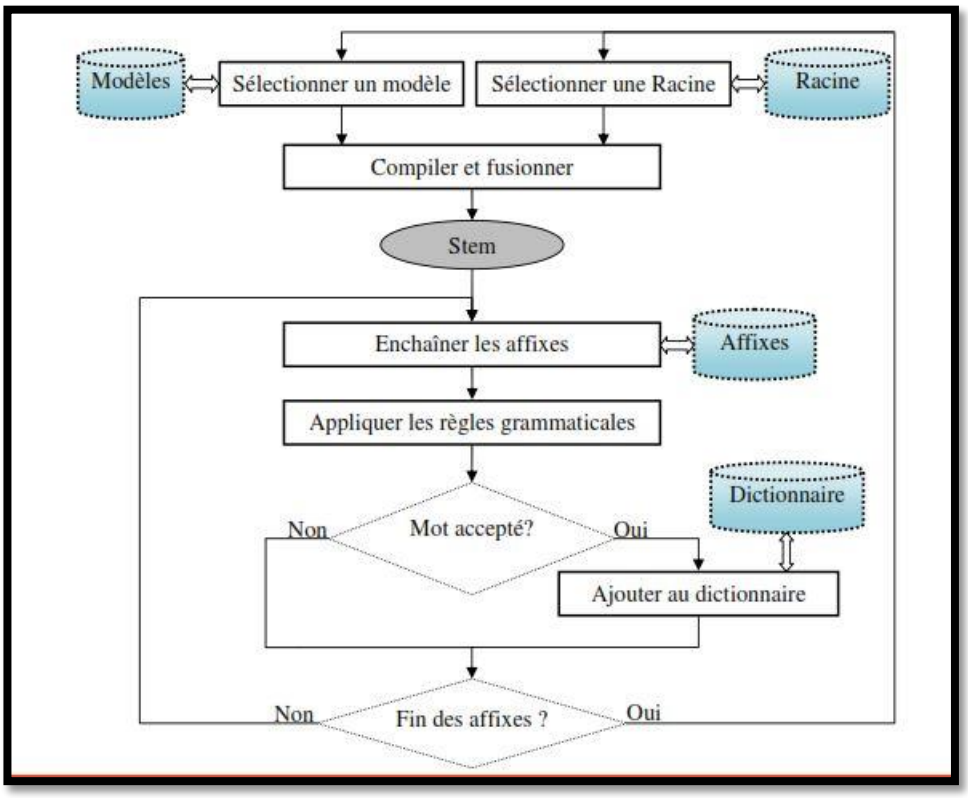


Figure 2.13: Lemmatiseur Xerox: Génération du dictionnaire. [3]

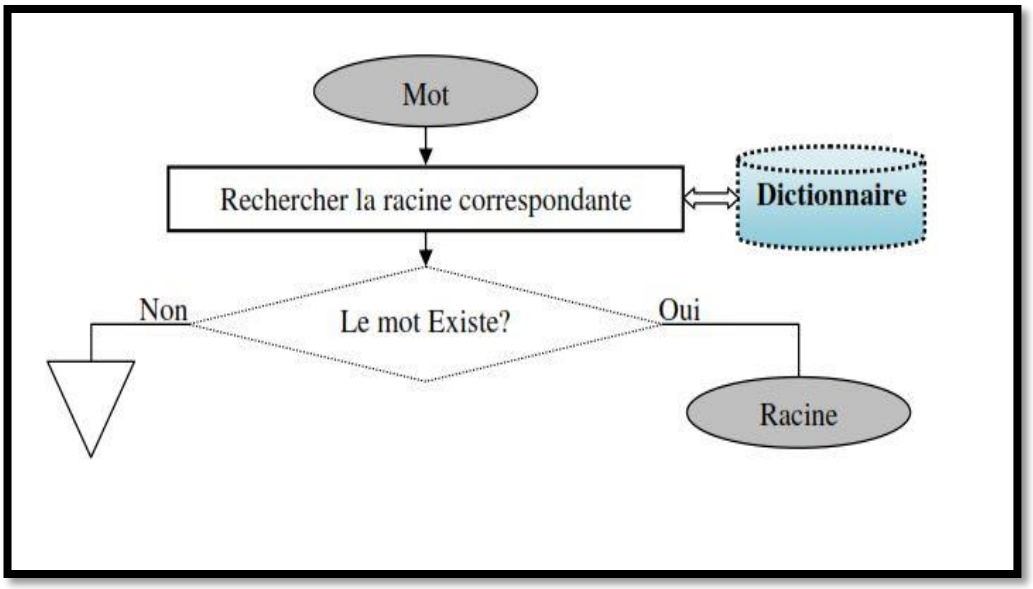


Figure 2.14: Lemmatiseur Xerox: Recherche. [3]

d.2. Lemmatiseur par génération systématique**:**

Kadri et Yagi [22] ont proposé une autre méthode pour construire un dictionnaire global. Les racines sont ordonnées et représentées selon le nombre des radicaux par une notation ensembliste: {F, M, L, Q}. Cet ensemble représente le premier radical (F), le radical médiale (M), le dernier radical dans une racine trilitère (L), et le dernier radical

dans une racine quadrilatère (Q). Un modèle est aussi représenté par un ensemble $\{x_1,\ldots,x_n\}$ ou x_i désigne la lettre à la $i^{ème}$ position du modèle. Pour générer un mot à partir d'une racine et d'un modèle, on doit faire correspondre la position de chaque élément de l'ensemble {F, M, L, Q} de la racine avec une position d'un élément x_i de l'ensemble $\{x_1,\ldots, x_n\}$ du modèle. Pour générer le dictionnaire global, le processus précédent est itéré sur tous les modèles et les racines. Dans cette méthode, plusieurs règles de transformation sont utilisées dans le processus de génération du mot.
(Figure 2.17)

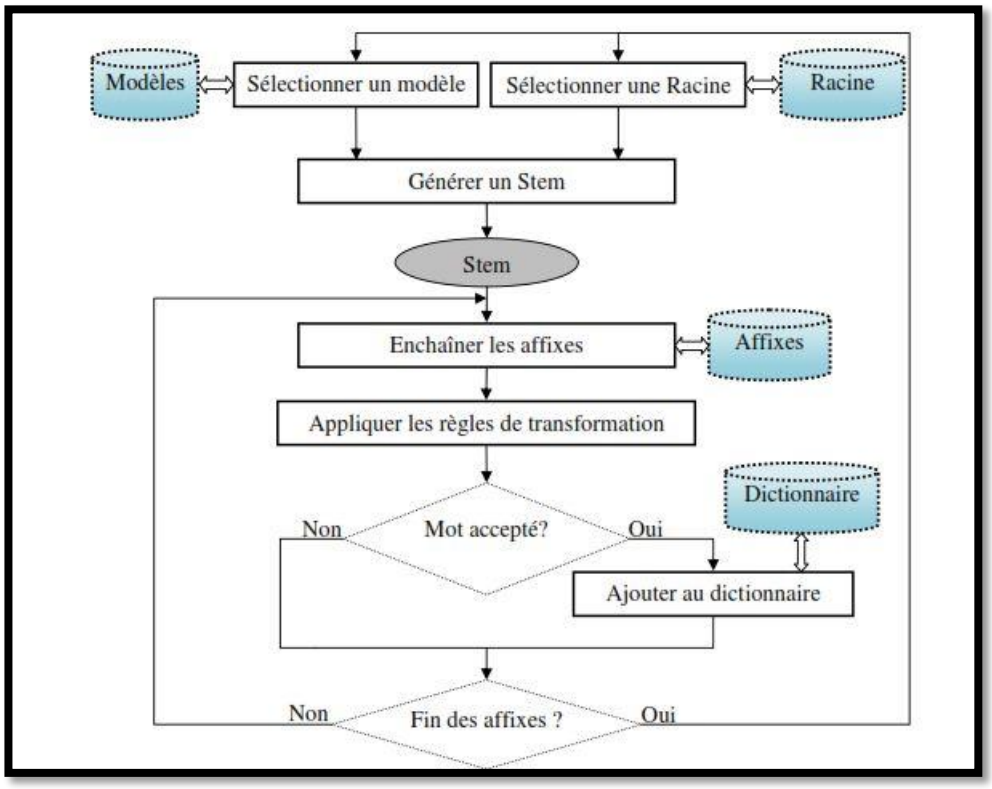


Figure 2.15: lemmatiseur par génération systématique. [3]

