République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université de Djilali BOUNAAMA KHEMIS MILIANA

Faculté des Sciences et de la Technologie

Département des Mathématique et Informatique

Mémoire Présenté

Pour l'obtention de diplôme

Master en « Informatique »

Option : « Ingénierie du logiciel »

Titre :

Détection Automatique De La Cohésion Lexicale Entre Phrases Dans Les Textes Arabes

Réalisé par :

M^elle^ CHOUCHAOUI Maissa

M^elle^ BRAHIMIA Yamna Affaf

Soutenue publiquement le : 01/06/2016

Devant le jury composé de :

Mr A.KHALFI ……..…Président

Mr D.BAHLOUL…..**.....**Encadreur

Mr SHADJ SADOK……..Examinateur1

M^me^ A.LABABOU…….....Examinateur2

Année Universitaire : 2015/2016

Remerciement

Nous tenons d'abord à remercier ALLAH le tout puissant de nous avoir donné la volonté, l'amour du savoir et surtout le courage et la patience pour effectuer ce modeste travail.

Il nous tient à cœur d'exprimer toute notre reconnaissance a ceux qui au long de notre travail nous ont apporté leurs aides, leurs conseils, et leurs encouragements.

Nos sincères remerciements à Mr BAHLOUL Djemel ...maître de conférence à l'université Khemis Miliana pour nous avoir encadrées et conseiller au cours de notre travail.

Sans oublier l'ensemble des enseignants et personnel l'administratifs du département de mathématique et informatique de l'université de Khemis Miliana _ Ain Defla _ qui nous ont apportées leur soutien de près ou loin.

En fin, nous remercions nos amies pour leurs encouragements et leurs soutiens et tous ceux et celles qui de près ou de loin ont contribué à la réalisation de ce travail.

Dédicaces maissa

Avec l'aide de Dieu le tout puissant que j'ai pu arriver au terme de ce travail que Je tiens à dédier à :

Mes très chers parents qui m'ont entouré de leur affection et qui m'ont toujours orienté avec leurs

Conseils judicieux. Que Dieu les gardes pour moi et les récompenser pour leur sacrifices.

A ma chère mère, la plus merveilleuse femme, je suis très reconnaissante à ses divers sacrifices,

Son soutien, et ses encouragements.

A ma sœur Meriem et à mes frères : Abd Djalil et Mohamed Yasser

A toute ma famille paternelle et maternelle

A CHOUCHAOUI Djamel

A mes amis : Imen, Yassmina, Meriem, Sbaa et mes amis de FB

A ma copine avec qui j'ai préparé ce Modest travail Affaf

Atous les étudiants da la promotion (2015/2016) de Informatique master Ingénierie du logiciel

A tous ceux ont attribué de près ou de loin à l'élaboration de ce Modes travail

Dédicace affaf

A la bougie de ma vie ma Mère

A mon père qui a toujours avec moi

Mes deux sœurs manel et marwa

A mon frère imad

A tout la famille BRAHMIA et la famille FADEL

A tout mes professeurs

A tout mes amis

Je dédié ce travail.

Sommaire

Liste des figures

La liste des tableaux

Liste des équations

Résumé

Détecter la cohésion lexicale entre des paires de phrases consiste à repérer les différentes relations lexicales et sémantiques. L'approche que nous proposons à travers ce travail consiste à comparer globalement les phrases d'un texte écrit en langue arabe en se basant sur une combinaison de critères distributionnels répétitifs et d'autres sémantiques basés essentiellement sur les relations de synonymie afin de calculer la similarité entre les paires de phrases.

Notre travail s'inscrit dans le cadre des travaux relatifs au traitement automatique de la langue arabe et particulièrement l'étude des relations entre phrases dans les textes arabes. Le système que nous avons développé peut être considéré comme un prétraitement nécessaire pour les applications d'aide à la lecture comme la segmentation thématique ou le résumé automatique.

Mots clés :

Traitement automatique de la langue arabe, Similarité distributionnelle, Synonymie, Arabic stemmer, Arabic Wordnet

الملخص

إن عملية تحديد وكشف التناسق المعجمي بين الجمل في النصوص العربية يتمثل في تحديد العلاقات المعجمية والدلالية المختلفة بينها. المنهجية المعتمدة من خلال هذا العمل هي إجراء مقارنة شاملة لجميع جمل النص وذلك بالاعتماد على طريقة تجمع بين معايير توزيعية تكرارية للكلمات وأخرى دلالية تقوم أساسا على علاقات الترادف من أجل حساب نسبة التشابه بين أزواج من الجمل.

يعتبر عملنا هذا جزءا من الأعمال المندرجة تحت المعالجة الآلية للغة العربية ولا سيما دراسة العلاقات بين الجمل في النصوص العربية.

نستطيع القول بأن النظام الذي أنجزناه من خلال هذا العمل يعتبر بمثابة تطبيق لازم ومهم لتطبيقات المعالجة الآلية الأخرى كالتجزئة الموضوعية والتلخيص الآلي للنصوص.

الكلمات المفتاحية :

المعالجة الآلية للغة العربية، التشابه التوزيعي، الترادف، أرابيك ستيمير، أرابيك وورد نيت.

Abstract

Detecting lexical cohesion between pairs of sentences is to identify the various lexical and semantic relations between them. The approach we propose through this work is to comprehensively compare the sentences of a text written in Arabic, the approach is based on a combination of repetitive distributional criteria and other semantic essentially based on synonymous relations where the object is to calculate the similarity between the pairs of sentences.

Our work is part of the work regarding the automatic processing of the Arabic language and particularly the study of relationships between sentences in Arabic texts. The system we have developed can be considered a pretreatment required for playback using applications such as thematic segmentation or automatic summarization.

Key words:

Arabic language processing, distributional similarity, Synonymy, Arabic stemmer, Arabic Wordnet

Introduction générale

Ces dernières années sont marquées par une augmentation énorme de la quantité d'information électronique rédigée en arabe et dont l'accès et la lecture des textes est devenu de plus en plus complexe et le besoin de développer des applications d'aide à la lecture est devenu incontournable.

Notre projet s'inscrit dans le cadre des travaux relatifs au traitement automatique de la langue arabe et particulièrement la segmentation thématique de textes arabes en sous thèmes exprimant les mêmes idées. L'objectif est de réaliser un système de détection automatique de la cohésion lexicale entre les différentes phrases d'un texte. Le travail est basé sur la recherche automatique des relations définissant une cohésion lexicale (mots liés), cette dernière peut être détectée en calculant par exemple des similarités distributionnelles répétitives ou de synonymie entre les mots des différentes phrases du texte.

En effet, la détection automatique de la cohésion lexicale entre les phrases d'un texte est une tâche très importante pour les systèmes d'aide à la lecture, le résumé automatique et la traduction automatique…

La mise en place du système d'identification nécessite l'utilisation de quelques outils de traitement automatique de la langue arabe tel que les analyseurs morphologiques et les bases de données lexicales.

La conception et réalisation des approches capables de traiter automatiquement la langue arabe et particulièrement l'étude des relations entre les phrases d'un texte écrit en langue arabe est très faiblement abordée par les informaticiens, alors que le besoin de développer de telles applications est devenu nécessaire par rapport aux attentes de la communauté arabe. Après finalisation, notre système constituera une bonne contribution dans ce sens.

Notre travail est organisé de la manière suivante : dans le premier chapitre nous introduisons une définition terminologique du traitement automatique des langues. Dans le deuxième chapitre, nous discutons le traitement automatique de la langue arabe et les problèmes liés ainsi que les différents outils disponibles. Dans le troisième chapitre, nous présentons les différentes techniques utilisées pour détecter automatiquement les relations de cohésion entre phrases. Le quatrième chapitre est consacré à la conception et la réalisation de notre système.

Chapitre I :

Traitement automatique des langues

I.1 Introduction :

Le Traitements Automatique des Langues (TAL) est une discipline qui associe étroitement linguistes et informaticiens. Il repose sur la **linguistique**, les **formalismes** (représentation de l'information et des connaissances dans des formats interprétables par des machines) et l'**informatique**.[1] Le but du TAL est réellement de "comprendre" le sens des phrases, les idées qui s'en dégagent et ce de manière à pouvoir "traiter" de la manière la plus optimale et la plus naturelle d'un point de vue humain ces phrases. [1]

Dans ce chapitre, nous présentons le traitement automatique des langues, leur niveaux de traitement et les problèmes , l'architecture d'un système de TAL, et les applications du TAL.

I.2 Bref historique du TAL :

L'année	Le traitement
Années 50	Traduction automatique – débuts du TAL
1964	Rapport ALPAC
Années 60	. Linguistique formelle (Chomsky, Montague) comme base pour le TAL. .Applications basées sur des techniques linguistiques (Eliza, shrdlu) – Chomsky (grammaires formelles, analyseurs syntaxiques); sémantique procédural (Woods) . Approches limitées à des domaines restreints. Non portables.
Années 70	Premières applications
Années 80	. Approches symboliques. Applications utilisent des connaissances linguistiques et encyclopédiques extensives. Manquent de robustesse.
Années 90 et plus	Premiers corpus, approches statistiques, apprentissage automatique. Applications utilisent corpus de grande taille et méthodes statistiques.

Tableau I.1 : Bref historique du TAL [2]

[1] http://www.technolangue.net/imprimer.php3?id_article=274

I.3 Les niveaux de traitement en TAL :

Nous introduisons dans cette section les différents niveaux de traitement nécessaires pour parvenir à une compréhension complète d'un énoncé en langage naturel. Chacun de ces niveaux suppose un traitement très spécifique permettant l'idée d'un système modulaire en traitement automatique.

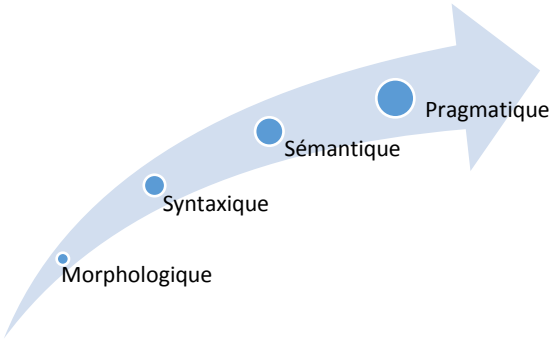


Figure I.1: Les niveaux de TAL

I.3.1 Niveau morpho lexical (morphologique) :

Dans ce niveau, le programme TAL étudie la formation des mots et leur variation de formes. Deux termes sont à étudier dans ce niveau, la flexion et la dérivation :

-La flexion

C'est l'ensemble de modifications que subit un mot dans sa terminaison. Selon le rôle qu'elle joue dans la phrase, la flexion est un processus qui consiste à modifier le radical d'un mot pour lui adjoindre à certains types d'éléments. [3]

-La dérivation

C'est un procédé de formation de mots nouveaux par addition, suppression ou remplacement d'un élément grammatical d'un mot simple. [3]

I.3.2 Niveau syntaxique :

On s'intéresse à ce niveau à l'agencement des mots et leurs relations structurelles. Les connaissances syntaxiques concernent la façon dont les mots sont agencés dans une phrase, c'est-à-dire, sa structure grammaticale. Ces connaissances ou règles sont décrites dans des grammaires, leur application permet la formation des phrases correctes et de lever les ambiguïtés.

I.3.3 Niveau sémantique :

La sémantique est une discipline qui à pour objectif la description des significations propres aux langues et leurs organisations théoriques. En TALN, la sémantique peut être définie comme l'étude de sens des mots, des phrases et des énoncés.

Le rôle de l'analyseur sémantique est donc d'attribuer un sens à la phrase structurée par l'analyseur syntaxique.

Les connaissances sémantiques nécessaires pour donner un sens aux noms sont seules qui explicitent non seulement la relation entres les mots et les objets, actions ou idées qu'ils désignent mais aussi les conditions qui permettent d'évaluer si une phrase à un sens ou non.

I.3.4 Niveau 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.

I.4 Les problèmes majeurs de TAL :

Les difficultés que l'on rencontre en TAL sont principalement de deux ordres, et ressortent soit de l'ambiguïté du langage, soit de la quantité d'implicite contenue dans les communications naturelles.

I.4.1 L'Ambiguïté :

Le langage naturel est ambigu, et ce à quelque niveau qu'on l'appréhende. Cette ambiguïté, loin d'être marginale, est un de ses traits caractéristiques. On peut d'ailleurs voir là le résultat d'un compromis inévitable entre d'un côté une capacité d'expression quasi illimitée, et de l'autre des contraintes liées à la limitation des ressources physiologiques mises en œuvre (taille de la mémoire à long et court-terme, densité de l'espace phonétique, contraintes articulatoires, etc). [4]

Cette ambiguïté se manifeste par la multitude d'interprétations possibles pour chacune des entités linguistiques pertinentes pour un niveau de traitement, comme en témoignent les exemples suivants :

– ambiguïté des graphèmes (lettres) dans le processus d'encodage orthographique : comparez la prononciation du i dans lit, poire, maison;
– ambiguïté dans les propriétés grammaticales et sémantiques (i.e. associées à son sens) d'une forme graphique donnée : ainsi manges est ambigu à la fois morpho-syntaxiquement, puisqu'il correspond aux formes indicative et subjonctive du verbe manger), mais aussi sémantiquement.
– ambiguïté de la fonction grammaticale des groupes de mots.
– ambiguïté de la portée des quantificateurs, des conjonctions, des prépositions.
– ambiguïté sur l'interprétation à donner en contexte à un énoncé. [4]

I.4.2 L'Implicite :

L'activité langagière s'inscrit toujours dans un contexte d'interaction entre deux humains, sensément dotés d'une connaissance du monde et de son fonctionnement telle que l'immense majorité des éléments de contexte nécessaires à la désambiguïsation mais aussi à la compréhension d'un énoncé naturel peuvent rester implicites. La situation change du tout au tout dès qu'une machine tente de s'insérer dans un processus de communication naturel avec un humain : la machine ne dispose pas de cette connaissance d'arrière-plan, ce qui rend la compréhension complète de la majorité des énoncés difficile, voire impossible, si l'on ne dispose pas de bases de connaissances additionnelles, donnant accès à la fois à un savoir sur le monde (ou le domaine) en général (connaissance statique) et sur le contexte de l'énonciation (connaissance dynamique). [4]

I.5 Architecture d'un système TAL :

De façon générale, les systèmes TAL peuvent se ramener à deux types différents d'architectures [5] :

I.5.1 Architecture startificationnelle (Appelée aussi séquentielle ou en série) :

C'est l'architecture la plus ancienne, elle utilise les connaissances linguistiques les unes après les autres. Dans un système pareil, il existe pour chaque type de connaissances un

module informatique différent qui prend en entrée les données qui lui sont fournies par le module précédent, sans aucune rétroaction ne soit possible.

La figure ci-dessous, montre un exemple d'une architecture startificationnelle d'un système qui à pour objet l'extraction de la représentation pragmatique d'un texte écrit :

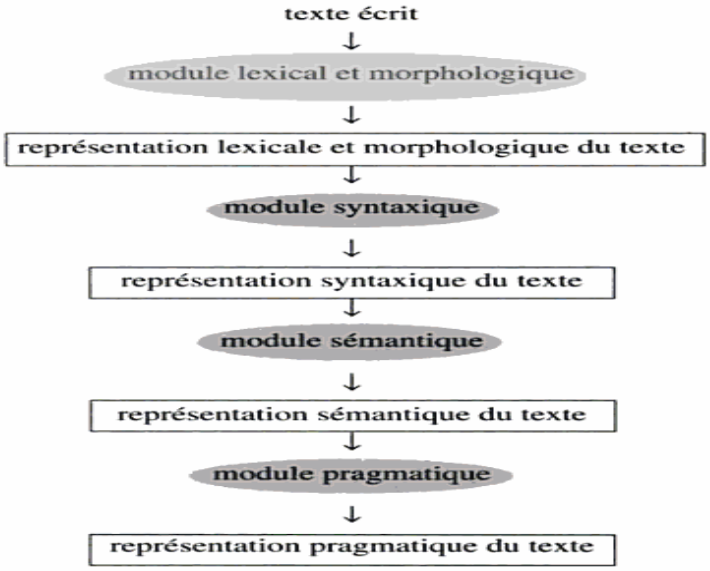


Figure I.2 : Architecture startificationnelle d'un système TALN [5]

I.5.2 Architecture moins hiérarchisée (Appelée aussi systèmes hétérarchiques, parallèles ou intégrés) :

Ce type d'architecture utilise les différentes connaissances en même temps, soit en intégrant les différentes connaissances au niveau du mot dans une représentation complexe, soit en faisant communiquer les différents modules.

