الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي والبحث العلمي
Ministère de l'enseignement Supérieur et de la Recherche scientifique

Université d'Adrar – Algérie
Faculté des sciences et de la technologie
Département des mathématiques et informatique

Mémoire,

présenté au ***Département des mathématiques et informatique*** *de l'Université d'Adrar pour satisfaire partiellement aux exigences du diplôme de Master académique (LMD),*

Option **:SITW**

Thème : Une approche de vocalisation automatique de texte arabe non vocalisé « MOSHAKIL »

Présenté par :
Siham Laroussi / Yamina Bourouba

Encadré par :
Mr. Mohamed Amine CHERAGUI

Pr : Dr .OMARI
Ex : Mr .MEDIANI
Ex : Mr .CHOGUER

Soutenu le : 09 Juin 2018

Année universitaire : 2017 / 2018

Remerciements

Nous tenons tout d'abord à remercier puissant « ALLAH » de nous avoir donné la force d'aboutir au terme de ce projet et d'y être arrivé en bonne santé.

Nous ne pouvons pas oublier de présenter notre gratitude à nos parents pour leur patience et les efforts inlassables qu'ils ne cessent de déployer pour nous.

On tien a remercier vivement nos enseignants pour leur présence et leur suivi tout au long de l'année : Mr.CHERAGUI Mohamed Amine., qui présenté ce sujet, avec ses conseils importantes qui nous ont permis de prendre la bonne direction dans le travail.

Sans oublier de Un grand merci Mr . madiani mohamed, pour leurs conseil et aides précieuses et Dr .Omar mohamed elgazzar De l'Egypte, qui n'a pas manqué de nous aider.

Nos remerciements et notre gratitude vont aux enseignants Faculté des sciences et sciences technologie d'Université d'ADRAR.

Dédicace

A mes parents, pour leur bienveillance et leur tendresse.

À ma grand-mère et tante Fatima– A toute ma famille.

Je dédie ce modeste travail en témoignage de ma reconnaissance à:

Tous mes amis, mes professeurs pour leur aide et soutien.

A mon binôme « yamina bourouba »

A tous ceux qui m'aiment.

Dédicaces

Je dédie ce modeste travail à :

A mes parents .Aucun hommage ne pourrait être à la hauteur de l'amour Dont ils ne cessent de me combler.

A toute ma famille, et mes amis,

A mon binôme Lroussi siham et toute la famille BOUROUBA.

Et à tous ceux qui ont contribué de près ou de loin pour que ce projet soit possible, je vous dis merci.

Table des matières

III. Chapitre 2 : La langue arabe vs TAL

IV. Chapitre 3 : Diacritisation automatique

Figure Liste des

Liste des Tableau

Introduction générale

La langue humaine n'est pas seulement un ensemble de signes et de règles structurales ou un simple moyen de communication, mais plus que ça, elle permet de décrire l'histoire de toute une race humaine ou une civilisation ; et un système de valeurs pour une nation permettant ainsi de conserver et de déterminer son identité et son existence, tel que la vie d'une nation est intimement liée à la vie de sa langue, et de ce fait on peut dire qu'on' a répondu amplement à notre question initial (À quoi sert la langue humaine?).

Au fil du temps, une science à part entière s'est forgée un chemin pour étudier cette langue sous le nom de « linguistique » afin de la préserver de toute diffamation (une concaténation de deux termes science et langue) dans le but de structurer la langue. Divers écoles linguistiques ont été bâties depuis celà, on partant de l'école de « Basra de Abou Isshak El Haddrami » et celle de « Koufa », en passant par « le Fonctionnalisme de Martinet » et « le Distributionnalisme de Harris et Bloomfield ».

Mais le temps passe, l'être humain évolue en terme de penser et de développement technologique d'une manière phénoménal. L'homme s'est développé dans tous les domaines, il a constaté que la langue peut entrer dans son champs de développement, et avec l'apparition de l'outil informatique cette alliance entre linguistes et scientifiques a donnée naissance à une nouvelle discipline connue sous l'acronyme TALN[1] dans le but est d'automatisé le langage naturel.

La langue arabe est une langue qui a fasciné plusieurs chercheurs, et celà grâce d'un côté à sa structure qui est fait d'une manière remarquable par nos anciens linguistes qui ont donné à cette langue une solide base et des règles de construction stable, mais aussi à la richesse morphologique de cette dernière en la comparant avec d'autres langues telles que : le français et l'anglais…etc.

Comme toute les langues de par le monde la langue à toujours suscitée son intérêt au processus d'automatisation, à travers le développement d'une pléthore d'applications touchant divers thématiques (la traduction automatique, la correction orthographique, le résumé automatique, ...etc.), mais l'un des axes qui suscite énormément d'intérêt vu l'impact que peut avoir sur le processus de compréhension d'un texte arabe est la vocalisation ou voyelllation automatique.

La vocalisation automatique, consiste à injecté à l'intérieur d'un mot des signes de diacritization tels que : Fetha «ً » ,dumma« ٌ », kesra «ٍ », skoun« ٌ », ainsi que d'autres, ces signes par leurs présence permette de lever l'ambiguïté dans énormément de scénarios permettant à d'autre programmes (comme : l'étiquetage morphosyntaxique, la lecture automatique, ….etc.) de donner de meilleur rendement et performance.

[1]TALN signifier : Traitement Automatique des Langues Naturelles

Le but de notre étude de contribuer au développement de l'automatisation de la langue arabe par l'élaboration d'un système de vocalisation automatique.

Pour cela, le mémoire est organisé comme suit :

- Chapitre 1 : permet de situer notre champs d'étude qu'est le Traitement automatique de la langue naturelle.
- Chapitre 2 : mettra la lumière sur la langue arabe, à travers sont historique, ces variantes, sa morphologie ainsi que les premiers travaux sur l'automatisation de la langue arabe ainsi que les grands facteur (techniques) qui ont laissé ce processus d'automatisation connaître un grand retard par rapport à d'autre langues.
- Chapitre 3 : est le noyau de notre mémoire, dans lequel on présentera l'architecture de notre vacaliseur de la langue arabe que nous avons baptisé MOSHAKIL, ainsi que les différentes phases de traitements.
- Chapitre 4 : sera dédié à la présentation de notre vocaliseur MOSHAKIL à travers l'environnement de développement, l'interface graphique et les différents tests est résultats obtenus.

Chapitre 1

TALN
TRAITEMENT AUTOMATIQUE DU LANGAGE NATUREL

1. Introduction

Le traitement automatique des langues naturelles (TAL) a pour objet la création des programmes informatiques capables de traiter automatiquement les langages naturels.

La langue naturelle désigne la langue parlée ou écrite par les êtres humains, par opposition aux langages artificiels, informatiques, mathématiques ou logiques, par exemple. En fait, le traitement ne concerne pas directement la langue. Mais il porte sur les données linguistiques, les textes codés dans une langue particulière. Sous cette dénomination générique. Nous regroupons aussi les dialogues, écrits ou oraux, et des unités plus petites, comme les paragraphes ou les phrases. [1].

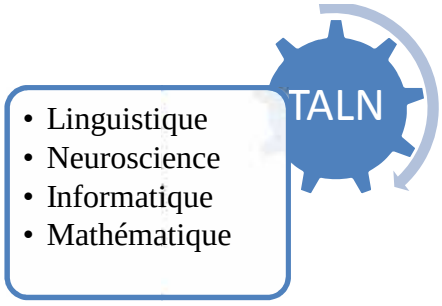

Figure 1: Domaines de recherche du TA

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

2. Définition

Le traitement automatique des langues naturelles (TALN) est un domaine à la frontière de la linguistique et l'informatique, il a pour objectif de développer des logiciels capables de traiter de façon automatique des données linguistiques exprimées dans une langue naturelle donnée et pour une application bien définie. Cet objectif passe nécessairement par l'explicitation des règles de la langue puis les représenter dans un formalisme calculable et enfin les implémenter à l'aide des programmes informatiques [2].

Parmi les applications les plus connues du TALN, on peut citer :

- la traduction automatique
- la correction orthographique
- la recherche d'information et la fouille de textes

- le résumé automatique
- la génération automatique de textes
- la synthèse de la parole
- la reconnaissance vocale
- la reconnaissance de l'écriture manuscrite .

3. Objectifs du TALN

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

Avec l'avènement des documents électroniques, des quantités phénoménales d'informations sont générées. Cette montée en volume de textes nécessite la production d'outils informatiques performants dont latâche est de trouver et d'extraire l'information pertinente sous une forme condensée [3].

4. Histoire du Traitement automatique des langages Naturelles

Historiquement, Le traitement automatique du langage naturel (TALN) est né à la fin des années quarante dans un contexte scientifique imprimé par les premiers travaux sur la traduction mais aussi dans un contexte politique qui s'explique par la seconde guerre mondiale. Le but de ce point est de donner quelques dates marquantes dans le développement du traitement automatique du langage naturel à travers le monde. [4]

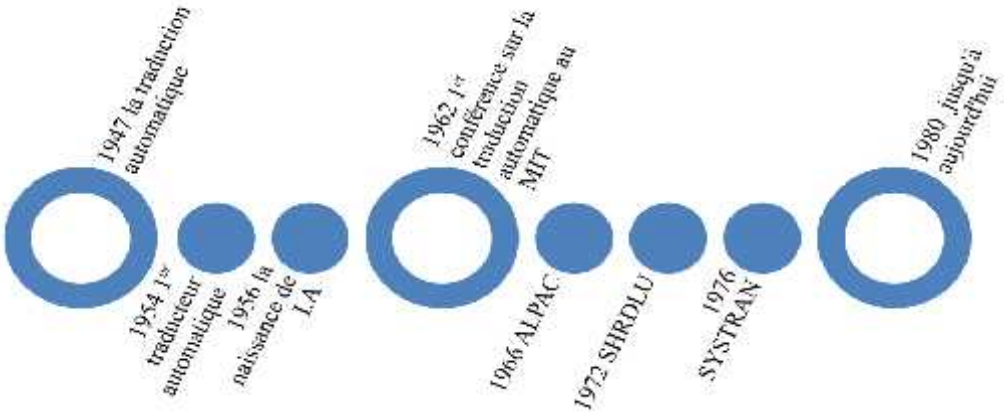

Figure 2: Les dates Marquantes dans l'histoire du TALN

❖ **1947** : Début des travaux sur la traduction automatique;

❖ Entre 1951 et 1954 : Zelling Harris publie ses travaux les plus importants de la linguistique (linguistique distributionnaliste) .

❖ **1954** : La mise au point du premier traducteur automatique (très rudimentaire) qui traduit du Russe à l'Anglais .

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

❖ **1957** : N. Chomsky publie ses premiers travaux sur la syntaxe des langues naturelles,et sur les relations entre grammaire formelles et grammaire naturelles .

❖ **1962** : la première conférence sur la traduction automatique est organisée au MIT (Institut Technologique du Massachussetts) par Y. Bar-Hillel .

❖ Entre 1961 et 1966 : beaucoup d'applications ont été mis en place tel que : BASBEL, SIR, STUDENT, ELIZA, …etc. Mettant en oeuvre des mécanismes de traitement simple, à base de mots clés .

❖ **1966** : L'histoire du TAL fait souvent celle des rendez-vous manqués et des désillusions cruelles Parmi ces faits marquants, on peut citer le rapport de la commission ALPAC (Automatic Language Processing Advisory Committee) en Anglais qui s'interroge sur l'utilité de poursuivre les recherches dans ce domaine .

❖ Dès lors, les crédits sont considérablement réduits et la recherche stagne jusqu'au début des années 70 .

❖ Depuis 1970, la plupart des recherches visent surtout la sémantique dans le cadre de la compréhension, mais aussi en parallèle les modèles syntaxiques connaissent en informatique des développements et des raffinements continus, et des algorithmes de plus en plus performants sont proposées pour analyser les grammaires les plus simples.

❖ **1972** : Terry Winogard, réalise le premier logiciel appelé SHRDLU capable de dialoguer en anglais avec un robot .

❖ **1976** : L'installation d'un système de traduction automatique commercial nommé SYSTRAN, la traduction automatique se fait connaître du grand public et suscite à nouveau l'intérêt des firmes privés que ce soit au Etats Unis ou au Japon .

❖ Entre 1980 jusqu'à aujourd'hui : La recherche en traitement automatique du langage naturel a connu depuis les années 80 jusqu'à nos jours une véritable progression, en termes de performance[1]qui se traduit d'un côté par la diversification des applications industrielles[2], mais aussi d'un autre côté par la création de plusieurs conférences internationales de renommées[3]et de laboratoires[4]de recherches à travers le monde.

5. Différents niveaux de Traitement

A partir des séquences de chaînes de caractères, différents niveaux d'analyses(Traitement) peuvent être envisagés. On parle dans la littérature : d'une analyse lexical, d'une analyse morphologique, d'une analyse syntaxique, d'une analyse sémantique et enfind'une analyse pragmatique.[2]

Le traitement automatique des langues se heurte à deux difficultés :

- ✓ L'ambiguïté de la langue : elle concerne les différents types d'ambiguïté propres à chaque niveau d'analyse. On parle souvent d'ambiguïté morphologique, d'ambiguïté syntaxique, d'ambiguïté sémantique, et d'ambiguïté pragmatique.
- ✓ La complexité des connaissances qui doivent être mises en œuvre à tous les niveaux d'analyse.

Dans ce qui suit nous allons décrire brièvement les différents niveaux d'analyse d'un texte en langue nature

[1] Le rendement des solutions proposées atteint aujourd'hui un certain seuil de fiabilité, qui se traduit par des pourcentages élevées de bon traitement.

[2] Beaucoup de grandes entreprises, se sont alliées à ce domaine tel que : IBM, XEROX, …etc.

[3] Exemple de conférences : ATALA, ACL-EACL, ANLP, ICASSP, …etc.

[4]Exemple de laboratoire : CERTAL (Centre d'Etude et de Recherche en Traitement automatique des Langues) en France. CRSTDLA (Centre de Recherche Scientifique et Technique pour le Développement de la Langue Arabe) en Algérie.

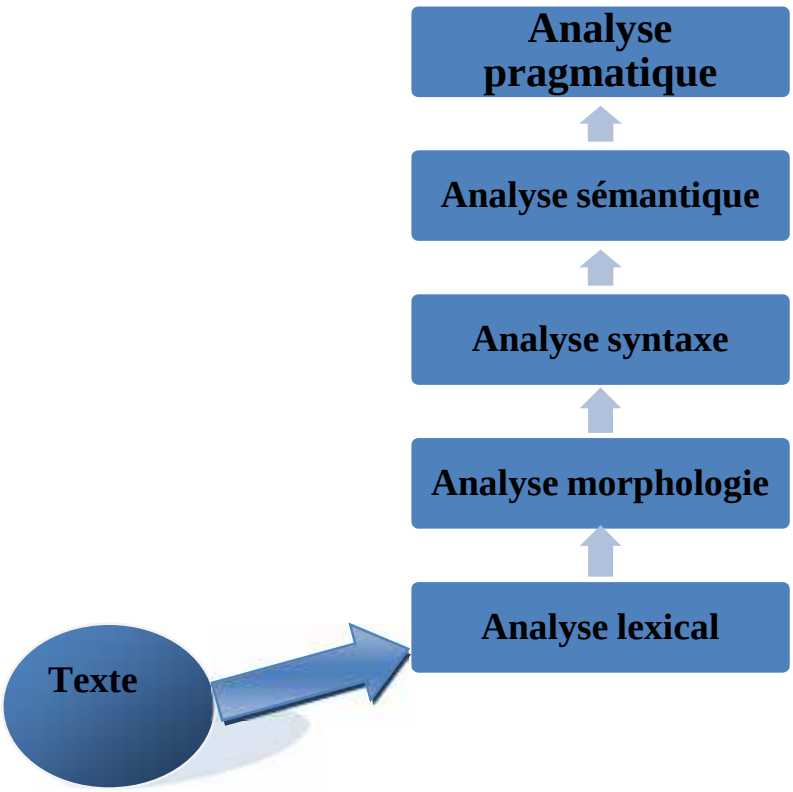

Figure 3: Les différents niveaux d'analyse d'un texte [2]

5.1 Analyse lexical

Le but de cette étape de traitement est de passer des formes atomiques identifiées par le segmenteur aux mots, c'est-à-dire de reconnaître dans chaque chaîne de caractère une (ou plusieurs) unité(s) linguistique(s), dotée(s) de caractéristiques propres (son sens, sa prononciation, ses propriétés syntaxiques, etc.).[5]

5.2 Analyse morphologique

L'analyse morphologique est indispensable pour tout système de traitement automatique de la langue naturelle, cette analyse permet de regrouper les mots en classes utilisables par les autres niveaux d'analyse. La définition de ces classes varie en fonction des traitements envisagés. A chaque classe on associe une étiquette appelée catégorie grammaticale ou catégorie lexicale. Il arrive qu'un même mot peut avoir différentes catégories grammaticales, on dit qu'il y a ambiguïté grammaticale ou une homographie.[2]

Exemple :

Soit le mot : WAEALOMADOPAOATI والمدرسة

Ensemble des préfixes du mot = {WA, EALO, MA}

Ensemble des suffixes du mot = {AT, I}

Les valeurs grammaticales associées sont :

- WA : conjonction
- EALO : Déterminant
- MA : Préfixe faisant partie du schème
- AT : Marque du féminin (désinence)
- I : Flexion (génitif)
- (DRS, R1OR2AR3) : Base du mot (analysable en racine et schème)

5.3 Analyse syntaxique

La syntaxe est l'étude des contraintes portant sur les successions licites de formes qui doivent être prisesen compte lorsque l'on cherche à décrire les séquences constituant des phrases grammaticalement correctes toutes les suites de mots ne forment pas des phrasesacceptables. Les contraintes envisagées sont de nature variée et correspondent à des propriétés sélectionnelles(telles que les règles d'accord en genre, en nombre, en cas, ...) ou positionnelles (telles que celles qui contrôlent les positions relatives des mots dans la phrase, ..).[6]

Exemple 1 :

Reprenons l'exemple précédant : Samy a mangé des pommes, et sa représentation morphologique [5] :

U1= Samy, U2 = a mangé, U3 = des, U4 = pommes.