4.2.1.3. La traduction

Les lemmatiseurs ont de meilleures performances dans la langue anglaise. Une seule méthode est proposée dans cette catégorie qui est celle du lemmatiseur par traduction. Cette méthode utilise le traducteur «Ajeeb Online» pour traduire le mot en anglais. Ensuite, elle applique une méthode d'extraction de la racine en anglais. Puis elle applique de nouveau le traducteur pour traduire la racine en arabe (Figure 2.18).

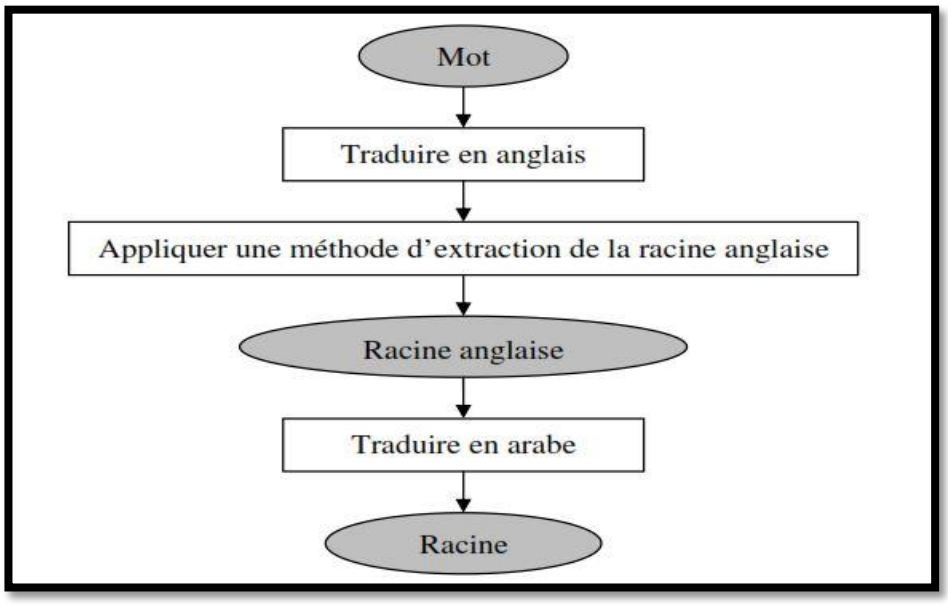


Figure 2.16: Lemmatiseur Par Traduction. [3]

4.2.2. Classe de l'analyse statistique

L'analyse statistique consiste à découper le mot et les racines d'un dictionnaire en N-grams, à calculer des coefficients de correspondance entre les N-grams du mot et des racines et à maximiser ou minimiser ces coefficients. Dans cette classe, il y a deux catégories : la première calcule les coefficients de la ressemblance et la deuxième calcule les coefficients de la dissemblance.

4.2.2.1. Calcul des coefficients de la ressemblance **:**

Adamson et Boreham [4] ont développé la première technique de classification automatique qui est basée sur la structure des mots. Le coefficient de ressemblance est calculé à partir du nombre de digrammes (2-gram) assortis dans de paires des sous-lettres. Un échantillon de mots d'une base de données chimique a été choisi. Cette base contient certains stems dérivés des noms des éléments chimiques. Chaque groupe est caractérisé par une racine et ses mots dérivés (Figure 2.19). Ahmed et Nürnberger [16] ont présenté le modèle N-gram qui peut être utilisé pour calculer la ressemblance entre deux chaînes de caractères en comptant le nombre des N-grams semblables qu'ils partagent. Le coefficient de ressemblance est donné par l'équation (1):

$$\delta_{\mathrm{n}}(a,b) = \frac{|\alpha \cap \beta|}{|\alpha \cup \beta|} \quad (1)$$

Où α et β sont les ensembles de N-gram.

Yousef et al. [36] ont proposé un nouvel algorithme en utilisant la technique des N-grammes de caractères. La similitude entre le mot et la racine dans cet algorithme est calculée par l'équation(2):

$$S = \frac{2C}{A+B} \quad (2)$$

Où : A:Nombre des bi-grammes uniques dans le mot (A),

B: Nombre des bi-grammes uniques dans la racine (B),

C: Nombre de paires uniques similaires entre le mot (A) et la racine (B). Le mot (A) et les racines potentielles (B) à comparer avec, puis la mesure de similarité est effectuée en calculant la valeur de S entre le mot (A) et chacunune des racines potentielles(B) à comparer avec, puis la mesure de similarité est effectuée en calculant la valeur de S entre le mot (A) et chacune des racines potentielles(B).

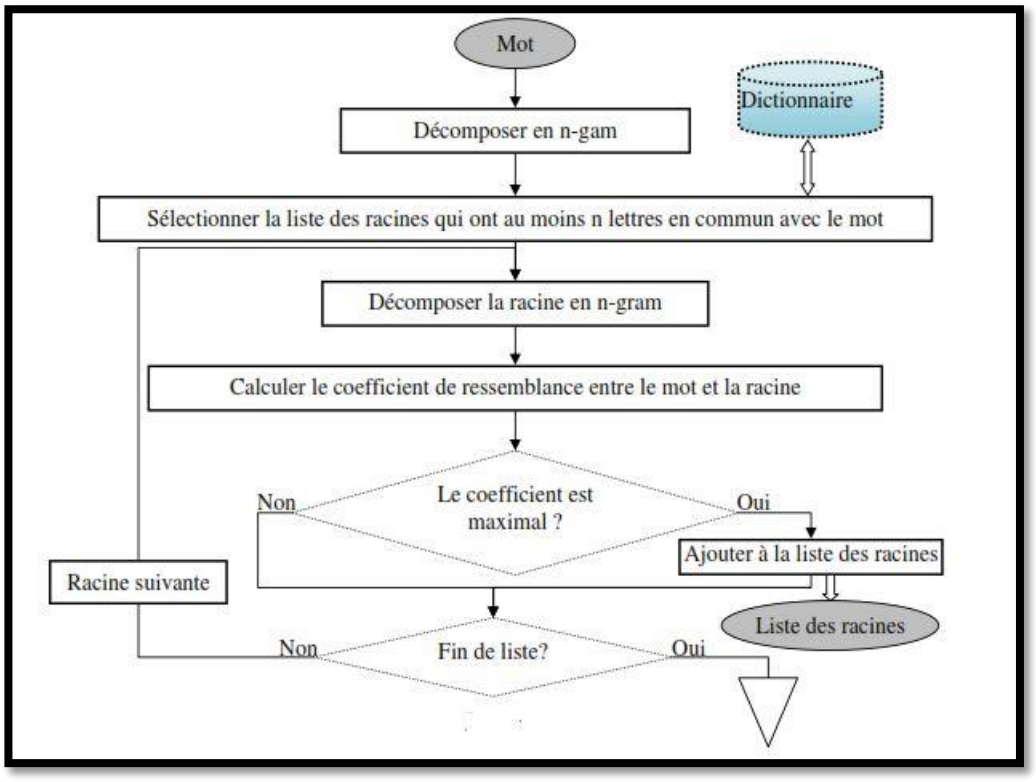


Figure 2.17: Analyse statistique (n-gram) basée sur le coefficient de ressemblance.[3]