L'avantage de ces architectures est évident, en intégrant les différentes connaissances, elles évitent les ambiguïtés artificielles, propres aux machines, et celles qui apparaissent lorsque les connaissances sont considérées les unes après les autres :

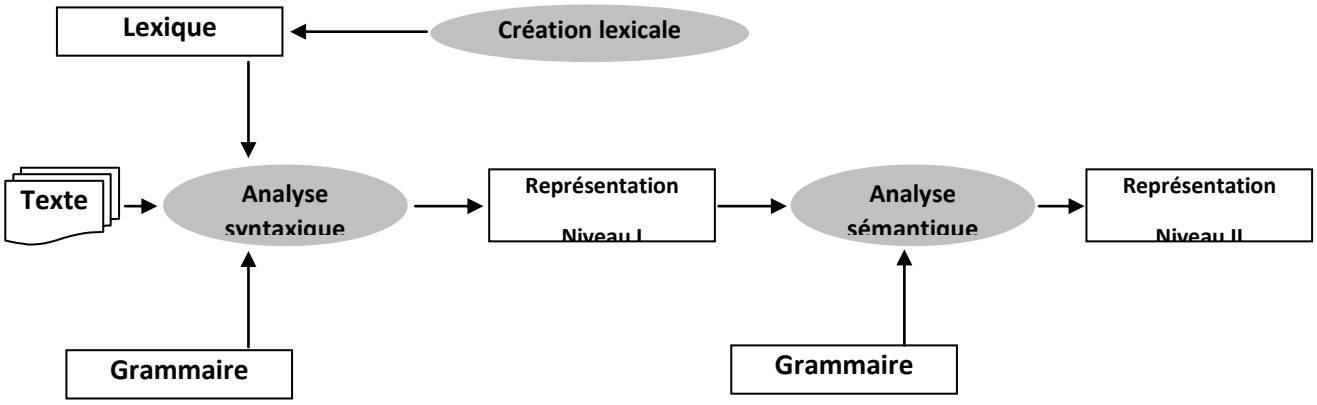


Figure I.3 : Architecture d'un système TALN intégré [5]

I.6 Les applications de TAL :

Concernant les applications, la demande de TAL provient, pour dire vite, de deux tendances « lourdes » : d'une part la nécessité de concevoir des interfaces de plus en plus ergonomiques, d'autre part la nécessité de pouvoir traiter (produire, lire, rechercher, classer, analyser, traduire) de manière de plus en plus « intelligente » les informations disponibles sous forme textuelle, de manière à pouvoir résister à leur prolifération exponentielle. Les applications des techniques de TAL sont donc nombreuses et variées. Nous avons regroupé ces applications en trois grandes familles, qui correspondent aux aides à la lecture de documents, aux aides à la production de documents, et enfin aux interfaces homme-machines [4].

I.6.1 Le traitement documentaire :

Les applications les plus immédiates du TAL sont celles qui visent à faciliter le traitement par l'humain des immenses ressources disponibles en langage naturel, comme par exemple :

I.6.1.1 La traduction automatique :

Cette application, qui a historiquement suscités les premiers efforts de recherche en TAL, reste un enjeu économique et politique de première importance. Si de tels traducteurs existaient, il serait sans doute beaucoup moins crucial de recourir, pour assurer une large diffusion à des documents, à une langue véhiculaire « universelle » telle que l'anglais ; Il est important de noter que même si la traduction complète indépendante du domaine est encore

hors d'atteinte, on peut obtenir de bons traducteurs spécialisés (domaines techniques), qui constituent un moyen efficace de préparer l'intervention manuelle d'un traducteur. Il existe également des environnements de travail fournissant des ressources lexicales (dictionnaires bilingues étendus) pour l'aide à la traduction. Enfin, il existe une application intéressante des systèmes complètement automatiques de traduction, qui se fonde sur le fait que la recherche documentaire, ou le filtrage manuel de document, qui peut nécessiter une lecture diagonale d'un grand nombre de documents, se fera toujours mieux dans la langue natale de l'utilisateur... [2]

De nombreux systèmes commerciaux sont aujourd'hui disponibles (Systran, Logos, Metal, Aleth-Trad...), et certains moteurs de recherche proposent une traduction automatique des pages html. [2]

I.6.1.2 La recherche documentaire :

La recherche de documents « intéressants » dans des bases documentaires. La prolifération des outils de recherche documentaire sur la toile, qui traitent quotidiennement des millions de requêtes, montrent bien l'importance de la demande en la matière. Les performances de ces moteurs témoignent du chemin qu'il reste à parcourir dans ce domaine. Si Google semble aujourd'hui sortir du lot, d'autres moteurs de recherche valent certainement la peine d'être connus [2].

De plus en plus d'outils fournissent également des outils de recherche spontanée d'adresses potentiellement intéressantes (à partir de profil utilisateurs), ou encore de surveillance automatique des publications dans des domaines donnés. [2]

I.6.1.3 Le Résumé automatique :

Le résumé automatique est une transformation réductrice d'un texte source vers un résumé par compression du contenu à l'aide d'une sélection et/ou généralisation de ce qui est important dans le texte source.[3]

Le Résumé Automatique de Texte semble être une bonne solution qui se trouve à la croisée de deux disciplines : traitement automatique de la langue (TAL) et recherche d'information (RI). Le Résumé de Texte Automatique consiste à produire une représentation courte d'un texte tout en conservant l'information pertinente. [6]

Pour élaborer des systèmes de résumé de texte automatique, la plupart des chercheurs se sont basés sur des systèmes à base de connaissances linguistiques. Ces systèmes utilisent essentiellement des techniques d'extraction dont le principe est de faire ressortir l'information pertinente par la sélection des phrases qui la caractérisent. [6]

Les techniques d′extraction s'appuient sur :

- La combinaison des mots du titre du texte en relation avec leur présence dans le texte source.
- L'analyse thématique du discours et de sa structure.
- La construction de relations de cohésion lexicale entre phrases de sorte à extraire celles qui sont le plus liées.
- L'utilisation de certains marqueurs représentant les relations rhétoriques comme la justification, la cause, la consécution, le contraste, la conséquence ,….[6]

I.6.1.4 La Reconnaissance de caractères :

Le principe général d'un système de reconnaissance optique de caractères est le suivant :

- Numérisation de documents écrits (scanner) en images
- Application de techniques de reconnaissance de formes (lettres) à l'aide d'apprentissage (réseaux de neurones, HMM)
- Exploitation d'un modèle de langage (dont des ressources : dictionnaires, grammaires, etc.) pour déterminer l'hypothèse la plus probable

Applications pratiques : dématérialisation de documents (bibliothèques), formulaires (chèques, administration), adresses pour le tri postal, identification d'immatriculation .
Industriels du domaine : Nuance (ScanSoft / Xerox), ABBYY, IRIS, Novo Dynamics, Datacap, EDT, Ligature. [7]

I.6.1.5 La Correction orthographique / grammaticale :

Les principaux traitements d'un correcteur orthographique sont :

- Identification des mots (tokenization)
- Correction orthographique : mots qui n'appartiennent pas au dictionnaire et qui ne sont pas en langue étrangère, ni des noms propres, ni des chiffres, ni des sigles...
- Correction grammaticale : déterminer la fonction des mots au sein de la phrase (déterminant, nom, verbe, adverbe, etc.) puis réaliser une analyse syntaxique à l'aide de grammaires.

Applications pratiques : correction de document rédigé par des étudiants dont le niveau de français est généralement bas.
Industriels du domaine : Synapse, Druide, Microsoft, Diagonal. [7]

I.6.1.6 D'autres applications sont issues du traitement documentaire, comme par exemple :

- Le routage, classement ou l'indexation automatique de documents électroniques sont des variantes applicatives du paradigme de la recherche documentaire.
- Plus complexe est la tâche de trouver (ou de produire à la demande) des réponses précises aux questions de l'utilisateur (tâche de "question-réponse").
- La lecture automatisée de documents, par exemple pour les stocker dans des structures formelles de données, ou pour en extraire des résumés.
- L'analyse d'un corpus de documents relatifs à un thème donné (histoire, stylométrie, veille technologique, etc). Une application typique de ce domaine consiste à fournir des outils de visualisation et d'exploration dynamique de champs disciplinaires (scientifiques, par exemple). [4]

I.6.2 La production de documents :

Si autant de documents électroniques sont aujourd'hui disponibles, c'est bien que quelqu'un les a écrits. Dans le domaine de l'aide à la production de texte (la génération de textes), les applications du TALN sont également nombreuses :

- les claviers « auto-correcteurs » (par exemple pour les handicapés) ;
- la reconnaissance optique de caractères. De nombreux systèmes commerciaux sont aujourd'hui disponibles, avec des performances très satisfaisantes : Recognita, Omnipage, ScanWorX... ;
- la génération automatique de documents à partir de spécifications formelles. En fait, de nombreux secteurs d'activité impliquent la production massive de textes très stéréotypés à partir de spécifications plus ou moins formelles (textes juridiques, compte-rendu d'exploration d'une base de données, rapports d'analyses statistiques, documentations techniques, etc). Pour cette classe de documents, il est parfaitement possible de générer automatiquement, sinon des textes complètement définitifs, du moins des versions préliminaires qui seront ensuite finalisés par des rédacteurs humains.

On retrouve dans ces applications la même dialectique que dans les applications destinées à faciliter la gestion de documents. D'un côté des applications à large couverture, qui utilisent essentiellement des ressources lexicales, avec des fonctions d'accès tolérant (permettant la correction d'erreurs) au lexique : c'est le cas des applications qui tournent autour de la correction orthographique. De l'autre, des applications qui intègrent des mécanismes de traitement de plus haut niveau (typiquement la génération), mais qui ne fonctionnent que pour des domaines beaucoup plus restreints. [4]

I.6.3 Les interfaces naturelles :

Dernier domaine d'application, qui est sans doute celui dans lequel la demande de traitements linguistiques est la plus forte, le domaine des interfaces naturelles (i.e. en langage naturel) telles que :

- l'interrogation en langage naturel de bases de données (traduction langage naturel ↔ SQL) ou de moteurs de recherche sur la toile. De multiples applications de ce type commencent à se mettre en place sur la toile.
- les interfaces vocales, qui mettent en œuvre de manière variable suivant les applications des modules de reconnaissance de parole, synthèse de parole, génération et gestion de dialogue, accès aux bases de connaissance,..., chacun de ces modules demandant des traitements spécifiques (désambiguïsation morpho-syntaxique et identification de syntagmes pour la synthèse, grammaires stochastiques pour la reconnaissance de la parole...).

Les premiers systèmes « grand-public » de dictée vocale commencent à véritablement arriver sur le marché (Via Voice, le produit d'IBM, Dragon Dictate, et bien d'autres encore), et l'intégration dans Windows d'une API de traitement de la parole devrait, dans les années qui viennent, faire littéralement exploser le marché des technologies vocales. De nombreux services commerciaux existent déjà, ou sont proches de la commercialisation qui font appel à l'ensemble de ces techniques (reconnaissance-dialogue-synthèse), pour des applications diverses telles que les ordinateurs « main-libre », la lecture téléphonique de courriers 21 électroniques, la réservation de billets (train, avion)... la liste est pratiquement sans limite. Le marché des technologies vocales représentait en 1995 quelques centaines de millions de dollars pour les États-Unis, et est estimé, pour l'an 2 000, à près de 3 Milliards de dollars. [4]

I.7 Conclusion :

L'étude des langues et des mécanismes nécessaires à la mise en œuvre à son traitement automatique par des machines est un domaine d'études foisonnant, et riche en applications potentielles ou émergentes. De nombreux progrès restent à accomplir pour mieux comprendre cette faculté et pour bâtir des systèmes capables de soutenir la comparaison avec l'humain, mais l'état des connaissances en permet aujourd'hui de proposer de nombreuses solutions efficaces à des problèmes et des demandes réels.

A travers ce premier chapitre, nous avons mis le point sur les concepts de base de TAL, nous avons ainsi exploré les principales générations des produits et systèmes TAL.

Dans ce qui suit, nous allons mettre le point sur le TAL Arabe et les différents problèmes liés particulièrement à la structure linguistique complexe de cette langue en essayant de recenser les différents outils existants et qui peuvent nous aider à accomplir les traitements nécessaires dans notre système, à savoir, la détection automatique des relations et cohésions lexicales entre différentes phrases d'un texte écrit.

Chapitre II :

Traitement automatique de la langue arabe

II.1 Introduction :

L'arabe est parlé dans plus de 22 pays, du Maroc jusqu'à l'Iraq et dans toute la péninsule arabe. C'est la première langue pour plus de 250 millions de personnes et la deuxième pour 40 millions. L'arabe, langue du Coran, est devenue la langue d'une civilisation et ne sert plus seulement à désigner les seuls habitants de la péninsule arabe qui la parlaient.[8]

Les outils du traitement automatique de la langue en général, et de la langue Arabe en particulier, se caractérisent par leur diversité au niveau des langages de développement utilisés, des entrées/sorties manipulées, des représentations internes et externes des résultats, etc. Cette diversité ne favorise ni l'interopérabilité entre ces différents outils, ni leur réutilisabilité dans de nouveaux contextes. [9]

Dans ce chapitre, nous présentons d'abord la langue arabe et sa morphologie. Ensuite, nous donnons un aperçu sur les problèmes de traitement automatique de la langue arabe et les principaux niveaux de TAL arabe et aussi nous présentons les principaux outils liés au TAL arabe écrit. Et en termine avec une conclusion.

II.2 La langue arabe et sa morphologie :

Par ses propriétés morphologiques et syntaxiques la langue arabe est considérée comme une langue difficile à maîtriser dans le domaine du traitement automatique des langues. L'arabe doit sa formidable expansion à partir du 7eme siècle grâce à la propagation de l'islam et la diffusion du Coran. Les recherches pour le traitement automatique de l'arabe ont débuté vers les années 1970. Les premiers travaux concernaient notamment les lexiques et la morphologie. [6]

La langue arabe appartient à la famille des langues sémitiques. Contemporaine la plus parlée de nos jours avec plus de 300 millions de locuteurs . C'est la cinquième langue parlée dans le monde . Dans la plupart des langues, comme le français ou l'anglais, la langue écrite et parlée est la même. La particularité de la langue arabe est que l'arabe écrit est différent de l'arabe parlé. [10]

La langue arabe se présente sous deux formes principales : l'arabe littéraire et l'arabe dialectal. L'arabe littéraire est la langue officielle du monde arabe, tandis que l'arabe dialectal – spécifique pour chaque pays – est la vraie langue parlée dans le monde arabe.[11]

D'après Farghaly et Shaalan (2009) « L'arabe littéraire se répartit en deux catégories : l'arabe classique et l'arabe moderne standard (AMS). L'arabe classique est utilisé dans les prières et les textes religieux, et constitue la base de l'arabe moderne standard. L'arabe moderne standard est une forme plus récente de l'arabe classique ; elle est utilisée dans les médias, les journaux, les salles de classe et l'administration » .

L'étude de la grammaire arabe a commencé très tôt au milieu du 11ème siècle de l'hégire et a donné lieu à d'énormes productions, avant de connaître une période de stagnation qui a duré plusieurs siècles. Ces dernières années, elle connaît un regain d'intérêt, entre autres dans le domaine du traitement automatique. [10]

Figure II.4 : Le monde arabe[2]

II.2.1 L'arabe dialectal :

كانت العرب و ان جمع جميعها اسم انهن عرب ،فهم مختلفون بالبيان متابينو المنطق و الكلام .

محمد بن جرير الطبري

L'arabe dialectal ou parlé est la langue parlée utilisé dans la vie quotidienne pour communiquer. Elle est apprise comme langue maternelle par les locuteurs arabophones.[11]

Dans le monde arabe, chaque pays dont la langue principale est l'arabe a son propre dialecte. Dans un même pays, le dialecte peut être différent d'une région à une autre avec de petites variations et quelques mots différents. Si deux pays arabes sont voisins, alors plus on se rapproche des frontières, plus le dialecte se rapproche du dialecte du pays voisin. Dans la région de Annaba qui se situe au Nord-Est de l'Algérie, par exemple, le

[2]http://fr.wikipedia.org/wiki/Monde_arabe

dialecte de Annaba est plus proche du tunisien que de l'algérien. Toutefois, les deux pays partagent exactement la même langue écrite qui est l'arabe moderne standard. Plusieurs classifications des dialectes ont été proposées pour le monde arabe. [11]

Avec l'apparition d'Internet et des nouvelles technologies, le dialecte devient de plus en plus une langue écrite dans les forums, les SMS, le chat ou aussi dans les messages électroniques. Le dialecte peut être écrit soit en caractères arabes, ou aussi en caractères latins (arabe translittéré) si l'utilisateur n'a pas l'habitude d'utiliser le clavier arabe ou a des problèmes d'encodage lors de l'utilisation des caractères arabes . Toutefois, même s'il est écrit, le dialecte reste de l'arabe informel. [11]

II.2.2 L'arabe moderne standard :

L'arabe moderne standard – appelé aussi arabe formel – est la forme écrite de l'arabe. L'arabe s'écrit de droite à gauche et les lettres sont liées entre elles. Il n'existe pas de distinction entre majuscules et minuscules en alphabet arabe. [11]

L'arabe son alphabet compte 28 lettres (Tableau 1) dont 25 sont des consonnes et 3 (ا و يا) des consonnes ou des voyelles longues selon leur contexte d'apparition. Le (ي/y/) est une voyelle longue dans دُلْفِينٌ /delfinun/ (« un Delphine ») et consonne dans بَيْتٌ /baytun/ (« un maison »). Toutes les lettres se lient entre elles sauf (ذ, وا, ر, ز, د,) qui ne se joignent pas à gauche. [6] [10]

Lettre arabe	Correspondant français	Prononciation
ا	a	Alef
ب	b	Baʾ
ت	t	Taʾ
ث	th	Thaʾ
ج	j	Jim
ح	h	Hhaʾ
خ	kh	Khaʾ
د	d	Dal
ذ	d	Thal
ر	r	Ra
ز	z	Zin
س	s	Sin
ش	sh	Shin
ص	s	Sad

Lettre arabe	Correspondant français	Prononciation
ض	d	Dad
ط	t	Tah
ظ	z	Zah
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
ي	y	Ya

Tableau II.2 : les 28 lettres arabes [6]

De plus certaines lettres comme ا Alef peuvent symboliser le آ ,أ ou إ; de même que pour les lettres ى et ه qui symbolisent respectivement ي et ة . La table des translittérations de l’alphabet arabe sont présentés en **Annexe1**.

La graphie des lettres est différente selon leur position dans le mot. Le Tableau II.2 montre les variations de la lettre م (Mim). Il résulte 78 formes graphiques à partir des 28 lettres.

