

Table des matieres

Introduction générale

CHAPITRE 1

Notions sur la langue arabe

1	Particularités de la langue arabe	2
2	L'Ambigüité des mots	4
3	Composition du lexique arabe : les formes de bases	5
3.1	Les verbes	5
3.2	Les noms	7
3.3	Les nombres.....	8
3.4	Mots isolés.....	9
3.5	La particule	10
4	Morphologie du mot arabe.....	11
4.1	Structure du mot arabe.....	11
4.2	Eléments essentiels de la morphologie du mot arabe	12
5	Conclusion	16

CHAPITRE 2

Analyse morphologique de la langue arabe

1	Analyse morphologique	17
1.1	Segmentation	17
1.2	Prétraitement morphologique	18
1.3	Stemming:.....	18
1.4	Analyse affixale.....	19
1.5	Analyse morphologique.....	20

2	Les méthodes d'extraction de la racine d'un mot arabe	21
2.1	Les méthodes basées sur l'analyse morphologique.....	21
2.2	Les méthodes basées sur la génération systématique	23
2.3	Les méthodes basées sur la transcription ou la traduction.....	24
2.4	Les méthodes basées sur l'analyse statistique	24
2.5	Les méthodes basées sur la position des lettres.....	25
2.6	Les méthodes basées sur l'analyse statistique et morphologique (Hybride).....	26
3	Conclusion	26

CHAPITRE 3

Conception et Réalisation

1	Description du système réalisé:	27
2	Présentation de l'Application.....	31
3	Les outils utilisés pour l'implémentation.....	42
3.1	L'environnement de programmation:	42
3.2	Langage de programmation	44
3.3	Test et Evaluation	44
4	Conclusion	45

Conclusion générale

Bibliographie

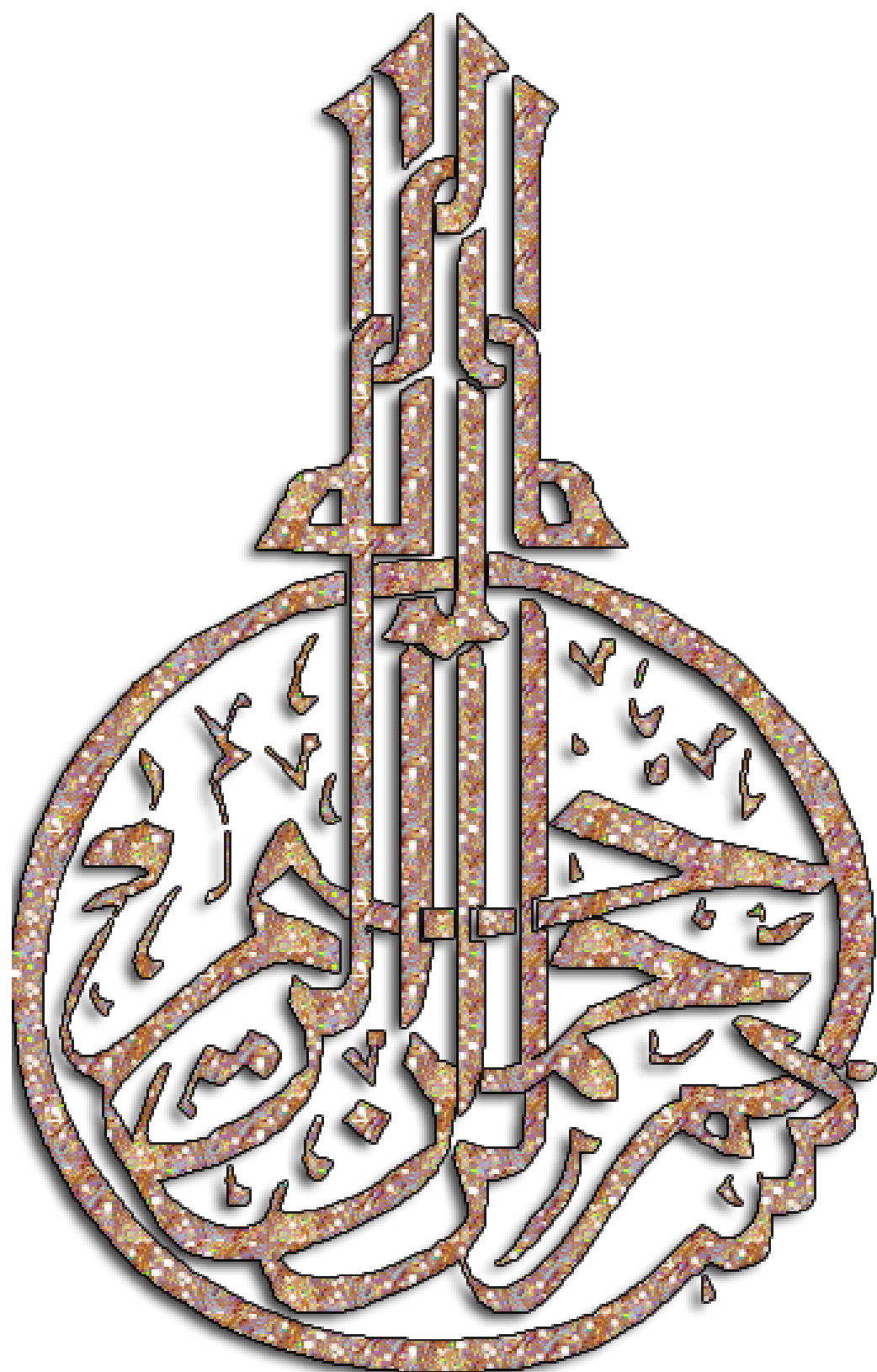
Liste des figures

Figure 2.1 : Exemple d'un Segmenteur	18
Figure 2.2 : Exemple de light Stemming	19
Figure 2.3 : Exemple de heavy Stemming.	19
Figure 2.4 : Analyse morphologie d'un mot.....	20
Figure 2.5 : Mesure de similarité (bigramme) entre deux mots.....	25
Figure 3.1 : Plan de travaille	27
Figure 3.2 : Extraction des données	28
Figure 3.3 : Extraction des Racines.....	29
Figure 3.4 : Extraction des Modelés	29
Figure 3.5 : Extraction des Racines.....	30
Figure 3.6 : Exemple d'extraction des racines	31
Figure 3.7 : Tableau de la base de données.....	31
Figure 3.8 : Tableau Racine	32
Figure 3.9 : Tableau Modèles.....	32
Figure 3.10 : Tableau Mots Vides.....	32
Figure 3.11 : Tableau des Préfixes.....	33
Figure 3.12 : Tableau Suffixe.....	33
Figure 3.13 : Plan du mode « Visiteur »	34
Figure 3.14 : Plan du mode « Utilisateur ».	34
Figure 3.15 : Plan de Mode « Administrateur ».....	35
Figure 3.16 : Interface de notre application	35
Figure 3.17 : Interface d'Accueil	36
Figure 3.18 : Interface de management.....	37
Figure 3.19 : Interface de notre BDD « Ma BDD ».....	37
Figure 3.20 : Interface de Gestion sur nos BDD « Gestion de BDD »1	38
Figure 3.21 : Interface de notre Application « Gestion de BDD »2	39
Figure 3.22 : Interface de notre Application « Suffixe »	39
Figure 3.23 : Interface de notre Application « Autre ».....	40
Figure 3.24 : Interface de notre Application « Ajouter Utilisateur »1	40

Figure 3.25 : Interface de notre Application «Ajouter Utilisateur» 2	41
Figure 3.26 : Interface de notre Application «A Propos»	41
Figure 3.27 : NetBeans.....	42
Figure 3.28 : Scene Builder.....	43
Figure 3.29 : Bases de données Derby	43

Liste des tableaux

Table 1.1 : Tableau de translittération de l'alphabet arabe	3
Table 1.2 : Exemple de variation de la lettre ع ayn et غ gayn.....	4
Table 1.3 : Les voyelles brèves	4
Table 1.4 : Ambiguïté causée par l'absence de voyelles pour les unités lexicales مدرسة et كتب	5
Table 1.5 : Les diacritiques doubles.....	5
Table 1.6 : Principaux schèmes verbaux augmentés.....	7
Table 1.7 : Dérivationnel irrégulier.....	10
Table 1.8 : Dérivationnel régulier	10
Table 1.9 : Structure du mot arabe	11
Table 1.10 : La segmentation du mot "أنتفكروننا"	12
Table 1.11 : quelques mots dérivant de la racine درس (a étudié)	13
Table 1.12 : Un exemple des préfixes	14
Table 1.13 : Un exemple des suffixes divisés selon leurs types.....	15
Table 1.14 : Un exemple de génération des Stems.....	16
Table 3.1 : Exemple de construction des mots à partir d'un schème.....	30



Introduction générale

La morphologie est un domaine de la langue qui permet la description des règles régissant la structure interne des mots (unités lexicales). Chez les grammairiens, la morphologie est l'étude des formes des mots (flexion et dérivation) ; en d'autres termes, la morphologie est l'étude des mots considérés isolément (hors contexte) sous le double aspect de la nature et les variations qu'ils peuvent subir.

L'étude des mots ou le traitement automatique du langage est très difficile à maîtriser étant donné que le langage humain est généralement très complexe et ne dispose d'aucunes règles fixes qui s'appliquent à la totalité des aspects de même nature. Le traitement de la langue arabe n'a vu le jour que quelques années plus tard après les premiers travaux. En effet, plusieurs tentatives ont été menées donnant des résultats plus ou moins satisfaisants. Ces tentatives ont clairement révélé l'aspect morphologique très complexe de cette langue, qui, à la différence des autres langues, possède une structure et des caractéristiques très spécifiques qui la rendent très particulière et augmentent ainsi considérablement l'effort nécessaire pour la maîtriser. Les techniques utilisées pour traiter la morphologie sont donc très diverses et variées, certaines sont inspirées des travaux menés sur les langues étrangères et d'autres reposent sur les propres caractéristiques de cette langue. Dans ce mémoire, nous allons exposer une contribution de notre part qui consiste à essayer d'appliquer l'une de ces méthodes et étudier la faisabilité d'un tel traitement sur une telle langue après bien sûr, l'exposition d'une bonne partie portant sur l'étude des différents aspects morphologiques de cette langue nécessaires pour la suite du travail. D'une façon plus claire, le mémoire sera organisé de la manière suivante :

Dans le premier chapitre intitulé " **Notions sur la langue arabe** ", nous allons donner un aperçu sur la langue arabe et exposer les différents aspects de sa morphologie ainsi que ses caractéristiques.

Puis, au sein du deuxième chapitre " **Analyse morphologique de la langue arabe** ", Nous allons présenter le processus général d'extraction de la racine d'un mot arabe et les différentes techniques utilisées pour ce faire.

Finalement, le troisième et dernier chapitre qu'on a nommé " **Conception et Réalisation** ", servira à exposer une conception de la méthode choisie, suivie de sa réalisation avec une présentation de quelques-unes de ses interfaces.

CHAPITRE 1

Notions sur la langue arabe

La langue arabe est une langue dérivationnelle et flexionnelle. A l'origine, la langue arabe est la langue parlée par les Arabes. En plus, elle est la langue sacrée du Coran et de l'Islam. Du fait de la propagation de l'Islam et la diffusion du Coran, cette langue est devenue une langue liturgique. Elle est parlée dans 22 pays alors que le nombre de ses locuteurs est plus de 280 millions.

Dans ce chapitre, nous commencerons par présenter les caractéristiques morphologiques de la langue arabe. Ensuite, nous décrirons le mécanisme de dérivation de ses mots. Puis, nous présenterons ces différentes catégories grammaticales. Aussi, nous préciserons les traits morphologiques des verbes et des noms. Enfin, nous donnerons un aperçu sur l'influence de la phonologie sur la morphologie de la langue arabe [4].

1 Particularités de la langue arabe

La langue arabe est une langue morphologiquement complexe qui possède des caractéristiques bien particulières par rapport aux autres langues dont :

❖ Alphabet arabe

La langue arabe est une langue sémitique qui s'écrit de droite à gauche et dont l'alphabet est un abjad [4]. L'écriture arabe comprend essentiellement les consonnes, elle comprend également les voyelles longues "ا", "ي" et "و". Les deux dernières sont des réalisations contextuelles des glides "ي" et "و".

L'écriture arabe comporte également des voyelles qui ne sont pas essentielles à l'écriture ainsi qu'un certain nombre de signes annexes dont l'emploi est facultatif hormis pour le coran servant à noter les trois voyelles brèves ("َ" (a), "ُ" (u) et "ِ" (i)).

Il existe, de plus, une série d'autres diacritiques de syllabation dont les plus courantes sont l'indication de l'absence de voyelle "ْ" (sukūn) et la gémiation des consonnes "ّ" (šadda). Notons aussi que si un mot arabe est indéfini (sans article ni complément du nom), il prend (sauf exceptions) les désinences "َ" (an), "ُ" (un) ou "ِ" (in), nommées nounation ou tanwīn.

Celles-ci sont notées par des diacritiques spéciales marquées par le redoublement du signe de la voyelle qui précède le suffixe « n » attendu en fin de mot.

