Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université d'Adrar

Faculté des Sciences et de la Technologie

Département des Mathématiques et de l'Informatique

Mémoire de fin d'étude

Présenté pour l'obtention du diplôme de Master en informatique

Option : Réseaux et Systèmes Intelligents

Thème

Réalisation d'un analyseur morphologique du texte arabe « AMTAR »

Réalisé par :
M[r] : Soulimani Mohammed

Encadré par :
M[r] : Cheragui Mohamed Amine

Année Universitaire 2013 / 2014

Remerciements

Je tiens tout d'abord à remercier Dieu qui m'a donné la force et la patience d'accomplir ce travail.

En second lieu je remercie mon encadrant Mr. Cheragui Mohamed Amine pour l'orientation, la confiance, la patience qui ont constitué un apport considérable sans lequel ce travail n'aurait pas pu être mené au bon port.

Mes vifs remerciements vont également aux membres du jury pour l'intérêt qu'ils ont porté à notre recherche en acceptant d'examiner notre travail et de l'enrichir par leurs propositions.

je tiens également à remercier toutes les personnes qui ont participé de près ou de loin à la réalisation de ce travail.

Enfin, un remerciement spécial à mon père, ma femme, mes frères mes sœurs, et toute ma grande famille.

ملخص

المعالجة الآلية للغات الطبيعية هي مجال من مجالات المعلوماتية يهتم بفهم و إنتاج الجمل اللغوية بواسطة الآلة. في هذه المذكرة نهتم بالمعالجة الآلية للغة العربية وخاصة على المستوى المورفولوجي حيث أن هدفنا هو إنشاء محلل صرفي للنصوص العربية غير المشكولة.

أمطار هو اسم المحلل الصرفي الذي قمنا بإنشائه حي ث يعتمد أساسا على قاعدة معطيات لغوية ونصوص برمجية تسمح بالقيام بمختلف خطوات المعالجة ابتداء من تقطيع النصوص إلى غاية تعيين الخصائص المورفولوجية للكلمات و إظهار النتائج.

قمنا باختبار برنامجنا باستعمال مدونة تسمى خليج حيث أعطىالاختبار نتائج جيدة وصلت إلى معدل92 %.

كلمات مفاتيح : المعالجة الآلية للغات الطبيعية، اللغة العربية، تحليل صرفي.

Résumé

Le traitement automatique des langues naturelles est un domaine de l'informatique qui s'occupe de la compréhension des énoncés linguistiques par la machine et la production de celle-ci par elle-même. Dans ce mémoire nous intéressons au traitement automatique de la langue arabe et surtout au niveau morphologique, dont l'objectif est de réaliser un analyseur morphologique de texte arabe non voyellé.

AMTAR est un analyseur morphologique du texte arabe que nous avons développé sur une base de données linguistiques arabe, et des modules de programmation permettant de réaliser les différentes étapes de traitement, commençant par la segmentation du texte, et finissant par l'affectation des traits morphologiques aux unités lexicales (mots) et l'affichage des résultats.

Nous avons évalué la qualité des résultats au moyen des testes de note système sur un corpus nommé « khaleej », et nous avons obtenu des bons résultats qui arrivent à un pourcentage moyen égale à 92%.

Mots clés : T.A.L.N, La langue Arabe, Analyse Morphologique.

Table des matières

Liste des tableaux

Liste des figures

Introduction Générale

1-Introduction

Le traitement automatique des langues naturelles (T.A.L.N.) est un domaine de recherches, qui fait collaborer les linguistes, les informaticiens, et qui appartient au domaine de l'Intelligence Artificielle (I.A.).

Le terme "langue naturelle" désigne les langues humaines. Pour l'être humain, ces langues sont en quelque sorte des créations et spontanées, mais pour la machine il encoure difficile de comprendre une énoncé langagière, ce qui fait le sujet de domaine de traitement automatique de langages naturels.

Le traitement automatique de la langue arabe (T.A.L.A.) connait des activités intensives au cours des dernières années ayants conduit à l'émergence d'un grand nombre de logiciels qui s'intéresse à trouver des solutions techniques qui conviennent avec les caractéristiques dérivationnelles et flexionnelles de l'arabe.

Le traitement en T.A.L.N. se fait en quatre (04) niveaux qui sont: l'analyse morphologique, l'analyse syntaxique, l'analyse sémantique, et l'analyse pragmatique.

Notre travail se déroule dans le premier niveau dont le titre c'est « Réalisation d'un analyseur morphologique du texte arabe »

2-Le but de travail

Le but principal de ce travail est de concevoir et de réaliser un analyseur morphologique pour l'Arabe non voyellé. Qui peut être utilisé ultérieurement par d'autres applications, tel que analyseur syntaxique, recherche de l'information, filtrage de documents, …etc. L'analyseur doit déterminer pour chaque mot ses différentes caractéristiques morphosyntaxiques.

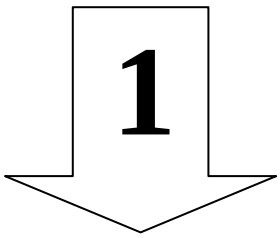
3-Le plan de travail

Notre travail s'articule au tour de quatre chapitres, à savoir :

- Dans le premier chapitre nous avons essayé de définir le domaine du traitement automatique de langues naturelles, ses différents niveaux d'analyse, domaines d'applications, et ses approches.
- Le second chapitre est dédié à la langue arabe et ses caractéristiques morphologiques.
- Le troisième représente les étapes de conception de note analyseur morphologique de texte arabe non voyellé, et ses modules.

- Le quatrième chapitre nous avons représenté l'environnement de développement du notre système, la description de l'interface graphique de notre analyseur morphologique, et les résultats des tests.

Chapitre



Le Traitement Automatique des Langues Naturelles (TALN)

1- Introduction

Le traitement automatique des langues naturelles (TALN) est un domaine qui regroupe tous les recherches en informatique qui se préoccupent de la compréhension des énoncés linguistiques par la machine, ainsi que leurs productions par elle même.

Les langues naturelles sont celle écrites ou parlées par des êtres humains. C'est encours une faculté innée pour toute l'humanité ; c'est elles qui les permettent de communiquer par un système de signes doublement articules.

Le TALN a pour objectif l'élaboration des systèmes (logiciels) capable de traiter automatiquement des langues naturelles. Ce traitement nécessite l'intervention des recherches dans différents domaines, tels que :

- **La linguistique théorique** : qui donne des descriptions et des théories cohérentes des connaissances linguistiques.
- **Informatique** : qui permet d'élaborer des algorithmes et des programmes adéquats pour le traitement.
- **L'intelligence artificielle** : qui donne des formalismes de représentation des connaissances.

- **Les mathématiques** : qui donnent des propriétés formelles pour les outils de traitements.

Dans ce chapitre nous allons définir le domaine du traitement automatique des langues naturelles (TALN) et ses domaines d'applications, ainsi, nous expliquerons les niveaux d'analyse et les différentes approches du traitement.

2- Bref historique

Les premiers travaux dans ce domaine étaient sur la traduction automatique (TA) par la création du premier traducteur qui fait la traduction mot à mot de quelques phrases russe vers l'anglais. C'était au début des années cinquante. C'est l'époque de la guerre froide, la compétition russes/américains bat son plein.

Les recherches sur le TALN ont commencé sérieusement dans les années 1950[Bouillon ,1998].

En 1952 : l'organisation de la première conférence sur la traduction automatique, organisée au MIT (Massachussets Institute of Technology) par Yehoshua Bar-Hillel.

C'est en 1953 que Bar-Hillel fait l'hypothèse d'on peut construire une machine capable de déterminer les structure de toute les phrase d'une langue, pourvu que la syntaxe de cette langue soit présentée sous forme opérationnelle à la machine.

En 1955 V.Yngve élabore une procédure basée sur les automates à états finis pour faire le traitement des phrases réduites en classes de mots. Le même chercheur, en 1960, élabore un modèle prédictif d'analyse syntaxique fondé sur une grammaire syntagmatique, et sur des hypothèses psycholinguistiques sur la mémoire à court terme pour fixer la profondeur de l'arbre de représentation des phrases, il se réfère au modèle de Chomsky [Cori et al, 2002]

En 1957 N. Chomsky a publié ses premiers travaux important la syntaxe des langues naturelles ; et sur les rapports entre grammaire formelle et grammaire naturelles. Chomsky a dit que l'analyse syntaxique d'une phrase peut se faire indépendamment de son sens. [Bouillon ,1998]

Cet époque a connu la création de plusieurs logiciel de traitement automatique de langage naturel, certain parmi eux n'est pas basé sur une grammaire générative tel que le système ELIZA qui simule un dialogue entre un psychiatre et son patient. Ce système n'utilise pas des connaissances linguistiques, mais une banque des données qui contient un stock des phrases indexés par des mots clés. Et d'autre système tel que LUNAR qui utilise une grammaire générative.

En 1961 et grâce à la conférence hebdomadaire organiser par David Ghats, le terme 'computationnel linguistique' est né pour designer les logiciels de traitement des langues.

Pendant les années 1970 la plupart des recherches dans le domaine du TALN sont dirigées vers la sémantique, alors que les recherches syntaxiques peuvent être considérées comme secondaires, surtout dans le cadre de la représentation des connaissances.

En 1972 T.Winograd a réalisé SHRDLU, c'est un logiciel qui permet de dialoguer, en anglais, avec un robot dans un monde réduit.

L'an 1976 est la date d'installation d'un système commercial pour la traduction automatique, nommé Systran. Le Succès de ce programme et sa grande diffusion fait le point de départ pour plusieurs systèmes commerciaux.

A partir des années 1980, le TALN commence à s'influencer par l'intelligence artificielle (la représentation des connaissances, le raisonnement, …) [Yvon, 2007].

De 1990 à nos jours et grâce au développement de l'informatique surtout au niveau de la capacité de stockage et la vitesse des calculs, le TALN est étendu vers l'utilisation des très grandes corpus (grande masse d'informations textuelles), pour élaborer ses règles, en utilisant des statistiques et des calculs numériques. Au niveau des applications, le TALN commence à utiliser l'apprentissage automatique, pour élaborer des programmes qui s'améliorent avec l'entraînement a travers des exemples.

Aujourd'hui, Le TALN est plus florissant. Beaucoup d'applications industrielles comme traduction automatique, recherche documentaire, interfaces en langage naturel ...etc., qui commencent à atteindre le grand public, sont là, pour témoigner de l'importance de ce domaine [Yvon, 2007].

3- Les domaines d'application du TALN

Le traitement automatique des langues naturelles (TALN) compte plusieurs domaines d'application, parmi lesquelles nous dénombrons :

3-1 Traduction automatique

Les premières recherches dans le domaine de TALN c'étaient dans le domaine de la traduction automatique (T.A.) qui se fait à partir d'un langage-source vers un langage -cible. Nous comptons deux types de traduction, le premier est la traduction assisté par l'ordinateur (TAO) qui nécessite l'intervention de l'homme. Et le deuxième est la traduction entièrement automatique qui se fait sans intervention de l'homme.

3-2 Reconnaissance de la parole

L'objectif principal de la reconnaissance de la parole est de décoder l'information portée par le signal vocal à partir des données fournies par l'analyse des signaux. On distingue fondamentalement deux types de reconnaissance, en fonction de l'information que l'on cherche à extraire du signal vocal : la

reconnaissance du locuteur, dont l'objectif est de reconnaître la personne qui parle, et la reconnaissance de la parole, où l'on s'attache plutôt à reconnaître ce qui est dit.

3-3 Synthèse de la parole à partir du texte

La synthèse de la parole à partir du texte désigne les traitements permettant à une machine de transformer un texte écrit en message oral. Le but est de créer des paroles artificielles à partir d'un texte quelconque [Baloul, 2003].

3-4 Recherche d'information

C'est de faire une correspondance entre l'information recherchée, qui se présente généralement sous forme d'une requête, et l'ensemble des documents disponibles. Il s'articule autour de deux étapes essentielles :

La phase d'indexation transforme les documents afin de créer une représentation de leur contenu textuel qui soit utilisable par le système de recherche de l'information.

La phase de recherche se base sur un formalisme précis défini par un modèle de système de recherche de l'information dont l'objectif est de présenter les documents répondu à la requête de l'utilisateur. Les documents présentés considérés comme les plus pertinents. [Boulaknadel, 2008]

3-5 Dialogue homme-machine

Les systèmes (logiciels) de dialogue homme-machine sont des systèmes qui permettent une interaction entre un être humain et une machine, par le biais d'un langage naturel. Les traitements sont faits grâce à deux modules qui sont : un module d'analyse et un module de génération de la réponse.

3-6 Résumé automatique

Le résumé automatique est un processus qui consiste à produire une représentation abrégée d'un texte dit original. Le résumé automatique de texte peut être fait selon deux approches : abstraction et extraction.

L'abstraction se base sur la réécriture du texte original dans une version plus courte. Cette approche nécessite une analyse morphologique et syntaxique, plus la génération d'un texte plus court et compréhensible par l'utilisateur. Ce qui la rend très difficile.

L'extraction consiste à extraire des énoncées appropriés du texte original et les enchaîner dans une forme plus courte [Douzidia, 2004].

3-7 Génération automatique de textes

C'est les traitements faits par la machine pour produire un texte qui donne des informations correctes dans une langue. La génération, généralement nécessite deux constituants :

- Un système expert de raisonnement qui traite la question " quoi dire ? "
- Un module de génération linguistique qui traite la question " comment le dire ? "

4-Les Niveaux d'analyse

Le traitement automatique des langues naturelles se fait en quatre (04) niveaux qui sont, le niveau morphologique, le niveau syntaxique, le niveau sémantique, et le niveau pragmatique.

4-1 Le Niveau Morphologique

Comme son nom l'indique l'analyse morphologique étudie la forme des mots c'est-a-dire comment ces derniers sont formés à partir des morphèmes[1], sans considéré les règles syntaxique selon lesquelles sont combinés, dont le but est d'affecter des informations morphologique (gendre, type, …) pour chaque mots. Les morphèmes sont les unités linguistiques minimales (c'est-à-dire non décomposables) porteuses de sens [Bernhard, 2006].