à la fin de mot	au milieu de mot	au début de mot
م	مـ	مـ
نُجُومٌ/nojomun / « des stars »	عَمِلَ/3amila/ « il a travaillé »	مَرَّ / marra/ « il est passé»

Tableau II.3 : Exemple de variation de la lettre م/mim

II.2.2.1 Les diacritiques :

Les voyelles brèves sont figurées par des symboles appelés signes diacritiques (les *harkāt*). L'opération qui consiste à les insérer par une machine dans un texte est appelée vocalisation automatique ou voyellation automatique. Au nombre de trois, ces symboles sont transcrits de la manière suivante:

- *La fetha* [a] est symbolisée par un petit trait sur la consonne (تَ/ta/)
- *La damma* [u] est symbolisée par un crochet au-dessus de la consonne (تُ /tu/)
- *La kasra* [i] est symbolisée par un petit trait au-dessous de la consonne (تِ /ti/)

Un petit rond ْ_ symbolisant *la soukoun* (سكون) est appose sur une consonne lorsque celle-ci n'est liée à aucune voyelle (سَوْفَ/ sawfa) . [10]

II.2.2.2 Le tanwin :

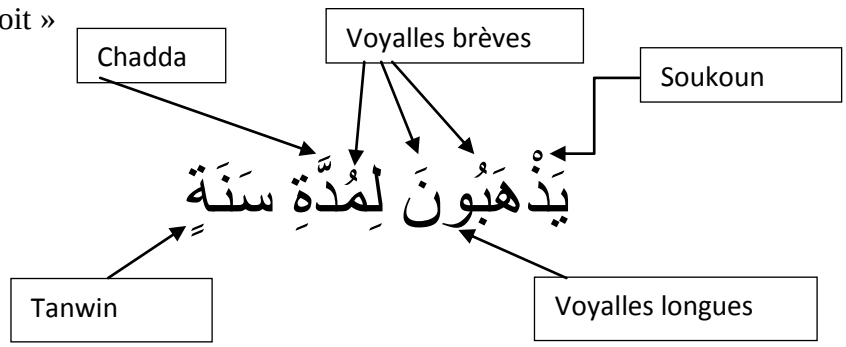
Le signe du tanwin est ajouté à la fin des mots indéterminés. Il est en relation d'exclusion avec l'article de détermination ال placé en début de mot. Les symboles du tanwin sont au nombre de trois et sont constitués par dédoublement des signes diacritiques ci-dessus, ce qui se traduit par l'ajout du phonème /n/ au niveau phonétique:[10]

- [an]: signe ً_ (بً/ban)
- [un]: signe ٌ_ (بٌ /bun)
- [in]: signe ٍ_(بٍ /bin)

II.2.2.3 La chadda :

Le signe de la chadda peut être placé au-dessus de toutes les consonnes en position non- initiale. La consonne qui la reçoit est alors analysée en une séquence de deux consonnes identiques: [10]

Signe الْحَقُّ_ /alha**kk**o « le droit »



II.2.3 Morphologie arabe:

Selon la grammaire traditionnelle, le lexique arabe comprend trois catégories de mots : verbes, noms (substantifs et adjectifs) et particules (adverbes, conjonctions et prépositions). Hormis les noms propres, les mots des deux premières catégories sont dérivés à partir d'une racine : un squelette de trois consonnes radicales le plus souvent.[12]

II.2.3.1 Structure d'un mot :

En arabe un mot peut signifier toute une phrase grâce à sa structure composée qui est une agglutination d'éléments de la grammaire, la représentation suivante schématise une structure possible d'un mot. Notons que la lecture et l'écriture d'un mot se fait de droite vers la gauche. [6]

Enclitique	**Suffixe**	**Cors schématique**	**Préfixe**	**Proclitique**

- **Proclitiques** sont des prépositions ou des conjonctions.
- **Cors schématique** représente la forme de base pour chaque mot.
- **Préfixes** et **suffixes** expriment les traits grammaticaux et indiquent les fonctions : cas du nom, mode du verbe et les modalités (nombre, genre, personne,…)
- **Enclitiques** sont des pronoms personnels.

Exemple :

أَتُعَلِمِينَهُمْ

Est-ce que tu peux les enseigner ?

Enclitique	**Suffixe**	**Cors schématique**	**Préfixe**	**Proclitique**
هُمْ **<hom'>**	ينَ **<yna>**	عَلِمِ **<EaLiMi>**	تُ **<tu>**	أ >
Objet masculin pluriel	Sujet féminin singulier	la racine ″ELM″	préfixe verbal du temps de l'inaccompli	conjunction d'interrogation

Tableau II.4 : Structure de mot أَتُعَلِمِينَهُمْ.

II.2.3.2 Catégories des mots :

L'arabe considère 3 catégories de mots :

- Le verbe : entité exprimant un sens dépendant du temps, c'est un élément fondamental auquel se rattachent directement ou indirectement les divers mots qui constituent l'ensemble.
- Le nom : l'élément désignant un être ou un objet qui exprime un sens indépendant du temps.
- Les particules : entités qui servent à situer les événements et les objets par rapport au temps et l'espace, et permettent un enchaînement cohérent du texte. [6]

II.2.3.2.1 Le verbe :

La plupart des mots en arabe, dérivent d'un verbe de trois lettres. Chaque verbe est donc la racine d'une famille de mots. Comme en français, le mot en arabe se déduit de la racine en rajoutant des suffixes ou des préfixes. [6]

La conjugaison des verbes dépend de plusieurs facteurs :

− Le temps (accompli, inaccompli).

− Le nombre du sujet (singulier, duel, pluriel).

− Le genre du sujet (masculin, féminin).

− La personne (première, deuxième et troisième)

− Le mode (actif, passif).

Voici un exemple qui illustre ces différents facteurs :

ف+ت+ح F+T+H donne le verbe فتح FaTaHa (ouvrir).Dans tous les mots qui dérivent de cette racine, on trouvera ces trois lettres F,T,H .

Comme on l'a dit précédemment, La conjugaison des verbes se fait en ajoutant des préfixes et des suffixes, ou les deux.

La langue arabe dispose de trois temps :

- L'accompli : correspond au passé et se distingue par des suffixes. Pour notre exemple, avec le féminin pluriel on a فَتحنَ /FaTaHna/ « *elles ont ouvert »,* et pour le masculin pluriel on a فَتحوا /FaTaHuu/ « *ils ont ouvert ».*

- L'inaccompli présent: présente l'action en cours d'accomplissement, les verbes conjugués à ce temps se distinguent par les préfixes. Pour notre exemple, يَفتح /yaFTaH/ « *ilouvre* »; تَفتح /taFTaH / « *elleouvre* ».
- L'inaccompli futur : correspond à une action qui se déroulera au futur et est marqué par l'antéposition de س*sa* ou سوف *sawfa* au début verbe conjugué à l'inaccompli. Pour notre exemple, En ajoutant l'antéposition س on obtient سَيَفتح /sayaFTaHu/ « *il ouvrira*» , qui désigne le futur ; on peut également ajouter l'antéposition سوف on obtient سوف يفتح « *il va ouvrir* ». [6]

II.2.3.2.2 Les noms :

Les substantifs arabes sont de deux catégories, ceux qui sont dérivés de la racine verbale et ceux qui ne le sont pas comme les noms propres et les noms communs. [6]
Dans le premier cas, le fait que le nom soit dérivé d'un verbe, il exprime donc une certaine sémantique qui pourrait avoir une influence dans la sélection des phrases saillantes d'un texte pour le résumé. [6]

La déclinaison des noms se fait selon les règles suivantes:

- Le féminin singulier: On ajoute le ة, **exemple** مسلم *musulman* devient مسلمة*musulmane* .
- Le féminin pluriel : De la même manière, on rajoute pour le pluriel les deux lettres ات , **exemple** كَاتب(écrivain) devient كَاتبات (écrivains).
- Le masculin pluriel : Pour le pluriel masculin on rajoute les deux lettres ين ou ونdépendamment de la position du mot dans la phrase (sujet ou complément d'objet) ,**exemple** الرّاجع (*revenant*) الرّاجعون ouالرّاجعين (*revenants*).
- Le Pluriel irrégulier: Il suit une diversité de règles complexes et dépend du Nom, e**xemple** طِفل (*un enfant*) أطفال (*des enfants*).

Comme l'a noté [48] « Le phénomène du pluriel irrégulier dans l'arabe pose un défi à la morphologie, non seulement à cause de sa nature non concaténative, mais aussi parce que son analyse dépend fortement de la structure comme pour les verbes irréguliers »
Certain dérivés nominaux associent une fonction au nom :

- Agent : celui qui fait l'action ,
- Objet : celui qui a subit l'action ,
- Instrument : désignant l'instrument de l'action ,

- Lieu : qui désigne en général un endroit (exemple : منزل « maison »)
- Ect…

Pour les pronoms personnels, le sujet est inclus dans le verbe conjugué. Il n'est donc pas nécessaire (comme c'est le cas en français) de précéder le verbe conjugué par son pronom. On distinguera entre singulier, duel (deux) et pluriel (plus de deux) ainsi qu'entre le masculin et féminin. [6]

II.2.3.2.3 Les particules:

Ce sont principalement les mots outils comme les conjonctions de coordination et de subordination.

Les particules sont classées selon leur sémantique et leur fonction dans la phrase, on en distingue plusieurs types (introduction, explication, conséquence, …). Elles jouent un rôle important dans l'interprétation de la phrase. Elles servent à situer des faits ou des objets par rapport au temps ou au lieu, elles jouent également un rôle clé dans la cohérence et l'enchaînement d'un texte.

Comme exemple de particules qui désignent un temps منذ,قبل,بعد*pendant, avant, après*, un lieu حيث*où*, ou de référence الذين*ceux*,….

Ces particules seront très utiles pour notre traitement à deux niveaux :

- Elles font partie de l'antidictionnaire qui regroupe les termes à ne pas prendre en considération lors de calcul de fréquence de distribution des mots,
- Elles identifient des propositions composant une phrase. Les particules peuvent avoir des préfixes et suffixes ce qui rajoute une complexité quant à leur identification. [6].

II.2.3.3 La structure des phrases en arabe :

En langue arabe, il y a deux types de phrases : les phrases verbales et les phrases nominales.

Les phrases verbales servent à indiquer un évènement ou une action. Elles commencent par un verbe suivi d'un sujet et d'un complément; ce dernier est optionnel. La phrase verbale est celle que l'on rencontre le plus souvent dans l'expression courante. Par exemple جلس الطالب est traduit par *L'étudiant s'est assis*.[11]

Les phrases nominales en arabe ne contiennent pas de verbe, il est sous entendu. En arabe, le verbe *être* est implicite, les verbes ne sont pas obligatoires pour construire une phrase. Les phrases nominales sont constituées d'un sujet et d'un attribut (adjectif qualificatif, complément circonstanciel, ...). Les phrases nominales sont affirmatives, par

exemple الطالب جميلest traduite par *L'étudiant est beau*, mais le verbe est absent de la phrase en arabe .[11]

L'utilisation des phrases verbales, comme les phrases nominales, dépend du contexte dans lequel la phrase est exprimée. Dans le style journalistique, les phrases verbales sont les plus utilisées. [11]

II.2.3.4 Les Schèmes :

L'Arabe est une langue **sémitique** dont la morphologie fonctionne sur le croisement des racines (généralement trilitères) avec un nombre de formes déterminées pour produire sa gamme de vocabulaire. Ces formes, appelées **moules** ou **schèmes**, sont essentiellement constituées par une suite de voyelles et consonnes qui viendront se placer avant, après et entre les lettres de la racine afin de former les différentes formes des flexions verbales et nominales.[3]

Une famille de mots peut ainsi être générée d'un même concept sémantique à partir d'une seule racine à l'aide de différents schèmes. Ce phénomène est caractéristique à la morphologie arabe. Le Tableau 3donne quelques exemples de schèmes appliqués aux mots عمل *travailler* et حمل *porter*. On peut ainsi dériver un grand nombre de noms, de formes et de temps verbaux. [6]

Schèmes	EML	عمل	Notion de *travailler*
R_1â-R_2i-R_3	EaMiL	عَامِل	ouvrier
R_1a-R_2a-R_3a	EaMaLa	عَمَلَ	a travaillé
maR_1R_2aR_3	maEMaL	مَعْمَل	atelier
R_1uR_2iR_3a	EuMiLa	عُمِلَ	aété travaillé
. . .			

HML	حمل	Notion de *porter*
HâMiL	حَامِل	porteur
HaMaLa	حَمَلَ	a porté
maHMaL	حْمَل	brancard
HuMiLa	حُمِلَ	a été porté
…		

Tableau II.5 : Exemples de schèmes appliqués aux mots عمل travailler et حمل porter

Les lettres en majuscule (R_i) désignent les consonnes de base qui composent la racine.

[3] http://ouiam91.tumblr.com/

Les voyelles (â, a, i,..) désignent les voyelles et les consonnes en minuscule (m,..) sont des consonnes de dérivation utilisées dans les schèmes.

La majorité des verbes arabes ont une racine composée de 3 consonnes. L'arabe comprend environ 150 schèmes ou patrons dont certains plus complexes, tel le redoublement d'une consonne ou l'allongement d'une voyelle de la racine, l'adjonction d'un ou de plusieurs éléments ou la combinaison des deux.

Exemple de schéma :

1/ Le Participe Actif إسم الفاعل

Ce schème désigne **celui qui fait l'action.** Il est toujours associé à un verbe de forme simple (non dérivé), il prend le sens de ce qui fait l'action du verbe issu de la racine.[4]

Signification المعنى	**Schème الوزن**	**Moule القالب**	**Verbe الفعل**
Celui qui fait	فاعل	ف+ا+ع+ل	فعل
Celui qui arrive	واصل	و+ا+ص+ل	وصل
Celui qui nage (nageur)	سابح	س+ا+ب+ح	سبح
Celui qui travaille (travailleur)	عامل	ع+ا+م+ل	عمل
Celui qui descend	نازل	ن+ا+ز+ل	نزل

Tableau II.6 : Exemples des schèmes إسم الفاعل

REMARQUE 1 :

a) Lorsqu'un verbe est **Concave** (أجوف) (sa deuxième letter est une voyelle), cette voyelle se transforme en همزة.

Exemple :

Signification المعنى	**Schème الوزن**	**Moule القالب**	**Verbe الفعل**
Celui qui jeune (jeuneur)	صائم	ص+ا+ي+م /ص+ا+ء+م	صام
Celui qui vend (vendeur)	بائع	ب+ا+ي+ع /ب+ا+ء+ع	باع

Tableau II.7 : les schèmes des mots concaves صائم et بائع.

b) Lorsqu'un verbe est **Défectueux** (ناقص) (satroisièmelettreestunevoyelle), la dernière lettre se transforme en ي

[4] http://ouiam91.tumblr.com/

Exemple :

Signification المعنى	**Schème الوزن**	**Moule القالب**	**Verbe الفعل**
Celui qui invite	داعي	د+ا+ع+ي	دعى
Celui qui marche	ماشي	م+ا+ش+ي	مشى

Tableau II.8 : les schèmes des mots défectueux دعى et مشى.

REMARQUE 2 :

Lorsque le verbe **n'est pas** trilitère, on obtient le participe actif de cette manière:

1. On le conjugue à la troisième personne du singulier à l'inaccompli
2. On change le ي en مُ
3. On met Kasra sur l'avant dernière lettre (si elle n'y est pas déjà)

Prenons le verbe سافر (voyager) comme exemple. Pour obtenir son participe actif, on suit les étapes précédentes :

1. Le conjuguer à le 3ème personne du singulier à l'inaccompli يُسافِر
2. On change le ي en م pour avoir:مُسافر
3. On s'assure qu'il y a *Kasra* sur l'avant dernière lettre : مُسافِر
4. On a obtenu le participe actif de سافر (voyager) qui est مُسافِر (voyageur)

Autres exemples :

le participe actif	**Le verbe الفعل**
مسلم (musulman)	أسلم
مغني (chanteur)	غنى (chanter)
مهاجر (émigré)	هاجر (émigrer)
مهاجم (attaquant)	هاجم (attaquer)

Tableau II.9 : le participe actif des exemples des verbes n'est pas trilitère

2/ Le Participe Passif إسم المفعول

Ce schème désigne **celui qui subit l'action, i**l est toujours associé aux verbes de forme simple (non dérivés), le mot qui est inséré dans ce schème prend le sens de ce qui subit l'action du verbe issu de la racine.[5]

Pour pouvoir appliquer ce Schème, le verbe doit être **transitif direct** (qui prend un complément d'objet directe). Les verbes intransitif comme ذهب(aller) ou سبح(nager) n'ont pas de participe passif.

Signification المعنى	Schème الوزن	Moule القالب	Verbe الفعل
Celui qui fait	فاعل	ف + ا + ع + ل	فعل
Celui qui arrive	واصل	و+ ا + ص + ل	وصل
Celui qui nage (nageur)	سابح	س + ا + ب + ح	سبح
Celui qui travaille (travailleur)	عامل	ع + ا + م+ ل	عمل
Celui qui descend	نازل	ن + ا + ز+ ل	نزل

Tableau II.10 : Exemples du participe passif إسم المفعول

II.2.3.5 Autres caractéristiques de la langue arabe :

Une autre caractéristique de la langue arabe est qu'elle est une langue pro-drop car elle permet l'omission des pronoms sujets. Le terme pro-drop a été introduit par Chomsky (1981) et provient de l'expression pronoun dropping . Ce pronom peut être soit un véritable pronom suffixe, soit il est sous-entendu (on dit qu'il est caché ضمير مستتر) , ce qui est e cas pour la troisième personne du singulier lorsque le sujet n'est pas énoncé sous la forme d'un nom après le verbe . [11]

Considérons l'exemple donné par تقدمت التكنولوجية (la technologie se développe) ; C'est une phrase verbale constituée d'un verbe suivi d'un nom. Le verbe est constitué de la racine تقدم (développe) suivie d'un suffixe تqui représente la marque de féminin . [11]

II.3 Problèmes de traitement automatique de la langue arabe :

La langue arabe rencontre quatre principaux problèmes en traitement automatique : la segmentation du texte, l'agglutination des mots et détection de racine, l'absence de

[5] http://ouiam91.tumblr.com/

voyelles à l'écrit et l'étiquetage grammatical. Pour chacun de ces problèmes, tout système de traitement automatique doit traiter et enlever certaine ambiguïté.