4.2.2.2. Calcul des coefficients de la dissemblance **:**

Khreisat [26] a proposé une approche statistique pour classer des documents arabes. La technique utilise une mesure de la dissemblance appelée «Distance Manhattan» et

une mesure de ressemblance appelée "Dice measure". Un corpus des documents des textes arabes a été collecté des journaux arabes en ligne. 40% du corpus a été utilisé pour l'apprentissage et le reste pour la classification. Une phase de normalisation a été utilisée. Il a aussi utilisé le 3-gram (Figure 2.20). Le coefficient de dissemblance entre les mots a et b est calculé en partitionnant chaque mot en 2-gram ou bien en 3-gram.

α ∩β présente l'intersection entre les deux ensembles et α U β l'union entre les deux ensembles (3):

$$\delta_n(a,b) = \frac{|\alpha \cup \beta| - |\alpha \cap \beta|}{|\alpha \cup \beta|} \quad (3)$$

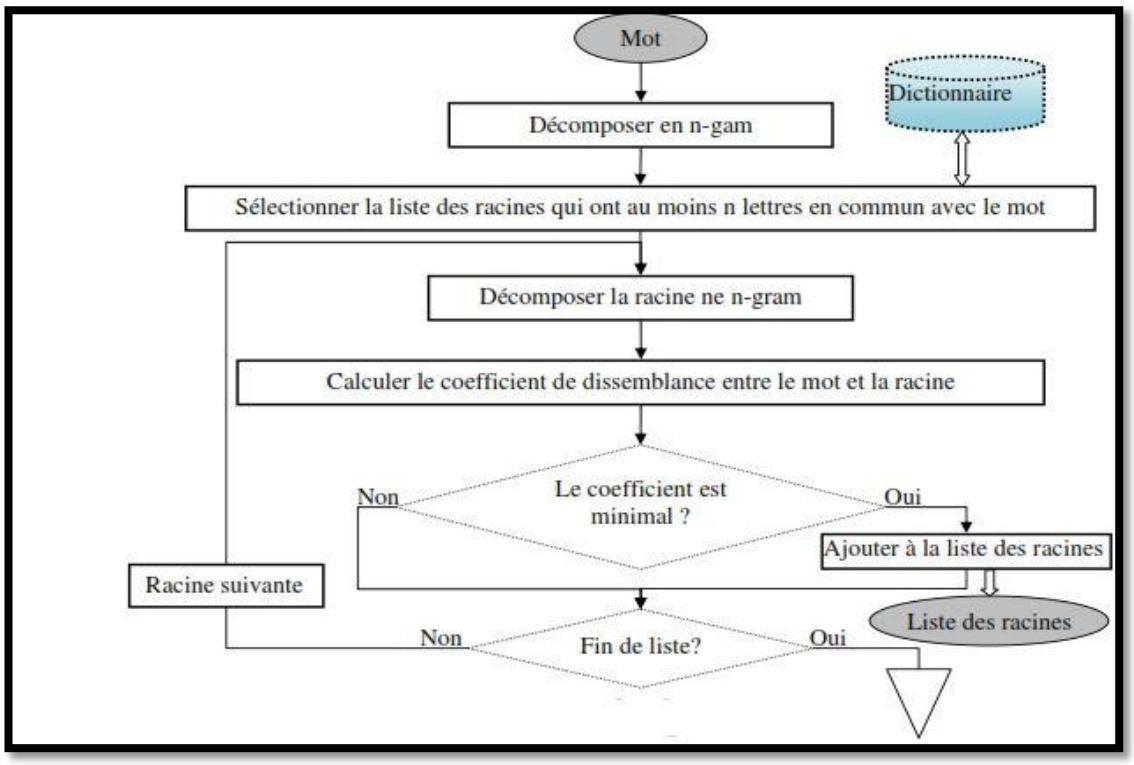


Figure 2.18: Analyse statistique (n-gram) basée sur le coefficient de dissemblance. [3]

4.2.3. Classe de l'analyse morphologique et statistique (hybride)

Les méthodes dans cette approche consistent à appliquer plusieurs étapes pour extraire la racine. La première étape consiste à appliquer l'analyse morphologique. La deuxième étape consiste à appliquer l'analyse statistique, en calculant les coefficients (de ressemblance et de dissemblance) entre le mot et une liste de racines d'un dictionnaire [26].

Roeck et Al-Fares [8] ont proposé la méthode lemmatiseur léger statistique. La première étape consiste à appliquer la méthode Lemmatiseur léger afin de supprimer les affixes. Dans la deuxième étape, on calcule les coefficients de ressemblance entre le stem donné par la première étape et une liste de racines sélectionnées dans un dictionnaire (Figure 2.21). Les racines correspondant aux coefficients de ressemblance maximum sont ajoutées à la liste de racines.

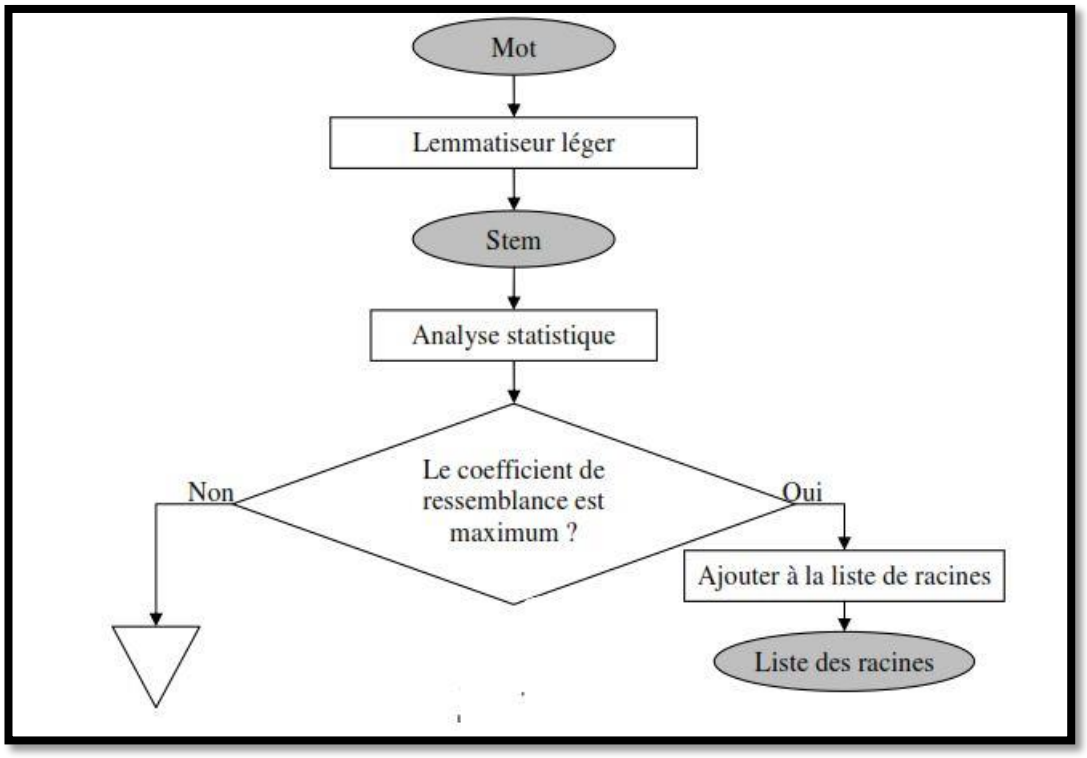


Figure 2. 19: Lemmatiseur Leger Statistique. [3]