On distingue les morphèmes libres, qui peuvent former un mot sans être associés à un autre morphème et les morphèmes liés qui sont toujours associés à d'autres morphèmes. C'est le cas des affixes notamment qui se combinent avec une base ou un radical [Bernhard, 2006].

Exemple : le mot arabe "يدرسون" est constitué de trois morphème qui sont : "يـ" ، "درس"، "ون"

A ce niveau, le traitement, d'un texte écrit dans une langue, commence par la segmentation du texte d'une façon à affecter des caractéristiques morphologiques pour chaque segment.

Exemple : la phrase arabe suivante "كَتَبَ التلميذُ الدرسَ" peut être analysée comme suit :

الوصف Désignation	المصدر Racine	نوع الكلمة Type de mot	الوزن Scheme	الكلمة Le mot
Verbe accompli actif 3e Personne Masculin فعل ماضي مبني للمعلوم للمفرد المذكر الغائب، مبني على الفتح	كتب	فعل Verbe	فَعَلَ	كَتَبَ
Déterminant للتعريف	//	سـابق Préfixe	//	ال
Nom إسم مرفوع	تلمذ	إسم Nom	//	تلميذُ
Déterminant للتعريف		سـابق Préfixe	//	ال
Nom اسم منصوب	درس	مصـدر	فَعْلَ	درسَ

Tableau 01 : Exemple d'analyse morphologique pour la phrase "كَتَبَ التلميذُ الدرسَ"

[1] La plus petite unité significative du langage.

4-2 Le Niveau Syntaxique

L'analyse syntaxique étudie les règles par lesquelles les mots peuvent se combiner pour former des phrases, c'est-à-dire les structures des phrases. Donc, à ce niveau on calcule la validité de certaines suites de mots pour construire des phrases grammaticalement corrects.

Pour faire ça, l'analyseur syntaxique utilise des règles grammaticales de la langue à traiter, et des statistiques pour choisir la règle la plus probable.

Dans le cadre de l'analyse syntaxique d'une phrase, il s'agit d'une segmentation en unités fonctionnelles appelées syntagmes. Par exemple, on peut citer les types de syntagme suivants : syntagme nominal, syntagme verbal, syntagme adjectival, …etc.

Tel qu'un syntagme est un ensemble de mots formant une seule unité catégorielle et fonctionnelle, mais dont chaque constituant conserve sa signification et sa syntaxe propres [Khelif, 2006].

4-3 Le Niveau Sémantique

L'analyse sémantique cherche le sens des mots et/ou des phrases. Elle utilise généralement des lexiques sémantiques [Khelif, 2006]. Elle étudie le sens des unités lexicales figurants dans une phrase exprimé dans une langue [Lison, 2006].

4-4 Le Niveau Pragmatique

L'analyse pragmatique étudie le sens des mots selon le contexte dont ils sont utilisés. Ce niveau d'analyse représente des rapports sémantiques entre les mots d'une langue et la réalité pour exprimer le sens le plus réal des mots [Khelif, 2006].

5- Les approches de développement d'une application en TALN

Généralement, il y a trois approches qui sont utilisées dans le TALN, à savoir :

5-1 L'Approche Symbolique

Cette approche se base sur des descriptions heuristiques basées sur des ressources linguistiques élaborées manuellement [Lison, 2006]. Malgré que cet approche a donné des résultats agréable dans certaines langues, il est encours difficile de formaliser tous les règles d'une langue, surtout pour l'arabe.

5-2 L'Approche Statistique

Cette approche construit des modèles de règles par l'apprentissage automatique à partir des données contenues dans un corpus, et le système sera ensuite entraîné pour s'améliorer. La bonne interprétation des unités lexicales s'opère à l'utilisation de l'interprétation la plus probable.

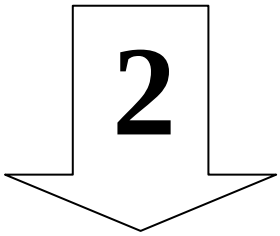
5-3 L'Approche Hybride

L'approche hybride fait combiner le meilleur des deux approches précédentes c'est- à- dire qu'elle permet d'intégrer des connaissances linguistiques à des modèles probabilistes du langage [lison, 2006].

6 Conclusion

Ce qui a été représenté dans ce chapitre n'est qu'une introduction au traitement automatique des langues, que nous avons opté pour définir le domaine dont nous allons élaborer notre analyseur morphologique du texte arabe. Alors que ce domaine reste très vaste par ses terminologies et ses outils propres.

Chapitre



La Langue Arabe : Entre Complexité et Richesse

1- Introduction

L'arabe est une langue sémitique tels que le phénicien, l'araméen, le syriaque, …etc. son nom est venu du peuple arabes qui parle cette langue.

L'arabe est la langue la plus parlée dans le groupe des langues sémitiques, et l'une des langues les plus diffusées dans le monde, parlée par plus de 422 millions de personnes. L'arabe est aussi la langue référentielle pour plus d'un milliard musulmans autour de monde.

Dans ce chapitre, nous allons étudier la langue arabe du point de vue informatique, en essayant du catégoriser ses ressources.

2- L'histoire de la langue arabe

L'histoire de la langue arabe compte trois grandes périodes [Baccouche, 2009] :

2-1 L'arabe ancien

Connu par des inscriptions remontent au VIIIème siècle .av.J.. Mais l'arabe littéraire connu ne remonte pas plus loin que le IIIème siècle .J..

La littérature est essentiellement orale (poèmes, chroniques, proverbes, etc.). La poésie ancienne est si élaborée qu'elle doit être le fruit d' une longue maturation dont l'évolution, comme pour toute tradition orale, est difficile à cerner. Comparée aux divers dialectes arabes, cette langue littéraire semble plus imprégnée par ceux du groupe ouest-arabique (Hijaz), allant de la mer rouge au plateau de Nedjd [Baccouche, 2009].

2-2 L'arabe classique

La naissance et la diffusion de l'islam jouent un rôle très important pour la diffusion de l'arabe, au début du 7ème siècle [Boulaknadel, 2008], liée au texte sacré(le Coran), l'arabe a très vite évolué vers une forme classique à la faveur de la codification des grammairiens. Sa graphie, empruntée sous une forme rudimentaire aux Nabatéens, s'est adaptée aux nouvelles exigences par l'adjonction de signes diacritiques (notamment les points et les signes des voyelles) afin d'assurer une bonne lecture du Coran dont la première version écriture officielle est établie sous le troisième calife Uthman au 1er s.H./VIIème s.J.

2-3 L'arabe moderne

C'est le fruit d'une évolution qui a duré plus d'un millénaire, avec une interaction entre l'arabe littéral et ses divers dialectes.

La renaissance de l'arabe, après une période de léthargie, s'est développée surtout dès le XIXème siècle, par un double effort : une relecture du patrimoine et une ouverture sur la culture européenne par la traduction, l'emprunt, les calques, …etc. Où cela le niveau de langue le plus perceptible est celui des médias. Aujourd'hui, le rapport littéral /dialectal est une donnée incontournable de la situation linguistique actuelle dans les pays arabes. Ceci nous amène à l'examen du cadre structurel, pour saisir le sens de l'évolution opérée du classique au moderne [Baccouche, 2009] .

D'autre part, et avec la diffusion de l'arabe sur le web, et la migration des arabe vers tous les régions du monde, les recherches et les études, sur le traitement automatique de la langue arabe, sont très nombreuse et surtout pour des raisons de traduction et de compréhension.

3- La morphologie arabe « الصرف »

C'est l'étude de la construction ou de la dérivation des unités lexicales et leur transformation selon le sens voulu, cette dérivation morphologique est décrite sur une base morpho-sémantique, tel qu'à partir de mêmes racine on dérive différents mots selon des schèmes prédéfinis [Boulaknadel, 2008].

Exemple : à partir de la racine « علم » on peut dériver les mots « عَلِمَ » (il a su), « عَالِمٌ » (savant), « عِلْمٌ » (savoir), « يَعْلَمُ » (il sait).

4 - Les caractéristiques morphologiques de l'arabe

L'arabe est une langue écrite dont l'alphabet est constitué de 29 lettres considérées toutes comme des consonnes (voire le tableau 02), et l'écriture est de droite à gauche. Toutes les lettres se lient entre elles sauf (ا,و,ر,ز,د,) qui ne sont pas joignables à gauche. D'autre part les lettres change ses formes selon leur position dans le mot (au début, au milieu ou à la fin du mot). Le tableau 02 illustre cette dernière notion sur la lettre « ه » (Ha).

A la fin précédé par une lettre non joignable à gauche	A la fin du mot	Au milieu du mot	Au début du mot
ه	ـه	ـهـ	هـ

Tableau 02: Exemple de variation de la lettre « ه » (Ha) selon sa position dans un mot

Lettres solaires		Lettres lunaires	
Lettre	Transcription	Lettre	Transcription
أ	Alef	ت	Ta
ب	Ba	ث	Tha
ج	Jim	د	Dal
ح	Hha	ذ	thal
خ	Kha	ر	Ra
ع	Ayn	ز	Zayn
غ	Ghayn	س	Sin
ف	Fa	ش	Shin
ق	Qaf	ص	Sad
ك	Kaf	ض	Dad
ه	Ha	ط	Tah
م	Mim	ظ	Zah
و	Waw	ل	Lam
ي	ya	ن	Nun

Tableau 03: l'alphabet de la langue arabe et leurs transcriptions

4-1 Voyellation

Les mots de la langue arabe sont écrits avec des consonnes et des voyelles, ces derniers sont ajoutés au-dessus ou bien au-dessous des lettres (ـَ ، ـِ ، ـُ ، ـْ), mais la plupart des textes arabe sont écrits

totalement ou partiellement sans voyelles, ce qui dirige les recherches vers ce type de textes. Les voyelles comptent deux types, les voyelles brèves et les voyelles longues, le tableau suivant présente la correspondance entre ces deux types, et la nonattion ou tanwïn. [Mesfar, 2008]

Voyelles brèves	Voyelles longues	La nonattion (tanwin)
"ـَـ " (a)	" ا "	" ً" (an)
"ـِـ " (i)	"ي "	" ٍ " (in)
" ـُـ" (u)	"و "	" ٌ " (un)
" ـْـ " (sukun)	sans voyelle long	sans tanwin

Tableau 04: Les voyelles de la langue arabe et leurs transcriptions

La plupart des textes arabe diffusés dans le monde sont non voyellés ce qui impose un problème d'ambiguïté parfois même pour l'homme, et pour la machine le ratio d'ambiguïté sera très élevé.

Les voyelles sont nécessaires pour la lecture et la compréhension correcte d'un texte arabe. Par ce qu'elles permettent de distinguer les mots ayants la même représentation morphologique, où l'absence des voyelles représente une sorte d'ambiguïté.

Exemple : le mot non voyellé كتب (ktb) peut avoir 17 voyellation différents, comme le montre le tableau04 ci aprés.

Voyellation	Translittération	traduction	Catégorie grammaticale
كَتَبَ	KaTaBa	a écrit	Verbe accompli, voix active, 3ème personne, masculin, singulier.
كُتِبَ	KuTiBa	a été écrit	Verbe accompli, voix passive, 3ème personne, masculin, singulier.
كَتَّبَ	KaTTaBa	a fait écrire	Verbe accompli, voix active, 3ème personne, masculin, singulier.
كُتِّبَ	KuTTiBa	fait écrire	Verbe accompli, voix active, 3ème personne, masculin, singulier.
كَتِّبْ			Verbe impératif, voix active, 2ème personne, masculin, singulier.
كُتُبُ	KuTuBu	Des livres	Substantif, masculin, pluriel, nominatif, déterminé.
كُتُبَ	KuTuBa		Substantif, masculin, pluriel, nominatif, déterminé.
كُتُبِ	KuTuBi		Substantif, masculin, pluriel, génitif, déterminé.
كُتُبٌ	KuTuBun		Substantif, masculin, pluriel, nominatif, déterminé.
كُتُبٍ	KuTuBin		Substantif, masculin, pluriel, génitif, déterminé.
كَتْبُ	KaTBu	Un écrit	Substantif, masculin, singulier, nominatif, déterminé.
كَتْبَ	KaTBa		Substantif, masculin, singulier, nominatif, déterminé.
كَتْبِ	KaTBi		Substantif, masculin, singulier, nominatif, déterminé.
كَتْبٌ	KaTBun		Substantif, masculin, singulier, nominatif, déterminé.
كَتْبٍ	KaTBin		Substantif, masculin, singulier, génitif, déterminé.
كَ + تَبِّ	Ka +tabbi	Comme le tranchement	Préposition + Substantif, masculin, singulier, génitif, déterminé.
كَ + تَبٍّ	Ka +tabbin	Comme le tranchement	Préposition + Substantif, masculin, singulier, génitif, déterminé.

Tableau 05 : Exemple de voyellation du mot non voyellé كتب (ktb) [Mesfar, 2008]

4-2 Flexion

Les mots de la langue arabe variaient en nombre, temps, Personne, gendre, … etc. on peut illustrer cette notion avec les exemples suivants :

- Pour le verbe "ذهب " (aller), on dire "ذَهَبَ" (il est allé) avec la 3ièmme Person du singulier, "ذَهَبْتُ" (je suis allé) avec la première Personne du singulier.
- Pour le nom "مسلم" (musulman) en masculin singulier, "مسلمان" (muslimani) pour le masculin dual ; ou "مسلمون" (muslimun) pour le masculin pluriel.

4-3 Agglutination

L'arabe est une langue agglutinative, tel que les morphèmes collent les unes aux autres pour constituer des unités lexicales très complexes où un mot peut porter le sens d'une phrase dans d'autres langues (le français par exemple) [Boulaknadel, 2008]. Ce phénomène pose un grand problème pour le traitement automatique de l'arabe où il augmente considérablement le taux d'ambiguïté surtout au niveau de la segmentation des mots.