II.3.1 Segmentation :

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

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

La reconnaissance de la fin de phrase est délicate car la ponctuation n'est pas systématique et parfois les particules délimitent les phrases.
Pour la segmentation de texte utilise :
• Une segmentation morphologique basée sur la ponctuation,
• Une segmentation basée sur la reconnaissance de marqueurs morphosyntaxiques ou des mots fonctionnels comme : أو ,و ,أي ,لكن ,حتى *ou, et, c.à.d., mais, quand.*
Cependant, ces particules peuvent jouer un autre rôle que celui de séparer les phrases. [6]

II.3.2 Agglutination des mots et Détection de racine :

La plupart des mots arabes sont composés par agglutination d'éléments lexicaux de base (proclitique + base + enclitique). Par exemple, la détermination peut s'exprimer par agglutination de l'article ال/?al/ avant le mot (المالية/?almaleya/ (« financement »)ou par agglutination d'un pronom personnel après celui-ci (ماله/malohu/ (« son argent »). [6]

Dans toute perspective de traitement automatique, le problème est donc de décomposer le mot en ces différentes parties. Cette décomposition nécessite des connaissances de niveau supérieur en cas d'ambiguïtés (si le mot accepte plusieurs segmentations). [6]

Pour détecter la racine d'un mot, il faut connaître le schème par lequel il a été dérivé et supprimer les éléments flexionnels (antéfixes, préfixes, suffixes, post fixes) qui ont été ajoutés. [6]

Nous utilisons la liste de préfixes et de suffixes voir Tableau II.11 . Plusieurs d'entre eux ont été utilisés par [Chen et Gey, 2002] pour la lemmatisation de mots arabes; ils ont été déterminés par un calcul de fréquence sur une collection d'articles arabes de l'Agence France Press (AFP). [6]

Préfixes							
لا	فبـ	للـ	كمـ	بمـ	وتـ	بتـ	والـ
با	وا	ليـ	فمـ	لمـ	ستـ	يتـ	فالـ
	فا	ويـ	الـ	و	نتـ	متـ	بالـ
Suffixes							
ا	ة	ين	ية	هم	ته	وه	ات
	ه	يه	تك	هن	تم	ان	وا
	ي	ية	نا	ها	كم	تي	ون

Tableau II.11 : Liste des préfixes et suffixes les plus fréquents [6].

L'analyse morphologique devra donc séparer et identifier des morphèmes semblables aux mots préfixés comme les conjonctions wa- و et fa- ف, des prépositions préfixées comme bi- ب et li- ل, l'article défini ال , des suffixes de pronom possessif. [6]

La phase d'analyse morphologique détermine un schème possible. Les préfixes et suffixes sont trouvés en enlevant progressivement des préfixes et des suffixes et en essayant de faire correspondre toutes les racines produites par un schème afin de retrouver la racine.[6]

Lorsqu'un mot peut être dérivé de plusieurs racines différentes, la détection de la racine est encore plus difficile, en particulier en absence de voyelles.

Par exemple, pour le mot arabe ايمان AymAn les préfixes possibles sont : "Ø","A ا" et "Ay اي " et les suffixes possibles sont : "Ø" et "An ان " (Tableau II.12), sans compter que ce mot peut aussi représenter un nom propre إيمَان *Imène*. [6]

Stem	**Préfixe**	**Schème**	**Suffixe**	**Racine**	**Signification**
AymAn ايمان	Ø	R1yR2aR3	Ø	Amn امن	croyance
ymAn يمان	A ا	R1R2aR3	Ø	Ymn يمن	convenant
mAn مان	Ay اي	R1R2R3	Ø	mAn مان	Va-t-il approvisionner
Aym ايم	Ø	R1R2R3	An ان	Aym ايم	Deux veuves

Tableau II.12 : Les stems possibles pour le mot ايمان [6].

Certains verbes sont considérés comme irréguliers, ce sont ceux qui portent des consonnes particulières dites faibles (و, ا, ي). Ils sont appelés ainsi parce que, lors de leur déclinaison, chacune de ces lettres est soit conservée, soit remplacée ou éliminée. Le Tableau II.13 donne un exemple de dérivation du mot قالdire. [6]

Caractère ا est remplacé par	**قال**	**dire**
ا	قال	Il a dit
و	يقول	Il dit
ي	قيل	Il a été dit
Ø	قل	dis

Tableau II.13 : Exemple de déclinaison du verbe irrégulier قال dire [6].

Une difficulté en traitement automatique de la langue arabe est l'agglutination par laquelle les composantes du mot sont liées les unes aux autres. Ce qui complique la tâche de l'analyse morphosyntaxique pour identifier les vrais composants du mot. [6]
Par exemple, le mot أَلَمُهُمْ ALaMuhum (*leur douleur)* dans sa forme voyellée n'accepte qu'une seule segmentation : هُمْ + أَلَمُ (ALaMu+hum). Dans sa forme non voyellée المهم (ALMHM), le même mot accepte au moins les trois segmentations présentées dans le Tableau II.14. [6]

Segmentation possible		Traduction en français
أ+ لم + هم	A+LM+hm	*Les a-t-il ramassés*
ألم + هم	ALM+hm ALM+hm	*Leur douleur* *Il les a fait souffrir*
أل + مهم	Al+MHM	*L'important*

Tableau II.14 : Exemple de segmentation du mot المهم. [6].

L'amplification de l'ambiguïté de segmentation s'opère selon deux façons :
− D'abord, il y a plus d'unités ambiguës dans un texte non voyellé que dans son correspondant voyellé,
− Mais aussi, les unités ambiguës acceptent plus de segmentations dans le texte non voyellé.

De plus le fait de précéder la lemmatisation par la troncature des préfixes avant les suffixes (et réciproquement) peut influencer les résultats. En considérant l'exemple dans le

Tableau II.14 , sur un texte où la notion de douleur est importante, le fait d'avancer la suppression des préfixes avant les suffixes les mots comme أَلَمُهُمْ *leur douleur (pour le pluriel),* **ألمهما***leur douleur* (pour le duel) exprimeront une touteautre notion.[6]

II.3.3 Voyellation :

Le problème de la voyellation réside dans l'absence quasi systématique de la voyellation dans les textes arabes. En effet, les signes de voyellation, lorsqu'ils sont notés, sous la forme de signes diacritiques placés au-dessus ou au-dessous des lettres, apparaissent dans certains textes (Coran, hadith) ou littéraires (poésie classique, notamment) : on dit qu'ils sont édités en graphie voyellée. [8]

L'absence de voyelles (la non-voyellation) dans les textes arabes génère plusieurs cas d'ambiguïtés et des problèmes lors de l'analyse automatique. En effet, l'ambiguïté grammaticale augmente si le mot est non voyellé. Cela est dû au fait qu'un mot non voyellé possède plusieurs voyellations possibles, et pour chaque voyellation est associée une liste différente de catégories grammaticales. [8]

L'exemple suivant II.15 du mot non-voyellé ktb | كتب possède 16 voyellations potentielles et qui représentent 8 catégories grammaticales différentes .

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

Tableau II.15 : Exemple de voyellation de mot non-voyellé ktb | كتب [8]

II.3.4 L'étiquetage grammatical :

L'étiquetage grammatical est l'opération qui consiste à attribuer à chacun des mots d'un texte la catégorie (nom, verbe, adjectif, article défini, etc.).

La difficulté de l'étiquetage grammatical s'amplifie lorsque les textes visés se présentent sous leur forme non pas voyellée, mais partiellement voyellée ou encore totalement non voyellée, ce qui correspond au cas le plus courant. [8]

II.4 Les Principaux Niveaux d'Analyse :

II.4.1 Analyse morphologique :

L'opération de l'analyse morphologique tient à étudier la forme d'un mot en faisant une analyse interne de la structure de ce dernier. Le but étant de décomposer un mot à des éléments plus petits (préfixes, suffixes, etc.) selon des règles de combinaison relatives à ces derniers. [8]

Le problème principal de cette analyse réside dans l'absence de voyellation. L'absence de voyellation pose un problème important. En effet, les mots non voyellés engendrent beaucoup de cas ambigus au cours de l'analyse (e.g. le mot non voyellé " فصل" prishors contexte peut être un verbe au passé conjugué à la troisième personne du singulier "il a licencié |فَصَلَ ", ou un nom masculin singulier "chapitre/ saison |فَصلٌ", ou encore une concaténation de la conjonction de coordination "puis |فَ" avec le verbe " صل" : impératif du verbelier conjugué à la deuxième personne du singulier masculin). [8]

II.4.2 L'analyse syntaxique :

L'analyse syntaxique permet d'associer à un énoncé sa ou ses structures syntaxiques possibles, en identifiant ses différents constituants et les rôles que ces derniers entretiennent entre eux. [8]

Toutefois, l'analyse syntaxique prend en entrée le résultat de l'analyse lexicale (éventuellement de l'étiquetage morpho-syntaxique) et fournit en sortie une structure hiérarchisée des groupements structurels et des relations fonctionnelles qui unissent les groupements. [8]

Enfin, il est à signaler que les ambiguïtés vocaliques et grammaticales, relatives à la non voyellation des mots, pose des difficultés au niveau de l'analyse syntaxique. Ainsi, une phrase, en absence de la voyellation, peut être interprétée et traduite selon plusieurs interprétations qui sont toutes syntaxiquement correctes. [8]

II.4.3 L'analyse sémantique :

L'analyse sémantique tente de découvrir de façon plus générale le sens des mots, des phrases ou des textes entiers. C'est la phase la plus laborieuse pour les machines, et pour cette raison elle reste encore assez peu employée.

L'absence de voyelles peut générer des défauts de sens dans le traitement automatique, par exemple, le mot (العلم) isolé peut avoir plusieurs interprétations (*la science* ou *drapeau*) alors que voyellé sera (العِلْمُpour *la science* et العَلَمُpour *le drapeau*).

Les outils qui opèrent cette analyse font souvent appel à de gigantesques thésaurus (comme Arabic Wordnet pour l'arabe), permettant de classer chaque terme dans une arborescence de concepts pour déterminer les thèmes dominants d'un texte, ainsi qu'à des algorithmes complexes permettant d'évaluer les relations entre les différentes idées d'un texte donné.

II.5 Les outils de T.A.L arabe :

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. Notre objectif dans cette section est de recenser les principaux outils de TAL en langue arabe. Les sujets abordés sont donc les analyseurs morphologiques, les concordanciers, les racineurs et les bases de données lexicales. [9]

II.5.1 Analyseurs morphologiques :

Le but principal de ce type d'analyseurs est de vérifier l'appartenance d'un mot donné au domaine linguistique choisi et de pouvoir disposer ainsi de tous les renseignements le concernant pouvant servir à l'analyse syntaxique. [9]
L'analyse morphologique est très développée pour les langues latines. Mais ce n'est pas le cas pour la langue arabe par manque de ressources linguistiques (eg. corpus, lexique de base, segmenteurs de textes en phrases,…). C'est pour cette raison, que la majorité des travaux se sont basés sur l'étiquetage morphologique s'appuyant sur des méthodes d'apprentissage et une légère analyse morphologique, par exemple khoja 2001 ; Diab et al. 2004. [9]

L'analyseur morphologique consiste après segmentation du texte, à étudier la forme d'un mot pris isolement (sans contexte) et à déduire les informations dérivationnelles et inflectionnelles . Ainsi, l'analyseur doit générer pour le mot traité une ou plusieurs solutions morphologique décrites par les informations suivante : les suffixes, les préfixes, le radical, la forme canonique (lemme) ainsi que d'autres informations comme le genre grammatical (féminin, masculin), le nombre (singulier, pluriel) ou le temps (verbe conjugué au présent, au passé parfait,...etc.). [9] .

II.5.1.1 AraMorph de Buckwalter :

L'analyseur de buckwalter developpé par Tim Buckwalter en langage Perl pour le compte du LDC (Linguistic Data Consortium) permet de segmenter chaque unité lexicale en une séquence du type préfixe-stem-suffixe. Le préfixe est une combinaison de 0-4 caractères, le suffixe est composé de 0 à 6 caractères et le stem comprend un à plusieurs caractères. Il est constitué principalement de trois lexiques : préfixes (548 entrées), suffixes (906 entrées), et stem (78839entrées). Les lexiques sont complétés par trois tables de compatibilité utilisés pour couvrir toutes les possibilités de combinaisons préfixe-stem (2435 entrées), suffixe-stem (1612 entrées) et préfixe-suffixe (1138 entrées). Ainsi, l'analyseur donne en sortie l'unité lexicale, sa catégorie morphosyntaxique et sa traduction anglaise. [9]

L'analyseur n'accepte pas du texte en arabe avec de l'alphabet romain dans le même document ; Un problème peut être rencontré lorsque par exemple, le texte contient des étiquettes grammaticales (Part of Speech) ou marqueurs XML en alphabet romain. [9]

II.5.1.2 L'analyseur morphologique Sebawi de Darwish

Développé par Darwish en une seule journée, Sebawi est un analyseur morphologique de la langue arabe. C'est un analyseur de surface utilisé dans des applications de recherches d'information. Il réalise seulement la recherche des racines possible d'un mot arabe. Cet analyseur morphologique arrive dans 84% des cas à trouver avec succès la racine. [9]

II.5.1.3 L'analyseur morphologique à états finis de Beesley 2001 (Xerox) :

L'analyseur morphologique de Xerox est basé sur l'approche de transducteur à états finis. La segmentation de la phrase en unités lexicales est réalisé par un transducteur à états finis. Ce transducteur découpe la chaîne d'entrée en une séquence d'unités lexicales qui peuvent correspondre à une forme fléchie, une marque de ponctuation, etc. La deuxième étape est l'analyse morphologique des unités lexicales produites par la segmentation de la phrase.

Cette étape est aussi réalisée par un transducteur qui relie la forme fléchie à la forme lexicale (et vice-versa). La forme lexicale est une séquence comprenant la représentation canonique de l'unité lexicale (le lemme), un ensemble d'étiquettes représentant le comportement morphologique de l'unité lexicale, et sa catégorie syntaxique. [9]

II.5.1.4 ASVM

L'analyseur de Diab (ASVM) est un logiciel libre, développé en Perl par l'équipe de Mona Diab à la Leland Stanford Junior University en 2004. Il s'agit d'une adaptation à l'arabe du système anglais YamCha basé sur les Support Vector Machines. Les données probabilistes ont été acquises pendant une phase d'entrainement sur le corpus annoté Arabic Tree Bank. Ci-dessous la description des fichiers d'entrée et de sortie de l'analyseur. [9]

II.5.1.5 Les analyseurs morphosyntaxiques (Pour détecter les catégories des mots) Exemple : Al Khalil

Après une recherche efficace et un travail intensif de l'équipe traitement automatique des langues naturelles du Laboratoire de Recherche Informatique « *LaRI* » de la Faculté des Sciences, Université Mohammed Premier, Oujda, en collaboration avec l'organisation Arabe de l'Education et la Culture et les Sciences (*ALECSO*) et la Cité du Roi Abdel Aziz pour la Science et la technologie (*KACST*) et avec le professeur «Al Moataz Bellah Assaid» université du Caire, cette équipe a développé un outil d'analyse morphologique batisé «AlKhalil Morpho Sys ».

AlKhalil Morpho Sys est un logiciel open source développé avec le langage de programmation orienté Java. Il consiste à faire une analyse morphologique permettant pour chaque mot du texte arabe pris hors contexte d'identifier ses différentes étiquettes morphosyntaxiques possibles, ainsi, il peut traiter les textes diacrités partiellement ou totalement. [14]

Alkhalil fournit des informations morpho-syntaxiques hors contexte du mot telles que les voyellations possibles du mot, les affixes qui s'ajoutent aux stems (préfixe et suffixe), le stem, la nature du mot (nom, verbe ou mot outil), et dans le cas des noms et des verbes le système fournit le schème, la racine et l'état syntaxique (POS tags).[13]

Alkhalil a remporté la première position, parmi les 13 systèmes morphologiques arabes à travers le monde, à un concours organisé par la Ligue arabe pour l'éducation la

culture et la science (ALECSO) (2010برنامج الخليل الصرف). Il est basé sur la modélisation d'un très large ensemble de règles morphologiques arabes, et sur l'intégration de ressources linguistiques qui sont utiles à l'analyse.[14]

Alkhalil contient environ 7000 racines obtenues à partir de Sarf ,(sarf 2007, an open source Arabic morphology system http://sourceforge.net/projects/sarf/) et NEMLAR corpus[1]. NEMLAR : Network for Euro Mediterranean LAnguage Resources. Ce corpus a été produit dans le cadre du projet NEMLAR. Le corpus écrit NEMLAR est constitué de 500 000 unités lexicales regroupés en 13 catégories différentes, visant à obtenir un corpus bien équilibré qui offre une représentation de la variété de traits syntaxiques, sémantiques et pragmatiques de la Langue arabe moderne. Chaque racine est reliée avec des procédures de dérivation spécifiques utilisées pour calculer les mots de cette racine.

Malgré les performances de cet analyseur, et de la plupart des analyseurs de la langue arabe, l'arabe est encore moins étudiée que les langues latines [14]

II.5.2 Les concordanciers :

La réalisation manuelle des concordances écrites était un travail de grande envergure envisageable uniquement pour les œuvres pérennes. Le traitement automatique a facilité la tâche et a étendu leurs champs d'application à de nombreuses disciplines scientifiques. Dans le cas de la langue arabe, l'aboutissement d'un concordancier électronique nécessite un travail préalable faisant appel à des ressources lexicales et des outils d'étiquetage morpho-syntaxique. L'approche classique de réalisation des concordanciers, basée sur une reconnaissance graphique des items dans les textes KWIC (KeyWord In Context), est inefficace dans le traitement de l'arabe, dont l'écriture est non-vocalisée, et dont les structures de l'unité lexicale peuvent être décrites comme agglutinantes et hautement flexionnelles. Ainsi, L'outil Ara- Conc développé pour l'arabe par [15] a pour objectif de donner les contextes et fréquences, et permettre l'exploration du corpus selon les traits proposés par l'analyse morphologique et selon les informations graphiques qui se trouvent dans le texte. La concordance finale arabe tourne autours du trio : unité lexicale, position et analyse morphologique. L'outil prend en entrée un texte ou un ensemble de textes. Il permet :

• La construction de listes de fréquences d'items, de racines ou tout autre trait de l'analyse morphosyntaxique, par ordre alphabétique ou par ordre fréquentiel.