5. Conclusion

Dans ce chapitre, nous avons présenté les méthodes d'extraction des racines arabes et nous avons présenté une classification de ces méthodes. Les classes principales sont basées sur des analyses morphologiques ou statistiques. Les analyses morphologiques consistent à identifier les morphèmes d'un mot, des affixes (préfixe, infixe, et suffixe), du modèle et de la racine. L'analyse statistique consiste à déterminer le coefficient de ressemblance et de dissemblance sémantique entre deux mots. Deux mots sont considérés semblables, s'ils ont en commun plusieurs sous-chaînes des N lettres. Dans les deux approches, le dictionnaire a un rôle important puisqu'il est utilisé pour valider les résultats obtenus.

Nous avons abordé dans le chapitre précédent une étude bibliographique en présentant les méthodes de traitement morphologiques et statistiques des mots arabes. Dans ce chapitre, nous allons mettre en œuvre l'implémentation de notre méthode que nous avons choisi et nous commençons par la description du notre système réalisé. Puis les différents outils utilisés, tels que: la présentation de l'environnement de développement, et nous expliquons le déroulement de l'application.

1. Description du système réalisé

1.1 Processus d'extraction du stem

Une étape préalable à toute analyse qui permet de vérifier si le mot à traiter est un mot outil (stopwords). Cette vérification est faite dans une base de données qui contient l'ensemble des mots outils. Puis appliquer notre méthode hybride pour l'extraction des stems.

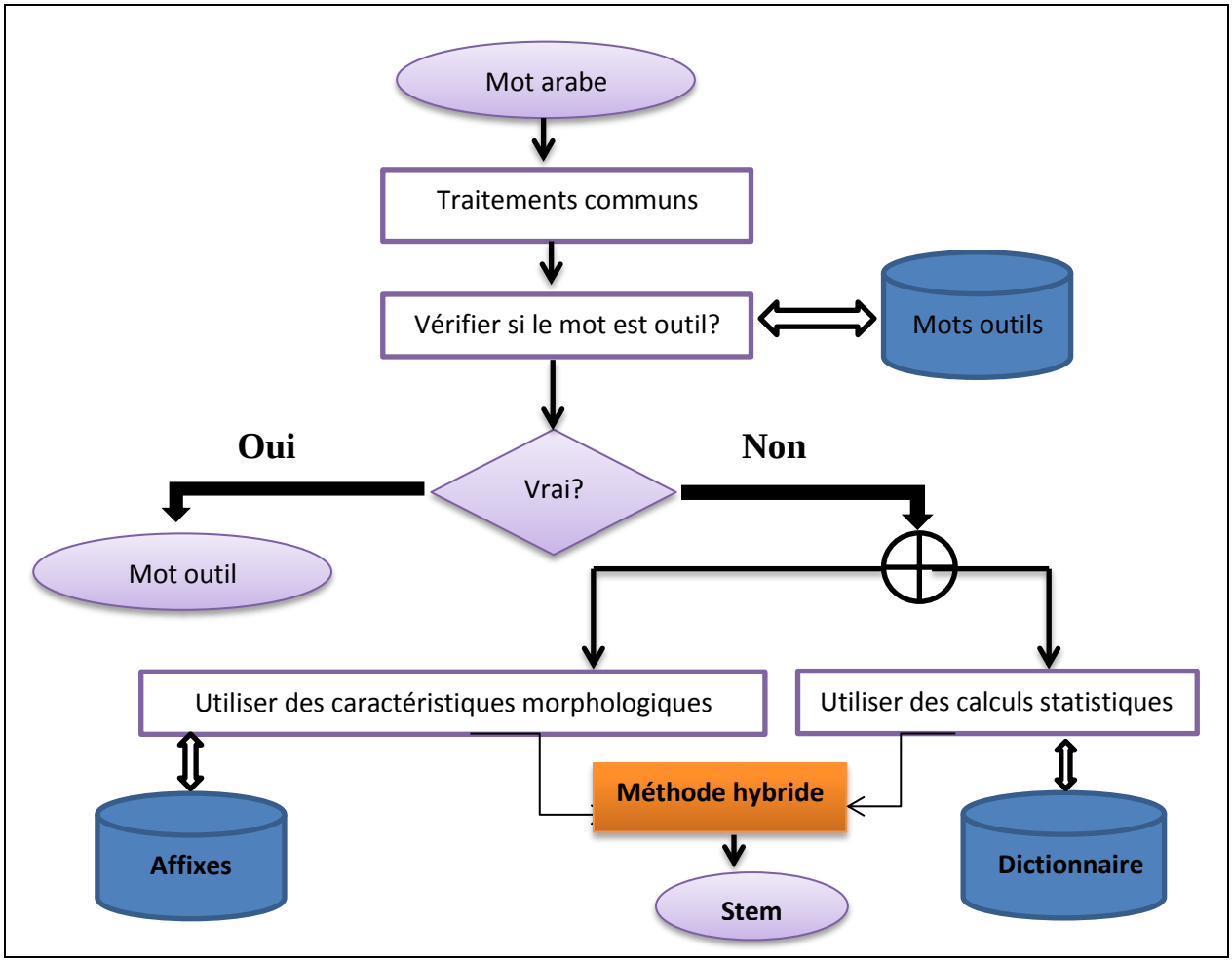


Figure 3.1: Processus d'extraction du stem d'un mot arabe.

1.2. Les étapes d'extraction

Le système d'extraction selon notre application est constitué des étapes décrites comme suivant:

1.2.1. Normalisation

La phase de normalisation est une phase commune à la plupart des méthodes d'extraction. Elle consiste à supprimer les signes diacritiques. De plus, certaines lettres sont remplacées par d'autres selon des règles prédéfinies comme exposé précédemment (par exemple : <أ, a> __ <ا, a>).

1.2.2. Analyse morphologique

L'analyse morphologique d'un mot arabe consiste à identifier la morphologie selon laquelle il est constitué, à savoir: les affixes, le schème et la racine.

a- Elimination des affixes:

Nous avons choisi la méthode Lemmatiseur léger qui est principalement basée sur la normalisation et l'élimination des affixes pour les mots de longueur supérieure ou égale à trois lettres et supprimer les affixes à partir des ensembles déjà définis.

Dans notre système, on teste d'abord si le mot entré contient un préfixe ou non (par exemple:{أ- ال- و- ك- ف- س- ل- ب- }), si c'est le cas, il sera éliminé. puis on teste s'il contient un suffixe { ني - نا - ك - كما - كم - كن - ه - هما - هم - هن- ون } on va le supprimer aussi.

1.2.3. Analyse statistique

Nous avons utilisé la technique des n-grammes. Un n-grammes est une séquence de N caractères issus d'une chaine de caractère : bi-grammes pour n=2.

