

Le traitement automatique du langage naturel est une discipline très importante qui ne cesse, depuis plusieurs années, de se développer et de gagner de plus en plus de terrain dans le domaine des différentes langues à travers le monde, bien que ce domaine soit 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. La majorité des travaux connus menés jusqu'ici concernaient les langues occidentales telles que l'anglais et le français. Le traitement de la langue arabe n'a vu le jour que quelques années plus tard après les premiers travaux. En effet, plusieurs tentatives dès lors, 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 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 "**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 de "**l'analyse morphologique de la langue arabe**", deuxième chapitre de ce travail, 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.

L'arabe (al arabiya en transcription traditionnelle) est la langue parlée à l'origine par les Arabes. C'est une langue orientale, sémitique (comme l'akkadien et l'hébreu), très structurée, dérivationnelle, flexionnelle, parlée par plus de 330 millions de locuteurs répartis sur 22 pays arabes et utilisée par plus de 1.5 milliards de musulmans et dont l'alphabet est utilisé dans plusieurs langues telles que le Persien, le malaisien et l'Urdu [35]. Elle est considérée comme la 5^{ème} langue au monde et figure parmi les six langues officielles des Nations Unies. Elle doit sa formidable expansion à partir du 7^{ème} siècle grâce à la diffusion du Coran et la propagation de l'islam qui a atteint toute l'Afrique du Nord et l'Asie mineure.

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 :

1.1. L'alphabet:

Il est constituée de 28 lettres (25 consonnes et 3 voyelles longues-ا، و، ي- a, w, y-) appelées حروف الهجاء Hufof alheaja.

Parmi ces lettres, 3 s'écrivent de différentes manières et posent certains problèmes pour le traitement de la langue. Ces lettres sont :

Hamza[ء]	أ	Exemple : أمل (espoir)
	إ	Exemple : جرى (téméraire)
	ؤ	Exemple : سؤال (question)
	ء	Exemple : ماء (eau)
	ئ	Exemple : دائم (permanent)
Ta marbuta[ة]	Utilisée à la fin des mots seulement. Exemple : سمية (somaia)	
Alif magsurah[ى]	Exemple : صدى (écho)	

1.2. L'écriture :

La langue arabe s'écrit et se lit de droite à gauche avec des lettres monocamérales (pas de minuscule et de majuscule) qui changent de forme de présentation selon leur position (au début, au milieu ou à la fin du mot). Le Table 1.2 montre les variations de la lettre ع (Ayn).

Toutes les lettres se lient entre elles sauf (ذ، د، ر، ز، و، لا) qui ne se joignent pas à gauche. [18].

Arabes	Français	Tra.	Arabes	Français	Tra.
أ	A	Alif	ض	D	Dad
ب	B	Ba'	ط	t	Tah
ت	T	Ta'	ظ	Z	Zah
ث	Th	Tha'	ع	‘ ‘	Ayn
ج	J	Jim	غ	Gh	Ghayn
ح	H	Hha'	ف	f	Fa
خ	Kh	Kha'	ق	Q	Qaf
د	D	Dal	ك	K	Kaf
ذ	D	Thal	ل	L	Lam
ر	R	Ra	م	m	Mim
ز	Z	Zayn	ن	N	Nun
س	S	Sim	ه	h	Ha
ش	sh	Shin	و	w	Waw
ص	S	Sad	ي	Y	Ya

Table1.1. : les 28 lettres arabes.

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

Table 1.2. : Exemple de variation de la lettre ع ayn

1.3. Voyellation :

Une unité lexicale arabe s'écrit avec des consonnes et des voyelles. Ces dernières sont différentes des consonnes et sont rarement notées pour lever les ambiguïtés et aider à la lecture et à la compréhension correcte d'un texte comme dans le Coran ou les livres didactiques. Elles permettent également de différencier les unités lexicales ayant la même représentation.

Deux types de voyelles existent en langue arabe :

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

Exemples :

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

Les voyelles courtes (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
—	فَتْحَة /fathatun/	A
—	كَسْرَة /kasratun/	I
—	ضَمَّة /dammatun/	U
—	سُكُون /sukûnun/	-

Table 1.3. : Les voyelles brèves

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

- كَتَبَ , il a écrit
- كُتِبَ , il a été écrit
- كُتُب , Livres

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

Table 1.4. : Ambiguïté causée par l'absence de voyelles
pour les unités lexicales كَتَب et مَدْرَسَة [12]

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 et qui consiste aussi à remplacer quelques lettres par d'autres, selon des règles prédéfinies.

✓ **La shadda :**

C'est le signe 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 : مدرّس a enseigné, درس a étudié.

✓ **Le tanwin:**

C'est un signe qui est utilisé 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.

اَ	[an]
اَوَّ	[on]
اِ	[in]

Table 1.5. : Les diacritiques doubles

2. Morphologie du mot arabe :

2.1. Structure du mot arabe :

En langue arabe, les mots sont généralement formés d'une agglutination de morphèmes lexicaux et grammaticaux. En effet, un mot peut représenter toute une phrase qu'on peut schématiser globalement selon le modèle suivant:

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

Table1.6. : Structure du mot arabe.

Où:

- ✓ Les proclitiques peuvent être 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 : soit le mot "اَتَتَعَلَّمُونَهَا" qui veut dire en français : « l'apprenez vous ? » Après segmentation, on obtient les éléments suivants :

Enclitique	Suffixe	Corps schématique	Préfixe	Proclitique
ها	ونـ	تعلم	تـ	أ
pronom suffixe complément du nom	suffixe verbal exprimant le pluriel	dérivé de la racine: علم	préfixe verbal du temps de l'inaccompli.	conjonction d'interrogation

2.2. Catégories du mot arabe :

Le lexique arabe comprend trois catégories d'unités : verbes, noms et particules.

2.2.1. Le verbe :

C'est un mot qui exprime un sens dépendant du temps, et auquel se rattachent directement ou indirectement les divers mots qui constituent une phrase.

Les traits flexionnels des verbes en arabe expriment les catégories suivantes :

- ✓ le temps : l'accompli (correspond au passé en français), l'inaccompli (correspond au présent en français),
 - ✓ le nombre du sujet (singulier, duel, pluriel),
 - ✓ le genre du sujet (masculin, féminin),
 - ✓ la personne (première, deuxième et troisième),
 - ✓ le mode (actif, passif).
 - ✓ La langue arabe dispose de trois temps. [11]
- L'accompli : correspond au passé et se distingue par des suffixes (par exemple pour le pluriel féminin on a كتبن KaTaBna, *elles ont écrit* et pour le pluriel masculin on a كتبوا KaTaBuu, *ils ont écrit*).
 - L'inaccompli présent: présente l'action en cours d'accomplissement, ses éléments sont préfixés (يكتب yaKTuBu *il écrit*; تكتب taKTuBu, *elle écrit*).
 - L'inaccompli futur : correspond à une action qui se déroulera au futur et est marqué par l'antéposition de س sa ou سوف sawfa au verbe (سيكتب sayaKTuBu *il écrira*, سوف يكتب sawfa yaKTuBu *il va écrire*). [18]

Comme nous l'avons vu, les verbes se divisent en deux catégories :

a. Les verbes sains (الأفعال الصحيحة) : Sont les verbes dont les lettres radicales ne contiennent pas de lettres défectueuses « alif-ا », « yā'-ي », « wāw-و ».

Exemple : جلس، حضر، ركب

Les verbes sains se répartissent en trois catégories : le verbe sain simple, le verbe sain hamzé « mahmouz-مهموز » et le verbe sain redoublé « muḍaaḡ-مضاعف » :

- Dans les verbes sains simples, les lettres radicales ne contiennent ni de « hamza » ni de redoublement de consonne. Exemple : سمع، شرب...

- dans les verbes sains hamzés « *mahmouz- مهموز* », l'une des lettres radicales contient une « *hamza* », au début, au milieu ou à la fin du mot.

Exemple : سَأَلَ أَخَذَ، أَمَرَ، أَذِنَ،

- les verbes sains redoublés « *mudaaḥ- المضاعف* », se divisent eux-mêmes en deux catégories :

- les verbes dont la deuxième et la troisième lettre sont identiques.

Exemple : مَدَّ، عَدَّ، سَدَّ،

- les verbes dont la première et la troisième lettre sont identiques ou dont la deuxième et la quatrième lettre sont identiques.

Exemple : زَلَزَلَ، وَسَّوَسَ،

b. Les verbes défectueux (الأفعال المعتلة) :

Sont les verbes dont l'une de ses lettres radicales est une lettre défectueuse « *alif-l* », « *yā'-ي* », « *wāw-و* ». Exemple : وَهَبَ، مَالَ، وَعَى،

Les verbes défectueux se divisent en quatre catégories :

- Les verbes assimilés « *mitā'al- مثال* » dont la première lettre est défectueuse.