• La construction d'une concordance, La consultation de la concordance peut se faire par item, par la racine, par la base ou par analyse morpho-syntaxique . [9]

II.5.3 Racineur :

Les racineurs se veulent d'abord un outil utile au TAL, ce type d'analyse « simpliste », traite de façon identique affixes flexionnels et dérivationnels. Les algorithmes de racinisation en arabe les plus connus sont ceux de [16] et [17] . Ci-dessous une description succinte de ces racineurs. [9]

II.5.3.1 Racineur de larkey :

L'approche de [16] est une analyse morphologique assouplie. Elle consiste à essayer de déceler les préfixes et les suffixes ajoutés à l'unité lexicale : par exemple

le duel (ان)dans (معلمان, deux professeurs), le pluriel des noms masculins (ون, ين) dans (معلمون, des professeurs) et féminins (ات) dans (مسلمات, musulmanes) ; la forme possessive (نا,كم, هم) dans (كتابهم, ses livres). [9]

II.5.3.2 Racineur de Khoja :

Le racine de Shereen khoja développé au sein de l'université de Lancaster, a été utilisé dans le cadre d'un système de recherche d'information développé à l'Université du Massachusetts . L'approche de Khoja [17] consiste à détecter la racine d'une unité lexicale, d'une part, il faut connaître le schème par lequel elle a été dérivée et supprimer les éléments flexionnels (préfixes et suffixes) qui ont été ajoutés, d'autre part comparer la racine extraite avec une liste des racines préalablement conçue. [9]

II.5.4 Bases de données lexicales :

Une base de données lexicales est un ensemble de mots (dits entrées lexicales ou données lexicales) catégorisés, c'est-à-dire associés à un certain nombre d'informations, généralement d'ordre linguistique (dites informations lexicales; par exemple, la catégorie grammaticale et les informations sur le comportement flexionnel des mots), et organisés en vue de leur utilisation par des programmes informatiques d'analyse de texte. La base de données lexicales sert ainsi de source de références et d'informations lors des procédures

d'annotation ou de catégorisation des mots d'un texte donné. On parle aussi de dictionnaire électronique.[6]

Wordnet Arabe est une base de données lexicale librement disponible pour l'arabe standard. Cette base de données suit la conception et la méthodologie du Princeton Wordnet pour l'anglais et d'EuroWordnet pour les langues européennes. Sa structure est celle d'un thésaurus, il est organisé autour de la structure des *synsets*, c'est-à-dire des ensembles de synonymes et de pointeurs décrivant des relations vers d'autres *synsets*. Chaque mot peut appartenir à un ou plusieurs *synsets*, et à une ou plusieurs catégories du discours. Ces catégories sont au nombre de quatre : nom, verbe, adjectif et adverbe. Wordnet et donc un réseau lexical dont les *synsets* sont les nœuds et les relations entre *synsets* sont les arcs. Il faut noter toutefois que Wordnet Arabe est une des rares ressources pour la langue générale arabe disponible en ligne. Il compte actuellement1 11269 *synset*set 23481 mots. [18]

II.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 (section 1), puis nous avons présenté la langue arabe et sa morphologie (section2) ,ensuite, nous avons mis le point sur les différents problèmes de traitement automatique de la langue arabe (section 3) , finalement, nous avons présenté quelques outils et produits de traitement automatique de la langue arabe .

[6] http://orthorelie.free.fr/bdlexicales.htm

Chapitre III :

Identification automatique de la cohésion lexicale

III.1 Introduction :

Détecter une cohésion lexicale entre deux phrases dans un texte, revient à trouver les relations qui peuvent lier ces dernières dans un même contexte, ainsi les relations basés sur la répétition de termes (distribution) ou celles basées sur les liens sémantiques (synonymie ou collocation) favorisent la détection automatique de la cohérence entre phrases.

Un système de détection automatique de la cohésion lexicale pourrait avoir un impact très important sur toute application de TAL nécessitant une vision allant au-delà d'une simple requête de recherche d'information, en effet, le repérage des liens entre phrases influe positivement sur les résultats retournés par des applications de segmentation thématique, de résumé automatique et de navigation intra-documentaire [19].

III.2 Cohésion :

Avant d'analyser la cohésion lexicale dans des textes écrit, nous devons tout d'abord définir le terme de cohésion (**التناسق**) pour pouvoir comprendre l'utilisation de la cohésion lexicale. Dans ce cadre, plusieurs définitions existent en littérature :

La cohésion peut être définit comme l'état des éléments de la langue qui se réconcilient et un crochet dans la sémantique [39].

Dans Larousse (2005), la cohésion est la propriété d'un ensemble dont toutes les parties sont intimement unies.

Dans la littérature arabe, le mot (**نسق** / coordoner)[7] peut avoir plusieurs sens mais qui tournent autour d'un même axe :

Par exemple, dans Lisan Alarab (**لسان العرب**) :

النَّسَقُ من كل شيء ما كان على طريقة نظام واحد عام في الأشياء.

Dans Al Mouadjam Alwassit (**المعجم الوسيط**) :

نَسَقَ الشيء نسقا أي نظمه ونسق الكلام أي عطف بعضه على بعض.

La cohésion est donc, une caractéristique linguistique qui garantit la continuité et la progression au sein du texte.

De leur part, [43] disent que la cohésion est la possibilité dans la langue pour que le texte devient une unité complexe ayant un sens à la cohésion grammaticale et lexicale.

[7] www.almaany.com/ar/dict/ar-ar/تناسق/

La signification d'une cohésion peut être vue aussi comme la relation entre la catégorie grammaticale d'un mot et sa forme lexicale.

III.3 Cohésion lexicale :

La cohésion lexicale (**التناسق المعجمي**) quant à elle peut avoir plusieurs définitions :

Selon [39] , la cohésion lexicale est la relation sémantique entre les éléments qui forment un texte en utilisant l'élément lexical élémentaire ou le mot.

D'après [44] , la cohésion lexicale concerne le lexique, le vocabulaire d'une langue.

L'utilisation d'un élément de cohésion lexicale dans un texte peut être repérée par réitération ou collocation.

La cohésion lexicale par réitération est repérée par la répétition, la synonymie et l'équivalence.

La cohésion lexicale par collocation est partagée en trois relations, l'antonymie, la métonymie et l'hyponymie.

Les marques de cohésion lexicales entre phrases peuvent être repérées alors par les six relations suivantes : Répétition, Synonymie, Hyponymie, Antonymie, Métonymie, Équivalence.

Notant que les cinq dernières relations peuvent être regroupées dans un seul ensemble, comme dans le cas de la base de données lexico-sémantique Wordnet, et dont les relations indiquées ci-dessus et bien d'autres sont regroupées dans des Synset.

Notant aussi qu'une expérience sur des textes français [20] a montré que 75.40% des relations de cohésion concernaient la répétition, 9.30% pour la synonymie et le taux restant est partagé entre les autres types de relations. Cela montre clairement que la détection de cohésion par réitération est majoritairement la plus adéquate dans un traitement automatique.

Nous pouvons dire alors que la détection de la cohésion lexicale se base principalement sur un calcul de similarité entres segments ou phrases. Selon les types de relations cités ci-dessus, la similarité peut être *distributionnelle* ou *sémantique*.

III.4 Types de similarité :

III.4.1 Similarité distributionnelle :

La similarité distributionnelle est une technique exploitée depuis longtemps pour la construction de classes sémantiques de mots. Elle consiste à rapprocher les mots sur la base de contextes qu'ils partagent, en faisant l'hypothèse que les mots les plus proches sémantiquement sont employés de manière similaire et tendent à apparaître dans les mêmes contextes.

La similarité distributionnelle et comme son nom l'indique est composée de deux concepts, la similarité et la distribution. Sachant qu'avant d'appliquer toute mesure de similarité basée sur la distribution (fréquences d'apparition) des termes dans un texte, nous devons tout d'abord mesurer ou quantifier cette distribution (ou importance), pour ce faire, plusieurs mesures existent :

III.4.1.1 Mesures de distribution :

a- TF-IDF :

Le $TF\text{-}IDF$ (de l'anglais *Term Frequency-Inverse Document Frequency*) est une méthode de pondération souvent utilisée en recherche d'information et en particulier dans la fouille de textes. Cette mesure statistique permet d'évaluer l'importance d'un terme contenu dans un document, relativement à une collection ou un corpus. Le poids augmente proportionnellement au nombre d'occurrences du mot dans le document. Il varie également en fonction de la fréquence du mot dans le corpus. Des variantes de la formule originale sont souvent utilisées dans des moteurs de recherche pour apprécier la pertinence d'un document en fonction des critères de recherche de l'utilisateur. [21]

Plusieurs fonctions existent pour le calcul de $TF\text{-}IDF$, par exemple, l'équation de [41] cité ci-dessous, permet de comparer le comportement d'un terme candidat dans le document analysé avec son comportement dans une collection de documents :

$$TF\text{-}IDF(terme) = TF(terme) \times \log\left(\frac{N}{DF(terme)}\right) \qquad \text{(III.1)}$$

TF représente le nombre d'occurrences d'un terme dans le document analysé. *DF* représente le nombre de documents dans lequel il est présent, *N* étant le nombre total de documents.

Au lieu de considérer un document par rapport à une collection, nous pouvons considérer une phrase par rapport à un texte, d'où : *TF* représentera le nombre d'occurrences d'un terme dans la phrase analysée. *DF* représentera le nombre de phrases dans lesquels il est présent, *N* étant le nombre total de phrases dans le texte.

b- OKAPI BM25 :

Okapi (ou BM25) [40] est une mesure alternative à TF-IDF. En Recherche d'Information (RI), elle est encore considérée comme l'une des méthodes à l'état de l'art dans ce domaine. Okapi est décrit comme un TF-IDF prenant mieux en compte la longueur des documents. Cette dernière est utilisée pour normaliser le *TF* (qui devient *TFBM*25) [21] :

$$Okapi(terme) = TF_{BM25}(terme) \times \log\left(\frac{N - DF(terme) + 0{,}5}{DF(terme) + 0{,}5}\right) \quad \text{(III.2)}$$

$$TF_{BM25} = \frac{TF(terme) \times (k_1 + 1)}{TF(terme) + k_1 \times \left(1 - b + b \times \frac{DL}{DL_{moyenne}}\right)} \quad \text{(III.3)}$$

*k*1 et *b* sont des constantes fixées à 2 et 0, 75 respectivement. *DL* représente la longueur du document analysé et $DL_{moyenne}$ la longueur moyenne des documents de la collection utilisée. Notant que Okapi peut aussi être adaptée pour considérer une phrase par rapport à un texte.

c- LIKEY :

Paukkeri et Honkela, [22] proposent Likey, une métrique qui sont deux méthodes qui comparent le comportement d'un terme dans le document analysé avec son comportement dans une collection de référence.

Le rang d'un terme dans le document et dans le corpus est obtenu à partir de son nombre d'occurrences, respectivement dans le document et dans le corpus de référence. Plus le rapport entre ces deux rangs est faible, plus le terme évalué est important dans le document analysé [22] :

$$Likey(terme) = \frac{rang_{document}(terme)}{rang_{corpus}(terme)} \quad \text{(III.4)}$$

III.4.1.2 Mesures de similarité :

a- Cosinus :

La mesure de similarité COSINUS est largement utilisée dans le domaine de la recherche d'information. Elle a été étendue pour prendre en compte les poids.

La mesure de COSINUS, initialement issue de l'algèbre linéaire, s'étend également aux vecteurs pondérés et est devenue la mesure standard des vecteurs pondérés dans le domaine de la recherche d'information. [23] préconise l'utilisation de la mesure COSINUS dans le domaine de la recherche d'information plutôt que le PRODUIT SCALAIRE, ou les DISTANCES DE MINKOWSKI, car COSINUS résout certains problèmes inhérents à ces deux dernières (favorisation des vecteurs longs, discrimination des vecteurs dont la différence entre les longueurs est significative, etc.).

Selon les besoins d'utilisation, plusieurs versions de Cosinus existent, parmi les versions proposées dans le calcul de similarité entre segments (ou phrases) est celle proposée par [24] qui l'utilisée dans son algorithme de segmentation thématique C99. La similarité entre deux phrases *x* et *y* est calculée par l'équation suivante :

$$sim(x,y) = \frac{\sum_j f_{x,j} \times f_{y,j}}{\sqrt{\sum_j f_{x,j}^2 \times \sum_j f_{y,j}^2}} \qquad \text{(III.5)}$$

$f_{i,j}$ dénote la fréquence du terme *j* dans la phrase *i*

b- Jaccard :

L'indice et la distance de Jaccard sont deux métriques utilisées en statistiques pour comparer la similarité et la diversité entre des échantillons. Elles sont nommées d'après le botaniste suisse Paul Jaccard . [25]

L'indice de Jaccard (ou *coefficient de Jaccard*) est le rapport entre le cardinal (la taille) de l'intersection des ensembles considérés et le cardinal de l'union des ensembles. Il permet d'évaluer la similarité entre les ensembles. Soit deux ensembles A et B, l'indice est :

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \qquad \text{(III.6)}$$

Alors que la *distance de Jaccard* mesure la dissimilarité entre les ensembles. Elle consiste simplement à soustraire l'indice de Jaccard à 1.

$$J_\delta(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} \quad \text{(III.7)}$$

c- l'information mutuelle (MI) :

La fonction de similarité qui est probablement très utilisée dans le domaine du TAL est l'information mutuelle. Il s'agit d'une information mutuelle ponctuelle qui compare la probabilité d'observer deux évènements aléatoires x et y ensemble (distribution jointe) aux probabilités de les observer indépendamment (distribution indépendante) [26]. MI est généralement calculé comme suit :

$$I(x,y) = \log\left(\frac{p(x,y)}{p(x)p(y)}\right) \quad \text{(III.8)}$$

Si l'association entre x et y est forte, la probabilité jointe $p(x,y)$ est plus importante que $p(x)p(y)$.

III.4.2 Similarité sémantique :

C'est un type de Similarité à base de connaissances est qui utilise des ressources lexicales et terminologiques. Un premier type de ressource qui peut être exploitée est l'inventaire de sens, c'est-à- dire une ressource qui, à chaque mot, lie une liste de sens possibles comme par exemple, un dictionnaire. D'autre part, des ressources telles que les thésaurus peuvent être utiles pour établir des liens entre les sens des différents mots.

Par ailleurs, des ressources lexicales telles que WordNet sont structurées et jouent le rôle d'inventaires de sens et de dictionnaires, mais donnent également accès à une hiérarchie de sens (en quelque sorte un thésaurus structuré). La majorité des mesures de similarité sémantique existantes se basent principalement sur Wordnet, du fait que c'est un outil exploitable et structuré au contraire des dictionnaires qui sont généralement fermés et leur utilisation dans des traitements automatiques reste encore très difficile. Les principales bases de données lexicales sont les suivantes :

III.4.2.1 Le réseau lexico-sémantique : Wordnet :

Wordnet est une base de données lexicale créée par une équipe linguistes et de psycholinguistes à l'université de Prinston. La version standard de WordNet est une base de données lexicale pour l'anglais. D'autres versions existent pour d'autres langues (Français, Espagnol, Allemand, Arabe,..), sachant que ces versions ont été développées indépendamment par des équipes extra-Prinston. La version arabe ou Arabic Wordnet (AWN) a été développée par une équipe de chercheurs espagnols et arabes [27] et qui est en cours d'enrichissement par d'autres équipes.

La base de données Wordnet est organisée en fonction de la signification des mots et non pas sur leur forme lexicale. L'unité de base de Wordnet et le Synset.

Il groupe des mots en jeux de synonymes appelés les synsets, Un synset est un ensemble de mots qui ont la même signification (Synset = ensemble de synonymes).

Wordnet fournit des définitions générales, et des diverses relations sémantiques entre ces ensembles de synonyme. Le but est double : pour produire une combinaison de dictionnaire et de thesaurus qui est plus intuitivement utilisable, et soutenir des applications d'analyse automatique des textes et d'intelligence artificielle.

Les Synsets sont organisés en quatre catégories, les noms, les verbes, les adjectifs et les adverbes. Les relations sémantiques entres les synsets de Wordnet sont :

- Hyperonyme (…est une généralisation de….)
- Hyponyme (... est une sorte de…).
- Troponyme (… est une manière de…).
- Meronyme (…fait partie …. Ou … membre de…)
- Holonyme (… contient …).
- Antonyme de monter en face est un antonyme de descendre
- Attribut (…est l'attribut de ….).
- Implication (…. Implique…).
- Cause (…à cause de…).
- Voir également (*verbe* liées à …) à déposer est liée à résider .
- Semblable à (…est semblable à…).
- Participe de (…est participe passé de…).
- Pertainyme (… réfère à…).

III.4.2.2 Les autres bases de données lexicales :

D'autres systèmes ont été développés pour des domaines particuliers et des langues restreintes, ces systèmes regroupent les concepts et ressources terminologiques utilisés dans un domaine donné.

- **Gene Ontologie** (The Gene Ontology Consortium, 2000) est une ressource terminologique dont l'objectif est de décrire le rôle des gènes dans les organismes (prokaryotes et eukaryotes) ainsi que leurs produits géniques.

Les termes de GO sont structurés en trois arbres hiérarchiques : processus biologiques, fonctions moléculaires et composants cellulaires. Le vocabulaire de GO est structuré à l'aide de trois types de relations : l'hypéronymie , méronymie et la synonymie .