Le mot obtenu lors de la phase morphologique ainsi que le stem avec lequel il sera comparé sont divisés en paires (appelées bi-grammes) puis la similitude entre le mot et le stem est calculée en utilisant l'équation (1). Ce processus est répété pour chaque stem dans la liste des stems: $S = \frac{\overline{A}+\overline{B}}{A+B}$ (1)

Où :

$\overline{A}$: Nombre de bi-grammes présents dans le mot traité (A) et ne sont pas présents dans le stem (B).

$\overline{B}$: Nombre de bi-grammes présents dans le stem (B) et ne sont pas présents dans le mot (A).

A: Nombre de bi-grammes uniques dans le mot (A).

B: Nombre de bi-grammes uniques dans le stem (B).

Nous utilisons l'équation (1) pour trouver le stem du mot correspondant , nous devons avoir le mot (A) et les stems potentiels (B) à comparer avec, puis la mesure de similarité est effectuée en calculant la valeur de S entre le mot (A) et chacun des stems potentiels (B).

Pour extraire le stem du mot par cette méthode, le mot (A) et le stem candidat(B) doivent être divisé en paires de séquence de lettres et ensuite, seulement les paires uniques seront prises en compte pour calculer la similarité (S). Le stem qui a la valeur minimale de (S) parmi la liste des stems est considéré comme le stem du mot.

Exemple :

Si nous avons le mot " الاستمرارية" .

- La première étape détermine que le mot " الاستمرارية " contient des préfixes et des suffixes donc, ils seront supprimés. Le nouveau mot obtenu est : استمرارية.
- On découpe alors les mots en sous-chaînes de caractères (chaque sous-chaîne contient 2 lettres), comme le montre la table 3.1.

استمرارية	**اس**	**ست**	**تم**	**مر**	**را**	**ار**	**ري**	**ية**
استمرار	اس	ست	تم	مر	را	ار	///	///
مستمر	مس	ست	تم	مر	///	///	///	///

Tableau 3.1.: Exemple de découpage des mots en 2-gram

Nombre de bi-grammes dans le mot "استمرارية": A=8.

Le stem candidat (B) : استمرار

Nombre de bi-grammes dans le stem "استمرار": B=6.

Nombre de bi-grammes présents dans le mot (A) et ne sont pas présents dans le stem(B), $\overline{A}$=2.

Nombre de bi-grammes présents dans le stem(B), et ne sont pas présents dans le mot(A), $\overline{B}$=0.

Ensuite, en utilisant l'équation (1), on peut calculer la similarité (S) entre le mot et le stem: $S = \frac{\overline{A}+\overline{B}}{A+B} \rightarrow S = \frac{2+0}{6+8} = \frac{2}{14} = \frac{1}{7}$

Le stem candidat (B) : مستمر

Nombre de bi-grammes dans le stem "مستمر ": B=4.

Nombre de bi-grammes présents dans le mot (A) et ne sont pas présents dans le stem(B), $\overline{A}$=5.

Nombre de bi-grammes présents dans le stem(B), et ne sont pas présents dans le mot(A), $\overline{B}$ =1.

Ensuite, en utilisant l'équation (1), on peut calculer la similarité (S) entre le mot et le stem:

$$S = \frac{\overline{A}+\overline{B}}{A+B} \rightarrow S = \frac{1+5}{4+8} = \frac{6}{12} = \frac{1}{2}$$

Le stem de mot "الاستمرارية" est donc " استمرار".

1.3. Les tables utilisées

Nous avons utilisé des tables pour les différents composants des mots pour réaliser notre système:

a. Table des stems:

Cette table sert à contenir l'ensemble des stems choisie qui seront utilisées pour vérifier si le stem obtenu après extraction est un stem valable ou non. Le nombre de stems initialement pris dans notre cas est de 1000, auquel on peut ajouter d'autres en cas de besoin.

1	أشار
2	اتبع
3	احتوى
4	اختلف
5	اساء
6	استحضر
7	استغل
8	استغنى
9	استقل
10	استمر

Tableau 3.2.: Table des stems.

b. Table des mots outils (stopwords):

Cette table contient les prépositions, les particules qui ont un effet de rection sur les verbes à l'accompli et à l'inaccompli, les particules des coordinations et d'autres particules; en résumé, tous les mots qui restent invariants quelque soit leur contexte.

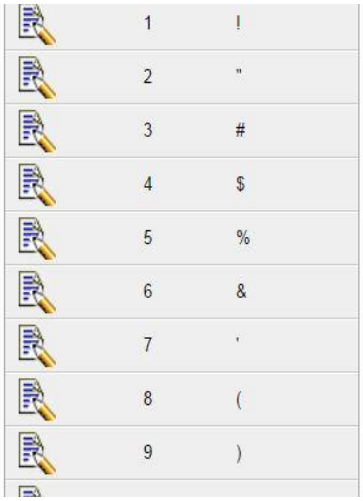
98	على
99	عليه
100	عليها
101	عن
102	عند
103	عندما
104	عنه
105	عنهم

ligne(s) 91 - 105 de 172

Tableau 3.3.: Table des mots outils.

c. Table des obstacles :

Cette table contient liste des ponctuations de la langue arabe.



1	!
2	"
3	#
4	$
5	%
6	&
7	'
8	(
9	)

Tableau 3.4. : Table des obstacles.

2. Les outils utilisés pour la programmation

Pour implémenter notre système nous avons a nécessité de l'utilisation d'un certain nombre d'outils :

2.1. Le langage de programmation (JAVA)

Java est un langage orienté objet permettant le développement d'applications complètes, créé par James Golsling et Patrick Naughton, employés de Sun Microsystems, avec le soutien de Bill Joy (cofondateur de Sun Microsystems en 1982), présenté officiellement le 23 mai 1995 au SunWorld. Ce langage de programmation est gratuit: on peut l'installer sur des serveurs, des postes de travail ou dans un Applet pour un navigateur. Il est également portable : il fonctionne sous tous les systèmes d'exploitation (Windows, Linux,…).

2.2. L'environnement de programmation (NetBeans IDE)

NetBeans est un environnement de développement intégré (EDI), placé en open source par Sun en juin 2000 sous licence CDDL (Common Development and Distribution Licence). En plus de java, NetBeans permet également de supporter

différents autres langages, comme C, C++, HTML et PHP. NetBeans comprend toutes les caractéristiques d'un IDE moderne (éditeur graphique d'interface et de pages web, refactoring, éditeur en couleur).

2.3. Le serveur de Base de données

Les tables utilisées dans notre système sont conçues à l'aide du serveur de base de données Oracle qui permet de mettre en place facilement un serveur Web. Il offre une bonne souplesse d'utilisation. Ainsi, il est à la portée d'un grand nombre de personnes puisqu'il ne requiert pas de connaissances particulières.

3. Description du système obtenu

Nous présentons dans cette section les différentes étapes de déroulement de notre application.

