République Algérienne Démocratique Et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Saad Dahleb Blida 1

Faculté des sciences

Mémoire

Présentée pour l'obtention du diplôme de Master

En : INFORMATIQUE

Spécialité **: Traitement Automatique des Langues**

Sujet

Vers un système de lemmatisation automatique pour la langue Arabe

Etabli par :

M^elle^ Ababsia Sara

M^elle^ Ferrouga Ihcene

Devant le jury composé de :

M^me^.Oukid lamia	PRESIDENTE	USDB1
M^r^. Kamach Hicham Abdallah	EXAMINATEUR	USDB1
M^r^.Amrouche Aissa	Encadreur	CRSTDLA
M^me^.Mezzi Melyara	Promotrice	USDB1

Année universitaire 2018/2019

Résumé

Notre recherche suggère l'un des programmes de base du traitement automatique de la langue et la langue Arabe en particulier, qui consiste à extraire les racines des mots Arabes et qui s'appelle lemmatisation. Nous avons opté pour une méthode qui utilise une liste de schèmes et l'avantage de cette méthode est qu'elle ne s'appuie pas sur les dictionnaires de mots. Au lieu de cela, il faut passer par une série de traitements linguistiques qui visent à trouver le bon schème pour chaque mot afin d'aboutir à la racine appropriée.

Mots clés : Lemmatisation, Traitement Automatique de la Langue, Extraction des Racines, Schèmes, Langue Arabe.

Abstract

Our research suggests one of the basic programs of automatic language processing and the Arabic language in particular, which consists in extracting the roots of Arabic words and it's called lemmatization. We have opted for a method that uses a list of schemes and the advantage of this method is that it does not rely on word dictionaries. Instead, one must go through a series of language treatments that aim to find the right scheme for each word in order to arrive at the appropriate root.

Keywords: Lemmatization, Automatic Language Processing, Word Roots extraction, Schemes, Arabic Language.

ملخص

يقترح بحثنا أحد البرامج الأساسية للمعالجة الالية للغة واللغة العربية على وجه الخصوص، والتي تتمثل في استخراج جذور الكلمات العربية والتي تسمى «lemmatisation». لقد اخترنا طريقة تستخدم قائمة من الأوزان وميزة هذه الطريقة هي أنها لا تعتمد على قواميس الكلمات. بدلاً من ذلك، يجب على المرء الاطلاع على سلسلة من العلاجات اللغوية التي تهدف إلى العثور على المخطط الصحيح لكل كلمة للوصول إلى الجذر المناسب.

كلمات مفاتيح:

Lemmatisation، معالجة الآلية للغة، استخراج الجذر, ، الأوزان. اللغة العربية.

Remerciements

En premier lieu, nous remercions Dieu, notre créateur, qui nous a donné la force et la persévérance pour réaliser ce travail.

Aussi, nous tenons à remercier infiniment

- *Nos chers parents pour leurs soutiens au long de nos études.*
- *Mr.Amrouche Aissa notre encadreur qui nous a accordé son soutien, son aide indéfectible et surtout sa patience et sa gentillesse.*

Nos remerciements les membres du jury pour leur présence, pour leur lecture attentive de notre mémoire ainsi que pour les remarques qu'ils nous adresseront lors de cette soutenance afin d'améliorer notre travail aussi à Mme.Mezzi Melyara d'avoir accepté d'examiner notre travail.

Nous tenons à exprimer notre profonde gratitude et nos sincères remerciements à tous ceux qui ont contribué, de près ou de loin à l'élaboration de ce mémoire de fin d'étude.

Tables des matières

Liste des Abréviations

ACR	Absolute Category Rating
TAL	Traitement Automatique des Langues
TALN	Traitement Automatique des Langues Naturelles
AS	Arabe Standard
NLTK	Natural Language Toolkit
VCS	version control system
IDE	Integrated Development Environment
API	Application Programming Interfaces
ISRI	Information Science Research Institute’s

Liste des figures

Liste des tableaux

Introduction générale

Le Traitements Automatique des Langues (TAL) est une discipline qui associe étroitement les linguistes et informaticiens. Il repose sur la linguistique, les formalismes (représentation de l'information et des connaissances dans des formats interprétables par des machines) et l'informatique. Le TAL a pour objectif de développer des logiciels ou des programmes informatiques capables de traiter de façon automatique des données linguistiques. Pour traiter automatiquement ces données, il faut d'abord expliciter les règles de la langue puis les représenter dans des formalismes opératoires et calculables et enfin les implémenter à l'aide de programmes informatiques [39].

Ce travail expose un procédé de lemmatisation dans l'analyse morpho-lexicale. Cette étape est cruciale car elle est la base de tous les applications de TALN (Traitement Automatique des Langues Naturelles) : la traduction automatique, la compréhension automatique des textes, la génération automatique de textes, la gestion électronique de l'information et des documents existants (GEIDE), l'indexation, la recherche d'information, la réponse aux questions, base pour la classification et la catégorisation des documents [28].

L'application du traitement automatique sur la langue Arabe pose des problèmes majeurs, dont : le problème de l'ambiguïté issue de l'absence des voyelles, ceci exige des règles morphologiques complexes. Le problème de reconnaissance des formes fléchies, car l'Arabe est une langue fortement flexionnelle. La lemmatisation est une procédure ramenant un mot portant des marques de flexion (par exemple, la forme conjuguée d'un verbe) à sa forme de référence (dite lemme), quelle que soit la forme sous laquelle le mot apparaît dans un texte. La lemmatisation sert ainsi à la reconnaissance morphologique des mots d'un texte [28].

Il existe Plusieurs approches de la lemmatisation qui sont appliquées à la langue Arabe, mais un lemmatiseur complet pour cette langue n'est pas disponible car la lemmatisation est difficile pour les langues avec des morphologies complexes comme l'Arabe.

L'objectif de notre travail est de développer un programme (Lemmatiseur) qui peut résoudre les problèmes possibles de traitement automatique de langue Arabe possibles.

Notre mémoire est composé des chapitres suivants :

- Le premier chapitre concerne un aperçu des particularités de la langue arabe, la morphologie de cette langue et les problèmes d'analyse du traitement automatique de cette dernière.
- Le deuxième chapitre est consacré aux différentes méthodes de lemmatisation et les difficultés de la lemmatisation des mots Arabes ;
- Le troisième chapitre concerne les étapes de la conception du système proposé ;
- Le quatrième chapitre traitera des aspects d'implémentation ;
- En dernier lieu, une conclusion générale permet de conclure notre travail et de présenter quelques perspectives pour la continuité et l'amélioration de ce travail.

Chapitre 1

Caractéristiques La langue Arabe

1.1 Introduction

La langue Arabe ne contient que 28 caractères, ce qui permet d'extraire environ 12, 122,912 mots. Et c'est l'une des caractéristiques qui ont fait la langue l'Arabe des plus difficiles langues dans le monde. Ce n'est pas la seule propriété que nous allons vous montrer dans ce chapitre où il se compose d'un aperçu des particularités de la langue l'arabe et système d'écriture de l'Arabe, en suite la morphologie Arabe et en fin les problèmes d'analyse du traitement automatique de la langue arabe.

Dans le cadre de notre travail, nous parlerons de la langue Arabe en référence à ce qui est communément appelé «**l'Arabe Standard** » (AS), c'est-à-dire, la langue de communication commune à l'ensemble du Monde Arabe. Il s'agit de la langue enseignée dans les écoles, donc écrite, mais aussi parlée dans le cadre officiel.

1.2 Un aperçu des particularités de la langue l'Arabe

1.2.1 Repères historiques de la calligraphie Arabe

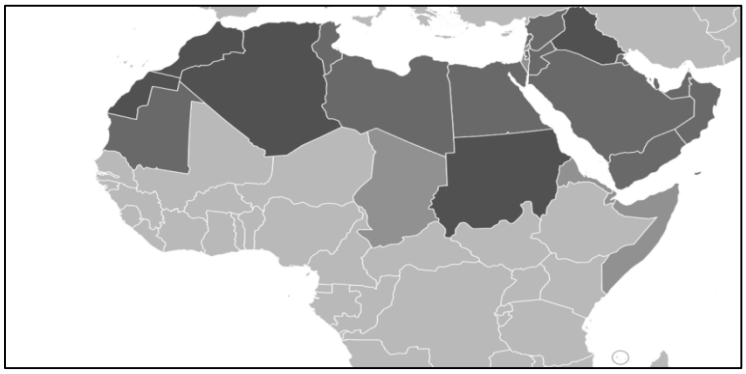


Figure 1.1: le monde Arabe.

La langue Arabe est l'une des langues les plus parlées et utilisées dans le monde [1]. Elle est la langue officielle de plus de 22 pays parlée par plus de 467 millions de personnes. C'est particulièrement important pour les musulmans, parce que c'est la langue du Coran et la prière est accomplie seulement par elle, et que de nombreux pays non Arabes l'utilisent, tels que: le Tchad, la Turquie, le Sénégal, le Mali, l'Éthiopie, et elle a également influencé de nombreuses autres langues à travers les âges, Connu sous plusieurs titres, notamment: la langue du Coran, et la langue [**الضاد**-al ḍād], car lettre [**الضاد**-al ḍād] se spécialise pour les Arabes dans leur langue [41].

Il est donc dit dans les mots d'Abou Tayeb al-Mutanabi :

وعَوْذُ الجاني وغَوْثُ الطَّرِيدِ وبِهِمْ فَخرُ كلِّ مَنْ نَطَقَ الضَّادَ [6] .

L'Arabe doit sa formidable expansion à partir du 7eme siècle grâce à la propagation de l'islam et la diffusion du Coran. Les recherches pour le traitement automatique de l'Arabe ont débuté vers les années 1970. Les premiers travaux concernaient notamment les lexiques et la morphologie [2], mais Les règles fondamentales de la langue Arabe, surtout celles morphosyntaxiques, n'ont pas changé depuis leur mise au point pour le Saint Coran. L'écriture Arabe s'est développée grâce à la révélation coranique [8].

L'écriture en Arabe est plus qu'un simple code de communication, c'est un art avec une calligraphie des plus raffinées, L'écriture Arabe, utilisant l'alphabet Arabe, a connu tour à tour l'introduction des éléments suivants :

- la séparation des mots : l'écriture se faisait en caractères attachés. Le document se présentait sous forme d'une seule ligne dont toutes les lettres sont liées.
- l'introduction de la césure, on la trouve dans certains anciens manuscrits écrit en Koufi ancien.
- l'introduction de la voyellisation : l'utilisation de signes de voyelle dans une couleur (rouge ou jaune) autre que celle utilisée pour les lettres (noires), sous forme d'un point au-dessus, au-dessous ou à gauche des lettres. Utilisation de deux points pour tanwyn.
- l'introduction des signes diacritiques : utilisation de points, un, deux ou trois, au-dessus ou au-dessous des lettres avec la même couleur que les lettres.
- le changement de la voyellisation : signes de voyelles actuels (FATHA, DAMMA, KASRA,..).

L'interdiction de la césure des mots et même de certaines expressions ;

- l'utilisation des signes d'arrêt et de commencement réservés au Saint Coran (Waqf, Wasl, . .).
- l'adaptation restreinte et progressive des signes de ponctuation ou de numérotation.

L'alphabet Arabe sert, moyennant de légers aménagements avec des points, comme alphabet d'écriture pour plusieurs langues à travers le monde, comme par exemple : le berbère, le farsi, le kirghize, le malais, le pashto, le persan, l'urdu, le sindhi, l'ouighour et d'autres langues africaines. Autrefois, le turque et l'espagnol étaient écrits également à l'aide de l'alphabet Arabe [8].

1.2.2 Système d'écriture de l'Arabe

L'écriture de la langue Arabe se distingue par ses différentes caractéristiques par rapport les autres langues dans le monde que nous allons vous présenter dans cette section de nous travaillons comme suit :

L'Arabe est classé sous le groupe des langues sémitiques contemporaines qui s'écrit de droite à gauche [1], et les lettres sont liées entre elles. Il n'existe pas de distinction entre majuscules et minuscules en alphabet Arabe [3], comme dans l'alphabet latin (l'écriture est donc monocamérale).

Son système graphique se compose d'un alphabet Arabe de type abjad constitué de 28 lettres (Tableau 1.1). Cet alphabet contient 25 consonnes et 3 voyelles longues «اٍ» , «و», «ي» [1].

Toutes les lettres se lient entre elles sauf (ذ،د،ز،ر،و،ا) qui ne se joignent pas à gauche. Exemple, Paris s'écrit باريس /bârîs/, [4].

Nous pouvons classer les consonnes selon plusieurs critères : des consonnes articulées avec une vibration des cordes vocales et des consonnes qui n'engendrent pas une vibration des cordes vocales, le franchissement de l'air à travers le conduit vocal donne naissance à d'autres variétés de sons. Mais pour les besoins de la transcription, les 28 consonnes arabes sont divisées en deux groupes :

- 14 consonnes solaires qui assimilent le «ل «de l'article.
- 14 consonnes lunaires qui n'assimilent pas le «ل «de l'article [22].

Les solaires se prononcent en double, comme par exemple avec le mot « soleil » شمس (chams), au lieu de prononcer, el-chams, on prononce ech-chams, car la lettre ش(chin), est une lettre solaire. Les lettres lunaires, se prononcent normalement et simplement pour elles-mêmes, c'est-à-dire sans les doubler. Par exemple avec le mot « lune », قمر) qamar - lune), on prononce القمر, el-qamar tout à fait normalement, parce que la lettre ق) qaf) est une lettre lunaire (Tableau 2.2) [24].

La Classification des consonnes tenant compte des contraintes de la transcription comme suit [22] :

- Solaires : **ت ث د ذ ر ز س ش ص ض ط ظ ل ن**.
- Lunaires :**أ ب ج ح خ ع غ ف ق ك م هـ و ي** .

Lettre Arabe	Correspondant français	Prononciation	Lettre Arabe	Correspondant français	Prononciation	Lettre Arabe	Correspondant français	Prononciation
ا	A	Alef	ز	z	zāy	ك	k	Kāf
ب	B	Bā	س	s	sīn	ل	l	lām
ت	T	Tā	ش	sh	chīn	م	m	mīm
ث	Th	Thā	ص	s	Ṣād	ن	n	nūn
ج	J	Jīm	ض	d	ḍād	ه	h	hā
ح	H	ḥā	ط	t	ṭā	و	w	wāw
خ	Kh	Khā	ظ	z	ẓā	ي	y	yā
د	D	Dāl	ع	‘‘	Ɛayn	ق	q	qāf
ذ	D	Dhal	غ	gh	ghayn			
ر	R	Rā	ف	f	fā			

Tableau1.1: Les 28 lettres Arabes.

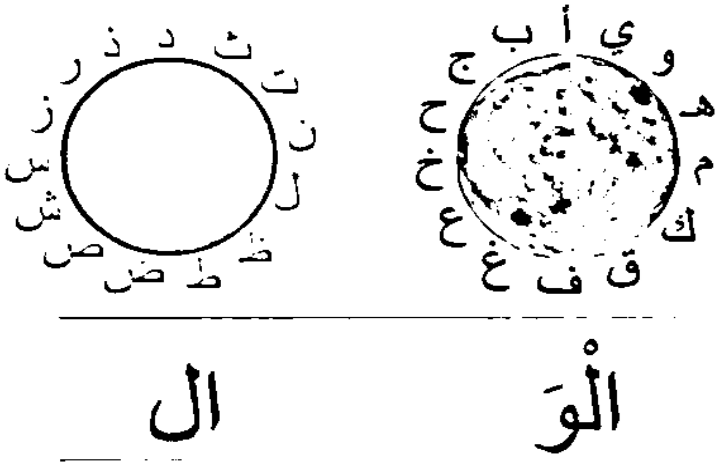


Figure1.2: les lettres solaires et les lettres lunaire [29].

Le système de signes diacritiques des points joue un rôle de premier ordre. Il y a des lettres qui ne se distinguent que par la présence, le nombre et la position de points. En effet, les lettres {ة, ـق, ـفـه ,ـن, ـث, ـب, ـت , ـيـ, ح, ج , خ, د, ذ, ر, ز, س, ش, ص, ض, ط, ظ, ع, غ} sont notées au moyen des glyphes{ـه, ع, ط, ص, س, ر, ح,ـ} [8].

De plus certaines lettres comme Alef ou Hamza peuvent symboliser le أ, ؤ,آ, ء, ئ ou إ de même que pour les lettres ى et ه qui symbolisent respectivement ي et ة [5].

Aussi, l'Arabe est semi cursive dans le sens où son alphabet est unique mais la forme des lettres change en fonction de la position qu'elles occupent dans le mot [1].

Chaque lettre possède une forme spécifique en fonction de sa position dans un mot (au début, au milieu ou à la fin) ou si elles sont utilisées de façon isolée [1], Il prend 78 différentes formes graphiques à partir des 28 lettres. Le Tableau 1.2 montre les variations de la lettre ض (ḍād) :

Isolée	**Finale**	**médiane**	**initiale**
ض	**ـض**	**ـضـ**	**ضـ**
افتراض/iftirāḍ/ Assomption	مريض/marīḍ/ patient	مفضل/mufaḍḍl/ favori	ضعيف/ḍa3īf/ pauvre

Tableau 1.2: Exemple de variation de lettre ض/ ḍād.

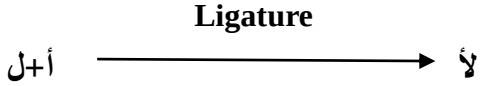
Les lettres sont attachées entre elles avec une *kashida*. La *kashida*' est pas une lettre en elle-même mais plutôt un allongement de certaines lettres en respectant très rigoureusement les règles de la calligraphie Arabe [9].

Mot	**sens**	**Kashida**
al-ḥamidu	la louange	الحميـــد
Rafiq	compagnon	رفـــيق

Tableau 1.3: Exemple sur kashida.

Les Ligatures sont l'union de deux graphèmes pour former un seul graphème est une ligature. Dans l'écriture Arabe on a des ligatures esthétiques, optionnelles qui se rencontrent surtout dans des compositions soignées. L'exemple ci dissout montre la ligature Lam-alif, qui est obligatoire [27].

Par exemple :



1.2.3 Les diacritiques

L'écriture Arabe comporte aussi des voyelles courtes, [1] sont figurées par des symboles appelés signes diacritiques (les harkāt), [4] qui sont généralement facultative mais essentielles dans les textes religieux (Coran, Hadith, etc.), [1] en peuvent représenter dans tableau suivant :

Nom	Transcription	Voyelle courte	singe
Fatha	[a]	est symbolisée par un petit trait sur la consonne	ـَ
Damma	[u]	est symbolisée par un crochet au-dessus de la consonne	ـُ
Kasra	[i]	est symbolisée par un petit trait au-dessous de la consonne	ـِ

Tableau 1.4: Les voyelles courtes en Arabe.

Il existe de plus, une série d'autres diacritiques dont les plus courants comme :

- **Le Sukun :** Un petit rond ◌ْsymbolisant la soukoun (سكون) est appose sur une consonne lorsque celle-ci n'est liée à aucune voyelle (سَوْف / sawfa) [4].
- **La Chadda :** Le signe de la chadda◌ّ peut être placé au-dessus de toutes les consonnes en position non- initiale. La consonne qui la reçoit est alors analysée en une séquence de deux consonnes identiquesالحقّ/alha**kk**o « le droit » [4].
- **Le Tanwin :** Le signe du tanwin est ajouté à la fin des mots indéterminés. Il est en relation d'exclusion avec l'article de détermination ال placé en début de mot. Les symboles du tanwin sont au nombre de trois et sont constitués par dédoublement des signes diacritiques ci-dessus, ce qui se traduit par l'ajout du phonème /n/ au niveau phonétique [4] :
- [an] : signe ◌ً (بً /ban).

- [un] : signeٌ (بٌ/bun).
- [in] : signe ٍ (بٍ /bin).

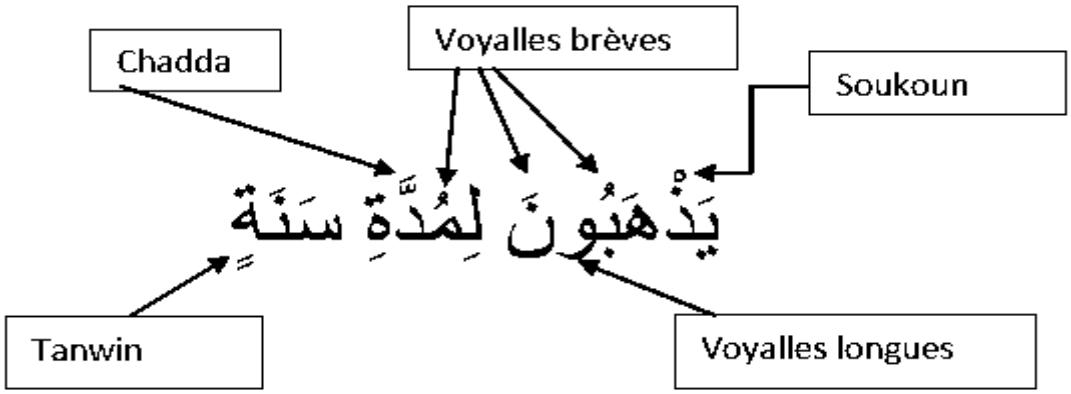


Figure1.3: exemple représenté différente Les diacritiques.

1.2.4 Les signes de ponctuation

- Les signes logiques : Il est à noter que la langue Arabe utilise deux types de virgules codées différemment : une virgule numérique identique à la virgule en caractères latins et une autre textuelle inversée et orientée à droite [10].
- Les signes séquentiels : Il est à noter que l'apostrophe n'existe pas en Arabe. Tout comme en français, le caractère qui représente le tiret est le même correspondant au signe mathématique de soustraction « - » [10].
- Les chiffres : Contrairement à ce qui est répandu, même dans certains pays Arabes, les chiffres Arabes sont ceux utilisés en occident et dans tout le Maghreb « 1, 2, 3, ... ». Les chiffres qu'on utilise dans certains pays Arabes du Moyen-Orient « ١, ٢, ٣, ... », sont en fait des chiffres indiens [10].

1.3 Morphologie Arabe

Selon la grammaire traditionnelle, le lexique Arabe comprend trois catégories de mots : verbes, noms (substantifs et adjectifs) et particules (adverbes, conjonctions et prépositions). Hormis les noms propres, les mots des deux premières catégories sont dérivés à partir d'une racine : un squelette de trois consonnes radicales le plus souvent [5].

1.3.1 Structure d'un mot

En Arabe un mot peut signifier toute une phrase grâce à sa structure composée qui est une agglutination d'éléments de la grammaire, la représentation suivante schématise une structure possible d'un mot. Notons que la lecture et l'écriture d'un mot se fait de droite vers la gauche [2].

Enclitique	Suffixe	Cors schématique	Préfixe	Proclitique

Exemple :

أَتُعَلِمِينَهُم

Est-ce que tu peux les enseigner ?

Enclitique	Suffixe	Cors schématique	Préfixe	Proclitique
هُم **(hom)**	ينَ **(yna)**	عَلِمِ **(EaLiMi)**	تُ **(tu)**	أ (a)
Objet masculin pluriel	Sujet féminin singulier	la racine « ELM »	préfixe verbal du temps de l'inaccompli	conjonction d'interrogation

Tableau 1.5: Structure de Mot Arabe أَتُعَلِمِينَهُم.