Exemple : le mot « يدرسونه » porte le sens de la phrase française « ils l'étudient » comme le montre le tableau suivant.

Enclitique	suffixe	Base
" ه " Pronom personale suffixe. (Complément)	"ونَ" Suffixe verbal exprimant le pluriel. (Sujet)	"يدرس " Verbe dérivé de la racine درس / DRS

Tableau 06 : exemple de l'agglutination de l'arabe

4-4 Mécanisme de dérivation

La plupart des mots arabes sont dérivés de leurs racines, selon des gabarits appelés les schèmes. Le processus de dérivation consiste à ajouter ou insérer des préfixes, des suffixes et/ou des infixes aux lettres de racine pour produire un nouveau mot, qui a une nouvelle fonction ou signification mais préserve le concept ou la signification principal portée par la racine. D'ailleurs, l'utilisation du mot dérivé dans un certain contexte exigera des clitiques à ajouter au début et à la fin du mot. Les proclitiques incluent des prépositions, des conjonctions et des articles définis, et les enclitiques incluent des pronoms relatifs. En outre, un ou plusieurs affixes ou clitique peuvent être ajoutés au mot dérivé [Sawalha, 2011].

4-4-1 Le Schème « الوزن »

Est un mot composé de trois lettres de base qui sont "ف" ، "ع" ، "ل" et qui peut être augmenté par d'autres lettres (préfixe, suffixe et infixe). Ces lettres correspondent à ses positions dans le schème comme suite :

- La lettre "فـا " dite lettre de la première position.
- La lettre "عـ " dite lettre de la deuxième position.
- La lettre "ل " dite lettre de la troisième position.

Le schème joue un rôle très important dans le processus de génération des formes dérivées à partir d'une racine. Autrement dit le schème peut être considéré comme un moule dans lequel on fait couler la racine, Pour produire un nouveau mot pourtant un sens.

4-4-2 La Racine « الجذر »

Est une suite de trois, quatre, ou cinq consonnes formant les lettres de base pour dériver des mots, En effet, à chaque racine correspond un champ sémantique et à l'aide de différents schèmes, on peut générer une famille de mots appartenant à ce champ sémantique [Khemakhem, 2006].

Exemple de dérivation

La dérivation du nom d'agent « إسم الفاعل » à partir d'un verbe trilitère en insérant la lettre « ا » (alif) après la première lettre, en utilisant le schème « فاعل »(voir la Figure 01).

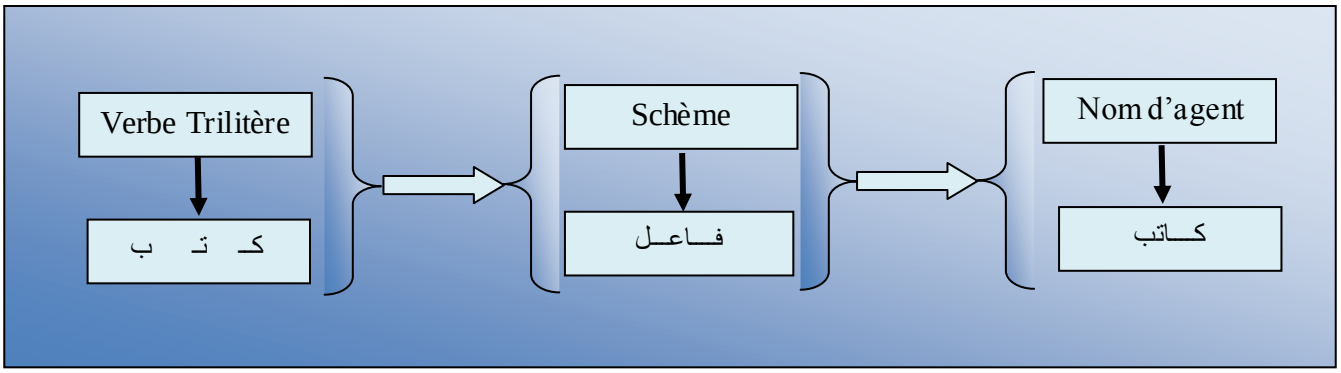


Figure 01: Exemple de dérivation d'un agent « كاتب »

4-5 Structure d'un mot arabe

Un mot arabe peut porter le sens d'une phrase complète à cause de sa structure complexe qui peut être décomposé en 5 éléments collés entre eux, ces éléments sont : **Proclitique**, **Préfixe**, **Base**, **Suffixe** et **Enclitique** [Boulaknadel, 2008].

Le tableau 07 suivant représente les différents composants d'un mot arabe (maximal) dont l'ordonnancement est de droite à gauche.

enclitique	suffixe	base	préfixe	proclitique

Tableau 07 : Les composants d'un mot arabe

- Les proclitiques sont des prépositions ou des conjonctions (coordonnants, déterminants,...etc.)
- Les préfixes et suffixes expriment des traits grammaticaux, tels que les fonctions des noms, le mode du verbe, le nombre, le genre, la personne…etc
- La base est une combinaison de lettres radicales (le plus souvent trois) et d'un schème.
- Les enclitiques sont des pronoms personnels dans la plupart des cas.

Exemple : la décomposition du mot arabe « أَتَتَذَكَّرونَهُم»

Enclitique	suffixe	base	préfixe	proclitique
" هم"	"ونَ"	"تَذَكَّر "	" ت"	" أ"

Tableau 08: Exemple de la décomposition du mot arabe « أَتَتَذَكَّرونَهُم»

4-6 Catégories des mots arabes

On distingue 3 grandes catégories des mots arabes : les **Verbes**, les **Noms** et les **Particules**.

4-6-1 Morphologie verbale

Les verbes sont des entités exprimants des sens dépendants du temps. Les verbes arabes sont formés à partir des radicaux de 3 consonnes dans la majorité des cas, et éventuellement de 4 consonnes.

4-6-1-1 Les temps des verbes arabe«الأزمنة»

La conjugaison des verbes arabe compte 3 temps :

- **L'accompli**

  Exprimant un événement qui a été fait au passé.

  Exemple le verbe كَتَبَ (kataba il a écrit).

  L'accompli est caractérisé par deux chose :

  - Il peut être combiné avec تاء الفاعل (ت Ta de l'agent) comme : كَتَبْتُ (j'ai écrit)
  - Il peut être combiné avec تاء التأنيث (ت Ta de féminin) comme : كَتَبَتْ (elle a écrit)

- **L'inaccompli**

  Indique que l'action est encours de réalisation ou sera réalisé (présent ou future). L 'inaccompli toujours commence par l'une des lettre suivant : « ن », « أ », « ي », « ت » qui sont regroupés dans le mot « نَـأَيْـتُ».

  Exemple : « يفكر» (il réfléchit)

- **L'impératif**

  Demande de réaliser une action dans la future.

  Exemple : « اغلق» (Ferme)

4-6-1-2 Les types de verbe

Lorsqu'aucune des consonnes radicales n'est pas une voyelle longue le verbe est dit « صحيح » (sain), dans le cas contraire le verbe est dit « معتل » (défectueux).

- **Le verbe sain « الفعل الصحيح »**

tous les verbes dont aucune des lettres radicales n'apparient pas a l'ensemble [ا , و , ي] (l'ensemble des voyelle longues). Les verbes sain sont de trois types [Chacha, 2008]:

- Si il y a deux consonnes radicales identiques en deuxième et troisième, le verbe est dit redoublé « مضاعف » (mudaaf)
- Si une des trios consonnes radicales est une " أ " (a hamza) indépendamment de sa position on dit que le verbe est « مهموز » (hamzé).
- Si le verbe sain n'est pas redoublé ou hamzé, il est donc un verbe régulier « سالم »

- **Le verbe défectueux « الفعل المعتل »**

Tous les verbes dont au moins une de ces lettres d'origine appartient à l'ensemble [ا , و , ي] (l'ensemble des voyelle longues). Le verbe défectueux, à son tour, se divise en quatre types selon la position des voyelles longues [Radjihi , 1973]:

- Dans la première position le verbe est dit « مثال » (assimilé).

  Exemple: « وَعَدَ »
- Dans la deuxième position le verbe est dit « أجوف » (creux).

  Exemple: « قَالَ»
- Dans la troisième position le verbe est dit «ناقص » (faible).

  Exemple: « رَمَى », «سَعَى »
- Si deux parmi les 3 consonnes radicales sont des voyelles longues, le verbe est dit « لفيف » (lafif). On distingue deux types de lafif :

  - Lafif joint (لفيف مقرون) : si les deux voyelles longues sont successives.

    Exemple : « رَوَى », et « أوى »
  - Lafif séparé (لفيف مفروق) : si les deux voyelles longue sont séparés[Radjihi , 1973].

    Exemple : « وَقَى», et « وَعَى »

4-6-2 Morphologie Nominale

Les noms sont des mots qui portent un sens indépendant du temps. Ils désignent un être ou un objet. Ils sont caractérisés par :

- La déclinaison Génitif « الجر » : la voyellation du nom comporte le cas du génitif marqué par le signe de voyellation "ـِ"
- L'appel «النداء » : c'est que le nom peut être précédé par une lettre d'appel.
  Exemple : « يا محمد » où ("يا": lettre d'appel) et ("محمد" : Nom Propre)
- La combinaison avec le déterminant " ال "
- La nonation « التنوين »: c'est un « n » qui suit la dernière lettre d'un nom, mais qui se prononce et n'est pas écrit, elle est remplacer par les doubles signes de voyellation (ً، ٍ، ٌ) dans l'écriture arabe.
  Exemple : كِتَابٌ (kitabo<u>n</u>)

On distingue trois catégories des noms arabes [Mesfar, 2008] :

4-6-2-1 Les noms primitifs

Ce sont des noms qui ne se dérivent pas d'une racine verbale.

Exemple : كبش , كرسي ,أخ

4-6-2-2 Les noms dérivés

Ce sont des noms qui se dérivent à partir d'une racine selon des schèmes prédéfinis, le nombre des noms dérivés de la même racine varie selon le statut de la racine dont ils se rattachent. Les noms dérivés de la langue arabe sont :

- **Le nom d'agent « اسم الفاعل»**

C'est un nom dérivé d'un verbe pour designer qui a fait l'action.

Exemple : كَتَبَ (il a écrit) ⟹ كَاتِبٌ (écrivain)

- **Qualificatif de supériorité «إسم التفضيل»**

C'est un nom signifie que deux choses ont un caractère commun mais un est plus qualifié que l'autre.

Exemple : « محمد <u>أَكْفَـأُ</u> من خالد »

- **Qualificatif assimilé «الصفة المشبهة»**

C'est un nom dérivé qui a le sens du non d'agent mais le qualificatif assimilé ne se dérive qu'a partir d'un verbe non transitif trilitère pour signifier que le qualificatif est permanant et confirmé.

Exemple : « جَـريح »

- **Nom de patient « اسم مفعول »**

C'est un nom dérivé d'un verbe pour designer qui a subi à l'action. Il ne se dérive qu'à partir des verbes transitifs.

Exemple : « مَسْحُوق »

- **Nom de temps « اسم الزمان »**

C'est un nom dérivé d'un verbe pour indiquer le temps de déroulement de l'action.

Exemple: « مَغْـرِب »

- **Nom de lieu « إسم المكان »**

C'est un nom dérivé d'un verbe pour indiquer le lieu du déroulement de l'action.

Exemple: «مَسْبَح »

- **Nom d'instrument « اسم الآلة »**

C'est un nom dérivé d'un verbe pour designer l'outil utilisé pour réaliser l'action.

Exemple : « مِنْجَرَة »

- **Qualificatif intensifié « صيغة مبالغة »**

C'est un nom dérivé d'un verbe. Il porte le sens intensifié du nom d'agent.

Exemple : « عَلَّام »

4-6-2-3 Les pronoms

Les pronoms sont considérés comme un ensemble des noms particuliers dans la langue arabe, et échappent à toute règle de dérivation. Dans cet ensemble, on distingue :

- **Les pronoms personnels :** Il existe deux types de pronom dans la langue arabe, les pronoms isolés et les pronoms liés (voir le Tableau 09). Le pronom séparé remplace le sujet dans une phrase arabe.

	les pronoms isolés			les pronoms liés		
Sémantique	arabe	prononciation	française	arabe	prononciation	française
1er singulier	أنا	Ana	je	ـي	-Y	je
2ème masculin singulier	أنتَ	Anta	Tu (masculin)	ـكَ	-Ka	toi (masculin)
3ème féminin singulier	أنتِ	Anti	Tu (féminin)	ـكِ	-Ki	Toi (féminin)
3ème masculin singulier	هو	Houwa	il	ـه	-Ho	lui
3ème féminin singulier	هي	Hya	elle	ـها	-Haa	elle
1er pluriel	نحن	Nahnou	nous	ـنا	-Naa	-
2ème masculin ou mixte pluriel	انتم	Antom	Vous (masculin ou mixte)	ـكم	-Kum	-
2ème féminin pluriel	انتن	Antonna	Vous (féminin)	ـكن	-Kunna	-
3ème masculin ou mixte pluriel	هم	Hom	ils	ـهم	-Hum	-
3ème féminin pluriel	هن	Honna	elle	ـهن	-Hunna	-

Tableau 09 : Les pronoms personnels arabes

- **Les pronoms démonstratifs « أسماء الإشارة» :** Ce sont des pronoms exprimant une idée de démonstration (voir le Tableau 10). Ils permettent d'indiquer que l'objet représenté se trouve, soit dans le texte, soit dans l'espace ou le temps [Mesfar, 2008].

Pronom	prononciation	français
هذا	Hadha	Ceci, Celui-ci
هذه	hadhi-hi	Celle-ci
هؤلاء	haoulai	Ceux-ci
ذلك	dhalika	Celui-là
تلك	tilka	Celle-là
أولئك	oulaika	Ceux-là

Tableau 10 : Les pronoms démonstratifs

- **Les pronoms Relatifs «الأسماء الموصولة » :** Ils se rapportent aux noms ou aux pronoms personnels qui les précèdent et qui nous désignons par antécédente (voir le Tableau 11).