3.1. L'interface principale

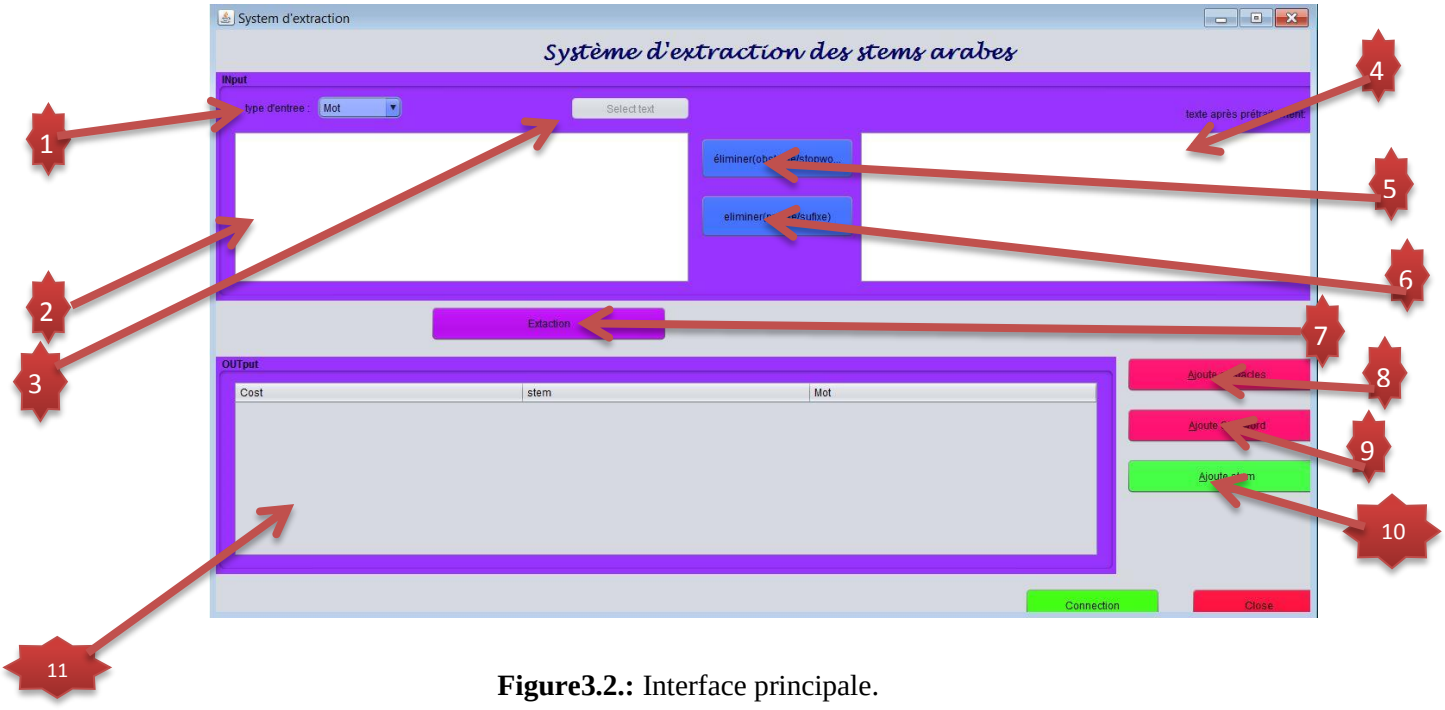


Figure3.2.: Interface principale.

1: bouton pour choisir le type d'entrée: mot ou bien texte.

2:zone dans laquelle est saisi le type d'entrée.

3: bouton pour charger un texte.

4: zone pour afficher le texte résultant après le prétraitement.

5:bouton servant à éliminer les obstacles et les stopwords.

6:bouton servant à supprimer les préfixes et les suffixes.

7:bouton pour lancer l'opération d'extraction des stems.

8:bouton pour ajouter des obstacles à la base de données.

9: bouton pour ajouter des stopwords à la base de données.

10: bouton pour ajouter des stems à la base de données.

11:zone résultat pour afficher les mots, leur stems et la valeur de similarité entre eux.

3.2. Prétraitements sur un texte

- ✓ obtenir un texte à partir d'un fichier sélectionné

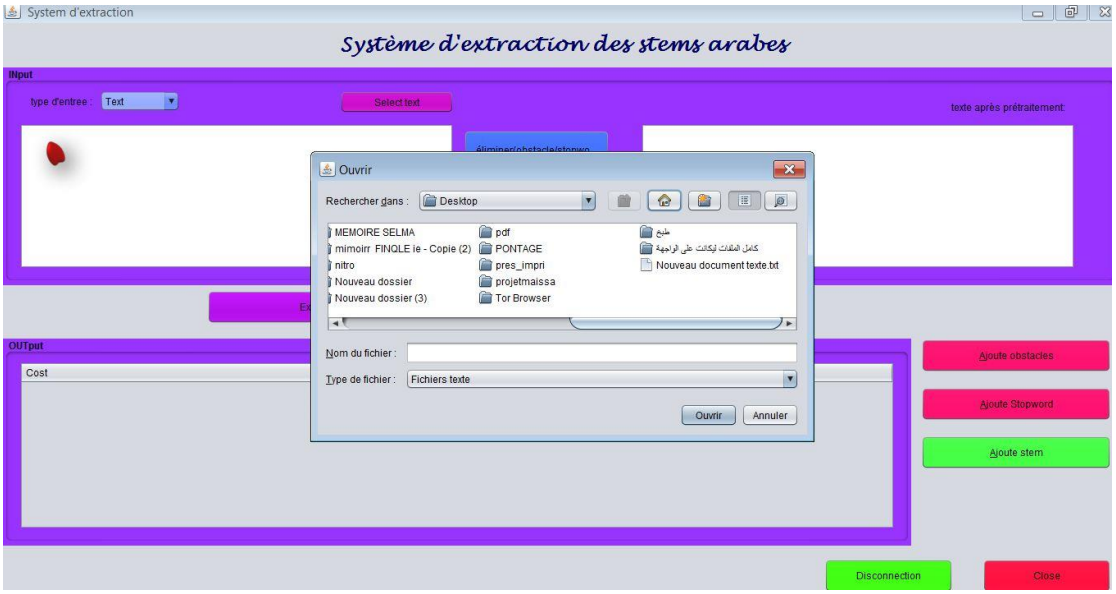


Figure 3.3.: Chargement de texte.

✓ Eliminer les obstacles et les mots outils (stopwords)

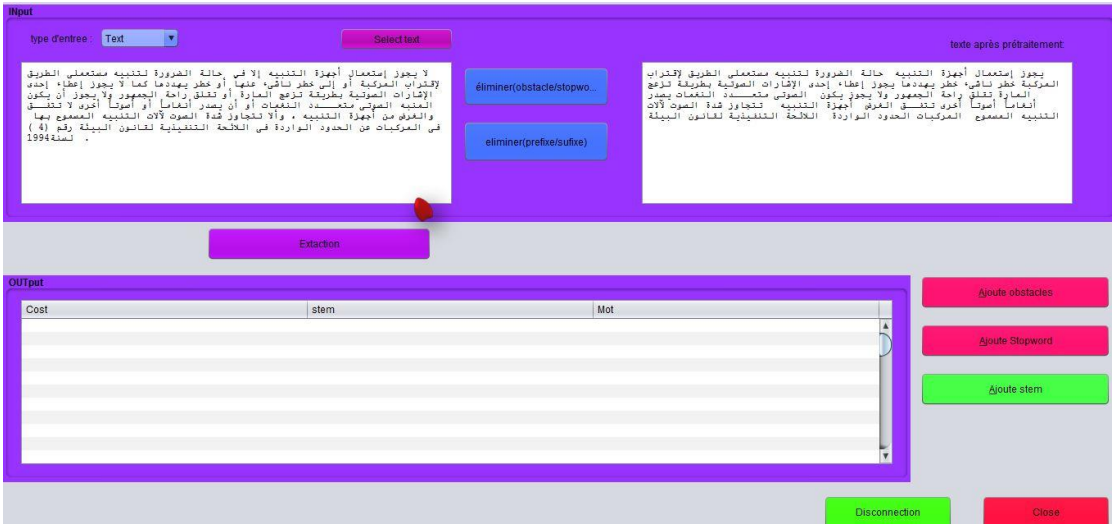


Figure 3.4.: Elimination des obstacles et les mots outils.