- ***Préfixe*** : Les préfixes sont représentés par un morphème correspondant à une seule lettre en début de mot, qui indique la personne de la conjugaison des verbes au présent. Les préfixes ne se combinent pas entre eux. Le tableau suivant présente la liste des préfixes verbaux en Arabe [14] :

Le préfixe	Signification
أ	Indique la première personne au singulier (je)
نَ	Indique la première personne au pluriel (nous)
تَ	Indique la deuxième personne féminine, masculine, singulière et duelle
يَ	Indique la troisième personne masculine au singulier, duel, pluriel, masculin et féminin pluriel.

Tableau 1.6: Liste des préfixes Arabe [14].

- ***Suffixe*** : Les suffixes en Arabe sont essentiellement utilises pour des terminaisons des conjugaisons verbales, ainsi que les marques du pluriel et du féminin pour les noms [14].

Les proclitiques : Les proclitiques sont en inventaire finis, et se combinent entre eux pour donner les traits syntaxiques (coordonnant, déterminant ...) qui peuvent accompagner le mot Arabe [15]. Dans le cas des verbes, les proclitiques dépendent exclusivement de l`aspect verbal. Dans le cas des noms et les déverbaux, le proclitique dépend du mode et du cas de déclinaison [14].

Nous pouvons répartir les proclitiques dans les catégories suivantes [1] :

Les proclitiques réserves aux noms et adjectifs :

L'article de définition« ال ».

Les prépositions : « بِـ », « لِ », « كَ ».

Les proclitiques réserves aux verbes :

- La particule du subjonctif « نصب ».
- La particule de la future : « سَ ».
- La particule d'occupation « جزم » : ل.

Les proclitiques généraux, utilises indépendamment de la catégorie des mots auxquelles ils s'attachent :

La conjonction de coordination : « فَـ » et « وَ ».

L'article d'interrogation : « أ ».

Le marqueur de corroboration « تأكيد » : « ل » [1].

Cette classification n'omet pas le fait qu'il existe certaines exceptions de proclitiques qui peuvent jouer différents rôles, comme pour le proclitique ' و ' (wa) utilisé généralement comme particule de liaison (conjonction de subordination et de coordination), mais également peut être utilisé comme particule d'accompagnement (واو المعية) ou de serment (واو القسم) [1].

Les proclitiques sont classés en quatre positions suivant :

- **1ere position :** L'article d'interrogation"أ".
- **2ième positon :** Les conjonctions de coordination" ف" et "و".
- **3ime positions :** -Les propositions :"بِـ"et"لِ"et "كَ".

  -La particule des surjections ≪نصب≫ : " لَ ".

  -La particule de l'apocope ≪جزم≫ : "لِ ".
- **4ime positions :** L'article de définition "الْ".

Par ailleurs, nous signalons que la fusion des proclitiques n'est pas faite de façon

Aléatoire, elle suit deux types de contraintes exprimées par une relation d'ordre et un ensemble de règle de compatibilité comme suit [1] :

Une relation d'ordre : chaque proclitique est incompatible, à cause de la relation d'ordre strict, avec un proclitique de même position, c'est le cas par exemple des proclitiques wa-و et fa coordonnants (فاء العطف et واو العطف) qui occupent la position 2 dans le vecteur d'ordre. Nous notons aussi qu'un proclitique occupant une position d'antériorité par apport à un autre n'a aucune chance de se retrouver placé après ce dernier dans la construction d'un mot graphique. Par exemple, l'interrogatif 'a- (همزة الاستفهام) occupe toujours la première position dans un mot graphique maximal et il est impossible de le trouver précédé par un autre proclitique [1].

Règles de compatibilité : pour des raisons syntaxiques et sémantiques, certains proclitique ne sont pas compatibles entre eux, c'est le cas par exemple des lettres « بetل » (bi- et li-) qui ne peuvent pas se combiner, car elles sont des prépositions (حروف جر) ayant des sens différent [1].

- ***Les enclitiques* :** Les enclitiques présentent les pronoms suffixes qui s'attachent toujours à la fin du mot graphique, leur liste est constituée des 17 éléments suivants : {نِي, يْنَ, اكَ, كِ, كُمَا, كُمْ, كُنَّ, هُ, هُمَا,هَا, هُمْ, هُنَّ , هِ, هِمَا, هِنَّ, هِمْ, نا}. Un mot graphique ne contient qu'un seul enclitique à la fois. Ils s'attachent aux verbes comme étant un complément-objet et aux noms et aux. Prépositions comme un complément du nom ou complément d'objet indirect. Leurs utilisations est régie par certaines restrictions [1]. Les enclitiques à la première personne tels que "نِي" (niy – moi / mon) ou "نا" (naa – nous/ notre) et ceux à la deuxième personne tels que "كَ" (ka – toi/ton) ou "كُمْ" (kum – vous/votre [masculin pluriel]) ont une forme invariable, mais ceux de la troisième personne sont variables et prennent différentes vocalisation suivants les règles suivantes :

  Dans le cas des verbes, l'enclitique peut varier en fonction de l'aspect du verbe et du pronom. La comptabilité entre les enclitiques et les verbes dépend de la propriété de transitivité du verbe. Ainsi, les verbes intransitifs et ceux conjugués à la forme passive ne prennent jamais des enclitiques. Par ailleurs, l'utilisation des enclitiques dans le cas des verbes peut être répartie selon l'aspect du verbe.

Dans le cas nominal, l'enclitique doit respecter une harmonie vocalique avec la voyelle casuelle de la forme à laquelle il se rattache, et dans le cas des noms se terminant par une voyelle double ou ***tanwin***, ces derniers ne prennent jamais des enclitiques. Seul le mode déterminé par annexion est susceptible de prendre des enclitiques selon les règles suivantes [1] :

Si le nom est fléchi au nominatif ou à l'accusatif, il nécessite l'utilisation des enclitiques suivants : هُ [PRON+3+m+s], هُمَا [PRON+3+m|f+d], هُمْ [PRON+3+m+p], هُنَّ [PRON+3+f+p]

Si le nom est fléchi au génitif, il nécessite l'utilisation des enclitiques suivants : هِ [PRON+3+m+s], هِمَا [PRON+3+m|f+d], هِمْ [PRON+3+m+p], هِنَّ [PRON+3+f+p].

Par ailleurs, certain mots nécessitent des transformations morphologiques avant de leur rattacher des enclitiques, c'est le cas des noms se terminant par une hamza, une " ى" ou une " ي" (y). Par exemple la forme مَلْهَى (*malha-* un manège), nécessite une transformation de celle-ci en " ا" avant sa suffixation pour produire la forme agglutinée مَلْهَاه (*malhAhu*– son manège) [1].

- **Les pré-bases :**

Les pré-bases sont obtenues par combinaison entre le(s) proclitique(s) et le préfixe. La génération des pré-bases se fait d'une manière automatique. Le tableau 1.11 représente un exemple de groupe de pré-base [14] :

Pré-base	Préfixe	Proclitique
أتَـ	تَـ	أ
سَتَّـ	تَّـ	سـ
أفَتـ	تَّـ	أفـ
أسَتَّـ	تَّـ	أسـ
وَسَتَّـ	تَّـ	وسـ
فَسَتَّـ	تَّـ	فسـ
أفَسَتَّـ	تَّـ	أفسـ

Tableau 1.7: Exemple de groupe de pré-base [14].

- **Les post-bases :**

En Arabe, les post-bases sont obtenues par combinaison entre le suffixe et le(s)enclitique(s). Les compatibilités dépendent des pronoms décrits par chacune des particules [15] :

- Les suffixes de la première personne se combinent très souvent avec les enclitiques de la deuxième et la troisième personne.
- Les suffixes de la deuxième personne se combinent très souvent avec les enclitiques de la première et la troisième personne.
- Les suffixes de la troisième personne se combinent très souvent avec les enclitiques de la première, la deuxième personne et la troisième personne. Enfin, il existe en Arabe des suffixes qui jouent le rôle de caractère terminal du mot. En effet ces types des suffixes ne se combinent avec aucun enclitique(s). Le tableau suivant présente un exemple de groupe de post-bases [14] :

Post-base	Enclitique	Suffixe
وك	ك	و
وهم	هم	و
وننا	نا	ونَ
ونني	ني	ون
أنكم	كم	أن
أنهم	هم	أن
تموه	ه	تمو

Tableau1.8: Exemple de poste-base [14].

- **Les racines «الجذر» :**

Une racine est purement consonantique, elle est formée par une suite de trois ou quatre consonnes (ou même cinq pour les noms) formant la base du mot. Elles sont aux alentours de 10000 racines dont la grande majorité (85%) sont trilatérales. Les restes sont des racines quadrilatérales ou quintilatérales. Une racine définit la signification fondamentale des mots dérivés en utilisant différents diacritiques et affixes avec les lettres de la racine pour créer l'inflexion de la signification. [18].

Par exemple, la racine « رسم» (Il a dessiné) a la signification de base « dessiner». Plusieurs mots sont dérivés à partir de cette racine et le tableau ci-dessus représente Quelques dérivation du verbe " رسم " :

	La racine " رسم "(dessiner)			
Verbes	رسم	Il a dessiné	يرسم	Il dessine
	رسما	Nous avons dessiné	يرسمون	Ils dessinent
	رسمت	Elle a dessiné	ترسم	Tu dessines
	ترسمون	Vous dessinez	نرسم	Nous dessinons
Noms	رسّام	peintre	مراسِم	la cérémonie

	رسومات	dessins	مرسوم	décret
	مرسم	Studio	راسمة	Traceur

Tableau1.9: Quelques dérivation du verbe " رسم ".

- **Les schèmes«الاوزان» :**

Un schème représente une forme ou modèle général composé par une séquence de caractères. Quelques caractères sont constants et d'autres sont variables. Les caractères variables sont destinés à être substitués par d'autres d'une racine pour générer le radicale. Les schèmes servent à produire la plupart des mots Arabes à partir d'une racine ou inversement à extraire la racine d'un mot. Ils sont aux alentours de 900 schèmes. Quel que soit le mot, il est donc issu d'une racine et inséré dans un schème. En fait le schème est une sorte de moule. De plus, ils permettent de déterminer la racine d'un mot Arabe .En général, les racines trilatérales sont représentées par le modèle«فَعَلَ, faire », les racines quadrilatérales sont représentées par le modèle«فَعْلَلَ». Par contre, on peut trouver le schème «عل » qui représente les mots qui ont perdu l'une des leurs lettres. Par exemple, le mot « زن, pèses » a pour racine « وزن » [16].

La racine	Le schème	Résultat
كتب	فِعَالٌ	كِتَابٌ
درس	مَفْعُولٌ	مَدْرُوسٌ
خرج	فَاعِلٌ	خَارِجٌ
روى	فَاعٍ	رَاوٍ

Tableau1.10: Exemple de construction des mots à partir d'un schème.

- **Les radicales « الجذوع » :**

Un radical est la dérivation obtenue à partir d'une racine donnée selon un modèle. L'Arabe classique a un grand nombre des Stems qui ne sont pas tous utilisables, 2%Seulementsont utilisables [12]. 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 « ل », « ع », « ف » qui sont les lettres de la racine de base « الفعل, faire » [13].

Racine	Schème	Radicale	Utilisable ?
كتب	فَعَلَ	كَتَبَ	Oui
رسم	فَاعِلٌ	رَاسِمٌ	Oui
وضع	مَفْعُولٌ	مَوْضُوعٌ	Oui

لعب	فَعْلاَءٌ	لَعْبَاءُ	Non

Tableau1.11: Exemple de génération des radicales.

1.3.2 Catégories des mots

L'Arabe considère 3 catégories de mots

- *Le verbe :* entité exprimant un sens dépendant du temps, c'est un élément fondamental auquel se rattachent directement ou indirectement les divers mots qui constituent l'ensemble.
- *Le nom :* l'élément désignant un être ou un objet qui exprime un sens indépendant du temps.
- *Les particules :* entités qui servent à situer les événements et les objets par rapport au temps et l'espace, et permettent un enchaînement cohérent du texte [2].

Verbe :

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» (dahrağa – glisser, faire glisser). 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) [17].

- **Verbe à racine simple :**

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

- *Le verbe sain (سالم, Sālim) :* qui ne contient aucune lettre radicale défectueuse, ni lettre hamza, ni lettre redoublée.

- *Le verbe de lettre Alif (مهموز, Mahmuz) :* qui contient une lettre radical hamza comme : « كلأ ».
- *Le verbe redoublé (*مضاعف*,mudaâ'if) : la* présence de deux consonnes identiques dans la deuxième et troisième position du radical de verbe nus trilitère et son augmenté comme : « شدُّ »ou la première et la troisième lettre identique dans le verbe quadrilatère comme : « زلزل ».

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

***Verbe Ramas (لفيف, lafif)* :**

Il contient deux longues voyelles au même temps, il est divisé en deux selon leur position :

- *Ramas séparé (لفيف المفروق, LafifMafruwk) :* la première et la troisième consonne sont des voyelles longues, comme « وفى ».
- *Ramas collé (لفيف مقرون, Lafifmakruwn) :* la deuxième et la troisième consonne sont des voyelles longues, comme « طوى ».

Verbe incomplet (ناقص, naâqis) :

La troisième consonne est une longue voyelle, il est nommé comme ça parce que dans leur conjugaison on supprime cette lettre comme : « رمى ».

- ***Verbe creux (أجوف, Ajwaf) :***

  La deuxième consonne est une longue voyelle, il est nommé comme ça parce que leur cavité est vidée d'une lettre saine ; par exemple : « قال ».

- ***Verbe assimilé (مثال, mithal) :***

  La première consonne est une longue voyelle, il est nommé comme ça parce qu'il a assimilé le verbe sain dans leur conjugaison au passé. Exemple : « صلو » [7].

- **Verbe à racine augmentée :**

Ce type de verbe est obtenu, comme indiqué ci-dessus, par des opérations morphologiques appliquées à des racines simples afin de donner un sens particulier. Il

existe différentes opérations utilisées, mais au final ces opérations intègrent une ou plusieurs lettres de l'ensemble rassemblé dans le mot (سَأَلْمُتُونِيهَا -saaltemouniha). Parmi les fonctions morphologiques utilisées, nous citons [1] :

- **le redoublement :** qui consiste généralement à redoubler la deuxième consonne radicale du verbe, les verbes obtenus suivent le schème « فَعَّلَ » *(fa''ala)*.
- **l'allongement :** cette opération est réalisée par l'ajout du glide " ا" (alif) à la première consonne radicale, ce qui donne le nouveau schème « فَاعَلَ » *(faâ'ala)*.
- **l'adjonction :** cette opération permet d'ajouter une ou plusieurs lettres à la racine radicale dans des positions différentes tel que :
  - adjonction d'un morphème des trois consonnes " اِسْتَ " *(ista)* au début de la racine radicale du verbe. Cette opération donne naissance à nouveau schème qui a la forme « اِسْتَفْعَلَ » *(istaf 'ala)*.
  - adjonction du glide " ا" *(alif)* au début de la racine radicale et l'ajout du morphème consonantique " ت" *(t)* après la première consonne, les verbes. obtenus suivent le nouveau schème « اِفْتَعَلَ » *(ifta'ala)*.
  - adjonction du morphème consonantique " ت" *(t)* pour les verbes à racine quadratique (racine de quatre lettres) donnant le schème « تَفَعْلَلَ » *(tafa'lala)*.
  - adjonction du glide "ا" *(alif)* au début de la racine quadratique et l'ajout du morphème consonantique " ن" *(n)* après la deuxième consonne. Cette opération morphologique produit le schème « اِفْعَنْلَلَ » *(if 'alala)* [1].

- **Flexions des verbes (conjugaison) :**

Il existe trois modes en Arabe pour la conjugaison des verbes :

- ***L'accompli (الماضي)* :** indique que l'action est achevée ce qui est impliqué le passé **.**C'est l'aspect le plus simple qui est utilisé avec la troisième personne du singulier pour représenter un verbe à l'infinitif, il se caractérise par une suffixation des marques [17].Le tableau suivant représenté l'ensemble des suffixes utilisés de manière générale, en utilisant le verbe ' شرب ' comme suivant :

' أَنَا'	' نَحْنُ'	' أَنْتَ'	' أَنْتِ'	' أَنْتُمَا'	' أَنْتُمْ'	' أَنْتُنَّ'	' هُو'	' هِيَ'	' هُمَا'	' هُمْ'	' هُنَّ'
(je)	**(nous)**	**(tu)**	**(tu)**	**(vous2)**	**(vous)**	**(vous)**	**(il)**	**(elle)**	**(ils2)**	**(ils)**	**(elles)**

شربتُ	شربنَا	شربتَ	شربتِ	شربتُمَا	شربتُمْ	شربتُنْ	شربَ	شربتِ	شربَا	شربُوا	شربْنَ

Tableau1.12: L'ensemble des suffixes de verbe ' شَرَبَ ***'***

- ***L'inaccompli (المضارع) : indique*** que l'action est en train de se réaliser, ce qui est impliqué le présent**.** Il permet la modification des lettres principales du verbe par une préfixation de ces éléments avec les lettres (أَنَيْتُ, *Anyt*) ainsi des infixations sous forme de duplication de lettre dans le cas de verbe redoublé ou de substitution d'une voyelle dans le cas d'un verbe défectueux. Par exemple ' شدّ ' (*$ada ~, se* souquer) se conjuguer avec le pronom elles par « شَددن » (*$edodena-* elles souquent) [7].

- L'inaccompli se caractérise par quatre modes flexionnelles :

  - ***L'inaccompli indicatif (مرفوع) :*** employé dans une proposition principale ou isolée. Il se caractérise par une désinence (ضَمَة) **[***dammat***]** et par des flexions longues [7].

  - ***L'inaccompli subjonctif (منصوب) :*** utilisé en proposition subordonnée s'il est précédé par une particule de subjonctif, il se caractérise par une désinence (فتحة**) [***fathat***]** et par des flexions courtes [7].

  - ***L'inaccompli apocopé (مجزوم) :*** il précède par une particule d'apocopé, Il se caractérise par l'absence de désinence (سكون**) [***sukun***]** et par des flexions courtes [7].

  - ***L'inaccompli futur :*** correspond à une action qui se déroulera au futur et est marqué par l'ajout de la lettre « س » « *sa* » ou de la particule « سوف » «*sawfa* » au début du verbe conjugué à l'inaccompli indicatif. Par exemple, pour le verbe « كتب » « *kataba,* écrire » nous obtenons « سيكتب, sayaktubu » pour « il écrira » ou « سوف يكتب, *sawfa yaktubu* » qui signifie « il va écrire » [1].

- ***L'impératif :*** il est utilisé pour exprimer un ordre, donner un conseil ou faire une suggestion ou une recommandation. Ce paradigme ne se conjugue qu'à la 2ème personne au singulier, duel et pluriel. La voyelle finale /i/ caractérise l'impératif (est structuré sur le soukoun) ou sur l'élimination du noun et de la lettre défectueuse du verbe non sain. Dans le tableau suivant nous donnons un exemple de conjugaison pour le verbe « كَتَبَ» (*kataba*', écrire) [1].

أَنْتَ (tu)	أَنْتِ (tu)	أَنْتُمَا (vous-2)	أَنْتُمْ (vous)	أَنْتُنَّ (vous)
أَكْتُبْ	أَكْتُبِي	أَكْتُبَا	أَكْتُبُوا	أَكْتُبْنَ

Tableau1.13: Conjugaison pour le verbe 'écrire' kataba', 'كَتَبَ' [1].

Nom :

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 [18] :

- ***Les primitifs :*** 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 :رأس (raás – tête)كُرسي (kursiyy – siège)كبش (kabš – bélier) ….etc. [18].

- ***Les dérivés :*** 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 [20]. Puisque ce sont des noms, alors ils peuvent recevoir les marques du genre (masculin et féminin), du nombre, d'outil, etc. Cette classe des noms peut contenir :

  - Les noms d'agent «أسماء الفاعِل » par exemple : كاتِب.
  - Les noms de patient «أسماء المفعول » par exemple : مكتوب.
  - Les noms adjectifs «الصفة » par exemple :جميلة .
  - Les noms d'outil « اسم الآلة» par exemple : ملعقة.
  - Les noms du nombre «العدد» par exemple : واحد.
  - les Noms de temps et de lieu « اسما الزمان والمكان » par exemple :مكتب .

- ***Les nombres :*** ce sont les numéros simples représentant les unités (de « صفر_sifr : Zéro » à « تسعة» tis'at_ neuf), les dizaines (عشرون_ishruwn : vingt) et les centaines (مئة : cent), etc. ; et les numéros composés comme les cardinaux, par exemple (ستة عشر– seize) [1].

Flexions des noms :

La déclinaison des noms en Arabe est concrétisée en trois principaux cas :

Nominatif (مَرْفُوع- *marfu3)*, accusatif *(*مَنْصُوب- *mansub)*, génitif *(*مَجْرُور- *majrur)*. Ces déclinaisons sont faites en fonction du rôle du mot dans la phrase, à l'exception de certains cas particuliers. Les noms qui sont déclinable en Arabe sont dits معربة (*mu'araba*).

Dans la suite de cette partie, nous classons la flexion des noms en trois catégories selon le nombre de la forme comme suit [1] :

- Les déclinaisons au singulier

Déclinaison de base à trois cas :

- C'est le cas le plus fréquent, il prend la voyelle « ضمَّة-(*dammat*-'u') » comme une marque du nominatif, la « فَتْحَة (*fathat*-'a')» à l'accusatif et la « كَسْرَة-(*kasrat*-i) au génitif. Quand le nom est indéfini le *tanwin* apparait marqué respectivement par les trois signes diacritiques : **«ٌ » (ũ_un)**, **«ً » (ã_an)** et **« ٍ » (ĩ_in).** A l'accusatif indéfini, excepté le cas des noms qui se terminent par " ة" (at) ou par " اء" (ā'), un álif « ا »(ā) vient renforcer le tanwîn "ً " (an) : par exemple, à l'accusatif indéfini, le nom « كِتَاب »(kitaāb – livre) produit « كِتَابًا »(kitaābãā–livre à l'accusatif indéfini) et le nom «جزيرة»ğaziyrat – île) produit« جَزِيرَةً»(ğaziyratã – île à l'accusatif indéfini)[18].
- *Le diptote (*الممنوع من الصرف- *al-mamnu' min aṣ-ṣarf) :* Les diptotes sont les noms qui, indéfinis grammaticalement, n'acceptent pas de tanwîn et prennent la même marque à l'accusatif et le génitif, soit la "َ " (فَتحة –*fathat* – a). Par contre, quand ils sont définis, ils suivent la déclinaison de base à trois cas. C'est le cas des noms féminins qui se terminent par " اء" (ā') tel que " صحراء" (sahraā' – désert), les adjectifs masculins de couleurs ayant pour schème " أفعل" (af'al) tel que " أحمر" (áhmar – rouge) et ceux qui sont féminins de schème " فعلاء" (fa'laā') tel que " بيضاء" (baydaā' – blanche) [18].
- *Déclinaison des cinq noms :* Sont des noms bilitères qui leur voyelle finale se prolonger quand ils sont définis par un complément : أب (père), أخ (frère), حم (beau-père), فو (bouche), ذو (possesseur) [7].

Indéfini	Annexion	Indéfini	défini par l'article	Annexion
أَبٌ	أَبُوبَكْرٍ	أَخٌ	الأَخُ	أَخُو مُحَمَّدٍ
أَبًا	أَبَابَكْرٍ	أَخًا	الأَخَ	أَخَا مُحَمَّدٍ
أَبٍ	أَبِي بَكْرٍ	أَخٍ	الأَخِ	أَخِي مُحَمَّدٍ

Tableau1.14 : Déclinaison des cinq noms pour أب (père), أخ (frère) [1].