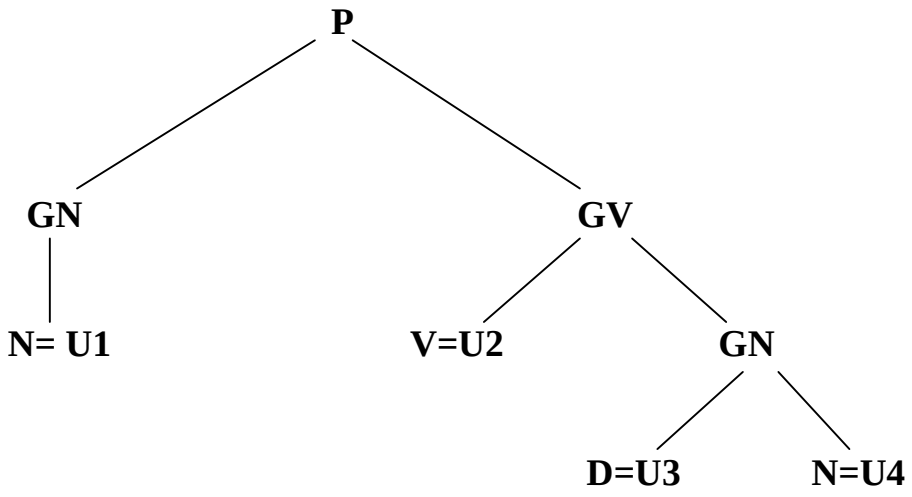

Figure 4. Arbre syntaxique d'une chaîne de caractères typographiques

P = Phrase => Samy a mangé des pommes

GN = Groupe Nominal

GV = Groupe Verbal

N = Nom => Samy

V = Verbe => a mangé

D = Déterminant => des

N = Nom => pommes

Exemple 2 :représentation arborescente de la phrase «شرب الولد الماء».

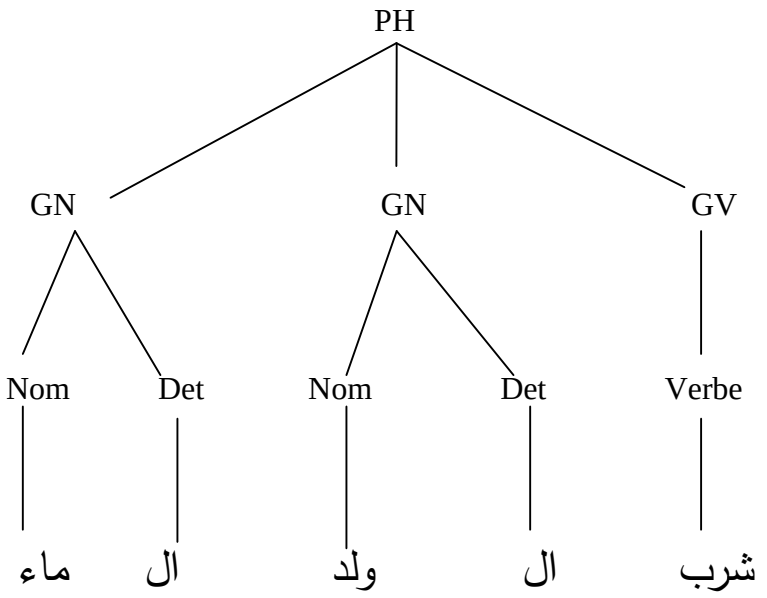

Figure 5 : Arbre syntaxique

5.4 Analyse sémantiques

La définition que nous donnons ici de la sémantique est assez proche de celle d'un modèle en logiqueformelle. La sémantique, en logique, repose sur le choix d'un ensemble appelé domaine. À chaque constanteintervenant dans les formules logiques, on associe un élément du domaine, et à chaque prédicat d'arité n onassocie une relation n-aire sur ce domaine. Comprendre le sens d'un énoncé linguistique revient, en premièreapproximation, à constituer une expression de type logique[5]qui

[5]La logique n'est pas de seul formalisme possible de représentation de ces relations conceptuelles, et d'autres modes de représentation sont également utilisés, en particulier les graphes conceptuels.

renvoie à une relation entre des objets dela situation considérée. La construction du sens correspondant à la phrase se fait de proche en proche, àpartir du sens trouvé pour les constituants. [7]

Exemple :

Proposition	**Négation**
شرب الولد الماء	لم يشرب الولد للماء

La logique des propositions ne s'intéresse pas au contenu des propositions maisseulement à leurs valeurs de vérité.

La logique des prédicats est une autre forme de représentation des énoncés. Parprédicat on entend une propriété telle que 'Homme', 'Mortel', etc.

Dans cette logique, pour créer des propositions, il faut combiner un prédicat avecun argument.

5.5 Analyse pragmatique

Pour pleinement comprendre un ou un texte dans son ensemble, il faut aussi avoir des connaissances pragmatiques, c'est-à-dire, celles qui permettent de situer le mot dans le contexte. Les connaissances pragmatiques précisent une représentation du monde référence qui constitue la culture commune nécessaire aux interlocuteurs.

Le niveau pragmatique est le niveau le plus difficilement accessible aux machines car certains énoncés ne se comprendre que dans un contexte géographique, historique ou culturel donné.[8]

6. Domaines d'application

Il est classique de présenter le domaine en l'organisant en grandes tâches, aux entrées/sorties biendentifiées la traduction automatique, la production de résumés, la génération d'énoncés ou de textes, l'interrogation en langage naturel de bases de données, la synthèse de la parole à partir du texte sont ainsi quelques exemples de ces tâches que l'on appelle tâches finalisées. En schématisant grossièrement, ces tâches finales s'organisent en trois types principaux : [4]

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

- Correction orthographique ou de syntaxe.
- Intégrée à toute application informatique impliquant la rédaction
- Correction basée sur des lexiques
- Ex : traitement de texte, courrier électronique, navigateur Internet (zone de saisie)
- Génération de texte à partir d'une description formelle.

- ♣ Atelier d'aide à la rédaction.
- ♣ Correction grammaticale.
- ♣ Reconnaissance de caractères (OCR pour Optical Character Recognition en Anglais).
- ♣ Apprentissage assisté par ordinateur des langues naturelles.

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

- ♣ Traduction automatique (ou l'aide à la traduction automatique).
- ♣ Résumé.
- ♣ Recherche et extraction d'information.
- ♣ Le routage, classement ou l'indexation automatique de documents électroniques sont des variantes applicatives du paradigme de la recherche documentaire.
- ♣ La recherche de documents « intéressants » dans des bases documentaires.
- ♣ L'analyse d'un corpus de documents relatifs à un thème donné (histoire, veille technologique, etc.).

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

- ♣ Agents dialoguant par téléphone.
- ♣ Assistants virtuels.
- ♣ Reconnaissance de la parole.
- ♣ Reconnaissance de la parole ou commande vocale (Reconnaissance vocale de Windows, Systèmes de navigation routière GPS, Smartphone...).
- ♣ Synthèse de la parole (Créer de la parole artificielle à partir d'un texte quelconque).
- ♣ Interfaces vocales (reconnaissance, synthèse, génération de dialogue, gestion du dialogue, accès aux bases de connaissance, etc.). [4]

7. Conclusion

Dans ce chapitre nous avons défini notre champ de recherche à savoir le TALN comme étant un domaine à la frontière de la linguistique et l'informatique, dont l'objectif est l'élaboration de programmes informatiques capables de traiter de façon automatique les langues naturelles,et nous avons présenté les différents niveaux et les domaines d'application du traitement automatique de langue naturelle.Dans le chapitre qui suit, nous mettre la lumière la langue arabeest sa position par rapport au processus d'automatisation.

Chapitre 2

LA LANGUE ARABE VS TALN

1. Introduction

La langue arabe est l'une des langues les plus parlées et utilisées dans le monde, elle occupe actuellement la cinquième place avec plus de 330 millions d'arabophones, tout en devenant la langue officielle de plus de 22 pays [8].

On distingue l'arabe classique et l'arabe moderne (MSA).L'arabe classique est la forme littéraire utilisée par tout pour les besoins de l'écriture et de l'imprimerie. C'est aussi la langue de la religion pour les musulmans, quelle que soit par ailleurs leur langue vernaculaire[1].

L'arabe moderne, dérivé de l'arabe classique, est la langue de la presse, des débats politiques, des textes scientifiques et de plus en plus celle des textes littéraires profanes. L'arabe est une langue très riche et différente des langues occidentales [9].

[1]C'est la langue locale parlée au sein d'une communauté.

.

Figure 6 : l'Expansion géographique de la langue arabe[9].

Dans ce chapitre, nous commencerons par présenter les caractéristiques de la langue arabe. Ensuite, les variantes travaux sur l'automatisation de la langue arabe. Enfin nous allons clôturer ce chapitre par un aperçu sur les différents problèmes que rencontre traitement automatique de la langue arabe.

2. Caractéristiques de la langue arabe

L'alphabet de la langue arabe compte 28 consonnes (Tableau 1). L'arabe s'écrit et se lit de droite à gauche les lettres changent de forme de présentation selon leur position (au début, au milieu ou à la fin du mot). Le Tableau 1 dans l'annexe montre les variations de la lettre.[10]

Par Exemple :

à la fin d'une lettre non joignable	à la fin	au milieu	au début
ع	ـع	ـعـ	عـ

Tableau 2: Exemple de variation de la lettre "ع"*A "y"*

Un mot arabe s'écrit avec des consonnes et des voyelles. Les voyelles sont ajoutées au-dessus ou au-dessous des lettres (◌, ◌, ◌, ◌). Elles sont nécessaires à la lecture et à la compréhension correcte d'un texte, elles permettent de différencier des mots ayant la même représentation. Le Tableau 3 donne un exemple pour les mots (كتب) "livres"et(مدرسة) "école".

Pendant, les voyelles ne sont utilisées que pour des textes sacres et didactiques. Les textes courants rencontrés dans les journaux et les livres n'en comportent habituellement pas.

De plus certaines lettres comme lAle peuvent symboliser le آ, أ ou إ; de même que pour les lettres ىet ەqui symbolisent respectivement(ي) "y"et (ة) "t" .

Mot sans voyelles	1ière Interprétation		2ième Interprétation		3ième Interprétation	
كتب	كتب	il a écrit	كتب	Il a été écrit	آتب	des livres
مدرسة	مدرسة	Ecole	مدرسة	enseignante	مدرسة	enseignée

Tableau 3: ambiguïté causée par l'absence de voyelles pour les mots (كتب) "livres" et (مدرسة)"école" .

2.1 Morphologies arabe

La morphologie traite la façon dont les mots sont construits. L'unité de base de la morphologie est le morphème. Les langues diffèrent au niveau des mécanismes par lesquels les morphèmes sont liés pour construire les mots .

La langue arabe a une structure morphologique très complexe comparativement aux langues indo-européennes. Les grammairiens arabes classiques ont adopté trois catégories pour classifier les mots de la langue Arabe, à savoir les noms (les adjectifs, les adverbes et les pronoms sont considérés comme des noms), les verbes et les particules (Fig. 5).[12]

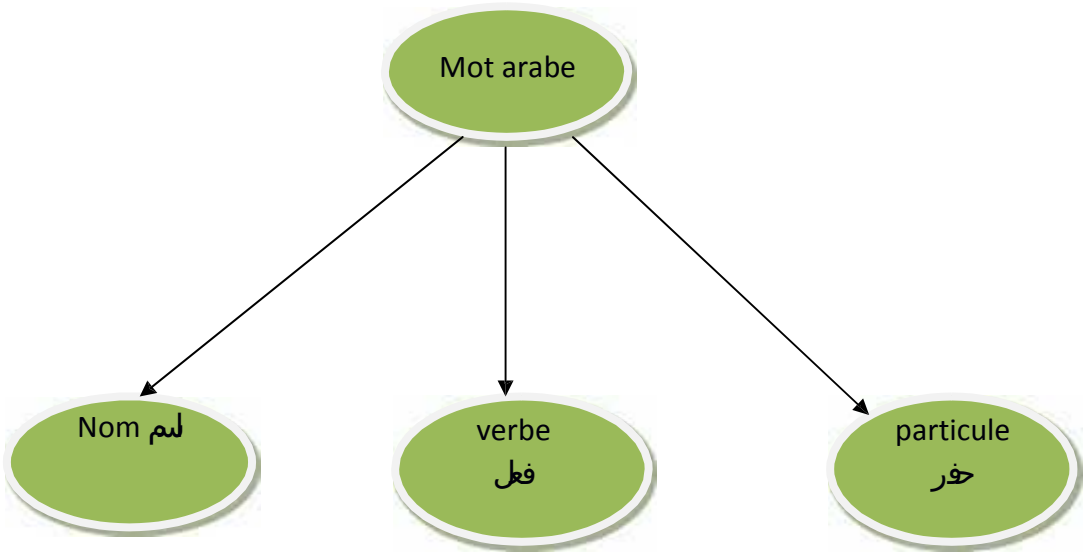

Figure 7: classification des mots de la langue Arabe selon la catégorie grammaticale .

Les noms (الأسماء) possèdent un sens propre indépendant du temps, tel que "homme" (رجل) ou "livre" (كتاب). Les verbes (الأفعال), par contre, ont un sens qui dépend du temps, par exemple "écrire" (كتب) ou "lire" (قرأ). Alors que Les particules (الحرفو)
possèdent un sens dépendant des autres mots de la phrase comme la particule "dans"(في)
Dans la morphologie arabe, nous pouvons distinguer deux systèmes principaux : la morphologie dérivationnelle (لإشقاق) et la morphologie flexionnelle (التصريف ا). La morphologie dérivationnelle s'intéresse à la création de nouveaux mots à partir de racines ou mots existants alors que la morphologie flexionnelle étudie les relations entre les catégories grammaticales des différentes formes fléchies (genre, nombre, temps, mode etc).[12]

2.1.1 Morphologie dérivationnelle

ل La majorité du vocabulaire de la langue Arabe est construite selon le formalisme racine-schème (جذر -نزو). En effet, pour exprimer certains termes sémantiques (i.e. les mots), une racine consonantique portant les informations sémantiques de base est associée à un ensemble limité de schèmes et en utilisant une séquence fixe de consonnes, de voyelles, de préfixes et de suffixes optionnels .