Exemple : وَعَدَ، وَلَدَ، وَجَدَ،

- Les verbes creux ou concaves « *aḡwaf- اجوف* » dont la deuxième lettre est défectueuse. Exemple : قَالَ، باع، نام،

- Les verbes défectueux « *naqess- ناقص* » dont la dernière lettre est défectueuse.

Exemple : رَمَى، سَعَى،

- Le verbe « *al laḥif- اللّيف* » est un verbe qui contient deux lettres défectueuses ; il se divise également en deux catégories :

- le verbe « *laḥif maqroun- اللّيف المقرون* » contient deux lettres défectueuses successives. Exemple : أَوَى، كَوَى،

- le verbe « *laḥif mafrouq- اللّيف المفروق* » contient deux lettres défectueuses non successives, autrement dit, une lettre saine sépare deux lettres défectueuses.

Exemple : [13]. وَشَى، وَعَى،

2.2.2. Le nom :

Est un concept qui désigne un être, un objet ou un état exprimant un sens indépendant du temps. Les noms arabes regroupent les substantifs, les adjectifs et les pronoms, ainsi que d'autres noms invariables [34] .

Ils sont classés selon deux types :

- les noms qui sont dérivés d'une racine verbale.
- les noms qui ne le sont pas, comme les noms propres et les noms communs.

La déclinaison des noms suit les règles suivantes:

- Le féminin singulier : on ajoute la lettre ة à la fin du mot. Exemple قليل *peu* devient قليلة.
- Le féminin pluriel : de la même manière, on rajoute pour le pluriel les deux lettres ات. Exemple : طالبات *étudiantes*.
- Le masculin pluriel : on rajoute les deux lettres ين ou ون en fonction de la position du mot dans la phrase. Exemple عاملون *travailleurs*
- Le pluriel irrégulier : il suit une diversité de règles complexes et dépend du nom. Exemple طلاب *étudiants*

Le phénomène du pluriel irrégulier dans l'arabe pose un défi à la morphologie, non seulement à cause de sa nature non concaténative , mais aussi parce que son analyse dépend fortement de la structure comme pour les verbes irréguliers. [19]

Certains dérivés nominaux associent une fonction au nom :

- Agent (celui qui fait l'action),
- Objet (celui qui a subi l'action),
- Instrument (désignant l'instrument de l'action),
- Lieu.

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 كَبْشُ [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]. table1.7.

- Celle qui regroupe les noms non conjugables et qui gardent la forme quelque soit le contexte. table 1.8.

Non conjugable		Conjugable	
Sous-catégorie	Exemples :	Sous-catégorie	Exemples :
Adverbe	حَيْثُ أَيْنَ، قَبْلَ،	Pronom relatif	الَّذان، اللواتي
Nom de voix	نَحْ أَخْ،	Nom de nombre	خمسة، عشرون، مئة
Nom de verbe	حيّ أَوْه، شَتَّانَ،	Pronom démonstratif	هذان، هؤلاء
Pronom Personnel (affixé ou isolé)	هما انت هن،	Nom propre	زبيدة، جمال، المسيلة،
Pronom interrogatif	مَا مَتَّى، كيف		
Pronom conditionnel	إِذَا مَنْ،	Nom commun	بيت بنت، قط،
Pronom allusif	كأي، كم		

Table 1.7 : Dérivationnel irrégulier

Conjugable			
Sous-catégorie	Exemples	Sous-catégorie	Exemples
Masdar	الانطلاق، الذهاب	Nom d'instrument	مفتاح فاس،
Participe actif	مجاهد، منطلق	Adjectif	بطل عظيم، حسن،
Participe passif	مكرم، مسموح	Elatif	اكبر فضلى،
Nom d'une fois	صُرْبَةً جَلَسَةً،	Nom diminutif	شجيرة قطييط
Nom de manière	خلسة نظرة،	Nom de relation	جزائري حضرمي،
Nom de temps	صباح، مشرق	Intensif	مغوار غفار،
Nom de lieu	مزرعة هنا،		

Table 1.8. : Dérivationnel régulier [6]

2.2.3. Les particules :

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 : ...، في، الى، من، على، ب، ...

Elles sont très utiles pour le traitement pour deux raisons :

- Elles font partie de l'antidictionnaire [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 [9].

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 [38].

On peut distinguer plusieurs types :

- Préposition : exemple (من ، الى، على، عَنْ، حَتَّى،)
- Particules de coordination : exemple (و، أَوْ ثُمَّ، ف، و)
- Particules interrogatives : exemple (أَ هَلْ، مَا،)
- Particules d'affirmation : exemple (أَجَلْ بَلَى، نَعَمْ،)
- Particules de négation : exemple (لَمْ لَنْ، لَا،)
- Particules distinctive : exemple (أَيُّ)
- Particules relatives : exemple (مَا)
- Particules de futur : exemple (سَ أَنْ، لَنْ،)
- Particules conditionnelles : exemple (كَيْفَمَا، مِنْ، إِذَا). [6]

2.3. 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). [3]

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

2.3.1. La racine :

Les racines sont à l'origine de la plupart des mots arabes. Ce sont des verbes formés de trois à cinq consonnes [28]. Elles sont aux alentours de 10000 racines dont la grande majorité (85%) sont trilitères. Les restes sont des racines quadrilitères ou quintilitères. Une racine

définit la signification fondamentale des mots dérivés en utilisant différentes diacritiques et affixes avec les lettres de la racine pour créer l'inflexion de la signification. [21].

Par exemple, la racine <درس, darasa, il a étudié>. A partir de cette racine, on peut former plusieurs mots sous plusieurs formes (présent, imparfait, futur simple, passé simple, impératif, etc.). Il y a aussi des formes supplémentaires telles que les noms verbaux (Table 1.9). [3]

Racine : <درس, darasa, a étudié>						
Conjugué	درس	Drs	Il a étudié	يدرس	Yidrs	Il étudie
	درسنا	Drsna	Nous avons étudié	يدرسون	Yidrswn	Ils étudient
	درست	Drst	Elle a étudié	تدرس	Tdrs	Tu étudies
	تدرسون	Tdrswun	Vous étudiez	ندرس	Ndrs	Nous étudions
Noms	دارس	Dars	Étudiant	تدريس	Tdris	Enseignement
	مدرسة	Mdras	École	مدروس	Mdrwus	Étudié
	درس	Drs	Cours	دراسة	Dirsh	Etude

Table1.9. : Quelques mots dérivés de la racine درس (a étudié)

2.3.2. le schème :

Un schème représente une forme ou modèle général composé de trois consonnes ف [f], ع [ʿ] 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 (voir Figure1.1). [15]

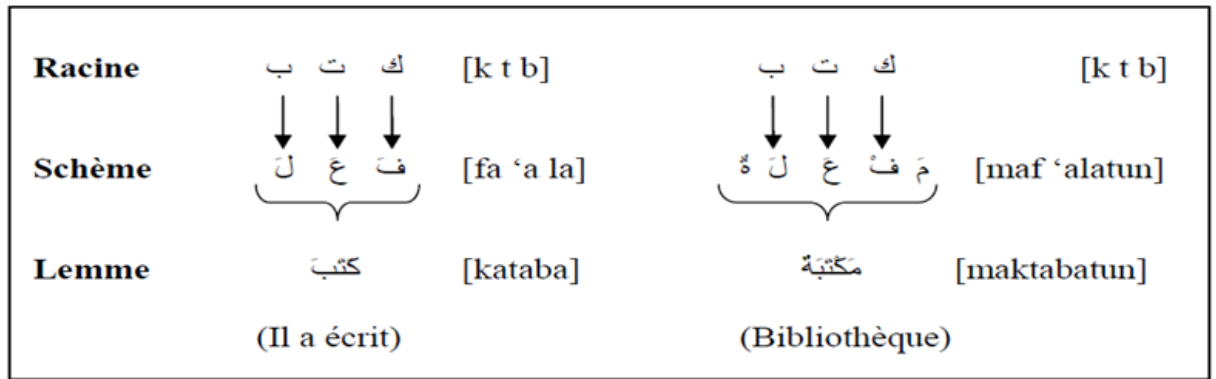


Figure1.1. : Exemples de dérivation de la racine [k t b]

Dans cette figure, le lemme est formé par le remplacement respectif des consonnes du schème par les consonnes de la racine ك [k], ت [t], ب [b], tout en gardant les autres composants du schème.

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. Par exemple, à partir de la racine ك ت ب [k t b] on peut générer le verbe كَتَبَ [kataba] par le schème verbal فَعَلَ [fa'ala] et le nom مَكْتَبَةٌ [maktabatun] par le schème nominal مَفْعَلَةٌ [maf'alatun] (voir Figure 1.1).

2.3.3. 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 [37].