- Les déclinaisons au duel

La forme du duel va dépendre de la nature du mot, car en langue Arabe un mot peut être nominatif, accusatif ou génitif. En règle générale le duel va se former de la façon suivante : il faut rajoute le suffixe alif noun « انِ » ou « يْنِ » tout dépendra de sa nature, à la fin du mot au singulier, pour avoir le duel.

- Au cas nominatif : Dans le cas où le mot à une place nominative dans la phrase le duel se formera de cette façon : Par exemple avec le mot masculin kitaboun, en Arabe et au singulier كيتبٌ, nous allons rajouter انِ ce qui va nous donner au duel kitabaniكتيبان.Quand le mot sera au féminin et au singulier, par exemple طَالِبَةٌ alors on transformera sa dernière lettre en ت et en rajouter انِ ce qui nous donnera طَالِبَتَانِ.

- Au cas génitif et accusatif : Par contre dans le cas où il sera plutôt génitif ou accusatif on ajoutera le suffixe يْنِ pour former le duel.

- Au masculin *:* Prenons l'exemple au masculin pour commencer avec le mot, étudiant en Arabeطالب.

- Au cas accusatif singulier on écrira : رأيت طالبًا, j'ai vu un élève. Toujours au cas accusatif mais au duel on écrira رأيت طالبَيْنِ j'ai vu deux élèves.

- Au cas génitif singulier on dira : سلمت على طالبٍ, j'ai salué un élève. Et au cas génitif duel nous dirons : سلمت على طالبَيْنِ, j'ai salué deux élèves.

Au féminin : nous allons utiliser le même mot étudiant en Arabeطالبة .Alors ce mot au cas accusatif singulier on dira : رَأَيْتُ طالبةً, j'ai vu un élève. Et au duel dans le même cas accusatif on dira : رأيت طالبتَيْنِ, j'ai vu deux élèves. Ensuite au cas génitif nous dirons : سَلَّمْتُ على طالبةٍ j'ai salué une élève. Et au duel cas génitif : سلمت على طالبتَيْنِ, j'ai salué deux élèves [30].

- Pour le mot se terminant par l'un des glides suivants ((ا) (alif – â), ' و' (wâw- w) et ((ي (yâ' - y), des transformations seront appliquées sur ces glides pour

obtenir le duel avec les suffixes décrit ci-dessus. Par exemple, le duel du mot مَلْهَى *(malha)* est obtenu en transformant la lettre (ى) « alif maksoura » en (ي-yaa) ensuite nous ajoutons les suffixations pour obtenir مَلْهَيَان (cas nominatif) et مَلْهَيَيْن pour les autres cas. Et pour le mot عصا (*'asaâ*– bâton) nous remplaçons d'abord le glide ' ا' par ' و' (wâw - w) ensuite nous ajoutons les suffixes ce qui donne pour le cas nominatif le mot عَصَوَان *('assawAn)* et pour les autres cas le mot عَصَوَيْن *('assawayn)* [1].

- Les déclinaisons au pluriel

Il existe deux grandes types de pluriels nous présentons comme suivants :

- Le pluriel externe ou régulier : Les pluriels externes sont formés par l'ajout d'un suffixe au singulier sans changement de la structure du mot. Nous distinguons [18] :

- Le pluriel externe masculin «الجمع المذكر السالم» : La flexion est réalisée par l'addition du suffixe 'ون' (uwna) dans le cas nominatif et du suffixe'ين'(iyna) dans les cas accusatif et génitif. Par ailleurs, nous notons que si le mot est défini par annexion, nous supprimons la lettre 'ن' (noun) dans tous les cas. Dans le tableau suivant nous exhibons quelques exemples de ce type de flexion [1].

المعرفة Défini	معرف بالإضافة Défini par annexion	النكرة Indéfini	الجمع المذكر السالم
المُعَلّمُونَ	مُعَلّمُو الرّياضيات	مُعَلّمُونَ	Nominatif
المُعَلّمِينَ	مُعَلّمِي الرّياضيات	مُعَلّمِينَ	Accusatif & génitif

Tableau1.15: Exemple sur Le pluriel externe masculin [1].

- Le pluriel externe féminin «الجمع المؤنث السالم » : Le pluriel externe féminin : De la même manière, nous rajoutons pour le pluriel féminin le morphème "ات"(āt) «سيارة » (*sayArah*- voiture) devient « سَيَّارَات » (*sayyaāraāt* _voitures des) [18].

المعرفة Défini	النكرة Indéfini	الجمع المؤنث السالم
السَيَّاراتُ	سَيَّاراتٌ	Nominatif
السَيَّاراتِ	سَيَّاراتِ	Accusatif & génitif

Tableau1.16: Exemple sur Le pluriel externe féminin [1].

- Le pluriel interne ou brisé «جمع التكسير» : La forme du nom au pluriel se différencier de leur singulier par infixation, ou par diminution de son origine, et se classer en deux groupes :
- Le pluriel de petit nombre : indique que le nombre de pluriel est entre 3 et 10 comme : « أحمال » (*AHemaAle*, Charges), ses schèmes est quatre : «أَفْعُل, فِعْلَة،أَفْعَال،أَفْعِلَة » (aaf-‘ilah, aaf-‘aal, fi‘-lah, aaf-‘ul).
- Le pluriel collectif : caractérise un nombre supérieur de 3 à l’infinité comme : « حمول » (*Humuwlun*, Charges).Il existe 16 schèmes pour ce type [7].

Les particules :

Sont des entités ou des particules clés employées pour situer des objets et des faits par rapport au temps et à l’espace et assurer ainsi un enchainement cohérent du texte. De plus, ils constituent des éléments importants dans l’interprétation du sens d’une phrase. Nous distinguons plusieurs types de ces mots, comme introduction, explication, conséquence, en fonction de leur sémantique et rôle dans la phrase. Parmi ces mots nous citons à titre d’exemple les éléments suivants [1] :

- Préposition **:** exemple (الى, على, عن, من, حتى).
- Particules de coordination : exemple (و, ف, ثم, او).
- Particules interrogatives : exemple (ما, هل, أ).
- Particules d’affirmation : exemple (نعم, اجل, بلى).
- Particules de négation : exemple (لا, لن, لم).
- Particules distinctive : exemple (اي).
- Particules relatives : exemple (ما).
- Particules de futur : exemple (لن, ان, س).
- Particules conditionnelles : exemple (كيفما, من, اذما) [11].
- Les conjonctions de subordination :(بَيْنَمَا, حَيْثُمَا).
- Les quantificateurs : exemple (كُلّ, بَعْض).
- Les adverbes : exemple (أَبَدًا, أَخِيرًا) [1].

- **Pronoms :** En Arabe, les noms invariables sont appelés des pronoms, et ils possèdent une structure et une flexion uniques quelle que soit leur place dans la phrase. Ils contiennent un type particulier et jouent une fonction syntaxique précise dans la langue [1]. Nous citons entre autres les types suivants :

  - **Les pronoms personnels :** sont des noms utilisés pour remplacer un nom ou désigner une personne ou un objet qu'ils soient absents, auditeurs ou locuteurs. Dans notre travail, nous étudions les pronoms personnels isolés [1], comme :أَنَا, نْحَن , أَنْتَ, أَنْتُمْ ….etc. et collés comme : Ha' El gha'yba (هاء الغائبة) et Kaf el khitab (كاف الخطاب), Ya' El Mota alim, Noun El motakallimine (نون المتكلمين).

  - **Les pronoms démonstratifs (أسماء الإشارة_asmaâ' ichaâra) :**

    Ces pronoms sont utilisés pour indiquer une ou plusieurs entités. Cette indication permet de situer l'entité dans l'espace, le temps ou tout simplement dans les textes et cela en fonction du contexte de la phrase [1], et elle comme suite :هذا , هُنَا, هؤلَاء, هُنَاك, ذاك, تلك, أولئك….etc.

1.4 Les problèmes d'analyse du traitement automatique de la langue Arabe

L'Arabe, comme toutes les langues naturelles, est caractérisée par un ensemble de phénomènes créant des difficultés et des problèmes qu'il faut prendre en considération lors d'un traitement automatique comme l'ambiguïté, l'absence de voyelles ,l'ordre des mots dans la phrase …etc., Dans la présente section, nous présentons les phénomènes que nous considérons les plus importants pour l'Arabe[1].

1.4.1 L'absence de voyelles

Le problème de la voyellation réside dans l'absence quasi systématique de la voyellation dans les textes Arabes. En effet, les signes de voyellation, lorsqu'ils sont notés, sous la forme de signes diacritiques placés au-dessus ou au-dessous des lettres, apparaissent dans certains textes (Coran, hadith) ou littéraires (poésie classique, notamment) : on dit qu'ils sont édités en graphie voyellée [7].

L'absence de voyelles (la non-voyellation) dans les textes Arabes génère plusieurs cas d'ambiguïtés et des problèmes lors de l'analyse automatique. En effet, l'ambiguïté grammaticale augmente si le mot est non voyellé. Cela est dû au fait qu'un mot non voyellé possède plusieurs voyellations possibles, et pour chaque voyellation est associée une liste différente de catégories grammaticales.

L'exemple suivant du mot non-voyellé ktb | كتب possède 16 voyellations potentielles et qui représentent 8 catégories grammaticales différentes [7] :

Mot voyellé	Pré-notion	Notion d'écrire
كَتَبَ	Kataba	Il a écrit
كُتِبَ	Kutiba	Il a été écrit
كَتب	Kutub	Des livres
كَتَّب	Katob	Un écrit
كَتَّبَ	Kattaba	Il a fait écrire
كُتِّبَ	Kuttiba	Faire écrire – forme factitive
كَتِّبُ	Kattibo	Fais écrire
كَتَبَّ	Katabba	Comme trancher

Tableau1.17: Exemple de voyellation de mot non-voyellé ktb | كتب [7].

1.4.2 Agglutination

Contrairement aux langues latines, en Arabe, les articles, les prépositions, les pronoms, etc. collent aux adjectifs, noms, verbes et particules auxquels ils se rapportent. Comparé au français, un mot Arabe peut parfois correspondre à une phrase française.

Exemple : le mot Arabe « أتتذكروننا » correspond en Français à la phrase "Est-ce que vous vous souvenez de nous ?".

Cette caractéristique peut engendrer une ambiguïté au niveau morphologique. En effet, il est parfois difficile de distinguer entre une proclitique ou enclitique et un caractère original du mot. Par exemple, le caractère "و" dans le mot "وصل" (il est arrivé) est un caractère original alors que dans le mot « وفتح » (il a ouvert), il s'agit d'une proclitique [26].

1.4.3 Irrégularité de l'ordre des mots dans la phrase

L'ordre des mots en Arabe est relativement libre. D'une manière générale, on met au début de la phrase le mot sur lequel on veut attirer l'attention et l'on termine sur le terme le plus long ou le plus riche en sens ou en sonorité. Cet ordre provoque des ambiguïtés syntaxiques artificielles, dans la mesure où il faut prévoir dans la grammaire toutes les règles de combinaisons possibles d'inversion de l'ordre des mots dans la phrase .Ainsi par exemple, on peut changer l'ordre des mots dans la phrase (Figure 1.4) pour obtenir deux phrases ayant le même sens [21].

l'ordre des mots dans la phrase
Est allé le garçon à l'école - ذهب الولدُ الى المدرسة
Le garçon est allé à l'école - الولد ذهب الى المدرسة
A l'école est allé le garçon – الى المدرسة ذهب الولد

Tableau1.18: Exemple de combinaisons possibles d'inversion de l'ordre des mots dans la phrase.

1.4.4 Mots étrangers translittérés en Arabe

Les translittérations en Arabe de mots étrangers posent un problème, puisqu'ils n'ont pas de racine en Arabe. Les mots translittérés sont considérés comme inconnus par l'analyseur.

Quelques items étrangers méritent une attention particulière en raison de leurs fréquences élevées. Exemple :أورو , دولار…etc. [7].

1.4.5 La segmentation de textes

La segmentation d'un texte est une étape fondamentale pour son traitement automatique ; son rôle est de découper le texte en unités d'un certain type qu'on aura défini et repéré préalablement. En effet, l'opération de segmentation d'un texte consiste à délimiter les segments de ses éléments de base qui sont les caractères, en éléments constituants différents niveaux structurels tels que : paragraphe, phrase, syntagme, mot graphique, mot-forme, morphème, etc.

Toutefois, les particularités de la langue Arabe, rend la segmentation Arabe toujours différente, il n'y a pas de majuscules qui marquent le début d'une nouvelle phrase. De plus, les signes de ponctuation, ne sont pas utilisés de façon régulière. Certaines particules

comme "et | و ", "donc | ف", etc. jouent un rôle principal dans la séparation de phrases et peuvent être déterminantes pour guider la segmentation [7].

1.4.6 Segmentation de phrase

La reconnaissance de la fin de phrase est délicate car la ponctuation n'est pas systématique et parfois les particules délimitent les phrases.

Pour la segmentation de texte utilise :

- Une segmentation morphologique basée sur la ponctuation.
- Une segmentation basée sur la reconnaissance de marqueurs morphosyntaxiques ou des mots fonctionnels comme : أي، لكن ،أو ، و، حتى *ou, et, c.à.d. mais, quand.*

Cependant, ces particules peuvent jouer un autre rôle que celui de séparer les phrases [23].

1.5 Conclusion

Après avoir vu les particularités de la langue Arabe, nous avons compris les difficultés rencontrées par les chercheurs dans le domaine du traitement automatique pour la langue Arabe Ce qui a conduit à l'absence de programmes informatiques complet à 100% pour l'utiliser dans différent applications et surtout le domaine TAL (traitement automatique de la langue). Pour le chapitre suivant, nous allons en apprendre davantage sur l'un de ces programmes appelé lemmatisation.

Chapitre 2
Les différentes méthodes de Lemmatisation

2.1 Introduction

Nous avons présenté dans le premier chapitre toutes les spécificités de la langue Arabe dans les différents niveaux d'analyse : morphologique, syntaxique, sémantique ainsi que les problèmes d'analyse du traitement automatique de la langue Arabe parmi lesquels l'extraction des racines ou lemmatisation qui présente l'un des traitements les plus importants pour la langue Arabe. Dans ce chapitre nous allons expliquer les notions de lemmatisations et quelques difficultés rencontrées ainsi que les différentes méthodes de l'extraction de racines.

2.2 Le lemme

Un lemme fait référence à un mot, de quelque nature que ce soit, qu'il soit composé ou simple, à condition qu'il puisse être référencé dans un dictionnaire. Un lemme est composé d'un radical, auquel peuvent s'ajouter un préfixe ou un suffixe [40].

Chaque mot est rapporté à son lemme qui est sa forme canonique qui dépend toujours de la catégorie grammaticale de ce mot, si c'est un nom il doit être au singulier et si c'est un verbe il doit être à la troisième personne du singulier.

2.3 Difficultés de la lemmatisation des mots Arabes

La complexité morphologique de la langue Arabe rend particulièrement difficile de développer des applications pour le traitement automatique en langue naturelle. Dans les langues sémitiques comme l'Arabe, la plupart des lemmes de noms, d'adjectifs et de verbes sont dérivés de quelques milliers de racines par l'insertion de nouvelles lettres, par

exemple, les mots («مكتبة»,bibliothèque), («كتاب», livre),(« كتب,keteb »,livres), (« كتبkataba »,il a écrit), et ((«نكتب» ,naketebou), nous écrivons), de la racine «كتب»[25].

Aussi, la langue Arabe est fortement productive, dérivationnelle et flexionnelle. Les articles définis, les conjonctions, les particules et d'autres préfixes peuvent être attachés au début d'un mot et un grand nombre de suffixes peuvent être attachés à la fin d'un mot. Un mot-clé donné peut être trouvé sous différentes formes. Les analyses du texte Arabe journalistique prouvent qu'il existe une variabilité lexicologique en Arabe plus que dans les langues européennes. Écrite de droite à gauche, la langue Arabe peut aussi être écrite avec ou sans les signes diacritique suivant le contexte de l'apparence des caractères, cette manière d'écriture orthographique contribue également à la variabilité qui peut brouiller les systèmes de traitement. Par exemple, les deux mots « تَبك » et « كتب » (kataba et tbk) semblent à priori identiques, ce qui n'est pas le même cas pour l'ordinateur, la différence n'apparaît pas [25].

Les textes Arabes diacritisés sont moins ambigu, tels que les livres d'enfants, les dictionnaires et le Quran mais ces signes diacritiques ne sont pas généralement inclus dans tous type de textes tels que les journaux. Pour cette raison, une normalisation comme la suppression des signes diacritiques, est typiquement représentée dans les systèmes de traitement.

Aussi, la forme plurielle des noms peut brouiller les systèmes de Recherche d'Information(RI). Dans ce cas-ci, un nom au pluriel prend une autre forme morphologique différente de sa forme initiale du singulier. Par exemple, le pluriel du mot «امرأة» (femme) est«نسوة» [25].

Si un mot qui commence par un préfixe possible, Nous devons vérifier si ce dernier ne fait pas partie du mot avant de le tronquer (décomposer). Par exemple : le mot "كامل", commence par le préfixe possible"كا" qui est une partie du mot d'origine. Retirer ce préfixe donne un faux mot "مل".Donc, nous devons le retourner tel qu'il est sans le décomposer [25].

2.4 Les différentes méthodes de lemmatisation

Dans cette partie, nous présentons quelques méthodes d'extraction de la racine d'un mot Arabe. Ces méthodes sont citées et réparties en trois grandes classes : les méthodes basées sur l'analyse statistique, les méthodes basées sur l'analyse morphologique et les méthodes hybrides (basées sur l'analyse morphologique et statistique) [13]. La figure 2.1 présente une classification générale des méthodes d'extraction de la racine d'un mot Arabe à trois niveaux.

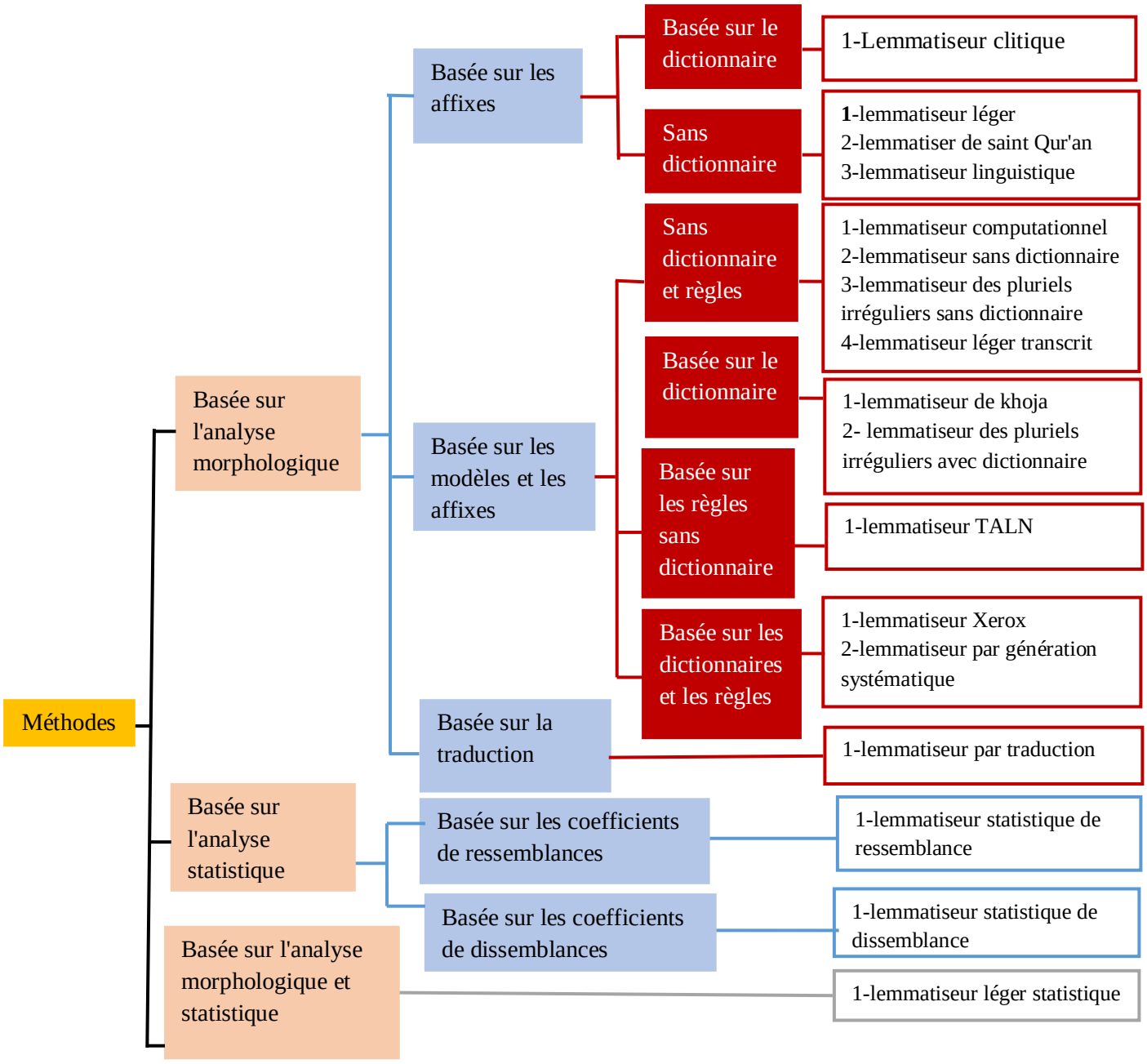


Figure2.1: Schéma générale de la classification des méthodes d'extraction de la racine en Arabe [13]

2.4.1 Classe de l'analyse morphologique

Les mots en Arabes démontrent une morphologie complexe, ils utilisent la morphologie racine-schème. La technique d'analyse morphologique est basée sur l'idée de la conformation du mot à des schèmes pour trouver la racine du mot. La racine est extraite après avoir retiré les affixes attachés à un mot donné sans les signes diacritiques. Parfois, cette technique pose quelques problèmes tels que la difficulté d'extraire quelques racines. Technique de traduction ; consiste à traduire le mot en anglais, à déterminer son lemme puis le retraduire dans sa langue d'origine. Cette technique pose des problèmes de traductions ambiguës [18].

2.4.2 Basée sur les affixes

Dans cette catégorie, les méthodes repèrent les différents types d'affixes du mot à traiter et ensuite les suppriment afin d'extraire la racine probable. Deux approches sont mises en place, la première utilise un dictionnaire pour valider la racine probable. Et la deuxième considère la racine extraite comme finale [13].

2.4.3 Basée sur les affixes avec dictionnaire

Dans cette approche, on trouve la méthode : lemmatiseur clitique. Cette méthode est basée sur les ressources linguistiques suivantes : un dictionnaire principal (5,4 millions d'entrées) contenant des mots avec leurs caractéristiques (genre, nombre, etc.), un dictionnaire de proclitiques (77 entrées) et un dictionnaire d'enclitiques (65entrées). L'extraction de la racine selon cette méthode se fait en quatre étapes. La première est la normalisation orthographique, la deuxième sert à éliminer les clitiques, en utilisant les dictionnaires de proclitiques et d'enclitiques, la troisième étape consiste à vérifier si le radical obtenu existe dans le dictionnaire principal, s'il existe, alors la racine sera ce radical obtenu, sinon, il faut appliquer les règles de réécriture sur le radical obtenu et vérifier s'il existe dans le dictionnaire principal [27], la Figure 2.2 explique cette méthode.

2.4.4 Basée sur les affixes sans dictionnaire

Dans cette approche, plusieurs méthodes sont proposées : le lemmatiseur léger, le lemmatiseur du saint Qur'an et le lemmatiseur linguistique [13].