En outre, il nous importe de signaler que les notions de lettre capitale et lettre minuscule n'existent pas, l'écriture arabe est dite monocamérale. En addition, l'arabe est une langue

semi-cursive ; la plupart des lettres s'attachent entre elles, leurs graphies diffèrent selon qu'elles soient précédées et/ou suivies d'autres lettres ou qu'elles soient isolées. Seulement six d'entre-elles ne s'attachent jamais à la lettre suivante, **elles sont notées par un astérisque dans le tableau suivant**

Lettre	Nom	Translittération	Commentaires	ك	kâf	k	Lettre lunaire
ء	hamzä	'		ل	Lâm	l	Lettre solaire
ا	âlif	ā	Voyelle longue	م	mîm	m	Lettre lunaire
ب	Bâ'	b	Lettre lunaire	ن	nûn	n	Lettre solaire
ت	Tâ'	t	Lettre solaire	و	wâw	w / ū	Lettre lunaire (*) / Voyelle longue
ث	Tâ'	t	Lettre solaire	ي	Yâ'	y / ī	Lettre lunaire / Voyelle longue
ج	Jîm	ğ	Lettre lunaire	ـَ	fathä	a	Voyelle brève
ح	hâ'	h	Lettre lunaire	ـُ	dammä	u	Voyelle brève
خ	Xâ'	h	Lettre lunaire	ـِ	kasrâ	i	Voyelle brève
د	dâl	d	Lettre solaire (*)	ـِـ	tanwîn	ã / an	Voyelle brève / Tanwîn
ذ	ḏâl	d	Lettre solaire (*)	ـِـ	tanwîn	ũ / un	Voyelle brève/ Tanwîn
ر	Râ'	r	Lettre solaire (*)	ـِـ	tanwîn	ĩ / in	Voyelle brève/ Tanwîn
ز	zây	z	Lettre solaire (*)	ـِـ	tâ'marbûtä	ä (at en annexion)	
س	Sîn	s	Lettre solaire	ى	Alifmaqsûrâ	ą	
ش	Sîn	š	Lettre solaire	آ	Alifmamdûdä	â	
ص	sâd	s	Lettre solaire	أ	hamzä	á	
ض	dâd	d	Lettre solaire	إ	hamzä	ù	
ط	tâ'	t	Lettre solaire	إ	hamzä	í	
ظ	zâ'	z	Lettre solaire	ؤ	hamzä	w	
ع	'ayn	'	Lettre lunaire	ئ	hamzä	ý	
غ	ğayn	ğ	Lettre lunaire	ـِـ	sukun		non transcrit
ف	Fâ'	f	Lettre lunaire	ـِـ	šaddä	lettre redoublée	Signe de gémation
ق	qâf	q	Lettre lunaire				

Table 1.1 : Tableau de translittération de l'alphabet arabe [16]

A la fin d'une lettre non joignable	A la fin	Au milieu	Au début
ع	ع	ع	ع
غ	غ	غ	غ

Table 1.2 : Exemple de variation de la lettre ع ayn et غ gayn

a) Voyellation

Deux types de voyelles existent en langue arabe, les voyelles longues et les voyelles courtes.

Les voyelles longues : ce sont les trois lettres ا،و،ي.

Exemples :

- (ا) : حال (état)
- (و) : فول (fèves)
- (ي) : فيل (éléphant)

b) Les voyelles brèves

Les voyelles brèves (diacritiques) : ce sont des signes qu'on ajoute au-dessus ou au-dessous des consonnes d'un mot pour spécifier sa prononciation ou pour en spécifier le sens.

Voyelle brève	Nom	Transcription
ـَ	فَتْحَة /fathatun/	a
ـِ	كَسْرَة /kasratun/	i
ـُ	ضَمَّة /dammatun/	u
ـْ	سُكُون /sukûnun/	-

Table 1.3 : Les voyelles brèves

2 L'Ambigüité des mots

Pour mieux comprendre, prenons l'exemple de « كُتِبَ » du tableau 1.4. Le dictionnaire nous renvoie les voyellations lexicales suivantes:

- كَتَبَ ,il a écrit
- كُتِبَ ,il a écrit
- كُتُبَ ,Livres

Unité lexicale	1 ^{ère} interprétation		2 ^{ème} interprétation		3 ^{ème} interprétation	
كتب	كُتِبَ	Il a écrit	كُتِبَ	Il a été écrit	كُتِبَ	Livres
مدرسة	مَدْرَسَة	Ecole	مُدْرَسَة	Enseignante	مُدْرَسَة	Enseignée

Table 1.4 : Ambiguïté causée par l'absence de voyelles pour les unités lexicales كتب et مدرسة [13]

La plupart des études dans le traitement automatique de la langue arabe ignorent les signes diacritiques et les suppriment durant une phase préalable qu'on appelle la normalisation qui consiste aussi à remplacer quelques lettres par d'autres selon des règles prédéfinies

➤ La Shadda :

C'est le signe " ّ " (šadda) de la gémation en arabe. il représente un doublement d'une consonne lors de sa prononciation et ne peut être utilisé dans la 1^{ère} lettre d'un mot

Exemple : درس « a enseigné », درّس « a étudié »

➤ Le tanwin:

C'est des signes qui sont utilisés à la fin des mots indéterminés consistant à un doublement des signes diacritiques et qui produisent le même son que les trois premières voyelles simples avec l'ajout du son « n » à la fin.

َ	[an]
ً	[on]
ِ	[in]

Table 1.5 : Les diacritiques doubles

3 Composition du lexique arabe : les formes de bases

La grammaire traditionnelle arabe ne connaît que trois sous-ensembles : verbes, noms et particules. [15] et [32].

3.1 Les verbes

Un verbe est une entité exprimant un sens dépendant du temps. La majorité des verbes arabes sont formés sur des radicaux de 3 consonnes ; tel est le cas du verbe « كتب » (kataba – écrire) et éventuellement 4 consonnes tel est le cas du verbe " " (دَحْرَجَ dahraġa – glisser, faire glisser). [16].

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:

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

Pour les pronoms personnels, le sujet est inclus dans le verbe conjugué. Il n'est donc pas nécessaire (comme c'est le cas en français) de précéder le verbe conjugué par son pronom. [31]

Plusieurs études linguistiques ont été menées sur le système verbal en arabe, voir [16]. Dans cette section, il nous importe d'introduire un classement des verbes selon leurs radicaux :

3.1.1 Les verbes à racine simple:

Les verbes à racine simple ont une base à trois consonnes appelées consonnes radicales. Ces verbes sont associés au schème verbal "فَعَلَ" (*fa'ala*). Lorsqu'aucune des consonnes radicales du verbe n'est une voyelle longue, il est dit *sain*. Ces radicaux peuvent être les causes de transformation ou défectuosité (عِلَّة), nous citons:

La présence d'un glide "أ" (*á – hamzä*), "ي" (*y – yâ'*) ou "و" (*w – wâw*) parmi les consonnes du radical. Selon la position de ce glide, nous distinguons différents types de verbes:

Si l'une des consonnes radicales est "أ" (*á–hamzä*), indépendamment des aposition

- Verbe hamzé «مَهْمُوز» (*mahmuwz*) ;
- la 1^{er} consonne radicale est un "و" (*w*) ou "ي" (*y*) → verbe assimilé «مِثَال» (*mithal*);
- La 2^{ème} consonne radicale est un "و" (*w*) ou "ي" (*y*) → verbe creux «أَجُوف» (*ağwaf*);
- la 3^{ème} consonne radicale est un "و" (*w*) ou "ي" (*y*) → verbe défectueux «نَاقِص» (*naāqis*);
- La présence de deux consonnes identiques en deuxième et troisième position du radical
→ verbe redoublé «مُضَاعَف» (*mudāā'af*).

Notons que, les causes de défectuosité peuvent être combinées pour avoir des transformations doubles ou triples au sein d'une même racine.

3.1.2 Les verbes à racine augmentée "الأفعال المزيدة" [30]

Aux schèmes simples, décrits ci-dessus, vient s'ajouter une dizaine de schèmes verbaux augmentés. Ces schèmes sont formés à partir de racines simples par un ensemble d'opérations morphologiques pour donner un sens particulier aux verbes résultats, nous citons :

Schème verbal augmenté	Opérations morphologiques associées
« فَعَّلَ » (<i>fa''ala</i>)	redoublement de la deuxième consonne radicale
« فَاعَلَ » (<i>faā'ala</i>)	allongement de la première consonne radicale par l'ajout d'un "ا" (ā – álif)
« أَفْعَلَ » (<i>áfa'ala</i>)	adjonction d'une "أ" (á – hamzā) au début de la racine
« تَفَعَّلَ » (<i>tafa''ala</i>)	adjonction d'une "ت" (t) au début de la racine + redoublement de la deuxième consonne radicale
« تَفَاعَلَ » (<i>tafaā'ala</i>)	adjonction d'une "ت" (t) au début de la racine + allongement de la première consonne radicale par l'ajout d'un "ا" (ā – álif)
« فَعَّلَلْ » (<i>ifta'ala</i>)	adjonction d'un "ا" (ā – álif) au début de la racine + insertion d'un morphème mono-consonantique "ت" (t) à la suite de la première consonne
« اِنْفَعَلَ » (<i>infa'ala</i>)	adjonction d'un morphème bi-consonantique "اِن" (in) au début de la racine
« اِسْتَفْعَلَ » (<i>ista'ala</i>)	adjonction d'un morphème tri-consonantique "اِسْت" (ista) au début de la racine

Table 1.6 : Principaux schèmes verbaux augmentés

3.2 Les noms

L'élément désignant un être, un objet ou un état qui exprime un sens indépendant du temps. La fonction du nom est sa relation avec un mot ou une expression de la phrase, elle change avec le changement de cette relation sans perdre son sens linguistique. Les substantifs arabes sont de deux catégories, ceux qui sont dérivés de la racine verbale et ceux qui ne les sont pas comme les noms propres et les noms communs.

Le système morphologique des noms arabes distingue des sous-catégories :

3.2.1 Les noms primitifs [30]

Ce sont des noms qui ne peuvent pas être rattachés à une racine verbale. Ils paraissent bien constituer le glossaire fondamental de la langue concrète.

Exemple :

- رَأْس (raás – tête)
- كُرْسِيّ (kursiyy – siège)
- كَبْش (kabš – bélier) etc ...

Dans cette catégorie, nous incluons aussi les noms à racine bilitère tels que :

- دم (dam - sang)
- فم (fam - bouche)
- أب (áab – père)
- أخ (áh–frère)etc...

3.2.2 Les noms dérivés ou déverbaux [19]

Ce sont les noms qui peuvent être dérivés à partir d'une racine verbale. Le nombre et la nature de ces formes varient selon le statut du verbe auquel ils se rattachent. Puisque ce sont des noms, alors ils peuvent recevoir les marques du genre, du nombre, du cas, etc. Cette classe des noms peut contenir:

- Les noms d'agent «أسماء الفاعل».
- Les noms de patient «أسماء المفعول»
- Les noms adjectifs «الصفة»
- Les noms d'outil «اسم الآلة»

3.3 Les nombres

Cette catégorie de noms est constituée par les numéraux simples représentant les unités: de " صفر " (sifr- zéro, 0) à " تسعة " (tis'at – neuf, 9) ;

Les dizaines :

- " عشرة " (aşarat–dix, 10),
- " عشرون " ('išruwn – vingt, 20),
- Et ... " تسعون " (tis'uwn – quatre-vingt-dix, 90) ;

Les centaines, etc. ainsi que les numéraux composés tels que les cardinaux de

- " أَحَدَ عَشَرَ " (áhada'aşara-onze, 11) à
- " عشر تسعة " (tis'at-'aşara-dix-neuf, 19).

Dans leur décomposition, les grammairiens arabes ont assimilé les adjectifs à des noms vu qu'ils y prennent presque toutes les formes morphologiques ; ils peuvent, par exemple, être définis ou indéfinis et se fléchir suivant le cas, le nombre et le genre [30].

3.4 Mots isolés

Les mots isolés sont les mots qui n'ont pas de racines, comme les noms propres, les noms communs et les particules.

Le nom propre désigne toute substance distincte de l'espèce à laquelle elle appartient. Il ne possède en conséquence aucune signification, ni aucune définition.

Exemple : Aissa, Alger, Msila, etc.

Le 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 : <نبات, plante>, <حشرة, insecte>, <قط, chat >, etc.»

D'un autre côté, on peut classer les noms en deux grandes classes:

Celle qui regroupe les noms conjugables ou semi conjugables pouvant avoir la forme duelle, plurielle etc. Ces noms sont soit des noms primitifs qui échappent à toute dérivation comme كَبْشٌ [kabšun] (bélier), soit des noms dérivationnels qui sont formés à partir d'une racine comme مَدْرَسَةٌ [madrasatun] (école) de la racine درس [d r s]. tableau 1.7.

Celle qui regroupe les noms non conjugables et qui gardent la forme quel que soit le contexte. Table 1.8.

Non conjugable		Conjugable	
Sous-catégorie	Exemples :	Sous-catégorie	Exemples :
Adverbe	قَبْلَ، أَيْنَ، حَيْثُ	Pronom relatif	الَّذَانِ، اللّوَاتِي
Nom de voix	نَحْ، أَحْ	Nom de nombre	خمسة، عشرون، مئة
Nom de verbe	شَتَّانَ، أَوْهَ، حَيَّ	Pronom démonstratif	هَذَانِ، هَؤُلَاءِ
Pronom Personnel (affixé ou isolé)	هَنَ، هَمَا، أَنْتَ	Nom propre	المسيلة، جمال، زبيدة
Pronom interrogatif	كَيْفَ، مَتَى، مَا		
Pronom conditionnel	مَنْ، إِذَا	Nom commun	قط، بنت، بيت
Pronom allusif	كَمْ، كَيْ		

Table 1.7 : Dérivationnel irrégulier

Conjugable			
Sous-catégorie	Exemples	Sous-catégorie	Exemples
Masdar	الانطلاق، الذهاب	Nom d'instrument	فأس، مفتاح
Participe actif	مجاهد، منطلق	Adjectif	حَسَنٌ، عَظِيمٌ، بَطْلٌ
Participe passif	مكرم، مسموح	Elatif	فُضِّلَ، اكبر
Nom d'une fois	جَلَسَ، ضَرَبَ	Nom diminutif	شَجِيرَةٌ، قَطِيط
Nom de manière	نَظَرًا، خَلَسَ	Nom de relation	حَضَرَمِي، جزائري
Nom de temps	صباح، مشرق	Intensif	غفار، مغوار
Nom de lieu	هنا، مزرعة		

Table 1.8 : Dérivationnel régulier [4]

3.5 La particule

Ce sont principalement les mots outils comme les conjonctions de coordination et de subordination. Elles servent à situer les événements et les objets par rapport au temps et l'espace, et permettent un enchaînement cohérent du texte. La particule est tout ce qui n'est ni un verbe ni un nom et qui n'a de sens que dans une phrase construite.