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 Table 1.10 présente des exemples de chaque type de préfixes [30]

Type	Les préfixes			
	Nom en français	Signification	Nom en arabe	Transcription
Préfixes nominaux	L'article de définition	Le	ال	Al (Lam altaarif)
	Les prépositions	Avec	ب	b
		Pour	ل	l
		Comme	ك	k

Préfixes verbaux	La particule du futur	Sera	س	s
	Les particules du subjonctif	Pour	ل	l

Préfixes généraux	Les conjonctions des coordinations	Et	ف	f
		Et	و	w
	L'article d'interrogation	Est-ce-que	أ	a

Table 1.10 : Un exemple des préfixes

Les préfixes peuvent s'enchaîner dans un mot pour former des préfixes composés qui peuvent atteindre jusqu'à cinq lettres (< اقبال, afabil, 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 (Table 1.10).

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 (Table 1.11). 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].

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 1.11 : Un exemple des suffixes divisés selon leurs types

c. 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, on y trouve : <مدارس, écoles>, il est obtenu à partir de la racine <درس, il a étudié> selon le modèle <مفاعل, mfaâl>. Les stems produits ne sont pas tous utilisables (Table 1.12). [3]

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, étudiant>	Oui
<أكل, akl, il a mangé>	<مفعول, mfawul>	<مأكول, makwul, mangeable>	Oui
<لعب, lab, il a joué>	<أفعلاء, afala'>	<العباء, alaba'>	Non

Table 1.12 : Un exemple de génération des Stems.

2.4. Morphologie dérivationnelle et flexionnelle :

Au niveau de la langue arabe, deux types de morphologie peuvent être constatés, la morphologie dérivationnelle et la morphologie flexionnelle.

2.4.1. La morphologie dérivationnelle :

C'est une branche de la morphologie qui consiste à construire de nouvelles primitives morphologiques à partir de celles qui existent selon des règles de dérivation adéquates. Tous verbe possède des formes dérivées qui lui sont associées et avec lesquelles il dispose de relations morphologiques, syntaxiques et sémantiques. Le nombre et la nature de ces formes varient selon le statut du verbe.

a. Dérivation verbale :

La dérivation verbale montre un mouvement progressif allant des formes simples (avec seulement trois consonnes) aux formes rendues de plus en plus complexes, par le redoublement des consonnes radicales, l'allongement de certaines voyelles brèves et l'ajout de certaines consonnes. Nous présentons ces formes par l'arbre suivant:

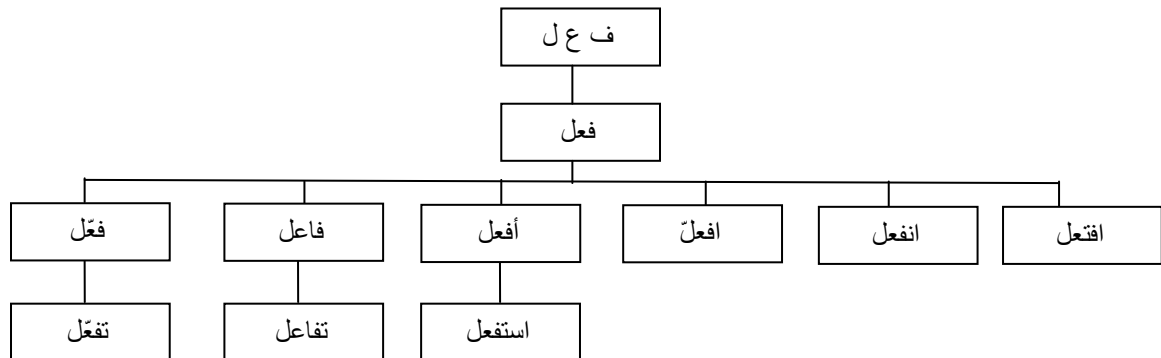


Figure 1.2 : Liste des schèmes verbaux

Il existe encore d'autres schèmes qui sont rarement utilisés, exemple : اَفْعَوْلْ، اَفْعَوْلْ، اَفْعَالْ. Les schèmes augmentent le sens de la racine d'un aspect causatif, actif, pronominal ou réfléchi, donc ces modèles (mots) expriment la même nuance de sens par rapport à la racine.[31]

b. Dérivation nominale :

Les dérivées nominales sont obtenues par l'ajout des préfixes aux formes verbales précédemment décrites pour obtenir les trois types de noms les plus intéressants, à savoir : le participe actif (اسم الفاعل), le participe passif (اسم المفعول) et le nom verbal (مصدر). Il existe à peu près 150 schèmes nominaux, exemple: "متفاعلون", "متفاعلين", "متفاعل", "متفاعلة", "متفاعلات", [31]

2.4.2. Morphologie flexionnelle :

Du point de vue flexionnel, La langue arabe emploie, principalement, pour la déclinaison des noms et la conjugaison des verbes, des indices d'aspect, de temps, de mode, de personne, de nombre, de cas, de mode, etc., qui sont présents généralement, sous forme de préfixes et suffixes.

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

2.5. mots dérivés :

Mot dérivé								
أَتَطْلِبُونَ			Attlbwun			Est-ce que vous demandez ?		
Préfixe			Stem			Suffixe		
أ	A	Est-ce-que	تطلب	tlb	Tu demande	ون	wun	ez
			Racine					
			طلب	tlb	demander			
Phrase								
Est-ce que vous demandez ?								

Table 1.13. : Un exemple de formation de mot

<أَتَطْلِبُونَ, attlbwun, Est-ce que vous demandez ?>

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 de racines. Ainsi, les mots qui dérivent d'une même racine ont des significations similaires. En effet, certains mots dérivés peuvent avoir la signification d'une phrase entière, comme c'est le cas du nom <اعلمون, Est-ce que vous savez?> (Table 1.13).

Il y a deux catégories de mots dérivés : Les verbes et les noms. Dans la plupart des cas, un verbe représente une racine.

Prenons l'exemple du mot <يعلمن, elles savent> qui dérive de la racine <علم, il a su> par l'ajout du préfixe <ي, yi> et du suffixe <ن, n>. Dans ce cas, le temps est présent, le nombre est pluriel, le mode est actif, le genre est féminin et la personne est troisième.

Dans le cas des noms, la dérivation est utilisée pour indiquer le genre, le nombre, etc. Par exemple, le féminin singulier nécessite d'ajouter le suffixe <ة, tah> comme <مدرسة, école>, mais le féminin pluriel nécessite d'ajouter le suffixe <ات, at> comme <مكتبات, mktbat, des librairies>. En général, le pluriel se fait en ajoutant quelques suffixes comme (<ات, at>, <ون, wun>, <ين, yin>, etc.). Par contre, il y a des mots qui ont des règles de composition plus complexes, comme le cas des pluriels irréguliers, comme le mot <اشجار, achjar, arbres> qui est le pluriel du mot <شجرة, chjrah, arbre> [30].

2.6. 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 : Nour, 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.
- La particule est un mot court qui ne représente qu'une expression grammaticale, comme les conjonctions de coordination et de subordination, des prépositions et des adverbes.

On peut constater plusieurs types de particules :

- de temps (<حينما, hhyinma, lorsque>, <بعد, bâd, après>, etc.),
- de lieu (<فوق, fwuq, au dessus>, <عند, ând, chez>, <حيثما, hhyithma, où>, etc.),
- de négation (<بلا, bla, sans>, <دون, dwun, sans>, <ليس, lyis, n'est pas>, etc.).

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 seront exposées au chapitre suivant.