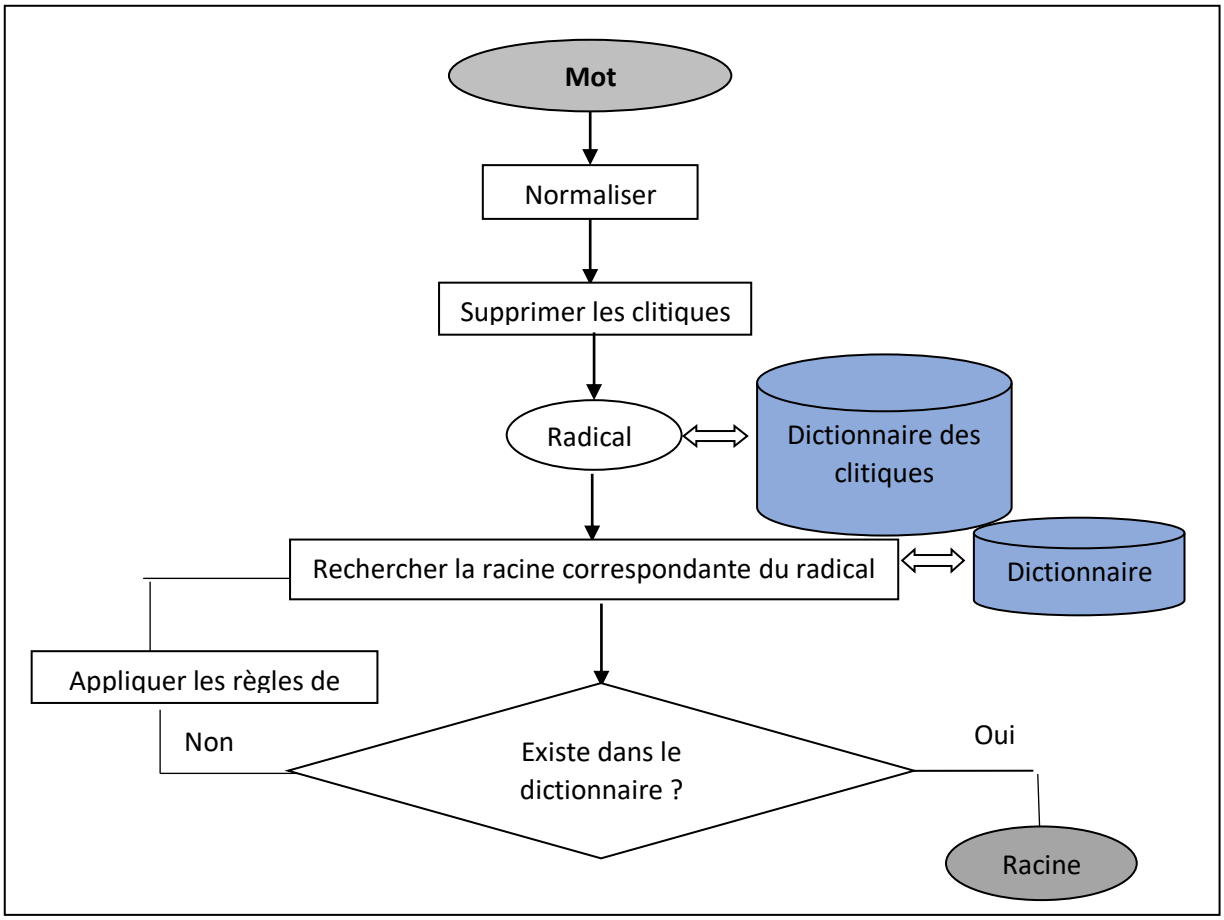


Figure2.2: Lemmatiseur clitique [13].

Lemmatiseur léger :

L'approche de suppression des affixes s'appelle généralement la lemmatisation assouplie ou légère «light stemming», quand elle est appliquée à la langue Arabe, elle se réfère à un processus de suppression d'un petit ensemble de préfixes et de suffixes, sans essayer de traiter les infixes, ou d'identifier les modèles (schèmes,اوزان) et de trouver les racines, dans la figure ci-dessous montrée ça [25].

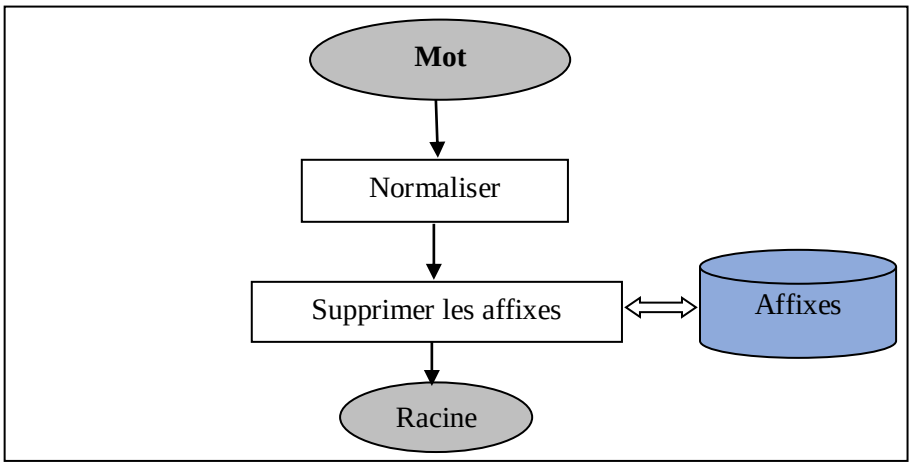


Figure2.3: Lemmatiseur léger [13].

Cette approche conçue par suppression des chaînes de caractères fréquemment trouvées comme préfixes ou suffixes [13].

Lemmatiseur du Coran :

Le Coran (القرآن al Qurān) est le texte sacré de l'islam. Il a une morphologie et une grammaire qui est différente de la langue Arabe standard a présenté une nouvelle méthode basée sur la méthode Lemmatiseur léger. Cette méthode a été appliquée sur une transcription du Coran. Elle élimine les préfixes et les suffixes qui sont distribués en six groupes selon leurs longueurs, pour extraire la racine transcrite. Ensuite, cette racine extraite est retranscrite en écriture Arabe [27], et la Figure 2.4 présente les étapes cette méthode.

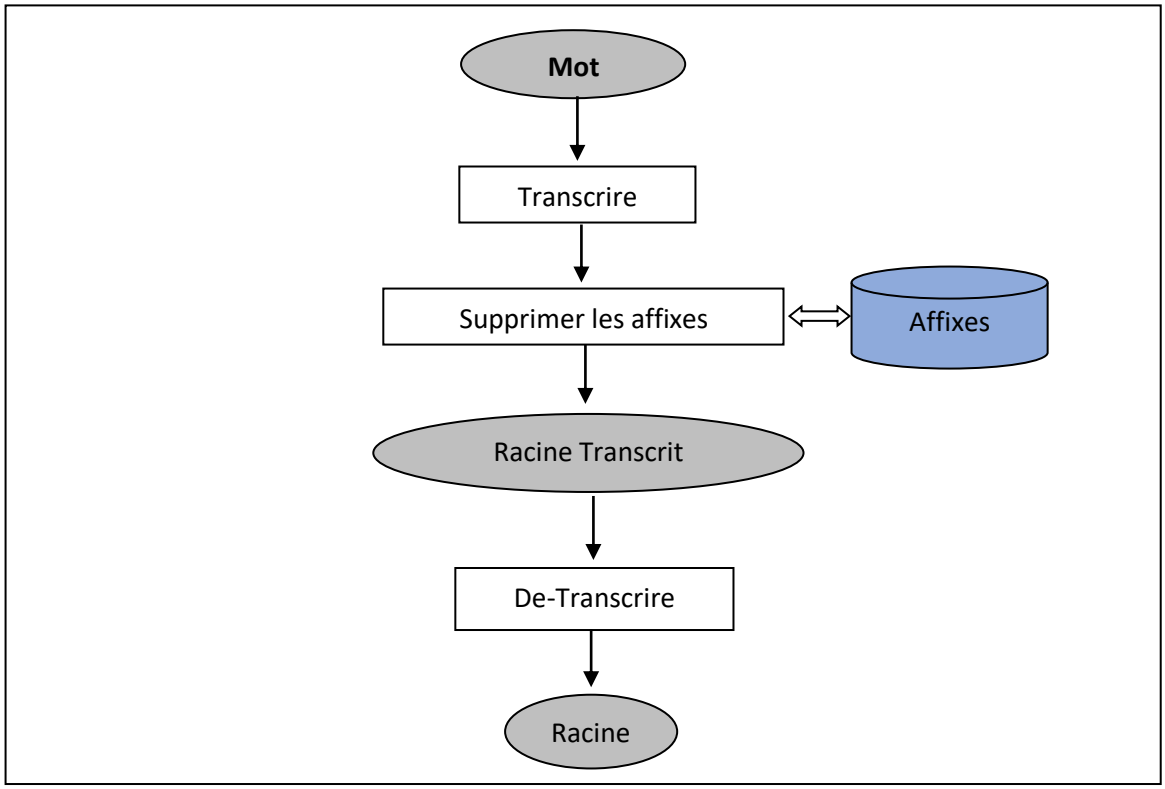


Figure2.4: Lemmatiseur du Saint Qur'an [13].

Lemmatiseur linguistique :

Cette méthode est basée sur la méthode lemmatiseur légèr. Les affixes sont divisés en 4 catégories : les antéfixes (qui sont avant les préfixes), les préfixes, les suffixes et les post-fixes (qui sont après les suffixes). Elle élimine d'abord les anté-fixes et les post-fixes, puis applique la méthode lemmatiseur légèr [27].

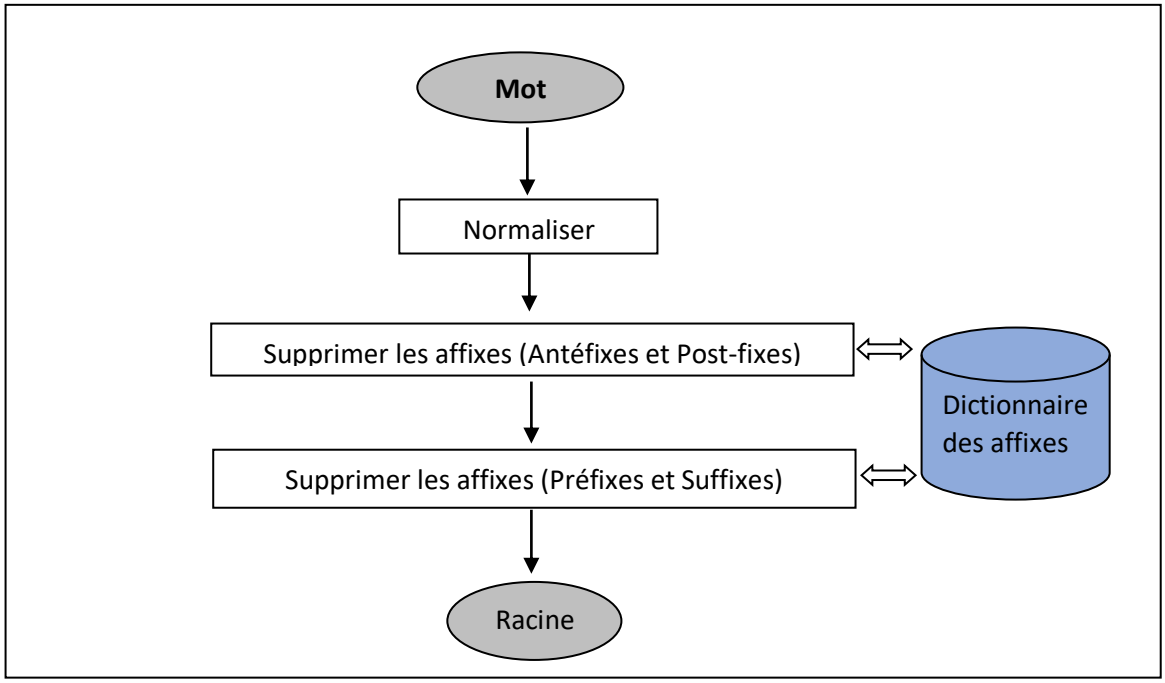


Figure2.5: Lemmatiseur linguistique [13].

2.4.5 Basée sur les modèles des affixes

Dans cette catégorie, Il y a quatre approches. La première consiste à traiter les affixes et rechercher les modèles correspondants. La deuxième consiste à utiliser un dictionnaire pour valider la racine extraite de la même façon que l'approche précédente. La troisième consiste à appliquer des règles spécifiques et prédéfinies sur la racine extraite de la même façon que la première. La quatrième consiste à générer dans une phase préalable, un dictionnaire global qui contient la plupart des mots Arabes et de leurs racines. Ensuite, pour extraire la racine d'un mot donné, une simple recherche est effectuée dans le dictionnaire ainsi construit [13].

Basée sur les modèles des affixes sans dictionnaire :
Dans cette approche, plusieurs méthodes sont proposées :

- **Lemmatiseur computationnel :**

Cette méthode sert à extraire seulement les racines trilittérales (composées de 3 lettres). La première étape de cette méthode est la suppression des préfixes les plus longs, la partie restante du mot (4 à 5 lettres) doit nécessairement contenir les trois lettres de la racine.

L'étape suivante consiste à extraire la racine en comparant tous les trigrammes possibles, dans les lettres restantes du mot, avec les schèmes sauvegardés dans une base de données [27], la figure suivant présente les étapes ci-dessus.

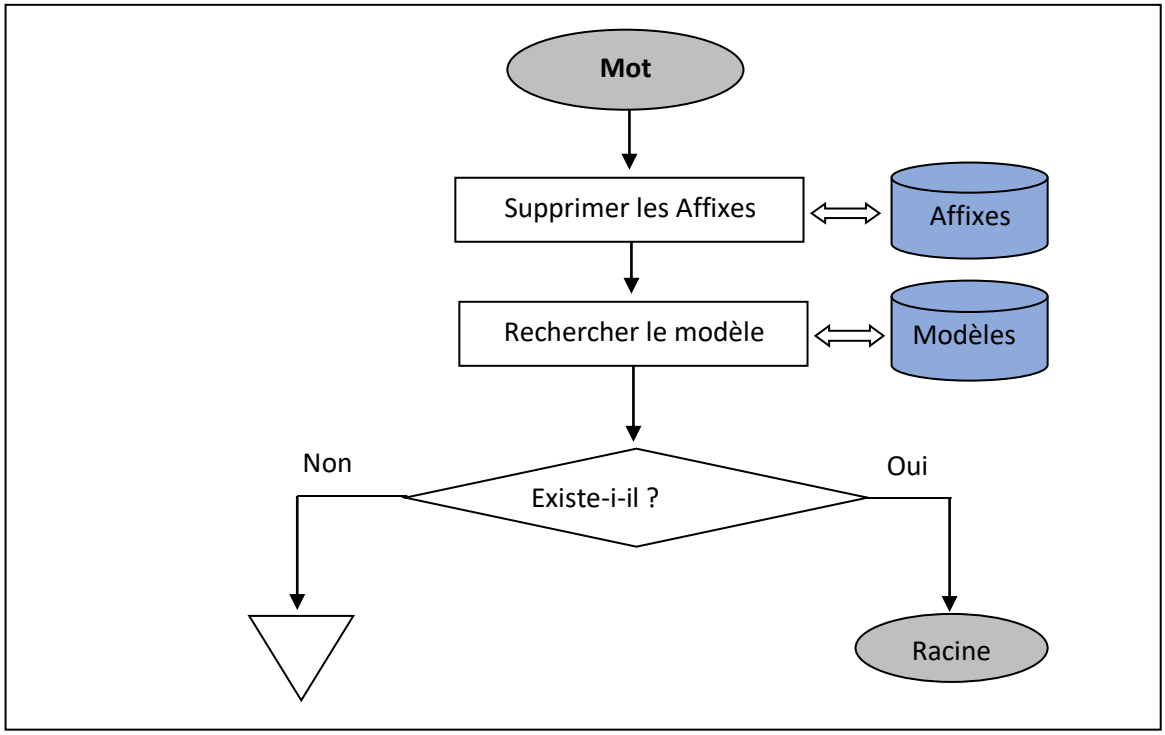


Figure2.6: Lemmatiseur computationnel [13].

- **Lemmatiseur sans dictionnaire :**

Dans cette méthode les ressources sont définies comme suit : D est l'ensemble des diacritiques. P3, P2 et P1 sont les ensembles de préfixes de longueur trois, deux et un. S3, S2 et S1 sont les ensembles de suffixes de longueur trois, deux et un. PR4 est l'ensemble de schèmes de longueur quatre. PR53 et PR63 sont les ensembles de schèmes de longueur cinq et six dont la racine est de longueur trois. PR54 et PR64 sont les ensembles de schèmes de longueur cinq et six dont la racine est de longueur quatre. La première étape est la normalisation orthographique. Si la longueur du mot est 4, alors on extrait le radical (radical) pertinent en supprimant les affixes de longueur 1 (S1, P1). Si sa longueur est 5, alors on extrait le radical (radical) de longueur 3 (selon les schèmes PR53), si aucun de ces schèmes ne correspond, alors on enlève les affixes pour avoir un radical (radical) de longueur 3. Si la partie restante du mot est de longueur 5, il faut extraire le radical (radical) de longueur 4 selon le schème PR54 et ainsi de suite [27].

- **Lemmatiseur des pluriels irréguliers sans dictionnaire :**

Pour extraire les racines à partir des pluriels irréguliers, Cette méthode procède en deux étapes et la figure ci-dessous montre ces étapes. La première utilise la méthode Lemmatiseur léger pour produire le Stem. La deuxième consiste à rechercher le modèle correspondant dans un ensemble de 39 modèles, de pluriels irréguliers, répertoriés par les auteurs à partir des livres de grammaire [13].

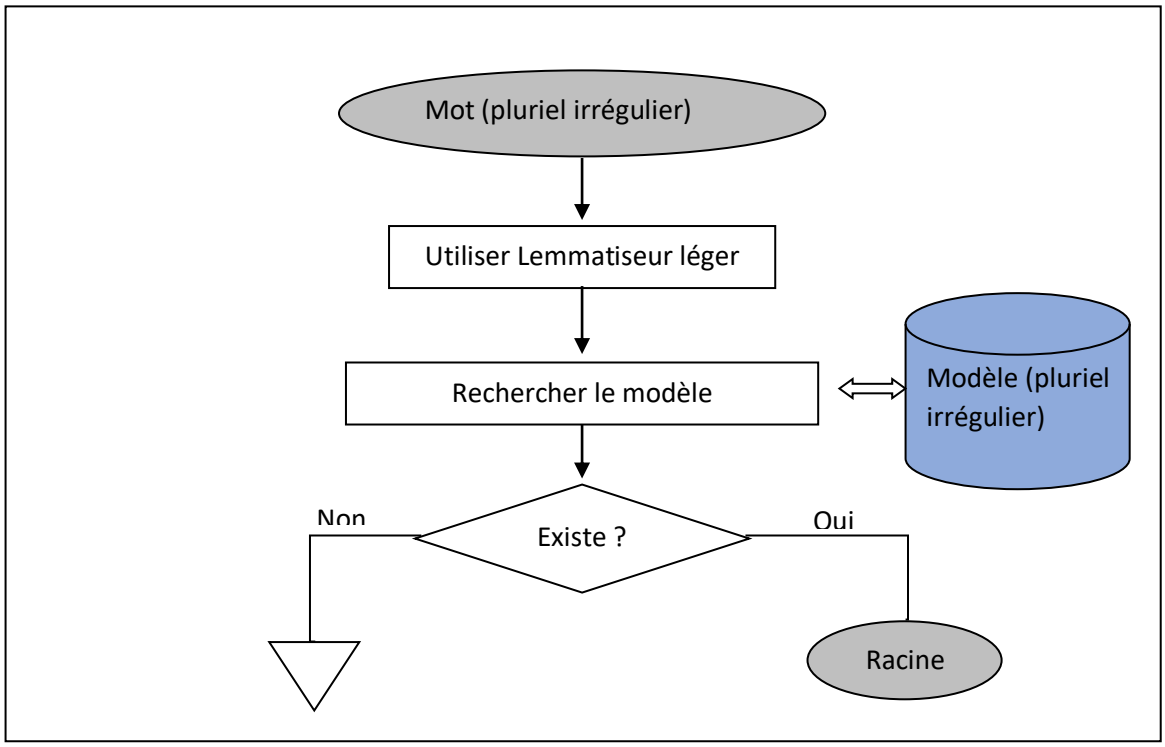


Figure2.7: Lemmatiseur des pluriels irréguliers sans dictionnaire [13].

- **Lemmatiseur léger transcript :**

Cette méthode calcule toutes les lemmatisations possibles. La phase préalable de cette méthode génère trois listes : la première pour les schèmes, la deuxième pour les préfixes et la troisième pour les suffixes. La construction des listes est basée sur un ensemble de couples (racine, mot dérivé). Un taux d'occurrence est calculé pour chaque élément des listes. Il y a cinq étapes principales dans cette méthode, la première étape est la normalisation orthographique, la deuxième est la transcription, la troisième supprime une séquence de n premières lettres consécutives ($n \leq 3$) si elle existe dans la liste des préfixes. La quatrième fait de même pour les n dernières lettres consécutives ($n \leq 3$) dans la liste des suffixes. La dernière étape consiste à extraire la racine selon la longueur du radical (radical). Si la longueur du radical (radical) est 3 alors il est examiné comme racine, sinon

il faut chercher un schème correspondant dans la liste des schèmes. Toutes les racines trouvées sont ajoutées à une liste qui présente la sortie de cette méthode [27].

Basée sur les modèles les affixes avec le dictionnaire :

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

- **Lemmatiseur de khoja :**

Cette méthode consiste à enlever les affixes après une première étape de normalisation. Ensuite, le résultat est comparé à une liste de modèles. Si une correspondance est trouvée, les lettres représentant la racine dans le modèle sont extraites. Ensuite, la racine ainsi extraite est validée dans un dictionnaire [13]. La figure 2.8 explique toutes les étapes du Lemmatiseur de khoja.