3.3. Traitement d'un texte

✓ Extraction des stems

Figure 3.5.: Extraction des stems.

4. évaluation des résultats:

Nous avons pris une liste simple de mots et fait l'extraction des stems selon notre méthode hybride, puis on a comparé les résultats obtenus avec ceux de la méthode n-grammes. Une part des résultats est montrée dans le tableau (3.5) qui suit :

Mot	Notre méthode hybride	Méthode n-grammes
تحاوليننني	حاول	تناول
بتأثيرها	تأثر	برهن
المصالحة	صالح	اصطلح
الأجهزة	جهز	انهزم
تذكاري	ذكر	تبارى
أستذكرونه	ذكر	استغنى
فذكرناهم	ذكر	فاهم
البرمجيات	برمج	استمرار
أعلمكم	علم	أعطى
يطعمون	اطعم	عمود

Tableau 3.5: extraction des stems avec la méthode hybride et n-grammes.

Et pour voir l'amélioration obtenue par la méthode hybride, en a utilisé un corpus constitué de 100 mots de test et les résultats montré dans la table (3.6) :

Méthode	Nombre de mots	Nombre des stems corrects	Taux d'extraction	Taux d'erreur
N-grammes	100	66	66%	34%
Hybride	100	70	70%	30%

Tableau 3.6.: Les résultats obtenu.

Comme nous le remarquons dans le tableau (3.6) la méthode hybride qui, supprime les affixes avant l'extraction des stems paraît plus satisfaisante par rapport à la méthode n-grammes.

5. Conclusion

Nous avons présenté dans ce chapitre quelques résultats pratiques sur l'extraction des stems en utilisant notre méthode hybride choisie, basée sur une analyse morphologique et des calculs statistiques. Et d'après ces résultats, nous pouvons dire que nous avons réussi à bien maitriser et implémenter la méthode choisie.

Conclusion générale

La langue arabe est une langue flexionnelle qui possède un système dérivationnel très riche. Nous avons abordé certaines de ses particularités et certaines de ses caractéristiques.

Nous avons utilisé dans ce mémoire, un algorithme hybride (morphologique & statistique) qui permet de trouver les stems des mots arabes; en utilisant une analyse morphologique basée sur la suppression des affixes, puis en appliquant la technique n-grammes de caractères pour analyser les mots arabes et éviter toute complexité causée par la richesse morphologique de cette langue.

Le présent travail rencontre plusieurs difficultés telle que : le manque de références et des travaux concernant les méthodes hybrides pour l'analyse des mots arabes, le manque du temps, ainsi que la difficulté liées au traitement automatique de la langue Arabe elle-même. Mais, nous pensons qu'on a quand même pu relever le défi et, par la même occasion, apprendre beaucoup de nouvelles connaissances tout au long de la réalisation de ce travail.

Notre étude montre que pour les méthodes d'analyse hybrides basées sur l'extraction des stems sont les plus appropriés par rapport aux autres méthodes d'extraction.

D'un point de vue général, le traitement automatique de la langue arabe et en particulier, l'extraction des stems, reste un domaine très ouvert et présente des marges de progression importantes, du fait de la richesse morphologique de cette langue qui, comme nous l'avons montré, reste un des problèmes majeurs de l'arabe, où de grandes améliorations peuvent encore être apportées.

[1] A. Chen, et F. Gey, "Building an Arabic stemmer for information retrieval". In Proceedings of TREC, Gaithersburg, pp. 631-639, 2002.

[2] A.Goweder, M. Poesio, A. De Roeck, et J. Reynolds," Identifying Broken Plurals in Unvowelised Arabic Text". Proceeding of EMNLP, pp. 246-253, 2003.

[3] Abd El Salam AL HAJJAR, "Extraction et gestion de l'information à partir des documents arabes", mémoire de Doctorat d'informatique, Université paris – saint dénis, Décembre 2010.

[4] Abd El Salam Al Hajjar, Mohamed Hajjar, et Khaldoun Zreik, "Classification of Arabic Information Extraction methods", 2nd International Conference on Arabic Language Resources and Tools Cairo (Egypt), pp. 22 – 23, 2009.

[5] Abd El Salam al Hajjar, Mohamed Hajjar, et Khaldoun Zreik, " A new system for evaluation of Arabic root extraction methods", The Fifth International Conference on Internet and Web Applications and Services. ICIW, Barcelona, 2010.

[6] Ahmed Al-hamlawi," شذا العرف في فن الصرف", livre publié en 17 décembre 2005.

[7] Aïda KHEMAKHEM, "ArabicLDB : une base lexicale normalisée pour la langue arabe", mémoire de MASTER en Systèmes d'Information et Nouvelles Technologies, Université de Sfax, Tunisie, 2006.

[8] A.N.De Roeck, et W. Al-Fares, " A morphologically sensitive clustering algorithm for identifying Arabic roots". Proceedings ACL-2000, Hong Kong, pp. 199- 206, 2000.

[9] Atef Ben Youssef**,** "Méthodes Mixtes pour la Traduction Automatique Statistique"**,** Mémoire de Master**,** Université STENDHAL Grenoble3, juillet 2008.

[10] baloul Sofiane, "Développement d'un système automatique de synthèse de la parole à partir de texte arabe standard voyelle", mémoire de doctorat d'informatique, Université du MarieFrance, 2003.

[11] BENZATER Nebia**,** " analyse morphologique de texte arabe pour son indexation sémantique", Mémoire de Magistère en Informatique, Université des Sciences et de la Technologie - Mohamed Boudiaf – Oran, 2014-2015.

[12] Boulaknadel Siham, "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", mémoire de doctorat d'informatique, Université de Nantes, octobre 2008.

[13] Char France, "Traitement Automatique de la langue arabe, la modération en arabe", mémoire de master 2 d'ingénierie linguistique En Traductique et gestion de

l’information, Institut National des Langues et Civilisations Orientales (I.N.A.L.C.O)-paris, 4 juillet 2012.

[14] DILEKH Tahar," Implémentation d’un outil d’indexation et de recherche des textes en arabe", mémoire de Magister en Système d’Information et de Connaissance (SIC), Université Hadj Lakhdar – Batna, 2011.

[15] F.A. Hawas, "Exploit relations between the word letters and their placement in the word for Arabic root extraction". Comput. Sci., pp: 327-431. 2013

[16] faraj ahmed et andreas Nürnberger,"n-grams conflation approach for Arabic text" proceedings of the international workshop on improving non English web searching (Inews07) in conjunction with the 30th annual international (ACM SIGIR) conference. amesterdam city, netherlands, pp 39-46,2007

[17] Fouad Soufiane Douzidia, "Résumé automatique de texte arabe", Mémoire présenté à la Faculté des études supérieures en vue de l’obtention du grade de M.Sc en informatique Septembre, 2004.

[18] G. Grefenstette, N. Semmar, F. et C. Fluhr, " Modifying a Natural Language Processing System for European Languages to Treat Arabic in Information Processing and Information Retrieval Applications". ACL computational approaches to semitic, languages workshop, USA, June 29, 2005.

