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 :

L'automatisation de la segmentation thématique de textes arabes

« Une Approche globale, non-linéaire, non supervisée »

Réalisé par :

MAROUF ARAIBI Mohammed el amine

ABDELMOUMENE Imene

Devant le jury composé de :

Mr S.HADJ SADOK……Président

Mr D.BAHLOUL…….....Encadreur

Mr I.BOUKADOUM..Examinateur1

Mr M.BOUZIANE…..Examinateur2

Année Universitaire : 2016/2017.

Remerciements

Nous tenons tout d'abord à remercier Dieu le tout puissant et miséricordieux, qui nous a donné la force et la patience d'accomplir ce Modeste travail.

Nous tenons à remercier notre encadreur Monsieur « Djamel Bahloul », pour l'orientation, la confiance, la patience qui a constitués un apport considérable sans lequel ce travail n'aurait pas pu être menée à bon port, Qu'il trouve dans ce travail un hommage vivant à sa haute personnalité.

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

Nous tenons à exprimer nos sincères remerciements à tous les profs de département mathématique et informatique.

Enfin, nous tenons également à remercier toutes les personnes qui ont participé de près ou de loin à la réalisation de ce travail

Merci à tous et à toutes.

Dédicace Amine

Un spécial dédicace à ma mère 'Allah Yarhamha ' et j'espère vraiment qu'elle est fière et heureuse de moi

Je dédie ce travail à mon père qui a sacrifié toute sa vie afin de me voir devenir ce que je suis.

A mes chères soeurs : imene et mazola pour leur soutien morale et leurs

Sacrifices le long de mes études. Avec toute ma tendresse.

A mes meilleurs amis Anes Adlane, Yassine, Hassane, Toufik,

Habib et Ayoub. A mes amis de notre section informatique de la promotion 2016/2017 et surtout à salah et fayçal

A mes enseignants qui nous ont beaucoup appris dans ces deux dernières

Années. Que Dieu les garde.

Dédicace Imene

Je Dédie Ce Modeste Travail :

A Mes Très Chères :

A Celui Qui M'a Toujours Appris Comment Réfléchir Avant D'agir, à Celui Qui M'a Soutenu Tout Au Long De Ma Vie Scolaire, à Celui Qui N'a Jamais épargner Un Effort Pour Mon Bien, Mon Cher Père.

A Celle Qui Est Toujours à Coté De Mon Cœur, à Celle Qui M'appris Le Vrai Sens De La Vie, à Celle Qui N'a Hésité Aucun Moment à M'encouragé, Ma Chère Mère.

A Mes Sœurs Abir Et Kaouther Et Mon Frère Ayoub.

A Mes Grands Parents Et Tous Mes Oncles Et Mes Tantes.

A Mes Cousins Et Mes Cousines.

A Tous Mes Amies Surtout Fouzia, Rania, Asmaa, Karima, Zola, Saida.

Et Tous Mes Amis Et Les Collègues De Promotion Informatique 2016/2017.

A Tous Ce Qui Enseigné Moi ont Long Ma Vie Scolaire.

A Tous Ceux Qui M'ont Aidé Dans La Réalisation De Ce Travail.

Et à Tous Ceux Que J'aime Et Qui M'aiment.

Sommaire

Liste des abréviations

TAL : Traitement automatique des langues.

TALA : Traitement automatique de la langue arabe.

TALN : Traitement automatique des langues naturelles.

Liste Des Tableaux

Liste Des Figures

Liste Des Equations

الملخص

غالبا ما يتم تنظيم النصوص المكتوبة باللغة العربية بالنظر إلى المواضيع والأفكار التي تحتويها وبالتالي فإن التجزئة التقليدية باستخدام علامات الترقيم فقط ليست ناجعة لتحديد الحدود الحقيقية للفقرات والجمل، الشيء الذي قد يؤثر سلبا على عملية تحديد أهمية بعض الفقرات على الأخرى.

تهدف التجزئة الموضوعية إلى تسهيل الوصول إلى المعلومات واسترجاعها وذلك بتحليل مضامين النصوص. يتم تحديد المواضيع الجزئية التي تشكل الموضوع الإجمالي للنص بتحليل ودراسة العلاقات بين الجمل والفقرات بالنظر إلى الكلمات والمفردات التي تحتويها. يتم في الأخير تشكيل مجموعة من الكتل النصية تحتوي كل واحدة منها على مجموعة من الجمل والفقرات التي تعالج موضوعا جزئيا مشتركا.

الكلمات المفتاحية: المعالجة الآلية للغة العربية، التجزئة الموضوعية، العلاقة بين الأجزاء النصية، التشابه التوزيعي.

Résumé

Les textes écrits en langue arabe sont souvent organises en thèmes d'où une segmentation classique en utilisant seulement la ponctuation, ne permet pas d'identifier correctement les vrais limites des segmentations d'un texte ce qui peut fausser les jugements de pertinence attribués à ces dernier. La segmentation thématique a pour rôle de faciliter l'accès aux informations, elle basée essentiellement sur l'analyse du contenu d'un texte, les segments du texte sont souvent identifies à travers des expressions clés qui définissent le sujet ou le thème du segment .Le groupement des passages textuels dans des blocs de thèmes est généralement base sur l'étude des relations entre les phrases.

Mots clés : Traitement automatique de la langue arabe, Segmentation thématique, Relation entre segments, similarité distributionnelle.

Abstract

Arabic texts are often organized into themes from which classical segmentation using only punctuation does not allow to correctly identify the true limits of the segmentations of a text, which can distort the judgments of relevance attributed to these latest. Thematic segmentation has the role of facilitating access to information, based essentially on the analysis of the content of a text, segments of the text are often identified through key expressions that define the subject or theme of the segment. The grouping of textual passages into blocks of themes is generally based on the study of the relationships between the sentences.

Keywords: Arabic language processing, Thematic Segmentation, Relationship between Segments, Distributional similarity.

Introduction générale

Introduction générale

Le traitement automatique de la langue arabe est une discipline en pleine expansion, dans laquelle on voit de plus en plus de recherches et de technologies se soucier des spécificités de cette langue et proposer des outils nécessaires au développement de son traitement automatique.

Le traitement automatique de la langue arabe touche plusieurs domaines, telles que, les applications de correction grammaticale, les applications de communication homme/machine, les applications de traduction automatique, etc.

La segmentation est une étape fondamentale dans le traitement automatique d'un texte, son rôle est de découper un texte en unités d'un certain type qu'on aura définies et repérées préalablement. La segmentation formelle d'un texte informatisé est l'opération de délimitation des segments de ses éléments de base qui sont les caractères, en éléments constituants de différents niveaux structurels : paragraphe, phrase, syntagme, mot graphique, mot-forme, morphème, etc. Or la segmentation thématique consiste à subdiviser un texte en bloc de phrases ou de paragraphes discutant un même sous thème ou une même idée.

Les spécificités du système d'écriture et de la morphologie arabes engendrent un grand nombre d'ambiguïtés effectives et virtuelles, morpho-lexicales ou syntaxiques, provoquant une importante explosion combinatoire, surtout au niveau de l'analyse morphologique, ce qui ajoute d'énormes difficultés à l'analyse automatique de l'arabe aussi bien au niveau lexical, morphologique ou syntaxique. La disjonction par exemple, au niveau du système d'écriture, entre l'ensemble des consonnes et celui des voyelles est une source de ces difficultés. Au niveau morphosyntaxique, le phénomène d'agglutination au sein du mot graphique arabe qui, dans certains cas, peut constituer une phrase complète, en est une autre.

Le segmenteur que nous décrivons dans ce mémoire est un outil de TAL arabe dont la fonction principale est de segmenter un texte arabe en sous thèmes en étudiant les relations lexicales et sémantiques qui peuvent exister entre ses segments.

Notre travail est organisé en 4 chapitres :

Chapitre 01 : consiste à discuter la langue arabe en général et ça particularité par rapport aux langues latines, les problèmes et difficultés liées au traitement automatique de la langue arabe et les outils qui sont déjà mis en place pour le TAL arabe.

Chapitre 02 : Dans ce chapitre nous présentons les différentes méthodes et algorithmes utilisées dans la segmentation thématique des textes.

Chapitre 03 : Concerne la conception. Nous détaillons l'approche adoptée pour segmenter thématiquement les textes arabes ainsi que l'architecture de notre système.

Chapitre 04 : Ce chapitre concerne la réalisation et l'implémentation de la solution proposée pour la segmentation thématique d'un texte écrit en langue arabe standard non-voyellé.

Chapitre I :

Traitement Automatique

De La Langue Arabe

I. Traitement Automatique De La Langue Arabe

I.1. Introduction

L'arabe est parlé dans plus de 22 pays, du Maroc jusqu'à l'Iraq et dans toute la péninsule arabe. 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.

Ce chapitre est consacré à la définition et à la présentation de la langue arabe et de ses spécificités et leurs problèmes liés au traitement automatique ainsi que les outils utilisés de TALA.

I.2. La langue arabe

I.2.1. Définition

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.

D'après Farghaly et Shaalan (2009), l'arabe littéraire se répartit en deux catégories : l'arabe classique et l'arabe moderne standard (MSA). 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.

La langue arabe, originaire de la péninsule arabique (Arabie Saoudite, Yémen, Oman, etc.), doit sa véritable expansion à la diffusion de l'islam. Auparavant, la langue arabe était utilisée comme moyen d'expression et d'échange dans la poésie, la prose et les histoires orales. La même langue parlée est écrite mais avec des accents différents et des variations linguistiques mineures au niveau de l'écrit. Elle était à cette époque dans une forme assez proche de l'arabe moderne standard.

C'est avec l'avènement de l'islam que la langue arabe a connu un véritable essor. La langue arabe classique est née lorsque l'islam fut révélé, en 610, par la révélation du Coran formulé en arabe. Cette époque était appelée par certains historiens et linguistes, la première métamorphose de la langue arabe. La langue arabe est devenue une langue officielle du monde musulman en 685 quand le calife Oumeya Abd Al Malik Ibn Marwan arriva à la capitale du monde musulman, Damas, avec pour objectif de centraliser son pouvoir politique : il a imposé donc l'arabe comme

unique langue officielle. Le calife entreprend des réformes de l'écriture par la suite et prend de grandes décisions concernant les signes écrits. À partir du VIIIe siècle une codification de la grammaire fixa la langue dans sa forme classique définitive et facilita la propagation de la langue par l'enseignement partout où l'islam a pu pénétrer. C'est à cette époque que les premiers traités et dictionnaires sont apparus. Entre le VIIIe et le Xe siècle, les sciences et techniques islamiques se sont développées.

Dans la maison de la sagesse1 à Baghdâd, des manuscrits grecs, de philosophie et de sciences, furent traduits en arabe : c'est la seconde métamorphose de la langue arabe. Farghaly (2010) présente un compte rendu détaillé et complet sur l'apparition de la langue arabe ainsi que sa structure.

L'arabe est une langue de civilisation qui a duré plus de quatorze siècles, et était arrivée jusqu'à l'occident entre le VIIIe et les Xe siècles. La langue Arabe s'est étendue sur plusieurs continents à des peuples non arabes, et est devenue la langue officielle de plusieurs pays. La figure 1.1 montre les pays du monde qui ont actuellement pour langue officielle la langue arabe. [1]

Figure I-1: Le monde arabe. [2]

I.2.2. L'alphabet et les caractères spéciaux

L'arabe s'écrit et se lit cursivement (écriture dont les lettres sont reliées les unes avec les autres) de droite à gauche, en utilisant un alphabet de 28 lettres (Tableau ***1.1***).

Lettre arabe	Prononciation	Lettre arabe	Prononciation
أ	Alif	ض	Dad
ب	Ba'	ط	Tah
ت	Ta'	ظ	Zah
ث	Tha'	ع	Ayn
ج	Jim	غ	Ghayn
ح	Hha'	ف	Fa
خ	Kha'	ق	Qaf
د	Dal	ك	Kaf
ذ	Thal	ل	Lam
ر	Ra	م	Mim
ز	Zayn	ن	Nun
س	Sin	ه	Ha
ش	Chin	و	Waw
ص	Sad	ي	Ya

Table I-1: Les 28 lettres arabes.

La représentation morphologique de l'arabe est assez complexe en raison de la variation morphologique et du phénomène d'agglutinement, les lettres changent de formes selon leur position dans le mot (isolée, initiale, médiane et finale). Le tableau 1.2 montre un exemple des différentes formes de la lettre «qaf » dans différentes positions:

Isolée	Initiale	Médiane	Finale
ق	قـ	ـقـ	ـق
	قرآن	القرآن	غسق

Figure I-2: Différentes écritures de la lettre « qaf » en différentes positions dans le mot et comme lettre séparée.

Pour une meilleure précision de la prononciation, des signes ont été inventés. Il s'agit de trois voyelles brèves et de sept signes orthographiques qui s'ajoutent aux consonnes. Ces trois voyelles brèves sont :

- **[Fatha] «ـَـ»** elle surmonte la consonne et se prononce comme un «a» français.
- **[Damma] «ـُـ»** elle surmonte la consonne et se prononce comme un «ou» français.
- **[Kasra] «ـِـ»** elle se note au-dessous de la consonne et se prononce comme un « i » français.