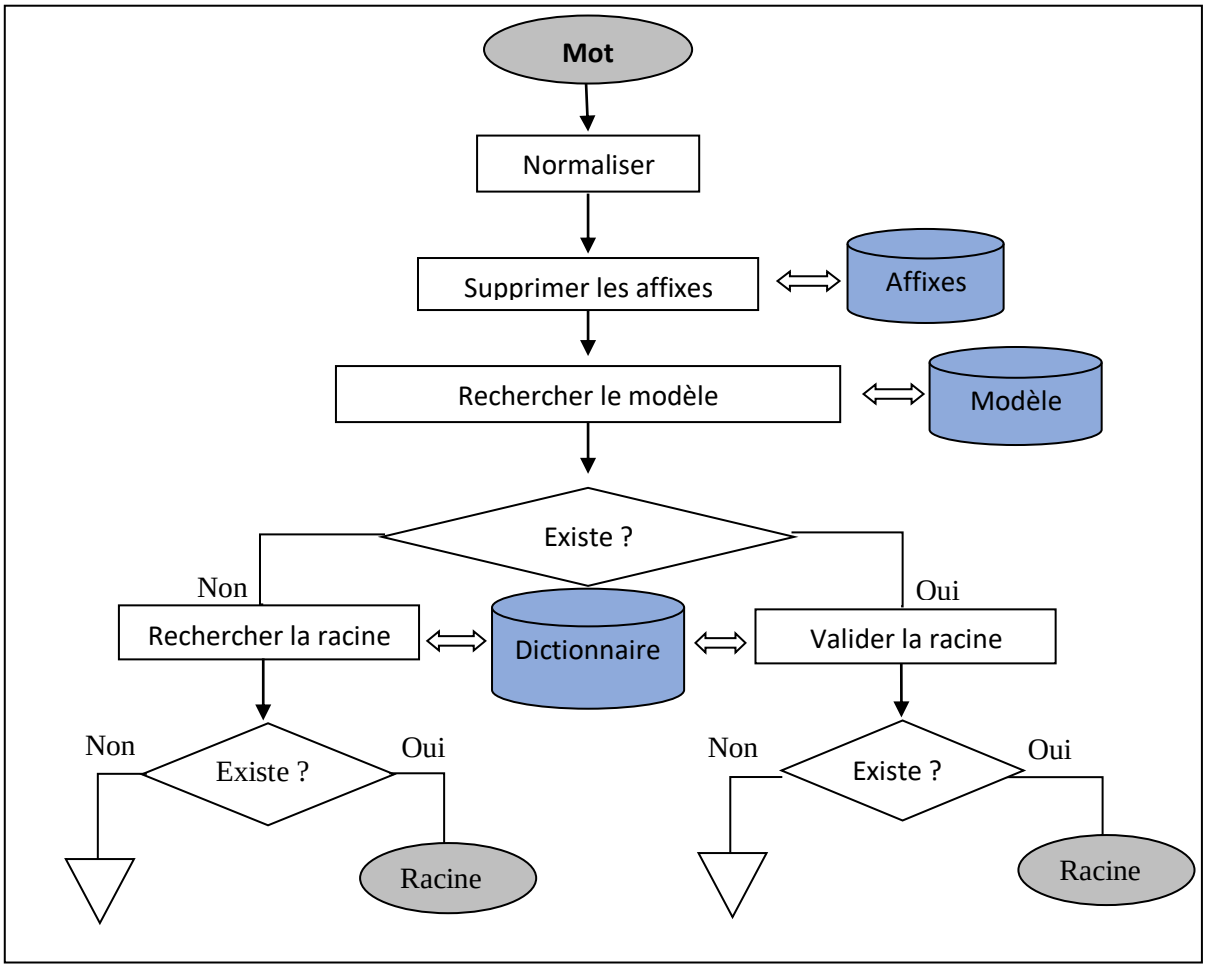


Figure2.8: Lemmatiseur de Khoja [13].

- **Lemmatiseur des pluriels irréguliers avec dictionnaire :**

Cette méthode sert à trouver la racine d'un mot qui est un pluriel irrégulier. Le principe de cette méthode est simple : rechercher simplement la racine dans une base de données. Le dictionnaire a été construit manuellement et validé par un linguiste à partir de 127.000 types de Stems, pour récupérer tous les types des modèles de pluriels irréguliers. Une liste d'environ 3600 Stems de pluriels irréguliers a été extraite et triée par ordre alphabétique et en fonction de chaque modèle du pluriel irrégulier [13], la présentation de cette méthode dans la figure suivant :

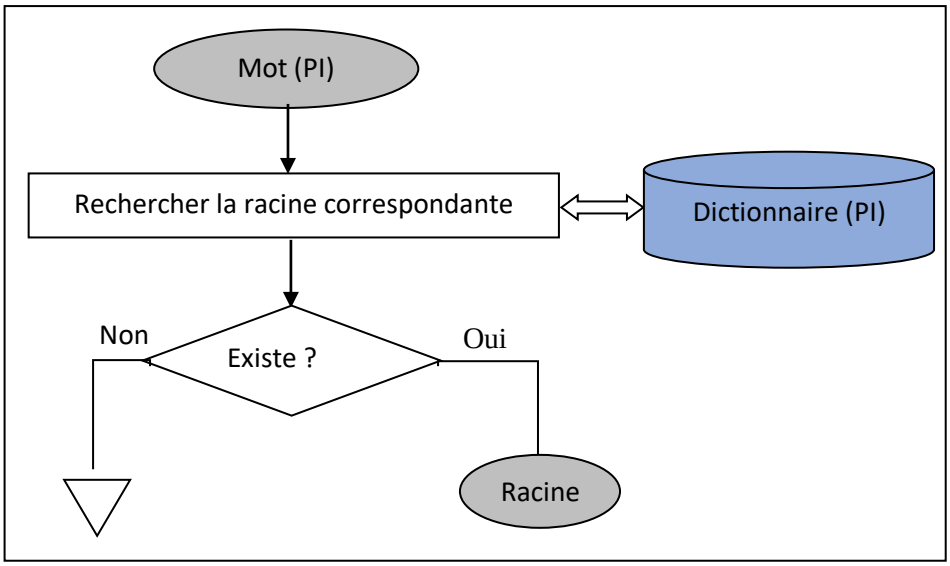


Figure2.9: lemmatiseur des pluriels irréguliers avec dictionnaire [13].

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

Dans cette approche, la méthode lemmatiseur TALN est proposée [13]. Nous avons trois étapes principales dans cette méthode. La normalisation orthographique est la première. La deuxième étape, consiste à supprimer les lettres qui représentent les éléments flexionnels (temps, nombre, personne, etc.) pour extraire le radical. La troisième étape consiste à trouver le schème correspondant au radical obtenu. La racine est extraite à partir du radical et du schème [27].La figure 2.10 montre ces étapes :

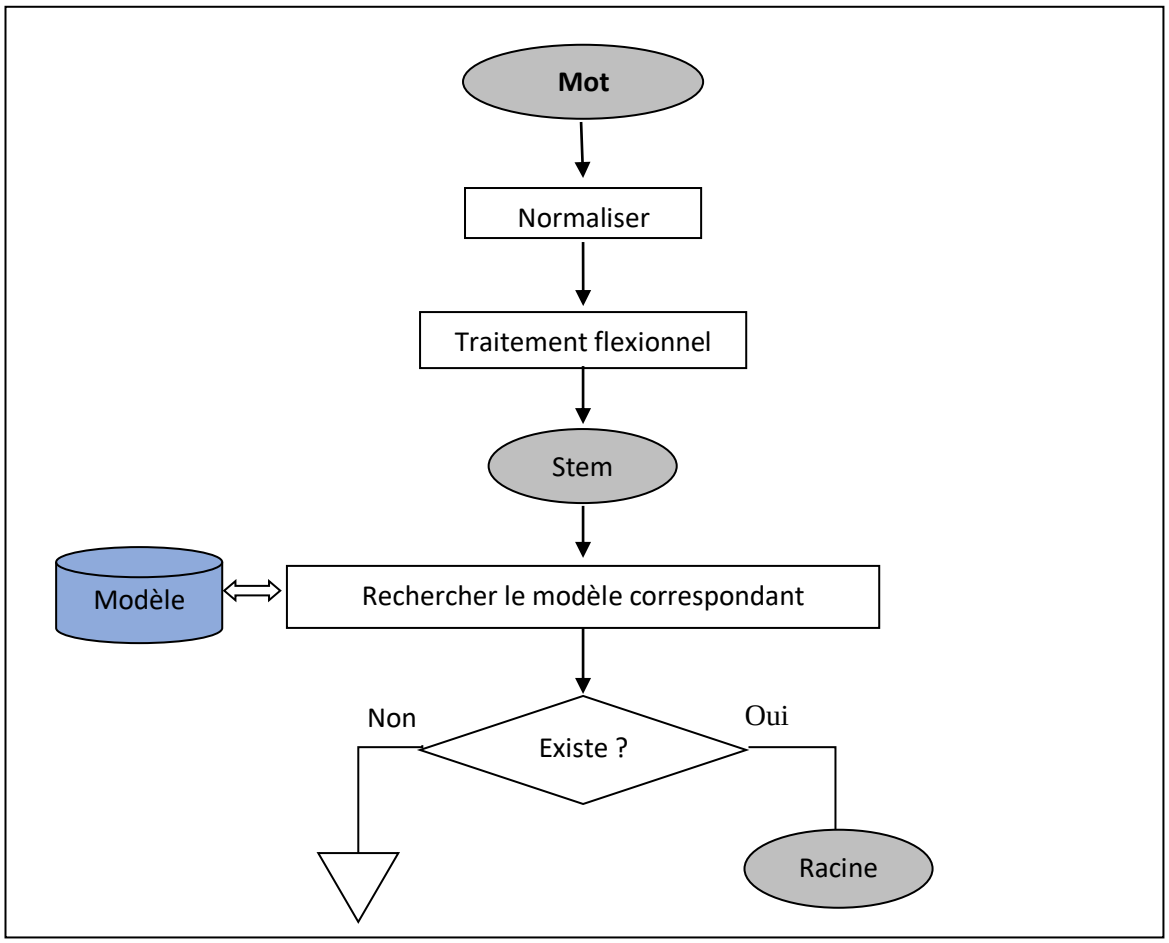


Figure2.10: Lemmatiseur TALN [13].

Basée sur les modèles les affixes avec des règles et dictionnaire :

Dans cette approche plusieurs méthodes sont proposées :

- **Lemmatiseur Xerox :**

Cette méthode procède en deux phases. Dans une phase préalable, on construit un dictionnaire global qui contient la plupart des mots Arabes et leurs racines. Ensuite, pour extraire la racine d'un mot donné, une simple recherche est effectuée dans le dictionnaire ainsi construit. La génération du dictionnaire est faite en utilisant des racines, des modèles, des règles et des affixes. Le centre Européen de recherche Xerox a proposé un analyseur morphologique pour l'Arabe standard moderne et qui est basé sur des dictionnaires. Le système a été largement modifié en utilisant la technologie d'états finis de Xerox [13].Dans les deux figures ci-dessous nous expliquerons cette méthode.

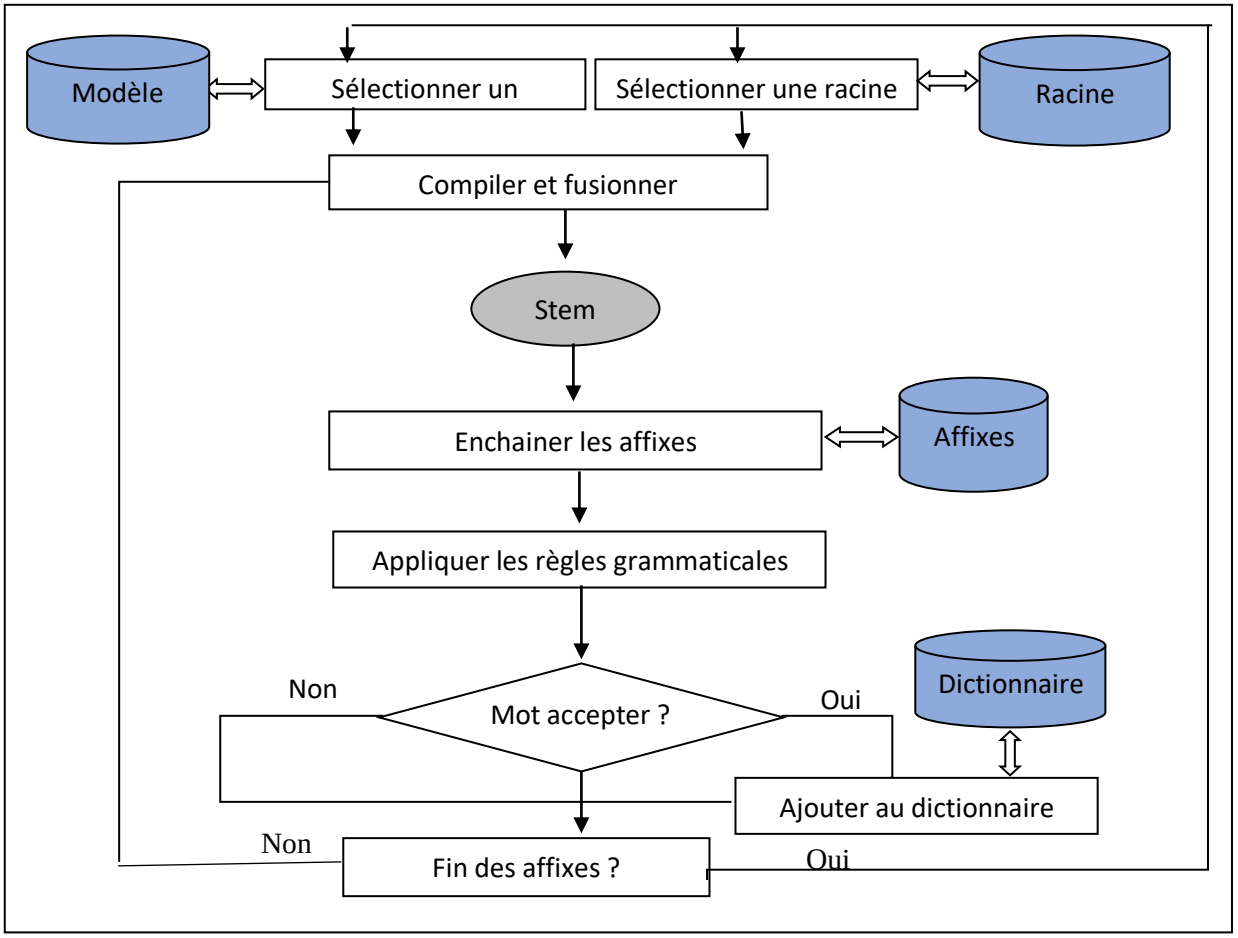


Figure2.11: Lemmatiseur Xerox : Génération du dictionnaire [13].

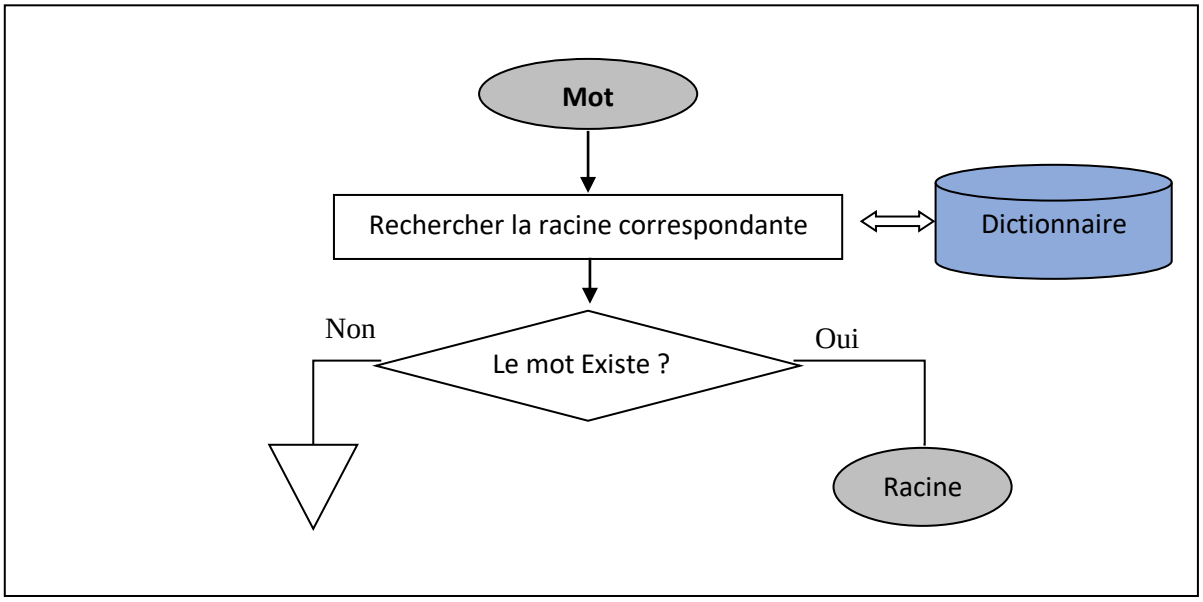


Figure2.12: Lemmatiseur Xerox : Recherche [13].

- **Lemmatiseur par génération systématique :**

Lemmatiseur par génération systématique propose une autre méthode pour construire un dictionnaire global. Les racines sont ordonnées et représentées selon le nombre des radicaux par une notation ensembliste : {F, M, L, Q}. Cet ensemble représente le premier radical (F), le radical médiale (M), le dernier radical dans une racine trilatérale (L), et le dernier radical dans une racine quadrilatérale (Q). Un modèle est aussi représenté par un ensemble {x1,..., xn} ou xi désigne la lettre à l'ième 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 xi de l'ensemble {x 1,..., 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 [15-27] la figure 2.13 présente cette méthode.

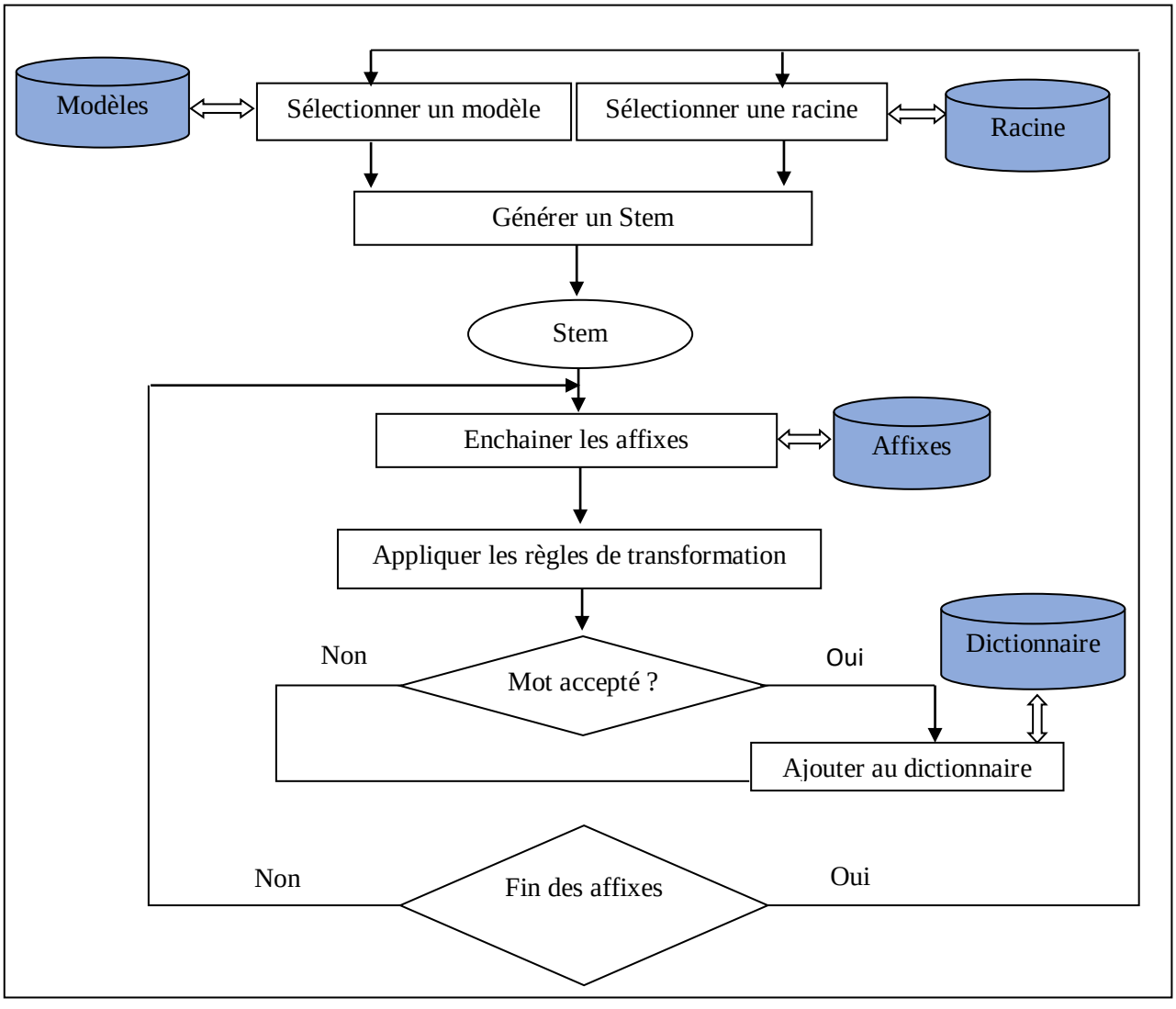


Figure2.13: lemmatiseur par génération systématique [13].

2.4.6 Lemmatisation par la traduction

L'idée est d'exploiter les algorithmes développés pour les langues latines et en particulier l'anglais vu que la plupart des algorithmes développés pour cette langue en donnés des résultats convaincants. Le principe de cette approche est de traduire un mot d'une langue source caractérisée par une richesse morphologique (comme l'Arabe) vers une langue cible, une fois le processus de lemmatisation accompli, l'opération inverse est effectuée. L'atout majeur de cette approche est de réduire l'inflexion élevée qui existe dans certaines langues [13], la figure suivant illustre cette méthode.

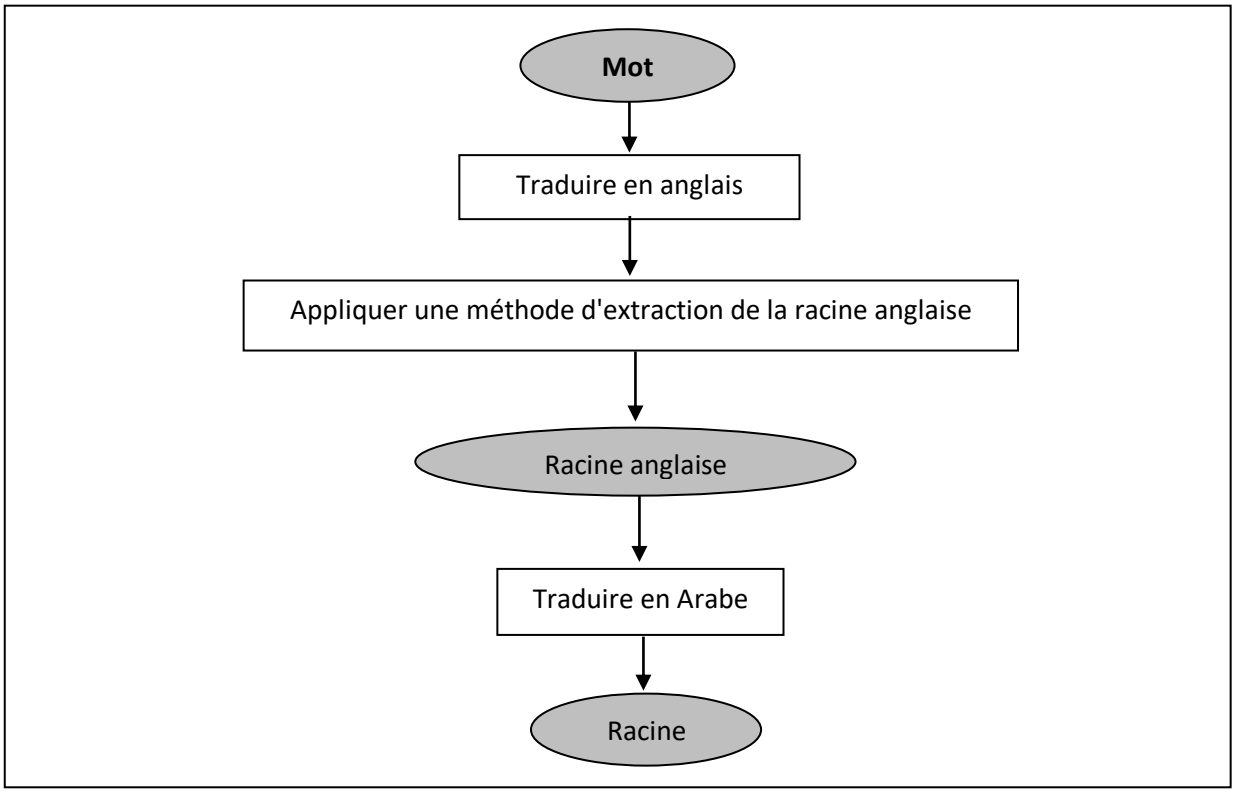


Figure2.14: Lemmatiseur Par Traduction [13].