- **UMLS** (Lindberg et al., 1993) Unified Medical Language System (UMLS) est une ressource terminologique biomédicale. Développé par la National Library of Médicine (NLM), elle regroupe plus d'une centaine de thésaurus de différentes langues dans un méta-thesaurus.

Celui-ci organise 700 000 concepts au sein d'un réseau sémantique composé de 134 types sémantiques et structuré par 54 relations sémantiques hiérarchisées par le lien is-a. Les types sémantiques associés aux termes de l'UMLS seront utilisés pour définir les contraintes Sémantiques sur les mots ou les termes pour lesquels nous avons acquis des relations de synonymie. [28]

III.5. Les travaux relatifs en langue arabe :

La plupart des systèmes de détection de la cohésion lexicale ou autrement dit étude des relations entre phrases traitent des textes des langues latines. Le peu de travaux enregistrés relativement en langue arabe, considère la cohésion lexicale comme une étape dans un processus de segmentation thématique ou de résumé automatique, la détection de la cohésion dans ces travaux est généralement statistique basée sur la répétition des termes. Les travaux recensés dans ce terme sont :

- [29] : Qui ont proposé une technique de segmentation basée sur une étude linguistique empirique. Ils ont utilisé un algorithme de détection de connecteurs passifs et actifs entre segments.
- [30], [31] et [32] ont fait une comparaison de deux algorithmes de segmentation thématique (C99 et TextTiling). Ils ont implémenté ces deux algorithmes avec deux

systèmes *ArabTiling et TopSegArab* sur des textes hétérogènes (médicaux, littéraires, scientifiques et islamiques).

• [33] et [8] : Proposaient une approche symbolique basée sur la relation entre segments noyaux et satellites en utilisant RST qui l'ont appliquée aux résumés automatiques.

• [34] : qui a implémenté la segmentation thématique de textes arabes pour la recherche d'information. Son travail est basé sur l'utilisation des méthodes probabilistes pour la détection de la cohésion lexicale.

III.6. Evaluation de la cohésion lexicale :

Évaluer un système de détection automatique de cohésion est délicat. De nombreux problèmes sont soulevés, et peuvent grosso modo être ramenés à deux questions :

(a) Quelle référence ? (b) Quel score d'évaluation ? [35]

Pour évaluer un tel système, il faut le comparer à une segmentation de référence.

III.6.1 Evaluation manuelle :

L'évaluation peut être faite manuellement en faisant appel à des annotations manuelles, mais cela retourne généralement un état d'accords inter-annotateurs très faibles. D'autres prennent le parti d'accoler bout à bout des séquences appartenant à des textes différents ; les ruptures lexicales sont alors les ruptures entre textes. L'évaluation manuelle est généralement très couteuses des évaluations automatique ou semi-automatique sont considérées comme des bonnes alternatives.

III.6.2 Evaluation semi-automatique :

Nous pouvons faire une évaluation semi-automatique en comparant les résultats produits automatiquement (segments cohérents) par le système et d'autres produits manuellement par un expert humain, dans ce cas nous faisons appel par exemple à la métrique F-mesure qui calcule les scores de rappel et de précision.

Cependant, les scores habituels de précision et de rappel ne sont pas toujours adaptés pour évaluer un système de détection de cohésion. En effet, ils ne permettent pas de rendre compte du fait qu'une rupture proche de la rupture de référence est meilleure qu'une rupture éloignée.

III.6.3 Evaluation automatique :

Pour évaluer de tel système de façon automatique, d'autres scores ont été proposés, dont les plus usités sont les mesures Pk [36] et WindowDiff [37]. La mesure Pk

consiste à compter le nombre de fois où deux mots pris au hasard à une distance k sont dans le même segment à la fois dans la référence et dans l'hypothèse. La mesure WindowDiff consiste à calculer la différence du nombre de ruptures dans une fenêtre glissante.

III.7 Conclusion :

La détection de la cohésion lexicale entre phrases a un rôle crucial dans plusieurs traitements automatiques, tel que la segmentation thématique ou le résumé automatique..etc. L'amélioration d'un tel système de détection de cohésion influe positivement les résultats rendus par les autres traitements automatiques des textes écrits.

En effet, nous pensons que la combinaison de plusieurs critères (par exemple, distributionnels et sémantiques) pour la détection des relations entre phrases améliorera majoritairement les résultats rendus par application d'un processus de détection automatique de la cohésion.

Dans ce qui suit, nous détaillerons notre approche pour la détection automatique de la cohésion lexicale entre phrases, nous travaillerons sur une collection de textes arabes non voyellé et écrits en arabe standard moderne.

Chapitre IV :

Conception du système de détection automatique de la-cohésion lexicale

IV. 1 Introduction :

Afin d'évaluer la pertinence des phrases ou de détecter les groupes de thèmes dans un texte, la détection de la cohésion lexicale constitue en effet une très bonne solution. Elle sert à identifier le maximum possible de relations entre différentes phrases afin de regrouper les plus cohérentes les unes avec les autres. Ainsi des relations distributionnelles réitératives ou sémantiques sont utilisées.

Notre objectif est de réaliser un système de détection automatique de cohésion dans les articles économiques arabes. Notre système est basé sur une combinaison de critères statistiques (distributionnels) et linguistiques (sémantiques). En effet, la plupart des méthodes et travaux existants même s'ils sont très peu relativement à la langue arabe, se basent seulement sur la répétition des termes. Nous pensons que l'ajout d'autres critères comme la synonymie, améliore tels systèmes de détection de cohésion.

IV.2 Description de l'approche :

L'approche que nous proposons pour détecter automatiquement la cohésion lexicale entre les phrases d'un texte est une approche globale, non-linéaire et hybride, elle combine un calcul statistique distributionnel (Répétition de termes) avec un modèle linguistique sémantique basé sur le calcul de la synonymie entre termes.

Notre approche est dite **globale**, parce qu'elle compare une phrase avec toutes les autres phrases du texte, au contraire des approches locales qui comparent seulement les phrases voisines.

L'approche est dite **non-linéaire**, parce qu'elle regroupe à la fin des phrases qui se situent dans des positions différentes dans le texte mais qui possèdent des liens de cohérence.

La catégorie de textes que nous étudions est une collection d'articles économiques arabes publiés dans une revue.

Pour calculer les fréquences de répétitions de termes, nous appliquons une métrique largement utilisée dans ce sens, la célèbre ***tf.idf*** qui calcul l'importance des termes dans le texte. Avant d'appliquer de telle technique sur les textes, nous devons faire une analyse

morphologique en éliminant les différents affixes dans les mots afin d'obtenir des tokens pour le calcul des occurrences des mêmes termes. Nous avons choisi comme analyseur morphologique « Arabicstemmer », un outil linguistique simple à intégrer et qui utilisent le célèbre algorithme de « Khoja stemmer ».

Les synonymes des termes sont calculés en utilisant la base de données lexicale ArabicWordnet, qui est un réseau lexico-sémantique qui regroupe pour chaque terme l'ensemble de ses synomymes dans des groupes, appelés **Synsets**.

Pour mesurer la similarité (taux de cohésion) entres les paires de phrases, nous calculons deux scores **Cosinus**, l'un mesure la similarité vis-à-vis la répétition des termes et le deuxième mesure la similarité par rapport à la distribution des synonymes des termes dans les phrases. La valeur finale de similarité est la somme des deux scores déjà calculés.

En fin, une matrice d'adjacence est construite et dont les éléments sont les valeurs de similarité Cosinus déjà calculées entre les paires de phrases. Cette matrice d'adjacence peut être représentés finalement par un graphe ***G=(S,A)***, S étant les sommets qui représentent les phrases et A les arêtes qui représentent les liens de cohésion entre les phrases.

IV. 3 Présentation du corpus :

Nous avons choisi corpus d'étude une collection d'articles économiques publiés dans la revue économique internationale « Economie contemporaine » (الاقتصاد المعاصر) éditée par la faculté des sciences économiques et de gestion de l'université Djilali Bounaâma – Khemis Miliana, Les articles sont écrits en arabe standard moderne et ils sont généralement non voyellé. Le corpus est constitué d'environ 50 articles d'une longueur moyenne de 15 pages Word. Le nombre moyen de mots par article est environ 2000 mots.

Ces articles sont convertit manuellement en .**TXT** et encodé en **UTF-8**. Ils ont été rapatriés sans restriction quant à leur contenu et leur volume. Nous estimons en effet que plus le corpus est varié, plus il sera représentatif et contiendra le plus important nombre de marqueurs et traits linguistiques.

Les articles économique et comme tout autres type d'articles spécialisés utilisent généralement un même vocabulaire dans plusieurs reprises dans le texte (favorise le calcul des racines au lieu d'une lemmatisation légère qui génère des défauts lors du calcul

des répétitions) d'où le choix d'un racineur (Arabicstemmer) comme analyseur morphologique.

D'une autre part, un même topic peut être discuté dans des segments ou phrases séparées, ce qui favorise l'utilisation d'une approche globale (comparer une phrase avec toutes les autres). La catégorie d'article traitée est caractérisée aussi par l'utilisation des termes proches pour décrire certains topics, ce qui favorise la prise en considération de la synonymie entre termes.

Avant d'effectuer toute analyse, le texte est segmenté en phrases. Cette dernière étant le segment minimal représentant un sens. Pour détecter les frontières des phrases nous nous sommes basé sur les signes de ponctuation ainsi que quelques expressions langagières qui rôle de frontière (..و، كما،ثم).

IV. 4 Les outils utilisés :

IV.4.1 L'analyseur ArabicStemmer :

Nous avons obtenu une version Java simplifiée de ShereenKhoja'sstemmer qui est Arabicstemmer. Arabicstemmer supprime le plus long suffixe et le préfixe. Il correspond alors le mot restant avec des motifs verbaux et nominaux, pour extraire la racine. Le stemmer utilise plusieurs fichiers de données linguistiques comme une liste de tous les caractères diacritiques, caractères de ponctuation, des articles précis, et 168 mots d'arrêt.[38]

Arabicstemmer suit cette procédure:

1. Retirer diacritiques représentant vocalisation.

2. Retirer stopwords, la ponctuation et des chiffres.

3. Retirez article défini "لا"

4. Retirer conjointement inséparables "و".

5. Retirer suffixes.

6. Retirer préfixes.

7. Résultat du match contre une liste de motifs. Si une correspondance est trouvée, puis extraire les caractères dans le motif représentant la racine.

8. Faites correspondre la racine extraite à une liste de racines connues "valides".

9. Remplacer les lettres faibles " ا,و,ي " avec "و"

10. Remplacer toutes les occurrences de hamza "ؤ, ئ, ء" avec "أ"

11. Deux racines lettre sont vérifiées pour voir si elles doivent contenir un double caractère. Le cas échéant, le caractère est ajouté à la racine. [1]

- Comment l'analyseur de Khoja analyse le mot : يشكركم?

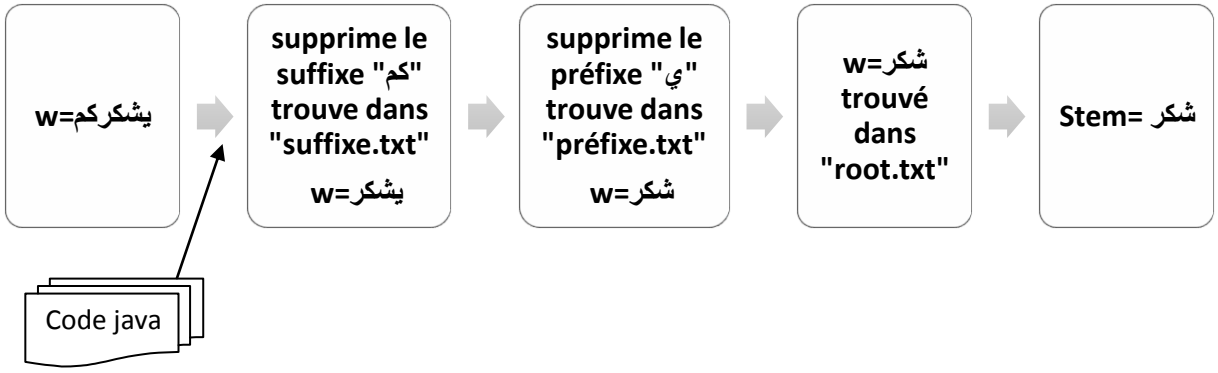


Figure IV.5: Analyse du mot - يشكركم -

Donc il trouve que le mot يشكركم il composé d'un préfixe, racine et suffixe :

يشكركم = كم + شكر + ي

IV.4.2 ArabicWordNet (AWN) :

L'ArabicWordNet est une base de données lexicale. Sa conception basé sur Princeton WordNet est construite suivant des méthodes développées pour EuroWordNet est reliée avec l'ontologie SUMO (SuggestedUpperMergedOntology). ArabicWordNet a été développé par DOI / REFLEX (2005-2007) . [39](voir Figure IV.6)

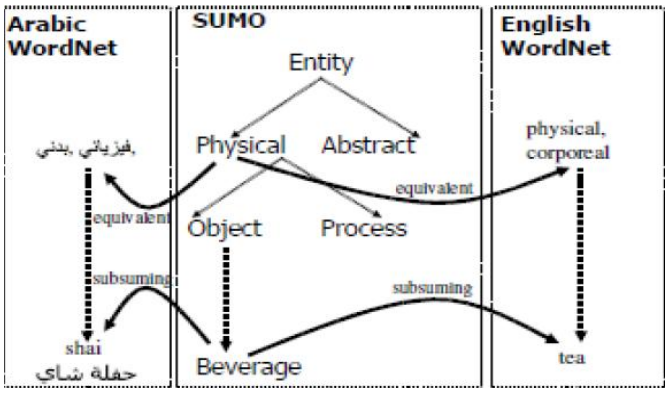


Figure IV.6 : Mapping de SUMO vers WordNet(s) (Structure et organisation de l'AWN)

SUMO est en pleine expansion afin d'offrir un fondement solide pour la formalisation sémantique de l'ArabicWordNet (AWN).

AWN ne vise pas à prendre tout texte arabe comme une entrée directement. Cependant, nous devons translittération convertir le texte du système arabe d'être nourri comme une entrée au système. Les résultats doivent être transcrits reconvertis en arabe à être compris. Cette technique a été introduite par Buckwalter (2002) , par exemple la lettre "و" est translittération pour "w". La lettre "ؤ" est translittération à "&". Tableau 1, montre quelques lettres arabes et leur translittération en Buckwalter .

الشدة	~	ألف ممدودة	|
السكون	o	ألف الوصل	{
حرف الباء	b	ألف مقصورة	Y
حرف العين	E	الفتحة	a
حرف الفاء	f	الضمة	u
		الكسرة	i

Tableau IV.16 :Représentation des lettres arabes dans Buckwalter

La base de données AWN est librement et publiquement disponible. L'ontologie ArabicWordNet contient 9228 concepts « Synsets » (6252 nominales et 2260 verbales, 606 adjectival, et 106 adverbales), contient 18,957 expressions et 1155 concepts nommés le fichier base de l'AWN sous format XML 2 contient les quatre balises :

• **Item** : Contient les concepts (Synsets), les classes et les instances de l'ontologie.

• **Word** : Contient les mots arabes vocalisés.

• **Form** : Contient les Racines des mots arabes « root ».

• **Link** : Contient les relations entre les concepts.

➢ L'utilisation de la base de données AWN :

On a les deux balises :

➢ <word wordid="$aAb_1" value"**شارِي**" synsetid="$aAb_n1AR" frequency="1" corpus="manchester20060717" authorshipid="11296" />

➢ <form value="**شري**" wordid="$aAb_1" type="root" authorshipid="11296" />

1. Récupérer le "wordid" du mot à partir de la balise "**form**"

2. Chercher le "synsetid" de ce mot dans la balise "**word**"

3. Récupérer toutes les balises "**word**" où le "synsetid" apparait

4. récupérer le "wordid" de ces dernières

5. Chercher le "wordid" dans la balise "**form**" puis prendre la valeur de "value"

Ou :

1. (à partir de 2) Récupérer les balises "**link**" dont le "synsetid" est cité

2. Récupérer les **link** de la même ligne et répéter 2.

Donc les balises nécessaires pour calculer la synonymie sont :

- ✓ Form
- ✓ Word
- ✓ link

⟹ (Type="has_hyponym" ou type="near_synonym")

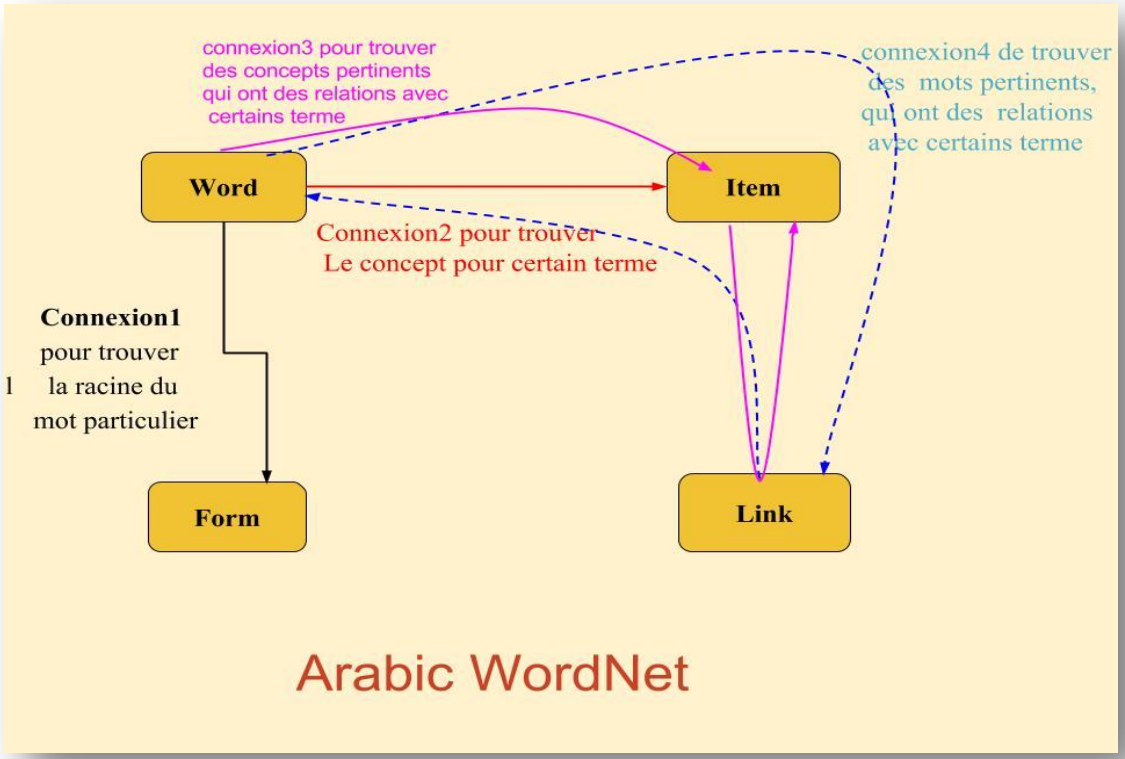


Figure IV.7: Rehcerche des synonymes dans AWN .xml

IV.5 Architecture du système :

Notre système de détection automatique de cohésion lexicale est basé principalement sur des techniques d'extraction. La mise en œuvre fonctionnelle du système est représentée par la figure IV.8.