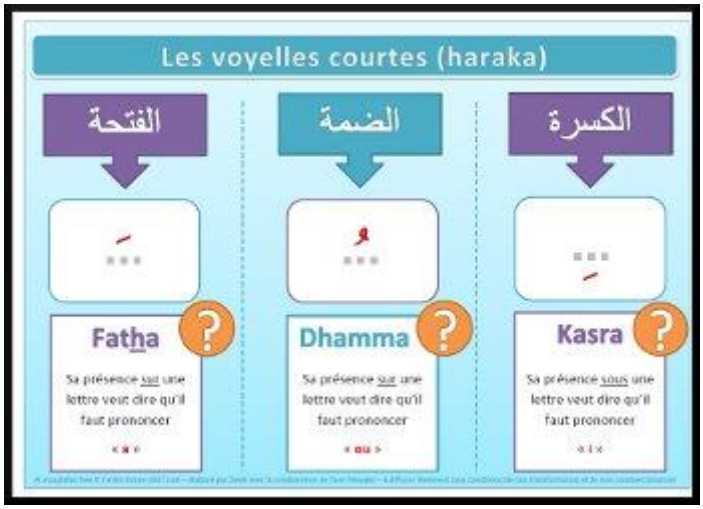


Figure I-3: Les trois brèves.

❖ Les sept signes orthographiques sont :

- **[Sukun] « ـْـ» :** ce signe indique qu'une consonne n'est pas suivie (ou muet) par une voyelle. Il est noté toujours au-dessus de la consonne.

❖ Les trois signes de Tanwin :

Lorsque (la Fatha, la Kasra et la Damma) sont doublées, elles prennent un son nasal, comme si elles étaient suivies de « n » et on les prononce respectivement :

- **[an] «ـًـ»** pour les Fathatan.
- **[in] «ـٍـ»** pour les Kasratan.
- **[un] «ـٌـ»** pour les Dammatan.

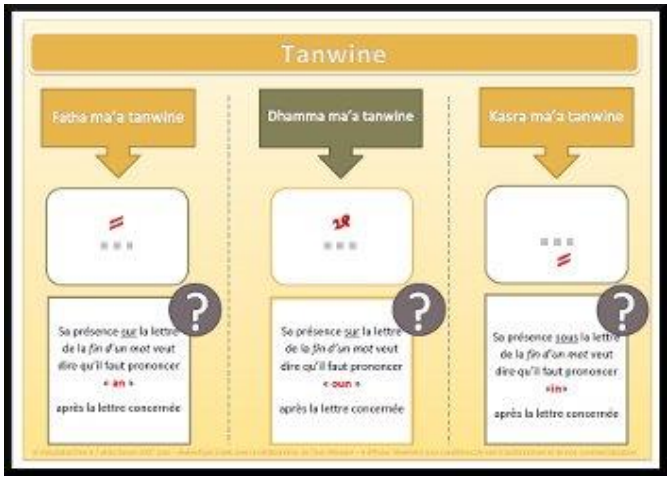


Figure I-4: Tanwin. [16]

- **[Chadda] «ـّـ» :** comme dans le français, l'arabe peut renforcer une consonne quelconque comme makkana/yomakkinou/makkine.

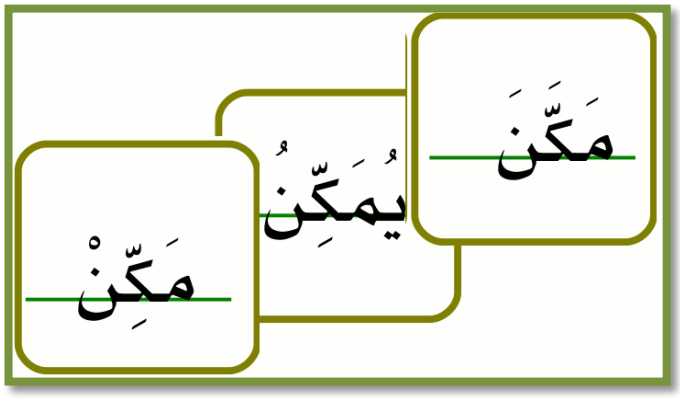


Figure I-5: Chadda. [17]

- **[Wasla] «ٱ »:** quant la voyelle d'un Alif au commencement d'un mot doit être absorbée par la dernière voyelle du mot qui précède.
- **[Madda] «آ »:** la madda (prolongation) se place sur l'Alif pour indiquer que cette lettre tient lieu de deux alifs consécutifs ou qu'elle ne doit pas porter le Hamza.

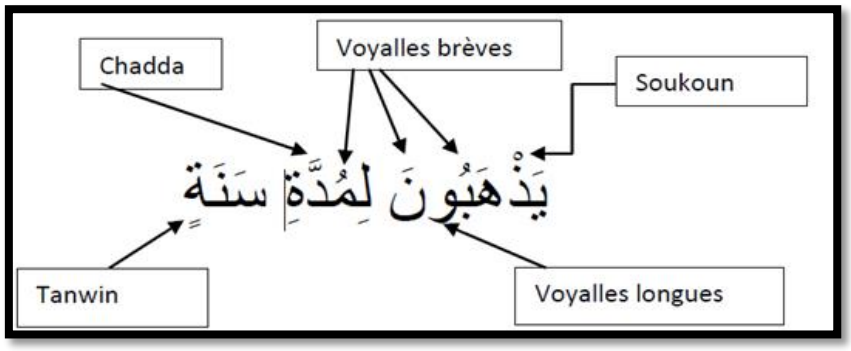


Figure I-6: Les signes orthographique/tanwin.

- Le tableau 1.3 montre un exemple pour les mots «كتب » et «شعر».

	Interprétation I		Interprétation II		Interprétation III	
كتب	كَتَبَ	Il a écrit	كُتِبَ	Il a été écrit	كُتُبٌ	Des livres
شعر	شَعَرَ	Il a senti	شِعْرٌ	Poème	شَعْرٌ	Chevelure

Table I-2: Ambiguïté causée par l'absence de voyelles pour les mots «كتب » et «شعر».

Cependant, les textes courants rencontrés dans les journaux et les livres ne comportent habituellement pas de voyelles. De plus, certaines lettres comme Alif «ا» peuvent symboliser le «آ », «إ», «أ», «ٱ » ou de même que pour les lettres «ى » et « ه » qui symbolisent respectivement «ي» et «ة». [3]

Les lettres lunaires initiales d'un nom n'assimilent pas l'article qui les précède et par « ا ب ج ح خ ع غ ف ق ك م و ي » : conséquent ne reçoivent pas le chadda. La lettre ال est prononcée.

Les lettres solaires initiales d'un nom assimilent l'article qui les précède et reçoivent « ت ث د ذ ر ز س ش ص ض ط ظ ل ن » : ainsi le chadda. La lettre ال est muette.

Les caractères de la langue arabe n'appartiennent pas au code ASCII, d'où la nécessité d'utiliser un autre code qui prend en charge la langue arabe, ce code est l'Unicode, ce dernier permet de coder tous les caractères utilisés par la langue arabe en mode 16 bits.

I.2.3. Les nombres

On emploie deux catégories de chiffres dans le monde arabe :

- Les chiffres dit arabes (utilisés en occident et plus au Maghreb) et transmis à l'Europe par les arabes à l'époque médiévale.
- Les chiffres dit indiens utilisés au Moyen-Orient et en Egypte, ils sont postérieurs aux premiers et fonctionnent exactement de la même façon.

Arabes	0	1	2	3	4	5	6	7	8	9
indiens	٠	١	٢	٣	٤	٥	٦	٧	٨	٩

Table I-3: Les chiffres arabes.

- Il est important de noter que les nombres représentent un chapitre complexe de la grammaire arabe, ils sont soumis à des règles strictes et complexes, d'où notre choix d'utiliser les chiffres arabes dans la langue contrôlée. Le choix des chiffres indiens a été écarté par souci de conformité aux normes de rédaction internationales. [4]

I.2.4. Morphologie arabe

Le lexique arabe comprend trois catégories de mots : verbes, noms et particules. Les verbes et noms sont le plus souvent dérivés d'une racine à trois consonnes radicales. 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. On dit donc que l'arabe est une langue à racines réelles à partir desquelles on déduit le lexique arabe selon des schèmes qui sont des adjonctions et des manipulations de la racine. [13]

Le Tableau 1.5 donne quelques exemples de schèmes appliqués aux mots كتب écrire et حمل porter. On peut ainsi dériver un grand nombre de noms, de formes et de temps verbaux.

schèmes	KTB	كَتَبَ	notion d'écrire
R_1â-R_2i-R_3	KâTiB	كَاتِب	écrivain
R_1a-R_2a-R_3a	KaTaBa	كَتَبَ	a écrit
maR_1R_2aR_3	maKTaB	مَكتَب	bureau
R_1uR_2iR_3a	KuTiBa	كُتِبَ	a été écrit
…			

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

Table I-4: Exemple de schèmes pour les mots كتب écrire et حمل porter.

Les lettres en majuscule (Ri) désignent les consonnes de base qui composent la racine. Les voyelles (â, à, 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. Une autre caractéristique est le caractère flexionnel des mots : les terminaisons permettent de distinguer le mode des verbes et la fonction des noms [13].

I.2.5. 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.

Plusieurs types d'affixes sont agglutinés au début et à la fin des mots : antéfixes, préfixes, suffixes et post fixes. La représentation suivante schématise une structure possible d'un mot. Note que la lecture et l'écriture d'un mot se fait de droite vers la gauche.

Post fixe	Suffixe	Corps schématique	Préfixe	Antéfixe

Table I-5: Tableau Montrer une structure possible d'un mot.

- Antéfixes sont des prépositions ou des conjonctions.

- Préfixes et suffixes expriment les traits grammaticaux et indiquent les fonctions : cas du nom, mode du verbe et les modalités (nombre, genre, personne,…)
- Post fixes sont des pronoms personnels.

Exemple 1 : اتتذكروننا

Ce mot exprime la phrase en français : "Est ce que vous vous souvenez de nous ?

- La segmentation de ce mot donne les constituants suivants **:**

Figure I-7: La segmentation de mot « اتتذكروننا ».

- **Antéfixe :** أ conjonction d'interrogation**.**
- **Préfixe:** ت préfixe verbal du temps de l'inaccompli.
- **Corps schématique :** تذكر dérivé de la racine : ذكر selon le schème taR1aR2aR3a.
- **Suffixe :** ون suffixe verbal exprimant le pluriel.
- **Post fixe :** نا pronom suffixe complément du nom. [5]

Exemple 2 :

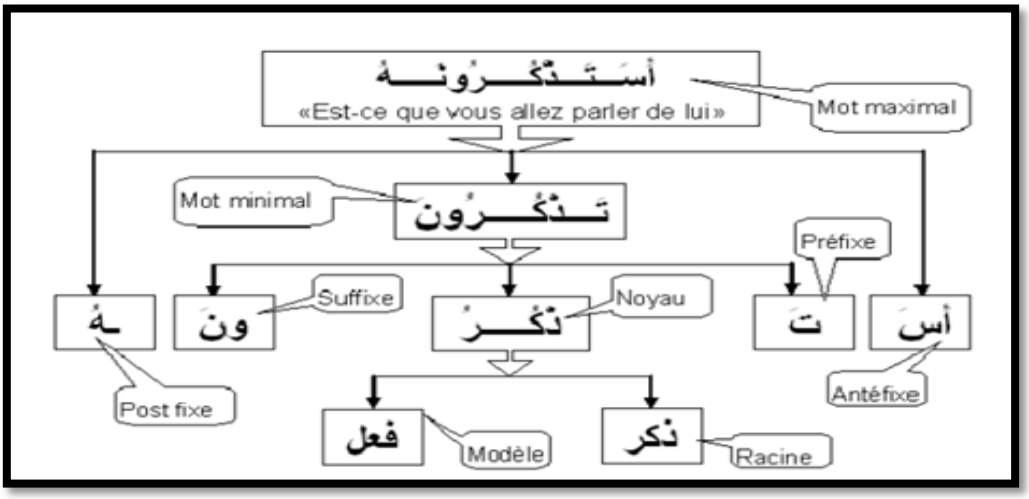


Figure I-8: La segmentation du mot «أَسَتَذْكُرُونَهُ» [25]

I.2.6. 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 le cas pour la troisième personne du singulier lorsque le sujet n'est pas énoncé sous la forme d'un nom après le verbe.