2.4.7 Classe de l'analyse statistique

La plupart des lemmatiseurs existant que ce soit à base de dictionnaires, élimination des affixes ou bien par analyse morphologique sont généralement spécifique à une langue particulière. L'objective de l'approche statistique est de palier à cette spécialisation en adoptant des techniques qui s'inspirent des calculs stochastiques (Probabilité et Statistique) [13].

Calcul des coefficients de la ressemblance :

Est une première technique de classification automatique qui est basée sur la structure des mots. Le coefficient de ressemblance est calculé à partir du nombre de digrammes (2-gram) assortis dans de paires des sous lettres. Un échantillon de mots d'une base de données chimique a été choisi. Cette base contient certains stems dérivés des noms des éléments chimiques. Chaque groupe est caractérisé par une racine et ses mots dérivés (Figure 2.19). [13], ont présenté le modèle N-gram qui peut être utilisé pour calculer la ressemblance entre deux chaînes de caractères en comptant le nombre des N-grams semblables qu'ils partagent. Le coefficient de ressemblance est donné par l'équation (1) :

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

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

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

$$\boldsymbol{s = \frac{2c}{A+B}} \boldsymbol{(2)}$$

Où :

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

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

C : Nombre de paires uniques similaires entre le mot (A) et la racine (B). Le mot

(A) et les racines potentielles (B) à comparer avec, puis la mesure de similarité est effectuée en calculant la valeur de S entre le mot (A) et chacun une des racines potentielles(B) à comparer avec, puis la mesure de similarité est effectuée en calculant la valeur de S entre le mot (A) et chacune des racines potentielles(B) [13].

Calcul des coefficients de la dissemblance :

Est une approche statistique pour classer des documents Arabes. La technique utilise une mesure de la dissemblance appelée «Distance Manhattan» et une mesure de ressemblance appelée "Dicemeasure". Un corpus des 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 de normalisation a été utilisée. Il a aussi utilisé le 3-gram (Figure 2.20). Le coefficient de dissemblance entre les mots a et b est calculé en partitionnant chaque mot en 2-gram ou bien en 3-gram.α∩β présente l'intersection entre les deux ensembles et α U β l'union entre les deux ensembles (3) [13] :

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

2.4.8 Classe de l'analyse morphologique et statistique (hybride)

Dans cette approche en trouve :

Lemmatiseur léger statistique :

Lemmatiseur léger statistique est une nouvelle méthode d'extraction de racine d'un mot Arabe (Figure 2.15). La première étape de cette méthode supprime les affixes en appliquant la méthode Lemmatiseur léger. La deuxième étape calcule des coefficients de ressemblance entre le radical obtenu dans la première étape et une liste des racines classées dans un dictionnaire. La dernière étape consiste à ajouter à la liste des racines, les racines correspondant aux coefficients de ressemblance maximum [27].

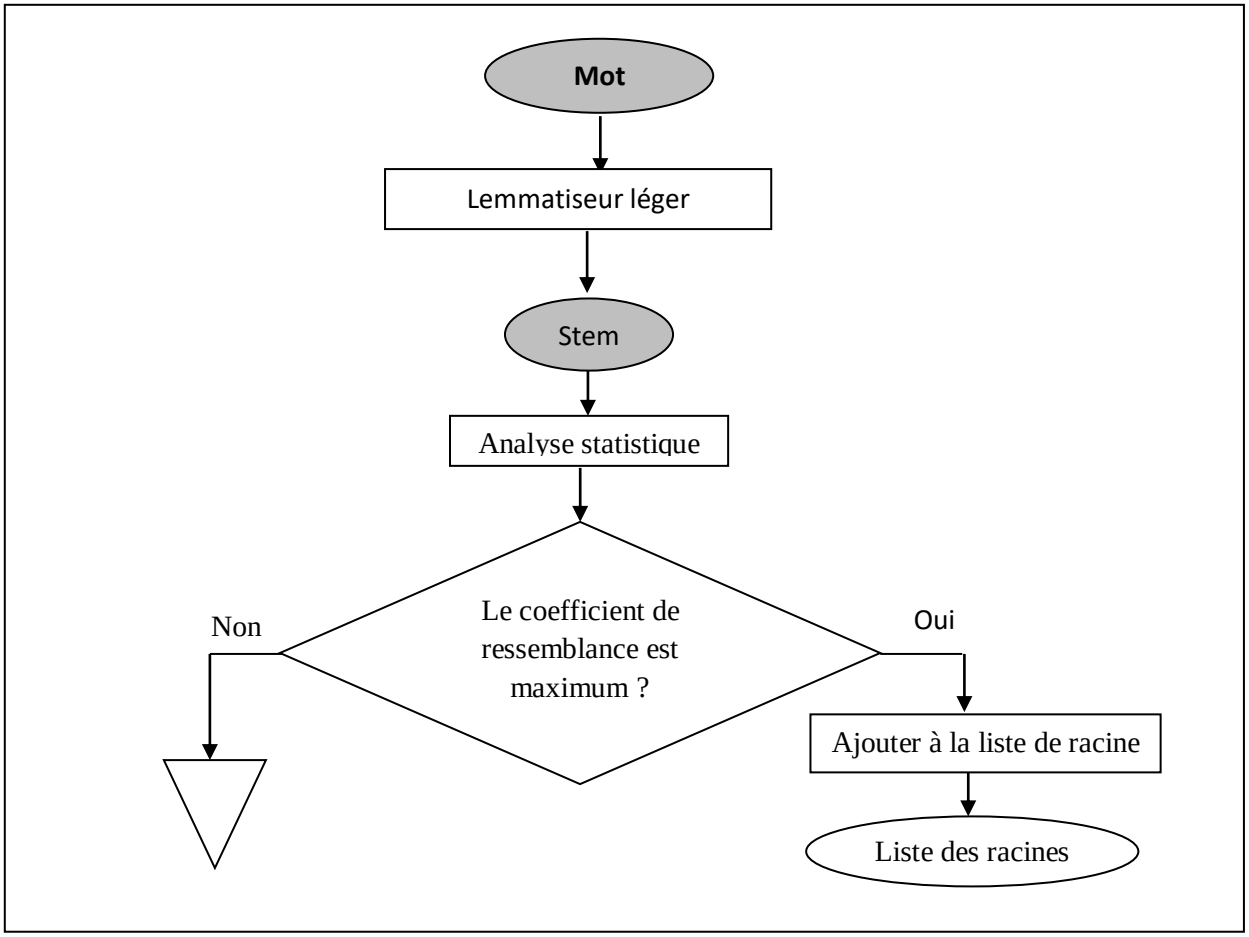


Figure2.15: Lemmatiseur Leger Statistique [13].

2.5 Comparaison entre quelques les méthodes de lemmatisation

Chacune des méthodes de lemmatisation arabe étudiées dans ce chapitre offre des avantages mais aussi souffre de certains inconvénients. Nous résumons ces les avantages et inconvénients de quelques méthodes de lemmatisation dans le tableau suivant :

Approche	**Avantages**	**Inconvénients**
Analyse Morphologique à base de Dictionnaire	-Basée sur des listes établies préalablement -Sortie soit un lemme ou racine -Traitant tous types des mots arabes	-Charge CPU élevée. -Gourmande en espace mémoire

	-Résultats précis - Adéquate pour l'apprentissage	
Analyse Statistique	-Basé seulement sur les calculs et la classification - Simple à implémenter - Pas besoin des grandes listes préalablement établies - Facile à gérer - Faible espace mémoire	- Difficulté de trouver des seuils de calcul - Résultats inexacts - Erreurs élevées pour le sur-stemming et le sous-stemming
Lemmatisation Légère	- Simple à implémenter - Sortie soit stem (tige) ou racine - Pas besoin de grandes listes préalablement établies - Facile à gérer - Adéquate pour la recherche d'information - Faible espace mémoire - Pas besoin de grande connaissance linguistique.	- Résultats inexacts. - Traitant seulement les préfixes et les suffixes

Tableau 2.1: Les avantages est les inconvénients de quelques méthodes de lemmatisation [16].

2.6 Conclusion

Dans ce chapitre, nous avons présenté les différentes méthodes pour l'extraction des racines (Lemmatisation), ces méthodes sont classées en fonction de leurs ressources. La classification des méthodes comprend trois classes principales : la première classe est basée sur l'analyse morphologique, la deuxième classe est basée sur l'analyse statistique et la troisième est basée sur l'analyse morphologique et statistique.

L'objectif du chapitre suivant est la présentation détaillée des différentes étapes de notre approche.

Chapitre 3

Méthode Développée

3.1 L’introduction

Dans ce chapitre, nous allons expliquer la méthode proposée en détail. Nous commençons par donner les étapes du prétraitement ; Par la suite, le processus de calcul de la similarité entre le mot et le schème. S’ensuivent plusieurs étapes pour conclure le processus en extrayant la racine appropriée.

3.2 Description du système réalisé

Dans le schéma ci-dessous, nous avons résumé toutes les étapes que nous avons suivis dans notre proposition :

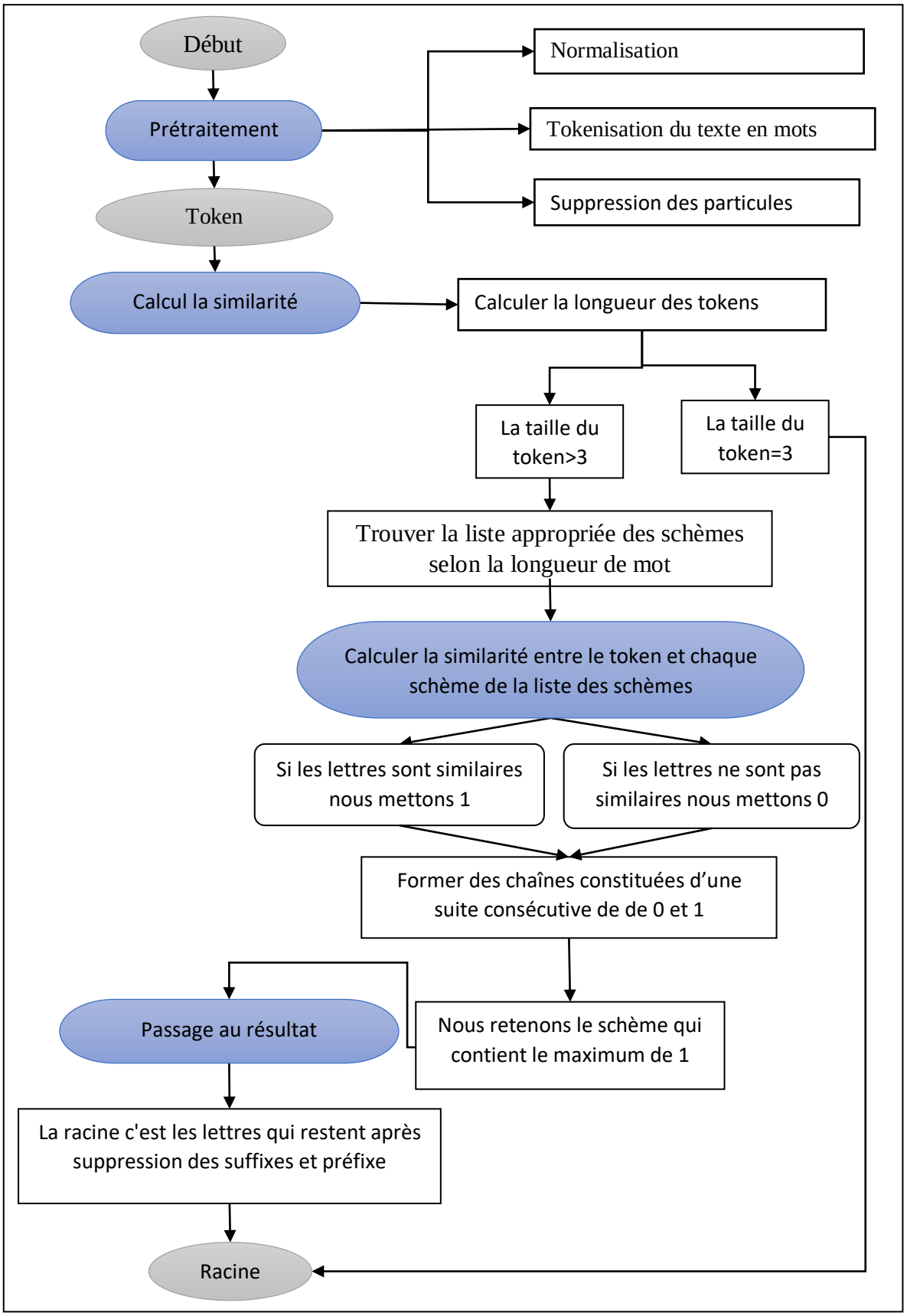


Figure 3.1: Schéma général de la méthode

3.2.1 Etape 1 : Prétraitement

Pour que les ordinateurs puissent comprendre le texte en langage naturel et appliquer dessus divers traitements, il faut utiliser les étapes de prétraitement comme suit :

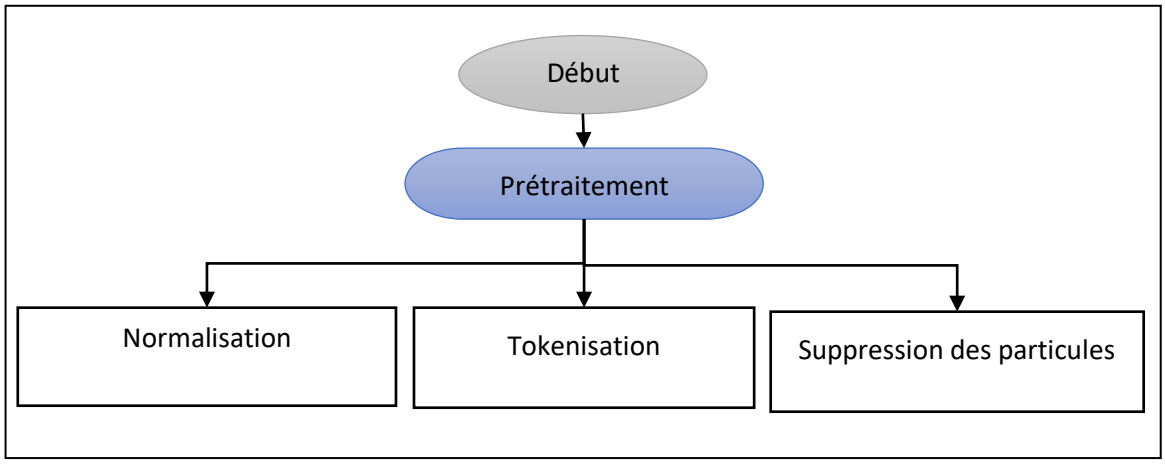


Figure 3.2: Schéma de prétraitement.

La normalisation :

La phase de normalisation vise à transformer une copie du document original dans un format standard plus facilement manipulable [19]. Cette étape est considérée nécessaire à cause des variations qui peuvent exister lors de l'écriture d'une même unité lexicale.

Le document est normalisé comme suit :

- Suppression des caractères spéciaux et chiffres.
- Suppression du caractère d'allongement (kashida ou tatouil), « جَمِيْلَـــةً→جميلة ».
- Remplacement de آ,إ etأparا.seulement si dans le début de token.
- Suppression les préfixes qui 2 caractères ("ال","وال","بال","كال","فال","لل").
- Remplacement des espaces multiples (plusieurs espaces consécutifs) par un seul.
- Suppression des diacritiques (ْ,ٌ,ً,َ,ّ,ِ,ٍ).

Les instructions de normalisation que nous avons mis en place, sont décrites par les pseudo algorithmes suivants :

Pseudo code de Remplacemnt de alif :

```
Algorithme norme Alif :
S : [] tableau de chaine de caractères, i : entier
Début :
Tant que (s[i] <>\0) faire
      Si s [0]= [ٱ ,إ ,أ]
        Alors remplace (s [0],ا )
      Fin si
 i++
Fait
```

Pseudo code d'Élimination les diacritiques :

```
Algorithme norme diacritique :
S : [] tableau de caractère, i=entier, s1=''
Début :
Tant que(s[i] <>\0) faire
    Si (voyelle(s[i] <>vrai)
Alorsconcatination (s1, s[i])
Fin si
    i++
Fait
Fin.
```

Pseudo code d'Élimination les ponctuations :

```
Algorithme norme ponctuation :
S : [] tableau de caractère, I=entier, s1= ' '
Début :
Tant que(S [I] <>\0) faire
    Si (ponctuation(S [I] <>vrai)
Alorsconcatination (S1, S [I])
Fin si
    I++
Fait
Fin.
```

Exemple de déroulement de ce processus :

Input :إِنَّ الْقُرَّاْء يَقْرَؤُوْنَ القرآن قِرَاْءَةً جَمِيْلَـــــةً

Output :ان القراء يقرؤون القران قراءه جميله

La tokenisation du texte :

La tokenisation d'un texte Arabe est une étape fondamentale pour son traitement automatique. Elle consiste à découper le texte en unités de type des mots appelées « tokens » et délimite ces des tokens par des espaces.

Le processus de tokenisation que nous avons utilisé est décrit par le pseudo algorithme suivant :

```
Algorithme tokenisation :
S : [] tableau de chaine de caractères ; I, J=entier, Ch.='' ; T : [] tableau de chaine de caractères
Début :
Tant que (s[i] <>\0) faire
Si s[i] =''
Alorsi++
Fin si
    Tant que(S[i] <>'') et (S [I] <>\0) faire
Concaténation (Ch, S [I])
        I++
Fait
Si (ch. <>'')
          Alors T[J ]←Ch.
J ++
        Fin si
Fait
```

Exemple de déroulement du processus de tokenisation :

Input : ذهب الولد الى المدرسة

Output :ذهب

الولد

الى

المدرسة

Elimination des particules :

Les particules sont généralement les plus fréquents dans un texte comme : في كل لم اللن له من هو هي قوة كما لها منذ وقد ولا لقاء مقابل هناك وقال وكان وقالت وكانت فيه لكن وفي ولم ومن وهو وهي يوم فيها منها يكون يمكن, حيث الا اما التي اكثر...

Les instructions d'elimination des particules sont décrites par le pseudo-algorithme suivant :

Algorithme supp particule :

Particule :[في,كل,لم,الى,لن,من, هو, هي, قوة, كما, لها, منذ , وقد, ولا لقاء, مقابل, هناك, وقال, وكان, وقالت , وكانت,فيه, لكن, وفي, ولم, ومن, وهو, وهي, يوم, فيها],i :entier

Début :

Tant que (s[i] <>'\0') faire

X= tokenisation(s [i])

i ++

Si x in particle

Alors suppression(x)

Fin si

Fait

Example:

Input : الولد ذهب الولد الى المدرسة

Output:الولد

ذهب

الولد

المدرسة

Les instructions de suppression des préfixes qui ont 2 caractères sont décrites par le pseudo-algorithme suivant :

Algorithme supp préfixes :

Prefixe2 : [**"ال","وال","بال","كال","فال","لل"**], i : entier, t : chaine des caractères

Début :

Tant que (s[i] <>'\0') faire

t= tokenisation(s [i])

i ++

Si t in prefixe2

Alors suppression(t)

Fin si

Fait

Exemple :

Input : ذهب الولد الى المدرسة

Output : ذهب

ولد

مدرسة

3.2.2 Etape 2 : calcul de la similarité

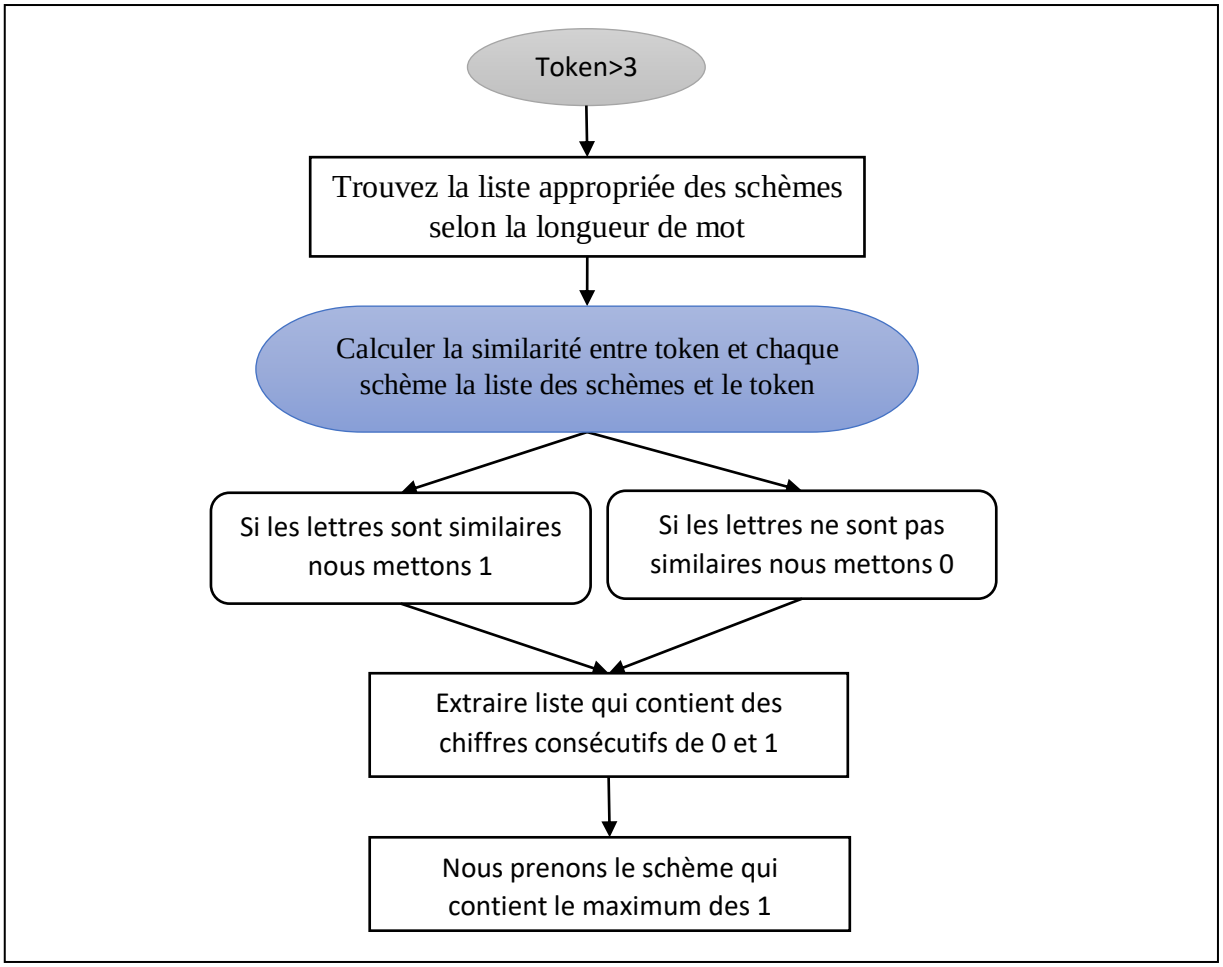


Figure 3.3: schéma calcule de similarité

Dans la deuxième étape nous devons d'abord créer une liste des schèmes où, tous les schèmes possèdent la même taille. Nous mettrons ces derniers dans une seule liste.