Les recherches sur le traitement automatique de l'arabe ont débuté vers les années 1970. Les premiers travaux concernaient notamment les lexiques et la morphologie. [9]

Par ses propriétés morphologiques et syntaxiques, le traitement automatique doit faire face à :

- la nature agglutinante de la langue : l'ensemble des morphèmes collés à l'unité lexicale véhiculent plusieurs informations morphosyntaxiques.
- la richesse flexionnelle de l'arabe
- l'absence de voyellation de la majorité des textes arabes écrits : ce phénomène entraîne un nombre important d'ambiguïtés morphologiques. En arabe, chaque lettre doit prendre un signe de voyellation et de surcroît. Les voyelles finales sont porteuses de certains traits morphosyntaxiques comme la déclinaison, le mode, le cas. [12]

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

En général, l'analyse morphologique a pour objectif de donner à chaque unité lexicale, en entrée, ses différents traits morphologiques, ainsi que ses différents constituants. Pour ce faire, une suite de traitements doit être effectuée, dont quelques uns sont communs et d'autres spécifiques à la méthode choisie. Les traitements communs du mot sont effectués tout au début, ensuite une vérification consiste à savoir si c'est un mot outil (stopword) ou non grâce à une table regroupant ce type de mots. Finalement, la méthode choisie pour extraire la racine est appliquée. [3]

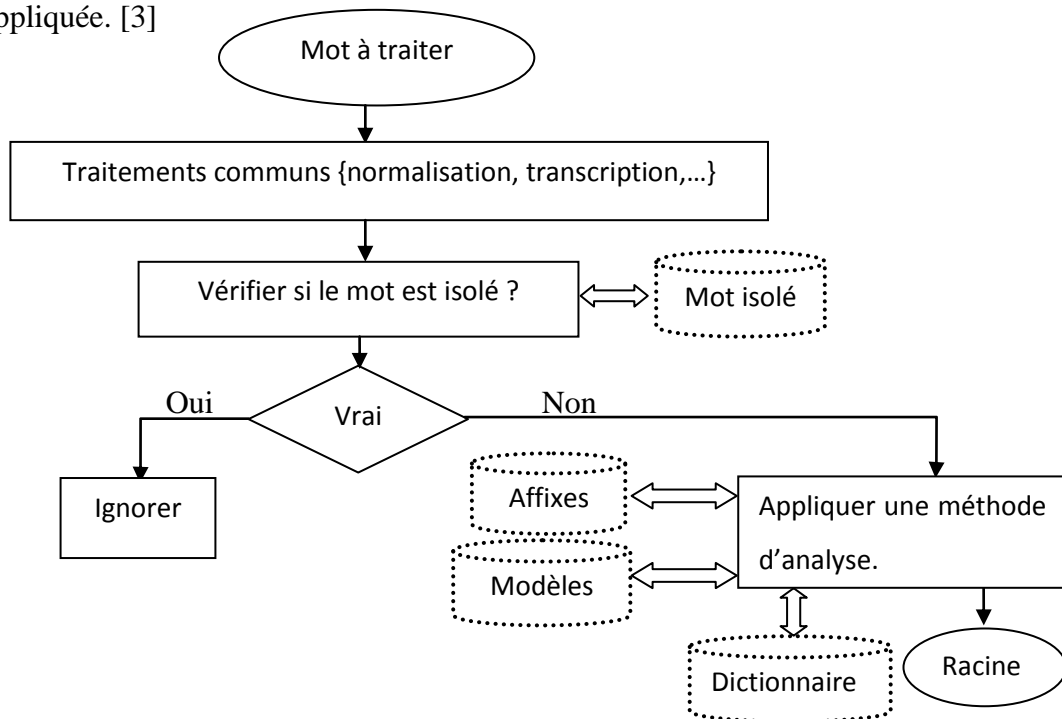


Figure2.1. : Processus d'extraction de la racine d'un mot

2. Traitements communs :

Ce sont les traitements qui sont appliqués en premier lieu pour la plupart des méthodes afin de préparer le mot pour l'analyse. Le premier est la spécification du codage utilisé. Ensuite une opération de normalisation est effectuée avec, éventuellement, une phase de transcription utilisée par peu de méthodes.

2.1. Normalisation :

La phase de normalisation consiste à prendre en entrée un mot et essayer de supprimer ou remplacer quelques lettres selon des règles prédéfinies. Chaque méthode a ses propres règles.

En général, un mot est normalisé :

- ✓ en supprimant : le symbole du tatweel « ~ », les diacritiques et le symbole de gémation, la ponctuation, les caractères spéciaux, les nombres...
- ✓ en remplaçant :
 - أ, إ et آ par Alif barre ا.
 - ى par ي à la fin du mot
 - ة par ة à la fin du mot....

Toutefois, cette normalisation peut entraîner beaucoup d'ambiguïtés étant donné que deux mots peuvent avoir la même forme normalisée, tout en ayant des sens différents.

2.2. Transcription :

La transcription d'un mot arabe, dans une autre langue cible (comme le français), consiste en l'écriture phonétique et symbolique de ce mot à l'aide des caractères et des chiffres latins de la langue cible. Par exemple, le mot <كتاب, ktab, livre> est transcrit en «ketab» en anglais, et en «ktab» en français. Elle peut se faire pour deux principales causes : l'absence des caractères arabes sur le clavier ou leur absence dans le site ou le service.

L'usage de l'arabe translittéré tend malgré tout à diminuer. La transcription est également utile pour résoudre le problème des signes diacritiques. Par exemple, considérons le mot <ملك, mlk> qui est écrit sans signes diacritiques. Ce mot pourrait avoir trois significations différentes (roi, royaume, ange) selon sa prononciation (مَلِك, مُلْك, مَلَك). A chaque signification et prononciation correspond une transcription distincte (malk, molk, malak). Un autre avantage de la transcription est qu'il est possible d'éviter l'élimination des suffixes et des préfixes du mot traité.

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

La langue arabe est une langue tellement complexe et d'une morphologie tellement variée que plusieurs méthodes d'analyse du mot arabe ont vu le jour durant ces dernières années.

Toutes ces méthodes sont plus ou moins efficaces selon l'ensemble des mots traités. Ces différentes méthodes peuvent être classées en cinq grandes classes :

- ❖ Méthodes basées sur l'analyse morphologique.
- ❖ Les méthodes basées sur les tables de correspondance (génération systématique).
- ❖ Les méthodes basées sur la transcription ou la traduction.
- ❖ Les méthodes basées sur l'analyse statistique.
- ❖ Les méthodes basées sur l'analyse hybride.

3.1. Les méthodes basées sur l'analyse morphologique :

3.1.1. Méthode basée sur les affixes :

Parmi ces méthodes, celle du lemmatiseur léger proposée initialement par Aljlayl et Frider [27], cette méthode est basée sur l'élimination des affixes pour tous les mots qui sont formés de plus de trois lettres. Plusieurs variétés de lemmatiseurs légers ont été développées. On peut citer :

Al-Stem qui est un lemmatiseur développé par Maryland et modifié par Larkey et al. [26], basé sur l'élimination du و s'il est au début du mot, des préfixes (ال،ل،فال،كال،بال) et les suffixes (ة،ة،ية،يه،ين،ون،ات،ان،ها،ي).

H. Al Ameen et al [20], ont par la suite, proposé un nouvel ensemble de préfixes (و،بال،فال،وال،كال،سي،وس،كأب،ل،و) et les suffixes (ا،ي،ة،ه، etc.)

Dans un autre travail, Chen et Gey [1] ont à leur tour, proposé d'autres ensembles d'affixes. ils ont recensé les mots qui commencent par un préfixe donné (par exemple, pour les préfixes و, ils ont trouvé 11732 mots, 94 043 mots qui commencent par la lettre ا, etc.) et les mots qui se terminent par un suffixe donné (par exemple : 52418 mots se terminent par ي, 19089 se terminent par ان, etc.). finalement, ils ont identifié les préfixes qui doivent être enlevés : 19 de trois lettres, 14 de deux lettres, et 3 d'une lettre, et les suffixes: 18 de deux lettres, 4 d'une lettre. Pour supprimer les préfixes et les suffixes des ensembles prédéfinis, chaque algorithme propose ses propres règles. Par exemple, dans le cas où un mot est composé d'au moins 5 lettres, un préfixe de trois lettres parmi l'ensemble (ال،بال،فال،كال،وال) est supprimé. Si le mot est formée d'au moins 4 lettres, les deux premières lettres parmi l'ensemble (ال،وا،بال،لل،وم،وت،وب،لا،سي،وس،وي،ول،كا،فال) [4]

PR63 : modèles de longueur 6 et d'une racine de longueur 3 (استفعال، مفعالة، افتعال...)

PR64 : modèles de longueur 6 et d'une racine de longueur 4 (افعال، متفعّل....) .

L'extraction de la racine repose sur l'utilisation d'un algorithme basé sur la longueur du lemme normalisé. [4]

3.2. Les méthodes basées sur les tables de correspondance (génération systématique) :

Dans cette approche, le principe est de construire une très grande table servant de dictionnaire global contenant la plupart des mots arabes classés par ordre alphabétique accompagnés, chacun, par sa radicale, sa racine et ses affixes. Ces mots peuvent inclure des mots fonctionnels, les mots étrangers, et les noms propres, où chaque mot utilise une entrée unique dans la table. Pour extraire la racine d'un mot quelconque, il suffit de faire une simple recherche dans la table construite.

Parmi les travaux réalisés dans ce contexte, on peut citer la méthode proposée par Goweder et al.[2], qui consiste à extraire les racines des pluriels irréguliers à partir d'un dictionnaire composé d'environ 3600 stems pluriels irréguliers, construit manuellement et validé par un linguiste.

En 2001, le centre Européen de recherche Xerox a proposé un analyseur morphologique pour l'arabe standard moderne basé sur des dictionnaires et utilisant la technologie d'états finis servant à créer des automates pour les expressions régulières et des formalismes d'équivalence pour réaliser les opérations [23].