Exemple:...، على، من، إلى، في،

Elles sont très utiles pour le traitement pour deux raisons :

- Elles font partie de l'anti dictionnaire [31] qui regroupe les termes à ne pas prendre en considération lors du calcul de fréquence de distribution des mots,
- Elles identifient des propositions composant une phrase.

Les particules peuvent avoir des préfixes et suffixes ce qui rajoute une complexité quant à leur identification [7].

Elles sont classées selon leur sémantique et leur fonction dans la phrase en plusieurs types (introduction, explication, conséquence, ...). Elles jouent un rôle important dans l'interprétation de la phrase [37].

On peut distinguer plusieurs types :

- Préposition : exemple (حَتَّى، عَنْ، مِنْ، إِلَى، عَلَى)
- Particules de coordination : exemple (و، فَ، ثُمَّ، أَوْ)
- Particules interrogatives : exemple (مَا، هَلْ، أَمْ)
- Particules d'affirmation : exemple (نَعَمْ، بَلَى، أَجَلْ)
- Particules de négation : exemple (لَا، لَنْ، لَمْ)
- Particules distinctives : exemple (أَيَّ)
- Particules relatives : exemple (مَا)
- Particules de futur : exemple (لَنْ، أَنْ، سَ)
- Particules conditionnelles : exemple (كَيْفَمَا، مَنْ، إِذَا). [4]

4 Morphologie du mot arabe

4.1 Structure du mot arabe

En langue arabe, les mots sont généralement formés d'une agglutination de morphèmes lexicaux et grammaticaux. En arabe, un mot peut représenter toute une proposition. La représentation suivante schématise une structure possible de mot complexe. Notons bien que la lecture se fait de droite à gauche.

Enclitique	Suffixe	Corps schématique	Préfixe	Proclitique
------------	---------	----------------------	---------	-------------

Table 1.9 : Structure du mot arabe

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

Exemple:

«أتتفكروننا» ce mot en arabe représente en français la phrase suivante:

«Est-ce que vous vous souvenez de nous ? », la segmentation correcte de ce mot se fait sous la forme suivante :

Enclitique	Suffixe	Corps schématiques	préfixe	proclitique
نا	ون	تفكر	ت	أ
pronom suffixe complément du nom	suffixe verbal exprimant le pluriel	dérivé de la racine: فكر	préfixe verbal du temps de l'inaccompli.	conjonction d'interrogation

Table 1.10 : La segmentation du mot "أتتفكروننا"

Cet exemple montre bien la richesse morphologique de la langue arabe. Pour identifier les différentes formes soudées par ces phénomènes d'agglutination, et envisager un traitement automatique, il va donc falloir mettre en œuvre une phase spécifique de segmentation. [19]

4.2 Eléments essentiels de la morphologie du mot arabe

La langue arabe a une morphologie riche et différente par rapport aux langues occidentales. L'analyse morphologique d'un mot arabe, consiste principalement à déterminer la structure générale de ce mot, s'il existe, et les autres éléments utilisés pour construire ce mot (les affixes, les modèles...). [1]

Les éléments essentiels de la morphologie de la langue arabe sont :

4.2.1 La racine

Les racines sont à l'origine de la plupart des mots arabes. Elles sont des verbes formés de trois à cinq lettres consonnes (le plus souvent trois) qui représentent une notion bien définie.

Exemple: "KTB" "كتب" «écrire».

Racine : <درس> darasa, a étudié						
Conjugué	درس	Drs	Il a étudié	يدرس	Yidrs	Il étudie
	درسنا	Drsna	Nous avons étudié	يدرسون	Yidrswn	Ilsétudient
	درست	Drst	Elle a étudié	تدرس	Tdrs	Tuétudies
	تدرسون	tdrswn	Vousétudiez	ندرس	Ndrs	Nous étudions
Noms	دارس	Dars	Étudiant	تدريس	Tdris	Enseignement
	مدرسة	Mdras	École	مدروس	Mdrwus	Étudié
	درس	Dars	Cours	دراسة	dirsh	Etude

Table 1.11 : quelques mots dérivant de la racine درس (a étudié)

4.2.2 Le schème

Un schème représente une forme ou modèle général composé de trois consonnes ف [f], ع [‘], 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 ou d’extraire cette dernière à partir d’un mot. Il peut être considéré comme un moule dans lequel on fait couler la racine pour en obtenir un mot. Leur nombre est estimé à presque 900 schèmes. En général, les racines trilatérales sont représentées par le modèle <فَعَلَ, faire>, les racines quadrilatérales sont représentées par le modèle <فَعَّلَ>. Le processus de génération consiste à remplacer la racine du schème par les consonnes de la racine en question, tout en gardant les mêmes voyelles et les mêmes lettres augmentées tout en respectant le même ordre des consonnes, autrement dit, le schème peut être considéré comme un moule sur lequel coule la racine [19].

4.2.3 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 [30].

a) Les préfixes

Les préfixes dépendent des mots auxquels ils s'attachent. En effet, la plupart des mots arabes commencent par le préfixe « ال التعريف », « al altâryif » « 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. La Tableau présente des exemples de chaque type de préfixes [33]

Type	Les préfixes			
	Nom en français	Signification	Nom arabe	Transcription
Préfixes nominaux	L'article de définition	Le	ال	Al (Lamltaarif)
	Les prépositions	Avec	ب	b
		Pour	ل	l
		Comme	ك	k
Préfixes verbaux	La particule du futur	Sera	سـ	s
	Les particules du subjonctif	Pour	ل	l
				
Préfixes généraux	Les conjonctions de coordinations	Et	ف	f
		Et	و	w
	L'article d'interrogation	Est-ce-que	أ	a
				

Table 1.12 : Un exemple des préfixes

Les préfixes peuvent s'enchaîner dans un mot pour former des préfixes composés qui peuvent atteindre jusqu'à quatre lettres (<وبال, wubal, et avec le|la|les>, <وال, wual, et le|la|les>, <بال, bal, avec le|la|les>, <كال, kal, comme le|la|les>, etc.).

Dans ce cas, certains préfixes ne peuvent prendre que la première position, l'article d'interrogation <أ, a> par exemple, d'autres peuvent prendre n'importe quelle position, l'article de définition <ال التعريف, al altâryif, l'article de définition> par exemple (Tableau 1.13).

b) les suffixes

Il y a deux types de suffixes, les suffixes verbaux et les suffixes nominaux. Les premiers dépendent de la transitivité et de la personne conjuguée (Tableau 1.14). 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. [1].

Type	Nombre	Suffixes		
		Signification	Nom en arabe	Transcription
Première personne	Singulier	Moi/mon	ني	Nyi
	Duel/Pluriel	Nous/Notre	نا	Na
Deuxième personne	Singulier	Toi/ton	ك	K
	Duel	Votre/vous	كما	Kma
	Pluriel	Votre/vous	كم	Km
		Votre/vous	كن	Kn
Troisième personne	Singulier	Lui/son	ه	H
	Duel	Eux/leur	هما	Hma
	Pluriel	Eux/leur	هم	Hm
		Eux/leur	هن	hn

Table 1.13 : Un exemple des suffixes divisés selon leurs types

4.2.4 le stem

Un Stem est la dérivation obtenue à partir d'une racine donnée selon un modèle. L'arabe classique à un grand nombre de Stems qui ne sont pas tous utilisables, 2% seulement sont utilisables selon Rashwan [29].

Le Stem correspond à un modèle si et seulement s'il possède le même nombre de lettres et les mêmes lettres dans les mêmes positions. Une exception est accordée aux consonnes <ف, f>, <ع, â>, <ل, l> qui sont les lettres de la racine de base <فعل, fâl, faire>.

Par exemple, le mot : <مدارس, écoles>, est obtenu à partir de la racine <درس, il a étudié> selon le modèle <مفاعل, mfaâl>. Les Stems produits ne sont pas tous utilisables. [1]

Racine	Modèle	Stem	Utilisable
<كتب, ktb, il a écrit>	<فعل, fal, faire>	<كتب, ktb, il a écrit>	Oui
<درس, drs, il a étudié>	<فاعل, faal>	<دارس, dars, synonyme d'étudiant>	Oui
<أكل, akl, il a mangé>	<مفعول, mfawul>	<مأكول, makwul, mangeable >	Oui
<لعب, lab, il a joué>	<أفعلاء, afala'>	<العباء, alaba'>	Non

Table 1.14 : Un exemple de génération des Stems.

5 Conclusion

La réalisation de ce chapitre nous a permis de cerner un nombre important de particularités et de caractéristiques propres à la langue arabe qui rendent plus complexe son traitement et exigent des approches spécifiques (non celles appliquées pour les langues occidentales). Une panoplie de techniques utilisées dans ce contexte sera exposée au chapitre suivant.

CHAPITRE 2

Analyse morphologique de la langue arabe

La morphologie est le domaine de la linguistique qui étudie la structure interne et les processus de formation des mots. Un morphème est souvent défini comme la plus petite unité significative d'une langue qui ne peut être décomposée en petites parties. [22], Il existe deux types de morphèmes : Le morphème principal qui s'appelle le mot Stem ou Racine, et les affixes. La racine est la forme de base du mot qui donne sa signification principale, tandis que les affixes ajoutent d'autres significations au mot.

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

1 Analyse morphologique

De plus, l'analyseur morphologique doit permettre l'analyse des différents types de mots arabes, à savoir, les noms, les verbes, les particules, etc. Pour réaliser cela, le système doit passer par des étapes, à savoir, la segmentation du texte en mots, le prétraitement morphologique, l'analyse affixale, l'analyse morphologique et le post traitement [10].

1.1 Segmentation

Cette partie peut être appelée aussi analyse lexicale, elle a pour but de segmenter le texte en mots. Pour la plupart des applications de traitement automatique des langues naturelles (e.g., l'analyse de texte, l'extraction d'information, le résumé automatique) la segmentation devient une phase importante pour repérer les segments contenant les informations recherchées. Ainsi par exemple, commencer une analyse d'un texte sans le segmenter en phrases conduit à des résultats peu fiables; de même, avoir un mauvais segmenteur conduit à accumuler les erreurs du traitement automatique du texte.

La segmentation consiste à désambiguïser les frontières des phrases et des paragraphes et se base généralement sur un ensemble de règles de segmentation. C'est une phase non triviale

pour toute application en TALN. En effet, segmenter un texte nécessite le repérage des frontières formelles marquées par des signes typographiques. Par ailleurs, dans les textes arabes actuels, les signes de ponctuation ne sont pas très utilisés et dans le cas où ils y figurent, ils ne sont pas gérés par des règles d'utilisation [9].

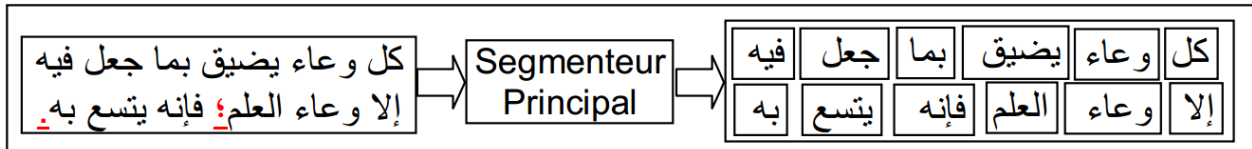


Figure 2.1 : Exemple d'un Segmenteur

1.2 Prétraitement morphologique

Cette étape consiste à supprimer les proclitiques et les enclitiques pouvant être agglutinés au mot en se basant sur les listes de proclitiques et d'enclitiques. Le mot restant sera filtré pour tester s'il s'agit d'une particule, d'un nom propre, d'un nombre ou d'une date. Si ce n'est pas le cas, il va subir l'analyse affixale dont le but est de déterminer toutes ses caractéristiques morphologiques possibles. [16]

1.3 Stemming:

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

Les algorithmes de racinisation en arabe les plus connus sont ceux de [15] et [27]. Le Stemming arabe présenté par Khoja est considéré par un certain nombre de chercheurs comme un Stemming standard [27] pour Modern Standard Arabe (MSA) et aussi pour les dialectes arabes, une des étapes principales dans le Stemming est la normalisation qui vise à transformer une copie du document original dans un format standard plus facilement manipulable. Cette étape est considérée nécessaire à cause des variations qui peuvent exister alors de l'écriture d'une même unité lexicale. Donnons comme exemple :

- Suppression des caractères spéciaux ;
- Remplacement de la lettre (ة) par (ه) ;

- Remplacement des lettres ا, آ, إ par ; ا
- Remplacement de la lettre finale ي par ى. [16]

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

- **Light Stemming** : s'intéresse seulement au Stem.

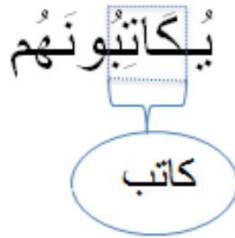


Figure 2.2 : Exemple de light Stemming

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

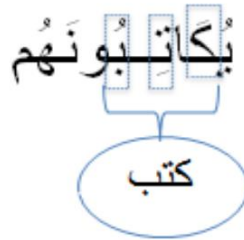


Figure 2.3 : Exemple de heavy Stemming.

1.4 Analyse affixale

La langue arabe possède des propriétés structurelles spécifiques. Un mot du vocabulaire est constitué d'une racine (trilitère ou quadrilatère) ou d'une forme canonique à laquelle on ajoute une combinaison affixale formée d'un préfixe, d'un infixé ou deux, et d'un suffixe. L'analyse affixale a pour objectif de reconnaître les éléments de base qui entrent dans la constitution d'un mot à savoir, la racine (R) ou la forme canonique et les affixes (préfixe (P), infixé (I) et suffixe(S)). Cette opération est effectuée en plusieurs étapes [12]. D'une étape à l'autre un mécanisme de filtrage permet d'éliminer les décompositions par asitesre connues. On distingue les principales étapes suivantes :

- identification des couples (P, S),
- identification des triades affixales candidates,

- filtrage lexical, contrôle des associations (R) et (P, I, S) et reconnaissance des transformations.