Exemple

Si la taille de schème=6

Alors la liste sera constituée comme suit :

["منفعلت منفعلة مفتعلي مفتعلت مفتعلة مفعلان مفعلين مفعلات فاعلون فاعلان فاعلات يفعلون مفعولي مفعولت مفوعلة مفعللي مفعللت مفعللة مفعالي مفعالت مفعالة مفعوعل مستفعل متفعلي متفعلت متفعلة متفاعل منفعلي مفعللة مفعنلي مفعنلت مفعنلة مفنعلي مفنعلت مفنعلة مفعيلي مفعيلت مفعيلة مفيعلي مفيعلت مفيعلة مفوعلي مفوعلت يفعلون مفعللت استفعل يفعلهن"]

Après, nous calculons la similarité comme suit :

- Calculer la longueur de token :

Si la longueur du token est 3, cela sera considéré comme une racine, parce que la racine arabe en général se compose de 3 consonnes.

Exemple :

Input :كتب

Output :كتب « Dans ce cas le token « كتب »est la racine ».

Sinon si token est supérieur à 3 :

Comme dans l'étape précédente, nous avons classé les racines en fonction de leur longueur dans des listes, nous allons à cette étape, l'utiliser pour trouver le schème approprié. Et ce, en calculant la longueur du token pour pouvoir choisir la liste de schèmes proportionnels à sa longueur. C'est ici que commence le processus de sélection du schème avec la comparaison les lettres du mot et les lettres de chaque schème de la liste. Ce processus est présenté ci-dessous :

Exemple :

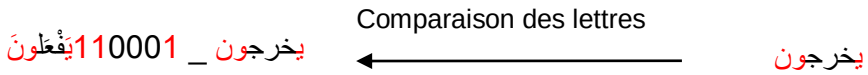
Si longueur (يخرجون)=6

Alors la liste des schèmes est égale à :

["منفعلت منفعلة مفتعلي مفتعلت مفتعلة مفعلان مفعلين مفعلات فاعلون فاعلان فاعلات يفعلون مفعولي مفعولت مفوعلة مفعللي مفعللت مفعللة مفعالي مفعالت مفعالة مفعوعل مستفعل متفعلي متفعلت متفعلة متفاعل منفعلي مفعللة مفعنلي مفعنلت مفعنلة مفنعلي مفنعلت مفنعلة مفعيلي مفعيلت مفعيلة مفيعلي مفيعلت مفيعلة مفوعلي مفوعلت يفعلون مفعللت استفعل يفعلهن"].

Nous effectuons par la suite, la comparaison suivante :

Si la lettre de token est similaire avec la lettre de schème nous mettons 1 sinon 0.

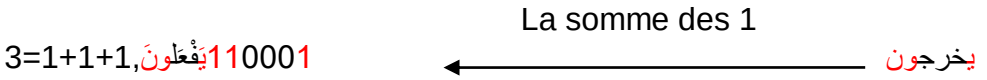


La liste des schèmes sera comme suit :

يخرجون

[[منفعلة [0, 0, 0, 0,0] , [منفعلت [0, 0, 0, 0, 0,, []مفنعلي [0, 0, 0, 0, 0, 0]] , [مفعنلة [0, 0, 0, 0, 0, 0, 0, 0]] , [مفعنلت [0, 0, 0, 0, 0]] , [مفعنلي [0, 0, 0, 0, 0]] , [متفعيل [0, 0, 0, 0, 0]] , [متفيعل [0, 0, 0, 0, 0]] , [يفعلون [1, 0, 0, 0, 1, 1]] , [مفعللة [0, 0, 0, 0, 0]] , [أفعالا [0, 0, 0, 0, 0]] , [مفعلون [0, 0, 0, 0, 1, 1], []يَفْعَلانِ,,[01, 0, 0, 0 ,1]]

- Apres avoir trouvé la suite de 0 et de 1 pour chaque schème de la liste, il faut calculer la somme des 1 comme suit :

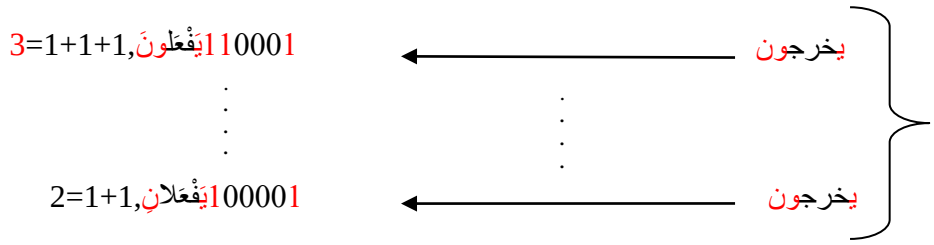


يخرجون

[[منفعلة,[0, 0, 0, 0, 0,0], [منفعلت [0, 0, 0, 0, 0,,0], []مفنعلي [0, 0, 0, 0, 0, 0] ,0, [مفعنلة [0, 0, 0, 0, 0,[0,0] , [مفعنلت [0, 0, 0, 0, 0] ,0, [مفعنلي [0, 0, 0, 0, 0] ,0, [متفعيل [0, 0, 0, 0, 0, 0, 0, 0] , [متفيعل [0, 0, 0, 0, 0] ,0, [يفعلون [1, 0, 0, 0, 1, 1] ,3, [مفعللة [0, 0, 0, 0, 0] ,0, [أفعالا [0, 0, 0, 0, 0] ,0, [مفعلون [0, 0, 0, 0, 1, 1] ,2, [يَفْعَلان,[0, 0, 0, 0, 1, 1] ,2[[

- L'étape suivante sert à extraire le max c.-à-d. la plus grande somme pour trouver le schème correspondant au token.

Exemple :



Le max =3 Donc le schème est يَفْعَلونَ

Le tableau suivant résume toutes les étapes précédentes :

Token	**يخرجون**		
Liste des schèmes contient même taille de token	يَفْعَلونَ	يَفْعَلانِ	منفعلة
Les chiffres des 0 et1	110001	100001	000000
La somme de la valeur de similarité	3	2	0
Le maximum somme de la valeur de similarité	3	2	0
Le schème correspondant	يَفْعَلونَ	يَفْعَلانِ	منفعلة

Tableau 3.1: Les étapes de notre méthode.

3.2.3 Etape 3 : passage au résultat

Après l'obtention de schème similaire au token, nous pouvons extraire la racine par suppression des suffixes et des préfixés.

Premièrement, il faut établir une liste des préfixes et des suffixes et en suite, comparer ces listes avec les préfixes et les suffixes qui sont composés dans le schème. Le tableau ci-dessous illustre cette méthode d'extraction la racine :

Token		يخرجون
Schème		يَفْعَلونَ
Le suffixe 1 caractère	= {'و', 'ن', 'ا', 'ت', 'ك', 'ي', 'ه', 'ة'}	يَفْعَلونَ
Le suffixe 2caractères	= {'ون', 'ات', 'ان', 'ين', 'تن', 'كم', 'نا', 'يا', 'نا', 'ها', 'تم', 'كن', 'ني', 'وا', 'ما', 'هم', 'هن', 'تك', 'يه', 'ته'}	
Le suffixe 3caractères	= {'كمل', 'تين', 'ننني', 'تان', 'همل', 'تمل'}	
Le préfixe 1 caractère	= {'ا', 'ن', 'ت', 'ي', 'و', 'س', 'ف', 'ب', 'ل'}	يَفْعَلونَ
Le préfixe 2caractères	= {'لن','لت', 'لي', 'بل', 'با', 'فا', 'كا', 'سن', 'ست', 'سا','سي','لل','ال', 'فن', 'في', 'فت'}	
Le préfixe 3caractères	= {'ولت','ولي','ولا','وسا','وسن','وست','وسي','وال','ولل','بال','كال', 'ولن'}	
La racine du schème		فْعَلَ
La racine du verbe		خرج

Tableau 3.2: Extraction de la racine

3.3 Conclusion

Nous avons présenté dans ce chapitre toutes les étapes de la méthode proposée pour l'extraction de racines. Cette méthode est en fait une hybridation des méthodes que nous avons étudiées lors de notre parcours bibliographique. Elle est basée sur l'analyse morphologique parce que cela dépend principalement des schèmes, mais aussi elle se base sur les techniques statistiques et celles utilisant les affixes.

Dans le chapitre suivant nous allons présenter les tests effectués avec notre méthode. En, commençant par un rappel de la description du système réalisé. Puis les différents outils utilisés (tels que : la présentation de l'environnement de développement…) en présentant quelques exemples d'application concrets pour illustrer les résultats.

Chapitre 4

Application et Résultats

4.1 Introduction

Dans ce dernier chapitre, nous allons présenter la partie de mise en œuvre de notre projet. Nous commençons tout d'abord par la présentation de langage de programmation ainsi que l'environnement de développement, en détaillant les différents outils et matériaux utilisés dans chaque étape. Puis nous montrons les différentes étapes de déroulement de l'application et enfin les résultats obtenus.

4.2 Environnement de développement

Dans cette section, nous allons d'abord expliquer ce qu'est un langage de programmation. Nous verrons ensuite brièvement l'histoire de Python, et puis nous allons présenter l'environnement de développement ainsi que les bibliothèques que nous avons utilisées.

4.2.1 Définition de python

Python [31] est un langage de programmation. Il est l'un des langages de programmation les plus intéressants du moment. Facile à apprendre, [32] Il s'agit d'un langage de programmation interprété, qui ne nécessite donc pas d'être compilé pour fonctionner. Un programme « interpréteur » permet d'exécuter le code Python sur n'importe quel ordinateur. Ceci permet de voir rapidement les résultats d'un changement dans le code. En revanche, ceci rend ce langage plus lent qu'un langage compilé comme le C.

Python [33] est un langage pour la programmation à usage général. Créé par Guido van Rossum et premièrement sorti en 1991, Python a une philosophie de conception qui met l'accent sur la lisibilité du code , notamment en utilisant des espaces importants . Il fournit

des constructions qui permettent une programmation claire sur les petites et grandes échelles.

Python dispose d'une dynamique de système automatique de gestion de la mémoire . Il prend en charge plusieurs paradigmes de programmation , y compris orienté objet , impératif, fonctionnel et procédural, et dispose d' une grande bibliothèque complète et standard .

Les interprètes Python sont disponibles pour de nombreux systèmes d'exploitation . CPython est l'implémentation de référence de Python. Elle est open source et dispose d'un modèle de développement communautaire. De même que presque toutes les autres implémentations de Python, Python et CPython sont gérés à titre lucratif Fondation Python Software .

4.2.2 PyCharm

PyCharm [36], est un environnement de développement intégré (IDE) utilisé pour la programmation en Python. Il fournit une analyse de code, un débogueur graphique, un testeur d'unité intégré, une intégration aux systèmes de contrôle de version (VCS) et prend en charge le développement Web avec Django. PyCharm est développé par la société tchèque JetBrains.

Il fonctionne sur plusieurs plates-formes Windows, Mac OS X et Linux. PyCharm a une édition professionnelle, publiée sous une licence propriétaire et une édition communautaire.

PyCharm permet de compléter le code de manière intelligente, d'inspecter le code, de mettre en évidence à la volée les erreurs et de les corriger rapidement, ainsi que la refacturation du code automatiquement et offre des fonctionnalités de navigation avancées.

Il s'intègre à IPython Notebook, dispose d'une console Python interactive et ainsi que de nombreux packages scientifiques, notamment matplotlib et NumPy.

Outre Python, PyCharm offre un support de premier ordre pour divers frameworks de développement Web Python, des langages de gabarit spécifiques, JavaScript, CoffeeScript, TypeScript, HTML / CSS, AngularJS, Node.js, etc.

Il a de nombreuses fonctionnalités telles que :

• Assistance et analyse du codage, avec complétion du code, mise en évidence de la syntaxe et des erreurs, intégration de l'interface et solutions rapides.

• Navigation de projet et de code : vues de projet spécialisées, vues de structure de fichiers et sauts rapides entre fichiers, classes, méthodes et usages.

• Test unitaire intégré, avec une couverture ligne par ligne.

4.3 Matériel utilisé

- **La machine 1**
  Nom de l'ordinateur : DELL-PC
  Processeur : Intel(R) Core(TM) i3-7020U CPU @ 2.30GHz 2.30 GHz
  Mémoire installée (RAM) : 4.00 Go
  Type du système : Système d'exploitation 64 bits
  Windows 10 professionnel
- **La machine2 :**
  Nom de l'ordinateur : TOSHIBA-PC
  Processeur : Intel(R) Core(TM) i3-4005U CPU @ 1.70GHz 1.70 GHz
  Mémoire installée (RAM) : 4.00 Go
  Type du système : Système d'exploitation 64 bits
  Windows 7.

4.4 Description du programme lemmatisation

Pour l'implémentation de notre solution, nous avons opté pour le langage python 3.7, et nous avons utilisé Pycharm comme environnement de développement qui nous permet d'utiliser différentes API (Application Programming Interfaces) à partir de plusieurs langages de programmation comme Java et C++...etc.
Nous avons utilisé dans notre application quelques bibliothèques pour assurer certaines fonctionnalités. Citons :

- **ArabicProcessingCog :**

  Est une bibliothèque destinée au traitement de texte dans python 3 particulièrement dédiée pour la langue Arabe. Nous l'avons utilisé notamment pour sa fonction de tokenization.

- **ISRI (Information Science Research Institute's**)

  Est un algorithme de stemming (radicalisation) [35].Elle partage de nombreuses caractéristiques avec l'organe de Khoja stemmer, la principale différence est qu'ISRI stemmer n'utilise pas de dictionnaire racine. De plus, si une racine n'est pas trouvée, ISRI stemmer va renvoyer une forme normalisée, plutôt que le mot d'origine non modifié. Dans notre programme, nous utilisons particulièrement les fonctions suivantes : fonction de normalisation et fonction de suppression de préfixes et suffixes. Par ailleurs, nous avons utilisé les listes de préfixes et de suffixes fournies par ISRI pour obtenir de bons résultats.

- **la bibliothèque nltk**

  Natural Language Toolkit (NLTK) [34],est une bibliothèque logicielle en Python permettant un traitement automatique des langues, développée par Steven Bird et Edward Loper du département d'informatique de l'Université de Pennsylvanie. En plus de la bibliothèque, NLTK fournit des démonstrations graphiques, des données-échantillon, des tutoriels, ainsi que la documentation de l'interface de programmation

4.4.1 Déroulement

Dans cette section, nous allons présenter et expliquer les différentes étapes de déroulement de notre programme qui débute par un prétraitement des textes Arabe et permet d'extraire en sortie, la racine de chaque mot occurrent dans ce texte et ce programme basé sur niveau lexicale.

La figure qui suit, illustre l'interface globale de notre système :

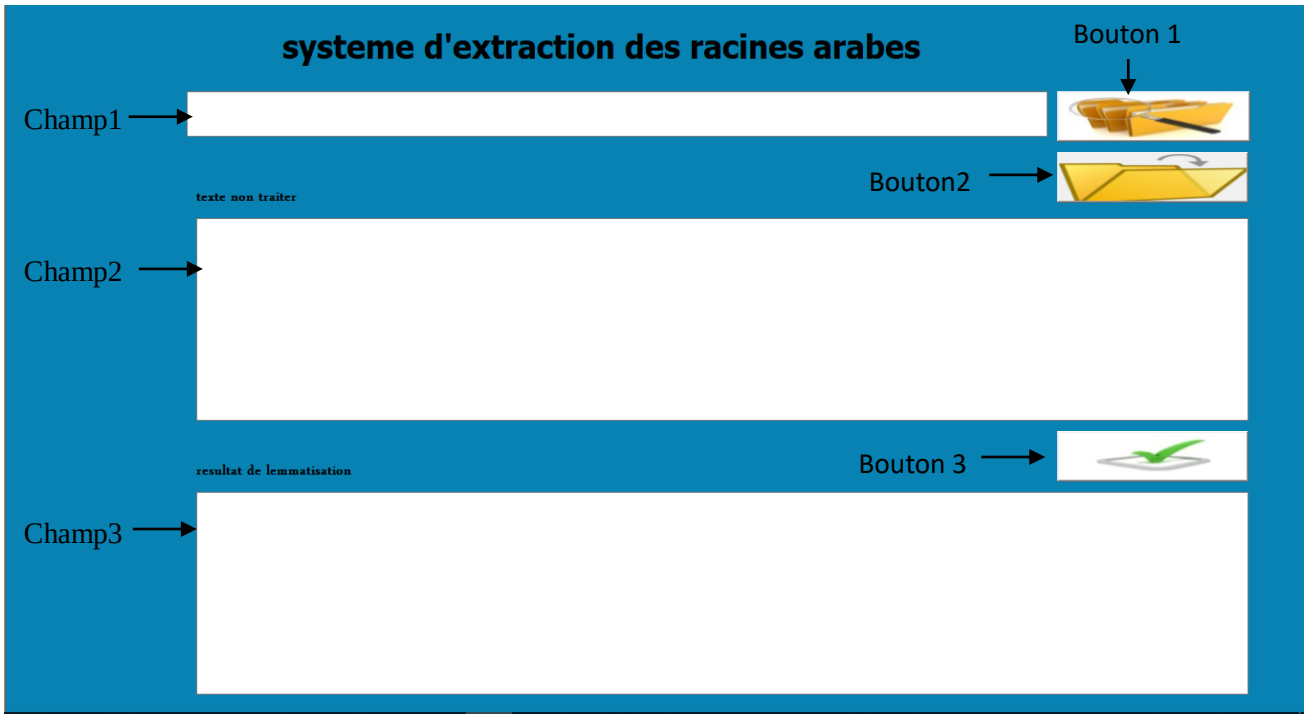


Figure 4.1: Interface globale

Bouton1 : pour choisir le fichier texte.

Bouton2 : pour ouvrir le fichier texte.

Bouton 3 : pour exécuter le processus lemmatisation.

Champ 1 : pour Capturer le lien de fichier texte.

Champ 2 : pour afficher le contenu de fichier texte.

Champ 3 : pour afficher le résultat de processus lemmatisation

- **Etape1 :** Choix un fichier texte

  Et pour l'étape suivante, nous devons cliquer sur le bouton 1 qui permet d'ouvrir la fenêtre que l'on voit dans la figure suivante. Et ce, pour choisir le texte approprié de notre programme. Il est à noter que le fichier doit être de type «*.txt», et la langue du texte doit obligatoirement être l'Arabe pour que les traitements puissent être effectués correctement.

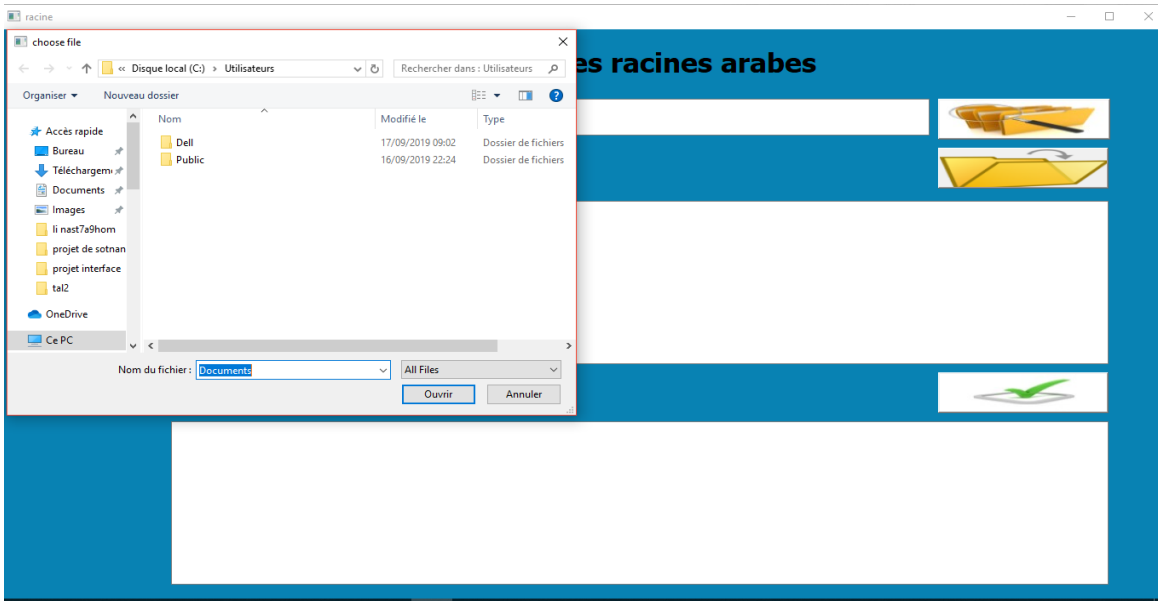


Figure 4.2: Choix du fichier texte.

- **Etape2 :** Capture du lien

  Après le choix du fichier, le lien du fichier que nous avons sélectionné est capturé dans le champ 1 comme montré dans la figure 4.3 :

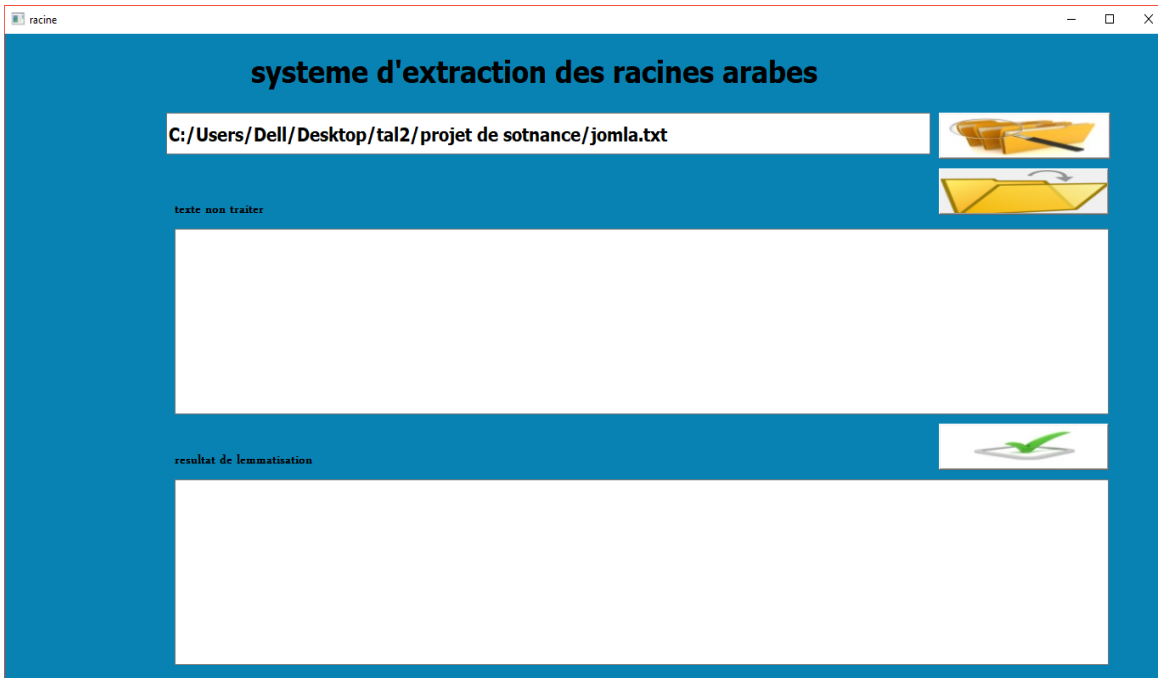


Figure 4.3: Capture du lien du fichier texte.