Afin de calculer le score de similarité (cohésion) entre les paires de phrases dans un texte (article), ce dernier passe par une série de traitements linguistiques et statistiques.

Nous décrivons ci-après les modules selon la numérotation de la figure IV.8.

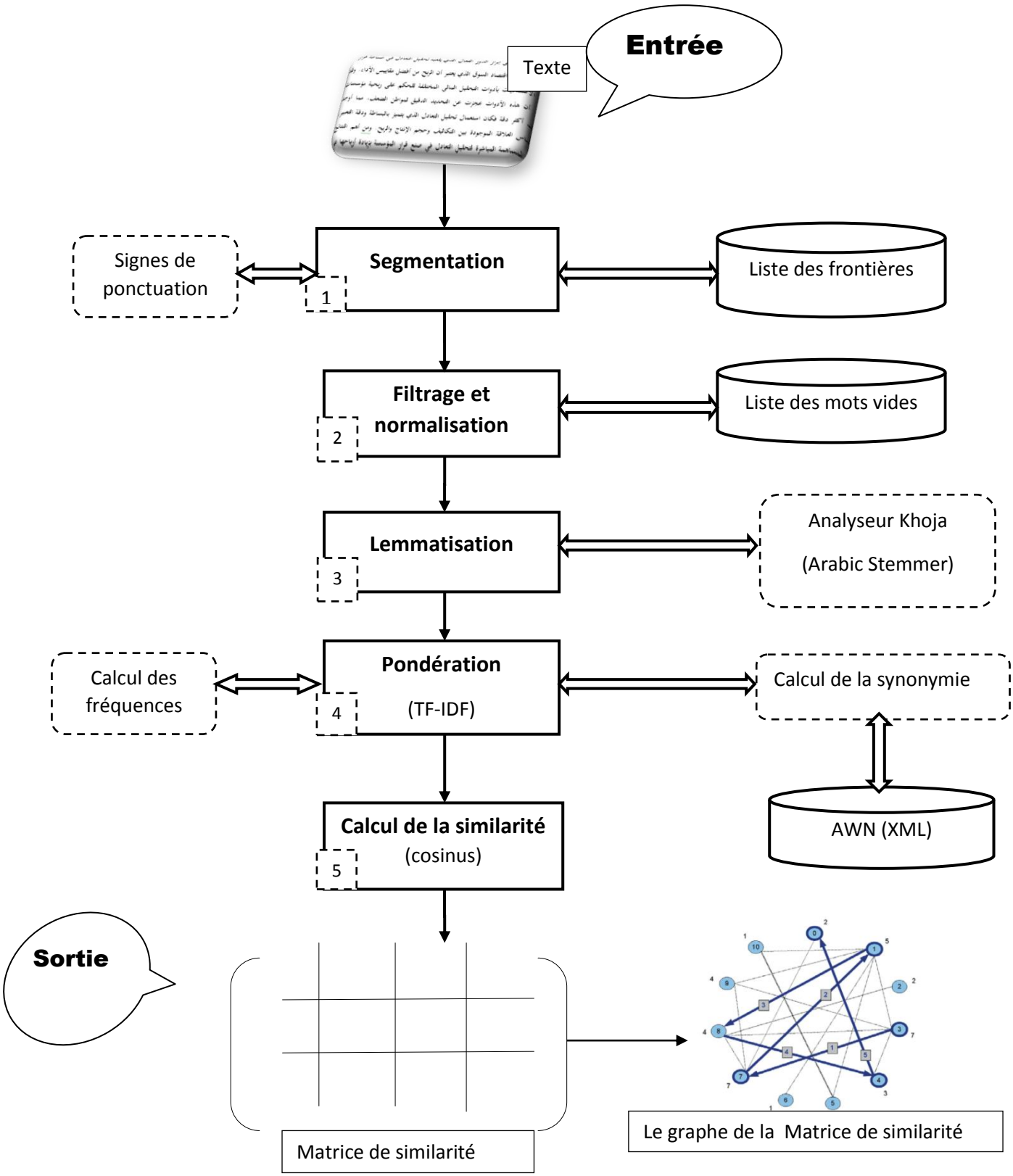


Figure IV.8: Architecture du système.

IV.6 Description des principaux modules du système :

IV.6.1 Segmentation:

Ce module consiste à segmenter le texte en phrases en utilisant les signes de ponctuation et des expressions langagières jouant le rôle de frontières. La liste des signes de ponctuation utilisée est celle définit par [42], cette liste est enrichi par d'autres signes spécifiques à l'arabe, tel que :

- la virgule renversée « ، ».
- le point-virgule renversé « ؛ ».
- le point d'interrogation inversé « ؟ ».

La liste des frontières langagières utilisée est celle définit par [33], cette liste contient les propositions comme (…,و ,ثم ,أو). Afin de standardiser les limites des phrases et lors du parcours du texte, toutes frontières rencontrée seront remplacées par le signe « | ».

Les éléments de la liste des frontières, ne jouent pas toujours un rôle de frontière. Pour cette raison nous proposons deux solutions, la premières consiste à définir une longueur minimale de phrases, nous utilisons pour cela, la longueur défini par [46] et qui est fixée à 6 termes. Notant que vis-à-vis notre corpus d'étude cette longueur reste expérimentale et peut être ajustée au fur et à mesure des tests.

La deuxième solution et particulièrement pour remédier aux ambiguïtés générées par le (و/et), consiste à utiliser un petit dictionnaire des mots indécomposables comme dans [6]. En effet, lors de la rencontre d'un mot qui commence par (و) on le compare avec la liste des mots indécomposables, s'il en fait partie on ne marque pas sa position par signe de frontière.

IV.6.2 Filtrage et normalisation :

Cette étape consiste à éliminer tous les mots non significatifs. Pour chaque mot reconnu, on le compare avec un des éléments dans la liste des mots vides ou « Stop words » comme (كان, بعد, أن,…), si un mot en fait partie en le supprime.

Cette étape consiste aussi à supprimer les caractères spéciaux et les chiffres et l'enlèvement des diacritiques dans le cas des mots partiellement voyellés.

Finalement et afin de préparer les mots au prochain traitement (lemmatisation), ces

derniers sont normalisés en effectuons les remplacements des lettres suivantes :

- Remplacer le إ ou le أ initial par l'alif nu ا.
- Remplacer le آ par le ا.
- Remplacer le ىء d'ordre par le ى
- Remplacer le ي final par le ى
- Remplacer le ة final par le ه

IV.6.3 Lemmatisation :

Pour un mot significatif normalisé, on applique une lemmatisation profonde en utilisant l'analyseur *ArabicStemmer* qui consiste à détecter le stem (racine) d'un mot et supprimer les éléments flexionnels (préfix et suffixes), ceci permet de retourner une liste d'items appelés *tokens* . Cette étape est en effet très importante pour le calcul de distribution des mêmes termes

IV.6.4 Pondération :

Cette phase consiste à associer un score à chaque terme en fonction de la distribution de ses occurrences et de ses synonymes dans le texte. Le score distributionnel répétitif est calculé par la métrique *tf_idf.* Le score de synonymie est quantifié en utilisant encore une fois *tf_idf* adaptée au calcul de distribution des synonymes d'un terme.

IV.6.4.1 Calcul des fréquences pondérées :

Notant que *tf_idf* dans sa version standard (voir chap. 3) calcule le poids (fréquence pondérée) d'un terme dans un document par rapport à une collection de documents. Cette métrique peut être adaptée pour considérer un terme d'une phrase par rapport à une collection de phrases (texte).

$$TF\_IDF_{dist} = TF(stem) \times \log \frac{N}{DF(stem)} \qquad \text{(IV.9)}$$

TF :représente le nombre d'occurrences d'un stem dans la phrase analysé.

DF :représente le nombre de phrases dans lequel il est présent.

N : étant le nombre total de phrases dans le texte.

Ce score présente l'importance d'un terme en fonction de ses répétitions dans le texte, comme le montre la figure suivante :

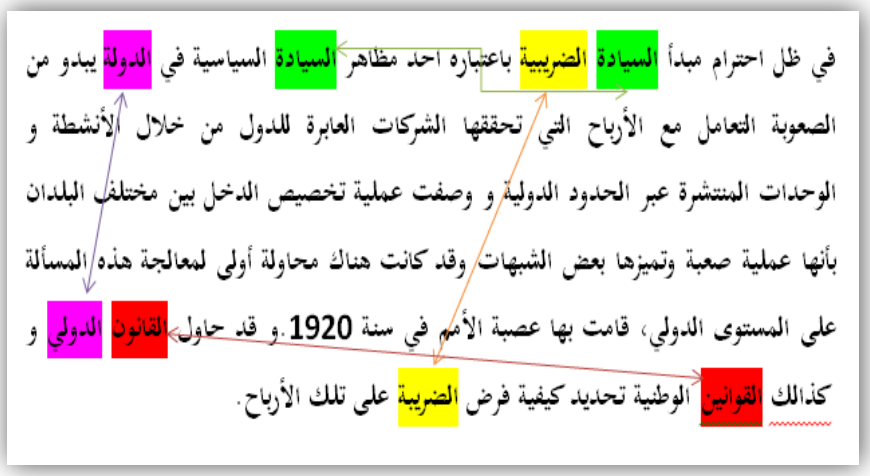


Figure IV.9:Liens de répétitions

IV.6.4.2 Calcul de la synonymie :

Après avoir extraire la liste des synonymes d'un terme à partir de la base de données lexicale AWN, un score de synonymie ($TF.IDF_{syn}$) est calculé en fonction de la distribution des synonymes de ce dernier. Dans ce cas nous calculons le nombre de synonymes d'un terme dans la phrases et le nombre de phrases ou apparait des synonymes de ce dernier.

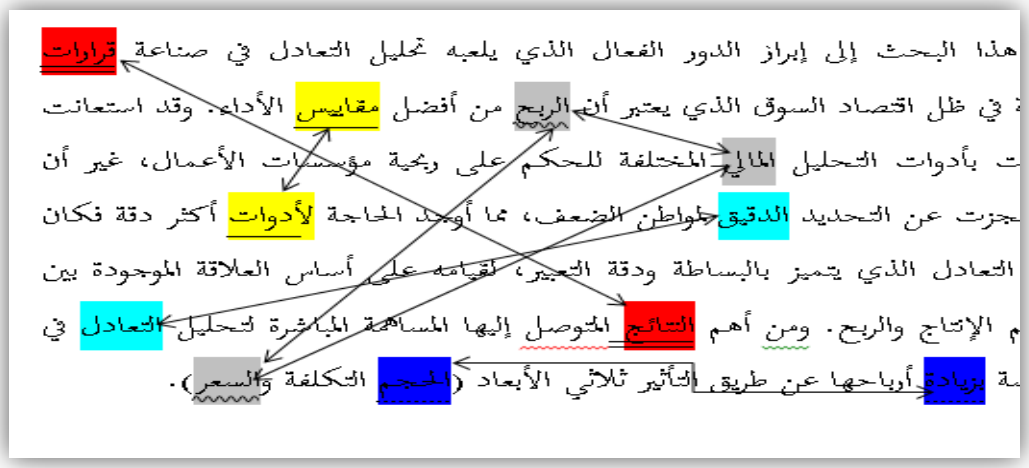


Figure IV.10 :Liens de synonymie

Le score $TF.IDF_{syn}$ quantifie l'importance d'un terme par rapport à la distribution de ses synonymes dans le texte, il est calculé comme suit :

$$TF\_IDF_{syn} = TF(syn) \times \log\left(\frac{N}{DF(syn)}\right) \qquad \text{(IV.10)}$$

IV.6.4.3 Score final :

Le score tf.idf final est la somme des deux scores, distributionnel répétitif $tf.idf_{dist}$ et celui de la synonymie $tf.idf_{syn}$.

$$TDF\_IDF_{final} = TF\_IDF_{dist} + TF\_IDF_{syn} \quad \text{(IV.11)}$$

IV.6.5 Calcul de la similarité :

Cette étape consiste à calculer la similarité entre les paires de phrases en utilisant la mesure Cosinus (voir chap. 3). La version standard de Cosinus calcule la similarité sur la base des fréquences d'apparition des termes d'une paire de phrases. Nous proposons une nouvelle version de Cosinus qui prend en considération les fréquences pondérées (tf.idf) au lieu de simples fréquence *f*. Cette nouvelle version mesure non seulement le nombre de relations répétitives entre phrases mais aussi l'importance de ces relations. Pour un système qui calcule la pertinence des phrases vis-à-vis les relations, une telle mise à jour de Cosinus améliore fortement les résultats.

L'équation du Cosinus mise à jour qui calcule la similarité entre une paire de phrases (x,y) est calculée par l'équation suivante :

$$cosinus(x,y) = \frac{\sum_{j \epsilon x \cap y} TF\_IDF_{x,j} \times TF\_IDF_{y,j}}{\sqrt{\left(\sum_{j \epsilon x} {TF\_IDF_{x,j}}^2\right) \times \left(\sum_{j \epsilon y} {TF\_IDF_{y,j}}^2\right)}} \quad \text{(IV.12)}$$

Après avoir calculé Cosinus entre les paires, une matrice de similarité SIM est construite. Cette dernière est une matrice d'adjacence carrée et dont les lignes et les colonnes sont les phrases (rangs des phrases dans le texte). Un élément SIM[i,j] représente la valeur de similarité Cosinus entre les deux phrases i et j.

	1	2	3	……	n
1	SIM[1,1]	SIM[1,2]	SIM[1,3]	…	SIM[1,n]
2	SIM[2,1]	SIM[2,2]	SIM[2,3]	. . .	SIM[2,n]
3	SIM[3,1]	SIM[3,2]	SIM[3,3]		SIM[3,n]
……					
n	SIM[n,1]				SIM[n,n]

Tableau IV.17 : Matrice de similarité

Cette représentation vectorielle de la matrice d'adjacence peut être illustrée par un graphe ***G=(S,A)***, S étant les sommets qui représentent les phrases et A les arêtes qui représentent les liens de cohésion entre les phrases, une arête entre les deux phrases i et j est tracée si SIM[i,j]≠0.

Afin d'évaluer le degré de la cohésion entre une phrase i et les autres phrases [i+1..n], nous calculons un seuil de cohésion *SC*, ainsi un groupe de cohésion de la phrase i regroupe seulement les phrases j ayant une valeur de similarité SIM[i,j] ≥ *SC*.

Un seuil de cohésion pour une phrase i, SC_i égal à la moyenne des valeurs de similarité de la phrase i avec les autres phrases. SC_i est calculé par l'équation suivante :

$$SC_i = \frac{\sum_{j=1}^{n} SIM[i,j] - SIM[i,i]}{n-1} \qquad \text{(IV.13)}$$

Afin de réduire le nombre d'arêtes dans le graphe *G(S,A)* et ne garder que les relations de cohésion les plus importantes, une arête entre les deux phrases i et j est tracée si**SIM[i,j] ≥*SC*$_i$**

Les groupes de cohésion résultants auront par exemple la forme suivante :

[1,2,5,7], [2,1,3], [3,2,5,8,11],…

Le premier groupe peut être interprété ainsi : la première phrase est plus liée avec la deuxième, la cinquième et la septième phrase qu'avec les autres phrases du texte.

IV.7 Implémentation :

IV.7.1 Le langage de programmation utilisé : Java :

Pour développer notre système de détection automatique de cohésion lexicale entre les phrases dans un texte arabe, nous avons utilisé le langage Java avec son environnement de développement NetBeans.

NetBeans est un environnement de développement intégré (EDI), placé en open source par Sun en juin 2000 sous licence CDDL et GPLv2 (Common Développent and Distribution License). En plus de Java, NetBeanspermet également de supporter différents autres langages, comme Python, C, C++, JavaScript, XML, Ruby, PHP et HTML. Il comprend toutes les caractéristiques d'un IDE moderne (éditeur en couleur, projets multi-langage, refactoring, éditeur graphique d'interfaces et de pages Web).

Le langage Java est actuellement le langage le plus utilisé pour développer des applications à fins linguistiques. La plupart des outils existants sont écrit en Java.

Java est caractérisé par sa puissance vis-à-vis le traitement des flux de données textuelles, sa facilité d'intégration des bibliothèques externes et sa pratique de déploiement.

Même si ces dernières années, les langages à script tel que PERL ou PYTHON cherche à prouver leur présence dans le champ des applications informatiques linguistiques, Java est toujours leader.