1.5 Analyse morphologique

Cette étape consiste à déterminer, à partir de la forme (R, P, I, S) obtenue pour chaque mot, toutes ses caractéristiques morpho-syntaxiques possibles (i.e. partie de discours, genre, nombre, temps, personne, etc.). La détection des caractéristiques morpho-syntaxiques se fait en trois phases :

- La première phase consiste à identifier la catégorie principale du mot (partie de discours), à savoir verbe, nom, adjectif, etc.
- La deuxième phase permet de déterminer pour chaque catégorie, identifiée au niveau de la phase 1, la liste de ses caractéristiques morphologiques.
- Un filtrage des listes de caractéristiques s'effectue en une troisième phase. [11]

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

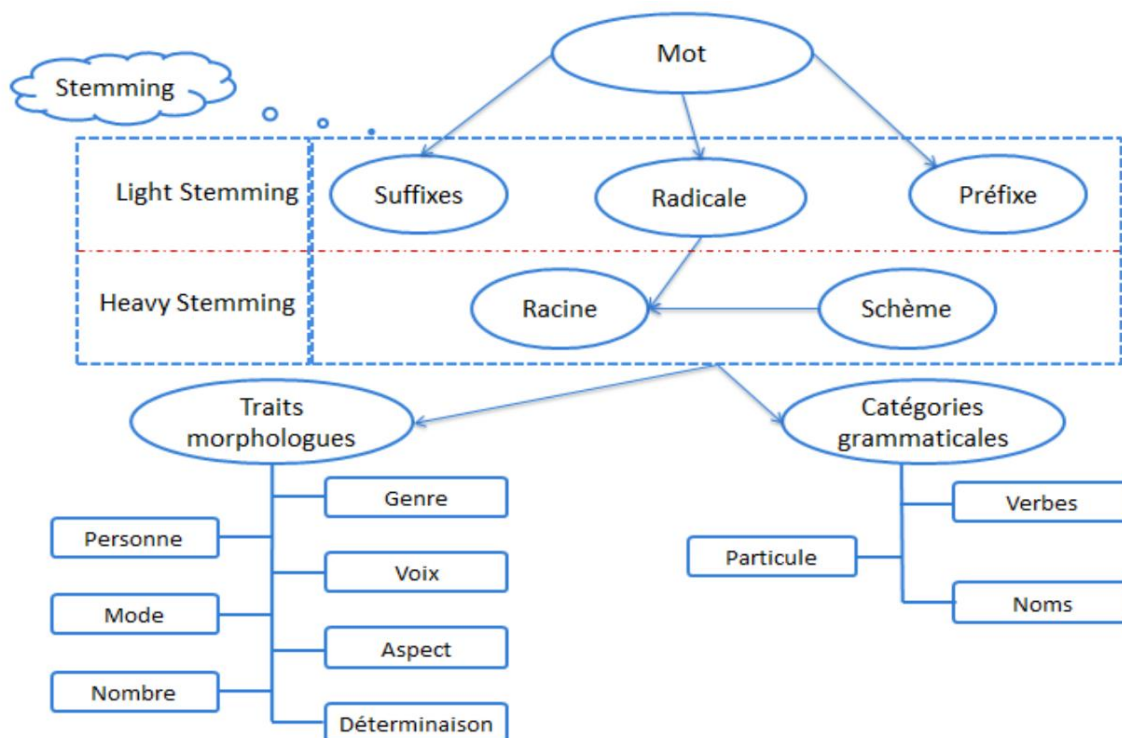


Figure 2.4 : Analyse morphologie d'un mot

2 Les méthodes d'extraction de la racine d'un mot arabe

Dans cette partie, nous présentons un survol des méthodes d'extraction de la racine à partir des mots arabes. Dans cette section, une vingtaine des méthodes sont citées et réparties sur cinq grandes classes :

1. les méthodes basées sur l'analyse morphologique.
2. les méthodes basées sur la génération systématique.
3. les méthodes basées sur la transcription.
4. les méthodes basées sur l'analyse statistique.
5. les méthodes basées sur la position des lettres.
6. les méthodes basées à la fois sur les analyses morphologique et statistique (Hybride).

2.1 Les méthodes basées sur l'analyse morphologique

2.1.1 Méthodes basées sur l'élimination des affixes avec utilisation de dictionnaire :

Pour extraire la racine d'un mot, on procède en deux étapes [20], [35]. 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 [17]. Le radical réécrit est vérifié à nouveau dans le dictionnaire principal. Cette méthode utilise les ressources linguistiques suivantes :

- ✓ Un dictionnaire principal (5,4 millions d'entrées) qui contient pour chaque mot ses caractéristiques possibles (genre, nombre, etc.).
- ✓ le dictionnaire proclitique (77 entrées).
- ✓ le dictionnaire enclitique (65 entrées).

Ces dictionnaires ont la même structure que le dictionnaire principal avec des versions vocalisées et non vocalisées pour chaque combinaison valide de clitiques.

2.1.2 Méthodes basées sur l'élimination des affixes sans utilisation de dictionnaire :

Une approche similaire qui est basée sur la normalisation et l'élimination des affixes, mais sans utiliser de dictionnaire, a été proposée initialement par Aljlayl et Frieder [6] et ensuite modifiée par d'autres. Chen et Gey [14] ont identifié d'autres ensembles d'affixes et de règles. Par exemple, si le mot est formé d'au moins cinq lettres ils enlèvent les préfixes :

<، lal>، <سال، sal>، <مال، mal>، <كال، kal>، <ولل، wull>، <ال، aal>، <فال، fal>، <بال، bal> et <ال، al>.

Kanaan et al. [24] 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. Al Ameen et al. [5] ainsi que Larkey et al. [28] ont développé un algorithme qui est basé sur la suppression de «و» au début du mot, des articles définis (<ال, al>, <وال, wual>, <بال, bal>, <كال, kal>, <فال, fal>, <الل, ll>) et des suffixes (<ي, yi>, <ه, h>, <ية, iyiah>, <يه, yih>, <ين, yin>, <ون, wun>, <ات, at>, <ان, an>, <ها, ha>), Kadri et Nie [23] ont introduit deux nouvelles catégories d'affixes : 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.

Khoja et Garside [25] ont proposé une méthode 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, cette racine est validée par rapport à une liste de racines connues (dictionnaires).

2.1.3 Méthodes basées sur l'élimination des affixes et les modèles sans utiliser de dictionnaire :

Cette méthode procède en trois étapes principales :

- ✓ La première est la normalisation.
- ✓ La deuxième, consiste à extraire le lexème, en supprimant les lettres qui représentent les caractéristiques (temps, nombre, personne, etc.) à partir du mot.
- ✓ La dernière étape consiste à rechercher le modèle qui a été utilisé pour construire le lexème obtenu dans l'étape précédente.

Taghva et al. [36] ont présenté une autre alternative pour identifier les modèles et pour extraire en plus des racines quadrilatérales. En plus des ensembles concernant les diacritiques et les affixes, ils ont défini plusieurs ensembles qui caractérisent les modèles selon leurs longueurs. Parmi ces ensembles:

PR4 : Modèles de longueur 4 (فعلة، فاعل، فعول)

PR53: Modèles de longueur 5 ET d'une racine de longueur 3 (...أفعال، افتعل، افتعل)

PR54: Modèles de longueur 5 ET d'une racine de longueur 4 (...مفعّل، افعلّ، افعلّ)

PR63: Modèles de longueur 6 ET d'une racine de longueur 3 (...استفعل، مفعالة، استفعل)

PR64: Modèles de longueur 6 ET d'une racine de longueur 4 (...متفعل، افعلّ، متفعل)

2.2 Les méthodes basées sur la génération systématique

Le principe général des méthodes de cette classe est assez simple. 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.

Dans ce cadre, Goweder et al. [21] ont proposé une nouvelle méthode qui sert à trouver la racine d'un pluriel irrégulier. Le dictionnaire global qui contient 3600 Stems pluriels irréguliers environ a été construit manuellement. Cette liste est ordonnée par ordre alphabétique des racines. Ce dictionnaire a été validé par un linguiste.

Le centre Européen de recherche Xerox a proposé un analyseur morphologique pour l'arabe standard moderne qui est basé sur des dictionnaires. En 2001 [8], le système a été largement modifié en utilisant la technologie d'états finis de Xerox pour construire le dictionnaire global. Cette technologie est basée sur des outils qui fournissent des algorithmes pour créer des automates des expressions régulières et des formalismes d'équivalence pour réaliser des opérations telles que l'union, la composition ou le remplacement. Ce système est utilisé dans beaucoup d'applications linguistiques telles que l'analyse morphologique et la tokenisation [8].

Kadri et Yagi 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 médial (M),
- ❖ le dernier radical dans une racine trilatérale (L),
- ❖ le dernier radical dans une racine quadrilatérale (Q).

Un modèle est aussi représenté par un ensemble $\{x_1, \dots, x_n\}$ ou x_i désigne la lettre à la $i^{\text{ème}}$ 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 $\{x_1, \dots, x_n\}$ 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. [1].

2.3 Les méthodes basées sur la transcription ou la traduction

Cette méthode donne 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
- 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. Lors de la phase de l'extraction de la racine, cette méthode procède en cinq étapes ;

- ✓ La deux premières sont la transcription et la normalisation.
- ✓ La troisième consiste à sélectionner de 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 de 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 [1].

2.4 Les méthodes basées sur l'analyse statistique

Adamson et Boreham [3] ont développé la première technique de la classification automatique basée sur la structure des mots. Le coefficient de la ressemblance est calculé à partir du nombre de digrammes (2-gram) assortis dans des paires de sous-lettres (Figure 3.4). Un échantillon des 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.

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

$$\delta_n(a, b) = \frac{|\alpha \cap \beta|}{|\alpha \cup \beta|} \quad (1) \quad \text{Où } \alpha \text{ et } \beta \text{ sont les ensembles de N-gram.}$$

Par exemple, prenons les deux mots <استمرارية, alistmraryiah> la continuité qui est désigné par (a) et <استمرار, istmrar, suite> qui est désigné par (b) dans l'équation (1). (a) et (b) sont décomposés en 2-gram, comme le montre la figure. Dans ce cas, les ensembles α et β sont donnés par ; {<إس, is>, <ست, st>, <تم, tm>, <مر, mr>, <را, ra> et <ار, ar>}, Et {<ال, al>, <لا, li>, <إس, is>, <ست, st>, <تم, tm>, <مر, mr>, <را, ra>, <ار, ar>, <ري, ryi> et <ية, ah>}, respectivement. Alors, le coefficient de ressemblance δ_2 ($n=2$) qui est calculé selon l'équation (1) est égal à 0.6

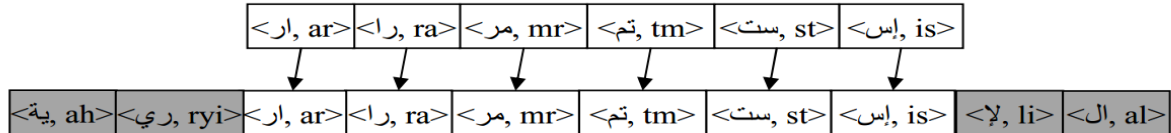


Figure 2.5 : Mesure de similarité (bigramme) entre deux mots.

Mustafa [1] 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. [34] 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. Ceci est dû au fait que les n-grams basés sur les mots peuvent être des affixes, ce qui induit des erreurs de similarité entre le document et la requête. 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 des documents du texte 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.

2.5 Les méthodes basées sur la position des lettres

Ce genre d'approches est considéré comme une approche aveugle qui tend à trouver la racine d'un mot sans même avoir besoin d'une table de racines, d'une liste de schèmes ou d'une liste de préfixes ou de suffixes des mots arabes. En effet, elle essaye de prédire les positions des lettres qui forment la racine du mot en classifiant les lettres arabes selon leurs positions dans le mot. Dans

cette approche, à chaque lettre est assignée une valeur entre 0 et 1 pour indiquer qu'elle appartient ou non à la racine. Les valeurs assignées aux lettres du mot dépendent de deux choses :

- La lettre fait partie de l'affixe ou non.
- Sa position si c'est une lettre d'affixe (au début, au milieu ou à la fin du mot).

Dans [2], Fatma Abu Hawas a proposé un système pour trouver les racines d'un mot en se basant sur le principe de la position des lettres. Son travail s'est articulé autour de deux parties principales. La première introduit quelques règles pour distinguer entre l'article de définition (ال) et le composant permanent (ال, al) qu'on peut trouver dans un mot quelconque. Dans la deuxième partie, elle classe les lettres arabes en groupes selon leurs positions dans les mots. Le système proposé est composé de plusieurs modules pour extraire les racines, et a été évalué en utilisant le saint Coran.

2.6 Les méthodes basées sur l'analyse statistique et morphologique (Hybride)

De Roeck et Al-Fares [18] ont proposé une méthode hybride qui est basée à la fois sur les analyses morphologique et 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 [3] entre le stem donné par la première étape et une liste des racines sélectionnées dans un dictionnaire. Les racines correspondantes aux coefficients de ressemblance maximum sont ajoutées à la liste des racines.

3 Conclusion

Dans ce chapitre, nous avons présenté un survol des méthodes d'extraction de la racine à partir d'un mot arabe. Les approches utilisées par ces méthodes sont basées sur des analyses morphologiques ou statistiques. En effet, les méthodes basées sur la transcription ou la génération systématique utilisent aussi les analyses morphologiques pour extraire la racine ou pour construire le dictionnaire global. Néanmoins, on peut facilement remarquer que les approches morphologiques sont plus simples à comprendre et à implémenter et reposent fortement sur les caractéristiques propres de la langue arabe ce qui nous pousse à considérer qu'elles sont plus appropriées pour notre approche.

L'avantage principal de l'approche statistique est qu'elle ne nécessite aucune connaissance préalable de la langue. Par contre, le premier avantage de l'approche morphologique est sa rapidité, le deuxième est qu'elle est plus appropriée dans le cas des mots complexes qui ne sont pas dans un dictionnaire.