ل Les dictionnaires arabes, traditionnels et modernes, sont majoritairement ordonnés par racine. En effet, au lieu de lister les entrées par ordre alphabétique, ces dictionnaires organisent les mots sous les rubriques de leurs racines. Un des dictionnaires arabes standards, (لسان العرب), comporte 6350 racines trilitères et 2500 quadrilatères.

- Les mots dérivants d'une même racine constituent ce que nous appelons traditionnellement un champ morpho-sémantique, où les attributs sémantiques sont assignés à travers des schèmes régis par des règles morphologiques. Le sens qui est inhérent à la racine est partagé par tous les mots appartenant au même champ. Cependant, les schèmes qui les produisent les rendent sémantiquement distinguables
- Traditionnellement, les grammairiens arabes ont utilisé les lettres "ع" "ق' et "ل' comme lettres génériques pour représenter la racine et les schèmes. Pour les mots dérivés, l'ordre de ces lettres est toujours le même:" قest première lettre, عest la seconde et لreprésente le reste des lettres. La majorité des mots arabes sont formés par des radicaux de 3 consonnes tel est le cas du verbe "كتب" (écrire) et éventuellement 4 consonnes tel est le cas du verbe "دحرج"(glisser). Ces Racine peuvent donner naissance à plusieurs schèmes à la suite d'une ou plusieurs transformations morphologiques (par exemple, le redoublement d'une consonne, l'allongement d'une voyelle, l'adjonction d'un morphème, etc.) . La figure suivante montre ce processus de dérivation pour quelques verbes et noms à partir des racinesrilitères et quadrilitères.

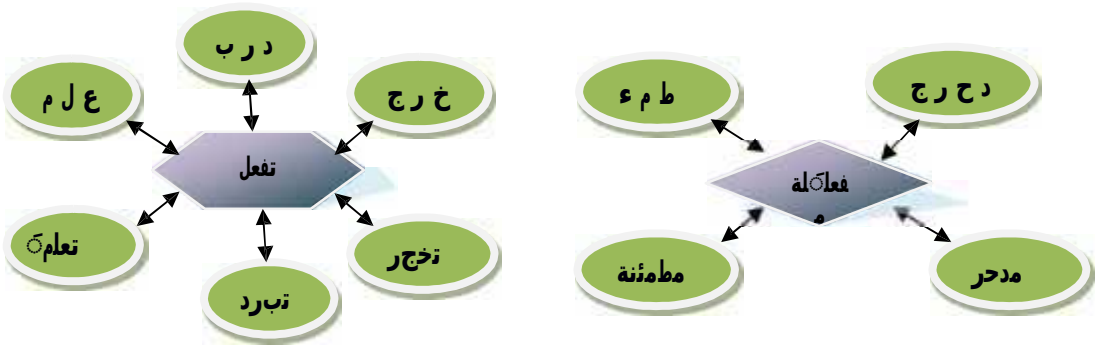

Figure 8 : Exemple de mots dérivés à partir du schème verbal «فعل»et schème nominal «مفعل»

Beaucoup de schèmes sont sémantiquement et morphologiquement ambigus. En effet, un seul schème peut être associé à différents concepts sémantiques et peut être utilisé tant pour les verbes que pour les noms et pour le singulier et le pluriel. Néanmoins, il y a aussi des schèmes qui sont utilisés exclusivement pour les verbes ou pour certaines catégories des noms. Par exemple, le schème فعيل correspond à la notion du diminutif, comme pour غصين (branchette).[11]

- **Dérivation des verbes**

La dérivation des verbes en arabe se fait à l'aide d'un nombre de schèmes limité. Les verbes simples à racine trilitère (الثلاثي المجرد) sont dérivés selon trois schèmesفعل (,فعل,فعل) ,tandis que les verbes simples à racine quadrilitère (الرباعي المجرد) suivent un seul schème (فعلل). Les transformations morphologiques, subies par ces verbes, produisent des verbes dits augmentés (الافعاال لمزيدة) qui suivent 15 schèmes dont 12 pour les verbes à racines trilitères et 3 pour les verbes à racines quadrilitère .

Il est à noter qu'une racine donnée n'est pas nécessairement compatible avec tous les schèmes. A titre d'exemple, le schème افعلest incompatible avec la racine دخل(en effet, le verbe اندخل n'est pas utilisé dans l'arabe standard). Les tableaux 4 et 5 présentent les schèmes verbaux simples et augmentés Voir le tableau 5 dans l'annex. [12]

	Les schèmes verbaux	Exemple
Trilitère simple (الثلاثيا لمجرد)	فعل	كتب (Ecrire)
	فعل	فهم(Comprendre)
	فعل	كبر (Grandir)
Quadrilatères simple (الرباعيا لمجرد)	فعلل	دحرج (Glisser)

Tableau 4 : Schèmes verbaux simples

- **Dérivation des noms**

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

a. Participe actif (اسم الفاعل)

Le participe actif est dérivé du verbe (à la voix active) pour désigner l'entité qui a accompli l'acte exprimé parce verbe. Ce nom est utilisé pour indiquer des actes ou des états temporaires, transitoires ou accidentels. Par exemple, le nom "كاتب"(écrivain) est un participe actif dérivé du verbe "كتب"(écrire), à la voix active, pour désigner l'entité qui a effectué l'acte d'écriture. Les participes actifs sont formés en utilisant un ensemble limité de schèmes nominaux (voir le tableau 6 dans l'annex).[12]

b. Participe adjectif semi-actif (الصفة المشبهة)

Le participe adjectif semi-actif est un nom dérivé qui a le sens du participe actif. Alors que le participe actif indique un acte ou un état temporaire, transitoire ou accidentel, le participe adjectif semi-actif désigne une action continue, un état habituel, ou une qualité permanente. Le tableau suivant montre des exemples du participe actif semi-actif dérivé du schème verbale فعل: [12]

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

Tableau 7: Les schèmes du Participe adjectif semi-actif

c. Forme d'exagération (صيغة المبالغة)

La forme d'exagération est un nom dérivé donnant la signification du participe actif, mais avec un sens supplémentaire. Elle dénote, en plus de l'exécution d'un acte, un haut degré de qualité, ou indique un acte qui est exécuté fréquemment ou de manière intensive. Ce nom dérivé est généré à partir d'un certain nombre de schème nominaux qui comprennent les formes suivantes .[12]

Schèmes	Forme d'exagération
فعال	توبا ,*(miséricordieux)*, فتاك, *(fatal)*
فعول	صدقو ,*(véridique)*, غفور , *(indulgent)*,شكور , *(reconnaissant)*
مفعال	مدرار, *(torrentiel)*
فعيل	بصير , *(lucide)*,لبيب , *(expert)*
فعل	حذر ,*(prudent)*,فطن, *(perspicace)*

Tableau 8: Les schèmes du Participe adjectif semi-actif

Le participe passif est un nom dérivé utilisé pour désigner l'entité qui a subit l'action exprimée par le verbe. Par exemple "مكتوب" *(*écrit*)* est un participe passif dérivé du verbe "كتب"*(*écrire*)* our désigner l'entité quia été affectée par l'action d'écrire. Les participes passifs sont générés à partir des schèmes nominaux qui correspondent aux schèmes verbaux, comme indiqué dans le Tableau 9 Voice dans l'annexe .

e. Nom d'instrument (اسم الآلة)

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

Schème	***Nom d'instrument***
مفعل	بضع ,*(lancette)*
مفعل	فتاح , *(clés)*
مفعلة	مطرقة , *(marteau)*

Tableau 10: Les schèmes du nom d'instrument .

f. Nom verbal (المصدر)

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

2.1.2 Morphologie flexionnelle

La flexion en linguistique est une opération de dérivation, qui ne crée pas de nouveaux mots, mais qui permettant d'appliquer des modifications sur un lemme afin de dénoter des

traits grammaticaux souhaités. Elle possède deux catégories : la déclinaison pour le système nominal et la conjugaison pour les verbes. Toute langue utilisant cette opération est appelée langue flexionnelle, et l'arabe en est une. En arabe, la flexion se concrétise par l'ajout des suffixes et préfixes ux lemmes pour refléter des indices
d'aspects, de mode, de temps, de personne, de genre, etc. Dans la suite de cette section nous détaillons ces opérations selon les deux axes : déclinaison et conjugaison.[12]

- **Inflexion des verbes**

Selon les grammairiens arabes classiques, les verbes de la langue Arabe sont classifiés en trois formes : accompli (ماضي), inaccompli (مضاعر) et impératif (أمر). Les verbes à l'accompli indiquent un acte achevé, tandis que les verbes inaccomplis désignent un acte inachevé qui vient de commencer ou en cours. Les verbes accomplis, inaccomplis et impératifs diffèrent dans leur inflexion selon la personne, le nombre et le genre. Cette différence apparaît dans les préfixes et les suffixes. Nous commençons par une présentation du système flexionnelle des verbes à l'accompli.[12]

- **Inflexion à l'accompli**

L'inflexion à l'accompli est réalisée en attachant aux verbes des suffixes qui indiquent la personne, le nombre et le genre. Ces suffixes sont identiques dans les deux voix : active
(المبني للمعلوم) et passive (المبني للمجهول) .
Les tableaux suivants [11], [12] et [13] présentent respectivement les suffixes de l'accompli, un exemple de l'accompli à la voix active et un exemple de l'accompli à la voix passive [11]

	Singulier		Duel		Pluriel	
	Mas	Fem	Mas	Fem	Mas	Fem
1ier	ت	ت	نا	نا	نا	نا
2éme	ت	ت	تما	تما	تم	تن
3éme		ت	ا	تا	او	ن

Tableau 11: Les suffixes de l'accompli

	Singulier		Duel		Pluriel	
	Mas	Fem	Mas	Fem	Mas	Fem
1ier	كتبت	كتبت	كتبنَا	كتبنا	كتبنَا	كتبنَا
2éme	كتبت	كتبت	كتبتما	كتبتما	كتبتم	كتبتن
3éme	كتب	كتبت	كتبَا	كتبَتَا	كتبوا	كتبن

Tableau 12: Le verbe كتَب conjugué à l'accompli (voix active)

	Singulier		Duel		Pluriel	
	Mas	Fem	Mas	Fem	Mas	Fem
1ier	أمرت	أمرت	أمرنا	أمرنا	أمرنا	أمرنا
2éme	أمرت	أمرت	أمرتما	أمرتما	أمرتم	أمرتن
3éme	أمر	أمرت	أمرا	أمرا	أمراو	أمرن

Tableau 13: Le verbe أمر conjugué à l'accompli (voix passive)

- **Inflexion à l'inaccompli**

L'inflexion des verbes à l'inaccompli est obtenue en ajoutant des préfixes et des suffixes qui varient selon la personne, le nombre et le genre. Le paradigme de l'inaccompli inclut 3 modes, l'indicatif (المضاعر المرفوع), le subjonctif (المضاعر المنصوب) et l'apocopé (المضاعر المجزمو).
Signalons que les préfixes utilisés à la voix passive s'obtiennent à partir de ceux de la voix active en changement de la voyelle courte «◌ُ » par «◌ِ ». De plus, les suffixes de l'inaccompli subjonctif et apocopé s'obtiennent à partir de ceux de l'indicatif en supprimant la consonne «ن». Les tableaux suivants [14], [15] et [16] présentent les préfixes, les suffixes ainsi que la conjugaison d'un verbe à l'inaccompli indicatif aux voix active et passive.

	Singulier		**Duel**		**Pluriel**	
	Mas	Fem	Mas	Fem	Mas	Fem
1ier	أ	أ	ن	ن	ن	ن
2éme	ت	ت	ت	ت	ت	ت
3éme	ي	ت	ي	ت	ي	ي

Tableau 14: Les préfixes de l'inaccompli voix active

	Singulier		**Duel**		**Pluriel**	
	Mas	Fem	Mas	Fem	Mas	Fem
1ier						
2éme		ين	ان	ان	ون	ن
3éme			ان	ان	ون	ن

Tableau 15: Les suffixes de l'inaccompli indicatif .

	Singulier		**Duel**		**Pluriel**	
	Mas	Fem	Mas	Fem	Mas	Fem
1ier	أنصر	أنصر	ننصر	ننصر	ننصر	ننصر
2éme	تنصر	تنصرين	تنصرانِ	تنصرانِ	تنصرونَ	تنصرن
3éme	ينصر	تنصر	ينصرانِ	تنصرانِ	ينصرونَ	ينصرن

Tableau 16: Leverbe" "نصر conjugué à l'inaccompli indicatif (voix active)

- **Inflexion à l'impératif**

L'inflexion des verbes arabes à l'impératif se fait en ajoutant des suffixes aux verbes à la deuxième personne. Les tableaux suivants [17] et [18] ressentent respectivement les suffixes à l'impératif ainsi qu'un exemple de conjugaison. [14]

	Singulier		**Duel**		**Pluriel**	
	Mas	Fem	Mas	Fem	Mas	Fem
2ième		ي	ا	ا	او	ن

Tableau 17: Les suffixes de l'impératif

	Singulier		**Duel**		**Pluriel**	
	Mas	Fem	Mas	Fem	Mas	Fem
2ième	سرد	رهي	رها	رها	رهوا	رهن

Tableau 18: Le verbe سردconjugué à l'impératif

- **Inflexion des noms**

L'inflexion des noms en langue Arabe comporte trois cas (حالة إعرابية) : le nominatif (المرفوع), l'accusatif (المنصوب) et le génitif (المجررو) suivant les trois nombres : singulier (مفرد), duel (مثنى), et pluriel (جمع). Les noms en arabe sont majoritairement déclinables (معرب) c'est-à-dire qu'ils se mettent à l'un de ces trois cas suivant leur fonction dans la phrase . Il diffère selon la nature de la forme (simple, diptote, etc.) et le nombre de celle-ci (singulier, duel ou pluriel).[14]

- **Inflexion du singulier (المفرد)**

- **Les mots déclinables aux trois cas** :C'est le cas le plus fréquent, il prend la voyelle «◌ » (ضمة) comme une marque du nominatif; la voyelle «◌ » (فتحة) à l'accusatif et la voyelle «◌ » (كسر) au génitif. Quand le nom est indéfini, le tanwîn apparaît marqué

respectivement par les trois signes diacritiques : «◌ », « ◌», et «◌ ». A l'accusatif indéfini, excepté le cas des noms qui se terminent par «ة»ou par «ى », le caractère «ا » est ajouté à la fin du mot pour renforcer le *tanwîn*: par exemple, à l'accusatif indéfini, le nom كتاب(livre) produitكتابا(livre à l'accusatif indéfini) et le nom جزيرة(île) produit جزيرة(île à l'accusatif indéfini).[14]

- **Les diptotes (الممنوع من الصرف):**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 voyelle «◌»(فتحة). 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** tel que "صحراء"(désert), les adjectifs masculins de couleurs ayant pour schème أفعل tel que " أحمر"(rouge) et ceux qui sont féminins de schème فعلاء tel que بيضاء(blanche).[1]
- **Les cinq noms (الأسماء الخمسة) :**Les noms أبو (père) ,أخو (frère), حمو (beau-père), فو (bouche) et ذو (possesseur) prennent la voyelle longue « و » au nominatif ; la voyelle longue « ا » à l'accusatif et la voyelle longue « ي » au génitif lorsqu'ils sont définis par un complément.[1]

- **Inflexion du duel (المثنى)**

L'inflexion du duel en langue Arabe est marquée par la voyelle longue « ا » dans le cas du nominatif et par la voyelle longue « ي » à l'accusatif et le génitif. Dans le cas du duel nom indéfini ou défini par l'article, la consonne «ن » est ajoutée aux marques de déclinaison. Par exemple, le duel du nom رجل (homme) prend la forme "رجلان " (deux hommes, au nominatif) et "رجلين " (deux hommes, à l'accusatif et au génitif). [11]

- **Inflexion du pluriel (الجمع)**

Il existe deux grandes catégories de pluriel en arabe [11]

- **Les pluriels réguliers (الجمع السالم):**Ces pluriels sont formés par l'ajout d'un suffixe au singulier sans changement de la structure du mot. Nous distinguons :
- **Le pluriel régulier masculin (جمع المذكر السالم) :** au nominatif, ce pluriel prend le suffixe « ونo » et dans le cas de l'accusatif et le génitif le suffixe « ين » est ajouté. Notons que dans le cas de la définition par annexion (التعريف بالإضافة) la consonne « ن » est supprimée du suffixe. Par exemple, le singulier " مدرس"(enseignant) devient "مدرسون"(des enseignants) au nominatif et مدرسين à l'accusatif et au génitif.