IV.7.2 Les Bases de données utilisés :

- APACHE DERBY :

Afin de stocker et gérer les résultats de l'exécution de notre système et aussi d'utiliser quelques dictionnaires (mots vides, frontières,..), nous avons créé une base de données Apache Derby qui peut être facilement embarquée avec l'application dans le cas d'un déploiement.

Apache Derby est un environnement intégré dans Java NetBeans, il permet de créer et de maintenir les bases de données relationnelles.

- AWN .XML

Une base de données .xml, générée à partir de AWN. Le fichier .xml généré sous forme d'un thésaurus regroupant les concepts et les relations entre ces concepts.

IV.7.3 JARs utilisés :

- Traitement des chaines et des flux de données :

  - Commons-io-1.3.2.jar
  - Commons-lang-2.4.jar

- Création et gestion des bases de données :

  - Java DB Driver-derby.jar
  - Java DB Driver-derbyclient.jar
  - Java DB Driver-derbynet.jar

- Création et visualisation des graphes : graphstream

  - gs-core-1.2.jar
  - gs-algo-1.3.jar
  - gs-ui-1.3.jar

IV.7.4 Test de l'application :

Afin de détecter les relations de cohésion entre les phrases d'un texte (article), l'utilisateur utilise une petite interface lui permettant d'entrer des textes pour traitement. Les principales interfaces du système sont les suivantes :

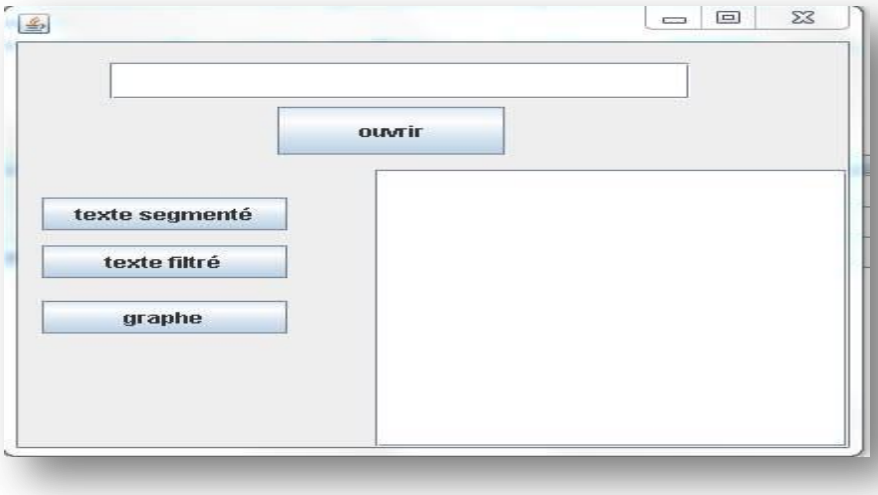


Figure IV.11: La fenêtre de Interface graphique

Figure IV.12: afficher l'emplacement de fichier.

Figure IV.13 : texte filtré affiché sous forme html

IV.8 Proposition d'une approche d'évaluation

Nous pensons que l'évaluation la plus adéquate avec notre système est celle faite d'une façon semi-automatique en calculant les métriques Rappel et Précision afin de valoriser la qualité des résultats de notre système. L'évaluation proposée consiste à :

- Présenter une collection de textes (e.g : 10 articles) préalablement segmentés par le système (phrases séparées) à un expert humain (spécialiste en économie contemporaine) et lui demander de regrouper avec chaque phrase les phrases les plus liées avec cette dernière afin de construire des groupes cohérents.
- Les groupes de phrases construits par l'expert humain sont comparés avec ceux produits automatiquement par le système. Pour chaque groupe de phrases on doit calculer les phrases correctement groupées, les phrases incorrectement groupées et les phrases oubliées. Cette technique est originalement utilisée pour évaluer des résumés automatiques [47] mais elle peut être facilement adaptée à diverses applications.

Les mesures Rappel, Précision sont calculées par les équations suivantes :

$$Rappel = \frac{Correctes}{Correctes+Oubliées} \quad \text{(IV.14)}$$

$$Précision = \frac{Correctes}{Correctes+Incorrectes} \quad \text{(IV.15)}$$

Correctes: Nombre de phrases correctement groupées par le système et par l'expert humain.

Incorrectes : Nombre de phrases groupées par le système et non pas par l'expert humain.

Oubliées : Nombre de phrases groupées par l'expert humain et non pas par le système

Afin de donner une valeur d'évaluation globale du texte, nous devons calculer la moyenne des valeurs Rappel et Précision de tous les groupes de phrases du texte. L'évaluation finale (F-Mesure) est calculée par l'équation suivante :

$$F - mesure = \frac{2.(Rappel.Précision)}{Rappel+Précision} \quad \text{(IV.16)}$$

Les groupes sont qualifiés informatifs et cohérents si la valeur Rappel t'envers 1 (Rappel ≈ 1), les regroupements sont dit précis si (Précision ≈ 1).

IV.9 Conclusion :

Dans ce chapitre, nous avons présenté l'approche utilisée pour concevoir et développer notre système de détection automatique de la cohésion lexicale dans les textes arabes, en décrivant les outils utilisés et l'architecture du notre système avec une description des différents modules de traitement. Nous avons testé notre système sur des articles économiques. L'approche que nous avons définie peut être généralisée pour d'autres catégories de texte en prenant en considération la particularité du vocabulaire et des relations qui peuvent exister dans des textes d'un domaine particulier (biologie, sociologie,..).

Conclusion générale

Actuellement, la détection automatique de la cohésion lexicale se base principalement sur les répétitions des termes ayant des structures morphologiques proches. Les nouveaux processus de détection visent à être plus intelligents en prenant en considération le sens des termes au lieu de considérer de simples chaines de caractères. Le travail que nous avons réalisé s'inscrit dans ce cadre et visait comme objectif de valoriser non seulement les relations qui peuvent exister entre les phrases mais aussi de valoriser l'importance de ces relations.

Par ailleurs, comme amélioration de ce travail, nous proposons l'étude d'autres relations telle que la collocation par exemple. Nous pensons que l'utilisation d'un analyseur morphosyntaxique puissant améliorera certainement les résultats. En effet, la détection des synonymes d'un terme sera plus adéquate si ce dernier est complètement ou partiellement voyellé.

Afin de faire une évaluation adéquate d'un système de détection de cohésion, nous prévoyons compléter notre travail par une approche de segmentation thématique linéaire ou hiérarchique.

Le travail réalisé nous a permis de découvrir l'importance de l'utilisation des outils impliqués dans le traitement automatique et spécialement pour le traitement de la langue arabe.

Nous espérons finalement par le modeste travail réalisé, apporter une contribution significative aux travaux liés au traitement de la langue arabe.

Bibliographie

[1] Benoît TROUVILLIEZ ,Traitement Automatique des Langues (TAL), Intelligence Artificielle (IA), Analyse sémantique et Clusterings , 31 mars 2010.

[2] Claire Gardent ,Traitement des Langues Naturelles (TAL) , Septembre 2011, ENS Cachan.

[3] J. Véronis : Informatique et linguistique, Support du cour pour licence en sciences du langage, Centre Informatique pour les Lettres et Sciences Humaines, Université de Provence, France, 2001.

[4] François Yvon, Une petite introduction au Traitement Automatique des Langues Naturelles , Support du cour.

[5] G. Sabah , L'intelligence artificielle et le langage, Hermès, Paris, 1998.

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

[7] Damien Nouvel , Nathalie Friburger, Jean-Yves Antoine, Traitement Automatique des Langues pour les Systèmes d'Information, support du cours pour Master SIAD–M2, Université FRANCOIS–RABELAIS .

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

[9] 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 en informatique, soutenue le 18 octobre 2008,Université de Nantes, France,.

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

[11] Souhir Gahbiche-Braham , Amélioration des systèmes de traduction par analyse linguistique et thématique Application à la traduction depuis l'arabe , thèse de doctorat en

informatique, soutenue le 30 Septembre 2013, École Doctorale d'Informatique, Université Paris Sud.

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

[13] Amine Chennoufi et Azzeddine Mazroui, Méthodes de lissage d'une approche morpho-statistique pour la voyellation automatique des textes arabes , article, 2014.

[14] K.Z Bousmaha [1,2], S. Charef_Abdoun[1], L. Hadrich_Belguith[2], M.K Rahmouni[1] , Une approche de désambiguïsation morpho_lexicale évaluée sur l'analyseur morphologique Alkhalil, article, [1]Université d'Oran, Faculté des sciences, Laboratoire RIIR, Algérie [2]ANLP Research Group, Laboratoire MIRACL, Université de Sfax, Tunisie, 2013.

[15] R. ABBES. La conception et la réalisation de concordancier électronique pour l'arabe. Thèse de Doctorat, Institut national des sciences appliquées de Lyon, Lyon, France, 2004.

[16] O. ALJLAYL, M.AND FRIEDER. On arabic search: Improving the retrieval effectiveness via a light stemming approach. In 11 the International Conference on Information and KnowledgeManagement (CIKM), pages 340–347, Virginia, USA, 2002.

[17] A. ARAMPATZIS, T. TSORIS et C. H. KOSTER. irena : Information retrievalengine based on natural language analysis. Rapport technique, Computing Science Institute, Nijmegen, Pays-Bas, 1996.

[18] Mohammed El Amine Abderrahim , Vers une interface pour l'enrichissement des requêtes en arabe dans un système de recherche d'information , thèse de doctorat en informatique , soutenue le 08 Juillet 2008 , Faculté des sciences de l'ingénieur , Université Abou Bekr Belkaid Tlemcen, Algérie.

[19] PÉRY-WOODLEY & SCOTT, Eds. (2006). Discours et Document : traitements automatiques. Numéro thématique, volume TAL 47

[20] Geri Onanda Sembiring Pandia : l'analyse des éléments de la cohésion lexicale dans les écritures des étudiants de la section française , 2010.

[21] Adrien Bougouin, État de l'art des méthodes d'extraction automatique de termes clés, TALN-RÉCITAL 2013, 17-21 Juin, Les Sables d'Olonne.

[22] PAUKKERI, M. et HONKELA, T. Likey : Unsupervised Language-Independent Keyphrase Extraction. In Proceedings of the 5th International Workshop on Semantic Evaluation, 2010.

[23] WITTEN, I., PAYNTER, G., FRANK, E., GUTWIN, C. et NEVILL-MANNING, C. : KEA : Practical Automatic Keyphrase Extraction. In Proceedings of the 4th ACM conference on Digital libraries. 1999

[24] Fred Y. Y. Choi, Advances in domain independent linear text Segmentation, Proceeding of NAACL-00, pp 26–33, 2000.

[25] Paul Jaccard (1901) : Bulletin de la Société Vaudoise des Sciences Naturelles 37, 241-272.Tanimoto, T.T. (1957) IBM Internal Report 17th Nov. 1957. Article Wiki : URL : https://fr.wikipedia.org/wiki/Indice_et_distance_de_Jaccard , consulté le : 02/05/2016

[26] Amandine Périnet, Thierry Hamon : Analyse et proposition de paramètres distributionnels adaptés aux corpus de spécialité, JADT 2014 : 12es Journées internationales d'Analyse statistique des Données Textuelles.

[27] Rodriguez, H., Black, W., Elkateb, S., Alkhalifa, M., Vossen, P., Pease, A. and Fellbaum, C., (2006). Introducing the Arabic WordNet Project, in Proceedings of the Third International WordNet Conference, Sojka, Choi, Fellbaum and Vosseneds.

[28] Mounira Manser, État de l'art sur l'acquisition de relations sémantiques entre termes : contextualisation des relations de synonymie, 2012.

[29] A. A. Touir, H. Makhtour, and W. Al-Sanea, "Semantic-Based Segmentation of Arabic Texts", Inf. Tech. J., 7(7)(2008), pp. 1009–1015.

[30] Harrag .F, Benmohammed .M, Étude Comparative des Algorithmes de Segmentation Thématique Pour la Langue Arabe , 2009.

[31] F. Harrag, A. Hamdi-Cherif, and A. S. Al-Salman, Applying Topic Segmentation Algorithms on Arabic Language, to be presented at The 7th ACS/IEEE Int. Conf. on Comp. Syst. and Appl. (AICCSA-2009), Rabat, Morocco, May 10–13, 2009.

[32] F. Harrag, A. Hamdi-Cherif, and A. S. Al-Salman, Comparative study of topic segmentation Algorithms based on lexical cohesion: Experimental results on arabic language, The Arabian Journal for Science and Engineering, Volume 35, Number 2C. P. 183-202. December 2010.

[33] KESKES I. Resumé automatique de textes arabes base sur une approche symbolique. 2011. Editeur : EUE. ISBN-13 : 978-3841780232

[34] Berrakem .F .Z, Segmentation thématique de textes arabes pour la recherche d'information, Thèse de doctorat, 2013. Université de Tlemcen

[35] Clémentine Adam, François Morlane-Hondère. Détection de la cohésion lexicale par voisinage distributionnel : application à la segmentation thématique. RECITAL'09, Jun 2009, Senlis, France.pp.article 6, 2009.

[36] BEEFERMAN D., BERGER A. & LAFFERTY J. (1999). Statistical models for text segmentation. Mach. Learn., 34(1-3), 177–210.

[37] PEVZNER L. & HEARST M. A. (2002). A critique and improvement of an evaluation metric for text segmentation. Computational Linguistics, 28, 1–19.

[38] Majdi Sawalhaet Eric Atwell "Comparative Evaluation of Arabic Language Morphological Analysers and Stemmers" School of Computing, University of Leeds, Leeds LS2 9JT, UK, Manchester, August 2008.

[39] Abdel Hamid Kreaa, Ahmad S Ahmad and KassemKabalan 'ARABIC WORDS STEMMING APPROACH USING ARABIC WORDNET' College of Information Engineering, Tishreen University, Latakia, Syria. International Journal ,November 2014.

[39] Kushartanti, et al. 2005. Pesona Bahasa Langkah Awal Memahami Linguistik. Jakarta :Gramedia Pustaka Utama .

[40] ROBERTSON, S. E., WALKER, S., BEAULIEU, M. et WILLETT, P. : Okapi at TREC-7 : Automatic Ad Hoc, Filtering, VLC and Interactive Track. 1999.

[41] JONES, K. : A Statistical Interpretation of Term Specificity and its Application in Retrieval. 1974.

[42] MOURAD G. (1999). La segmentation de textes par l'étude de la ponctuation; Acte Nde colloque international, CIDE'99, Document Electronique Dynamique, pp. 155-171, Damas, Syrie.

[43] Halliday, M.A.K et Ruqaiya Hasan. 2002, Cohesion in English Revision VI. London : Longman.

[44] Larousse. 2005. Le Petit Larousse. Illustré (100r) édition. Paris : Hachette

[45] Slim MESFAR , Analyse morpho-syntaxique automatique et reconnaissance des entités nommées en arabe standard, thèse de doctorat en informatique, soutenue le 24 novembre 2008, Université de FRANCHE-COMTE.

[46] M. A HEARST, Text-tiling : segmenting text into multi-paragraph subtopic passages, Computational Linguistics ,1997.

[47] J.M Torres-Moreno, Résumé automatique de documents, une approche statistique, Lavoisier, septembre 2011.

[48] G. A. Kiraz : Analysis of the Arabic Broken Plural and Diminutive, In Proceedings of the 5th International Conference and Exhibition on Multi Lingual Computing (ICEMCO96), Cambridge, UK.

Annexe A

Tableau des translittérations de l'alphabet arabe **[45]**

Lettre	Nom	Translittération	Commentaires
ء	*hamzä*	'	
ا	*álif*	ā	Voyelle longue (*)[54]
ب	*Bâ'*	b	Lettre lunaire
ت	*Tâ'*	t	Lettre solaire
ث	*Tâ'*	t̲	Lettre solaire
ج	*Jîm*	ǧ	Lettre lunaire
ح	*ḥâ'*	ḥ	Lettre lunaire
خ	*Xâ'*	h̲	Lettre lunaire
د	*dâl*	d	Lettre solaire (*)
ذ	*d̲âl*	d̲	Lettre solaire (*)
ر	*Râ'*	r	Lettre solaire (*)
ز	*zây*	z	Lettre solaire (*)
س	*Sîn*	s	Lettre solaire
ش	*Sîn*	š	Lettre solaire
ص	*ṣâd*	ṣ	Lettre solaire
ض	*ḍâd*	ḍ	Lettre solaire
ط	*ṭâ'*	t	Lettre solaire
ظ	*ẓâ'*	ẓ	Lettre solaire
ع	*'ayn*	'	Lettre lunaire
غ	*ġayn*	ġ	Lettre lunaire

ف	*Fa'*	f	Lettre lunaire
ق	*qāf*	q	Lettre lunaire
ك	*kaf*	k	Lettre lunaire
ل	*Lām*	l	Lettre solaire
م	*mīm*	m	Lettre lunaire
ن	*nūn*	n	Lettre solaire
ه	*Ha'*	h	Lettre lunaire
و	*wāw*	w / ū	Lettre lunaire (*) / Voyelle longue
ي	*Ya'*	y / ī	Lettre lunaire / Voyelle longue
ـَ	*fatḥa*	a	Voyelle brève
ـُ	*ḍamma*	u	Voyelle brève
ـِ	*kasra*	i	Voyelle brève
ـً	*tanwīn*	ã / an	Voyelle brève / Tanwîn
ـٌ	*tanwīn*	ũ / un	Voyelle brève/ Tanwîn
ـٍ	*tanwīn*	ĩ / in	Voyelle brève/ Tanwîn
ـة	*tā' marbūṭā*	ā (at en annexion)	
ى	*Alif maqṣūra*	ą	
آ	*Alif mamdūda*	â	
أ	*hamza*	á	
ؤ	hamzā	ù	
إ	*hamza*	í	
ئ	*hamza*	ẁ	
ئ	*hamza*	ý	
ـْ	*sukun*		non transcrit
ـّ	*šadda*	lettre redoublée	Signe de gémination