CHAPITRE 3

Conception et Réalisation

Après avoir pris connaissance, dans les chapitres précédents, du domaine de la langue arabe avec toute sa complexité et les différentes méthodes de son traitement morphologique avec tous les défis rencontrés pour extraire les composants essentiels d'un mot quelconque, nous allons exposer dans ce qui suit, notre contribution qui consiste à choisir l'une de ces méthodes et essayer de l'implémenter avec quelques améliorations.

1 Description du système réalisé:

En général, notre méthode consiste à rechercher le radical à partir du mot et après on va extraire la racine (cette figure représente notre travail en général) .

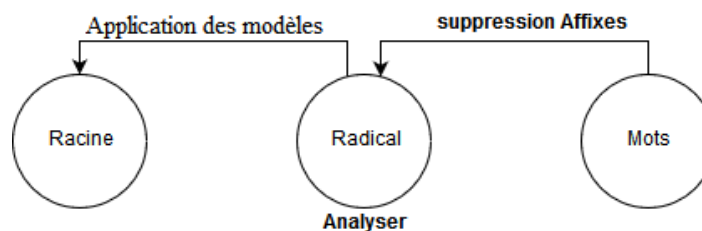


Figure 3.1 : Plan de travail

❖ Le système d'extraction

a) Le filtrage général

- ✓ Suppression des mots vides
- ✓ Suppression des caractères spéciaux
- ✓ Remplacer les lettres
 - Remplacer la lettre (ة) par (o) ;
 - Remplacer les lettres ل, ا, ؤ, ع par ; ا
 - Remplacer la lettre finale ي par ى.

b) Élimination des affixes

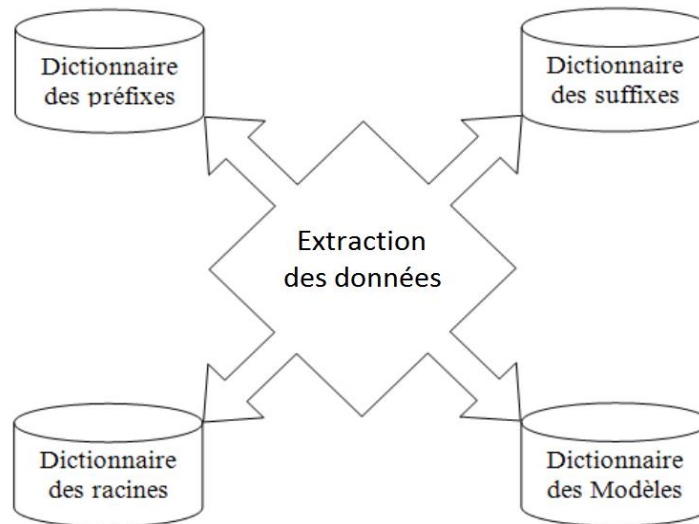


Figure 3.2 : Extraction des données

Nous avons utilisé une méthode qui permet d'éviter les mots qui ne contiennent pas des préfixes ou des suffixes, On teste la première lettre (Préfixe) si elle appartient à l'ensemble des lettres :

{ "ه", "ق", "غ", "ع", "ظ", "ط", "ض", "ص", "ش", "ز", "ر", "د", "ذ", "خ", "ح", "ج", "ث" }

et la dernière lettre (suffixe) si elle appartient à l'ensemble des lettres :

{ "ف", "ل", "ق", "غ", "ع", "ظ", "ط", "ض", "ص", "ش", "ز", "ر", "د", "ذ", "خ", "ح", "ج", "ث", "ب", "س" }.

Puis, On passe à des méthodes d'élimination des affixes

a. Elimination des préfixes

Nous avons comparé les préfixes détectés avec les préfixes prédéfinis (Dictionnaire des préfixes)

b. Elimination des suffixes

Nous avons éliminé les suffixes détectés avec les suffixes prédéfinis (Dictionnaire des suffixes)

Après l'élimination, la chaîne qui reste est appelée « Radical ».

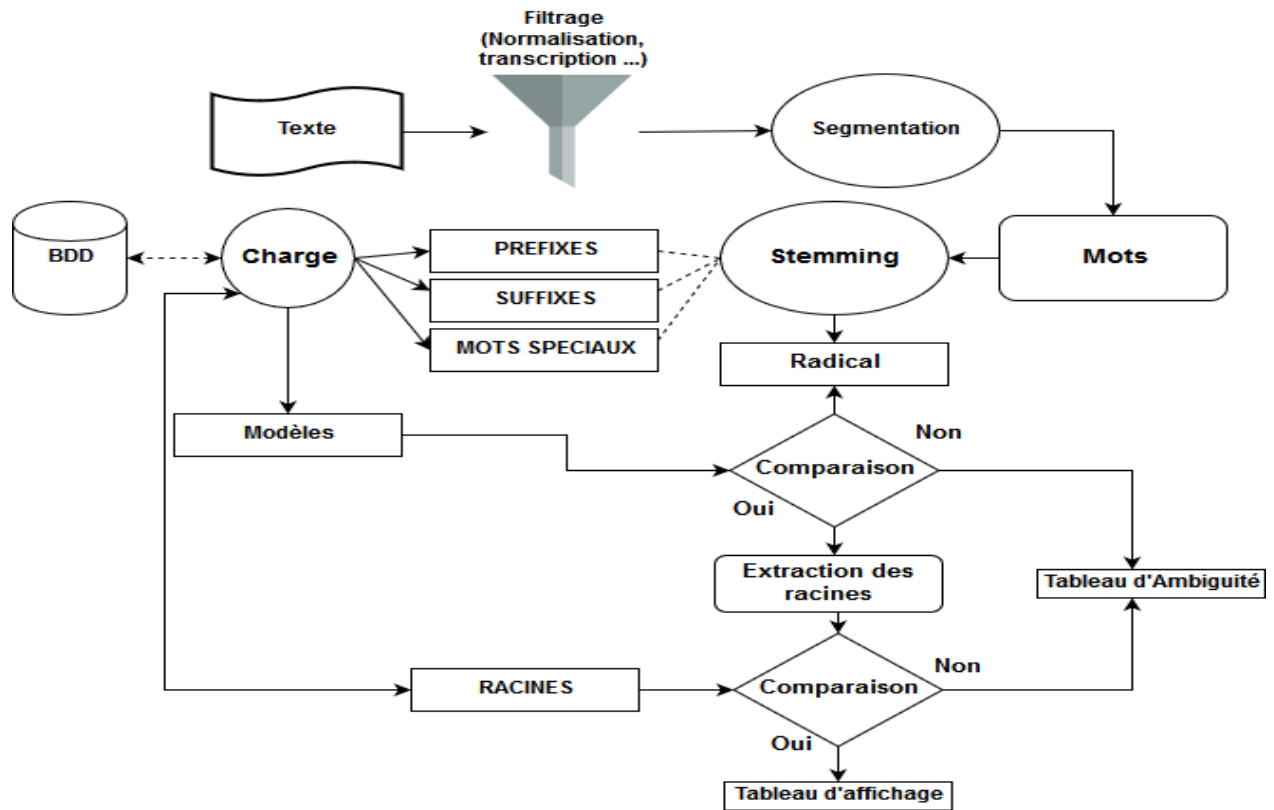


Figure 3.3 : Extraction des Racines

c) Extraction des modèles et des radicaux à partir du Mot

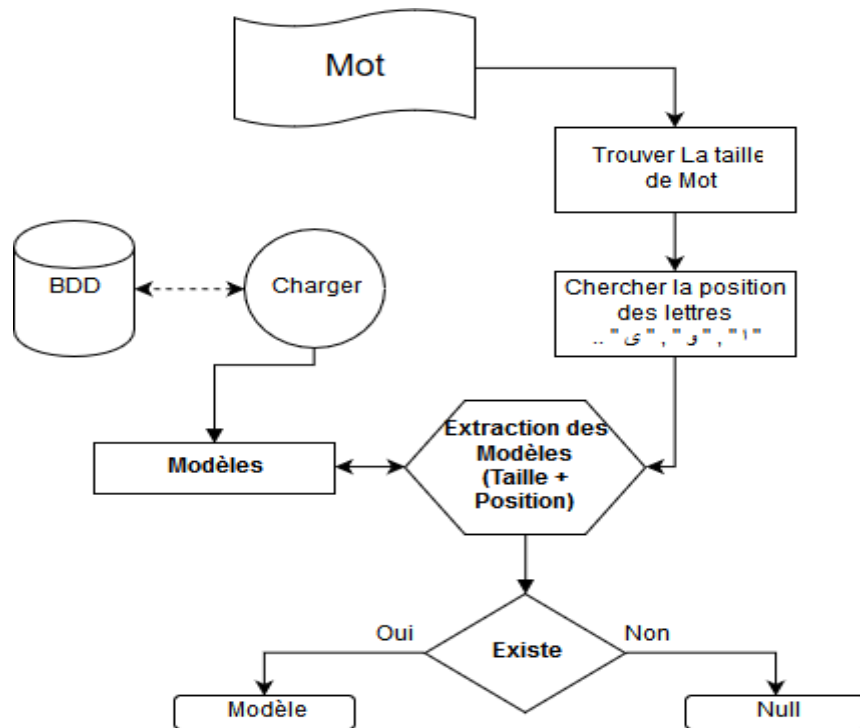


Figure 3.4 : Extraction des Modelés

Nous avons comparé les positions des lettres " ا ", " و ", " ي " dans le mot avec celles des modèles ayant la même taille.

- ❖ Si la taille du mot est égale à 3 Alors le mot est la racine
- ❖ Si la taille est supérieure à 3 donc on va vérifier la position des lettres " ا ", " و ", " ي " pour extraire le radical ; On utilise la méthode générale qui calcule la position de chaque voyelle dans le mot traité et après on extrait le modèle qui contient les même voyelles dans la même position de notre mot.

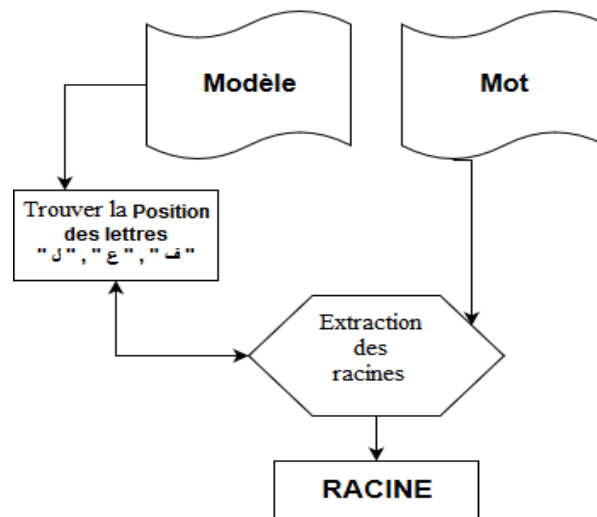


Figure 3.5 : Extraction des Racines

Par exemple : pour le mot « كاتِب » nous avons la lettre " ا " qui est à la 2^{ème} position, alors on va chercher dans notre base de données tous les radicaux qui contiennent la lettre " ا " à la même position que notre mot et ayant la même taille ; alors, le radical est « فاعِل ».

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

Par exemple :

La racine	Le schème	Résultat
كتب	فعال	كتاب

Table 3.1 : Exemple de construction des mots à partir d'un schème

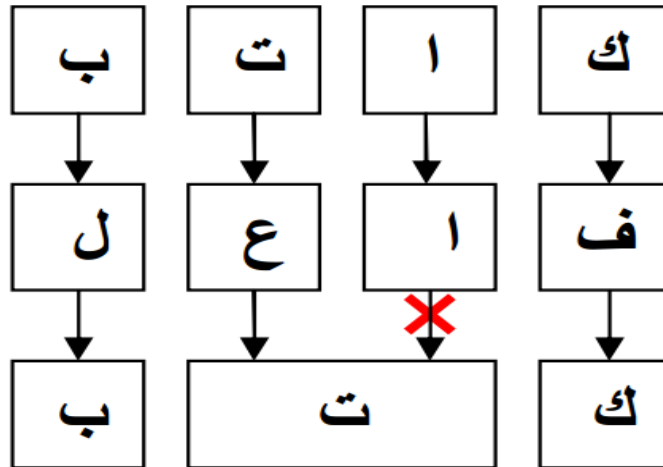


Figure 3.6 : Exemple d'extraction des racines

Nous avons trouvé la position des lettres " ل ", " ع ", " ف " puis, on a extrait la racine par rapport aux positions.

2 Présentation de l'Application

❖ Présentation de la BDD

a) Les tables du système utilisées :

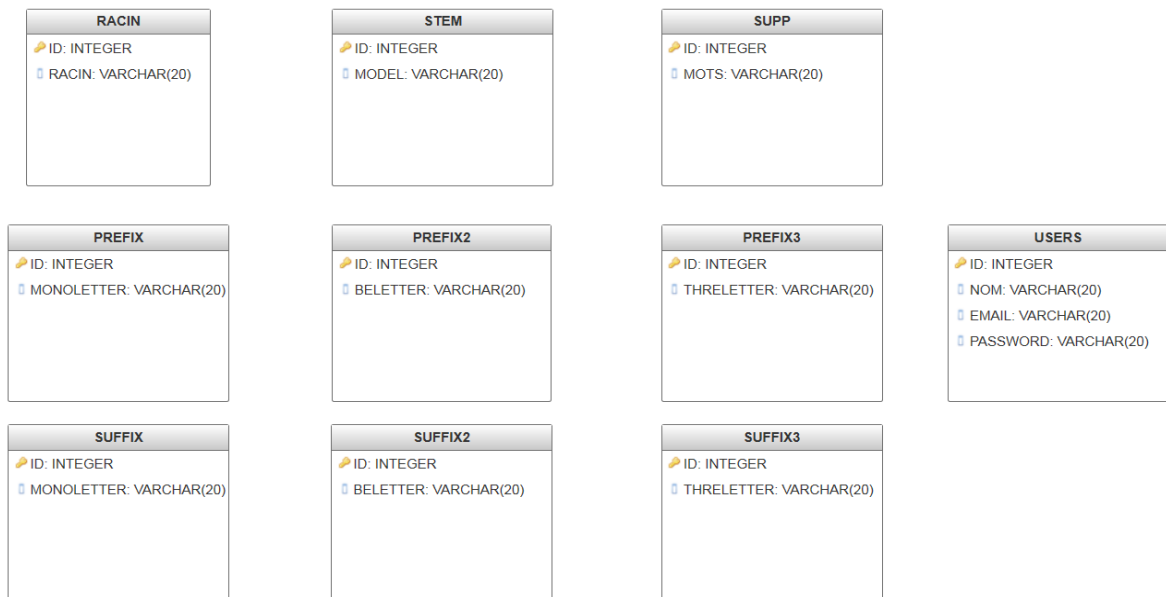


Figure 3.7 : Tableau de la base de données

a. Les racines

La table 4.1, sert à contenir l'ensemble des racines choisies qui seront utilisées pour vérifier si la racine obtenue après extraction est une racine valable ou non. Le nombre de racines initial pris dans notre cas est de 3823 racines.