Pronom	prononciation	Français
الذي	alla-dhi	Qui
الذي	alla-dhi	Lequel
التي	alla-ti	Laquelle
الذين	Alla-dhin	
اللواتي	alla-lawati	Lesquelles

Tableau 11 : Les pronoms relatifs

4-6-3 Les particules

Ce sont des mots qui ne portent aucun sens, s'ils ne sont pas combinés avec un verbe ou un nom. Donc, il y on a des particules verbales, nominales, et communs (verbale et nominales). Certaines particules peuvent être des suffixes ou préfixes dans un mot complexe [Cheragui, 2010].Il existe plusieurs sous-catégories des particules dans la langue arabe parmi lesquelles, on peut citer :

- **Les particules du génitif « حروف الجر»**

  elles sont en nombre de dix neuf (19) : « من ، إلى، عن، على، في، و، ب، تـ ، كـ ، لـ، عدا، رُبَّ،خلا،حاشا،حتى،منذُ ،مُذ، لولا، كي، »

  Exmeple : « خرجت من المنزل »

- **Les particules du conditionnel « حروف الشرط »**

  elles sont en nombre de huit (08) : « لوما، لولا،لو، على، إنْ، أمّـا، ما، إذما »

  Exemple : «لو زرعت، لحصدت»

- **Les particules d'exception « حروف الإستثناء »**

  Ils sont en nombre de six (06) : « عدا، سِوى، خالا، حاشا، غيرَ، إلا »

  Exemple : « قل غير ذلك »

- **Les particules du futur « حروف الإستقبال »**

  Appelées aussi lettres du futur, elles sont « س, سوف », qui s'utilise avec l'inaccompli.

  Exemple : «سنضحي »

- **Les particules de conjonction « حروف العطف »**

  Ils sont de nombre neuf (09) : « و, و، لا، لكن ، ف، بل، حتى،ثم، أو »

  Exemple : «فستبصر و يبصرون »

- **Les particules de serment « حروف القسم »**

  Ils sont en nombre de quatre (04): « ب, ت , ل , و »

  Exemple : « والـلّــه » (Je jure par Dieu).

5 - Conclusion

Les informations exposées dans ce chapitre, ne représentent que l'essentiel de la morphologie arabe, sachant que cette langue est très riche morphologiquement, ce qui complique les traitements. Dans le chapitre qui suit, nous expliquerons les difficultés confrontées pondant la conception et la réalisation de notre analyseur morphologique.

Chapitre

Conception du système AMTAR (Analyseur Morphologique du Texte Arabe)

1-Introduction

L'importance des outils de traitement automatique de la langue arabe a considérablement augmenté dans la dernière décennie en raison de développement énorme du contenu numérique arabe sur internet. Ce fait augmente l'importance de créer les outils de traitement automatique qui peuvent traiter ce contenu [Sonbol et al, 2011].

L'analyse morphologique est une étape importante dans le traitement automatique de l'arabe en raison de la structure morphologique complexe de l'arabe où nous avons des infixes avec des préfixes et des suffixes. En outre, chaque préfixe ou suffixe peut avoir sa propre étiquette syntaxique ; ceci signifie que nous devons employer le résultat de l'étape d'analyse morphologique à des étapes plus élevées du traitement arabe comme l'étiquetage, et l'analyse syntaxique.

L'analyse morphologique de la langue arabe, comme les autres langues, étudie la forme des mots et les variations qu'ils subissent dans la phrase [Barakat, 2006].

L'arabe est une langue dont la morphologie est riche comparée à d'autres langues, est une langue particulièrement, basé sur la morphologie dérivative et flexionnelle. La richesse de la morphologie arabe

rend le processus d'analyse difficile. D'une part, le processus d'analyse morphologique est employé dans la plupart des applications de TALN telles que la recherche documentaire, la vérification d'orthographe et la traduction automatique. Et d'autre part, elle est la première étape avant l'analyse syntaxique, et l'analyse sémantique. [Gridach et al, 2011]

La phase du conception reste une étape essentielle pour n'importe quel système en informatique. Dans le chapitre actuel, nous allons proposer une conception pour notre système AMTAR (Analyseur Morphologique du Texte Arabe), et ses différents modules de traitement. Mais avant de commencer nous allons citer quelques travaux antérieurs qui s'occupent de l'analyse morphologique.

2-Quelques travaux antérieurs

Les systèmes d'analyse morphologique de l'arabe sont nombreux, dans ce point, nous allons présenter trois systèmes d'analyse morphologique, à titre d'exemple.

2-1 L'analyseur morphologique AlKhalil

Est un programme open source, proposé par L'organisation arabe pour la culture et les sciences, développé dans le langage de programmation java, et peut être utilisé sur des défirent environnement (Windows, Linux et Mac OS et Solaris).

Il est caractérisé par une indépendance entre les bases de données et le programme, la possibilité de mettre à jour les bases de données (ajouter / supprimer / modifier), le programme indexe tous les mots du texte en sélectionnant le nombre de fréquence et l'emplacement de mots dans texte. Il analyse 92% de la totalité des mots des textes, et permet d'enregistrer les résultats en même temps en html et csv[1].

2-2 L'analyseur Morphologique de Sakhr

Est un analyseur synthétiseur morphologique qui fournit l'analyse de base pour tout mot arabe. Cet analyseur couvre toute la langue arabe moderne et classique. L'analyseur identifie toutes les bases possibles d'un mot, c'est-à-dire trouver sa forme de base après l'extraction des suffixes et préfixes. Il n'existe pas, à notre connaissance, de version d'essai pour cet analyseur ; une version est disponible à la vente.

2-3 L'analyseur Morphologique Aramorph

Aramorph est un analyseur morphologique qui est entièrement développé en orienté objet, ce programme informatique est donc une collection de classes indépendantes et d'objets. AraMorph identifie la plupart des traits morpho-syntaxiques inclus dans le mot graphique. Il segmente les unités lexicales, repère les différents composants et atteste son appartenance à la langue. Pour cela, le système exploite

[1] Est un fichier dont l'extension est csv (Comma Separated Values), et peut être exploré par un tableur tel que Excel.

lexique DINAR.1[2] pour éviter les analyses théoriquement possibles et inexistantes dans la langue. Par la suite, l'analyseur donne une liste des traits associés à l'unité lexicale en entrée. Il offre deux types d'options. Le premier vise les traits morphosyntaxiques, le second concerne l'analyse des préfixes et suffixes. En plus des étiquettes morphosyntaxiques, il donne en sortie d'autres informations comme la base, l'unité lexicale minimale vocalisé ou non ainsi que la forme complète supposée vocalisée ou non. Analyser les préfixes revient à décrire ses découpages possibles et d'examiner les compositions des clitiques. Ceci amène le système à faire la distinction entre les clitiques ayant la même forme mais appartenant à des catégories syntaxiques différentes [Ramzi, 2004].

3- Architecture générale du système AMTAR

Notre système (AMTAR) est un analyseur morphologique de texte arabe non voyellé et ses constituants élémentaires pour arriver à déterminer leurs caractéristiques morphologiques. La figure 02 représente l'architecture générale de notre système AMTAR.

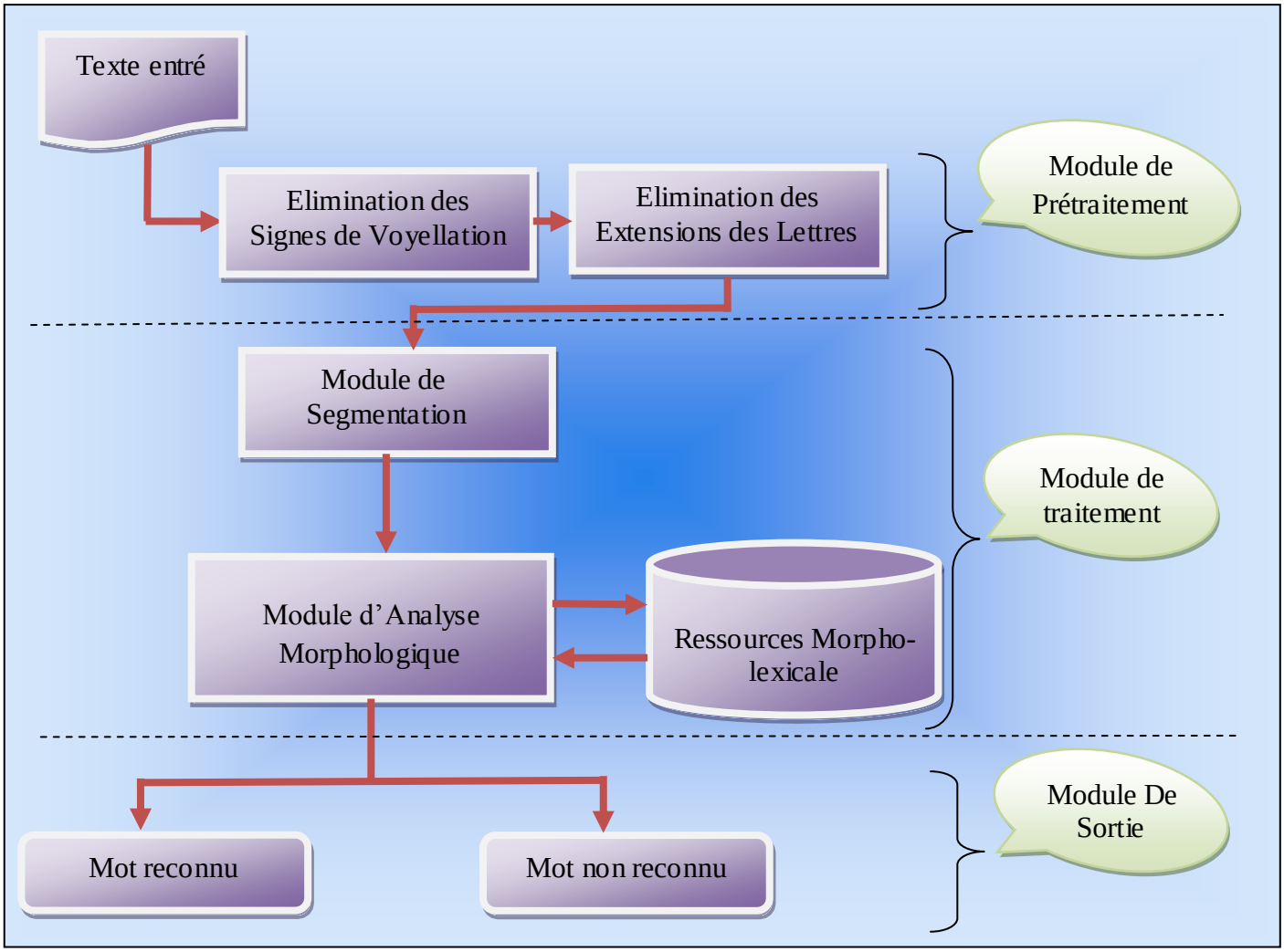


Figure 02 : architecture générale du système AMTAR

[2] Dictionnaire Informatisé de l'Arabe version 1

4- La conception de la base de donné linguistique

Une phase essentielle dans l'implémentation d'un analyseur morphologique de la langue l'arabe est la conception et la création d'une base de données linguistiques, qui inclut la majorité des primitives linguistiques arabes. Cette base de données doit comporter toutes les primitives linguistiques de l'arabe telles que les particules, les schèmes des verbes, les schèmes des noms dérivés, les affixes et les mots spéciaux.

La base de données linguistique que nous avons conçue comporte six (06) tables qui sont : Racines, Schèmes, Préfixes, Suffixes, Mots outils, et Mots spéciaux.

4-1 La table des Racines

On appellera ici "racine" l'ensemble de 3 consonnes (parfois 4 ou 5) qui donne une notion générale. Une famille de mots peut ainsi être générée d'un même concept (notion générale sémantique à partir d'une seule racine à l'aide de différents schèmes).

Exemple : - Racine des 3 consonnes ك ت ب "ktb" porte la notion de « écrire »

- Racine des 4 consonnes ز ل ز ل" zlzl " porte la notion de « tremblement du terre »

4-2 La table des Schèmes

Les schèmes sont des gabarits qui permettre de dériver une famille de mots quand ils sont combinés avec une racine. Ils seront utilisés pour analyser les verbes et les noms dérivés (voire le tableau 12).

Schème	Description arabe	Description français
فَعَلُوا	فعل ماضي مبني للمعلوم للجمع المذكر الغائب، مبني على الرفع.	Verbe Accompli Actif 3e Personne Masculin Pluriel, Invariable Indicatif.
فَعَلَتْ	فعل ماضي مبني للمعلوم للمفردالمؤنث الغائب، مبني على الفتح	Verbe Accompli Actif 3e Personne Féminin Singulier, Invariable Accusatif.
فَعَلَتَا	فعل ماضي مبني للمعلوم للمثنى المؤنث الغائب، مبني على الفتح.	Verbe Accompli Actif 3e Personne Féminin Duel, Invariable Accusatif.

Tableau 12: Exemple des lignes de la table des schèmes

4-3 La table Préfixes

La table des préfixes contient la plupart des préfixes arabes qui peuvent être combinés avec les bases pour ajouter un sens.

les préfixes arabes			
بال	لكَ	يـَ	أ
أللْ	فسَ	يـُ	بـِ
أفل	فلْ	لِلْ	فـَ
أول	وسَ	وَبـِ	وَ
أوبال	ولَ	فكَ	كَ
أفبال	ولْ	فل	لِ
وبال	فلَ	لبـِ	لْ
فبال	فللْ	ابـِ	لَ
وكال	وال	ال	سَ
ابال	فال	اكَ	تـَ
اكال	كال	اوَ	نـُ
وللْ	اوللْ	افَ	تـُ
//	أفللْ	ال	وكَ
//	أفبـِ	أسَ	وَل

Tableau 13: Liste des préfixes arabes.

4-4 La table des Suffixes

Cette table contient la plupart des suffixes arabes qui peuvent être combinés avec les verbes et les noms et ne se combinent pas entre eux.