- **Le pluriel régulier féminin (جمع المؤنث السالم) :** le suffixe ajouté dans le cas de ce pluriel est « ات» auquel s'ajoute la voyelle «ِ » au nominatif et la voyelle« ِ» à l'accusatif et génitif. Par exemple, le mot " شجر"(arbre) devient اشجرت(arbres) au nominatif et اشجرتà l'accusatif et au génitif.
- **Les pluriels brisés (جمع التكسير) :** ces pluriels doivent ce nom aux modifications et infixations qu'ils causent par rapport à la forme du singulier, à la différence des pluriels réguliers (masculin et féminin). Les formes du pluriel brisé sont très nombreuses et généralement imprévisibles. Par exemple : le nom مفتاح(*clé*) se transforme pour donner les deux formes plurielles مفَاتيح et مفاتح(*clés*).[11]

3. Les variantes de langues arabes

Langue arabe est un terme vague qui fait référence aux nombreuses variétés existantes de la langue arabe. En effet, l'arabe possède plusieurs variantes depuis ses débuts. Il est à noter de ce fait que même à l'époque préislamique, l'arabe possédait déjà des dialectes distincts en un nombre considérable, comme c'était le cas entre des dialectes des tributs de Qahtane, Adnane et Himyar. Selon il n'y pas d'accord sur le nombre de variétés réellement utilisées aujourd'hui, et par conséquent il existe plusieurs classifications pour ces variétés. Par exemple définit deux variétés : la variété élevée ou l'arabe classique et la variété basse utilisée dans la communication quotidienne des arabophones (les dialectes). Nous citons aussi certaines classifications faites de manière locale comme celle du sociolinguiste réalisée pour l'arabe en Egypte qui met en avant les cinq variétés suivantes :

1. L'arabe classique patrimonial (fuSha al-turaa∀,الترات الفصحى)
2. L'arabe classique contemporain (fuSha al-9aSr - ,العصر الفصحى)
3. Le familier des éduqués (9aamiyyat al-mu∀aqqafiin – العامية المثقفين)
4. Le familier des éclairés (9aamiyyat al-mutanawwiriin - العامية المتنورين)
5. Le familier des analphabètes (9aamiyyat al-?ummiyyiin – العامية الاميين) [13]

4. Travaux sur l'automatisation de la langue arabe

En terme de travaux de référence concernant le traitement automatique de la langue arabe , il est impératif de citer l'étude pionnière de David Cohen « Vers un traitement automatique de l'arabe » qui datede 1960concernaient notamment le lexique et la morphologie[4]. Dès le milieu des années 1970, les travaux de chercheurs tels que Yahya Hlal, puis ceux de Fathi Debili dans les années 1980, ont montré la possibilitéd'un traitement automatique de la langue arabe. Dans les années 1990, on peut également citer aussi, toujours les travaux de

Joseph Dichy notamment dans le cadre du projet européen DIINARMBC (Dictionnaire informatisé de l'arabe, multilingue et basé sur corpus) [9] . Depuis plusieurs travaux ont vu le jour touchants différents thématiques : l'analyse morphologique (*AraMorph de Buckwalter etSebawi de Darwish*), l'analyse morphosyntaxique (*Al Khalil, de LaRI*[2]), la lemmatisation automatique (Khoja, Isri, JIdr, ….etc.), Résumeurs automatique (Lakhas de SofianDouzidia), la correction orthographique (les travaux de Khaled Shaalan), ains que d'autres projets.

5. Problèmes du traitement automatique de la langue arabe

L'arabe, comme toutes les langues naturelles, est caractérisée par un ensemble de phénomènes créant des difficultés et des problèmes qu'il faut prendre en considération lors d'un traitement automatique. En plus des phénomènes classiques, comme l'ambiguïté, la coordination ou l'anaphore, nous trouvons aussi dans le cas de l'arabe d'autres phénomènes propres aux langues sémitiques tel que l'absence de voyelles, l'agglutination et l'ordre des mots dans une phrase. Dans la présente section, nous présentons les phénomènes que nous considérons les plus importants pour l'arabe.[14]

5.1 L'absence de voyelle – voyellation

Nous trouvons plusieurs définitions pour décrire le phénomène de la voyellation
qui est concrétisée par l'absence des voyelles courtes, appelées aussi les diacritiques, dans les textes en arabe. Cette absence génère plusieurs cas d'ambiguïté compliquant ainsi le traitement automatique. Ces ambiguïtés lexicales sont du essentiellement au fait que chaque consonne peut prendre l'une des sept voyelles de l'arabe, ce qui crée des combinaisons de mots dont le nombre diffère d'un mot non voyellé à un autreen fonction de l'existence de la combinaison obtenue dans le vocabulaire ou pas. Selon, l'absence de diacritiques en arabe entraîne une complexité de calcul d'un ordre de grandeur plus grand que la manipulation de ses homologues langues latines. [14]

[2]Laboratoire de Recherche Informatique « *LaRI*» de la Faculté des Sciences, Université Mohammed Premier, Oujda,

5.2 Agglutination

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

La base est une combinaison de lettres radicales (le plus souvent trois) et d'un schème. La base – avecpréfixe et suffixe - forme le noyau lexical, éventuellement entouré d'extensions . Comme le montre l'exemple suivant : ولبضربها Les éléments clitiques sont séparés par le symbole "+" :

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

Cet exemple révèle la complexité morphologique de l'arabe. Il s'agit du verbe يضرب employé au présent du subjonctif, 3ème personne du masculin pluriel, la base verbale est /ضرب/et la racine /ضرب/. Le pronom sujet n'est pas réalisé. En position proclitique, on utilise la conjonction de coordination "wa" وet la conjonction "li" ل.En position enclitique, on utilise le pronom complément d'objet 3ème personne du féminin singulier "haA" ها"elle".[13]

5.3 Ambiguïté lexicale et syntaxique

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

Par ailleurs, le problème ne réside pas dans l'analyse d'un langage ambigu en soi; mais c'est plutôt au niveau de son traitement de façon robuste et réaliste. En effet, après une première phase de segmentation du texte en unités lexicales, il est convenu de chercher dans le lexique

les interprétations correspondant à chacune d'entre elles. A chaque interprétation, nous associons une catégorie syntaxique reconnue par la grammaire.

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

5.4 Irrégularité de l'ordre des mots dans la phrase

La construction des phrases en arabe est flexible, dans le sens où l'ordre des mots dans une phrase donnée est relativement libre. Généralement, un mot placé au début de la phrase est un terme sur lequel nous voulons attirer l'attention, s'en suit le terme le plus long ou le plus riche en sens ou en sonorité. Cette flexibilité provoque des ambiguïtés syntaxiques artificielles due à la prise en compte de toutes les règles de combinaison possibles des composants d'une phrase. Pour illustrer cette propriété prenons les phrases suivantes :[14]

- Verbe + sujet + complément :
  (- L'Algérie s'est qualifiée pour la coupe du monde) تأهلت الجزائر إلى كأس العالم
- Sujet + verbe + complément :
  (- C'est l'Algérie qui s'est qualifiée en coupe du monde) الجزائر تأهلت إلى كأس العالم
- Complément + verbe + sujet :
  (- C'est pour la coupe du monde que l'Algérie s'est qualifiée) إلى كأس العالم تأهلتا لجزائر

6. Conclusion

Les différentes recherches effectuées sur l'analyse de la morphologie de la langue arabe montrent que c'est une langue très difficile à traiter à cause de l'agglutination et des ambiguïtés graphiques.

Dans ce chapitre, nous avons explicité les différentes connaissances liées au traitement automatique de la langue arabe. On a commencé par une petite introduction , puis nous avons présenté la langue arabe et sa morphologie ,ensuite, nous avons mis le point sur les différents problèmes de traitement automatique de la langue arabe , finalement, nous avons présenté quelques travaux et produits de traitement automatique de la langue arabe .

Chapitre 3

DIACRITISATION AUTOMATIQUE

1. Introduction

L'absence de vocalisation dans les textes arabe est l'un de plus grand défis du traitement automatique de la langue arabe vu l'importance que jeu la voyellation dans le processus de compréhension global de l'énoncé. De ce fait, l'ordinateur a besoin de mécanismes et d'algorithmes pour restaurer les différentes formes de vocalisation.

Dans ce chapitre qui constitue le noyau de notre travaille nous présentons en détaille la démarche employé pour mettre en place notre outil de vocalisation que noua avons baptisé MOSHAKIL .

2. Outils de vocalisation de la langue arabe

2.1 sHarakat de Multillect[1]

- **Auteur** : fondation de Multilect
- **Principe** : Harakat de Multillect fournira des signes diacritiques pour les textes en arabe qui vont les rendre plus facile à lire et à comprendre. Cette application aidera à la fois des locuteurs natifs et ceux qui étudient l'arabe. Notre application vous aidera à économiser votre temps pour pouvoir développer les aptitudes de lecture et d'écriture. Elle peut être utilisée à la fois dans les écoles et les universités, et pendant l'étude personnelle de la langue arabe.Harakat de Multillect est un outil éducatif utile et une plate-forme unique pour résoudre des tâches pratiques en travaillant avec les textes en arabe.

2.2 Mishkal[2]

- **Auteur :**Taha Zerrouki
- **Principe** : Le programme fournit le service de traitement de texte arabe automatiquement, l'utilisateur peut corriger la composition ou l'effacer pour tout le texte.

2.3 AlKhalilDiacritizer

- **Auteur :**Taha Amine Chennoufi et Azzeddine Mazroui

[1] https://harakat.ae/fr (19/04/2018)
[2] http://tahadz.com/mishkal(23/05/2018)

♣ **Principe :** Nous présentons dans ce travail un nouveau système de voyellation automatique des textes arabes en utilisant trois étapes. Durant la première phase, nous avons intégré une base de données lexicale contenant les mots les plus fréquents de la langue arabe avec l'analyseur morphologique AlKhalil Morpho Sys pour fournir les voyellations possibles pour chaque mot.

Le second module dont l'objectif est d'éliminer l'ambiguïté repose sur une approche statistique dont l'apprentissage a été effectué sur un corpus constitué de textes de livres arabes et utilisant les modèles de Markov cachés (HMM) où les mots non voyellés représentent les états observés et les mots voyellés sont ses états cachés. Le système utilise les techniques de lissage pour contourner le problème des transitions des mots absentes et l'algorithme de Viterbi pour sélectionner la solution optimale. La troisième étape utilise un modèle HMM basé sur les caractères pour traiter le cas des mots non analysés.[15]

2 .4 ArabDiac2 .0

♣ **Auteur :**société d'ingénierie pour le développement de systèmes informatiques (RDI)

♣ **Principe :** est un outil de vocalisation de texte arabe fourni par RDI, il est construit sur l'infrastructure de traitement automatique de langue naturel de RDI (Analyseurs morphologiques, étiqueteurs …etc.), comme le projet Sakhr[4]. le système est totalement ferméa précision est supérieur à 95% mesurée au niveau du mot [16]

3. Approches de vocalisation

En se référant aux travaux de recherches antérieurs, nous pouvons diviser les tentatives de voyellation automatique des textes arabes en trois parties: approches fondées sur des règles, approches statistiques et approches hybrides.

3.1 Approches fondées sur les règles

Dans ce cadre, certains travaux ont eu recours à la programmation des règles linguistiques vocales, morphologiques et syntaxiques pour la voyellation des mots arabes. El-Sadany et Hashish [22] ont mentionné une méthode se basant sur des règles morphologiques pour la voyellation semi-automatique des verbes arabes. Aussi, Debili et Achour [18] ontétudié l'impact de l'analyse lexicale, l'analyse morphologique et l'étiquetage syntaxique pour dissiper l'ambigüité dans le processus de voyellation des textes arabes. L'absence d'un système de voyellation des textes arabes basé uniquement sur les règles est due aux taux élevé d'ambigüités, de l'existence d'un nombre important de règles morphosyntaxique set l'absence d'un analyseur syntaxique efficace.

3.2 Approches statistiques

L'approche statistique est apparue en 2002 avec les travaux de GAL [17]qui a présenté une approche markovienne pour la voyellation du Coran pour la langue arabe et les textes de l'Ancien Testament pour la langue hébraïque. Schlippe, Nguyen et Vogel [19] ont mis au point un système de voyellation des textes arabes basé sur la traduction automatique. Al Ghamdi, Muzaffar et Alhakami [20] en 2010 ont présenté le système de voyellation KAD(The Arabicdiacritizer) basé sur les 4-grammes au niveau des lettres. Enfin on pour a citer le travail de Hifny [21] qui a présenté une méthode purement statistique basée sur les n-grammes et utilisant quelques techniques de lissage qui prélèvent une masse de probabilité sur les transitions observées, et cette masse est redistribuée sur les événements non observés.

3.3Approches hybrides

Ce sont les approches qui combinent les règles linguistiques et les traitements statistiques afin d'exploiter les points forts des deux méthodes. Parmi les travaux importants, on peut citer le système de voyellation ArabDiac développé par la société RDI [7]. Ce système utilise l'analyseur morphologique Arab Morpho et l'étique teurArab Tagger puis les n-grammes. En 2009 Zitouni et Sarikaya ont présenté un système de voyellation utilisant un classificateur statistique basée sur l'entropie maximale. Leurs caractéristiques sont basées sur des caractères simples du mot, segments morphologiques et l'état syntaxique pour atteindre le meilleur classement de mots.

4. Architecture du Système « MOSHAKIL »

Dans cette section on présentera les grands axes concernant notre système de voyellation que nous avons baptisé «MOSHAKIL», c'est un outils qui permettra d'intégré dans un texte arabe non vocalisé des signes diacritiques comme : Fetha « ◌َ », Kesra « ◌ِ », dumma « ◌ُ », ainsi que d'autres signes. L'originalité de notre travail, réside dans le fait de s'inspirer du processus de traduction automatique qui consiste à traduire un texte d'une langue source vers une langue cible, dans notre cas la langue source va être considérée comme étant le texte sans vocalisation et la langue cible est le texte arabe vocalisé.

Pour atteindre cet objectif on va suivre la même démarche qu'un processus de traduction automatique statistique

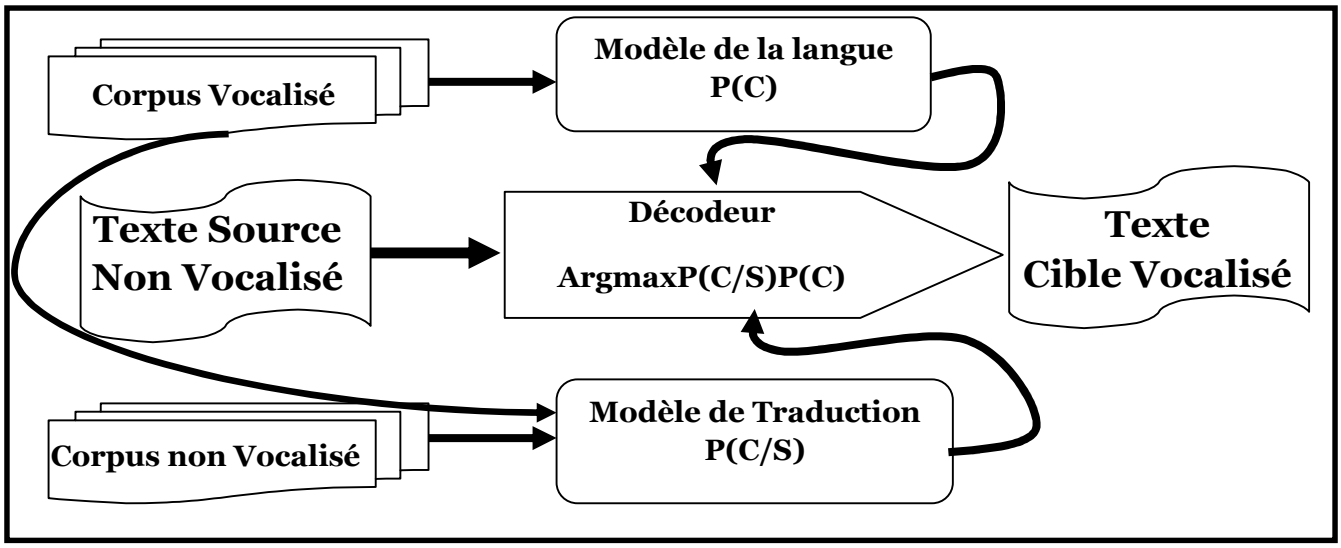

Figure 9 : Architecture générale du Modèle de Vocalisation

4.1 Pré traitement (Construction d'un Corpus parallèle)

4.1.1 Présentation

La phase de pré traitement sera dédiée, à la mise en place d'un corpus parallèle dédié à la fois pour création du modèle de vocalisation, mais aussi au test et à l'évolution de ce modèle.