Racins
كتب
ذهب
اكل
درس
علم
أبى
أبد
أبر
أبض
< >

Figure 3.8 : Tableau Racine

b. Les modèles

Sert à rassembler les différents modèles choisis auxquels vont être comparés les radicaux obtenus après l'élimination des suffixes pour voir quel est le modèle adéquat à partir duquel on peut obtenir le modèle cherché. Le nombre initial de modèles de notre table est 50.

Modeles	Taille
فعل	3
فاعل	4
فعول	4
افتعل	5
افعال	5
مفعول	5
افعلل	5
< >	

Figure 3.9 : Tableau Modèles

c. Les mots vides

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

Mots Vide	Taille
وله	3
ذات	3
اي	2
بدلا	4
اليها	5
انه	3
الذين	5
< >	

Figure 3.10 : Tableau Mots Vides

d. Les préfixes

Contient l'ensemble de préfixes possibles qui peuvent être rencontrés au début d'un mot arabe. Elle sert à vérifier si une séquence de lettres au début d'un mot est un préfixe ou non. Leur nombre est de 37.

Prefixe	Prefixe	Prefixe
ل	ال	وال
س	سن	كال
ب	فا	بال
ن	كت	وست
م	يا	فال
ت	مت	افي
ف	وت	افت

Figure 3.11 : Tableau des Préfixes

e. Les suffixes :

C'est une table similaire à la table précédente mais qui concerne la séquence de lettres se trouvant à la fin des mots. Ils sont au nombre de 28.

Suffixe	Suffixe	Suffixe
ا	وه	هما
ه	ته	تما
ى	هم	كما
و	ين	
ء	وا	
ة	ان	
ن	تم	
	هن	

Figure 3.12 : Tableau Suffixe

b) Les plans de l'interface

Ce plan, qui est représenté l'interface au mode visiteur.

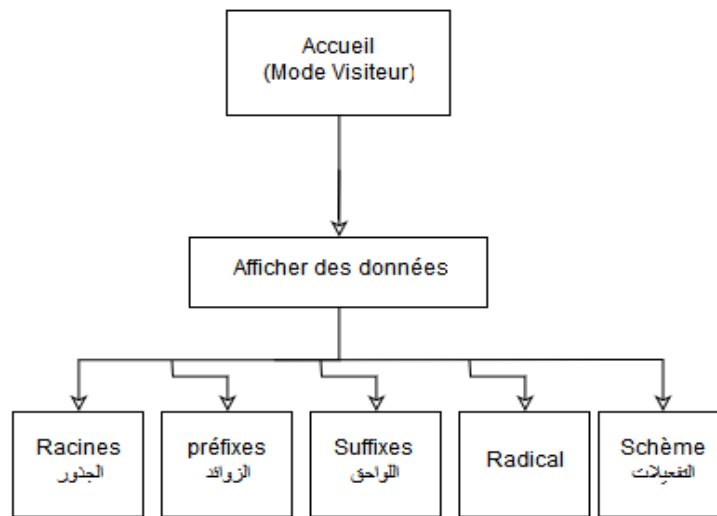


Figure 3.13 : Plan du mode « Visiteur »

Ce plan, qui est représenté l'interface au mode Utilisateur.

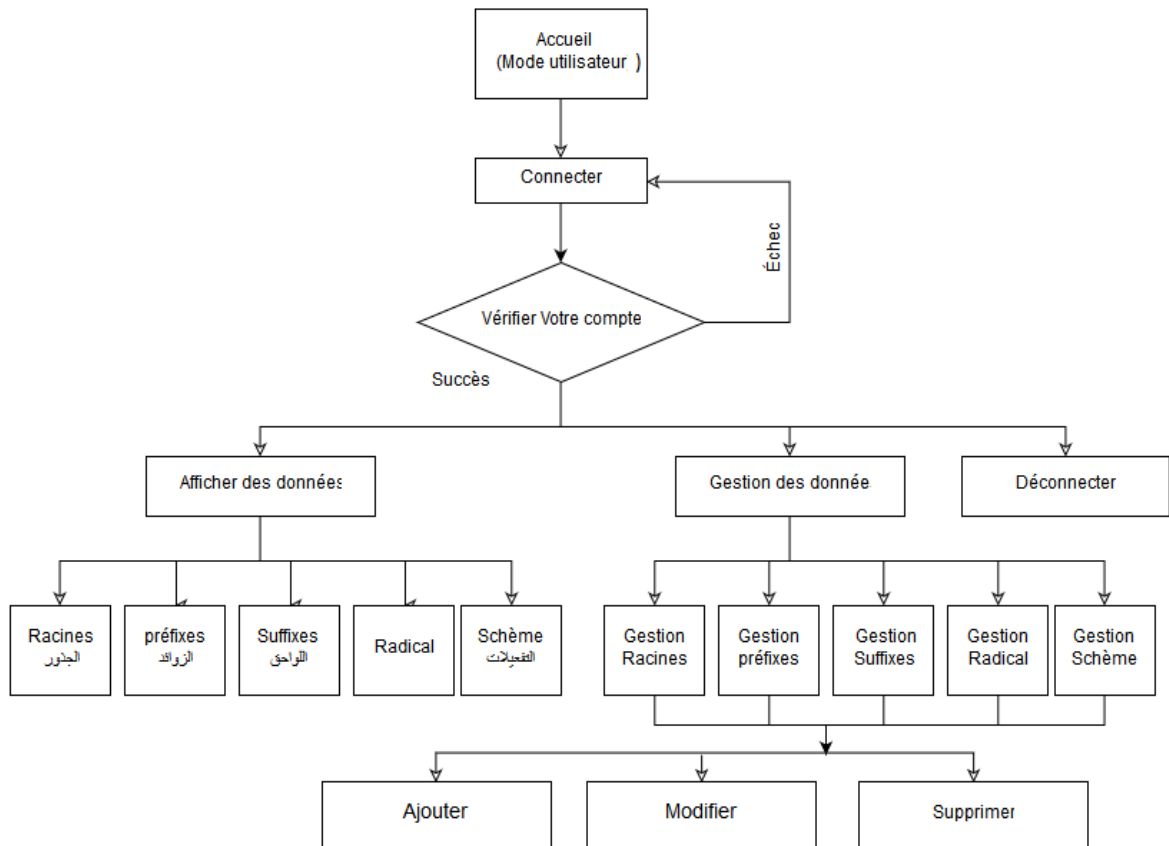


Figure 3.14 : Plan du mode « Utilisateur ».

1. Bouton « Accueil » : Entrer à l'interface principale.
 2. Bouton « Ma BDD » : Entrer à l'interface d'affichage de notre base de données.
 3. Bouton « Gestion BDD » : Entrer à l'interface de Gestion de notre Base De Données.
 4. Bouton « A Propos » : Entrer à l'interface de description de notre application.
 5. L'interface de connexion à l'application : ici pour spécifier les zones de modification ; Quelques zones pour les utilisateurs seulement, d'autres pour les clients et d'autres pour l'administrateur.
 6. Bouton « Exit » : Sortie du programme.
- II. Bouton « Accueil » : il montre l'interface principale, qui contient tous les éléments de notre Programme

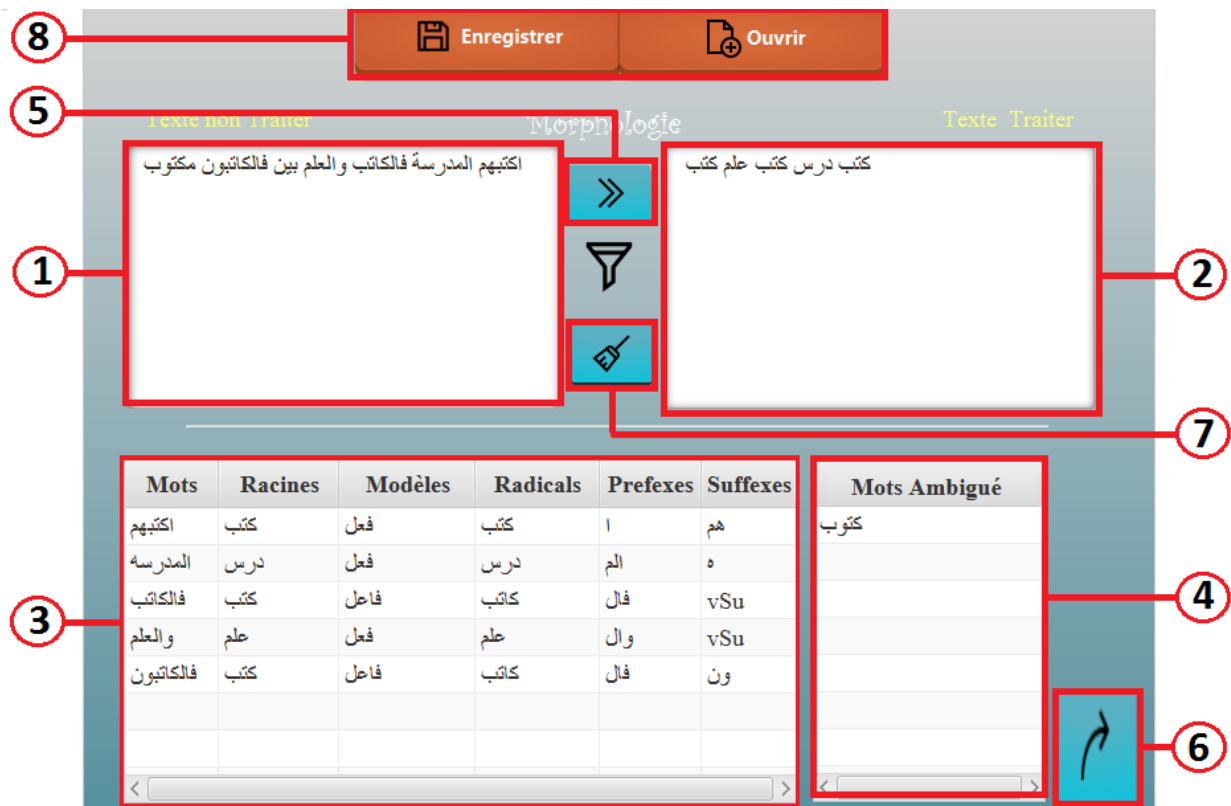


Figure 3.17 : Interface d'Accueil

1. Champ de saisie de notre texte
2. Champ pour Afficher le résultat du filtrage
3. Tableau montrant la segmentation de notre texte « Mots, Racines, Modelés, Radicales, préfixes, suffixes ».
4. Tableau des mots ambigus sur notre dictionnaire
5. Bouton de filtrage
6. Bouton pour passer à l'interface de manipulation des mots ambigus
7. Bouton pour vider tous les champs
8. Bouton pour ouvrir ou enregistrer le texte

Bouton 6) l'interface de manipulation des mots ambiguë

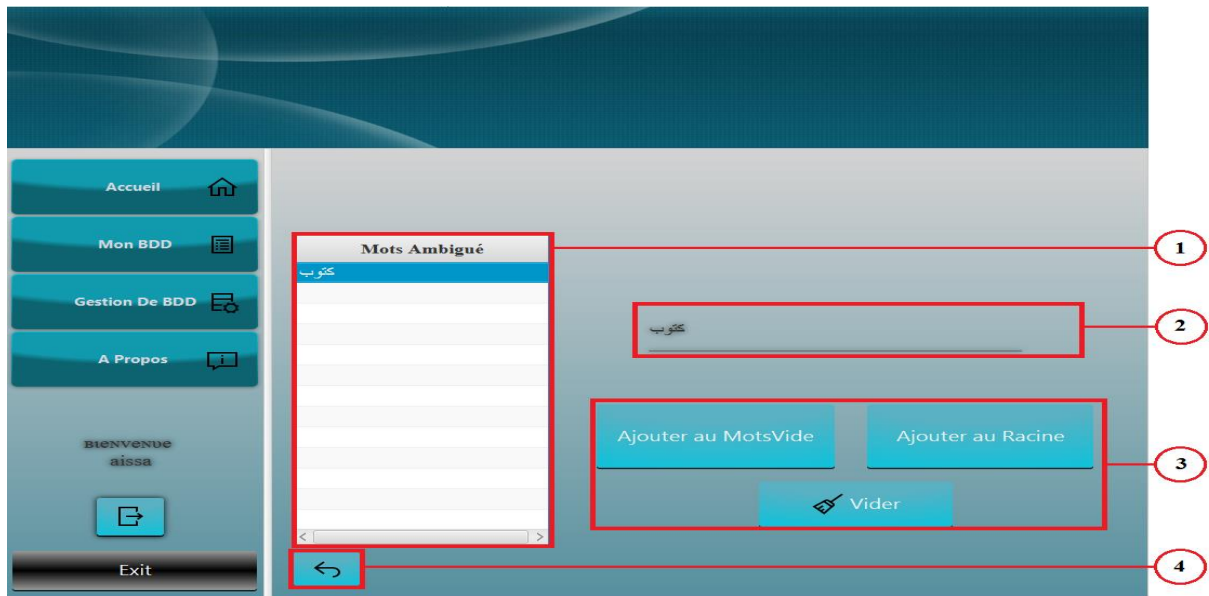


Figure 3.18 : Interface de management

6-1 Table des mots ambiguë.