Suffixes	Préfixes	Racine	radical	Le mot
	ال	ح ل ل	تحليل	التحليل
			
		ح ل ل	تحليل	تحليل
			
هـ		ح ل ل	تحليل	تحليله
			
	وال	ح ل ل	تحليل	والتحليل
			
ات		ح ل ل	تحليل	وبتحليلات
			

Table 2.1. : Exemple de table de correspondance

Ce type d'analyseurs a l'avantage d'être précis, de permettre l'analyse des mots qui n'ont pas une origine arabe et de fournir la possibilité d'avoir plusieurs entrées pour les mots ambigus. Mais, en contre-partie, il nécessite beaucoup de travail pour construire le dictionnaire de tous les mots arabes et énormément de temps pour chercher un mot.

3.3. Les méthodes basées sur la transcription ou la traduction :

La traduction est une technique utilisée pour appliquer des approches déjà développées pour des langues latines telles que l'anglais sur des langues fortement flexionnelles comme l'Arabe. L'inflexion élevée de quelques langues peut être réduite par une application de cette technique [1] [14].

Cette méthode fût proposée à l'origine pour l'anglais et adaptée au langage arabe par A. Chen et F. Gey [1]. Le principe consiste à traduire le mot arabe à analyser vers l'anglais. Après l'élimination des mots outils qui peuvent surgir, il lui est appliqué une méthode d'extraction de sa racine anglaise. Après obtention de la racine, celle-ci est traduite à son tour vers la langue arabe. Pour ce faire, Chen et Gey ont utilisé le traducteur «Ajeeb Online» [7]

Exemple : le mot دروسهم (leurs leçons), est initialement traduit en anglais vers le mot « their lessons ». le terme « their » étant un mot outil, il est supprimé. En anglais, la racine du mot obtenu est « lesson ». Ce terme retraduit en arabe, donne le mot درس.

Cette technique pose quelques problèmes, tels que :

- Les traductions ambiguës, par exemple, la traduction des deux mots arabes différents tels que (مريض , patient) et (صبور , patient) au même mot français « patient », ainsi que la traduction du mot arabe (ذهب , or) à un mot français incorrect « aller ».
- L'efficacité des algorithmes de lemmatisation avec les langues latines elles-mêmes : Certains lemmatiseurs tels que **Porter** échoue à lemmatiser quelques mots. Par exemple, le mot « children » en anglais (أطفال , enfants) n'est pas lemmatisé par Porter pour engendrer le mot « Child » (طفل , enfant).
- L'échec de choisir le lemme correct : Cette technique choisit le plus court terme parmi un ensemble de termes groupés pour représenter un lemme candidat. Dans certains cas, cette technique rencontre une difficulté pour choisir un lemme candidat. Par exemple, les mots (وثيقة , document) et (وثائق , documents) tous les deux ont la même longueur. [14]

3.4. 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 [16], 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 classifie 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.

3.4. Les méthodes basées sur l'analyse statistique :

L'analyse statistique est une technique indépendante du langage basée sur la structure des mots. Elle consiste à découper les mots ainsi que les racines du dictionnaire en un ensemble de segments de n lettres consécutives chacun, appelés n -grammes. Ensuite, à calculer une mesure d'association entre les n -grammes du mot traité et les n -grammes des racines candidates. Pour essayer de la maximiser ou la minimiser. Selon la nature de cette mesure (appelée également coefficient), deux techniques peuvent être utilisées :

3.4.1. Technique basée sur le *coefficient de similarité* :

Deux mots sont considérés similaires s'ils ont en commun plusieurs n -grammes (généralement on considère des 2-grammes où $n=2$). Pour extraire la racine d'un mot, on calcule le coefficient de similarité entre ce mot et une liste de racines du dictionnaire. Les racines qui ont le coefficient le plus élevé sont considérées comme racines potentielles.

Dans les travaux de [32], la mesure de similitude (S) est calculée selon la formule suivante: $S = 2 \times C / (A + B)$ (4.1)

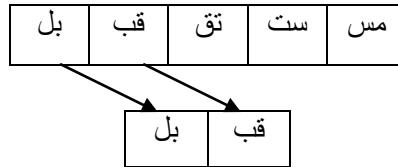
Où :

A : représente le nombre de bi-grammes uniques dans le premier mot.

B : représente le nombre de bi-grammes uniques dans le deuxième mot.

C : représente le nombre de bi-grammes uniques partagés par les deux mots.

Par exemple : prenons le mot « مستقبل » avec la racine « قبل », le mot et la racine sont décomposés en 2-grammes, comme le montre la figure suivante :



Ahmed et Nürnberger ont proposé un modèle n-grammes pour calculer la ressemblance entre deux chaînes de caractères en comptant le nombre des n-grammes semblables qu'ils partagent [17]. Le coefficient de ressemblance est donné par l'équation (1):

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

Où α et β sont les ensembles de n-grammes.

3.4.2. Technique basée sur le *coefficient de dissimilarité* :

Deux mots sont considérés similaires s'ils ont en commun plusieurs sous-chaînes de n-grammes ou s'ils n'ont pas de sous-chaînes différentes). Pour extraire la racine d'un mot, on calcule le coefficient de dissimilarité entre ce mot et une liste de racines du dictionnaire. Les racines qui ont le coefficient le plus bas sont considérées comme racines potentielles.

Parmi les travaux réalisés dans ce domaine, celui de Khreisat expose une autre approche statistique pour classer des documents arabes. Cette approche consiste à calculer une mesure de dissimilarité appelée «Distance Manhattan» et une mesure de dissimilarité appelée «Dice measure». Un corpus de documents de 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 normalisation a été utilisée [25].

Le coefficient de dissimilarité entre les mots a et b est calculé en partitionnant chaque mot en 2-grammes ou bien 3-grammes. $\alpha \cap \beta$ présente l'intersection entre les deux ensembles et $\alpha \cup \beta$ l'union entre les deux ensembles(3) [25].

$$\delta_n(a, b) = \frac{(\alpha \cup \beta) - (\alpha \cap \beta)}{(\alpha \cup \beta)} \quad (2)$$

3.5. Les méthodes basées sur l'analyse hybride :

C'est les méthodes utilisant les deux techniques précédentes (statistique et morphologique). Parmi ces méthodes, celle proposée par De Roeck et Al-Fares [8] où une première étape consiste à utiliser un lemmatiseur léger pour supprimer les affixes afin d'obtenir un stem et une autre pour calculer les coefficients de similarité entre ce stem et une liste de racines sélectionnées dans un dictionnaire. Les racines ayant obtenu un coefficient de similarité élevé sont ajoutées à la liste de racines.

Conclusion :

Dans ce chapitre, nous avons exposé un éventail de méthodes utilisées dans le domaine du traitement de la langue arabe qui diffèrent en efficacité et en vitesse de traitement. 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.

Après avoir pris connaissance, dans les chapitres précédents, du domaine de la langue arabe avec toutes les complexités qu'il englobe, 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 pour enfin l'évaluer en la comparant avec un travail déjà réalisé.

1. Description du système réalisé :

Pour réaliser notre système, notre choix s'est orienté vers la méthode morphologique avec racines et affixes vue précédemment, qui consiste à utiliser des tables pour les différents composants des mots arabes.

1.1. Les tables utilisées :

Notre système comporte cinq tables essentielles :

a. Table des racines :

Cette table 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 initialement pris dans notre cas est de 200, auquel on peut ajouter interactivement d'autres en cas de besoin.

1	les racines
2	بَلَع
3	يَأْس
4	بَغَض
5	أَوَّل
6	يَتَق
7	أَتَر
8	أَتَم
9	أَسَر
10	أَكَل
11	تَكِن
12	تَحَف
13	بَيَع
14	يَعَل
15	يَلد
16	بَلَع
17	بَلِل
18

Table3.1. : table des racines

b. Table des schèmes :

Sert à rassembler les différents schèmes choisis auxquels vont être comparés les radicaux obtenus après l'élimination des affixes pour voir quel est le modèle adéquat à partir duquel on peut obtenir la racine cherchée. Le nombre initial de schèmes de notre table est de 54.

1	Les schèmes
2	فَعَال
3	فَعِيل
4	فَعِل
5	أَفْعَال
6	تَفْعَل
7	فَوَاعِل
8	مَفْعَل
9	تَفْعِيل
10	فَعَالَاء
11	مَفَاعِل
12	تَفَاعِل
13	مُسْتَفْعِل
14	فَاعِلَات
15	فَوَاعِيل
16	مَفَاعِيل
17	تَفَاعِيل
18	إِفْعَال
19	اِسْتَفْعَل
20

Table3.2. : Table des schèmes

c. Table des préfixes :

Contient l'ensemble des 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.

d. Table des suffixes :

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

e. Table des mots outils :

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 de coordination et d'autres particules; en résumé, tous les mots qui restent invariants quelque soit leur contexte.