Considérons l'exemple donné par Attia, اكلت الدجاجة (la poule a mangé). C'est une phrase verbale constituée d'un verbe suivi d'un nom. Le verbe est constitué de la racine اكل (manger) suivie d'un suffixe ت qui représente la marque de féminin. Si on considère que le deuxième mot qui apparaît après le verbe est le sujet, alors la phrase est traduite par La poule a mangé. La marque de féminin attachée au verbe fait référence donc à la poule. Par contre, si on considère que le deuxième mot est le complément d'objet, alors la phrase devient Elle a mangé le poulet. Dans le dernier cas, la phrase en arabe ne contient pas de sujet (qui est le pronom personnel Elle), il est sous-entendu. La marque de féminin attachée au verbe fait référence dans ce cas à elle. Il est à noter que pour les deux phrases, la diacritisation est exactement la même, et qu'il est difficile de traduire cette phrase sans savoir le contexte dans lequel elle apparaît. [6]

I.3. Catégories des mots

L'arabe considère 3 catégories de mots :

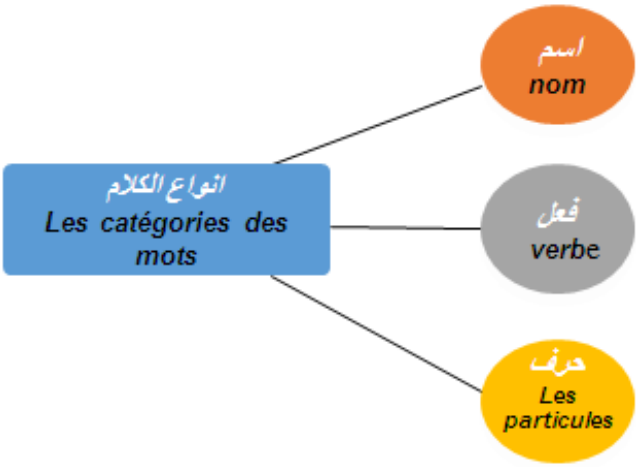


Figure I-9: Les Catégorie des mots.

I.3.1. 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. 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.

❖ La déclinaison des noms se fait selon les règles suivantes:

- **Le féminin singulier :** On ajoute le, ة exemple صغير petit devient صغيرة petite
- **Le féminin pluriel** : De la même manière, on rajoute pour le pluriel les deux lettres, ات exemple صغير petit devient صغيرا ت petites.
- **Le masculin pluriel :** Pour le pluriel masculin on rajoute les deux lettres ين ou ون dépendamment de la position du mot dans la phrase (sujet ou complément d'objet), exemple : الراجع revenant devient الراجعين ou الراجعون revenants.
- **Le Pluriel irrégulier:** Il suit une diversité de règles complexes et dépend du nom. exemple : طفل un enfant devient أطفال des enfants.

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).

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.

I.3.2. 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.

❖ 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).

Par exemple : ب + ت + ك *K+T+B donne le verbe* كتب *KaTaBa. (Écrire).*

Dans tous les mots qui dérivent de cette racine, on trouvera ces trois lettres K, T, B La conjugaison des verbes se fait en ajoutant des préfixes et des suffixes, un peu comme en français.

❖ **La langue arabe dispose de trois temps :**

- **L'accompli :** correspond au passé et se distingue par des suffixes (par exemple pour le pluriel féminin on a كتبن KaTaBna, elles ont écrit et pour le pluriel masculin on a كتبوا KaTaBuu, ils ont écrit).
- **L'inaccompli présent:** présente l'action en cours d'accomplissement, ses éléments sont préfixés (يكتب yaKTuBu il écrit : تكتب TaKTuBu, elle écrit).
- **L'inaccompli futur :** correspond à une action qui se déroulera au futur et est marqué par l'antéposition de س sa ou سوف sawfa au verbe (سيكتب SayaKTuBu il écrira, سوف يكتب sawfa yaKTuBu il va écrire).

I.3.3. Les particules

Ce sont principalement les mots outils 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 حيت où, ou de référence الذين ceux,….

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

- Elles font partie de l'anti dictionnaire 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. [7]

I.4. La Classification d'un mot

Dans cette figure on va montrer la classification lexicale d'un mot :

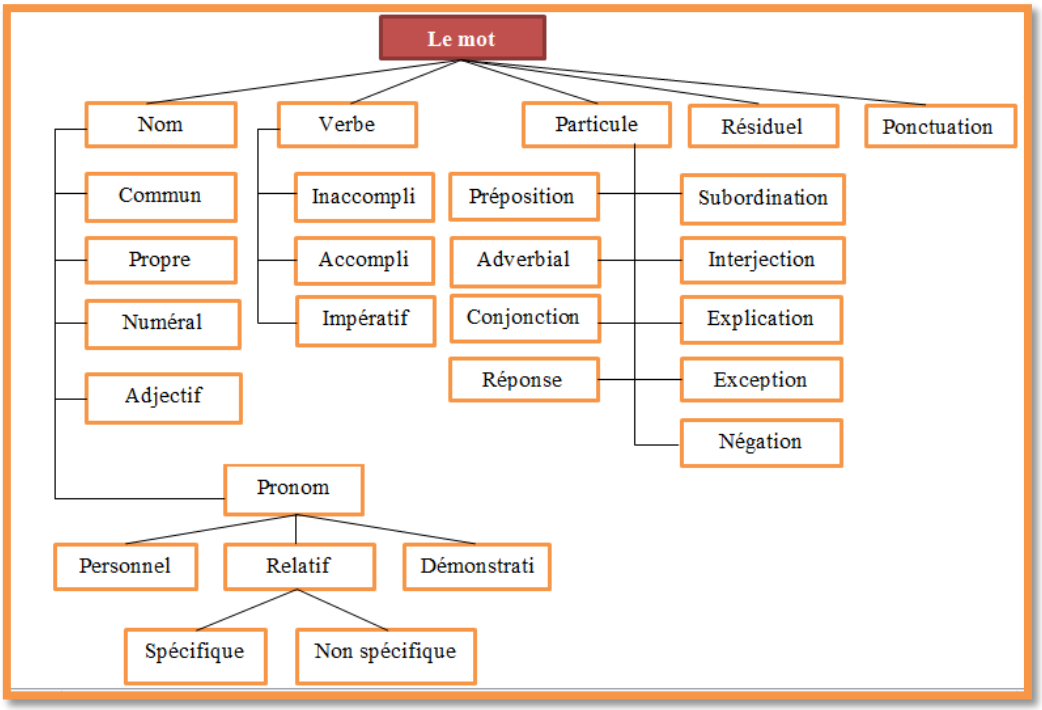


Figure I-10: Classification lexicales d'un mot. [15]

I.5. La structure d'une phrase arabe

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

I.5.1. 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 .Elle est composée de 3 parties [19] :

I.5.1.1. Le Verbe : Appelé en arabe « **الفِعل** »

C'est la partie qui décrit une action effectuée dans le passé, le présent, ou le futur.

I.5.1.2. Le Sujet : Appelé en arabe « **الفاعِل** »

C'est la personne ou la chose effectuant l'action décrite par le Verbe.

I.5.1.3. Le Complément : Appelé en arabe **« المفعول ».**

Et il y a deux genres :

- **Direct :** Il correspond à la personne ou la chose sur laquelle s'applique le Verbe.
- **Indirect :** C'est le complément qui indique le temps, le lieu, les circonstances…etc.

Remarque :

Le Verbe est la seule partie OBLIGATOIRE pour former une phrase verbale. Les deux autres parties peuvent ne pas être présentes.

Exemple : « يذهب سامي إلى الجامعة »

Figure I-11: Exemple d'une phrase verbale.

- Le verbe ici est « يذهب » qui est le verbe allé au présent.
- Le sujet (celui qui exécute l'action d'aller) est Samy.
- Le complément ici est « à l'université : إلى الجامعة », qui est donc un complément indirecte.

I.5.2. Les phrases nominales

En langue arabe, une phrase nominale suit les règles suivantes:

- Elle commence soit par un nom ou soit par un pronom.
- Elle est composée de deux parties: la première partie appelée mubtada' (اَلْمُبْتَدَأُ) est le sujet de la phrase. La deuxième partie, appelée khabar (اَلْخَبَرُ), apporte une information sur le sujet.

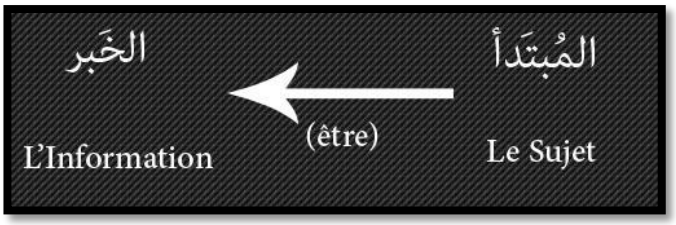


Figure I-12: La structure d'une phrase nominale.

I.5.2.1. Le mubtada : اَلْمُبْتَدَأُ

Le sujet en arabe al-mubatada est toujours composé d'un nom ou d'un pronom. Il est toujours au cas nominatif مَرْفوعٌ.

Le mubtada est le sujet de la phrase. Il peut représenter une personne ou une chose. Il est décrit ou précisé par le khabar de la phrase. Par exemple, dans la phrase: Mohammed est debout. « Mohammed » est le mubtada – sujet et « est debout » est le khabar – description ou précision sur le sujet. Le mubtada peut être composé de :

- **اِسْمٌ مرفوعا** : *un nom défini :* **جَالِسٌ** اَلْكَلْبُ – *le chien est assis.*
- **ضَمِيرٌ :** un pronom : **هُو طَالِبٌ جَيِّدَ** – Il [est] un bon étudiant.
- **عَلَمٌ :** un nom propre : **أَحْمَد طَالِبٌ** – Ahmed est un étudiant.
- **اِسْمُ الْإِشارَة :** un pronom démonstratif : **هَذَا كِتابٌ** – c'est un livre.
- **مُضافٌ** : un annexant (dans le cas d'une annexion) : **كِتابُ النَّحْوِ مُفِيد** – le livre de grammaire est profitable.

I.5.2.2. Le khabar :اَلْخَبَر

Le khabar quant à lui, n'a pas l'obligation d'être composé d'un nom ou d'un pronom mais lui aussi est toujours au cas nominatif – مَرْفوعٌ bien qu'il puisse être composé de :

- **اِسْمٌ مفردا** : *un seul nom:* **طَالِب** أَحْمَد – Ahmed est un étudiant
- **شِبْهُ جُمْلَةٍ** : *une phrase ou un complément introduit par une préposition :* **أَحْمَدُ** فِي الْفَصْل – Ahmed est dans la classe.
- **جُمْلَةٌ فِعْلِيَّةٌ** *: une phrase verbale :* **دَخَلَ فِي الْفَصْل** أَحْمَد – Ahmed est entré dans la classe.

- **جُمْلَةٌ اسْمِيَّةٌ** : *une phrase nominale :* رَشِيد **أُمُّهُ صَالِحَة** – Rachid, sa mère est pieuse.

Remarque :

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 جميل est traduite par L'étudiant est beau**,** mais le verbe est absent de la phrase en arabe. [19]

I.6. Traitement automatique de la langue arabe

Le traitement automatique de la langue arabe est une discipline en pleine expansion, dans laquelle on voit de plus en plus de recherches et de technologies se soucier des spécificités de cette langue [14] et proposer des outils nécessaires au développement de son traitement automatique. Par ses propriétés morphologiques et syntaxiques la langue arabe est considérée Comme une langue difficile à maîtriser dans le domaine du traitement automatique de la langue.

I.7. Problèmes de T.A.L arabe

La langue arabe est considérée comme une langue difficile à maîtriser dans le domaine du TALN.

Le traitement automatique de l'arabe ont toujours fait face aux problèmes variés de la langue arabe issus de la nature agglutinante de cette langue, sa richesse flexionnelle, l'absence de v voyellation de la majorité des textes arabes écrits, etc.

Dans la section suivante, nous essayerons de donner une brève présentation de ces problèmes, qui rendent le traitement automatique de la langue arabe une tâche difficile à maîtriser.

I.7.1. Absence de voyelles

Le problème de la voyellation réside dans l'absence quasi systématique de la voyellation dans les textes arabes. En effet, les signes de voyellation, lorsqu'ils sont notés, sous la forme de signes diacritiques placés au-dessus ou au-dessous des lettres, apparaissent dans certains textes

(Coran, hadith) ou littéraires (poésie classique, notamment) : on dit qu'ils sont édités en graphie voyellée.

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

L'exemple suivant 1.6 du mot non-voyellé « ktb كتب » possède 16 voyellations potentielles et qui représentent 9 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
كَتِّبْ	kattibo	Fais écrire
كَتَّبَ	katabba	Comme trancher
كَتَّب	katabb	Comme 'tranchement'
...	...	...

Table I-6: Exemple de voyellation.

❖ L'absence des voyelles génère une certaine ambiguïté à deux niveaux:

- Sens du mot.
- Difficulté à identifier sa fonction dans la phrase.

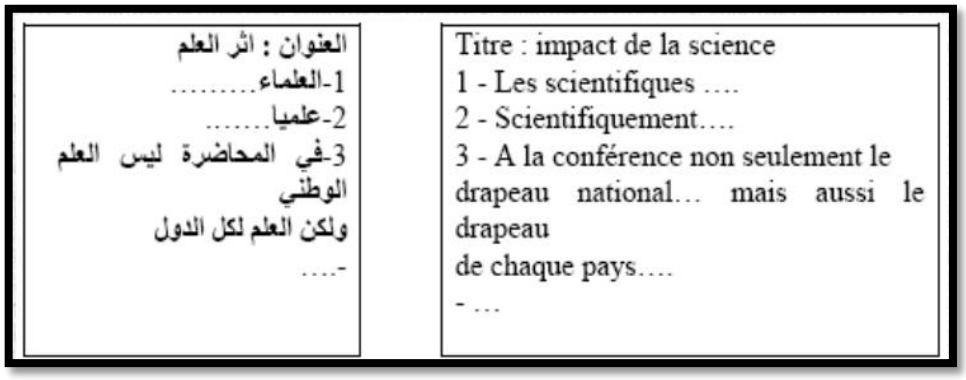


Figure I-13: Effet du mot non voyellé العلم sur les extraits.