Dans le cas des verbes, les suffixes ne dépendent pas uniquement de l'aspect et du pronom mais aussi du type de verbe [Ramzi, 2004].

les suffixes arabes		
يْنِ	نِي	هُ
اة	نَا	هْ
تا	هَا	كَ
وه	انِ	كِ
هُمَا	كُمْ	يٌّ
هُنَّ	هُمْ	نَّ
تُمَا	انِ	تِ
تُنَّ	ونَ	و
كُمَا	ينَ	ة
كُنَّ	هِمْ	ي
وها	وهم	//

Tableau 14: Liste des suffixes arabes

4-5 La table des Mots outils

Cette table contient la plupart des particules que nous avons mentionnées dans le chapitre précédent. Le tableau 14 ci-dessus représente quelques exemples.

Mot spécial	Désignation arabe	Désignation français
أنّى	ظرف مكان ، يفيد الشرط الإستفهام	Nom Circonstanciel de lieu.
حَتَّى	حرف جر، يفيد الابتداء - الاستثناء - التعليل - العطف - الغاية.	Particule de Subordination.
أُولَئِكَ	اسم اشارة جمع المذكر او المؤنث ، يستعمل للمتوسط	Nom Démonstratif

Tableau 15: Exemples des mots outils

4-6 La table Mots spéciaux

Contient les noms propres, les noms des objets, les noms des animaux les jours de la semaine, les mois, les noms des pays, …etc.

Exemple : الجزائر (Algérie) : إسم دولة (nom de pays)

5- Le module de prétraitement

Le texte analysé par AMTAR est un texte non voyellé, donc un module de prétraitement est nécessaire pour éliminer les signes de voyellation (si elle existe) et permettre le traitement des textes voyellés, en plus n'importe quel texte doit subir a une normalisation pour éliminer les extensions des lettres.

6- Le module de traitement

Se divise en deux phases qui sont :

- La phase de segmentation
- La phase d'analyse.

6-1 La phase de segmentation

La segmentation est une phase primordiale dans le traitement automatique d'un texte, son rôle est de découper un texte en unités lexicales [Mouelhi, 2008] [Ghassan, 2002]. Cette opération consiste à diviser le texte en mot pour qu'on puisse ensuite affecter les informations morphologiques (verbe, nom, adjectif, adverbe, déterminant, …etc.) pour chaque mot. Notre système fait segmenter le texte selon trois niveaux qui sont:

6-1-1 La segmentation du texte en phrase

La segmentation d'un texte en phrase sur la base des ponctuations. C'est à dire l'élimination des signes des ponctuations majeurs.

Exemple : « أخذ محمد قلمه، وبدأ بالكتابة»

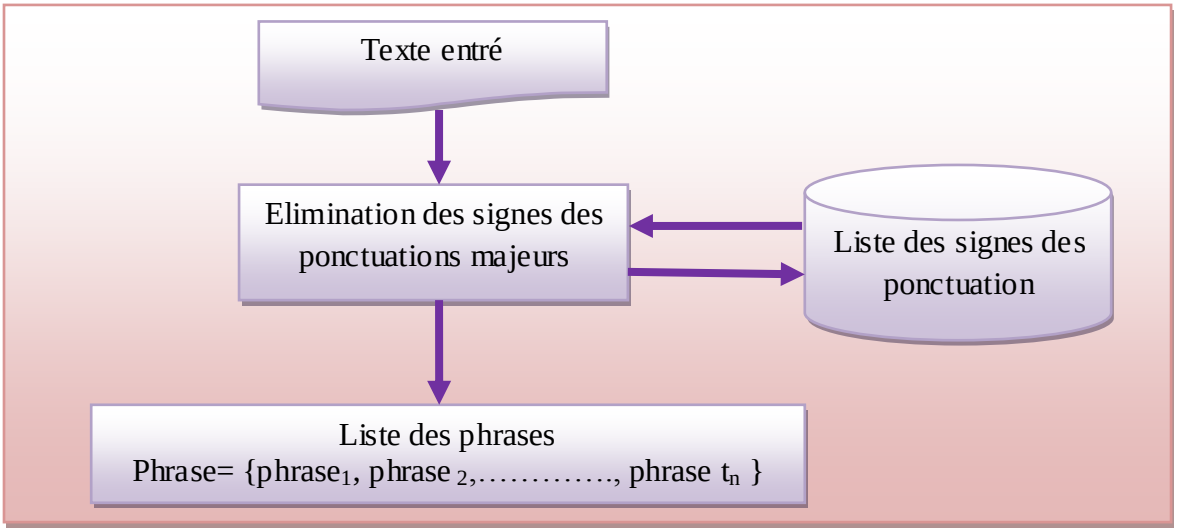


Figure 03 : architecture du module de segmentation du texte en phrase

6-1-2 La segmentation des phrases en mots

La segmentation sur la bas de l'élimination des blancs et des signes de ponctuation mineurs.

Exemple: on prend la phrase « أهدى الأب إلى ولده كتابا فبدأ بقراءته » la segmentation de ce dernier donne les unités morpho-lexicales représentés dans la figure suivante:

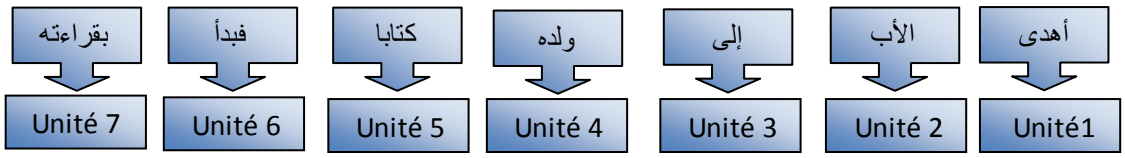


Figure 04 : Exemple de la segmentation du la phrase arabe « أهدى الأب إلى ولده كتابا فبدأ بقراءته »

6-1-3 La segmentation du mot

La segmentation au sien du mot va plus loin que la segmentation de phrase en mots, en terme de découpage. Elle consiste à isoler les différents constituants des mots arabes (proclitiques, préfixes, base, suffixes et enclitiques).

Exemple: la segmentation du mot « سندرسه » donne le résultat représenté dans le tableau suivant:

Suffixe لاحق	Base جذع	Préfixe سابق
« ه »	« ندرس »	« س »

Tableau 16 : Exemple de segmentation du mot arabe « سندرسه »

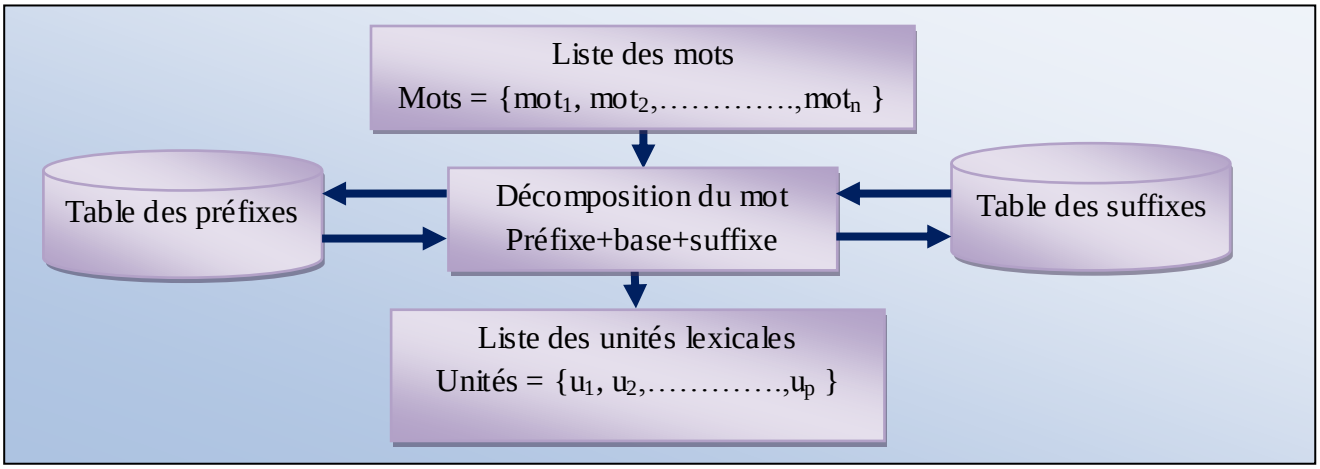


Figure 05 : Organigramme du module de segmentation du mots

6-2 la phase d'analyse

Le programme cherche les mots considérés comme des mots non décomposable (les mots qui ne nécessitent pas une segmentation) quelque soit leurs types (noms propres, particules, noms dérivés, racine) et les traiter directement.

6-2-1 Traitement d'un mot spécial

Le traitement des mots spéciaux se fait directement par une projection sur la table des mots spéciaux, selon le processus représenté dans la figure suivante (voire la figure 06)

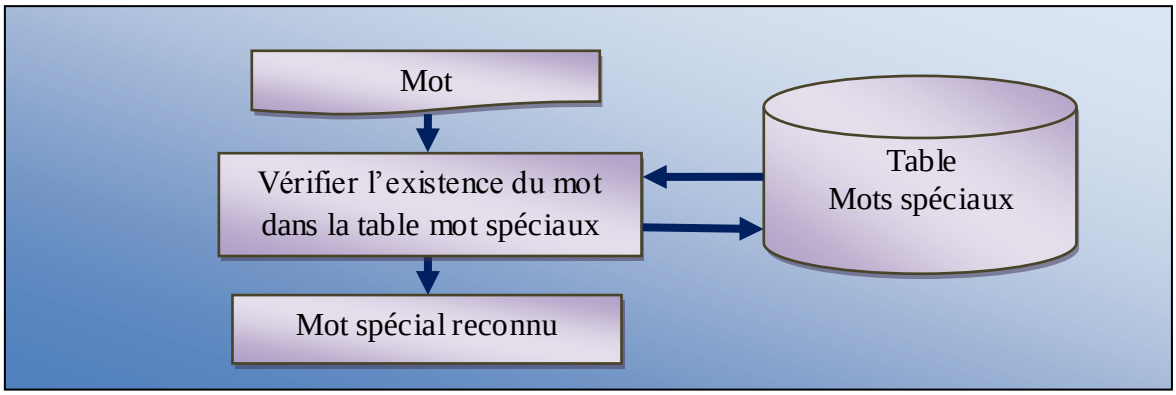


Figure 06 : Organigramme du traitement des mots spéciaux

6-2-2 Traitement des mots outils

La catégorie des mots outils contient les particules dont nous avant mentionner au chapitre précédent (voire chapitre 2, les particules) le traitement des éléments de cette catégorie est similaire au traitement des mots spéciaux et se fait directement par projection sur la table des mots outils.

6-2-3 Traitement d'un nom dérivé

Le traitement d'un nom dérivé se fait selon les étapes suivantes :

- Extraire les schèmes ayants la même longueur que ce dernier.

- Extraire la racine du nom dérivé par un lemmatiseur.
- Vérifier la concordance des lettres radicales et additives.
- Affecter les traits morphologiques du schème au nom dérivé.

Exemple: le mot dérivé « كاتب » peut être analysé comme suite :

- Extraire le schème qui a la même longueur « فاعل »

- Extraire la racine « كتب »

- vérifier la concordance des lettres additives « ا »
- Affecter les caractéristiques morphologiques du schème au mot.

	Lettre radical	**Lettre radical**	**Lettre additive**	**Lettre radical**
Le mot arabe de la droite à gauche	**ب**	**تـ**	ا	**كـ**
schème	**ل**	**ع**	ا	**ف**
racine	**ب**	**تـ**		**كـ**
Caractéristiques morphologiques	**اسم فاعل من الفعل الثلاثي | Participe Actif du Verbe Trilitère.**			

Tableau 17: Exemple de l'analyse du nom dérivé « كاتب »

6-2-4 Traitement des affixes

La catégorie des affixes comprend les préfixes, les suffixes, et les infixes[3], L'analyse des préfixes, et des suffixes commence par segmenter le mot en ses différents composants (préfixes, base, et suffixes), puis extraire les préfixes et les suffixes tel que pour chaque mot décomposable nous prenons l'affixe le plus long, et nous cherchons dans les tables des affixes (préfixes et suffixes).

affixes	types	Description arabe	Description français
كُمْ	suffixe	ضمير متصل للجمع المذكر المخاطب	Pronom Affixé 2e Personne Masculin Pluriel
أ	préfixe	حرف استفهام	Particule d Interrogation

Tableau 18 : Exemple du traitement des affixes

6-2-5 Traitement des verbes

Le traitement dépond à son type (sain ou défectueux). Dans ce qui suit nous allons détailler chaque traitement à part.

3 les infixes sont les lettres insérés au milieu d'un mot pour ajouter un sens

6-2-5-1Traitement des verbes sains

Les verbes sains sont analysés par le biais d'un module de programmation, selon les étapes suivantes :

- Extraire les schèmes ayants la même longueur du verbe.
- Extraire sa racine (ses lettres radicales).
- Vérifier la concordance des lettres radicales et additives.
- Affecter les traits morphologiques du schème au verbe.