1	mots outils
2	أدما
3	من
4	كيفما
5	لئن
6	إن
7	ما
8	أي
9	لا
10	لثم
11	لئن
12	أجل
13	نعم
14	بلى
15	ما
16	هل
17

Table3.3. : Table des mots outils

1.2. Le système d'extraction :

1.2.1. Elimination des affixes :

Comme dans la plupart des systèmes d'extraction de racines, une première phase consiste à normaliser certaines lettres des mots arabes comme exposé précédemment (i.e. :...ي،ة،أ)

Ensuite, on passe à la phase d'élimination des affixes (préfixes et suffixes) où il est essentiel de considérer les affixes réels. Par exemple : dans le mot فالحون, si en considère que la chaine فال est un préfixe, on obtient le mot حون auquel on ne trouvera pas de racine étant donné que فال fait partie du mot original à traiter.

Dans notre système, on teste d'abord la 1^{ère} lettre du mot.si elle appartient à l'ensemble des lettres { "ث","ج","ح","خ","ذ","د","ر","ز","ش","ص","ض","ط","ظ","ع","غ","ق","ه" }, alors le mot ne contient pas de préfixe. Puis, on teste la dernière lettre qui, si elle appartient à l'ensemble des lettres { "ل", "س","ب","ث","ج","ح","خ","ذ","د","ر","ز","ش","ص","ض","ط","ظ","ع","غ","ق","ف" }, alors le mot ne contient pas de suffixe. Dans le cas échéant (probabilité de présence d'un

préfixe et/ou suffixe), on procède au principe suivant : on identifie le plus long préfixe (ou suffixe) du mot et on vérifie dans la table s'il fait partie de l'ensemble des préfixes (ou suffixes) sinon, on cherche le préfixe ou suffixe de moindre taille et on reprend le même processus jusqu'à trouver une correspondance.

Exemple : la décomposition du mot "المفلحون" d'après le processus précédent donne :

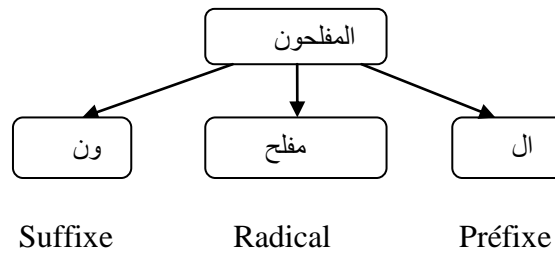


Figure3.1. : Décomposition correcte en préfixe et suffixe

Ce qui est une décomposition correcte.

Par contre, la décomposition du mot "فسمعهم" donne :

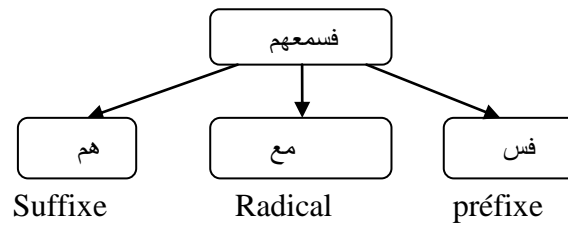


Figure3.2. : Décomposition erronée.

C'est une décomposition fausse bien que le mot "مع" existe en arabe, car la racine de "سمعهم" est "سمع" et non pas "مع". Le problème qui s'est posé dans ce cas est dû au fait que "س" est une lettre du radical "سمع" et peut jouer le rôle d'un préfixe en même temps. La prise en compte de tous les cas envisageables de décomposition garantit la rencontre de la racine exacte du mot.

1.2.2. Identification du schème et de la racine :

Le principe de la méthode de recherche est très simple : pour un mot M, un schème S de la table des schèmes correspond au mot M si la taille du schème est égale à la taille du mot M et si toutes les lettres correspondant aux infixes du schème se trouvent dans le mot M aux mêmes positions.

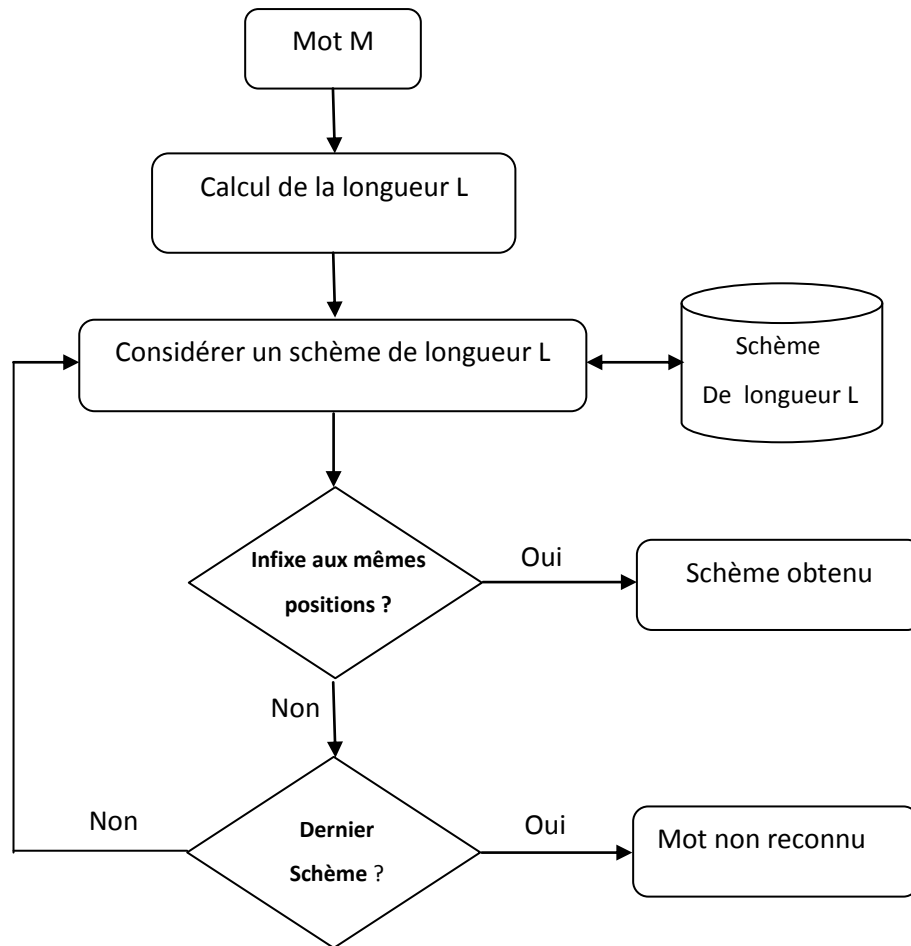


Figure 3.3. : Recherche du schème et de la racine.

Ainsi, l'obtention de la racine se fait en éliminant les lettres infixes du mot M.

Exemple :

Soit le mot "مقاليد". Le processus de recherche de schème parcourt la table des schèmes en vérifiant tous les enregistrements qui ont la même taille que le mot (ici, 6) jusqu'à rencontrer le schème "مفاعيل". La position des infixes dans ce schème sont "1" la lettre "م", "3" la lettre "ا" et "5" la lettre "ي" qui sont leurs mêmes positions dans le mot "مقاليد" donc, c'est le bon schème.

L'élimination de la lettre "م" de la position 1, la lettre "ا" de la position 3 et la lettre "ي" de la position 5 du mot "مقاليد" donne la racine correcte du mot "مقاليد" qui est "قلد". (figure 3.4)

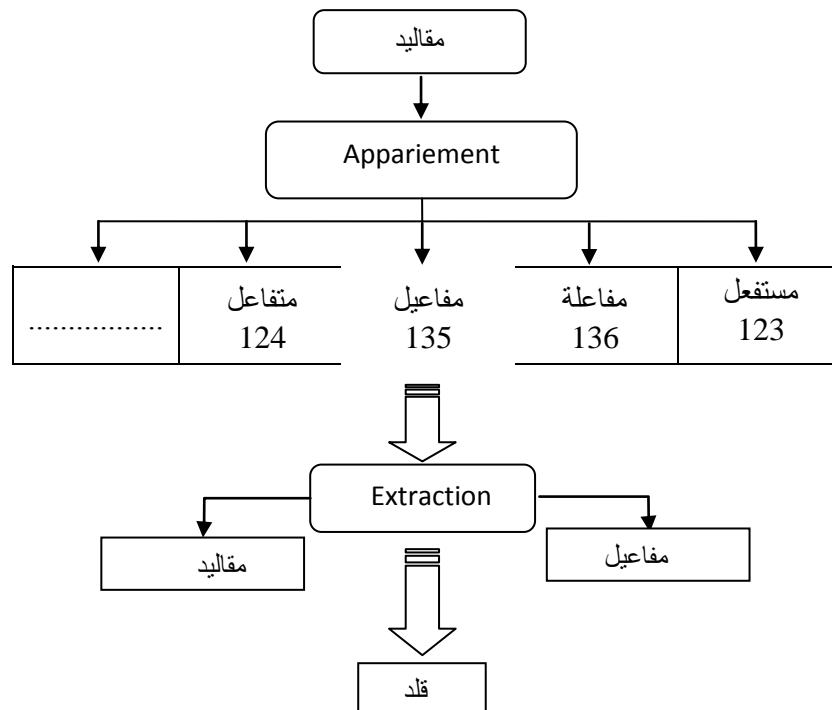


Figure 3.4. : Recherche du schème et de la racine.