6-2 champs de modification de mot

6-3 les boutons de manipulation sur nos BDD.

6-4 Bouton de retour au page d'accueil.

III. Bouton « Mon BDD » : Affichage de notre BDD



Figure 3.19 : Interface de notre BDD « Ma BDD »

1. Bouton « RACIN » : Affichage nos racines.
 2. Bouton « PREFIX » : Affichage nos préfixes.
 3. Bouton « SUFFIX » : Affichage nos suffixes.
 4. Interface des racines
 5. Cette partition pour la recherche
 6. Tableau de nos racines.
- IV. Bouton «Gestion BDD »:Affiche l’interface de Gestion sur nos BDD (Pour les utilisateurs)

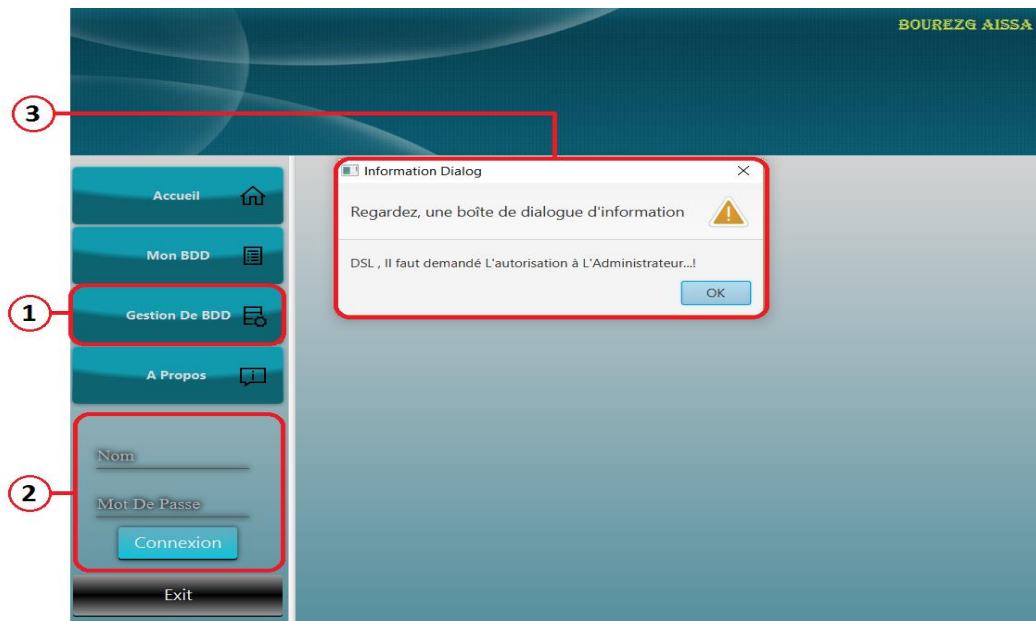


Figure 3.20 : Interface de Gestion sur nos BDD « Gestion de BDD »1

1. Bouton « Gestion de BDD » ; si le client est cliqué sur le Bouton, l’Application empêcher et afficher (3)
2. Interface pour rentrer les informations des utilisateurs
3. Boite de dialogue que signifier cette espace pour les utilisateurs.



Figure 3.21 : Interface de notre Application « Gestion de BDD »2

Description la figure 3.21 :

1. Bouton « RACIN » : manipulation de nos racines
2. Bouton « PREFIX » : manipulation de nos préfixes.
3. Bouton « SUFFIX » : manipulation de nos suffixes.



Figure 3.22 : Interface de notre Application « Suffixe »

3-1 Bouton « Suffixe»

3-2 Tableaux des suffixes, chaque'un son taille

3-3 Champ de saisie suffixe

3-4 Les Boutons de manipulation sur les suffixes

4. Bouton « Autre » : manipulation des autres tableaux « Modelés, Mots Vides,... ».



Figure 3.23 : Interface de notre Application « Autre »

4-1 pour Afficher l'interface des tableaux « Modelés, Mots Vides, Radicaux ».

4-2 tableaux qui représentent nos Bases de données.

4-3 Champ de saisie dans les tableaux.

4-4 Les boutons de manipulation de nos données.

5. Bouton « Ajouter des utilisateurs » : manipulation de nos utilisateurs (partition pour l'administrateur seulement).

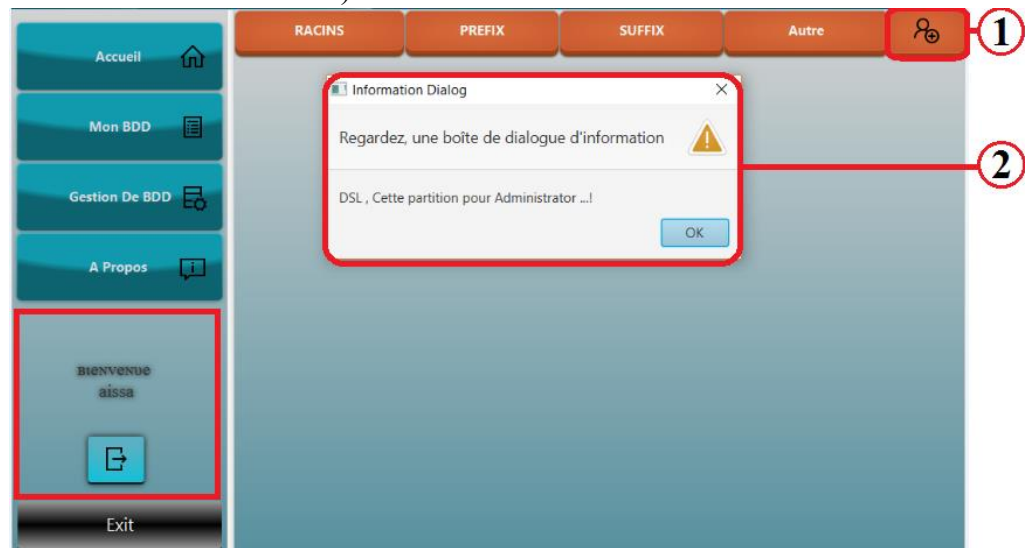


Figure 3.24 : Interface de notre Application « Ajouter Utilisateur »1

5-1 Bouton « Ajouter utilisateur »

5-2 Boite de dialogue pour informer l'utilisateur cette espace pour l'Administrateur seulement.

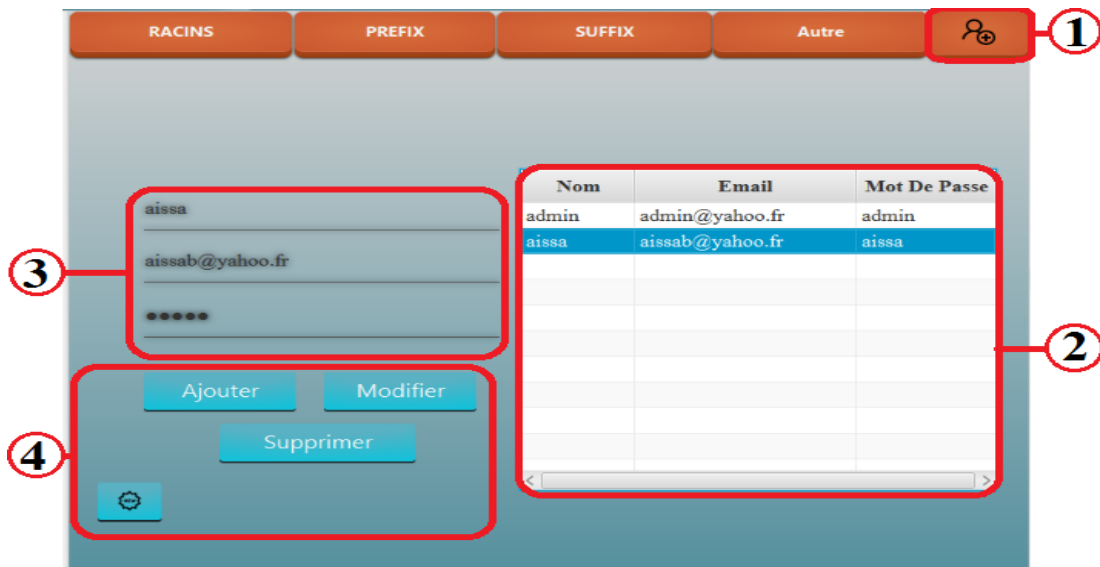


Figure 3.25 : Interface de notre Application «Ajouter Utilisateur» 2

5-1 : Bouton « Ajouter utilisateur ».

5-2 : Tableau des utilisateurs.

5-3 Et 5-4 : Les champs pour manipuler nos informations.

6. Pour ajouter ou supprimer ou modifier nos racines.

7. Tableau de nos racines avec leur longueur.

8. Cette interface représente le nom de l'utilisateur qu'il va connecter.

V. Bouton « A Propos » : qui Affiche des informations sur notre application.

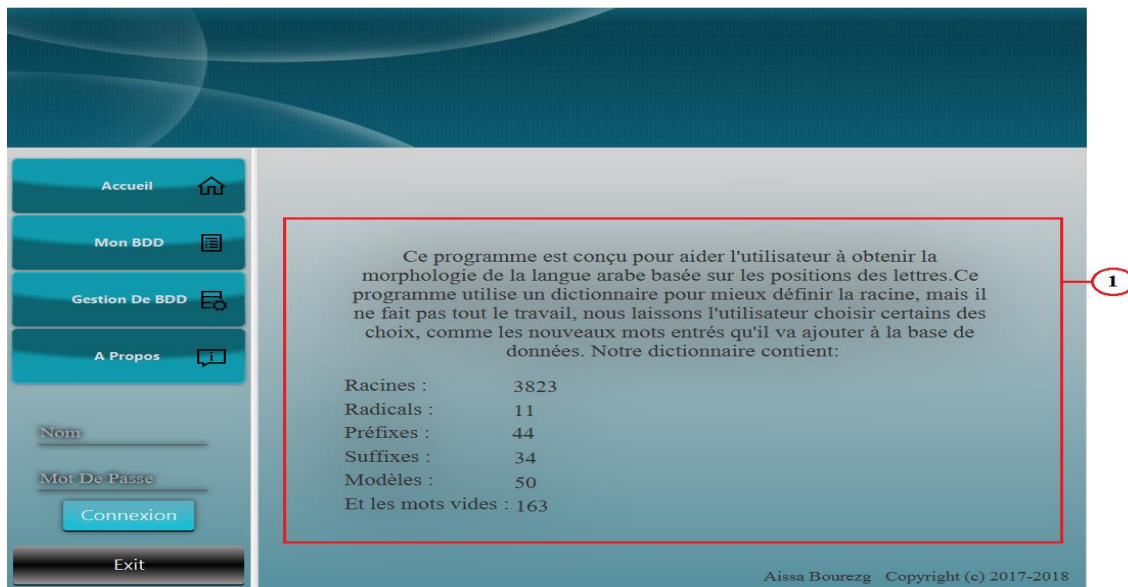


Figure 3.26 : Interface de notre Application «A Propos»

1. Statistique sur notre Application

3 Les outils utilisés pour l'implémentation

3.1 L'environnement de programmation:

3.1.1 NetBeans IDE 8.2:



Figure 3.27 : NetBeans

NetBeans IDE est un environnement de développement intégré gratuit, open source (IDE) qui vous permet de développer des applications bureautiques, mobiles et Web. L'IDE prend en charge le développement d'applications dans diverses langues, y compris Java, HTML5, PHP et C++. L'IDE offre un support intégré pour le cycle de développement complet, de la création de projet au débogage, au profilage et au déploiement. L'IDE s'exécute sous Windows, Linux, Mac OS X et d'autres systèmes basés sur UNIX.

L'IDE fournit un support complet pour les technologies JDK 7 et les améliorations Java les plus récentes. C'est le premier IDE qui prend en charge JDK 7, Java EE 7 et JavaFX 2. L'IDE prend pleinement en charge Java EE en utilisant les dernières normes pour Java, XML, Web services et SQL et prend pleinement en charge le GlassFish Server, l'implémentation de référence De Java EE.Scene Builder [39].

3.1.2 Scene Builder



Figure 3.28 : Scene Builder

Scene Builder est gratuit et open source, mais il est soutenu par Gluon .Il est un outil de mise en page visuelle qui permet aux utilisateurs de concevoir rapidement les interfaces utilisateur de l'application JavaFX, sans codage. Les utilisateurs peuvent faire glisser et déposer des composants UI dans une zone de travail, modifier leurs propriétés, appliquer des feuilles de style, et le code FXML pour la mise en page qu'ils créent est automatiquement généré en arrière-plan (the background). Le résultat est un fichier FXML qui peut ensuite être combiné avec un projet Java en liant l'interface utilisateur (UI) à la logique de l'application. [40]

3.1.3 Serveur de base de données

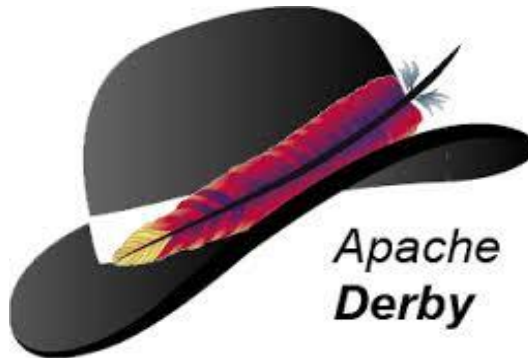


Figure 3.29 : Bases de données Derby

Est une base de données relationnelle open source intégrée entièrement en Java et disponible sous Licence Apache, version 2.0. Voici quelques avantages importants:[38]

- ✓ Derby a une petite empreinte (Small footprint) - environ 2,6 méga-octets pour le moteur de base et le pilote JDBC intégré.
- ✓ Derby est basé sur les normes Java, JDBC et SQL.

- ✓ Derby fournit un pilote JDBC intégré qui vous permet d'intégrer Derby dans n'importe quelle solution basée sur Java.
- ✓ Derby prend également en charge le mode client / serveur le plus familier avec le pilote JDBC Derby Network Client et Derby Network Server.
- ✓ Derby est facile à installer, à déployer et à utiliser.