Exemple : Un des résultats possibles pour le verbe « وقف » est représenté dans le tableau suivant

Le mot	schème	Description arabe	Description français
وقف	فعل	فعل ماضي مبني للمعلوم للمفرد المذكر الغائب	Verbe Accompli Actif 3e Personne Masculin Singulier

Tableau 19 : Exemple de l'analyse de verbe « وقف »

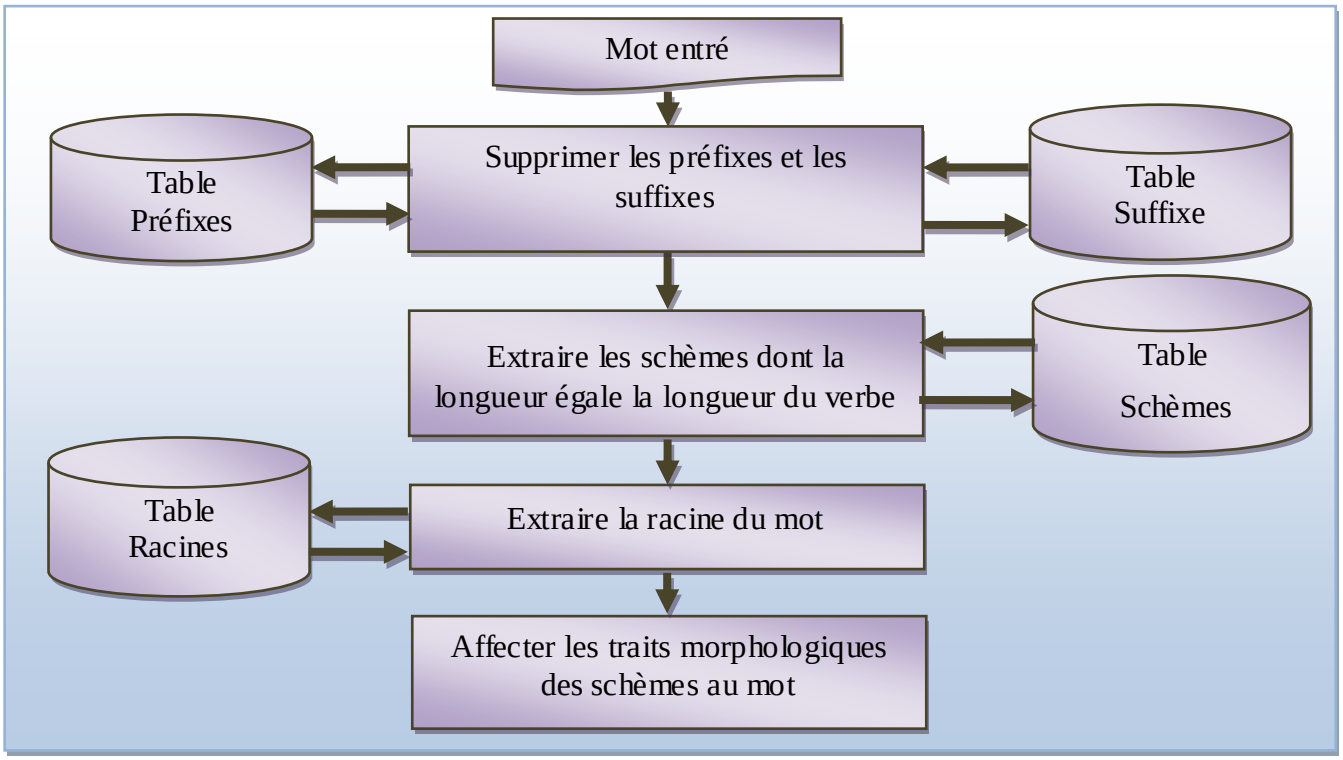


Figure 07 : Organigramme du traitement d'un verbe sain

6-2-5-2 Traitement des verbes défectueux

Pour les verbes défectueux[4], le traitement commence par une phase de normalisation qui dépond du type des verbes défectueux :

[4] au moins une de ces lettre d'origine appartient à l'ensemble [ا , و , ي]

- Pour le verbe assimilé «المثال » quand il est conjugué dans l'inaccompli «المضارع », il doit subir une normalisation qui consiste à supprimer la voyelle longue.

Exemple : le tableau 19 présente un exemple de suppression de la voyelle longue du verbe assimilé " المثال "

Verbe dans l'accompli	Verbe dans l'inaccompli	Verbe dans l'inaccompli
وجد (il a trouvé)	يوجد (il trouve) (Forme non utilisable)	يجد (il trouve) (Forme utilisable)

Tableau 20 : Exemple de la suppression de la voyelle longue du verbe assimilé « المثال »

- Dans le cas du verbe creux " أجوف " la normalisation consiste à transformer la voyelle longue « ا » à ses origines « و » ou « ي ».

  Exemple : « قال »⟶ « قول »

  « صار » ⟶ « صير »

- La suppression de la voyelle longue du verbe creux " أجوف " , quand il est conjugué dans l'accompli avec la troisième personne féminin plurielle, la deuxième personne (singulier et pluriel). Donc, la normalisation consiste à insérer la voyelle supprimée.

  Exemple : pour le verbe « قال », quad il est conjugué avec la troisième personne féminin plurielle, il s'écrit « قلن » (elles disent), et avec la deuxième personne singulier s'écrit « قلتَ » (Tu dis)

- Dans le cas du verbe faible " ناقص " la normalisation consiste à transformer les voyelles longues « ا » et « ى » à ses origines « و » ou « ي ».

  Exemple : سعى (il démarche) ⟶ سعي

- La suppression de la voyelle longue du verbe faible " ناقص ", quand il est conjugué avec la troisième personne masculin plurielle. Donc, la normalisation consiste à insérer la voyelle supprimée.

  Exemple : محا (il a effacé) ⟶ محوا (ils ont effacé)

Note : Après les normalisations, les verbes défectueux seront traité comme des verbes sains.

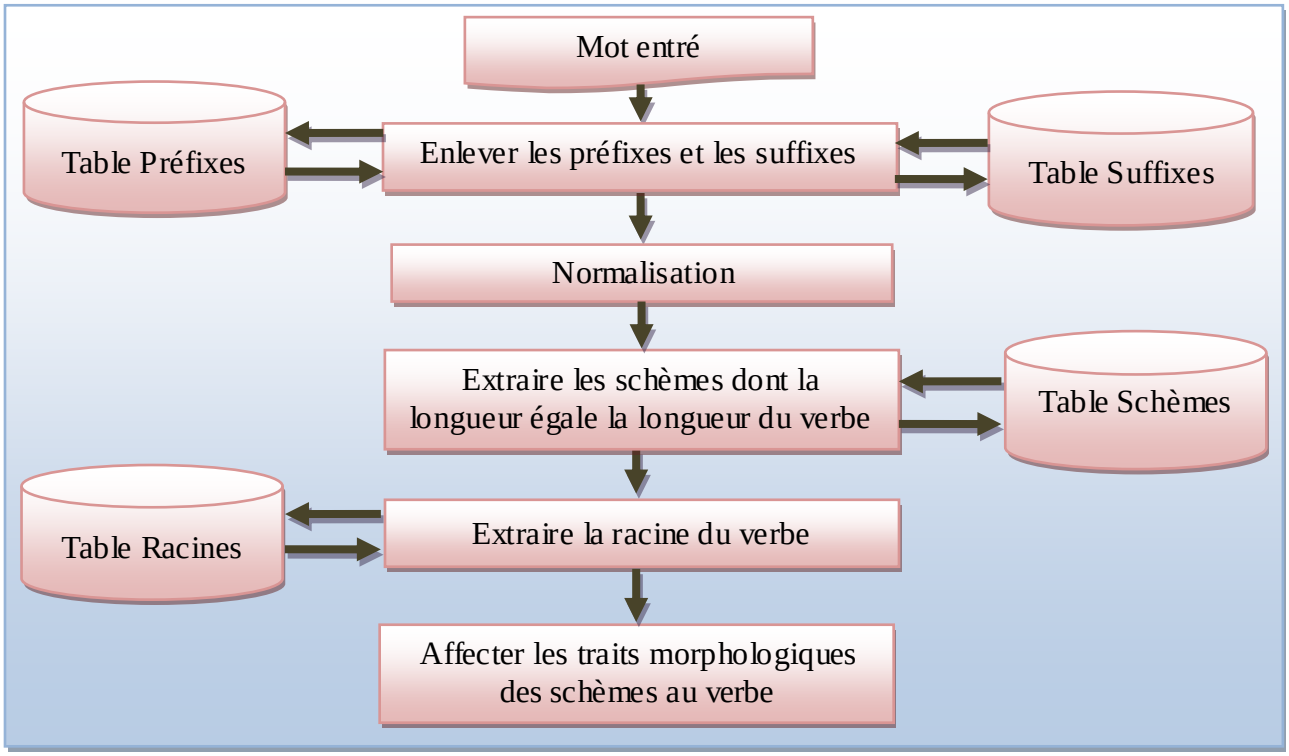


Figure 08 : Organigramme du traitement d'un verbe défectueux

7- Conclusion

La conception reste une phase primordiale dans la réalisation de n'importe quel logiciel. Pour notre système AMTAR, cette phase consiste à représenter les différents modules de traitement, en associant avec chaque module une description de son comportement, et le résultat voulu de chaque module.

Chapitre 4

Application et Résultats

1 Introduction

Dans ce quatrième chapitre nous allons présenter l'enivrement de développement du système AMTAR, en précisant le système d'exploitation et le langage de programmation. En plus nous expliquerons l'interface graphique de notre outil. Et nous finissons par les tests et les résultats obtenus.

2 L'environnement de développement

Dans ce point, nous allons expliquer L'environnement de développement du notre programme. Sachant qu'il est développé sur le plate forme Windows et le langage de programmation python.

2-1 Le système d'exploitation Windows 7

Windows 7 est un système d'exploitation de la société Microsoft, sorti le 22 octobre 2009 et successeur de Windows Vista.

Cette version de Windows reprend l'acquis de Windows Vista tout en apportant de nombreuses modifications, notamment par divers changements au niveau de l'interface et de l'ergonomie générale, un effort particulier pour la gestion transparente des machines mobiles et le souci d'améliorer les

performances globales du système (fluidité, rapidité d'exécution même sur des systèmes moins performants, tels lesnetbooks) par rapport à son prédécesseur.

Cette version contient de nombreux changements dans l'interface, notamment une nouvelle barre des tâches, un menu démarrage amélioré et l'absence du volet Windows. Le volet Windows a certes disparu, mais les gadgets sont toujours là et on peut les placer où l'on veut sur le bureau.

De plus, Windows 7 utilise la mémoire vive de façon bien plus intelligente, en utilisant beaucoup de mémoire s'il y en a, tout en sachant fonctionner correctement sur des configurations n'ayant que 512 Mo de mémoire vive. Windows 7 utilise également moins de données pour démarrer.

2-2 Le langage de programmation

Python est un langage de programmation objet interprété, portable, dynamique, extensible, gratuit, qui permet une approche modulaire. Il est développé depuis 1989 par Guido van Rossum. Il offre un environnement complet de développement comprenant un interpréteur performant et de nombreux modules.

La version du langage de programmation que nous avons utilisé est : Python 2.7.3

2-2-1 Caractéristiques du langage python

Les caractéristiques principales du langage Python sont résumées dan les points suivants [Swinnen, 2000] :

- Python est portable, non seulement sur les différentes variantes d'Unix ou de Windows, mais aussi sur les autre systèmes d'exploitation : MacOS, BeOS, NeXTStep, MS-DOS.
- Python est libre et gratuit même pour les usages commerciaux
- 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.
- Qualité L'utilisation de Python permet de produire facilement du code évolutif et maintenable.
- Python gère ses ressources (mémoire, descripteurs de fichiers...) sans intervention du programmeur, par un mécanisme de comptage de références.
- Python est orienté-objet. Il supporte l'héritage multiple et la surcharge des opérateurs. Dans son modèle objets, et en reprenant la terminologie de C++, toutes les méthodes sont virtuelles.
- Python intègre, comme Java ou les versions récentes de C++, un système d'exceptions, qui permettent de simplifier considérablement la gestion des erreurs.
- Python est dynamique. l'interpréteur peut évaluer des chaînes de caractères représentant des expressions ou des instructions Python.

- Il est orthogonal. un petit nombre de concepts suffit à engendrer des constructions très riches.
- Il est réflectif. il supporte la métaprogrammation, 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.
- Il est introspectif. un grand nombre d'outils de développement, comme le debugger ou le profiler, sont implantés en Python lui-même.
- Python est dynamiquement typé. Tout objet manipulable par le programmeur possède un type bien défini à l'exécution, qui n'a pas besoin d'être déclaré à l'avance.
- Python possède actuellement deux implémentations. L'une, interprétée, dans laquelle les programmes Python sont compilés en instructions portables, puis exécutés par une machine virtuelle (comme pour Java, avec une différence importante: Java étant statiquement typé, il est beaucoup plus facile d'accélérer l'exécution d'un programme Java que d'un programme Python). L'autre génère directement du bytecode Java.
- Python est extensible : comme Tcl ou Guile, on peut facilement l'interfacer avec des bibliothèques C existantes. On peut aussi s'en servir comme d'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, protocoles Internet (Web, News, FTP, CGI, HTML...), persistance et bases de données, interfaces graphiques.

2-2-2 Le choix de python

Nous avons choisi le langage python parce qu'il permet de traiter les chaines de caractères d'une façon simple et efficace, grâce aux fonctions prédéfinis destinés à ce traitement. Ainsi notre choix est dirigé par les caractéristiques de python que nous avons cité précédemment.

2-2-3 Les bibliothèques python utilisés

Pour réaliser notre analyseur, nous avons utilisé les packages suivants :

- **Sqlite** : Ce module permet de faire la gestion des bases de données.
- **Tkinter** : est une librairie basique mais très simple d'utilisation pour construire rapidement des interfaces graphiques avec Python.
- **Pmw** : est une boîte à outils pour la construction des fenêtres composites de haut niveau en Python en utilisant le module Tkinter.

- **Tashaphyne** : package écrit en python destiné pour le traitement automatique de la langue arabe.il est développé par monsieur Taha Zarrouki.

2-2-4 Les fonctions python pour la manipulation des chaines de caractères

- **Len (argument) :** Renvoi la longueur d'une chaîne de caractères.
- **Split (argument) :** Convertit une chaîne de caractères en une liste de sous chaînes.
- **Join (argument) :** Rassemble une liste de chaînes en une seule chaîne.
- **Find (argument) :** Cherche la position d'une sous chaîne dans la chaîne de caractères.
- **Stripe () :** Enlève les espaces éventuels au début et à la fin de la chaîne.
- **CH [:x] :** Enlève les « x-1 » derniers caractères de la chaîne « CH ».
- **CH [x:] :** Enlève les « x » premiers caractères de la chaîne « CH ».

3 Description de l'interface de AMTAR

Dans cette section nous allons représenter l'interface graphique de notre système.