L'ambiguïté vient du mot العلم la science ou drapeau alors que voyellé on aura العِلْمُ pour la science et العَلَمُ pour le drapeau. Cette ambiguïté pourrait, dans certains cas, être levée soit par une analyse plus profonde de la phrase ou des statistiques (par exemple il est plus probable d'avoir « العلم الوطني » le drapeau national que la science nationale). [18]

I.7.2. Agglutination

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

Exemple : le mot arabe" اتتذكروننا " correspond en Français à la phrase "Est-ce que vous vous souvenez de nous ?".

Cette caractéristique peut engendrer une ambiguïté au niveau morphologique. En effet, il est parfois difficile de distinguer entre une proclitique ou enclitique et un caractère original du mot. Par exemple, le caractère "و " dans le mot "وصل" (il est arrivé) est un caractère original alors que dans le mot "وفتح " (il a ouvert), il s'agit d'une proclitique. [20]

I.7.3. Irrégularité de l'ordre des mots dans la phrase

L'ordre des mots en arabe est relativement libre. D'une manière générale, on met au début de la phrase le mot sur lequel on veut attirer l'attention et l'on termine sur le terme le plus long ou le plus riche en sens ou en sonorité. Cet ordre provoque des ambiguïtés syntaxiques artificielles, dans la mesure où il faut prévoir dans la grammaire toutes les règles de combinaisons possibles d'inversion de l'ordre des mots dans la phrase.

Ainsi par exemple, on peut changer l'ordre des mots dans la phrase (Tableau 1.7) pour obtenir deux phrases ayant le même sens.

Verbe + sujet + complément	فعل + فَاعل + متمم	Est allé le garçon à l'école	ذهب الولدُ إِلَى المدرسة
sujet + verbe + complément	فَاعل + فعل + متمم	Le garçon est allé à l'école	الولدُ ذهب إِلَى المدرسة
complément + verbe + sujet	متمم + فعل + فَاعل	A l'école est allé le garçon	إِلَى المدرسة ذهب الولدُ

Table I-7: Exemple de combinaisons possibles d'inversion de l'ordre des mots dans la phrase.

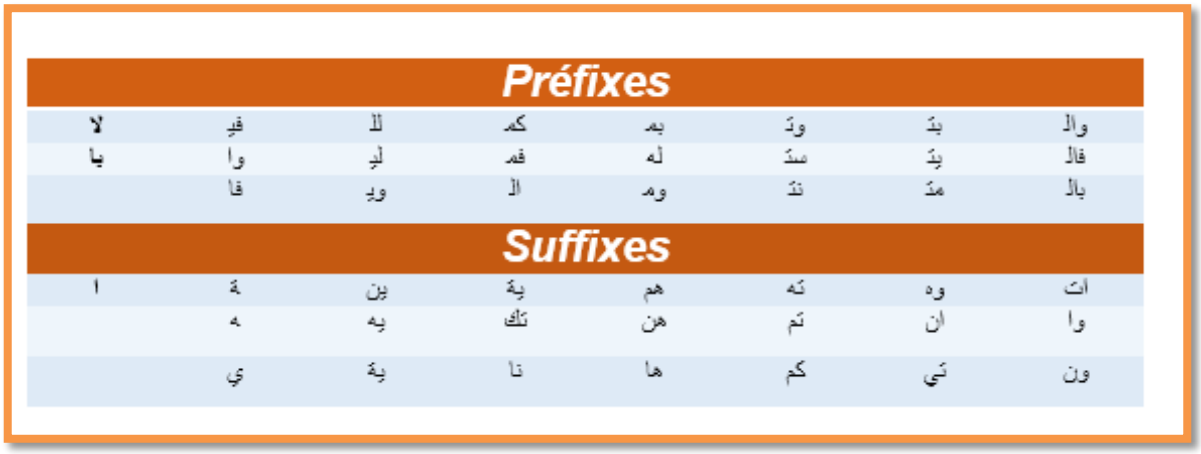
I.7.4. Absence de ponctuation régulière

La langue arabe n'est pas basée principalement sur les signes de ponctuations et les marqueurs typographiques ; il est à noter que ces derniers ne sont pas utilisés de façon régulière dans les textes arabes actuels, et même dans le cas où ils y figurent, ils ne sont pas gérés par des règles précises d'utilisation.

Par ailleurs, nous pouvons trouver tout un paragraphe arabe ne contenant aucun signe de ponctuation à part un point à la fin de ce paragraphe. Ainsi, il convient de noter que la présence des signes de ponctuation ne peut pas guider la segmentation comme c'est le cas pour d'autres langues latines, telles que le français ou l'anglais. Ainsi, la segmentation de textes arabes doit être guidée non seulement par les signes de ponctuations et les marqueurs typographiques mais aussi par des particules et certains mots tels que les conjonctions de coordination, etc. [21]

I.7.5. Détection de racine

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. J'utilise la liste de préfixes et de suffixes voir (Tableau). Plusieurs d'entre eux ont été utilisés 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).



Préfixes							
والـ	بـ	وَ	بم	كم	لـ	فِ	لا
فال	بِ	سَ	له	فم	لِ	وا	يا
بال	مَ	نَ	وم	الـ	وِ	فا	

Suffixes							
ات	وه	ته	هم	هَ	ين	ة	ا
وا	ان	تم	هن	تك	يه	ه	
ون	ني	كم	ها	نا	يهَ	ي	

Table I-8: La liste de Préfixes et suffixes les plus fréquents.

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-‘بِ’ et li-‘لِ’, l'article défini ‘ال’, des suffixes de pronom possessif. [22]

I.7.6. Mots étranger translitérée en arabe

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

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

I.7.7. Ambiguïtés dérivationnelles et flexionnelles

La flexion est la variation de la forme des mots en fonction de facteurs grammaticaux tel que la conjugaison pour les verbes (exemple : le mot " يتأثّرون " (ils s'influencent) est le résultat de la concaténation du préfixe " ي" indiquant le présent et du suffixe " ون " indiquant le masculin pluriel du verbe " تأثّر "). Le problème en analyse morphologique de l'arabe se rapporte surtout au niveau de la dérivation qui est un phénomène plus complexe que la flexion. En effet, la dérivation est la formation de nouveaux mots à partir de mots existants. Dans le cas de la langue arabe, la plupart des mots sont dérivés à partir de racines trilitères ou quadrilatères. Le mot arabe n'est pas le résultat d'une simple concaténation de morphèmes comme c'est le cas pour l'anglais (exemple : unfailingly = un+fail+ing+ly), mais c'est à partir d'une racine, d'une combinaison de voyelles, de préfixes, d'infixes, de suffixes et d'un schème morphologique qu'on obtient un mot (exemple :

à partir de la racine " أثر " (choisir/citer à) on peut dériver plusieurs verbes tel que " تأثّر " (s'influencer) et plusieurs noms tel que " متأثّر. [18]

I.8. Difficultés de traitement automatique de l'arabe

I.8.1. La segmentation de textes

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

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

D'après l'étude réalisée par Belguith [21], certaines particules comme "et | و ", "donc | ف ", etc. jouent un rôle principal dans la séparation de phrases et peuvent être déterminantes pour guider la segmentation.

I.8.2. L'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.

À proprement parler, l'analyse morphologique ne fait que la séparation et l'identification des morphèmes semblables aux mots préfixés (comme les conjonctions " wa | و " et "fa | ف", etc.), des prépositions préfixées (comme " bi | ب" et " li | ل", l'article défini " ال ", etc.), des suffixes de pronom possessif.

La phase d'analyse morphologique détermine un schéma 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.

Le problème principal de cette analyse réside dans l'agglutination et l'absence de voyellation. Pour l'agglutination et contrairement aux langues latines, en arabe, les pronoms, les

prépositions, les articles, les conjonctions, et autres particules collent aux noms, verbes, adjectifs et particules auxquels ils se rapportent. Comparé au français, un mot arabe peut parfois correspondre à une phrase française.

Cette caractéristique engendre une ambiguïté morphologique au cours de l'analyse. Ainsi, la reconnaissance des unités lexicales qui composent une unité morphologique n'est pas toujours facile à détecter. Le problème est de reconnaître que la bonne segmentation réside ainsi, dans la difficulté de distinction entre un proclitique ou enclitique et un caractère original du mot.

Par exemple, le caractère "و" dans le mot "il est arrivé | وصل" est un caractère original alors que dans le mot "et il a ouvert | وفتح ", il s'agit plutôt d'une proclitique.

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" pris hors 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 verbe lier conjugué à la deuxième personne du singulier masculin). [21]

I.8.3. L'étiquetage grammatical

L'étiquetage grammatical est l'opération qui consiste à attribuer à chacun des mots d'un texte la catégorie (non, verbe, adjectif, article défini, etc.) qui est la sienne dans le contexte où il apparaît.

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

Dans ces conditions, le but général de l'étiquetage grammatical consiste à répondre à la question suivante : Comment associer aux différents mots qui composent un texte l'étiquette qui leur convient, compte tenu du contexte où ils occurrent ? Ainsi, le problème des étiquettes grammaticales est souvent posé lorsque les textes désirés sont sous leurs formes non ou partiellement voyelles, plutôt qu'à leurs formes voyelles.

Le problème de la voyellation d'un mot est ainsi posé du fait que le choix de l'accentuation qui convient au mot est difficile et dépend essentiellement du contexte.

Le tableau 1.9 présente le problème d'ambiguïté grammaticale rencontrée lors de l'attribution catégorique d'un mot non voyéllé "ktb | كتب»»", qui admet au moins cinq étiquettes grammaticales qui sont les suivantes : [21]

Exemple de voyellation	Étiquettes grammaticales
كُتُبٌ | kutubun : des livres	substantif, masculin, pluriel
كَتبٌ | katbun : un écrit	substantif, masculin, singulier
كَتَبَ | kataba : il à écrit	verbe, 3ème personne masculin, singulier de l'accompli actif
كُتِبَ | kutiba : il a été écrit	verbe, 3ème personne masculin, singulier de l'accompli passif
كَتِّبْ | kattib : fais écrire	verbe à l'impératif, 2ème personne masculin, singulier

Table I-9: Exemple de d'étiquettes grammaticales attribuées selon la voyellation.

I.8.4. 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.

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

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]

I.8.5. 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 (ِ العِلْمُ pour la science et العَلَمُ pour 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.

I.9. 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 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.

I.9.1. Analyseurs morphologiques

L'analyseur morphologique segmente les unités lexicales, repère les différents composants et atteste leur appartenance à la langue et de pouvoir disposer ainsi de tous les renseignements le concernant pouvant servir à l'analyse syntaxique.

I.9.1.1. Buckwalter

L'analyseur de Buckwalter développé par 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 (78839 entrées). 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.

I.9.1.2. Analyseur de shaalan

Shaalan, 1989 signale un projet de maîtrise au Caire Université. Est un moteur de règles écrites dans SICStus Prolog et a besoin de fond en Prolog qui est difficile à atteindre par un linguiste typique. Elle est antérieure normes de codage modernes, en utilisant un système de translittération.

I.9.2. Aramorph

L'analyseur morphologique Aramorph [23] segmente les unités lexicales, repère les différents composants et atteste son appartenance à la langue. Pour cela, le système est assisté par le lexique DINAAR.1 pour éviter les analyses théoriquement possibles et inexistantes dans la langue. Par la suite, l'analyseur donne une liste des traits associés à l'unité lexicale en entrée. Il offre deux types d'options. Le premier vise les traits morphosyntaxiques, le second concerne l'analyse des préfixes et suffixes. En plus des étiquettes morphosyntaxiques, il donne en sortie d'autres informations comme la base, l'unité lexicale minimale vocalisé ou non ainsi que la forme complété supposée vocalisée ou non.

Analyser les préfixes revient à décrire ses découpages possibles et d'examiner les compositions des clitiques. Ceci amène le système à faire la distinction entre les clitiques ayant la même forme mais appartenant à des catégories syntaxiques différentes.

Exemple : «والتلوث » "et la pollution" l'analyseur découpera le proclitique «و » et dira que «و » est celui de la liaison.

I.9.3. 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 réalisée 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.

I.9.4. 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 TreeBank.Ci-dessous la description des fichiers d'entrée et de sortie de l'analyseur.