- **Cas de la traduction :**Un corpus de données bilingues « parallèle » est un corpus qui contient des paires de documents bilingues ou des paires de phrases bilingues qui sont la traduction directe l'un de l'autre. Les deux phrases (ou documents) parallèles sont souvent de longueur similaire. L'ordre des phrases parallèles est maintenu dans deux documents parallèles . Une phrase dans la langue L1 est souvent traduite par une phrase dans la langue L2 (nous appelons cela une correspondance). [27]
- **Cas de la vocalisation :** L'objectif de cette étape est la normalisation du corpus que nous avons choisi pour élaborer notre démarche de vocalisation, à avoir : Tashkeela [28] Ce corpus est composé de 75 million de mots arabe entièrement vocalisé manuellement , le contenu de ce corpus est issu de la bibliothèque « المكتبة الشاملة[3] » .

Vue la taille du corpus et la quantité de mots qu'il contient nous avons opté, pour sélectionner qu'une partie de ce corpus. Le tableau ci-dessous, donne des indication sur la partie du corpus que nous avons utilisé pour l'élaboration du modèle de vocalisation, ainsi que les tests :

[3]Constitué de 97 livres de la langue arabe classique et moderne.

Corpus	Propriété
Numéro de linge	500000
Nombre de caractères	9967954
Nombre de mots	2501914

Tableau 19 : parte de la corpus tashkila

4.1.2 Normalisation du Corpus

Il important de préciser que le corpus dans son état initial ne sera pas exploitable par rapport à notre démarche de résolution, pour cette raison une série d'opération seront effectuées, à savoir :

- **Etape 1 :** Supprimer les lignes vides ;
- **Etape 2 :** Supprimer lignes avec un seul caractère ;
- **Etape 3 :** ignorer les lignes sans signe de vocalisation ;
- **Etape 4 :**segmenter chaque mot de telle manière d'obtenir une paire (lettre, voylle) ;

Exemple : Soit la phrase suivante :

« كِتَابُ الْوُضُوءِ مَا جَاءَ فِي الْوُضُوءِ قَبْلَ سَحْنُون»

«كِ تَ ابُ # الْ وُ ضُ وءِ #مَ ا #جَ ا ءَ #فِ ي # الْ وُ ضُ وءِ #قَ بْ لَ #سَ حْ نُ ون»

- **Etape 5 :** Générer la version non-vocalisée : nous faisons supprimer la vocalisation de la partie dédié à l'élaboration du modèle.

Exemple :

«آتِ يَ ك #بِ السلْ عَةِ # عَ دًّ ا #أَ وْ #بَ عْ دَ # عَ دٍ #فَ هَ ذَ ا #لَ ا #بَ أْسَ #»

«آت ي ك #ب السل ع ة # غ د ا # أ و #ب ع د # غ د #ف ه ذ ا #ل ا #ب أس #»

Ala fin de cette phase de prétraitement, nous allons obtenir deux corpus parallèles (l'un contient un texte vocalise et l'autre texte non vocalise), c'est deux corpus vont être scindés en deux partie une partie pour l'élaboration et l'apprentissage du modèle est l'autre pour le test

Exemple : soit l'extrait du corpus Parallèle

Texte non vocalisé	Texte vocalisé
ث م # ب ل غ # ا ل ص ب ي # و ع ت ق # ا ل ع ب د	ثُ م # بَ لَ غَ # ا ل ص بِ ي # وَ عَ تَ قَ # ا لْ عَ بْ د
ل م # ت ن ف س خ # ا ل إ ج ا ر ة	لَ مْ # تَ نْ فَ س خْ # ا لْ إِ جَ ا رَ ةُ
ه ذ ا # ا ل م ذ ه ب	هَ ذَ ا # ا لْ مَ ذْ هَ بُ
و ع ل ي ه # ا ل أ ص ح ا ب	وَ عَ لَ يْ هِ # ا لْ أَ صْ حَ ا بُ
و ق ط ع # ب ه # ك ث ي ر # م ن ه م	وَ قَ طَ عَ # بِ هِ # كَ ثِ ي رٌ # مِ نْ هُ مْ
م ن ه م # ص ا ح ب # ا ل ه د ا ي ة	مِ نْ هُ مْ # صَ ا حِ بُ # ا لْ هِ دَ ا يَ ةِ
ذ ك ر و ه # ف ي # ب ا ب # ا ل ح ج ر	ذَ كَ رُ وهُ # فِ ي # بَ ا بِ # ا لْ حَ جْ رِ

Tableau 20 : Exemple de l'extrait du Corpus parallèle.

4.2 Traitement (Elaboration du Modèle)

Une fois la préparation du corpus achevé, on passera à l'étape suivante à savoir mettre en place le modèle qui va générer les signes de vocalisation.

Rappel : nous utilisons démarche similaire à celle de la traduction automatique statistique qui utilise des modèle purement stochastiques (Statistiques | probabilité) est dont le procédé s'appuie sur deux modèles, le premier est le modèle de traduction et le second est le modèle de langue.

- le modèle de traduction (qui est représenté par la probabilité) :

$$Pr(C|S) \qquad (1)$$

Qui peut être interprétée par le fait de calculer la probabilité d'avoir une phrase (un mot) vocalisé « C » sachant la Phrase (mot)non vocalisé « S ».

- le modèle du langage (qui est représenté formellement par la probabilité conditionnelle) :

$$Pr(C) \qquad (2)$$

Qui indique la probabilité d'avoir une suite de caractères et de voyelles correcte

4.2.1 Modèle d'alignement (Modèle de traduction)

- **Cas traduction :** Le rôle du modèle de traduction est de déterminer statistiquement qu'une phrase source donnée se traduise en une phrase cible équivalente. L'apprentissage d'un tel modèle repose sur

l'exploitation de corpus d'apprentissage parallèles (dits « alignés » ou « bilignues »). Ces corpus parallèles sont des corpus bilingues qui couplent deux ensembles de données textuelles alignés au niveau de la phrase et tels que l'un est la parfaite traduction de l'autre. De ce couple, on peut alors extraire des correspondances (ou alignements) entre deux langues comme l'illustre

- **Cas vocalisation :**Dans notre cas le procéder d'alignement se fait caractère par caractère entre une phrase non vocalisée et la phrase vocalisée, ainsi le but de notre modèle d'alignement consiste à extraire des relations d'appariement entre les caractères de la phrase (ou du mot) non vocalisé et les caractères de la phrase (ou du mot) vocalisé.

Pour accomplir cette tâche de correspondance on doit définir une variable « A » appelée alignement. Au sein d'une paire (phrase (mot) non vocalisé | phrase (mot) vocalisé), l'alignement précise les correspondances existantes entre les caractères de la phrase non vocalisé et ceux de la phrase vocalisé. Il définit donc pour chaque lettre source les différentes variations de vocalisation qui vont avec. Les correspondances sont symbolisées par des liens. Il important à préciser qu'a la différence entre la traduction automatique ou en peut avoir des alignements diagonaux, dans notre cas l'alignement se fait en vertical.

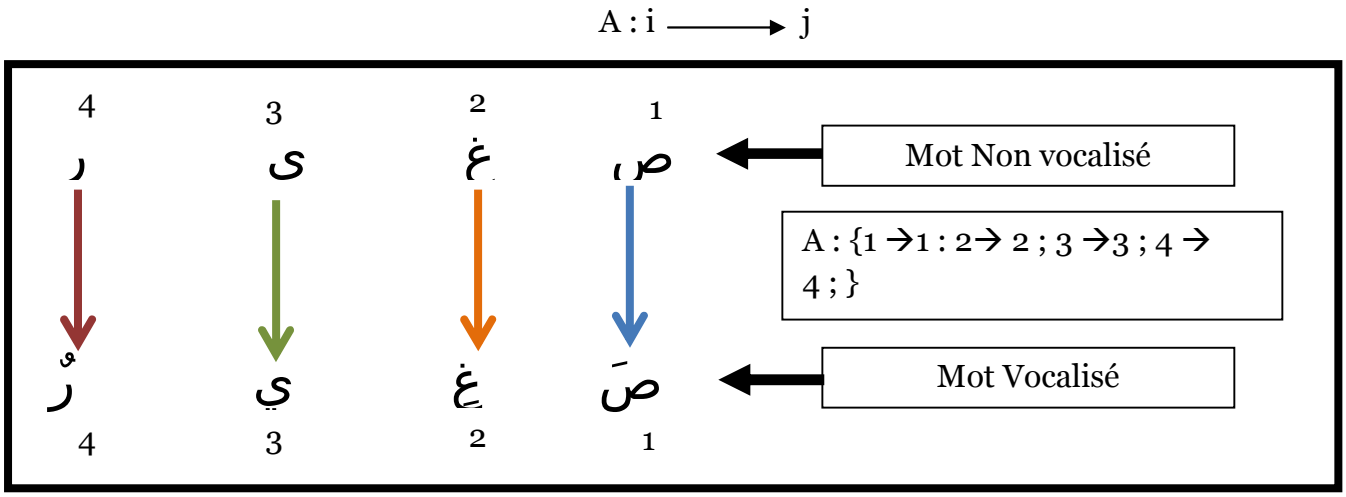

Figure 10 : Exemple d'un processus d'alignement

La formule «1 » concernant le modèle d'alignement sera réécrite en prend cette fois ci le paramètre d' l'alignement.

$$Pr(C|S) = \sum Pr(C, A|S) \qquad (3)$$

Pour effectuer cette alignement nous avons eu recours à l'utilisation Giza + + est une extension d'un programme plus ancien, Giza++. Cet outil effectue l'alignement statistique, il mit en œuvre plusieurs

modèles de Markov cachés et des techniques de pointe qui permettent d'améliorer les résultats d'alignement [33].

4.2.2 Word (Phrase) extraction

- **Cas traduction :** dans étape le but est de définir les frontières des phrases en prenant en considération le résultat de la phase précédente à savoir l'alignement.
- **Cas vocalisation :** l'idée de base est de limité la frontière du mot, pour cela nous avons utilisé la ploteforme MOSES pour accomplir cette tâche.

Moses : est une boîte à outils très performante qui implémente les algorithmes d'apprentissage et de décodage pour les systèmes de traduction automatique statistique. Dans cette boîte à outils, il y a aussi des scripts « prêts à l'emploi » qui réalisent toutes les étapes avec des options par défaut. Cet outil est gratuitement disponible sur le web et est toujours en cours de développement.[32]

Exemple : Extraire un mot d'un alignement de caractère

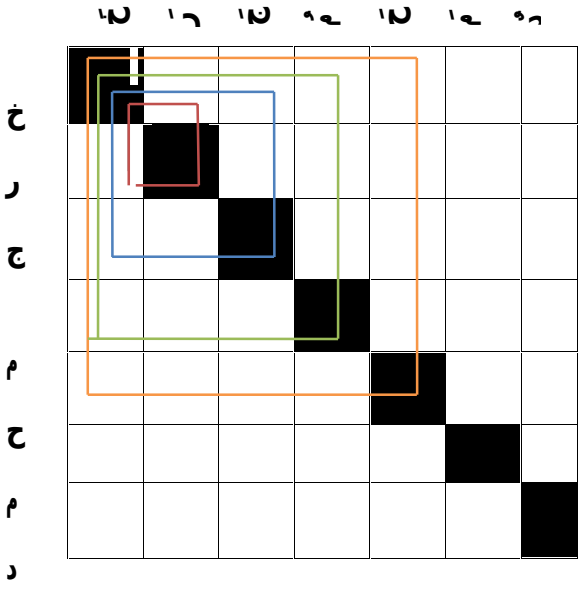

Figure 11 : Extraction un mot d'un alignement de caractère

خ- خَ

ر خ–رَ خَ

ج ر خ–جَ رَ خَ

م ج ر خ - مُ جَ رَ خَ

ح م ج ر خ – حُ مُ جَ رَ خَ

4.2.3 Word (Phrase) table

- **Cas traduction** :le but de cette phase est d'associé à chaque paire (Mot, et sa traduction) une probabilité, cette dernière prend en considération la fréquence de cette paire dans le corpus d'apprentissage.
- **Cas vocalisation :** même principe que la traduction automatique sauf dans notre cas la paire sera : (lettre no vocalisé, lettre vocalisé).

Remarque Importante : Pour implémenté cette phase nous avons utilisé l'outil GIAZ++.

4.3 Le modèle de langue

- **Cas de la traduction :**Le modèle de langue n'est pas une propriété de la traduction automatique statistique, mais on peut la trouver dans divers domaines d'application du traitement automatique du langage naturel (comme : les systèmes de recherche d'information, la reconnaissance de la parole, …etc.). Le but de cette modélisation consiste à trouver le mot le plus probable sachant ceux qui le précèdent. Cette tâche est réalisée lors de la phase d'apprentissage sur le corpus de la langue cible. On va définir la phrase « C » qui est constituée par la séquence de mots « c_1, c_2, ..., c_m ». La probabilité de l'avoir est donnée par l'équation suivante :

$$P(C=c_1,c_2,...,c_n) = P(c_1)\,P(c_2|c_1)\ \ldots\ P(c_n|c_1,c_2,...,c_{n-1}) \quad (8)$$

- **Cas le cas de vocalisation :** Suivant la même démarche en vocalisation on cherchera à trouver une suite de pair (lettre / voyelle) la plus probable possible.