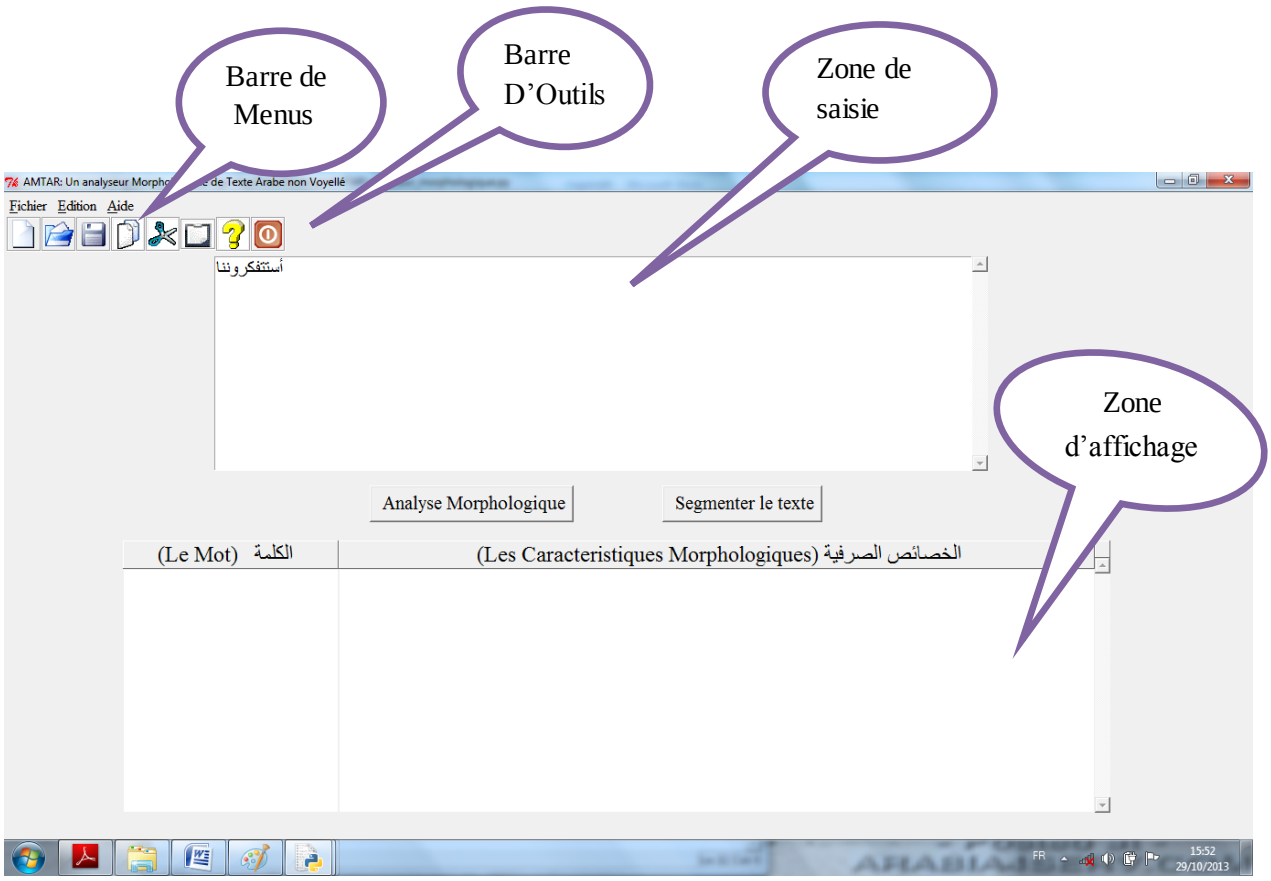


Figure 09 : Interface graphique du système « AMTAR »

La fenêtre principale de notre analyseur est représentée dans la figure 09. Elle contient les éléments suivants :

- Une zone de texte qui permet de saisir un texte ;
- Une barre de menu ;
- Une barre d'outils qui contient des boutons de raccourcis.
- Une zone d'affichage qui permet d'afficher les résultats de traitement.
- Deux boutons de traitement :
  - Analyse morphologique ;
  - Segmenter le du texte.

3-1 Barre de menu principal et barre de boutons

La barre de menus principaux est composée de trois (03) POPUP (voir la figure 10):

- Fichier;
- Edition ;
- Aide.

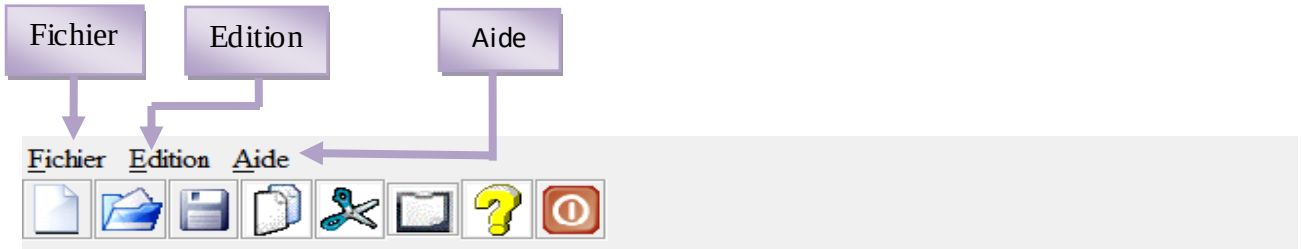


Figure 10 : Barre de menu principal du système « AMTAR »

3-1-1 Pop-up Fichier

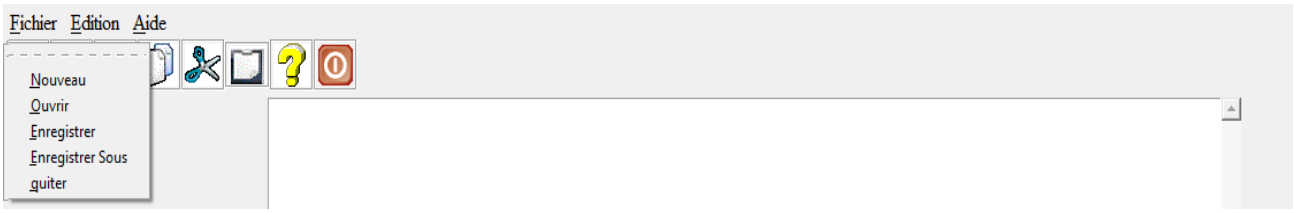


Figure 11 : Représentation du pop-up Fichier.

Item	Description
Nouveau	Créer une nouvelle zone de texte.
Ouvrir	Ouvrir un fichier texte qui contient des données textuelles.
Enregistrer	Enregistrer les données et les résultats dans un nouveau fichier texte.
Enregistrer sous	Enregistrer les données et les résultats dans un fichier texte qui existe déja.
Quitter	Sortir de l'application.

Tableau 21 : Description des items que contient le pop-up Fichier.

3-1-2 Pop-up Edition

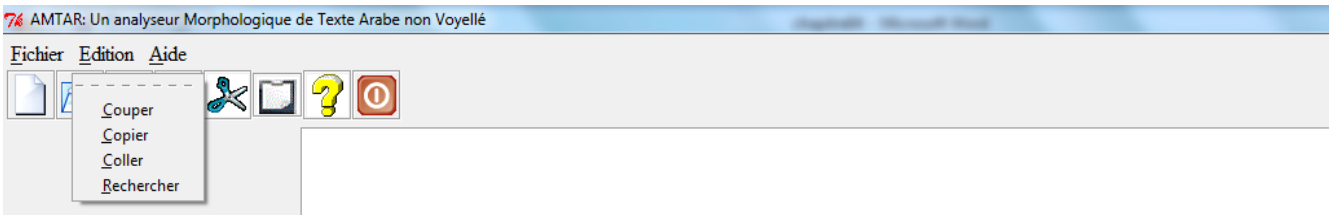


Figure 12 : Représentation du pop-up Edition.

Item	Description
Couper	Enlever la partie sélectionnée.
Copier	Dupliquer la partie sélectionnée.
Coller	Intégrer la partie sélectionnée
Rechercher	Trouver la chaîne de caractères voulue.
Remplacer	Remplacer une chaîne de caractères par une autre.

Tableau 22 : Description des Items que contient le pop-up Edition.

3-1-3 Pop-up Aide

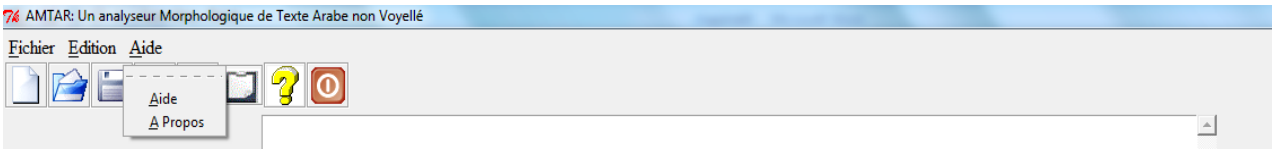


Figure 13 : Représentation du pop-up Aide.

Item	Description
Aide	Cet item propose de outils d'aide à l'utilisateur.
A propos	Cet item contient un des informations concernant le système AMTAR.

Tableau 23 : Description des Items que contient le pop-up Aide.

3-2 La barre d'outils

La barre d'outils contient des boutons raccourcis (voir figure 14). Elle est utilisée pour raccourcir les items du menu principal, et cela par un simple clic sur un bouton.

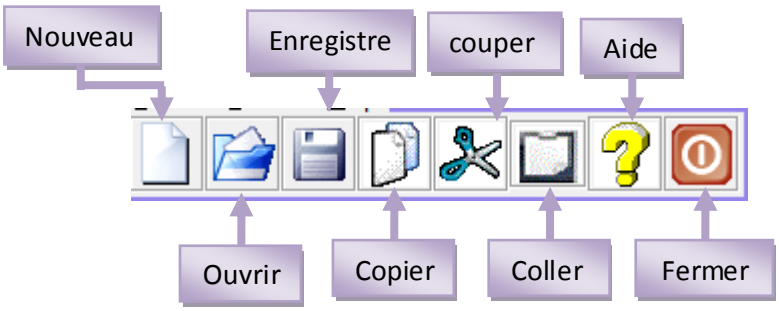


Figure 14 : Barre d'outils (boutons des raccourcis) du système « AMTAR »

3-3 Boutons de traitement

Les traitements faits par l'application sont basé sur deux boutons (Voire la Figure 15et le Tableau 24)

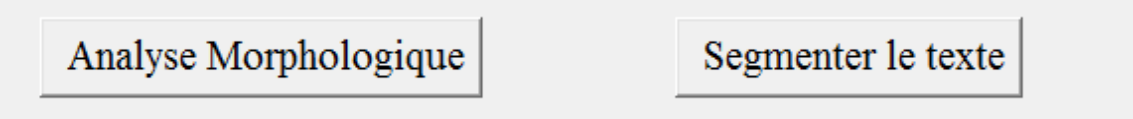


Figure 15 : Représentation des boutons de traitement.

Bouton	Fonction
Analyse Morphologique	Attribuer à chaque unité lexicale reconnue un ensemble des traits morphologiques possibles.
Segmentation du texte	Découper un texte en un ensemble des unités lexicales.

Tableau 24 : Description des boutons de traitement.

4 Présentation du corpus de test

Le système AMTAR a été testé sur le corpus « khaleej », préparé par Abbas Mourad, pour réaliser des expériences sur l'identification du sujet pour la langue arabe. Il a été extrait de milliers d'articles qui ont été téléchargés à partir d'un journal en ligne.

Ce corpus contient plus de 5000 articles qui correspondent à près de 3 millions de mots, sur quatre catégories de sujets (voir le tableau 25).

sujet	Taille du corpus (Nombre des documents)
Economie	909
International News	953
Local News	2398
Sports	1430
Nombre total des documents	5690

Tableau 25 : description du corpus « Khaleej »

5 Tests et résultats

Les résultats de test du système AMTAR sont représentés dans les tableaux et les figures suivants, selon les catégories du corpus de test tel que :

- Reconnu : Représente le nombre des mots reconnus par l'analyseur.
- Non Reconnu : Représente le nombre des mots non reconnus par l'analyseur.