[19] G. Kanaan, R. Al-Shalabi, J. Jaarn, M. Al-Kabi, et A. Hasnah, " A New Stemming Algorithm to Extract Quadri-Literal Arabic Roots", 2004.

[20] GHOUL Dhaou, "Outils génériques pour l'étiquetage morphosyntaxique de la langue arabe : segmentation et corpus d'entraînement", Mémoire de master, Université STENDHAL Grenoble3, 2010-2011.

[21]H. Al Ameed, S. Al Ketbi, A. Al Kaabi, K. Al Shebli, N. Al Shamsi,

, N. Nuaimi, Al et S. Al Muhairi, " Arabic Light Stemmer: A new Enhanced Approach", The Second International Conference on Innovations in Information Technology (IIT'05), 2005.

[22]J. Yaghi, et S. Yagi, "Systematic Verb Stem Generation For Arabic". Workshop on Computational Approaches To Arabic Script-Based Languages. 2004.

[23] K., Darwish," Building a Shallow Arabic Morphological Analyzer in One Day". The ACL-02 Workshop on Computational Approaches to Semitic Languages, Philadelphia, USA, 2002.

[24] K. R Beesley," Finite-State Morphological Analysis and Generation of Arabic at Xerox Research: Status and Plans in 2001". In The ACL 2001Workshop on Arabic Language Processing: Status and Prospects, Toulouse, France, 2001.

[25]K Taghva, R. Elkoury, et J.Coombs, "Arabic Stemming without a root dictionary". International Conference on Information Technology: Coding and Computing (ITCC'05), pp. 152-157, 2005.

[26] L. Khreisat, " Arabic Text Classification Using N-gram Frequency Statistics a Comparative Study". The 2006 International conference on Data Mining Part of the 2006 World Congress in Computer Sciences DMIN, pp. 78-82, 2006.

[27] L. Larkey, L. Ballesteros, et M. E. Connel, " Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis", Proc. of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 275 - 282, 2002.

[28] L. Larkey, L. Ballesteros, et M.Connell, "Light Stemming for Arabic IR, Arabic Computational Morphology: Knowledge-based and Empirical Methods", A. Soudi, A. Van Bosch, and G. Neumann Editors. Kluwer/Springer's series on Text, Speech, and Language Technology, 2005.

[29] M. Aljlayl, et O. Frieder, "On arabic search: Improving the retrieval effectiveness via a light stemming approach". In Proceedings ofACM Eleventh Conference on Information and Knowledge Management, Mclean, 2002.

[30]M. Mustafa, H. AbdAlla, et H. Suleman, "Current Approaches in Arabic IR: A Survey". In Proceedings The Annual International Conference on Asia-Pacific Digital Libraries (ICADL), Bali, Indonesia. 2008.

[31] M. Sanan," Etude Des Méthodes De La Recherche D'information Et De L'indexation Sur Les Documents Electroniques : Cas De La Langue Arabe", thèse de Doctorat, UNIVERSITE PARIS VIII - SAINT DENIS, 2008.

[32] M. Sanan, M. Rammal, et K.Zreik, " Arabic documents classification using N-gram", Conférence ICHSL6, Toulouse, 2008.

[33] Mostapha Al-Glayini, "جامع الدروس العربية", livre édité en 2007 en Bierut, Lebanon.

[34] N. Semmar, F. Elkateb-Gara, et C. Fluhr, "Using a Stemmer in a Natural Language Processing system to treat Arabic for Cross-language Information Retrieval". International conference on Machine Intelligence, Tozeur, TUNISIE. 2005.

[35] N. Thabet, "Stemming the Qur'an". WORKSHOP ON Computational Approaches to Arabic Script-based Languages, University of Geneva, Switzerland. 2004.

[36] N. Yousef, A. Abu-Errub,A Odeh, et H. Khafajeh, "An improved Arabic word's roots extraction method using n-gram technique", Journal of computer Science 10 (4): 716-719,2014.

[37] S. Khoja, et R.Garside, "Stemming Arabic text". Computing Department, Lancaster University, Lancaster, www.comp.lancs.ac.uk/computing/users/khoja/stemmer.ps, 1999.

[38] Slim MESFAR, "Analyse morpho-syntaxique automatique et reconnaissance des entités nommées en arabe standard", mémoire de doctorat d'informatique, Université de Franche Comté, novembre 2008.

[39] Y. Kadri,et J. Nie, "Effective Stemming for Arabic Information Retrieval". Proceedings of the Challenge of Arabic for NLP/ MT Conference, Londres, Royaume-Uni, 2006.

[40]Wikipedia: www.wikipedia.org/wiki/مقدمة/بوابة : اللغة العربية

Consulté le : 20/02/2017.

[41] Wikipedia: traitement automatique des langues : http://en.wikipedia.org/ wiki/ traitement automatique des langues. consulté le: 10.03. 2017.

CHAPITRE 2

LE TRAITEMENT AUTOMATIQUE DE LA LANGUE ARABE

CHAPITRE 1

PRESENTATION DE LA LANGUE ARABE

CHAPITRE 3

DESCRIPTION DU SYSTEME REALISE

ملخص:

نقدم في هذا المشروع برنامجا يتيح تحليل كلمات اللغة العربية التي تعتبر لغة معقدة بنيويا وفهم كلماتها يتطلب الوصول إلى جذع أو جذر الكلمة. والوصول إلى الجذع أو الجذر ليس بالأمر الهين و يظل نوعا من التحدي بخصوص تطوير تطبيقات المعالجة الآلية للغات الطبيعية.

و لإنجاز هذا العمل قمنا باستعمال طريقة هجينة تتكون من تقنيتين مختلفتين وهما : تقنية التحليل الصرفي مع تقنية احصائية (n-grammes).وقد جاء هذا العمل ليقدم مساهمة إضافية في إثبات فعالية الطريقة الهجينة المقترحة مقارنة مع الطرق الأخرى.

الكلمات المفتاحية: جذع، جذر، المعالجة الآلية للغات الطبيعية، طريقة هجينة ،التحليل الصرفي، n-grammes

Abstract:

In this project we introduce a program that allows the analysis of words in Arabic because the Arabic language is structurally complex and understanding its words requires access to the stem or the root of the word. Access to the stem or root is not easy and remains a challenge for the development of automatic natural language processing (ANLP) applications.

In order to accomplish this work, we used a hybrid method consisting of two different techniques: morphological analysis and n-grammes. This work was an additional contribution for demonstrating the effectiveness of the proposed hybrid approach compared to other methods.

Key words: Stem, Root, ANLP, hybrid method, morphological analysis, n-grammes.,

Résumé:

Nous présentons dans ce projet la réalisation d'un programme qui permet d'analyser les mots arabes car la complexité morphologique da la langue arabe la rend particulièrement difficile, et la compréhension des mots nécessite l'accès au stem ou la racine du mot. Et l'accès au stem ou à la racine n'est pas facile et reste un défi en ce qui concerne le développement des applications de traitement automatique de la langue naturelle(TALN).

Pour accomplir ce travail, nous avons utilisé une méthode hybride qui se compose de deux différentes techniques, à savoir: La méthode d'analyse morphologique et la méthode n-grammes. Ce travail est venu pour apporter une contribution supplémentaire pour prouver l'efficacité de la méthode hybride proposée par rapport aux autres méthodes.

Mots clés:

Stem, Racine, TALN, méthode hybride, analyse morphologique, n-grammes.