2. Les outils utilisés pour l'implémentation :

L'implémentation du système proposé a nécessité l'utilisation d'un certain nombre d'outils :

2.1. L'environnement de programmation (NetBeans) :

Comme environnement de programmation, nous avons choisi l'environnement intégré (IDE) pour Java, NetBeans, placé en open source par Sun en juin 2000 sous licence CDDL (Common Development and Distribution License) et qui, en plus de Java, permet également de supporter différents autres langages, comme Python, C, C++, XML et HTML. NetBeans IDE 7.1 est le premier IDE supportant les tout derniers standards et spécifications de la plateforme Java, avec notamment le support complet du développement pour JavaFX 2.0 et JDK7.

2.2. Le langage de programmation :

Le langage choisi est Java qui est un langage de programmation informatique orienté objet créé par James Gosling 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. La particularité et l'objectif central de Java est que les logiciels écrits dans ce langage doivent être très facilement portables sur plusieurs systèmes d'exploitation tels que

UNIX, Windows, ... avec peu ou pas de modifications. Pour cela, divers plateformes et frameworks associés visent à guider, sinon garantir, cette portabilité des applications développées en Java.

2.3. Le serveur de Bases de données :

Toutes les tables utilisées dans notre système sont conçues à l'aide du serveur de bases de données XAMPP qui est un ensemble de logiciels permettant de mettre en place facilement un serveur Web et un serveur FTP. Il s'agit d'une distribution de logiciels libres (X Apache MySQL Perl PHP) offrant une bonne souplesse d'utilisation, réputée pour son installation simple et rapide. Ainsi, il est à la portée d'un grand nombre de personnes puisqu'il ne requiert pas de connaissances particulières et fonctionne, de plus, sur les systèmes d'exploitation les plus répandus.

3. Présentation de l'interface :

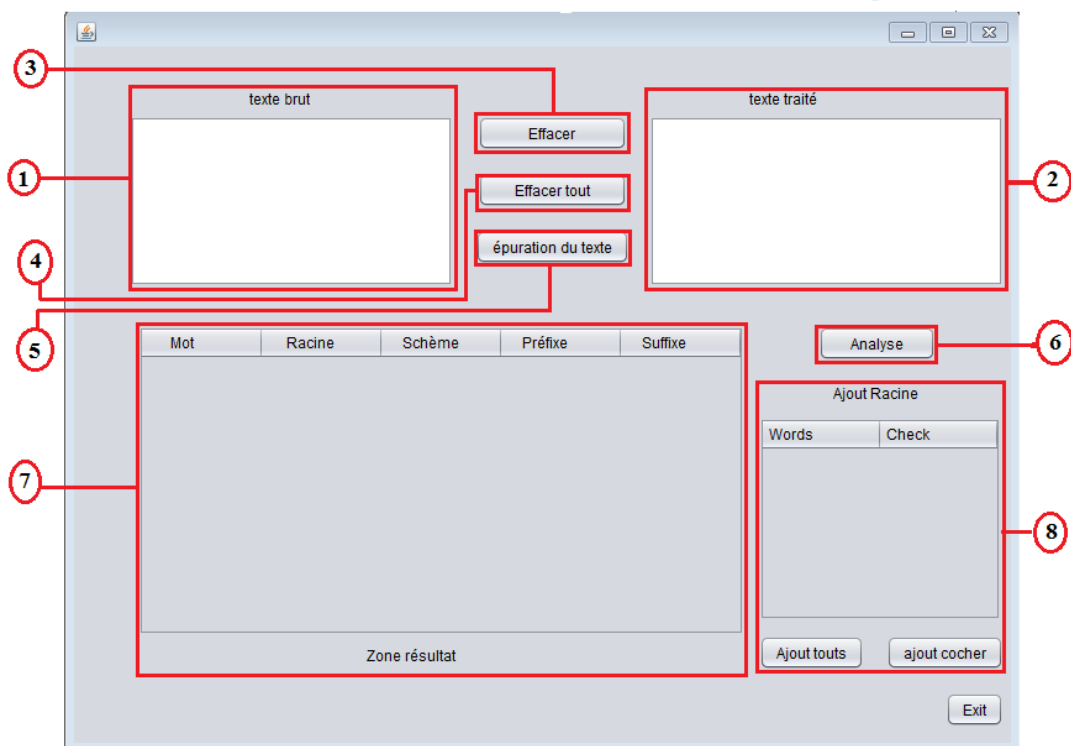


Figure 3.5. : L'interface de l'application

- ① Texte brut : zone dans laquelle est saisi le texte à traiter.
- ② Texte traité : sert à contenir le texte résultant de la phase d'épuration.
- ③ Effacer : bouton servant à effacer la zone texte brut
- ④ Effacer tout : sert à effacer toutes les zones.

L'utilisation de la langue arabe comme moyen de communication à travers le support informatique a été longtemps appréhendée avec beaucoup d'hésitation par la communauté scientifique, notamment celle du monde arabe où cet outil trouvera beaucoup d'utilisations importantes. En effet, la langue et les différentes difficultés qui s'y rattachent, notamment le problème de l'ambiguïté issue de l'absence des voyelles, le problème de reconnaissance des formes fléchies (la langue arabe étant fortement flexionnelle) et le problème du manque de diversité des sujets traitant le domaine du traitement morphologique de la langue arabe se limitant à juste une partie de ce dernier, tout cela pose un énorme défi difficile à surmonter.

Malgré tout cela et malgré la courte durée consacrée à la réalisation de ce sujet, nous avons osé nous aventurer dans ce domaine et on peut dire que, vu les résultats obtenus, nous pensons qu'on a quand même pu relever ce défi et par la même occasion apprendre beaucoup de nouvelles connaissances tout au long de la réalisation de ce travail telles que la programmation objet (java) et les concepts du traitement automatique du langage.

Toutefois, le sujet étant très vaste, il reste beaucoup à faire pour améliorer ce travail, on peut donc proposer comme perspectives, l'introduction d'un plus grand nombre de racines et de schèmes ainsi que l'extension du traitement pour prendre en compte les mots défectueux.

Bibliographie :

- [1] A. Chen, et F. Gey, "Building an Arabic stemmer for information retrieval" .TREC 2002. Gaithersburg: NIST, pp 631-639, 2002.
- [2] A. Gower, M. Poesio, A. De Roeck et J. Reynolds, "Identifying Broken Plurals in Unvowelised Arabic Text". Proceedings of EMNLP, pp. 246-253, 2003.
- [3] Abd El Salam AL HAJJAR, "extraction et gestion de l'information a partir des documents arabes", thèse de Doctorat, Université Paris VIII, Saint Denis ,2010.
- [4] Abd El Salam AL HAJJAR, Mohammad HAJJAR, 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, Mohammad HAJJAR, Khaldoun ZREIK, "A system for evaluation of arabic root extraction methods", Fifth International Conference on Internet and Web Applications and Services, Barcelone, 2010.
- [6] Aïda KHEMAKHEM, "ArabicLDB : une base lexicale normalisée pour la langue arabe", Mémoire de master, université de Sfax, Tunisie, 2006.
- [7] Al Ajeeb Al Ajeeb, Sakher Company, website: <http://lexicons.ajeel.com>. 2010.
- [8] A. N. De Roeck, et W. Al-Fares, "A morphologically sensitive clustering algorithm for identifying Arabic roots". Proceedings ACL-2000. HongKong, pp. 199 – 206, 2000.
- [9] Atef Ben Youssef, "Méthodes Mixtes pour la Traduction Automatique Statistique", Mémoire de master, université Stendhal, Grenoble3, 2008.
- [10] B. Hammo, H. Abu-Salem, S. Lytinen, et M. Evens , "A Question Answering System to Support the Arabic Language ", Proceedings of the ACL-02 workshop on Computational approaches to Semitic languages Philadelphia, Pennsylvania, pp. 1 – 11, 2002.
- [11] Ben Taamallah Sahnoun, "Prétraitement de données et création d'un segmenteur de l'arabe pour un système de traduction probabiliste vers le français", Mémoire de master, université Stendhal, Grenoble3, 2012.
- [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", thèse de doctorat, Université de Nantes, 2008.