- **Etape3 :** Ouverture du fichier

  En appuyant sur le Bouton 2, le contenu de ce dernier est affiché dans le champ 2 comme montré dans la figure suivante :

Figure 4.4: Affichage du contenu du fichier texte.

- **Etape4 :** Processus de prétraitement

  Avant de passer à l'étape suivante, nous devons suivre le processus de prétraitement afin de préparer le texte sélectionné. Les états de ce processus sont montrés ci-dessous :

La Normalisation :

- Suppression des caractères spéciaux et chiffres.
- Suppression du caractère d'allongement (kashida ou tatouil), « جَمِيْلَـــةً→جميلة ».
- Remplacement de آ,إ et أpar ا.seulement si dans le début de token.
- Suppression les préfixes qui 2 caractères ("ال","وال","بال","كال","فال","لل").
- Remplacement des espaces multiples (plusieurs espaces consécutifs) par un seul.

On peut écrire ces les instructions dans le programme suivant :

```
defde (text) :
    noise = re.compile ("" ٥    | # Tashdid
٥    | # Fatha
٥    | # Tanwin Fath
٥    | # Damma
٥    | # TanwinDamm
٥    | # Kasra
٥    | # TanwinKasr
٥    | # Sukun
                          [0-9]    | # arkam
                      [a-zA-Z]     | # alpha
    [*+-/=,;:°><|//?؟¿!_(@#$%^&*)~ø/$£}{`'']    | # amaliyat
                                [[] | # hadina
                                []] | # hadina1



_     # Tatwil/Kashida
                                "", re.VERBOSE)
text = re.sub(noise, '', text)
text = re.sub(r'(.)\1+', r'\1', text)
return text
```

Le code suivant concerne le remplacement du ٱet le ٱet le آ initial par l'alif nu ا

```
defnorm(word, num=3):

if num == 1:
word = re_short_vowels.sub('', word)
elifnum == 2:
if word[0]:
word = re.sub("[ااآٱإ]", "ا", word)
return word
elifnum == 3:
word = re_short_vowels.sub('', word)
if word[0]:
word = re.sub("[ااآٱإ]", "ا", word)
return word
return word
```

Le code suivant concerne la suppression les préfixes :

```
defpre2 (word) :
    p2 = [u"ال", u"وال", u"بال", u"كال", u"فال", u"لل"]
"""normalize short prefix"""
for sp1 in p2:
if word.startswith(sp2):
return word[2:]
return word
```

Nous exécutons les programmes en même temps :

Input :

أكتب خَبِيرَاتْ خُبَرَاءْ حَتَّى يَسْأَلُهُمْ كَاتِبْ نُبَلَاءْdfggمَصِيرُ هُمَا المَكْتُوبْ

Output :

مصير هما مكتوب اكتب خبيرات خبراء حتى يسألهم كاتب نبلاء

Tokenization et élimination les mots vides :

Le programme suivant permet de fractionner le texte en des tokens, Par la suite, nous procédons par la vérification si le token est un mot vide ou non avec listes des mots vides.

```
deftokenize(text):
return text.strip().split()
stopwords = [u"بلا", u"فلا", u"ولا", u"ان", u"بعد",u"ضد",u"يلي",u"الى", u"في", u"من",u"حتى",
u"وهو", u"يكون", u"به", u"وليس", u"أحد",u"على", u"وكان", u"تلك",u"كذلك", u"التي", u"وبين",
u"فيها", u"عليها", u"إن", u"وعلى", u"لكن", u"عن",u"مساء",
u"ليس",u"منذ",u"الذي",u"أما",u"حين", u"ومن", u"ليست", u"وكانت",u"أي", u"ما",u"عنه",
u"حول", u"دون",u"مع", u"لكنه", u"ولكن", u"له",u"هذا",u"والتي", u"فقط", u"ثم"u"هذه",u"أنه",
u"تكون",u"قد",u"بين",u"جدا",u"لن",u"نحو",u"كان",u"لهم",u"لأن",u"اليوم",u"لم", u"هؤلاء",
u"فإن",u"فيه",u"ذلك",u"لو",u"عند",u"اللذين",u"كل",u"بد", u"لدى", u"وثي",u"أن", u"ومع",
u"فقد", u"بل",u"هو", u"عنها", u"منه", u"بها", u"وفي", u"فهو",u"تحت", u"لها", u"أو", u"إذ",
u"علي", u"عليه", u"كما", u"كيف", u"هنا", u"وقد",u"كانت", u"لذلك",u"أمام",u"هناك",u"قبل",
u"معه", u"يوم", u"منها", u"إلى",u"إذا", u"هل", u"حيث", u"هي", u"اذا", u"او", u"و", u"لا",
u"الي", u"إلي",u"مازال", u"لازال",u"لايزال",u"ما يزال", u"اصبح", u"أصبح", u"أمسى",
u"امسى",u"أضحى", u"اضحى", u"ظل", u"ما برح", u"ما فتئ", u"ما انفك", u"بات", u"صار",
u"وليست",u"إن",u"كأن", u"ليت", u"لعل", u"لاسيما", u"ولايزال", u"الحالي", u"ضمن",
u"اول",u"وله", u"ذات", u"اي", u"بدلا",u"اليها",u"انه",u"الذين",u"فانه",u"وان",u"والذي",u"وهذا",
u"لهذا", u"الا", u"فكان",u"ستكون",u"مما",u"أبو",u"بإن",u"الذي", u"اليه", u"يمكن", u"بهذا",
u"لدي",u"وأن",u"وهي",u"وأبو",u"آل",u"الذي",u"هن",u"الذى",u"وكذلك",u"وما",u"ولهذا",u"وقبل"
,u"اما",u"بتلك",u"بهذه",u"بذلك",u"لماذا",u"ماذا",u"لما",u"ولما",u"ولماذا",u"وماذا",u"ولم",u"مثل",
u"ابن",u"مثلا",u"بعض",u"كلا",u"قال",u"قالت",u"لقد",u"ولقد",u"ولعل",u"انهم",u"ايضا",u"كل",u"وكل
",u"بينهم",u"بين",u"بينما",u"بينهن",u"وبينما",u"وبين",u"بيننا",u"وبينهم",u"وايضا",u"وان",u"والى
",u"وفي",u"واذا",u"لذا",u"ولذا",u"بحيث",u"وهل",u"ومنها",u"بما",u"وبما",u"سوف",u"لسوف",u"
ولسوف",u"وسوف",u"أثناء",u"وأثناء",u"التي",u"والتي",u"بالتي",u"فالتي",u"فيهم",u""]

f = open('C:/Users/Dell/Desktop/tal2/projet /normalisatio.txt', encoding='utf-8')
a=f.read()
aa=de(norm(a,3))
for token in tokenize(aa):
if token not in stopwords:
print(token)
```

Input :

مصير هما على مكتوب الى من اكتب خبيرات خبراء حتى يسألهم كاتب نبلاء

Output :

مصير هما

مكتوب

اكتب

خبيرات

خبراء

يسألهم

كاتب

نبلاء

A ce stade, nous avons terminé le processus de prétraitement. Le texte est prêt pour la prochaine étape.

Etape3 : Exécution finale du programme

Le Bouton 3permet d'exécuter la fonction de lemmatisation qui donne pour chaque mot du fichier texte sélectionné une racine. Ce résultat est affiché dans le champ3. La figure suivante montre ceci :

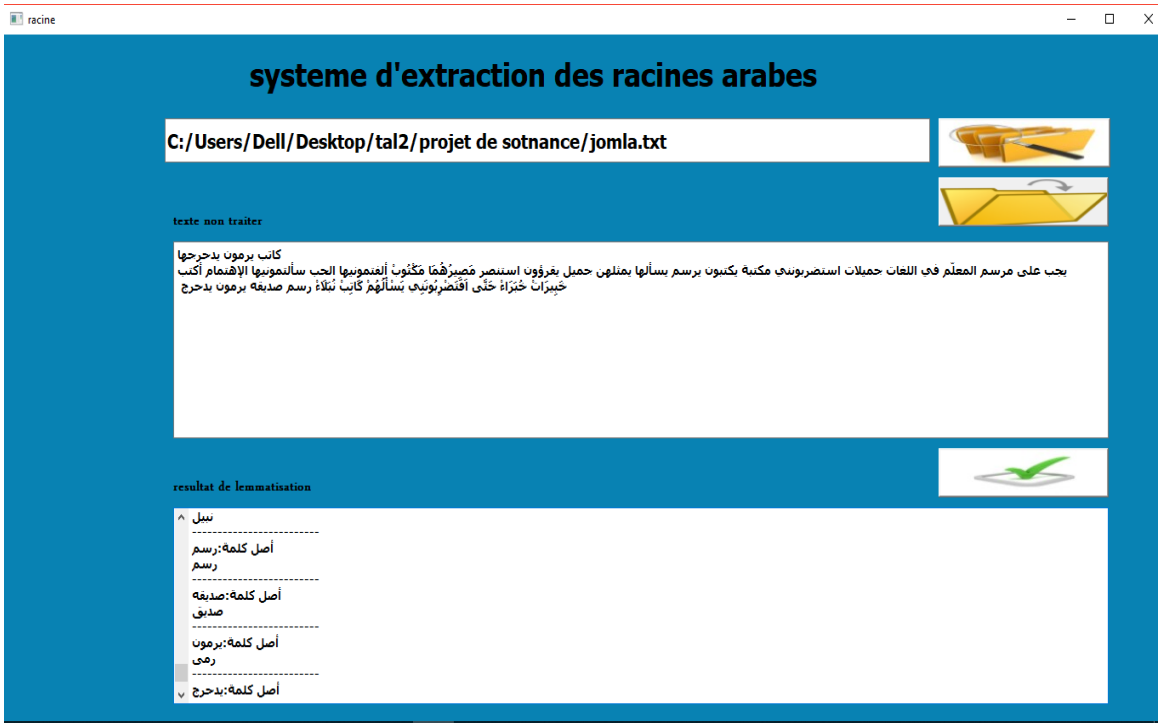


Figure 4.5 : Résultat du programme de lemmatisation.

4.5 Comparaison avec d'autres lemmatisations

Après avoir extrait les racines, nous allons maintenant comparer ces résultats obtenus par la méthode que nous avons utilisée avec d'autres méthodes qui sont :

- Light 10[1] [37].
- ISRI[2].
- ARLSTEM[3] [38].

La méthode ARLSTEM et la méthode Light 10 sont toutes les deux disponibles en ligne et en mode open-source et pour ISRI, il s'agit d'un programme dans la bibliothèque nltk dans python.

Le tableau suivant montre la comparaison que nous avons effectuée :

Les mots	Notre approche	Light 10	ISRI	ARLSTEM
سألتمونيها	سأل	سألتمون	سألتموني	لتموني
خَبِيرَاتْ	خبيرة	خَبِيرَاتْ	خبر	خبير
استنصر	نصر	استنصر	نصر	ستنصر
يمثلهن	مثل	يمثلهن	مثل	مثله
جميل	جميل	جميل	جمل	جميل
يقرؤون	قرأ	يقرؤ	قرؤ	يقرؤ
مكتبة	مكتبة	مكتب	كتب	مكتب
خُبَرَاءْ	خبير	خُبَرَاءْ	خبراء	خبراء
كَاتِبْ	كاتب	كَاتِبْ	كتب	كاتب
يرمون	رمى	يرم	يرم	يرم
مَصِيرُهُمَا	مصير	مَصِيرُهُمَا	صير	مصير
اللغات	لغة	لغ	لغت	لغ
المعلّم	معلّم	معلّم	علم	معلم
ألفتمونيها	ألف	ألفتمون	الفتموني	فتموني
يدحرج	دحرج	يدحرج	دحرج	دحرج

[1] Nous pouvons voir le code source de Light 10 : https://github.com/disooqi/ArabicProcessingCog
[2] Nous pouvons voir le code source d'ISRI : https://www.nltk.org/_modules/nltk/stem/isri.html
[3] Nous pouvons voir le code source d'ARLSTEM : https://www.nltk.org/_modules/nltk/stem/arlstem.html

يدحرجها	دحرج	يدحرج	دحرج	دحرج
رسم	رسم	رسم	رسم	رسم
صديقه	صديق	صديق	صدق	صديق
يكتبون	كتب	يكتب	كتب	يكتب
الاهتمام	إهتمام	اهتمام	همم	هتمم
نسوة	نسوة	نسو	نسة	نسو
الحب	لحب	حب	لحب	لحب

Tableau4.1: Comparaison de notre méthode avec les autres méthodes.

Nous remarquons que notre méthode a donné des résultats avancés pour certains mots comme les mots : « سألتمونيها », « يقرؤون» et « خُبَرَاءْ », «اللغات», « يرمون» les racines elles correctes par rapport aux autres méthodes et aussi il y a des résultats similaires pour d'autres mots. Par ailleurs, il y a quelques cas que nous n'avons pas pu résoudre car il y a des mots qui changent complètement quand on change leur état morphologique comme le mot « نسوة» notre système ne pouvait pas extraire la racine de ce mot, par ce que la racine de cette mot est « مرأة».

Bien que nous ayons réussi à extraire les racines des mots arabes avec notre système comme montré dans le tableau ci-dessus, mais la particularité morphologique de la langue arabe reste toujours le plus gros obstacle pour construire un programme de lemmatisation complet de cette langue.

4.6 Conclusion

Tout au long de ce chapitre, nous avons présenté l'environnement de développement ainsi que le langage de programmation utilisé pour implémenter et développer notre programme. Nous avons également mentionné les différentes bibliothèques qui ont été utilisées dans le programme ainsi que quelques captures. En dernier, nous avons comparé les résultats obtenus par notre système par rapport aux résultats donnés par d'autres systèmes de lemmatisation.

Nous concluons que notre système d'extraction de racines donne de bons résultats car il extrait la racine d'un mot donné quel que soit son état morphologique sans avoir à utiliser des dictionnaires.

Conclusion générale et perspective

Le traitement automatique de l'Arabe est un domaine de recherche stimulant. Il combine en effet plusieurs défis intéressants, parmi lesquels on peut citer la complexité morphologique de la langue, L'objectif du traitement automatique des langues est la conception de logiciels capables de traiter de façon automatique des données exprimées dans une langue dite « naturelle ».

Les langues naturelles se fondent sur des règles grammaticales, syntaxiques et morphologiques. Le niveau de difficulté et de complexité dépend de la langue elle-même. L'arabe est une langue hautement flexionnelle et a besoin particulièrement d'une lemmatisation efficace.

La lemmatisation des mots arabes a été un sujet central de nombreuses recherches en traitement des textes qui ont tenté de trouver les racines des mots arabes qui sont beaucoup plus abstraits que les lemmes. On fait appel à une lemmatisation de textes pour des fins de recherche d'informations, de classification, de traductions ou autres.

Plusieurs approches de lemmatisation sont appliquées à la langue Arabe, mais un lemmatiseur complet pour cette langue n'est pas disponible. La lemmatisation est difficile pour les langues avec des morphologies complexes comme l'Arabe, c'est la raison pour laquelle nous avons cherché à mettre en œuvre une solution qui donne de bons résultats. Cette méthode s'inspire de la méthode khalilienne des signes diacritiques dont nous avons modifié certaines étapes pour atteindre notre objectif souhaité.

La méthode repose fortement sur les schèmes. Au cours de la mise en œuvre de la méthode, nous avons rencontré plusieurs difficultés, à savoir : la suppression des signes diacritiques (manque de voyellation) ce qui entraîne parfois un changement de la signification du mot par exemple dans le mot عَلَم و عِلْم.Aussi, la forme plurielle des noms qui peut brouiller le traitement. Dans ce cas-ci, un nom au pluriel prend une autre forme morphologique différente de sa forme initiale du singulier. Par exemple, le mot « امرأة » (femme) au singulier prend la forme « نسوة » au pluriel.

Au final, nous sommes satisfaites de notre expérience, des connaissances acquises et des résultats obtenus jugés encourageants, mais il y a toutefois des lacunes que nous n'avons pas pu résoudre et auxquelles nous n'avons pas pu trouver de solutions. Ces

lacunes concernent principalement l'ambiguïté morphologique spécifique à la langue Arabe.

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

Références bibliographiques

[1] Saadane Houda, «Le traitement automatique de l'Arabe dialectalisé : aspects méthodologiques et algorithmiques »UNIVERSITÉ GRENOBLE ALPES, 14 décembre 2015.

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

[3] Souhir Gahbiche-Braham, Amélioration des systèmes de traduction par analyse linguistique et thématique Application à la traduction depuis l'arabe, thèse de doctorat en informatique, soutenue le 30 Septembre 2013, École Doctorale d'Informatique, Université Paris Sud.

[4] Soufian Baloul, Développement d'un système automatique de synthèse de la parole à partir du texte arabe standard voyellé, thèse de doctorat en informatique, soutenue le 27 mai 2003, Université du Maine, France.

[5] S. Baloul, M. Alissali, M. Baudry, P. Boula de Mareüil : Interface syntaxe-prosodie dans un système de synthèse de la parole à partir du texte en arabe, 24es Journées d'Étude sur la Parole, 24-27 juin 2002 Nancy, pp.329-332.

[6] نضال أحمد الشريف, « The sound of aldadu (ض) in Arabic Language », The Islamic University–Gaza Research and Postgraduate Affairs, juillet 2017.

[7] Mohamed Hédi Maâloul, Approche hybride pour le résumé automatique de textes. Application à la langue arabe, Thèse de doctorat en informatique, soutenue le 18 décembre 2012, Université Aix –Marseille.

[8] Azzeddine Lazrek, « Vers un système de traitement du document scientifique Arabe », Université Cadi Ayyad, 18 février 2002.

[9] R. Boite, H. Bourlard, T. Dutoit, J. Hancq et H. Leich. Traitement de la parole, chapitre, Synthèse de la parole à partir d'un texte, pp. 345-441, Collection Electricité, Presses polytechniques et universitaires romandes, 2000.

[10] Mouelhi Zoubeïr« Essai de lexicométrie d'une œuvre Arabe classique :Al-'Imtâ' wa-l-Mu'ânasa de Tawhîdî », Université Lumière Lyon 2,22 novembre2008.

[11] Aïda KHEMAKHEM, "ArabicLDB : une base lexicale normalisée pour la langue arabe", Mémoire de master, université de Sfax, Tunisie, 2006.

[12] 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.

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

[14] Dhaou Ghoul, « Outils génériques pour l'étiquetage morphosyntaxique de la langue arabe : segmentation et corpus d'entraînement », Mémoire de master, UNIVERSITE STENDHAL, 12/10/2011.

[15] Abbes, Ramzi, la conception et la réalistion d'un concordancier électronique pour l'arabe. Thèse de doctorat en sciences de l'information, Lyon,ENSSIB/INSA. . (décembre 2004).

[16] Nejmeddine khalfallah. عناصر اولية من النحو العربي. Lorraine : Département d'Arabe - Université de Lorraine. (2008).

[17] Mohsen Maraoui, Mounir Zrigui, Georges Antoniadis," Un système de génération automatique de dictionnaires étiquetés de l'arabe",CITALA 2007, Rabat, Maroc.

[18] Selim Mesfar, « Analyse morpho-Syntaxique Automatique et Reconnaissance Des Entités Nommée en Arabe Standard », Université de Franche_Comte Ecole Doctorale (Langages, Espaces, Temps, Societies), 24 novembre 2008.

[19] Larkey L. S., Ballesteros L. and Connell M., Improving Stemming for Arabic Information Retrieval: Light Stemming and Co occunence Analysis, In Proceedings of the 25th Annual International Conference on Research and DeveÏopment in hformation Retrieval (SIGIR 2002), Tampere, finÏand August 2002, pp. 275-282.

[20] Rahma Boujelbane, Mariem Ellouze, Frédéric Béchet, Lamia Belguith, « De l'arabe standard vers l'arabe dialectal : projection de corpus et ressources linguistiques en vue du traitement automatique de l'oral dans les médias tunisiens», 8 Sep 2015

[21] Lamia Hadrich Belguith, Leila Baccour et Mourad Ghassan. Segmentation de textes arabes basée sur l'analyse contextuelle des signes de ponctuations et de certaines particules. Actes de la 12ème conférence sur le Traitement Automatique des Langues Naturelles TALN'2005 - Dourdan France, vol. Vol. 1, 2005.

[22] Tahar SAIDANE (1), Mounir ZRIGUI (2) et Mohamed BEN AHMED(3), La Transcription Orthographique-Phonétique De La Langue Arabe. (1) Société Tunisienne d'Electricité et du Gaz, Centre de production de Sousse, Tunisie. (2) Laboratoire RIADI, Unité Monastir Faculté des Sciences de Monastir, Tunisie. (3) Laboratoire RIADI, Ecole Nationale des Sciences de l'informatique, Tunis, Tunisie. 19-22 avril 2004.

[23] Atef Ben Youssef, Méthodes Mixtes pour la Traduction Automatique Statistique, Université STENDHALGrenoble3, 01 juillet 2008

[24] A.Chentir, Etude de la Microprosodie en vue de la Synthèse de la parole en Arabe Standard. Thèse de Doctorat, Ecole Nationale Supérieure Polytechnique, Algérie, 2009.

[25] J.Wightwick, M.Gaafar, Arabic verbs and essentials of grammar. Chicago: Passport Books, 1998.

[26] Lamia Hadrich Belguith, Mme.NouhaChaâben, « Analyse et désambiguïsation morphologiques de textes Arabes non voyellés », Faculté des Sciences Économiques et de Gestion de Sfax – Laboratoire LARIS, 10-13 avril 2006.

[27] Georges Lebboss, « Contribution à l'analyse sémantique des textes Arabes », Université Paris 8 Vincennes à Saint-Denis, le 8 juillet 2016.

[28] Groblink,M., Mladenic,D.: 2004,Text mining tutorial Slovenia (2004).

[29] https://i.pinimg.com/564x/db/fb/fb/dbfbfb406791d3032e59bb811cf83d64.jpg, 27/06/2020, 22:09, pinimg.

[30] https://lArabefacile.fr/le-duel-en-Arabe-explication/ ,24/02/2019, heure : 22 :45, L'Arabe facile.

[31] https://python.doctor/ ,08/09/2019, heure : 14:00, Python Doctor France.

[32] https://www.lebigdata.fr/python-langage-definition08/09/2019, heure : 13 :15, Lebigdata.

[33] https://www.tresfacile.net/introduction-au-langagepython/?fbclid=IwAR1jqYRlI7BjlNU92hNQUTSk9VtS65CPUUkDn-R_0rVRbkPYFokXuPH_3HM 08/09/2019, heure : 13 :20, Python Très Facile

[34] https://www.techopedia.com/definition/30343/natural-language-toolkit-nltk, 08/09/2019, heure : 09:00, Techopedia.

[35] https://kite.com/python/docs/nltk.isri,19/09/2019, heure : 22 :00, Kite.

[36] https://www.jetbrains.com/pycharm/features/,19/09/2019, heure : 21 :05, Jetbrains.

[37] https://github.com/disooqi/ArabicProcessingCog/tree/master/arabic_processing_cog , heure : 19 :25, 20/09/2019 Github.

[38] https://www.nltk.org/_modules/nltk/stem/arlstem.html#ARLSTem.20/09/2019, heure : 19 :15 , Nltk Project.

[39] http://www.technolangue.net/imprimer.php3?id_article=274,20/09/2019, heure : 20 :25, Technolangue.

[40] https://www.linternaute.fr/dictionnaire/fr/definition/lemme/ ,16/06/2020, heure : 22 :00, linternaute.

[41] https://islamqa.info/ar/20953, واي باك مشين, 16/06/2020,21 :45.