3.2 Langage de programmation

❖ JavaFX

Le langage de programmation JavaFX Script est basé sur la plateforme de développement et de ce fait requiert que JDK 8 (cette version est plus rapide) soit installé sur votre système.

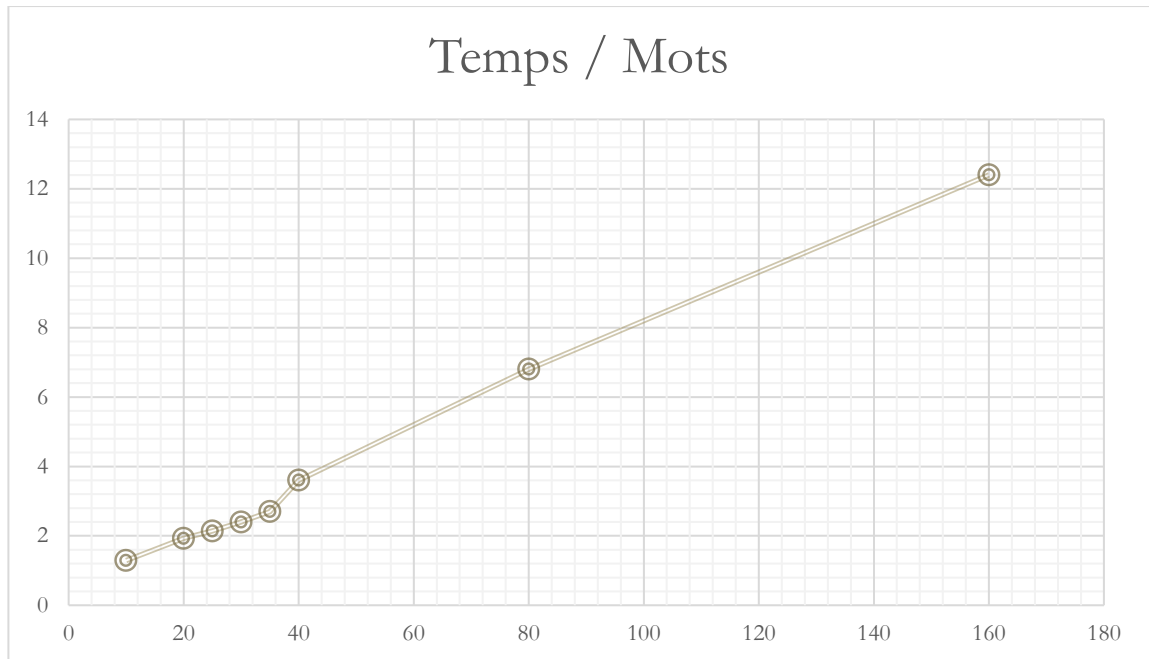
❖ SQL

Le **SQL** (Structured Query Language) est un langage permettant de communiquer avec une base de données. Il est indispensable en informatique pour stocker et lire des données [41].

3.3 Test et Evaluation

Nous avons fait une teste de nombre des mots sur le temps :

Mots	10	20	25	30	35	40	80	160
Temps	1.29	1.93	2.14	2.4	2.7	3.6	6.8	12.4



4 Conclusion

D'après les résultats obtenus, nous pouvons dire que nous avons réussi à bien maîtriser et implémenter la méthode morpho-lexicale basée sur la position des lettres qui a donné de bons résultats avec les données utilisées.

Nous ajoutons une nouvelle chose qui est l'apprentissage de l'algorithme chaque fois que nous l'utilisons, donc, à chaque fois, la base de données devient riche avec de nouveaux mots (racines), que nous pouvons utiliser plus tard, ce qui a rendu nos derniers résultats meilleurs. L'un des plus grands problèmes que nous avons traversé dans notre implémentation est que nous ne pouvons pas distinguer que la lettre appartient au préfixe ou à la racine.

Conclusion générale

La compréhension et l'utilisation de la langue arabe comme moyen de communication à travers le support informatique depuis longtemps avec une grande réticence par la communauté scientifique, en particulier dans le monde arabe, où cet outil trouvera de nombreuses utilisations importantes. En effet, la langue et les difficultés qui y sont associés, y compris le problème de l'incertitude après l'absence de voyelles, ont commencé le problème de l'identification des motifs et le problème du manque de diversité des sujets qui traitent dans la lutte contre la morphologie arabe est limitée à seulement une partie de celui-ci, tout cela est un grand défi difficile à surmonter.

Malgré tout cela, Nous avons essayé, à travers ce mémoire, de présenter brièvement les différentes approches de racinisation de la langue arabe ainsi que la nôtre. La technique que nous proposons se base positionnement des lettres dans le processus d'extraction de racines, dans le but de minimiser le taux d'erreur et d'ambiguïté, dû généralement à l'enlèvement des affixes.

Et comme perspective à ce travail:

- L'amélioration de la modélisation de ces méthodes est possible au niveau de l'extraction des modèles, racines ...
- Essayez de trouver la racine d'un mot sans même avoir besoin d'une table de racines

Bibliographie

- [1]: Abd El Salam AL HAJJAR, "extraction et gestion de l'information a partir des documents arabes", thèse de Doctorat, UNIVERSITE PARIS VIII, SAINT DENIS ,2010.
- [2]: Abu Hawas, F., "Exploit relations between the word letters and their placement in the word for arabic root extraction ", computer Science, vol. 14, no. 2, pp. 327-341, 2013.
- [3]: Adamson, G.W. and Boreham, J., The use of an association measure based on character structure to identify semantically related pairs of words and document titles. Information Storage and Retrieval, pp. 253-260, 1974.
- [4]: Aïda KHEMAKHEM, "ArabicLDB : une base lexicale normalisée pour la langue arabe", Mémoire de master, université de Sfax, Tunisie, 2006.
- [5]: Al Ameen, S., Al Ketbi, A., Al Kaabi, K., Al Shebli, N., Al Shamsi, N. and Al Nuaimi, S., Al Muhairi, Arabic Light Stemmer: A new Enhanced Approach , The Second International Conference on Innovations in Information Technology (IIT'05),pp. 1-9, 2005.
- [6]: Al jlayl, M. and Frieder, O., on arabic search: Improving the retrieval effectiveness via a light stemming approach. In Proceedings of ACM Eleventh Conference on Information and Knowledge Management, Mclean, VA, pp. 340 - 347, 2002.
- [7]: Atef Ben Youssef, "Méthodes Mixtes pour la Traduction Automatique Statistique", Mémoire de master, université Stendhal, Grenoble3, 2008.
- [8]: Beesley, K. R., 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.
- [9]: Belguith, Baccour et Mourad, Segmentation de textes arabes basée sur l'analyse contextuelle des signes de ponctuations et de certaines particules, Laboratoire de recherche LARIS– Faculté des Sciences Economiques et de Gestion, Sfax, Tunisie, TALN 2005,10 juin 2005
- [10] : Belguith, & Chaâben. , Analyse et désambiguïsation morphologiques de textes arabes non voyellés. In Actes des Troisièmes journées scientifiques des jeunes chercheurs en Génie Électrique ET Informatique (GEI'2003).

- [11]: BELGUTH HADRIC L., Traitement des erreurs d'accord de l'arabe basé sur une analyse syntagmatique étendue pour la vérification et une analyse multicritères pour la correction, Thèse de doctorat en informatique, Faculté des Sciences de Tunis. . (1999)
- [12] : BEN HAMADOU A. Vérification et correction automatique par analyse affixale des textes écrits en langage naturel : le cas de l'arabe non voyellé. Thèse d'État en informatique, Faculté des sciences de Tunis. . (1993)
- [13]: 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", thèse de doctorat, Université de Nantes, 2008.
- [14]: Chen A. and Gey F., Building an Arabic stemmer for information retrieval. In Proceedings of TREC, Gaithersburg, Maryland, pp. 631—639, 2002
- [15]: D.E. Kouloughli, Lexique fondamental de l'arabe standard moderne, Paris: le Harmattan, 287 p., (1991).
- [16]: dariouache Adnane "Etude et réalisation d'un analyseur morphologique de la langue arabe" Projet de Fin d'Etudes Master Sciences et Techniques, université de Maroc Sidi Mohamed Ben Abdellah, 2015.
- [17]: Darwish, K., Building a Shallow Arabic Morphological Analyzer in One Day. The ACL-02 Workshop on Computational Approaches to Semitic Languages, Philadelphia, USA, 2002.
- [18]: De Roeck, A.N., and Al-Fares, W., A morphologically sensitive clustering algorithm for identifying Arabic roots. In Proceedings ACL-2000. Hong Kong, pp. 199 – 206, 2000.
- [19] : Ed-dariouache Adnane, "Etude et réalisation d'un analyseur morphologique de la langue arabe", mémoire de master, UNIVERSITE Sidi Mohamed Ben Abdellah, Fès, 2015.
- [20]: Grefenstette, G., Semmar, N., F. and Fluhr, C., 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 ANN ARBOR USA, June 29 2005.
- [21]: Goweder, A., Poesio, M., De Roeck, A. and Reynolds, J., Identifying Broken Plurals in Unvowelised Arabic Text". Proceedings of EMNLP, pp. 246-253, 2003.

- [22]: Jurafsky D., Martin. J. H.: Speech and Language Processing: An Introduction to Speech Recognition. Natural Language Processing, and Computational Linguistics, and Speech Recognition, Prentice-Hall, 2007.
- [23]: Kadri, Y. and Nie, J., Effective Stemming for Arabic Information Retrieval. Proceedings of the Challenge of Arabic for NLP/ MT Conference, Londres, Royaume-Uni, 2006.
- [24]: Kanaan, G., Al-Shalabi, R., Jaarn, J., Al-Kabi, M. and Hasnah, A., A New Stemming Algorithm to Extract Quadri-Literal Arabic Roots, 2004.
- [25]: Khoja, S. and Garside, R., Stemming Arabic text. Computing Department, Lancaster University, Lancaster, www.comp.lancs.ac.uk/computing/users/khoja/stemmer.ps, 1999.
- [26]: Khreisat, L., 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]: LARKEY, BALLESTEROS, & CONNELL. Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis. . (2002).
- [28]: Larkey, L., Ballesteros, L. and Connell, M., Light Stemming for Arabic IR, Arabic Computational Morphology: Knowledge-based and Empirical Methods, A. Souidi, A. Van Bosch, and G. Neumann Editors. Kluwer/Springer's series on Text, Speech, and Language Technology, 2005.
- [29]: M. Rashwan, M. Al-Badrashiny, M. Attia, S.M. Abdou, "A hybrid system for automatic arabic diacritization», The 2nd International Conference on Arabic Language Resources and Tools, Egypt, 2009.
- [30] : Mesfar, S, "Analyse morpho-syntaxique automatique et reconnaissance des entites nommees en arabe standard". Thèse de doctorat, université de Franche-Comte, 2008.
- [31] : Mohsen Maraoui, Mounir Zrigui, Georges Antoniadis, " Un système de génération automatique de dictionnaires étiquetés de l'arabe", CITALA 2007, Rabat, Maroc.
- [32] : S. Khoja, R. Garside, G. Knowles, A tagset for the morpho-syntactic tagging of arabic, Actes de la conférence internationale Corpus Linguistics 2001, Lancaster, (2001).

- [33] : Sanan, M., " Etude Des Methodes 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.
- [34]: Sanan, M., Rammal, M. and Zreik, K., Arabic documents classification using N-gram, Conférence ICHSL6, Toulouse, 2008.
- [35]: Semmar, N., Elkateb-Gara, F. and Fluhr, C. 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.
- [36]: Taghva, K. Elkoury, R. and Coombs J., "Arabic Stemming without a root dictionary", International Conference on Information Technology: Coding and Computing (ITCC'05) - Volume I pp. 152-157, 2005.
- [37] : Y. Kadri, A.Benyamina, "Système d'analyse syntaxicosemantique du langage arabe", mémoire d'ingénieur, université d'OranEs-sénia, 1992.

Webographie

- [38]: Apache, <https://db.apache.org/derby/>. Consulté le : 15/05/2017.
- [39]: Oracle, https://docs.oracle.com/cd/E40938_01/doc.74/e40142/gs_nbeans.htm. Consulté le : 18 /05/2017.
- [40]: Oracle, <http://www.oracle.com/technetwork/java/javase/downloads/javafxscenebuilder-info-2157684.html>. Consulté le : 15/05/2017.
- [41]: SQL, <http://sql.sh/2143-cours-sql-pdf> . Consulté le: 22/05/2017.

ملخص

بواسطة ثروتها الصرفية والنحوية، تعتبر العربية بين اللغات الأكثر صعوبة للتعامل في مجال المعلومات للبحث. ويرجع ذلك إلى الصعوبات التي تمت مواجهتها خلال الجذعية، والتي ليس لديها نهج موحد. أعمالنا هي على هذه الأبحاث. نحن نحاول إعداد نهج جديد لاستخراج الجذر (الجذعية) للغة العربية بطرق فعالة، وذلك استنادا بمواقع الحروف.

الكلمات المفتاحية: الجذعية، اللغة العربية، استرجاع المعلومات، المعالجة التلقائية للغة العربية.

Résumé

Par sa richesse morphologique et syntaxique, la langue arabe est considérée parmi les langues les plus difficiles à traiter dans le domaine de recherche d'information. Cela est dû, notamment, aux diverses difficultés rencontrées dans sa racinisation, qui n'a pas encore connu une approche standard. Nos travaux se situent sur cet axe de recherche. On essaye de développer une nouvelle approche de racinisation de la langue arabe simple et efficace, et il est basé sur la position des lettres.

Mots-clés: Racinisation, la langue arabe, extraction de l'information, traitement automatique de la langue arabe,

Abstract

By its morphological and syntactic richness, the Arabic language is considered as one of the most difficult languages to deal within the search information domain. This is due to the difficulties met during stemming which doesn't have a standard approach. Our work is about this search area. We try to develop a simple and efficient stemming approach of the Arabic language, and it's based on position of the letters.

Keywords: Stemming, Arabic language, information retrieval, automatic treatment of Arabic language.