- [13] CHAAR France," Traitement automatique de la langue arabe : La modération en arabe ",
Mémoire de master, Institut National des Langues et Civilisations Orientales, Paris, 2012.
- [14] Dilekh Tahar, "Implémentation d'un outil d'indexation et de recherche des textes en
arabe", Mémoire de Magister, université Hadj Lakhdar , Batna,2011.
- [15] Ed-dariouache Adnane, "Etude et réalisation d'un analyseur morphologique de la langue
arabe", mémoire de master, UNIVERSITE Sidi Mohamed Ben Abdellah, Fès, 2015.
- [16] F. Abu Hawas, "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.
- [17] Farag 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. Amsterdam City, Netherlands, pp. 39-46, 2007.
- [18] Fouad Soufiane Douzidia , " Résumé automatique de texte arabe", Mémoire de M.Sc,
Université de Montréal,2004.
- [19] G. A. Kiraz, "Analysis of the Arabic Broken Plural and Diminutive", Proceedings of the
5th International Conference and Exhibition on Multi-Lingual Computing (ICEMCO96),
Cambridge, UK, 1996.
- [20] H. Al Ameen, S. Al Ketbi, A. Al Kaabi, K. Al Shebli, N. Al Shamsi, N. Al Nuaimi, et S.
Al Muhairi, "Arabic Light Stemmer: A new Enhanced Approach" , The Second
International Conference on Innovations in Information Technology (IIT'05), 2005.
- [21] H. Wehr," Dictionary of Modern Written Arabic". Publié par Harrassowitz -Germany,
1961.
- [22] K. Darwish," Al-stem: A light Arabic stemmer, 2002". Available: [http :
//www.glue.umd.edu/~kareem/research](http://www.glue.umd.edu/~kareem/research).
- [23] K. R. Beesley, "Finite-State Morphological Analysis and Generation of Arabic at Xerox
Research: Status and Plans in 2001". The ACL 2001Workshop on Arabic Language
Processing: Status and Prospects, Toulouse, France, 2001.
- [24] K. Taghva, R. Elkoury , et J. Coombs, "Arabic Stemming without a root dictionary",
International Conference on Information Technology: Coding and Computing (ITCC'05)
- Volume I pp. 152-157, 2005.

- [25] 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.
- [26] L. Larkey, L. Ballesteros, et M. Connell," Light Stemming for Arabic IR Arabic Computational Morphology: Knowledge-based and Empirical Methods", A.Soudi, A. van en Bosch, and Neumann, G., Editors. Kluwer/Springer's serieson Text, Speech, and Language Technology, 2005.
- [27] M. Aljlal et Frieder Aljlal, et O. Frieder," On arabic search: Improving the retrieval effectiveness via a light stemming approach". Proceedings of ACM Eleventh Conference on Information and Knowledge Management, Mclean, VA, pp. 340 - 347 , 2002.
- [28] M. Mustafa, H. AbdAlla, et H. Suleman, "Current Approaches in Arabic IR: A Survey". Proceedings The Annual International Conference on Asia-Pacific Digital Libraries (ICADL), Bali, Indonesia. 2008.
- [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] 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.
- [31] Maraoui Mohsen , Zrigui Mounir, Antoniadis Georges," Un système de génération automatique de dictionnaires étiquetés de l'arabe",CITALA 2007, Rabat, Maroc
- [32] 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, pp. 716-719, 2014.
- [33] R. Al Shalabi, et N. Evens, "A Computational Morphology System for Arabic", Proceedings of COLING-ACL, New Brunswick, NJ, pp. 66-72, 1998.
- [34] R. Blachère, M. Gaudefroy-Demombynes, "Grammaire de l'arabe classique", Edition Maisonneuve-Larose, Paris, 1975.
- [35] S. Al-Fedaghi et H. Al-Sadoun, "Morphological compression of arabic text," in Information Processing & Management, pp. 303-316, 1990.

- [36] S. Khoja and R. Garside, "Stemming Arabic text". Computing Department, Lancaster University, Lancaster, www.comp.lancs.ac.uk/computing/users/khoja/stemmer.ps, 1999.
- [37] S. Mesfar, "Analyse morpho-syntaxique automatique et reconnaissance des entités nommées en arabe standard". Thèse de doctorat, université de Franche-Comté, 2008.
- [38] Y. Kadri, A. Benyamina, "Système d'analyse syntaxicosémantique du langage arabe", *mémoire d'ingénieur, université d'Oran Es-sénia, 1992*

TABLE DES MATIERES

INTRODUCTION GENERALE

CHAPITRE 1 : LA LANGUE ARABE

1. Particularités de la langue arabe :	01
1.1. L'alphabet:	01
1.2. L'écriture :.....	01
1.3. Voyellation :.....	02
2. Morphologie du mot arabe :.....	04
2.1. Structure du mot arabe :	04
2.2. Catégories du mot arabe :.....	05
2.2.1. Le verbe :.....	05
2.2.2. Le nom :	07
2.2.3. Les particules :	08
2.3. Eléments essentiels de la morphologie du mot arabe :.....	09
2.3.1. La racine :.....	09
2.3.2. le schème :.....	10
2.3.3. les affixes :	11
2.4. Morphologie dérivationnelle et flexionnelle :	14
2.4.1. La morphologie dérivationnelle :	14
2.4.2. Morphologie flexionnelle :.....	15
2.5. mots dérivés :	15
2.6. mots isolés	16
Conclusion	16

CHAPITRE 2 : ANALYSE MORPHOLOGIQUE DE LA LANGUE ARABE

1. Processus d'extraction de la racine d'un mot arabe :.....	17
2. Traitements communs :.....	18
2.1. Normalisation :.....	18

2.2. Transcription :	18
3. Les méthodes d'extraction de la racine d'un mot arabe :	19
3.1. Les méthodes basées sur l'analyse morphologique :	19
3.1.1. Méthode basée sur les affixes :	19
3.1.2. Méthodes basées sur les racines et les affixes :	20
3.2. Les méthodes basées sur les tables de correspondance (génération systématique) :	21
3.3. Les méthodes basées sur la transcription ou la traduction :	22
3.4. Les méthodes basées sur la position des lettres :	23
3.5. Les méthodes basées sur l'analyse statistique :	23
3.5.1. Technique basée sur le <i>coefficient de similarité</i> :	23
3.5.2. Technique basée sur le <i>coefficient de dissimilarité</i> :	24
3.6. Les méthodes basées sur l'analyse hybride :	25
Conclusion	25

CHAPITRE 3 : CONCEPTION ET REALISATION

1. Description du système réalisé :	26
1.1. Les tables utilisées :	26
1.2. Le système d'extraction :	27
1.2.1. Elimination des affixes :	27
1.2.2. Identification du schème et de la racine :	28
2. Les outils utilisés pour l'implémentation :	30
2.1. L'environnement de programmation (NetBeans) :	30
2.2. Le langage de programmation :	30
2.3. Le serveur de Bases de données :	31
3. Présentation de l'interface :	31
Conclusion :	32
Conclusion Générale	33

BIBLIOGRAPHIES

Liste des tableaux

Table1.1. :	Les 28 lettres arabes.....	02
Table 1.2. :	Exemple de variation de la lettre ع ayn.....	02
Table 1.3. :	Les Voyelles brèves	03
Table 1.4. :	Ambiguïté causée par l'absence de voyelles pour les unités lexicales مدرسة et كتب.....	03
Table 1.5. :	Les diacritiques doubles.....	04
Table1.6. :	Structure du mot arabe.....	04
Table 1.7. :	Dérivationnel irrégulier	08
Table 1.8. :	Dérivationnel régulier.....	08
Table1.9. :	Quelques mots dérivés de la racine درس (a étudié)	10
Table 1.10 :	Un exemple des préfixes.....	12
Table 1.11. :	Un exemple des suffixes divisés selon leurs types.....	13
Table 1.12. :	Un exemple de génération des Stems.	13
Table 1.13.:	Un exemple de formation de mot <أطلبون>, attlbwun, Est-ce que vous demandez ?>.....	15
Table 2.1. :	Exemple de table de correspondance.....	21
Table3.1. :	Table des racines	26
Table3.2. :	Table des schèmes	26
Table3.3. :	Table des mots outils.....	27

Listes des figures

Figure 1.1. :	Exemples de dérivation de la racine [k t b]	11
Figure 1.2. :	Liste des schèmes verbaux	14
Figure 2.1. :	Processus d'extraction de la racine d'un mot arabe	17
Figure 3.1. :	Décomposition correcte en préfixe et suffixe	28
Figure 3.2. :	Décomposition erronée	28
Figure 3.3. :	Recherche du schème et de la racine	29
Figure 3.4. :	Recherche du schème et de la racine	30
Figure 3.5. :	L'interface de l'application	31
Figure 3.6. :	Exemple de L'interface	32