Exemple :

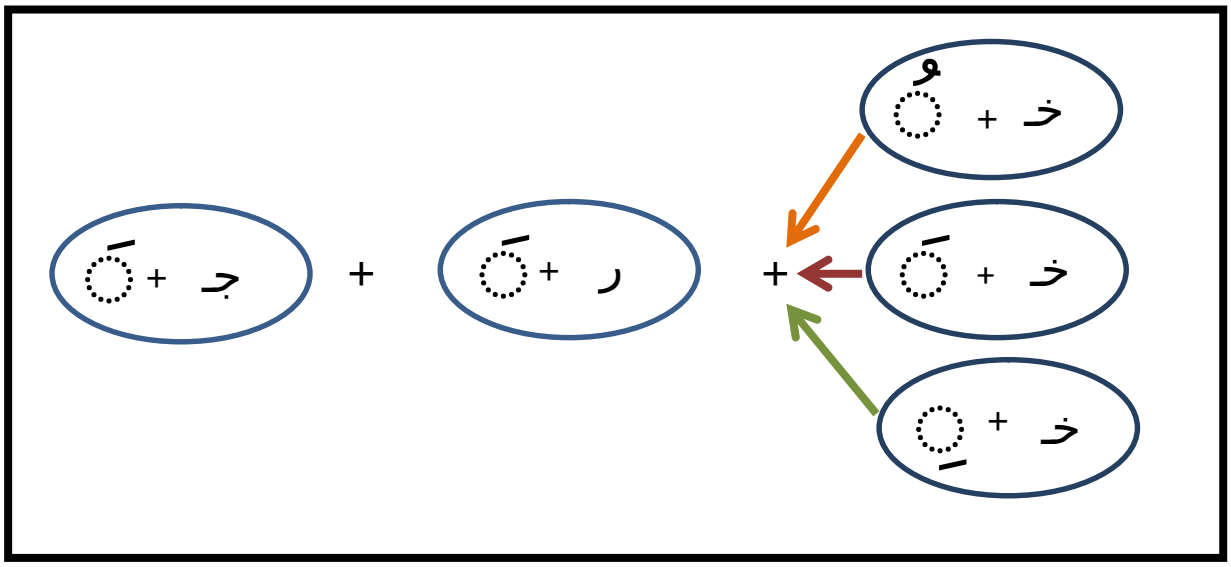

Figure 12 : Exemple du Principe général du Modèle de langue

Pour le calcul des probabilités, il y a différentes méthodes. Ces probabilités sont estimées sur des corpus d'apprentissage de grande taille. Dans notre cas nous avons adopté l'approche baptisée N-Gramme, à savoir une sous-séquence de « n » éléments construite à partir d'une séquence donnée. Dans notre cas, à partir d'une séquence de paire (lettre | voyelle) donnée il est possible d'obtenir la fonction de vraisemblance de l'apparition de la paire suivante. À partir d'un corpus d'apprentissage, il est facile de construire une distribution de probabilité pour le prochaine pair (lettre | voyelle) avec un historique de taille «n ». Cette modélisation correspond en fait à un modèle de Markov d'ordre « n » où seules les « n » dernières observations sont utilisés pour la prédiction de la paire suivante. Ce type de modèle de langage est souvent utilisé sous sa forme 3-gramme qui ne compte que 2 mots comme historique. Le calcul de la probabilité d'apparition d'une paire« m » sachant les « n »paires qui le précèdent (historique « h ») en utilisant une fonction « Fré » qui, pour une séquence de paire (lettre | voyelle), donne la fréquence où cette séquence a pu être observée dans le corpus d'apprentissage, s'exprime sous la forme suivante :

$$Pr(m|h) = \frac{Fré\ (m,h)}{Fré\ (h)} \qquad (9)$$

Dans notre cas et vu que nous travaillons sur les lettres de la langue arabe, nous avons pu constater que la taille moyenne d'un mot arabe est de 5 caractères, de ce fait N=5.

Exemple : Estimons la probabilité de la suite de lettres des mots « الحَيَاةُ»

Uni-gramme (n = 1)

$$P(t) = \prod_{i=1}^{1} p(l_i)$$

P(الحَيَاةُ)=p(ا)*p(ل)*p(ح)*p(ي)*p(ا)*p(ةُ)

Bi-gramme (n= 2)

$$P(t) = \prod_{i=1}^{1} p(l_i \mid l_{i-1})$$

P(ا)*p(ا|ل)*p(ل|ح)*p(ح|ي)*p(ي|ا)*p(ا|ةُ)

Trigramme (n = 3)

$$P(t) = \prod_{i=1}^{1} p(l_i \mid l_{i-2} l_{i-1})$$

P(ا)*p(ا|ل)*p(ال|ح)*p(لح|ي)*p(حي|ا)*p(يا|ةُ)

Pour mettre en place notre modèle de langue dédié à la vocalisation nous avons utilisé **kenAlm** est une boite à outils utilisée pour la construction des modèles de langage statistiques. L'avantage de cette boite à outils est de réduire les besoins de stockage ainsi que la mémoire lors de décodage. Par conséquent, cet outil nous permet de gagner du temps pour le chargement du modèle de langage. [31]

4.4 Décodeur

- **Cas de la traduction :** Le décodage est un procédure dynamique qui cherche à qui maximise la probabilité du procéder du traduction (choix de la meilleure traduction). En combinant les trois scores : modèle de traduction, modèle de langue et le score de distorsion (pénaliser les réordonnances dans la traduction).
- **Cas vocalisation :** pour faire le décodage, et choisir la meilleure paire (lettere / voyelle), nous avons utilisé la plateforme MOSES, cependant cet outil prendra on considération que deux facteur à savoir : le modèle d'alignement (modèle de traduction) et le modèle de langue, vu que l'ordonnancement n'est pas important.

  Il aussi important de préciser que MOSES utilise la formule ci-dessous pour calculer ce score [26].

$$t = \text{argmax } Pr(C/S)\ Pr(C) \qquad (10)$$

- ✓ c est le caractère à vocaliser
- ✓ t est une vocalisation possible du caractère c
- ✓ Le modèle de langage Pr(C)
- ✓ Le Modèle d'alignement Pr(C/S)

- ✓ L’algorithme de recherche (argmaxt).

5. Conclusion

Dans ce chapitre, nous avons présenté la conception notre système de vocalisation automatique de texte arabe non vocalise basé sur la boite à outils Moses, construit à l’aide d’un modèle de langage 5-grammes et en utilisant différents corpus parallèles que nous avons décrits. Nous envisageons d’exploiter notre système pour vocalisation de grands corpus de texte arabe vocalise.

Chapitre 4

Implementation ET Resultants

1. Introduction

Après avoir décortique les phases importante de l'architecture de notre vocaliseur «**MOSHAKIL** », arrive maintenant la partie dédiée à la présentation de notre outil, ainsi dans ce chapitre on parlera de l'environnement de développement, l'interface graphique de notre vocaliseur et la parité la plus importante dans ce chapitre à savoir l'expérimentation et les résultats obtenus accompagné d'une autocritique.

2. Enivrement de développement

2.1. Langage de programmation

Le langage de programmation que nous avons adopté pour implémenter notre application est le python est un langage développé en 1989 par Guido van Rossun [26] très utilisé dans la programmation WEB (accès aux basesde données, programmation objet), comme langage de scripts (manipulation de fichiers, administration de systèmes, configuration de machines), le calcul scientifique (librairies mathématiques).[17]

2.1.1. Pourquoi choisir PYTHON ?

- ✓ Python est entièrement gratuit.
- ✓ C'est un langage complet et puissant dans de nombreux domaines.
- ✓ Raccourcit le cycle de développement par rapport aux langages compilés et permet un prototypage rapide des projets.

2.1.2. Caractéristiques du langage python

Un peu les principales caractéristiques de Python, plusprécisément, du langage et de ses deux implantations actuelles:

⁾**Python est portable**, non seulement sur les différentes variantes d'Unix, mais aussi sur les OSpropriétaires: MacOS, BeOS, NeXTStep, MS-DOS et les différentes variantes de Windows. Un nouveau compilateur, baptisé JPython, est écrit en Java et génère du bytecode Java.

⁾**Python est gratuit,** mais on peut l'utiliser sans restriction dans des projets commerciaux.

⁾ Python convient aussi bien à des **scripts** d'une dizaine de lignes qu'à des **projets complexes** de plusieurs dizaines de milliers de lignes.

- **La syntaxe de Python est très simple** et, combinée à des **types de données évolués** (listes,dictionnaires,...),conduit à des programmes à la fois très compacts et très lisibles.
- Il n'y a **pas de pointeurs** explicites en Python.
- **Python est dynamique**(l'interpréteur peut évaluer des chaînes de caractères représentant des expressions ou des instructions Python),**orthogonal** (un petit nombre de concepts suffit à engendrer des constructions très riches), **réflectif** (il supporte la méta programmation, par exemple la capacité pour un objet de se rajouter ou de s'enlever des attributs ou des méthodes, ou même de changer de classe en cours d'exécution)et i**ntrospectif**(un grand nombre d'outils de développement, comme le debugger ou le profiler, sont implantés en Python lui-même).
- Python est extensible : comme Tcl ou Guile, on peut facilement l'interfacer avec des bibliothèques C **existantes**. On peut aussi s'en servir commed'un langage d'extension pour des systèmes logiciels complexes.
- La bibliothèque standard de Python,et les paquetages contribués, donnent accès à une grande variété de services: chaînes de caractères et expressions régulières, services UNIX standards(fichiers,pipes, signaux, sockets, threads...),protocoles Internet (Web,News, FTP,CGI,HTML...), persistance et bases de données, interfaces graphiques.[18]

3. Caractéristiques technique

Notre vocaliseur est développé dans un environnement LINUX Ubuntu en utilisant un ordinateur dont les caractéristiques techniques, sont les suivantes :

N°	Composant	Description
1	Processeur	Intel «I3»
2	RAM	3.9G
3	HDD	500 Go

Tableau 21: Représentation des Caractéristiques Techniques de L'Ordinateur de Développement.

4. Interface graphique du vocaliseur

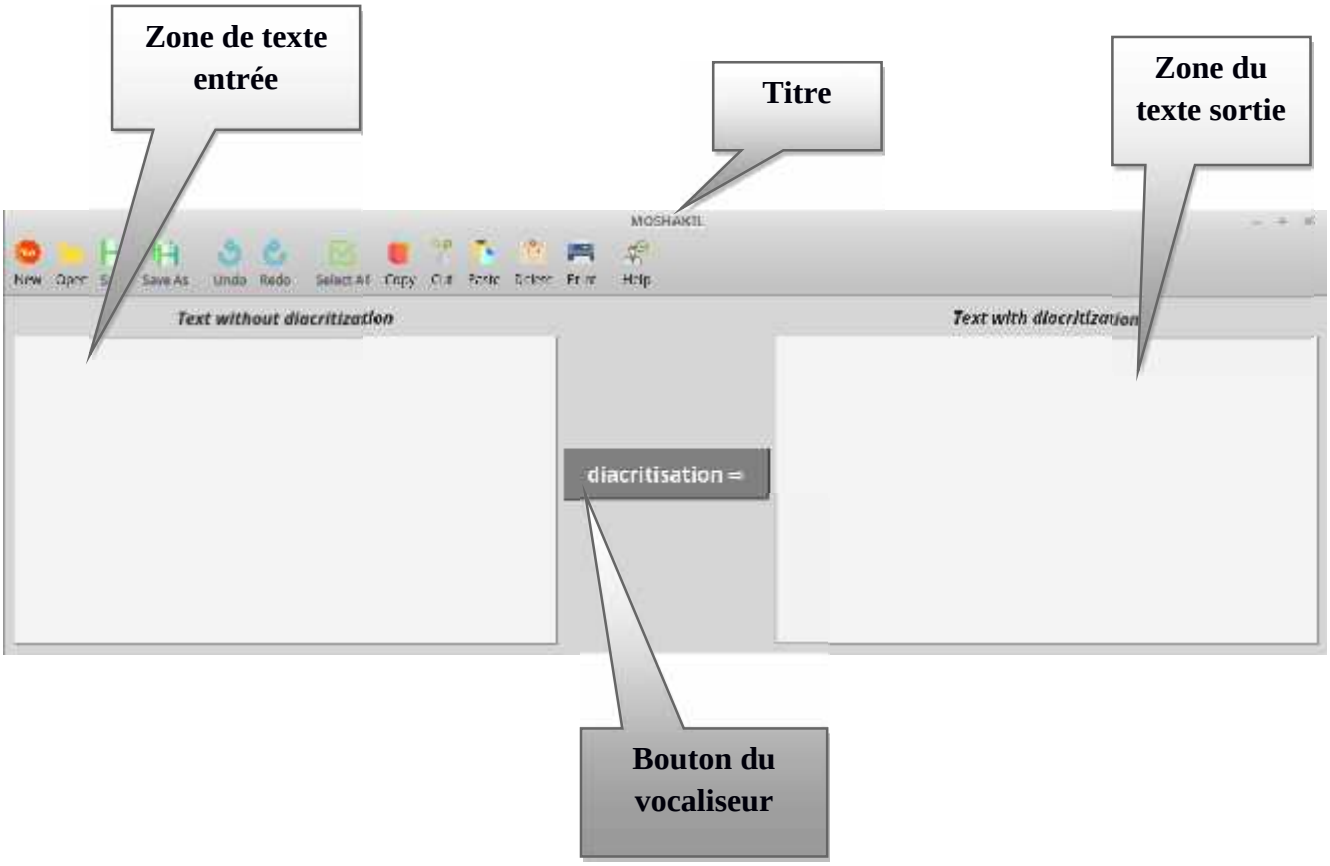

Figure 13: l'interface graphique du vocaliseur**«MOSHAKIL »**

4.1. Barre d'outils

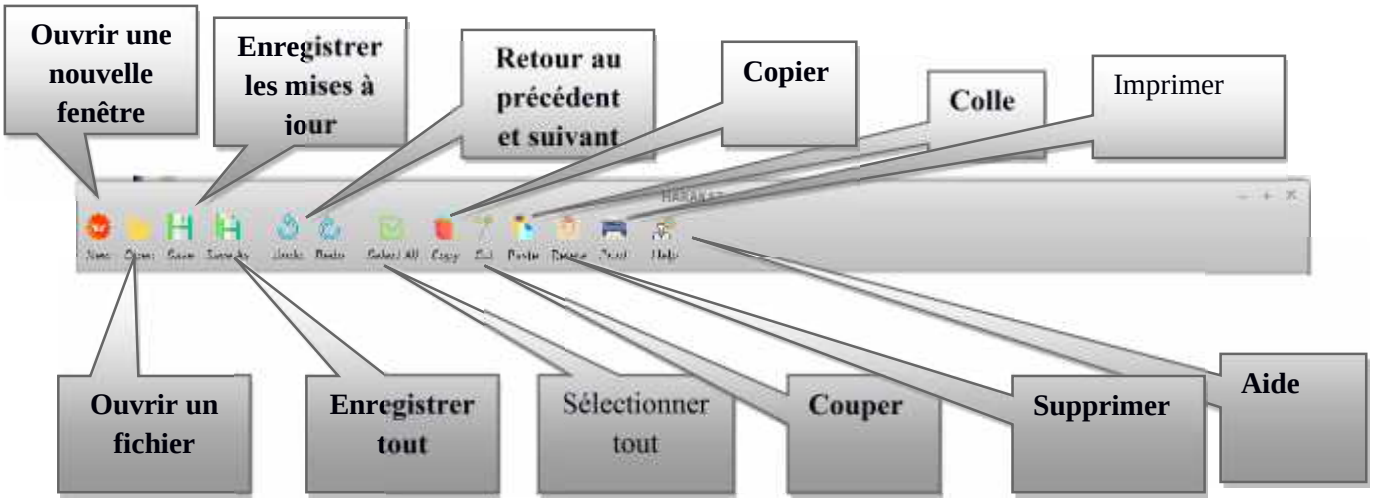

Figure 14: Les boutons de raccourses.

4.2. Représentation des boutons de raccourses

 **New:** pour l'ouverture d'une nouvelle fenêtre

 Open : ouvrir un fichier texte.

 **Save** : sauvegarder les mises à jour.

 **Save as :** sauvegarder les mises à jour avec un nom.

 **Undo**: retour au précédent.

 **Redo:** retour à la suivant**.**

 **Sellect all :** Serve à sélectionner tout**.**

 **Copy** : partager l'écriture.

 **Cut** : abstraire l'écriture.

 **Past** : cimenter l'écriture.

 **Delet**: Sert à la suppression**.**

 **Print :** pour imprimer les fichiers.

4.3. Le bouton du système «MOSHAKIL»

Il fait apparaitre le texte arabe vocalisé

Figure 19: Le bouton de la vocalisation.

4.4. Exemple sur la fonction de bouton «MOSHAKIL »

Figure 15: Exemple de la vocalisation.

5. Expérimentation et Résultats

5 .1 test de corpus «Tashkila»

Nous avons pris 972935 mots de la corpus tashkila et l'avons divisé en 10 parties nous avons effectué le test sue le mot, la lettre et la dernière lettre du mots.