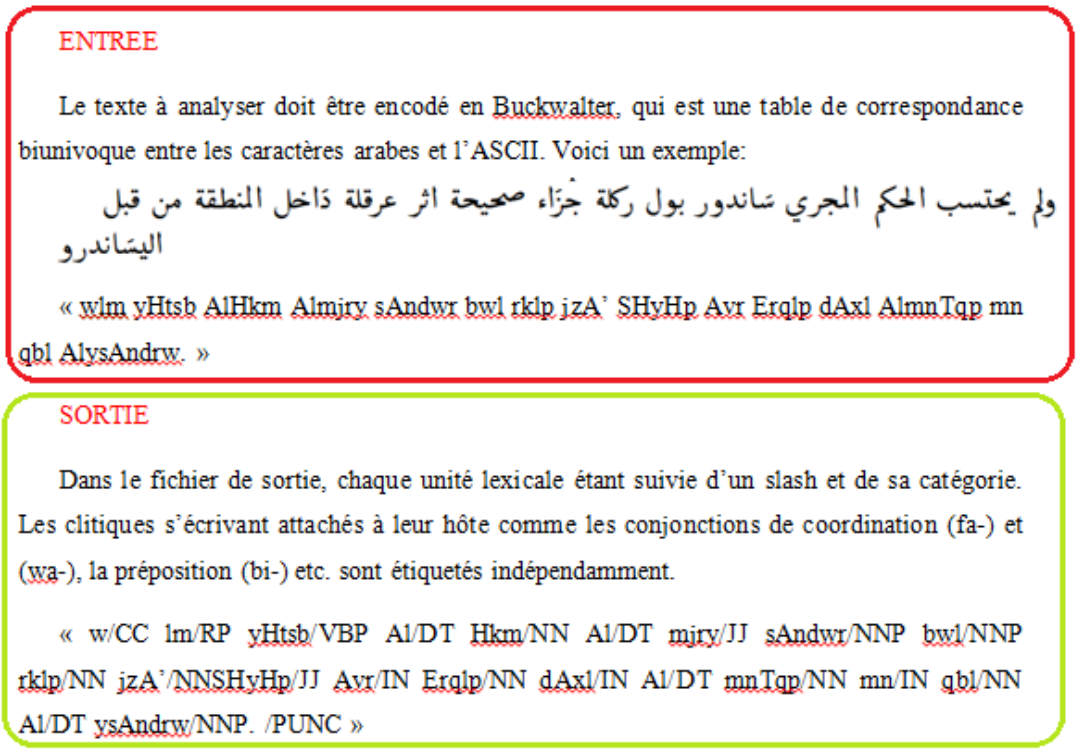


Figure I-14: Exemple d'analyse avec ASVM.

I.9.5. Analyseur de Sakhr

La Société Sakhr [24] a également produit un analyseur morphologique, qui est désigné par Multi-Mode processeur morphologique (MMMP). Le site Sahhr http://www.zahhr.com/ prétend que leur programme couvre moderne et l'arabe classique, et il identifie la forme de base en supprimant tous les affixes et lui donne le modèle morphologique.

I.9.6. Sebawai de Derwish

Sebawai est un analyseur morphologique arabe (développé par Darwish 2005). L'analyseur morphologique utilise des modèles orthographiques arabes pour trouver les racines. La couverture est pas parfait .cet analyseur morphologique trouve la racine avec succès 84% du temps".

I.9.7. L'analyseur morpho-syntaxique Alkhalil

Alkhalil Morpho Sys qui est un analyseur morphologique textes arabes standards. Il peut traiter les textes non diacrités, ainsi, il peut traiter les textes diacrités partiellement ou totalement. En fait, 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) (2010 برنامج الخليل الصرفي).

الدخل INPUT	الخرج OUTPUT							
	الكلمة المشكولة Voweled Word	السابق Prefix	الجذع Stem	نوع الكلمة Type	الوزن Pattern	الجذر Root	الحالة الإعرابية POS Tags	اللاحق Suffix
هذا	هذا	#	هذا	اسم إشارة	#	#	#	#
	هذًا	#	هذا	مصدر أصلي	فعًا	هذ	مفرد مذكر منصوب نكرة	#
	هَذَا	#	هذا	فعل ماض مبني للمعلوم	فعا	هذ	ثلاثي مجرد مسند إلى الغائب(هو) متعد ولازم	#
	هَذَا	#	هذا	فعل ماض مبني للمجهول	فُعَا	هذ	ثلاثي مجرد مسند إلى الغائبين(هما) متعد ولازم	#
	هَذَا	#	هذا	فعل ماض مبني للمعلوم	فَعَا	هذ	ثلاثي مجرد مسند إلى الغائبين(هما) متعد ولازم	#
	هَذَا	#	هذا	فعل أمر	فَعَا	هذ	ثلاثي مجرد مسند إلى المخاطبين(أنتما) متعد ولازم	#
ما	ما	#	ما	اسم شرط	#	#	#	#
	ما	#	ما	اسم موصول	#	#	#	#

Table I-10: Analyse de mot « hadha » par l'analyseur alkhalil.

I.9.8. Le système AraParse d'analyse morphosyntaxique de l'arabe non-voyellé

AraParse est un système capable d'analyser les textes arabes tels qu'ils se présentent habituellement, c'est-à-dire sous leurs formes voyellées, non voyellées ou partiellement voyellées [9]. L'objectif est de réaliser un noyau de système d'analyse morphosyntaxique de grandeur réelle qui peut être réutilisé dans d'autres applications à grande échelle.

I.9.9. 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 (Key Word 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 AraConc développé pour l'arabe par Aramorph 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 autour 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 morphosyntaxique.

I.9.10. Racineurs

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 larkey et de Shereen khoja [22]. Ci-dessous une description succincte de ces racineurs.

I.9.10.1. Racineur de larkey

L'approche de larkey [9] 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 (معلمين, des professeurs) et féminins (ات) dans (مسلمات, musulmanes) ; la forme possessive (نا, كم, هم) dans (كتابهم, ses livres) et les préfixes dans les articles définis (ال, وال, بال, كال, فال). L'ensemble des préfixes et suffixes à supprimer sont présentés dans le tableau 1.11

Préfixes			Suffixes	
1- Caractère	2- Caractère	3- Caractère	1- Caractère	2- Caractère
ت	ال	وَال	ة	اة
ل	بت	فَال	ه	ان
ا	تَ	بَال	ي	تَا
ي	يت		ا	تك
م	لت			تي
	مت			ته
	وت			تم
	ست			هم
	نت			هن
	بم			هَا
	لم			كم
	وم			وَا
	كم			ون
	فم			وه
	لَ			ية
	وي			نَا
	لي			ين
	سي			يه
	في			
	وَا			
	فَا			
	لَا			
	بَا			

Table I-11: Liste des préfixes et suffixes.

I.9.10.2. Racineur de Khoja

Le racineur 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 consiste à détecter la racine d'une unité lexicale, d'une part, il faut connaître 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. Arabic stemmer est une version open source simplifiée de l'algorithme développé par Shereen khoja, il est très utilisé dans la détection les racines des mots arabes. [9]

I.9.11. Les 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 lexicale 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.

Arabic Wordnet 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 synsets et 23481 mots. [22]

I.10. Conclusion

Le but de ce chapitre était de présenter des informations relatives à la langue arabe et sa morphologie, ces caractéristiques, ainsi que les problèmes de traitement automatique de la langue arabe tels que le non voyellation, l'agglutination, l'irrégularité de l'ordre des mots et l'absence de ponctuation. Nous avons mis le point sur quelques outils de TALA et nous avons choisi comme outils à utiliser dans notre projet, le racineur Arabic Stemmer qui nous aidera à calculer les fréquences de distribution des termes et la base de données lexicale Arabic WordNet pour le calcul de la synonymie.

Chapitre II :

La Segmentation Thématique

II. La Segmentation Thématique

II.1. Introduction

La segmentation thématique est une nouvelle technique pour l'amélioration de l'accès à l'information, elle peut être définie comme la tâche de subdivision d'un document en plusieurs paragraphes thématiquement cohérents. En recherche d'information par exemple, avoir des documents thématiquement segmentés peut résulter en la récupération des segments de texte courts et pertinents qui correspondent directement à la requête d'un utilisateur au lieu de longs documents examiné avec soin par l'utilisateur pour trouver l'objet de son intérêt.

Dans ce chapitre, nous avons présentons les approches utilisées dans la segmentation thématique (linaires, non-linaires, globales/locales..) ainsi que la présentation des algorithmes les plus utilisés dans ce cadre comme C99, TextTiling, Dot Plotting**,..** et nous finissons par un petit recensement des travaux relatifs en langue arabes.

II.2. Définition De La Segmentation

Avant d'analyser la segmentation thématique dans des textes écrit surtout arabe, nous devons tout d'abord définir le terme de segmentation (التجزئة) pour pouvoir comprendre l'utilisation de la segmentation thématique Dans ce cadre, plusieurs définitions existent :

II.2.1. Dans Larousse

- Action de segmenter, fait d'être segmenté.
- Division d'une unité anatomique en plusieurs éléments.
- Ensemble des divisions successives de l'œuf fécondé qui aboutissent
- Division embryonnaire du corps des annélides et des arthropodes en une suite de segments, ou anneaux pluricellulaires, d'abord semblables, puis différenciés. [34]

II.2.2. Dans Reverso

Division en segments, en parties délimitées [35]

II.2.3. La segmentation en arabe :

هو تقسيم شيء وربطه بعضه ببعض حسب المواضيع وترابطها

II.3. La segmentation thématique

II.3.1. Définition De La Segmentation thématique

La segmentation thématique (التجزئة الموضوعية) peut avoir plusieurs définitions :

Selon [39] la segmentation thématique est la relation sémantique entre les éléments qui forment un texte en utilisant l'élément lexical élémentaire ou le mot. D'après avoir segmenter ce texte.

La segmentation thématique est une étape nécessaire et signifiante dans le traitement de la langue naturelle. La fonction d'un segmenteur est de couper un texte courant en segments, de sorte qu'ils puissent être introduits dans un capteur morphologique ou dans un étiqueteur de position pour un traitement ultérieur. Le segmenteur est responsable de définir des limites de mots, les clitiques délimitantes, les expressions pluri termes, les abréviations et les nombres.

La segmentation thématique est un sujet important dans le traitement de la langue naturelle car elle est étroitement liée « à l'analyse morphologique » .C'est beaucoup plus avec le cas des langues riches et complexes morphologiquement telle que l'arabe. Dans le cas de celle-ci où un mot simple peut comporter un lemme et jusqu'à trois clitiques, la connaissance morphologique doit être incorporée au segmenteur. [40]

La plus petite unité textuelle thématiquement cohérente en son sein et thématiquement distincte des unités textuelles précédentes et suivantes. L'unité atomique du segment thématique est la phrase. [28]

Or il est très difficile de détecter les frontières correctes des phrases dans un texte. D'où la segmentation formelle en paragraphes en s'appuyant sur certains signes semble moins ambiguë.

II.3.2. Le système d'écriture arabe

En TAL, on présente généralement les langues, quant à leur système d'écriture, comme appartenant à deux familles différentes : les langues « avec séparateurs » et les langues « sans séparateurs ». Les langues dites « avec séparateurs » sont celles qui ont des systèmes d'écritures segmentées c'est-à-dire des écritures délimitées par des espaces (space delimited writings) et où les mots sont nettement séparés par des délimiteurs (espace, signes de ponctuation, caractères spéciaux, ...). A ce type de langues on oppose les langues dites « sans séparateurs ». Elles présentent des systèmes d'écritures non segmentées (unsegmented writings) où les mots ne sont

pas séparés par des espaces et où les frontières des mots ne sont pas nettes. Le japonais, le chinois et le thaï sont les représentants parfaits de cette deuxième famille de langues [29].

La langue arabe présente un système d'écriture à l'intersection des deux familles. C'est un système d'écriture qui combine une écriture segmentée et une écriture non segmentée. En effet, une partie des mots graphiques arabes correspondent à des mots minimaux séparés par des délimiteurs. En revanche, une bonne partie des mots graphiques arabes sont composés d'une suite d'unités lexicales agglutinées analysable en termes de mots minimaux et de clitiques et qu'il faut donc segmenter si l'on veut arriver aux unités de base les composants. [29]

.

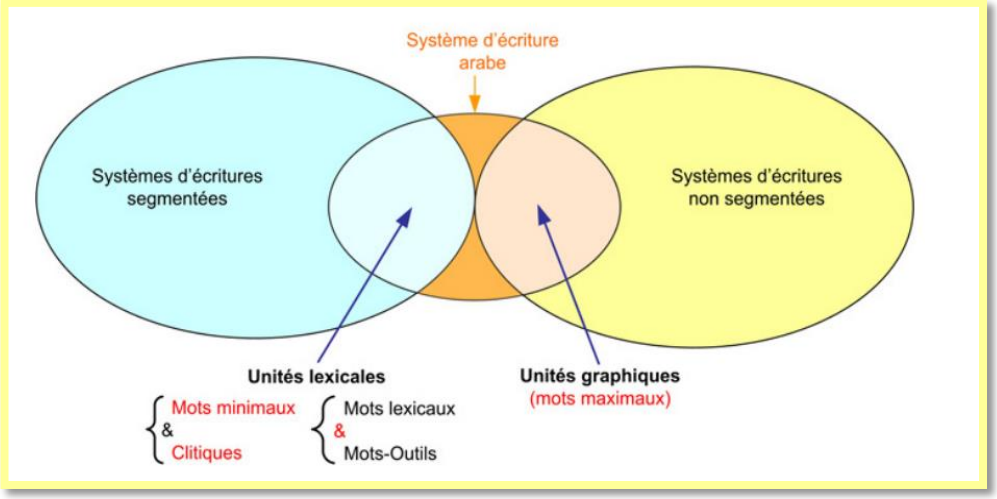


Figure II-1: Écritures segmentées et non segmentées.