	Economie		
	Reconnus	Non Reconnus	Total
Nombre des mots	23155	1944	25099
Pourcentage	92,25%	7,75%	100%

Tableau 26 : Résultat d'analyse des textes de la catégorie « Economie »

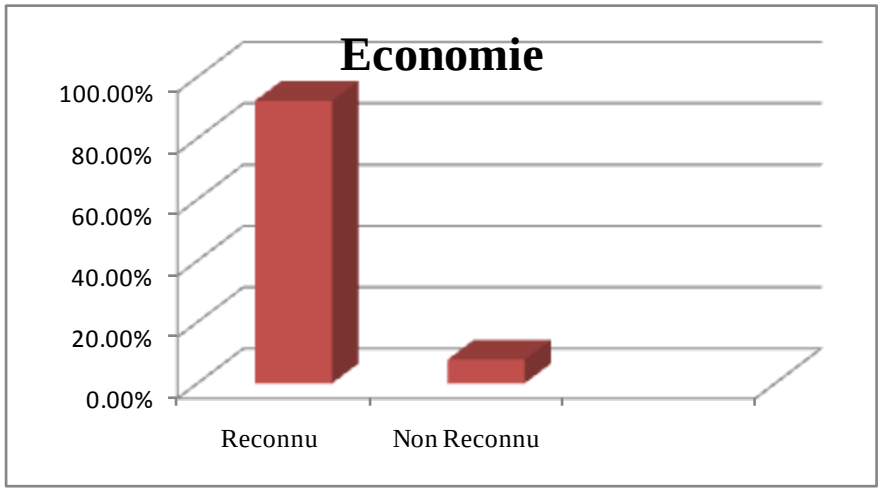


Figure 16 : Graphe des résultats de la catégorie « Economie »

	International news		
	Reconnus	Non Reconnus	Total
Nombre des mots	28302	2634	30936
Pourcentage	91,49%	8,51%	100%

Tableau 27 : Résultat d'analyse des textes de la catégorie « International news »

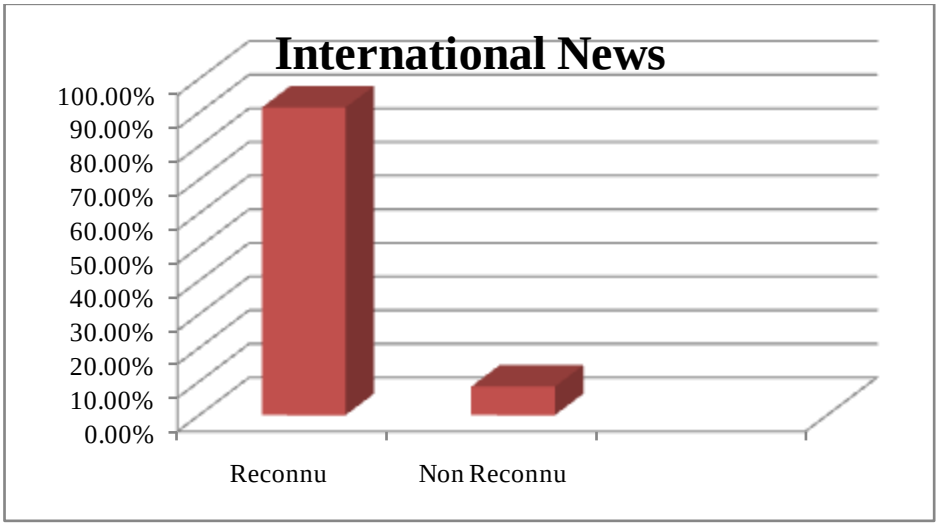


Figure 17 : Graphe des résultats de la catégorie « International News »

	local news		
	Reconnus	Non Reconnus	Total
Nombre des mots	42572	3828	46400
Pourcentage	91.75%	8.25%	100%

Tableau 28 : Résultat d'analyse des textes de la catégorie « local news »

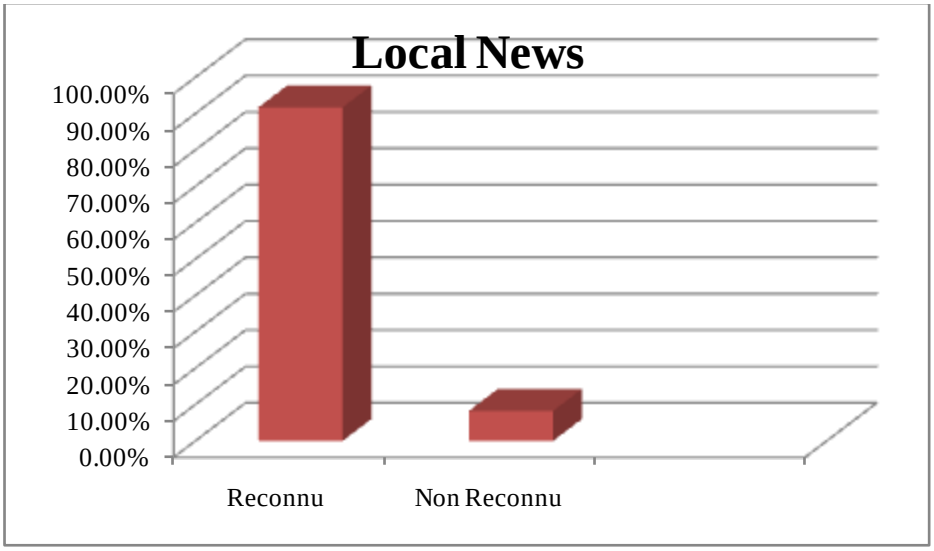


Figure 18 : Graphe des résultats de la catégorie « Local News »

	Sport		
	Reconnus	Non Reconnus	Total
Nombre des mots	67682	6266	73948
Pourcentage	91.53%	8.47%	100%

Tableau 29 : Résultat d'analyse des textes de la catégorie « Sport »

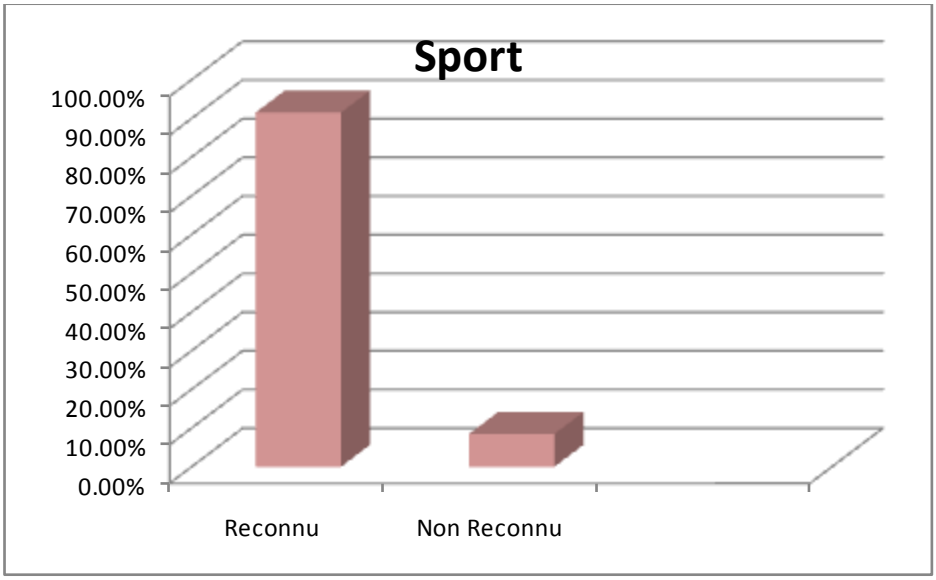


Figure 19 : Graphe des résultats de la catégorie « Sport »

6 Discussion

A partir des résultats de test obtenus, et par des calculs simples, on peut déterminer la moyenne des pourcentages des mots reconnus égale à 92%. C'est un bon chiffre, mais il nous amène à poser la question, pourquoi 8% des mots sont non reconnus ?

La réponse est résumée dans les points suivants :

- Les textes collectés dans le corpus du test comportent des noms et des abréviations non arabes.
- Il y a des mots arabes qui admettent plusieurs possibilités de la segmentation ce qui fait un sorte d'ambiguïté pour le module de segmentation qui peut donner des segmentations fausse.
- La base de données de l'analyseur nécessite une mise à jour pour atteindre le maximum des ressources linguistiques.

7 Conclusion

Le système « AMTAR » est un analyseur morphologique de bonne qualité. Cette proposition est justifiée par le pourcentage des mots reconnus exprimé dans les résultats des tests. Mais il admet des améliorations qui peuvent lui faire atteindre un pourcentage plus grand.

Conclusion Générale (Bilan et perspectives)

1 Bilan

Le traitement automatique de la langue arabe est devenu inévitable, vu de la grande masse des informations en arabe diffusées sur l'internet et dans les médias.

La conception et la réalisation d'un outil d'analyse du texte arabe non voyellé reste un engagement pour nous les arabes, car la pluparts de nos écritures sont non voyellés. Dans ce mémoire, nous avons décrit notre système AMTAR (Analyseur Morphologique du Texte Arabe), et nous arrivons à conclure que l'informatisation de la langue arabe nécessite la compréhension des caractéristiques de cette langue, et l'élaboration des modèles propre de l'arabe.

En plus, nous avons présenté les détailles de la conception et la réalisation de notre système AMTAR, et les résultats des tests de ce système, ces derniers son fait sur un corpus nommé « Khaleej ».

Notre système AMTAR analyse des textes arabe non voyellés, et donne des bons résultats exprimé en arabe et en français. A parti de ces résultats, nous avons conclu que notre système est fiable performant.

2 Perspectives

L'analyse des textes arabes par le système « AMTAR » arrivent à un pourcentage moyenne égal 92% ; c'est un chiffre bon par rapport aux autres analyseurs trouvés dans le monde, tel que Alkhalil, mais « AMTAR » admet des améliorations qui peuvent être citées comme des perspective, a savoir :

- Enrichir la base de donnés par autres ressources linguistiques et surtout la table des mots spéciaux.
- Ajouter une description des résultats, en anglais.
- Elaborer des applications qui se basent sur l'analyseur AMTAR, tel qu'un analyseur syntaxique ou un système de recherche de l'information…etc.
- Pousser le traitement des verbes faibles.

Bibliographie

[Alrahabi et Dichy, 2009] Alrahabi M. et Dichy J., « Levée d'ambigüité par la méthode d'exploration contextuelle: la séquence 'alif-nûn (ان) en arabe », in Ghenima, Malek, Ouksel, Aris et Sidhom, Sahbi (eds.), Systèmes d'Information et Intelligence Economique, 2ème conférence Internationale (SIIE 2009), organisée par l'université de Nancy, France et l'université de la Manouba, École supérieure de commerce électronique (ESCE), Tunis, Tunis, Hammamet, 12-14 février 2009, IHE éditions, p. 573-585

[Baccouche, 2009] Baccouche T. « Dynamique de la langue arabe » Synergies Tunisie n°1-2009 pp.17-24

[Baloul.2003] Baloul S. « développement d'un système automatique de synthèse de la parole à partir du texte arabe standard voyellé » thèse université de Maine France, 2003

[Belgacem et al, 2007] Belgacem M., Mars M., Antoniadis G., et Zrigui M., « Analyseur morphologique pour l'arabe » CITALA2007, 18-19 juin 2007, Rabat, Maroc, 2007.

[Bernhard, 2006] Bernhard D., « Apprentissage de connaissances morphologiques pour l'acquisition automatique de ressources lexicales », thèse, Universit´e Joseph Fourier – Grenoble I, 2006

[Bessou et al, XXXX] Bessou S., Louail M., Refoufi A., Kadem Z.& Touahria M. « Un système de lemmatisation pour les applications de TALN » Université de Ferhat Abbess, Algérie.

[Boulaknadel, 2008] Boulaknadel S. « TAL et R.I en langue arabe dans un domaine de spécialité : Apport des connaissances morphologiques et syntaxiques pour l'indexation »thèse, niversité de Nantes, 2008

[Bouillon ,1998] Bouillon P. , « Traitement automatique des langues naturelles », édition de Boeck Larcier s.a.,1998

[Chacha, 2008] فارس شاشة, « المعالجة الآلية للغة العربية :إنشاء نموذج لساني صرفي إعرابي للفعل العربي » , مذكرة لنيل شهادة الماجستير في علم المكتبات و التوثيق، 2008. جامعة الجزائر

[Cheragui, 2010] Cheragui M. A. « Un modèle d'analyse multicritère de la levée de l'ambiguïté associé à un Tagger pour le Traitement Automatique de l'arabe » Mémoire de Magistère, Ecole nationale Supérieure en Informatique Oued Smar Alger, 2010.

[Cori et , 2002] Cori M., Léon J., « La constitution du TAL : Étude historique des dénominations et des concepts », TAL. Volume 43 – n° 3/2002, pages 21 à 55

[Douzidia, 2004] Douzidia F.S. « Résumé automatique de texte arabe », Mémoire de Master, Université de Montréal, 2004.

[Ghassan, 2002] Ghassan M. « La segmentation de textes par exploration contextuelle automatique, présentation du module SegATex » Inscription Spatiale du Langage : structure et processus ISLsp. 29 & 30 janvier 2002. IRIT, Université Paul Sabatier, Toulouse, 2002.

[Gridach et Chenfour, 2011] Gridach M., Chenfour N. « Developing a New System for Arabic Morphological Analysis and Generation » article,Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), IJCNLP 2011, , pages 52–57, Chiang Mai, Thailand, November 8, 2011.

[Khelif, 2006] Khelif M. K. « Web sémantique et mémoire d'expériences pour l'analyse du transcriptome » thèse, université de Nice-Sophia Antipolis - UFR Sciences,2006

[Khemakhem, 2006] Khemakhem A. « ArabicLDB : une base lexicale normalisée pour la langue arabe » mémoire, Université de Sfax, 2006.

[Lison, 2006] Lison P. « Implémentation d'une Interface Sémantique-Syntaxe basée sur des Grammaires d'Unification Polarisées » mémoire Université catholique de Louvain, 2006.

[Mesfar, 2008] Mesfar S. « analyse morpho-syntaxique automatique et reconnaissance des entités nommées en arabe standard, thèse, université de FRANCHE-COMTE,2008.

[Mouelhi, 2008] Mouelhi Z. « AraSeg : un segmenteur semi-automatique des textes arabes » article JADT 2008 : 9es Journées internationales d'Analyse statistique des Données Textuelles, 2008.

[Radjihi , 1973] .1973 الدكتور عبده الراجحي، « التطبيق الصرفي » , دار النهضة العربية للطباعة والنشر

[Ramzi, 2004] Ramzi A. « la conception et la réalisation d'un concordancier électronique pour l'arabe » thèse, institut national des sciences appliquées de Lyon, 2004.

[Sawalha, 2011] Sawalha M. S. S., « Open-source Resources and Standards for Arabic Word Structure Analysis:Fine Grained Morphological Analysis of Arabic Text Corpora » these, The University of Leeds School of Computing, UK, 2011.

[Sonbol et al, 2011] Sonbol R., Ghneim N., et Desouki M. S. « An Application Oriented Arabic Morphological Analyzer »article, Journal de l'université du Damascus Vol. (27) - No. (1) 2011

[Swinnen, 2000] Swinnen G., « Apprendre à programmer avec Python », livre électronique du Licence de Documentation Libre, 2000

[Tahir et al, 2004] Tahir Y., Chenfour N., et Harti M. « Modélisation à objets d'une base de données

morphologique pour la langue arabe » article, JEP-TALN 2004, Traitement Automatique de l'Arabe, Fès, 20 avril 2004

[Yvon, 2007] Yvon F., « Une petite introduction au traitement Automatique du langage naturel, support de cours », Ecole Nationale Supérieur des télécommunications, Avril 2007.