- **Pour les lettres :**

Test	Nombre de Mot	Nombre de lettres	Nombre de lettre Diacritizé Correctement	Nombre de lettre Diacritizé Non Correctement	Accuracy
Test 1	99231	427886	405377	22509	0.932591840585
Test 2	98634	417188	395885	21303	0.934880692552
Test 3	97582	425638	404357	21281	0.936121100058
Test 4	99577	419627	398632	20995	0.936394400162
Test 5	97013	409322	386029	23293	0.927743844748
Test 6	95868	228057	13581	214476	0.924640013317
Test 7	92547	373941	47056	326885	0.843805806165
Test 8	97935	377665	329316	48349	0.842440062439
Test 9	99781	359980	312405	47575	0.836267835879
Test 10	94757	373983	321201	52782	0.824673642252

Tableau 22 : résultat de test corpus tashkila pour lettre

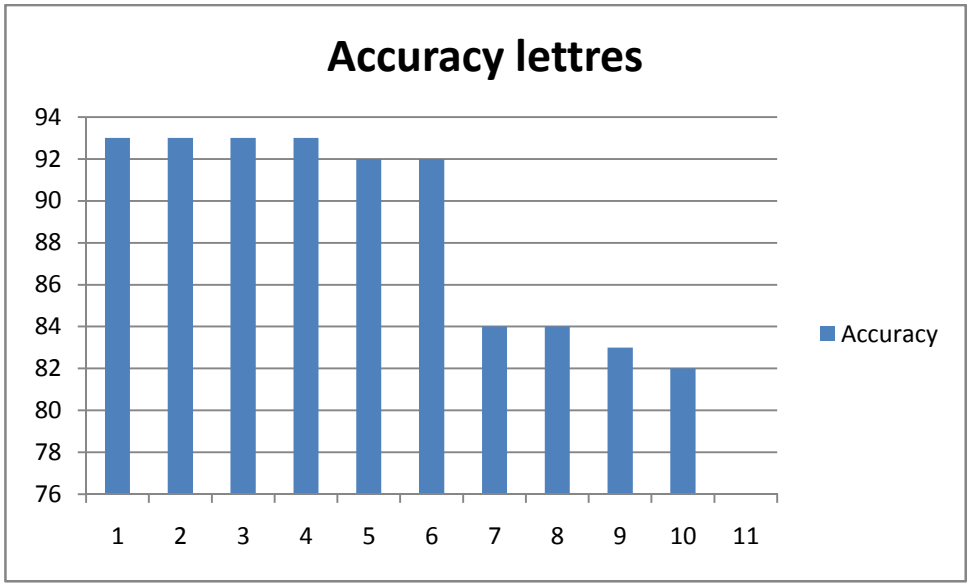

Figure 16 : pourcentage d'accuracy corpus tashkila pour lettre

- **Dernier Caractère (lettre) :**

Test	Nombre de dernière lettre	Nombre de dernière lettre Diacritizé Correctement	Nombre de dernière lettre Diacritizé Non Correctement	Accuracy
Test 1	99231	84097	15110	0.847692199139
Test 2	98634	83463	14119	0.855311430387
Test 3	97582	85320	14257	0.856824367073
Test 4	99577	83135	13873	0.856991175985
Test 5	97013	81315	14552	0.848206369241
Test 6	95868	45425	8631	0.840332248039
Test 7	92547	74600	23314	0.761893089854
Test 8	97935	74213	24254	0.753683975342
Test 9	99781	70377	24365	0.742827890482
Test 10	94757	73121	25514	0.741329142799

Tableau 23 : résultat de test corpus tashkila pour dernier caractère

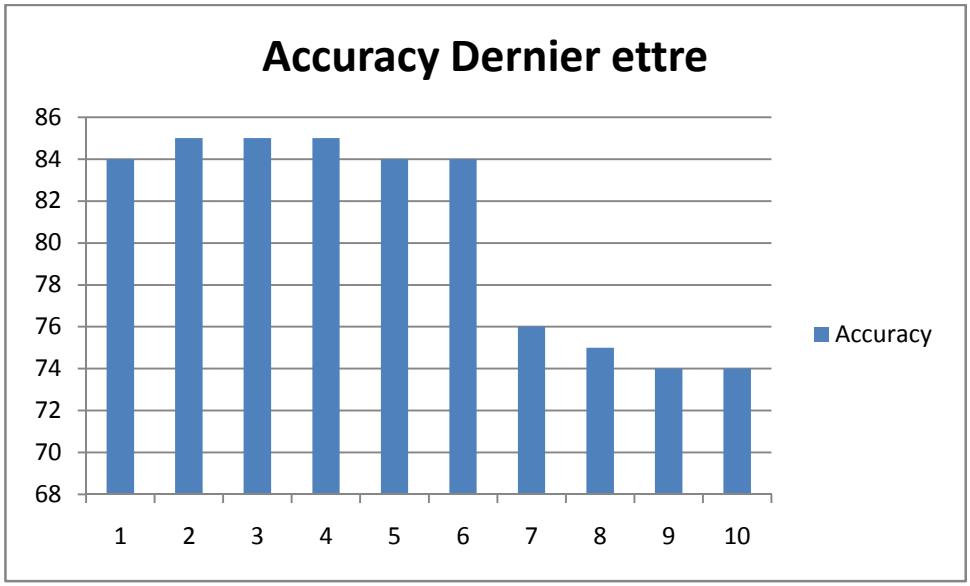

Figure 17 : pourcentage d'accuracy corpus tashkila pour dernier caractère

- **Pour les mots**

Test	Nombre de Mot	Nombre de lettres	Nombre de mots Diacritizé Correctement	Nombre de mots Diacritizé Non Correctement	Accuracy
Test 1	99231	427886	81051	18156	0.816988720554
Test 2	98634	417188	80544	17038	0.825398126704
Test 3	97582	425638	82445	17132	0.827952237967
Test 4	99577	419627	80207	16801	0.826808098301
Test 5	97013	409322	78014	17853	0.813773248354
Test 6	95868	228057	43611	10445	0.80677445612
Test 7	92547	373941	63058	34856	0.644014134853
Test 8	97935	377665	63027	35440	0.63894811656
Test 9	99781	359980	59760	34982	0.630765658314
Test 10	94757	373983	60559	38076	0.613970700056

Tableau 24 : résultat de test corpus tashkila pour mots

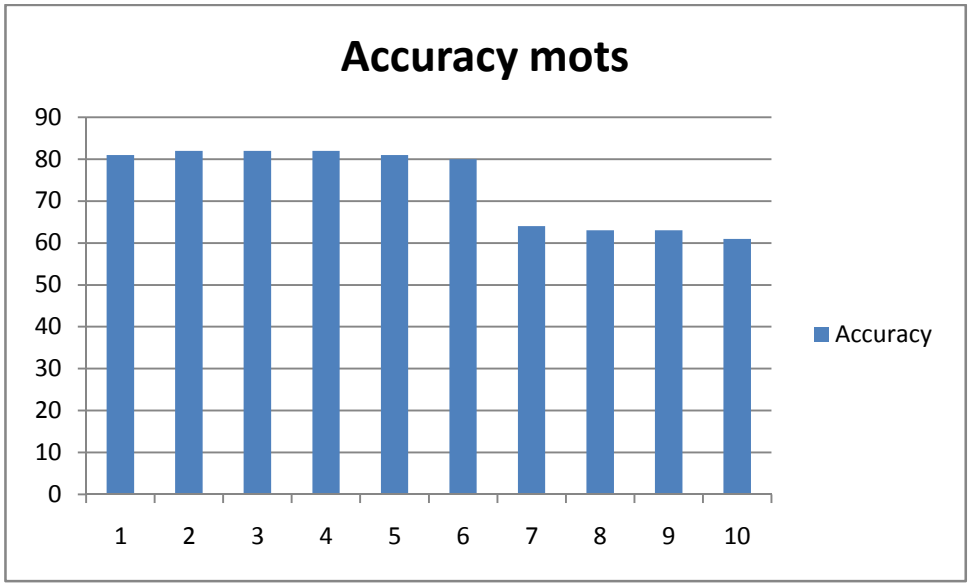

Figure 18 : pourcentage d'accuracy corpus tashkila pour dernier caractère

5.2 Résultats de teste de corpus «WikiNewsEvaluatio»

- **Pour les lettres**

Test	Nombre de Mot	Nombre de lettres	Nombre de lettre Diacritizé Correctement	Nombre de lettre Diacritizé Non Correctement	Accuracy
Test 1	8158	39192	31849	7343	0.76155994285
Test 2	8090	39397	31448	7949	0.741294018095

Tableau 25 : résultat de test corpus WikiNewsEvaluatiopour caractère

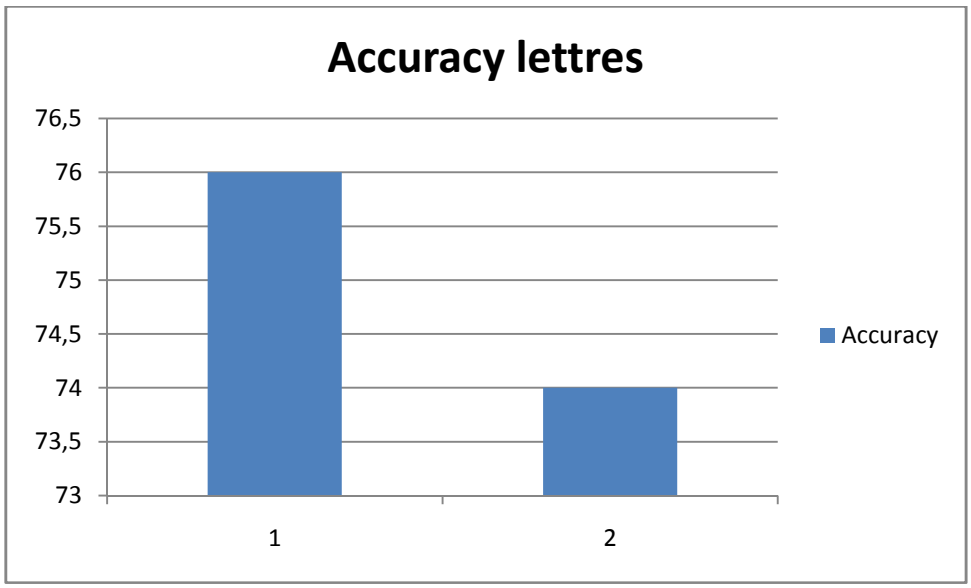

Figure 19 : pourcentage d'accuracy corpus WikiNewsEvaluatio pour caractère

- **Dernier Caractère (lettre) :**

Test	Nombre de dernière lettre	Nombre de dernière lettre Diacritizé Correctement	Nombre de dernière lettre Diacritizé Non Correctement	Accuracy
Test 1	8158	5137	3021	0.629688649179
Test 2	8090	4827	3263	0.596662546354

Tableau 26 : résultat de test corpus WikiNewsEvaluatio pour dernier caractère

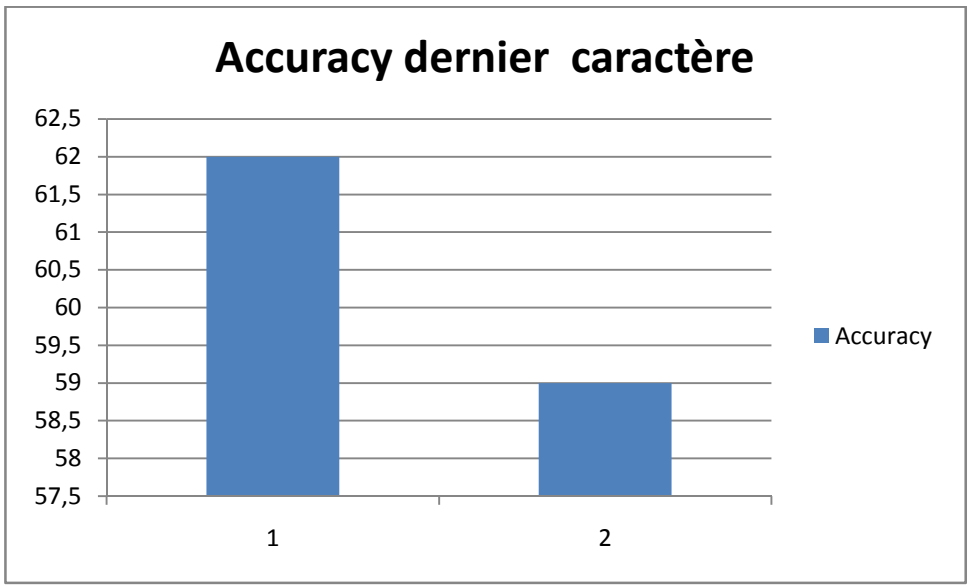

Figure 20 : pourcentage d'accuracy corpus WikiNewsEvaluatio pour dernier caractère

- **Pour les mots :**

Test	Nombre de Mot	Nombre de lettres	Nombre de mots Diacritizé Correctement	Nombre de mots Diacritizé Non Correctement	Accuracy
Test 1	8158	39192	3962	4196	0.485658249571
Test 2	8090	39397	3665	4425	0.453028430161

Tableau 27 : résultat de test corpus WikiNewsEvaluatio pour mots

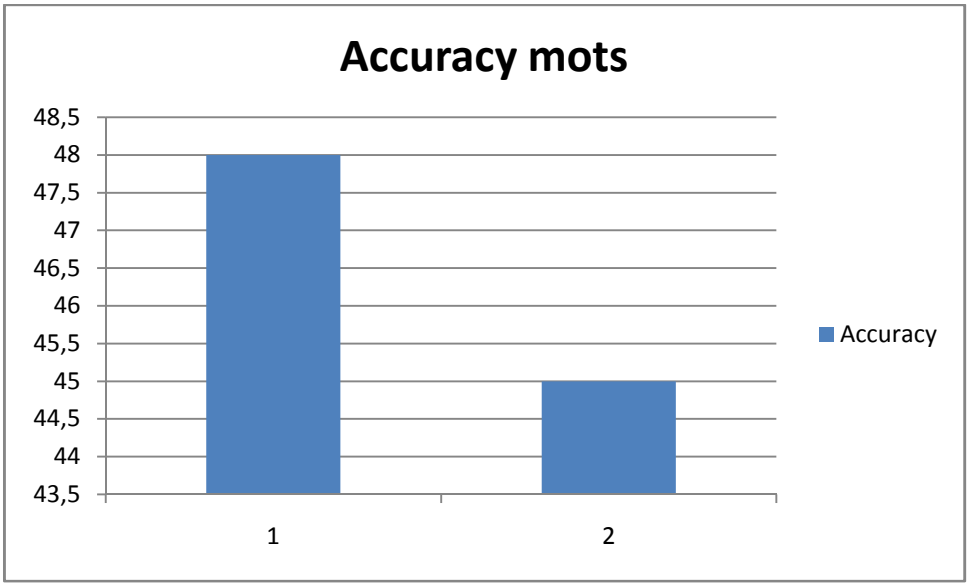

Figure 21 : pourcentage d'accuracy corpus WikiNewsEvaluatio pour mots

5.3 Résultats de teste de corpus «tashkila divisé »

Ici, nous nous assurons qu'il n'y a pas de mots communs entre le corpus de test et corpus d'entrainement.

- **Pour les lettres :**

Test	Nombre de Mot	Nombre de lettres	Nombre de lettre Diacritizé Correctement	Nombre de lettre Diacritizé Non Correctement	Accuracy
Test 1	3958	15336	14520	816	0.934145750948

Tableau 28 : résultat de test corpus tashkila divisé pour lettres

- **Dernier Caractère (lettre) :**

Test	Nombre de dernière lettre	Nombre de dernière lettre Diacritizé Correctement	Nombre de dernière lettre Diacritizé Non Correctement	Accuracy
Test 1	3958	3432	526	0.867104598282

Tableau 29 : résultat de test corpus tashkila divisépour dernier lettre

- Pour les mots

Test	Nombre de Mot	Nombre de lettres	Nombre de mots Diacritizé Correctement	Nombre de mots Diacritizé Non Correctement	Accuracy
Test 1	3958	15336	3323	635	0.839565437089

Tableau 30 : résultat de test corpus tashkila divisépour mots

6. Analyse Critique

Tester un système est un point très important où ses limites sont révélées afin d'ouvrir le champ pour l'améliorer.

Notre corpus de travail global est divisé en deux sous-corpus: le corpus système d'entraînement et le corpus de test pour évaluer notre application. Nous avons effectué le test sur 10 parties de corpus tashkila ,et test sure corpus tashkila divisé .cette expérience à aboutir à un résultat satisfaisant dont le taux est de 93% au niveau lettres, 83% au niveau mots et dernier caractère 62% , Nous avons également effectué un test sur notre corpus WikiNewsEvaluatio , Les résultats étaient moyen dont le taux est de 75% au niveau lettres, 46% au niveau mots et dernier caractère 62% .

La différence qui est de pour assurer une vocalisation plus correcte est due en grande partie à l'absence signe de vocalisation des mots arabesques et des noms propres du corpus système d'entraînement.