II.3.3. Les types de la segmentation

Il existe plusieurs niveaux d'analyse signes d'intérêt permettent de repérer les différents éléments constituant le texte et en définir les frontières. On peut s'arrêter au niveau mot graphique, au niveau des unités lexicales ou aller au-delà de celles-ci pour arriver aux unités de base (les morphèmes).

Selon la visée de l'analyse à entreprendre : lexicale, morphologique, syntaxique ou thématique, on peut généralement trouver quatre types de segmentation :

- La segmentation lexicale (tokenization) qui est la segmentation d'un texte en segments lexicaux (tokens) ou mots. Ce type de segmentation est aussi appelé itémisation.
- La segmentation morphologique cherche à isoler les différents constituants des items lexicaux en unités distinctes, plus petites, qui sont les morphèmes.

- La segmentation syntaxique qui consiste à isoler les différents constituants du texte en unités indépendantes, supérieures aux mots, comme les propositions, les syntagmes … etc. Ce type de segmentation est aussi appelé chunking. [30]
- La segmentation thématique consiste à subdiviser un texte en un ensemble de segments thématiquement cohérents en se basant sur les relations lexicales et sémantiques entre ces derniers.

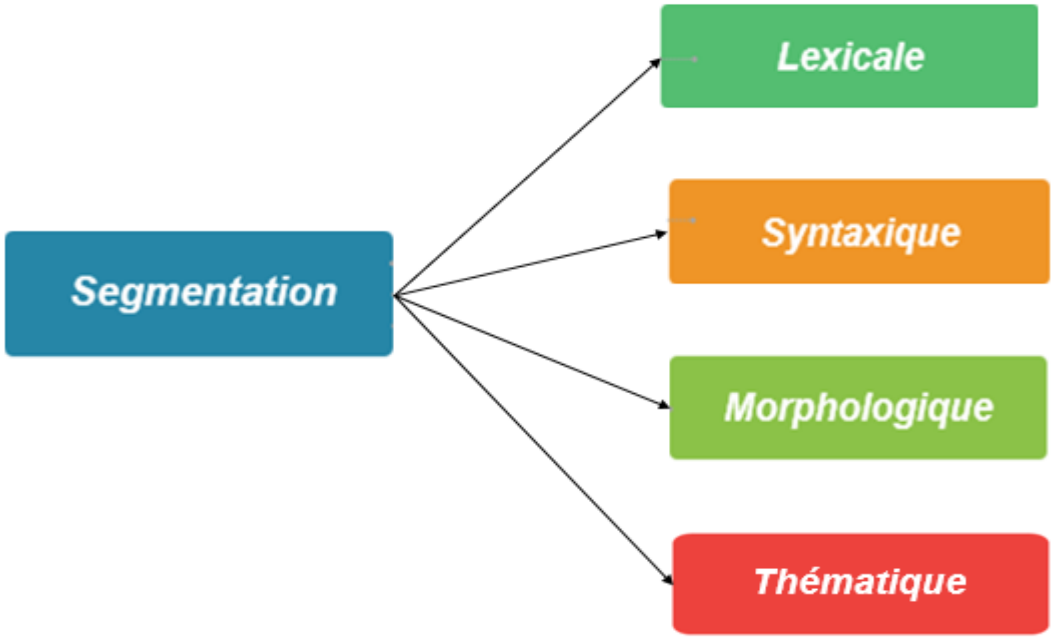


Figure II-2: Les types de segmentation.

II.4. Les approches de la segmentation

Selon plusieurs points de vue, une segmentation thématique peut être passive ou active, locale ou globale, linéaire ou non-linéaire, supervisée ou non-supervisée :

II.4.1. Segmentation passive et active

Lorsque l'on doit catégoriser des méthodes en TALN, on applique souvent le schéma classique de l'opposition supervisé contre non-supervisé. Cette catégorisation se base sur une différence méthodologique entre les approches et on peut la rapprocher des travaux de [Ferret, 2006] qui définit deux grandes catégories d'approches : celles exploitant des connaissances externes au texte (les approches exogènes) et celles n'exploitant que le contenu du texte (les approches endogènes).

Bien que satisfaisante, cette vision de la segmentation thématique ne s'intéresse qu'à la manière dont on procède, et non à l'objectif que s'est fixé le créateur d'une méthode. Nous avons choisi de nous intéresser de plus près à cet objectif.

Notre distinction entre segmentation active et passive vient donc de cette volonté de catégoriser les méthodes de segmentation thématique, non pas sur la base de leur méthodologie, mais sur la base de la motivation initiale, l'objectif de leur créateur. Au-delà de vouloir segmenter thématiquement le texte, on peut isoler deux visions de la segmentation thématique :

- Il y a les méthodes qui cherchent à regrouper les phrases similaires entre elles, formant ainsi des segments thématiques. Dans ces méthodes, les frontières apparaissent comme une conséquence de ce regroupement, elles sont donc retrouvées « par défaut », passivement. Ce sont les méthodes que nous appelons « à détection passive » des frontières.
- Il y a les méthodes qui cherchent à identifier les propriétés des frontières pour localiser ces dernières. Dans ces dernières, les frontières thématiques n'apparaissent pas par défaut, mais sont activement recherchées. D'où notre dénomination de « détection active » des frontières. [41]

II.4.1.1. Les approches passives

Les méthodes passives cherchent à regrouper les phrases en segments thématique et déduisent de ce regroupement la position des frontières. Parmi les plus représentatives des approches passives, mais aussi les plus originales des méthodes de segmentation thématique, on citera toute un groupe que nous appellerons : Méthodes graphiques. [41]

II.4.1.1.1. Les méthodes graphiques

En passant par une représentation graphique des termes, il est plus facile de visualiser leur répartition le long du document étudié. Ainsi la méthode du nuage de points, présentée par [Helfman, 1994] emploie cette représentation pour la recherche d'information. Le principe est de positionner sur un graphique chaque occurrence des termes du document. Dans cette représentation, un terme apparaissant à une position i et une position j du texte, sera représenté par les 4 couples (i, i), (i, j), (j, i) et (j, j). Les portions du document où les répétitions de termes sont nombreuses apparaîtront alors sur le graphique comme les zones de forte concentration de points.

Cette approche visuelle de la représentation d'un texte a été reprise et adaptée à la segmentation thématique par [Reynar, 1998] dans son algorithme DotP lotting. L'idée est d'identifier les segments thématiquement cohérents sur le graphique en cherchant les limites des zones les plus denses. La densité d'une région du graphique est calculée en divisant le nombre de points présents dans la région par l'aire de cette dernière. L'objectif de DotP lotting est d'isoler les segments thématiques soit en maximisant leur densité, soit en minimisant la taille des zones « vides » entre les segments. On notera que, dans son principe, cette méthode est très proche des méthodes utilisant des matrices de similarité présentées plus loin.

Cette approche a même inspiré des méthodes originales, comme celle proposée par [Ji & Zha, 2003], qui consiste à remplacer le problème de segmentation thématique par un problème de segmentation d'image. Cette méthode utilise une technique de diffusion anisotropique sur la représentation graphique de la matrice de distance afin de renforcer les contrastes entre les zones denses et les frontières. La transformation du texte en un ensemble de points, pour ensuite retrouver les segments thématiques en regroupant ces points en nuages correspond totalement à la description que nous avons faite des méthodes à détection passive. On regrettera que ces méthodes aient une approche très peu « linguistique » de la segmentation thématique. En effet, en réduisant le texte à une représentation graphique de ses termes, on perd toute notion de compréhension et on se contente alors de compter des mots. [42]

Plus traditionnelles que les méthodes graphiques et surtout plus répandues, les approches s'appuyant sur un calcul de distance ou de similarité sont surement parmi les plus populaires en segmentation thématique, Selon les types de relations cités ci-dessus, la similarité peut être distributionnelle ou sémantique.

II.4.1.1.2. 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.[43]

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 : [43]

II.4.1.1.2.1. Mesures de distribution

1. TF-IDF :

Le TF-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. [44]

Plusieurs fonctions existent pour le calcul de TF-IDF par exemple, l'équation de [45] cité ci-dessous, permet de comparer le comportement d'un terme candidat dans le document analysé avec son comportement dans une collection de documents ***:***

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

- **TF :** représente le nombre d'occurrences d'un mot dans le document analysé.
- **DF :** représente le nombre de documents dans lequel il est présent.
- **N :** étant le nombre total de documents. Plus le score TF-IDF d'un terme candidat est élevé, plus celui-ci est important dans le document analysé.

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.

2. OKAPI BM25

Okapi (ou BM25) [46] 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 TFBM25) :

$$\text{Okapi(terme)} = \text{TF}_{\text{BM25}}(\text{terme}) \times \log\left(\frac{\text{N} - \text{DF(terme)} + 0{,}5}{\text{DF(terme)} + 0{,}5}\right) \qquad (\mathbf{II.2})$$

$$\text{TF}_{BM25} = \frac{\text{TF(terme)} \times (k_1 + 1)}{\text{TF(terme)} + k_1 \times \left(1 - \text{b} + \text{b} \times \frac{\text{DL}}{DL_{\text{moyenne}}}\right)} \qquad (\mathbf{II.3})$$

- Dans la formule (II.3) :

- **k1 et b :** sont des constantes fixées à 2 et 0, 75 respectivement.
- **DL :** représente la longueur du document analysé.
- **DL moyenne :** la longueur moyenne des documents de la collection utilisée.

Notant qu'Okapi peut aussi être adapté pour considérer une phrase par rapport à un texte.

3. LIKEY :

Paukkeri et Honkela, [47] 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 [47]

$$\text{Likey (terme)} = \frac{\text{rang}_{\text{document}}\,(\text{terme })}{\text{rang}_{\text{corpus}}(\text{terme })} \qquad (\mathbf{II.4})$$

II.4.1.1.2.2. Mesures de similarité

1. 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. [48] 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 [49] qui l'utilisée dans son algorithme de segmentation thématique C99. La similarité entre deux phrases x et y est calculée par l'équation suivante :

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

- **fi,j** : dénote la fréquence du terme j dans la phrase **i** .

2. 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. [50]

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 et l'indice est :

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

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

$$J_\delta(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} \qquad \textbf{(II.7)}$$

3. 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) [51]. MI est généralement calculé comme suit :

$$I(x,y) = \log\left(\frac{p(x,y)}{p(x)p(y)}\right) \qquad \textbf{(II.8)}$$

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

II.4.1.1.2.3. 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).

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 [52] 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…). [52]

II.4.1.2. Les approches actives

Les approches actives cherchent à identifier les propriétés des phrases frontières pour les localiser. Cette démarche de recherche volontaire des frontières a bien entendu eu une grande influence sur les méthodologies employées. Aussi, peut-on s'attendre à une majorité de méthodes utilisant des ressources externes aux textes pour définir les propriétés des frontières thématiques (des approches exogènes selon la définition de [Ferret, 2006]).

Toutefois, nous allons voir que certaines méthodes tentent de retrouver les frontières thématiques, sans s'appuyer sur des données externes (et sont donc endogènes toujours selon [Ferret, 2006]).

Lorsque l'on cherche à identifier un type de phrase particulier, comme les phrases frontières par exemple, le premier réflexe que l'on a est d'apprendre à quoi ressemblent ces phrases sur un corpus pour construire un modèle, puis d'utiliser ce modèle pour les identifier. [41]

II.4.2. Les approche non-supervisée

Une approche « non-supervisée », quant à elle, s'appuie seulement sur l'information fournie par les données qu'elle doit traiter, et parfois, dans le TALN, sur des ressources linguistiques généralistes externes (thésaurus, dictionnaire, analyseur syntaxique, etc.). De telles approches sont plus généralistes et plus adaptables, mais rarement aussi efficaces qu'une méthode supervisée (lorsque cette dernière est appliquée à des données proches de ses données d'apprentissage). [31].

L'exemple le plus courant dans cette approche est la Segmentation à partir de chaines lexicales**,** cette dernière relie les occurrences multiples des mots dans un document et estime qu'une chaine

est rompue si la distance entre deux occurrences du même mot est trop importante. Cette distance est généralement exprimée en nombre de phrases. Ainsi, la méthode Segmenter présentée par Kan (Kan et al. 1998), procède selon ce principe pour effectuer une segmentation thématique du document étudie. On notera tout de même une subtilité. La distance à partir de laquelle l'algorithme considère qu'il y a rupture dépend de la catégorie syntaxique du mot implique dans le lien. Une autre approche basée sur les chaines lexicales est proposée par Hearst (M. A. Hearst 1997) avec son algorithme Text Tilling. Un score de cohésion est attribué à chacun des blocs de texte en fonction du bloc qui le suit. Il est quant à lui calculé sur la base d'un premier score dit (lexical) attribue à chaque paire de phrases en fonction de la paire de phrases qui la suit. Ce score lexical est lui-même calcul à partir des paramètres que sont le nombre de mots en commun, de mots nouveaux et de chaines Segmentation thématique par calcul de distance sémantique lexicale active dans les phrases considérées. Le score de chaque segment de texte est alors le produit scalaire normalise des scores de chacune des paires de phrases qu'il Contient. Si un segment présente un score différent des segments précédents et suivants, alors la rupture thématique se situe au sein de ce segment. Ces méthodes ne résolvent pas le problème de la taille variable des frontières et / ou de la localisation précise de ces dernières. [42]

II.4.3. Les approches supervisées

Les méthodes supervisées s'appuient sur un apprentissage utilisant une base de données d'apprentissage pour en déduire des règles. En segmentation thématique, ces méthodes utilisent un corpus d'apprentissage composé de textes dont les frontières thématiques ont été annotées. En utilisant ces annotations les méthodes supervisées construisent un modèle permettant de retrouver les frontières thématiques d'autres textes. Les principales différences entre ces méthodes viennent donc de ce qu'elles apprennent et de comment elles l'apprennent.