7. Conclusion

Dans ce chapitre, nous avons présenté l'implémentation de notre application de vocalisation notamment le langage de programmation choisi ainsi que les caractéristiques techniques de l'ordinateur qui a servi pour le développement de cet outil. Parmi les points de ce chapitre figure la description de l'interface graphique.

Conclusion (Bilan et Perspectives)

1. Bilan

Les outils de traitement automatique de la langue arabe sont l'ensemble des recherches et développements visant à modéliser et reproduire, à l'aide de machines, la capacité humaine à produire et à comprendre des énoncés linguistiques dans des buts de communication.

Plusieurs travaux, en ce qui concerne l'automatisation de la langue arabe ont démontrés que l'absence de signes de vocalisation, est l'un des problèmes major qui laisse le procéder d'automatisation en retard. Par conséquent, une forme de mot en arabe peut être voyellée de multiples façons, avec des significations différentes en fonction du contexte où elle apparaît. Ce problème du non vocalisation des textes peut engendrer un haut degré d'ambiguïté affectant les systèmes de reconnaissance des entités nommées. En effet, les vocalisations acceptées pour une forme d'un texte peuvent désigner des mots (déclencheurs) introduisant différents types d'entités nommées.

L'objectif de notre mémoire été de proposer une démarche qui contribue de près à la compréhension d'un énoncé exprimé en langue arabe, par le fait de mettre en place un procédé orignal qui s'inspire étroitement de celui de la traduction automatique statistique.

MOSHAKIL, c'est le nom que nous avons donné, à notre outil de vocalisation automatique dédié à la langue arabe, cet outil qui fait collaborer apprentissage et calcul probabiliste permettra d'affecter à chaque lettre une voyelle qui va avec.

D'après les premiers tests expérimentaux, effectuer sur plusieurs fragments de corpus arabe, les résultats obtenus son assent encourageante, laissant pensé que le choix en ce qui concerne la démarche été correcte.

2. Perspectives

Certaines améliorations peuvent être apportées à l'architecture de notre vocaliseur de texte arabe, parmi ces modifications, on peut évoquer :

- L'utilisation d'un étiqueteur morphosyntaxique, ce qui va permettre d'augmenter le pourcentage de la vocalisation de la dernière lettre.
- Augmenté sensiblement la taille du corpus d'apprentissage.
- Augment le nombre de fragment lors du calcul des N-gramme (de 5 jusque 7) afin de toucher la totalité des mots arabe (en terme de morphologie).
- Utiliser d'autres outils comme : SRILM (Modèle de langue).

Références

[1] Zied Arbi,« Traitement automatique des langages naturels», École Supérieure des Communications de Tunis , MQTT protocol for iot applications Viewproject March18 April 2015.

[2] Mohammed El Amine ABDERRAHIM, «Reconnaissance des unités linguistiques signifiantes», thèse de doctorat,Université Abou Bekr BELKAID TLEMCEN ,Spécialité : Informatique ,08 Juillet 2008.

[3] Fouad SoufianeDouzidia, « Résumé automatique de texte arabe» , Mémoire Master en informatique , Département d'informatique et de recherche opérationnelle Faculté des arts et des sciences , Université de Montréal , Septembre, 2004.

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

[5] BOUGHERARA TABOURI Yasmina.,«L'utilisation du traitement automatique des langues (T.A.L.) pour l'étude des adverbes français dans les textes journalistiques: modalisation et subjectivité», Mémoire Magister,,Université d'Oran 2, 2016.

[6] F. YVON, «Une petite introduction au Traitement Automatique des Langues Naturelles», support de cours, Ecole Nationale Supérieur des télécommunications, 26Avril 2007.

[7] François Yvo«introduction au traitement Automatique du langage naturel », support de cours, Ecole Nationale Supérieur des télécommunications, 19 février 2006.

[8] CHOUCHAOUI Maissa , BRAHIMIA YamnaAffaf« Détection Automatique De La Cohésion Lexicale Entre Phrases Dans Les Textes Arabes», Mémoire Présenté Pour l'obtention de diplôme Master en Informatique , Option : Ingénierie du logiciel ,

Faculté des Sciences et de la Technologie Département des Mathématique et Informatique, Université de Djilali BOUNAAMA KHEMIS MILIANA, 01/06/2016.

[9] HoudaSaadane, Mathieu GUIDERE, Christian Fluhr,«La reconnaissance automatique des dialectes arabes à l'écrit», Article LIDILEM, Université de Grenoble, 5 /2/2014.

[10] SouhirGahbiche-Braham, «Amélioration des systèmes de traduction par analyse linguistique et thématique Application à la traduction depuis l'arabe», THÈSE de Doctorat Spécialité : Informatique, UNIVERSITÉ PARIS SUD, Septembre 2013.

[11] GASMI Mounira «Utilisation des ontologies pourl'indexation automatique des sites Web en Arabe» ,MémoiredeMAGISTERSpécialité : Informatique ,UNIVERSITE KASDI MERBAH OUARGLA,mai 2009 .

[12] Mohamed OuldAbdallahiOuldBebah ,«Contribution à l'analyse morpho-syntaxique de la langue Arabe et application à la voyellationautomatique» ,thèse de Doctorat en informatique, Faculté des SciencesOujda, Université Mohamed Premier , Octobre 2013.

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

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

[15] Amine Chennoufi ,AzzeddineMazroui «Méthodes de lissage d'une approche morpho-statistique pour la voyellationautomatique des textes arabes» conférences du TALN,université Mohamed , Faculté des Sciences , Département de Mathématiques et Informatiques Oujda, Maroc,2014.

[16] ندغنيم .د,غيداءربداوي .د
« منهجيات التشكيل الالي لنصوص العربية بهدف وضع خطة عمل لبناء مشكل الي مفتوح المصدر **دراسة**»
مجلة جامعة دمشق للعلوم الهندسية المجلد التاسع والعشرون –**العدد الثاني** - 2013

[17] Gal Y. (2002). An HMM Approach to Vowel Restoration in Arabic and Hebrew. In ACL-02 Workshop onComputationalApproaches to SemiticLanguages.

[18] Debili F., Achour H. (1998). Voyellation automatique de l'arabe. In Proceedings of the workshop on Computation approaches to Semitic languages, COLING-ACL '98. Montréal.

[19] Schlippe T., Nguyen T. , Vogel S. (2008). Diacritization as a Machine Translation Problem and as a Sequence Labeling Problem. In 8th AMTA conference, Hawaii. Pages 21-25.

[20] Alghamdi M., Muzaffar Z., Alhakami H. (2010), "Automatic Restoration of Arabic Diacritics: A Simple, Purely Statistical Approach," The Arabian Journal for Science and Engineering, vol. 35, 2010.

[21] Hifny Y.(2013). Restoration of arabic diacritics using dynamic programming. In The 8th International Conference on Computer Engineering & Systems (ICCES'2013), 26-28 Nov. 2013, Cairo, Egypt, pages 3-8 ISBN:978-1-4799-0078-7.

[22] El-Sadany T., Hashish M. (1988). Semi-automatic vowelization of arabic verbs. In 10th NC Conference, SaudiArabia.

[23] Rashwan M., Al-Badrashiny M. , Attia M., Abdou S.M., Rafea A. (2011), "A Stochastic Arabic Diacritizer Based on aHybrid of Factorized and Unfactorized Textual Features," Audio, Speech, and Language Processing, pages. 166-175 .

[24] Zitouni I., Sarikaya R. (2009), "Arabic Diacritic Restoration Approach Based on Maximum Entropy Models," Computer Speech & Language, vol. 23, no. 3, pp. 257-276, 2009.

[25] F. J. Och and H. Ney, «A systematic comparison of various statistical alignment models,» Computational Linguistics, vol. 29, pp. 19-51, 2003.

[27] P. Koehn, et al., «Moses: Open source toolkit for statistical machine nslation,»2007, pp. 177-180.

[26] Loic Gouarin «les base du langage python »laboratoire de mathématiques d'Orsay , 6 décembre 2010

[27] Thi-Ngoc-DiepDO«Extraction de corpus parallèlepour la traductionautomatique depuis et versune langue peudotée»thèse Pour obtenir le grade de docteur en sciences délivréparl'université de grenoble, Spécialité : MSTII / INFORMATIQUE,7 août 2006

[28] TahaZerrouki n, AmarBalla«Tashkeela: Novel corpus of Arabic vocalized texts, data for auto-diacritization systems»journal homepage: www.elsevier.com/locate/dib, The National Computer Science Engineering School (ESI), Algiers, Algeria,2017

[29] Frédéric Blain,« Modèles de traduction évolutifs» thèse Doctorat de l'Université´e du Maine, spécialité informatique, 23 septembre 2013

[30] Dr. Philipp Koehn,« Statistical Machine Translation» Textbook Graduate Level 400 Pages, University of Edinburgh, November 20, 2007

[31] Marwa Hadj Salah, Didier Schwab, Hervé Blanchon, Mounir Zrigui«Système de traduction automatique statistique Anglais-Arabe» Univ. Grenoble Alpes, France,6 Feb 2018 .

[32] Haïtem Afli « Approche mixte pour la traduction automatique statistique» Mémoire de master 2 recherche - 30 crédits- Mention sciences du langage , Spécialité: Modélisation et traitement automatique en industrie de la langue : Parole, Ecrit, Apprentissage (TALEP) Université d'Avignon, 2011. Français , Année universitaire 2009-2010 .

[33] Och, F. and Ney, H. 2000.Improved statistical alignment models.Proceeding ACL 440–447.

ANNEXE

Lettre arabe	Correspondant Français	Prononciation
ا	A	Alef
ب	B	Ba’
ت	T	Ta’
ث	Th	Tha’
ج	J	Jim
ح	H	Hha’
خ	Kh	Kha’
د	D	Dal
ذ	D	Thal
ر	R	Ra
ز	Z	Zayn
س	S	Sin
ش	Sh	Shin
ص	S	Sad

Lettre arabe	Correspondant français	Prononciation
ض	d	dat
ط	t	tah
ظ	z	Zad
ع	‘’	Ayn
غ	gh	Ghayn
ف	F	Fa
ق	Q	Qaf
ك	K	Kaf
ل	L	Lam
م	M	Mim
ن	N	Nun
ه	H	Ha
و	W	Waw
ى	Y	Ya

Tableau 1: les 28 lettres arabes.

	Les schèmes verbaux	***Exemple***
Trilitère augmenté (الثلاثي المزيد)	فعّل	قَدَّم Présenter
	فاعل	شارَك Participer
	أفعل	أعطى Donner
	انْفَعَل	انْطَلَق se lancer
	افْتَعَل	احْتَرَم Respecter
	افْعَلّ	احْمَرّ Rougir
	تَفَاعَل	تَلاءَم Convenir
	تَفَعَّل	تَقَدَّمَ se présenter
	اسْتَفْعَل	اسْتَخْرَج Extraire
	افْعَوْعَل	اخْضَوْضَر
	افْعَوَّل	اعْلَوَّط *monter sur*
	افْعَالَّ	اصْفَارَّ *Jaunir*
Quadrilitères augmenté (الرباعي المزيد)	تَفَعْلَلَ	تَدَحْرَج *Rouler*
	افْعَلَلَّ	اطْمَأَنّ *se rassurer*
	افْعَنْلَل	احْرَنْجَم *s'entasser*

Tableau 5: Schèmes verbaux augmentés

Schèmes verbaux	Schèmes du participe actif	Exemples
فعل	فاعِل	قَارئ , *(lecteur)*
فعّل	مُفَعِّل	مُدَرِّس , *(enseignant)*
فَاعَل	مُفاعِل	مُشارك ,*(participant)*
انْفَعَل	مُنْفَعِل	مُنْسَجِم‘ ,*(cohérent)*
افْتَعَل	مُفْتعِل	مُنْتَسِب,*(adhérent)*
افْعَلّ	مُفْعَلّ	مُصْفَرّ ,*(jaunâtre)*
أفْعَل	مُفْعِل	مُصلِح, *(réformateur)*
تَفَاعَل	مُتفاعِل	مُتلازم, *(corrélatif)*
اسْتَفْعَل	مُسْتَفْعِل	مُسْتَثْمِر , *(investisseur)*
افْعَالّ	مُفْعَالّ	مُحْمَارّ , *(rougeâtre)*
افْعَوْعَل	مُفْعَوعِل	مُخْضَوَضِر , *(verdoyant)*
افْعَوّل	مُفْعَوّل	مُعْلَوّط , *(montant sur)*
فَعْلَل	مُفَعْلِل	مُزخرف ,*(décorateur)*
تَفَعْلَل	مُتَفَعْلِل	متَعَجْرف , *(arrogant)*
افْعَلّ	مُفْعَلِلّ	مُطْمَئِنّ , *(calme)*
افْعَنْلَل	مفْعَنلِل	مُحرَنْجم , (congestionné)

Tableau 6: Les schèmes du participe actif

Schèmes verbaux	*Schèmes du participe passif*	*Exemples*
فعل	مَفْعُول	مَقْروء, *(lu)*
فعّل	مُفَعَّل	مُدَرَّس , *(enseigné)*
فَاعَل	مُفاعَل	مُطالَب , *(revendiqué)*
انْفَعَل	مُنْفَعَل	مُنْتَظَر , *(attendu)*
افْتَعَل	مُفْتَعَل	حْتَرَم, *(respecté)*
افْعَلّ	مُفْعَلّ	مُصْفَرّ , *(jaunis)*
أفْعَل	مُفْعَل	مُدْمَج , *(intégré)*
تَفَاعَل	مُتفاعَل	مُتَعارَف , *(reconnu)*
اسْتَفْعَل	مُسْتَفْعَل	مُسْتَخْدَم , *(utilisé)*
افْعَالّ	مُفْعَالّ	مُحْمَارّ, *(rougis)*
افْعَوْعَل	مُفْعَوعَل	مُعْرَوْرى , *(monté -sans selle-)*
افْعَوّل	مُفْعَوَّل	مُعْلوط , *(monté)*
فَعْلَل	مُفَعْلَل	مُزَخْرَف , *(décoré)*
تَفَعْلَل	مُتَفَعْلَل	مُتَسَربَل , *(porté)*
افْعَلّ	مُفْعَلّ	مُطْمْأن , *(rassuré)*
افْعَنْلَل	مفْعَنلَل	مُحرَنْجَم , *(bondé)*

Tableau 9: Les schèmes du participe passif

Résumé

Le traitement automatique de la langue arabe est une discipline en pleine expansion, dans laquelle on voit de plus en plus de recherches et de technologies se soucier des spécificités de cette langue et proposer des outils nécessaires au développement de son traitement automatique comme : la traduction automatique, correction automatique d'orthographie, le résumé automatique, l'interrogation de base de données en langage naturel, …etc.

Dans cette même optique, le but de ce mémoire est de contribué au développement de l'automatisation de la langue arabe, par la mise en place d'un outil d'une grande important dans la compréhension de l'énoncé arabe, à savoir la vocalisation automatique, ou l'originalité de ce travail réside dans le fait de s'inspirer du procéder de traduction automatique statistique pour mettre en place une démarche de vocalisation qui associé à la fois apprentissage et calcul probabiliste.

Mots clé : Vocalisation, langue arabe, Traduction automatique statistique, Moses, GIZA++.

ملخص

ان المعالجة الألية للغة العربية هي ميدان في تطور متزايد، حيث نلاحظ **ان** المزيد والمزيد من الأبحاث والتقنيات التي تختص وتهتم باللغة العربية و هذه باقتراح وتطوير العديد من البرامج مثل: الترجمة الآلية ، التصحيح الإملائي الآلي ، التلخيص الآلي ، استعلام قاعدة بيانات باستعمال **اللغة** الحية ، ... **إلخ**.

وفي نفس السياق ، فإن الهدف من هذ البحث هو المساهمة في تطوير المعالجة الألية للغة العربية ، من خلال وضع أداة ذات أهمية كبيرة في فهم السياق **العربي** ، وهو التشكيل الآلي.أين التميز في هذا العمل يكمن في كونه مستلهمًا من عملية الترجمة الآلية الإحصائية من أجل تنفيذ عملية التشكيل **التي** تجمع بين كل **من** التعليم الألي **والاحتمالات**.

الكلمات المفتاحية : التشكيل الآلي ,اللغة العربية ,الترجمة الآلية