Dans la ligné des marqueurs linguistiques, on citera les travaux de [Passonneau & Litman, 1993, Passonneau & Litman, 1997] qui se sont attachés à construire des modèles pour trois grandes catégories d'indice linguistiques que sont les propositions nominales de références, les mots indicateurs et les pauses. Leur travail portant sur la segmentation de discours oraux, la notion de pause dans le discours, qui nous est inconnue dans un cadre purement textuelle, est importante. En dehors de cette particularité, ils construisent des modèles pour chacun de ces indices linguistiques en se basant sur des textes segmentés par des juges humains puis utilisent ces modèles pour retrouver les frontières thématiques. Leur originalité vient du fait que chacune des catégories de

marqueurs linguistiques est traitées différemment. Leurs travaux utilisant beaucoup d'information prosodique, ils sont difficilement applicables à un cadre textuelle n'incluant pas cette information.

Le problème de ces méthodes, c'est qu'elles héritent du travers reconnu des méthodes supervisées en général (quel que soit le domaine) : performantes sur des textes proches de leur corpus d'apprentissage, elles deviennent inefficaces dès qu'elles sortent de ce cadre. De plus les phases d'apprentissages peuvent parfois être lourdes et couteuses, surtout si peu de ressources sont disponibles pour l'apprentissage comme c'est souvent le cas en français. [31]

- Certaines approches combinent des ressources externes ou un apprentissage supervisé et une méthode endogène pour segmenter thématiquement. Nous appelons c'est méthodes « hybrides ».

II.4.4. Les approches hybrides

Parmi les méthodes hybrides, on citera [Ferret et al,2001] qui combinent une ressources externes, des marqueurs linguistiques génériques et un modèle statistique plus « classique » pour retrouver les frontières thématiques du texte. Cette approche utilise des expressions qui marquent le début des cadres thématique ([Charolles, 1997]) avec soit une analyse typiquement endogène comme celle proposée par [Hearst, 1997] (décrite plus loin), soit une méthode utilisant des ressources externes pour enrichir le texte et généraliser la notion de cohésion lexical à une notion plus étendue de cohésion du champ lexical. La force de cette approche est que les marqueurs ne sont pas juste identifiés par leur lexie, mais également grâce à un ensemble de règles (sept précisément) décrivant les conditions dans lesquelles le marqueur doit apparaître. [41]

II.4.5. Les approches linaires et hiérarchiques

Il existe également deux manières de segmenter un texte, soit linéairement (des portions de texte adjacentes), soit hiérarchiquement (on cherche à repérer les phrases correspondant à un même thème qui ne sont pas nécessairement adjacentes). [42]

II.5. Algorithmes de segmentation thématique

II.5.1. TextTiling

La méthode TextTiling est un algorithme de détection automatique des changements de sujet au sein d'un texte. L'objectif de ce dernier est de permettre un découpage du texte motive d'un point de vue sémantique.

II.5.1.1. Principe général de la méthode TextTiling

Comme son nom l'indique, la méthode TextTiling propose un découpage du texte en unités continues qui ne se superposent pas 2. L'idée générale dont découle cette méthode est qu'un texte traitant d'un sujet donne est en réalité divise en une articulation de sous-sujets. Chacun de ces sous-sujets est évoqué grâce à un vocabulaire spécifique. Par conséquent, la transition entre ces sous-sujets correspond à un changement d'une partie relativement importante du vocabulaire. Retrouver les frontières entre les diverses thématiques du texte équivaut des lors à détecter les changements importants en ce qui concerne le vocabulaire employé. Les seuls éléments pris en compte par TextTiling pour détecter un changement de thématique sont les schémas de co-occurrente et de distribution lexicale. Les éventuels autres indices figurant dans le texte (sa disposition typographique, la référence des pronoms, etc.) ne sont pas pris en compte. [32]

II.5.1.2. Les trois étapes de la méthode TextTiling

Les trois étapes essentielles qui constituent les méthodes TextTiling sont :

1-La tokenisation du texte en unités de la taille d'une phrase. La tokenisation peut se définir comme l'opération de découpage du texte à analyser en groupes de mots, dont on a éliminé les éléments annexes, tels que par exemple les signes de ponctuation. Dans le cadre de TextTiling, cette tokenisation doit englober une opération de lemmatisation.

2-La deuxième étape majeure des méthodes TextTiling est constituée du calcul de deux types de scores, à savoir les scores de similarité lexicale et les scores de profondeur. Un score de similarité lexicale est calculé pour chaque intervalle entre deux portions contigües de texte. Plus ce score sera ´élevé, plus le nombre d'unités lexicales communes entre les deux parties de texte sera important. Quant au score de profondeur, il correspond à une brusque chute du

score de similarité lexicale. En d'autres termes, si l'on représente sur un graphique l'ensemble des scores de similarité lexicale, ceux-ci vont dessiner une courbe plus ou moins accidentée. Les scores de profondeur correspondent aux vallées dessinées par cette courbe.

3- La détection des frontières entre les différents sous-sujets qui constituent le texte, grâce aux scores précédemment calculs. [32], A partir de ce score, le calcul d'un score de cohésion (ou de profondeur) est effectué qui quantifie la similarité entre un bloc et les blocs voisins. En terme de graphe de score de similarité, un score de cohésion peut etre représente comme la somme des différences entre le sommet du pic et les creux des valles voisines cf. Illustration 1. Le calcul des scores de cohésion procède comme suit :

- On commence au premier creux entre 2 blocs et, on mémorise le score de similarité associée avec les blocs de chaque cote du creux.
- On vérifie le score de similarité du creux précédant, si c'est plus haut, on continue et on examine le score de similarité du creux précédant.
- On continue jusqu'à ce que le score soit plus bas que celui déjà examiner.
- Ensuite, on soustrait le score de similarité du creux initial avec le score maximum de similarité rencontre.
- Cette procédure est répétée pour les creux entre les blocs suivant le premier Creux.
- Enfin, la somme des deux différences est calculée. [32].

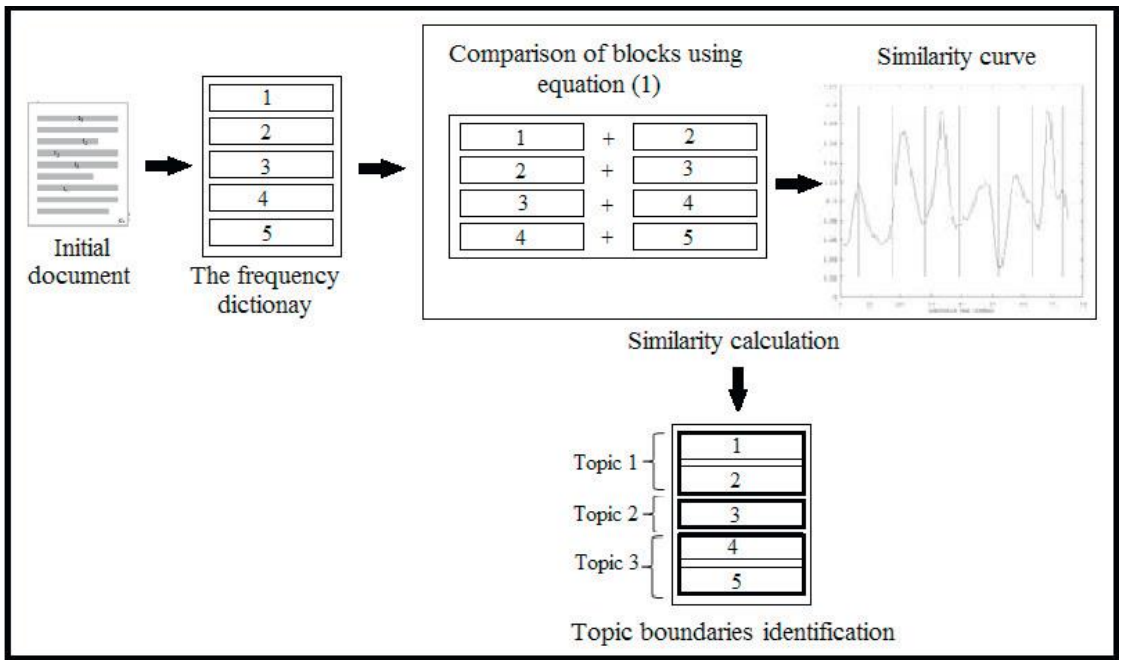


Figure II-3: Les étapes de TextTiling.

II.5.1.3. Avantages

L'algorithme de TextTiling est robuste et ne demande pas beaucoup de ressources pour être exécute. Son utilisation dans des systèmes d'extraction d'information donne de bons résultats. La détermination des ruptures se fait dans un contexte local, c'est-à-dire que l'algorithme détermine la rupture en fonction de ce qu'il y a avant et après et pas dans sa globalité. [53]

II.5.1.4. Inconvénients

L'utilisation de la distribution (idf) dans la mesure de la similarité ne permet pas de traiter des textes sous forme de flux. De plus, la présence dans le texte de beaucoup de titres, de petits paragraphes engendrent de mauvais résultats car l'algorithme de TextTiling ne prend pas en compte l'organisation hiérarchique. [53]

II.5.2. Dot Plotting

L'algorithme que nous utilisons est proposé par (Reynar, 2000), et est en fait une adaptation pour la segmentation de la méthode des nuages de points présentée par (Helfman, 1994) il se base sur la cohésion lexicale pour la recherche d'information. Et Il se base sur une représentation graphique du texte par les positions des occurrences des termes du texte à segmenter. Lorsqu'un terme apparaît à deux positions du texte x et y, les quatre points (x, x),(x, y),(y, x) et (y, y) sont représentés sur un graphe, ce qui permet de déterminer visuellement les zones du texte où les répétitions sont nombreuses. Cette méthode est adaptée par (Reynar, 2000) à la segmentation thématique de textes. Les positions de début et de fin des zones les plus denses du graphe sont les limites des segments thématiquement cohérents. La densité est calculée pour chaque unité d'aire en divisant le nombre de points d'une région par l'aire de cette région. A partir de là, deux algorithmes peuvent déterminer les frontières thématiques : identifier les limites en maximisant la densité au sein des segments, ou repérer la configuration qui minimise la densité des zones entre les segments. [41]

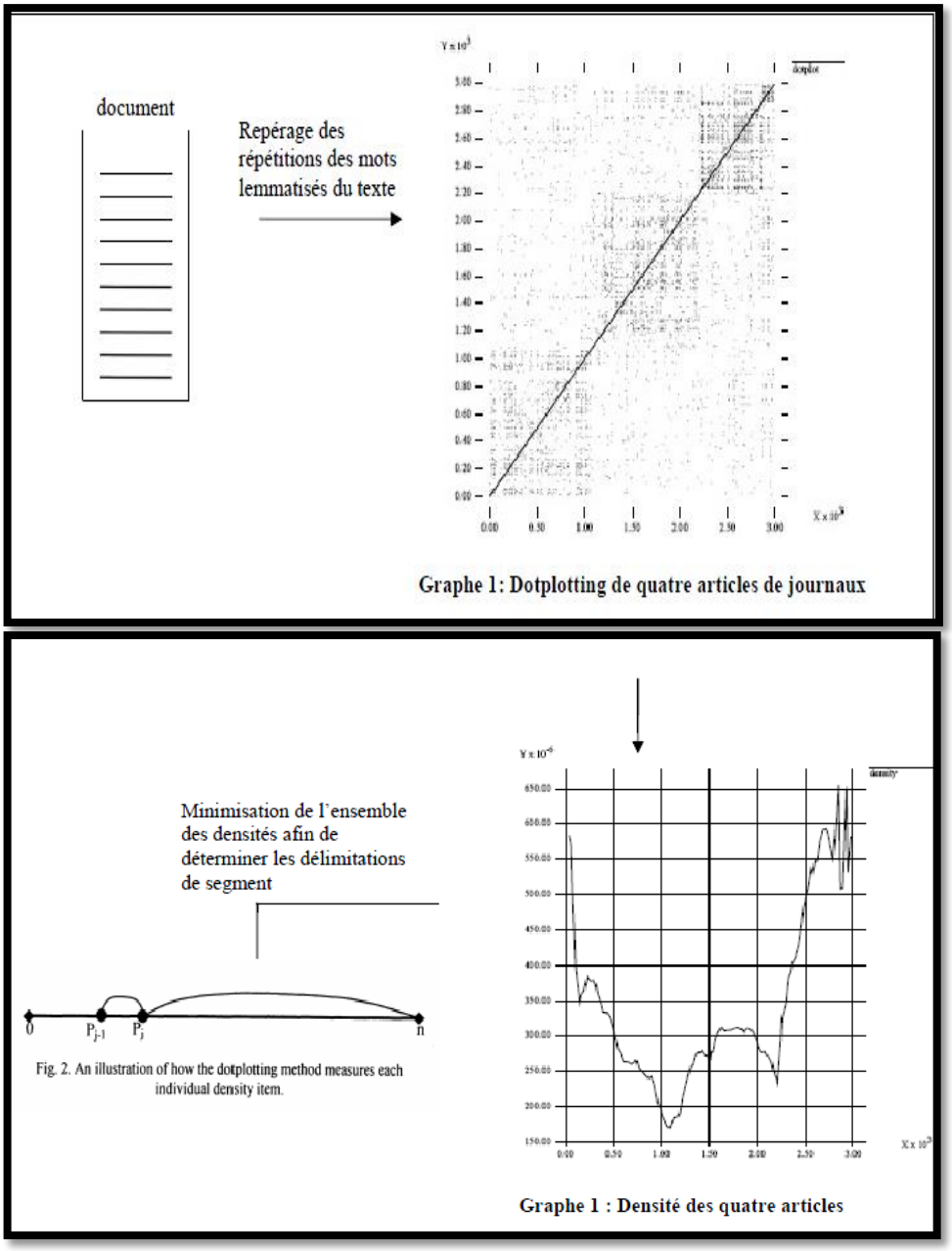


Figure II-4: Les étapes de Dot Plotting.

II.5.2.1. Les étapes de Dot Plotting

1- le découpage physique, Le texte a subi le prétraitement et les fins de phrases ou de paragraphe sont conservées.

2- Le texte est parcouru et à chaque répétition de mot, quatre points sont dessines sur le Graphique. Par exemple, pour un mot répète à la ligne x et la ligne y, les coordonnées des quatre points sont (x, x), (x, y), (y, x), (y, y), Pour plus de lisibilité, le texte n'a pas été traite c'est-à-dire que tous les mots ont été gardés et non lemmatises.

3- Un algorithme de maximisation peut être mis en place pour déterminer les délimitations des segments bien qu'elles soient visibles très nettement sur le graphe lorsque la densité est forte ce qui revient à mettre en place un algorithme de minimisation de la densité des régions extérieures aux quarts de forte densité.
4- Une fois, ces densités calculées, l'algorithme commence par sélectionner une délimitation qui aboutit à la densité extérieure aux régions de forte densité la plus faible. Des délimitations supplémentaires sont ajoutées jusqu'à ce que soit la densité extérieure croit soit le nombre de segments soit atteint. Les délimitations possibles sont sélectionnées à partir d'une liste soit de phrases soit de fin de paragraphe au choix de l'utilisateur.

II.5.2.2. Avantages

Il offre une représentation graphique du texte et permet de mettre en évidence les différents segments thématiques. Le calcul de la similarité n'utilise pas la distribution ce qui permet avec quelques adaptations de l'appliquer à des flux. L'algorithme permet le choix de l'utilisation de la structure du texte comme des paragraphes pour définir les ruptures de thème. [53]

II.5.2.3. Inconvénients

La détermination d'une rupture se fait en fonction de la limite précédemment déterminée et le reste du texte, de ce fait la rupture n'est pas déterminée ni de manière globale ni de manière locale. [53]

II.5.3. L'algorithme C99

Cet algorithme proposé par (Choi, 2000) utilise une mesure de similarité entre chaque unité textuelle. L'idée de base de cette méthode est que les mesures de similarité entre des segments de textes courts sont statistiquement insignifiantes, et que donc seul des classements locaux sont à considérer pour ensuite appliquer un algorithme de catégorisation sur la matrice de similarité. Dans un premier temps, une matrice de similarité est donc construite, représentant la similarité entre toutes les phrases du texte à l'aide de la mesure de similarité proposée par (Rijsbergen, 1979), calculée pour chaque paire de phrases du texte, en utilisant chaque mot commun entre les phrases, et après « nettoyage » du texte : suppression des mots vides et lemmatisation. On effectue ensuite un « classement local », en déterminant pour chaque paire d'unités textuelles, le rang de sa mesure de similarité par rapport à ses $m \times n - 1$ voisins, $m \times n$ étant le masque de classement choisi. Le rang est le nombre d'éléments voisins ayant une mesure de similarité plus faible, conservé sous la forme d'un ratio r afin de prendre en compte les effets de bord. [33]

$$r = \frac{\text{rang}}{\text{nombre de voisins dans le masque}} \qquad \textbf{(II.9)}$$

Enfin, la dernière étape détermine les limites de chaque segment de la même manière que l'algorithme Dotplotting emploie la maximisation. En effet on cherche à déterminer quelle configuration offre la plus grande densité, en recherchant une nouvelle limite thématique à chaque étape. Les segments sont alors représentés par des carrés le long de la diagonale de la matrice de similarité modifiée avec les classements locaux. Pour chaque segment de la répartition proposée à une étape de la segmentation on considère son aire notée **ak** et son poids **sk** qui L.Sitbon, P. Bellot est la somme des tous les rangs des phrases qu'il contient. On calcul alors la densité **D** de la configuration avec :

$$D = \frac{\sum_{k=1}^{m} s_k}{\sum_{k=1}^{m} a_k} \qquad \textbf{(II.10)}$$

L'algorithme s'arrête lorsque la densité de la meilleure répartition proposée est suffisamment faible, ou si le nombre de frontières thématiques est déjà déterminé, lorsqu'il est atteint. [33]

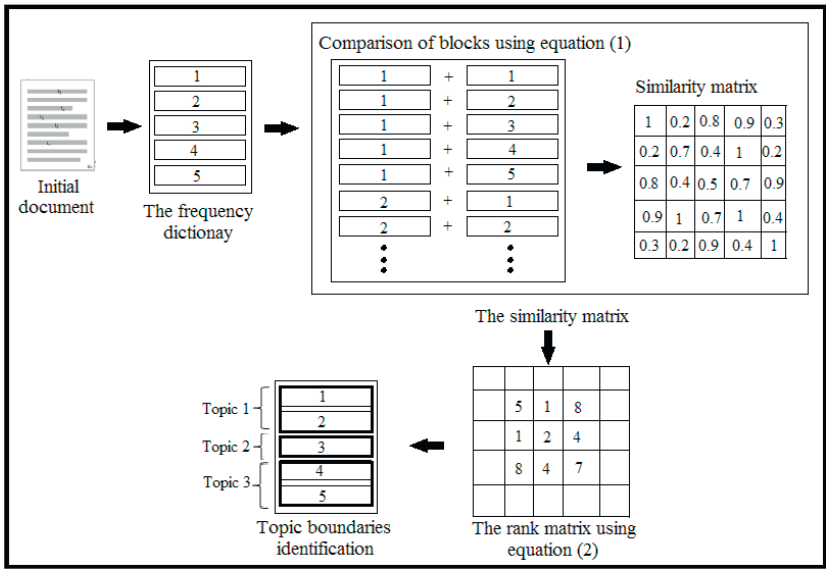


Figure II-5: Les étapes de C99.

II.5.3.1. Avantages

L'algorithme C99 donne aussi une représentation du texte afin de visualiser les différents segments thématiques. C99 donne des bons résultats sur des textes contenant de petits paragraphes car il ne prend pas la mesure de similarité de façon absolue mais de façon relative. La distribution des mots dans le texte n'est pas prise en compte ce qui permet la gestion des flux.

II.5.3.2. Inconvénients

La détermination des changements de thèmes est fondée sur le principe de maximisation utilise par Reynar, les mêmes problèmes résolus dans l'amélioration se posent notamment la dissymétrie selon dans quel sens est pris le texte. Cependant la mise en place de la matrice de rang attenue le problème de prise en compte du contexte global.

II.5.4. Segmenter

Segmenter (Kan et al.1998) effectue une segmentation linéaire basée sur les chaînes lexicales présentes dans le texte. Ces chaines relient les occurrences des termes dans les phrases. Une chaîne est rompue si le nombre de phrases séparant deux occurrences est trop important. Ce nombre dépend de la catégorie syntaxique du terme considéré. Une fois tous les liens établis, un poids leur est assigné en fonction de la catégorie syntaxique des termes en jeu et de la longueur du lien. Un score est ensuite donné à chaque paragraphe en fonction des poids et des origines des liens qui le traversent ou qui y sont créés. Les marques de segmentation sont alors apposées au début des paragraphes ayant les scores maximaux. Etant donné qu'un concept peut être désigné par un ensemble de mots, le concept de chaînes lexicales a été élargi aux chaînes conceptuelles à l'aide de WordNet (Fellbaum, 1998) ou d'autres ressources sémantiques. (Kan et al.1998) montre que l'amélioration est très peu significative.

II.5.5. Minimum cut segmenter

Proposé par [Malioutov et Barzilay, 2006], Minimum cut segmenter utilise la théorie définie par [Shi et Malik, 2000] qui représente un texte par un graphe. La première étape consiste à détecter la similarité entre les phrases en utilisant l'algorithme DotPlotting, ensuite, les relations entre unités (phrases) sont représentées par un graphe de similarité ***G={V,E}*** :

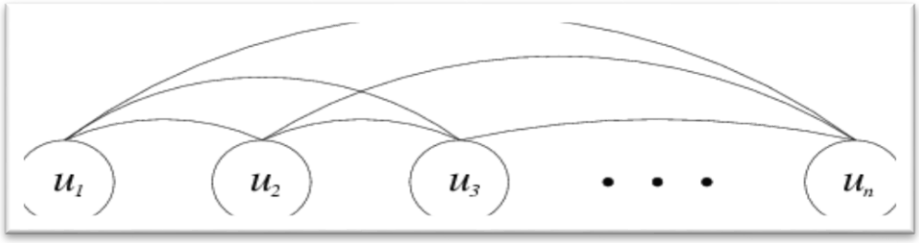


Figure II-6: Graphe de similarité.

$\boldsymbol{V}$ représente l'ensemble des nœuds (phrases), $\boldsymbol{E}$ représente l'ensemble des frontières pondérées. Une frontière pondérée $\boldsymbol{w(u,v)}$ définie la mesure de similarité entre deux nœuds$\boldsymbol{u}$ et $\boldsymbol{v}$.

La procédure de segmentation consiste à segmenter le graphe en deux classes disjointes de nœuds **A** et **B** en maximisant leur dissimilarité avec l'équation ***cut*** suivante :

$$cut(A,B) = \sum_{u\in A, v\in B} \omega(u,v) \qquad \textbf{(II.11)}$$

.

Cette procédure est répétée récursivement pour chaque classe jusqu'à obtenir une segmentation minimale (petites classes disjointes).

II.5.6. Segmenteur Bayésien

Bayesian segmenter [Eisentein & Barzilay 2008], est un algorithme basé sur la théorie d'allocation latente de dirichlet (Latent Dirichlet allocation) implémentée par [Blei et al., 2003]. L'objectif est d'utiliser la cohésion lexicale dans un contexte bayésien. La théorie utilise les modèle linguistique multinomial pour effectuer une segmentation Formellement, un modèle linguistique multinomial est défini comme suit :

Si une phrase t est dans un segment j alors le sac de mots X_t est tiré du modèle linguistique θ_j .Les pauses thématiques se produits alors dans les frontières de phrases. Chaque phrase t est alors assignée à un segment thématique z.

La probabilité de cette observation pour toutes les phrases du texte est donnée par :

$$p(X|z,\theta) = \prod_t^T p\left(X_t\middle|\theta_{z_t}\right) \qquad \textbf{(II.12)}$$

X est l'ensemble de tous les T phrases, z est le vecteur d'assignement des phrases aux segments et Θ est l'ensemble de tous les K modèles linguistiques.

II.5.7. HAPS (Hierarchical Affinity propagation for segmentation)

Introduit par [Kazantseva et Szpakowicz, 2014], HAPS est un algorithme de segmentation linéaire de textes. Il utilise la théorie *Affinity Propagation* définie par les mêmes auteurs [Kazantseva et Szpakowicz, 2011]. L'objectif et de construire un arbre thématique basé sur un calcul de similarité. Une segmentation multi-niveaux est faite et dont les nœuds de haut niveau représentent les changements thématiques les plus saillants. Les nœuds du bas niveau représentent les changements thématiques faibles (vers la stabilité).

Le principe est d'identifier pour chaque segment dans l'arbre son centre (phrase ou paragraphe) qui donne la meilleure description de son contenu.

Le processus de segmentation HAPS se déroule comme suit :

- Calculr une matrice de similarité entre unités atomiques (phrases ou paragraphes) dans les segments.

- Introduire le Nombre de niveaux souhaités dans l'arbre thématique.

- Introduire une valeur de préférence pour chaque niveau de l'arbre (cette valeur permet déterminer le centre des segments).

- Calculr la somme des similarités entre les différents centres de segments dans les différents niveaux en utilisant l'algorithme *k-means*.

- Dérouler l'algorithme *HAPS* en utilisant les graphes de facteurs.

- La figure suivante représente un fragment de deux niveaux d'un graphe de facteurs **:**

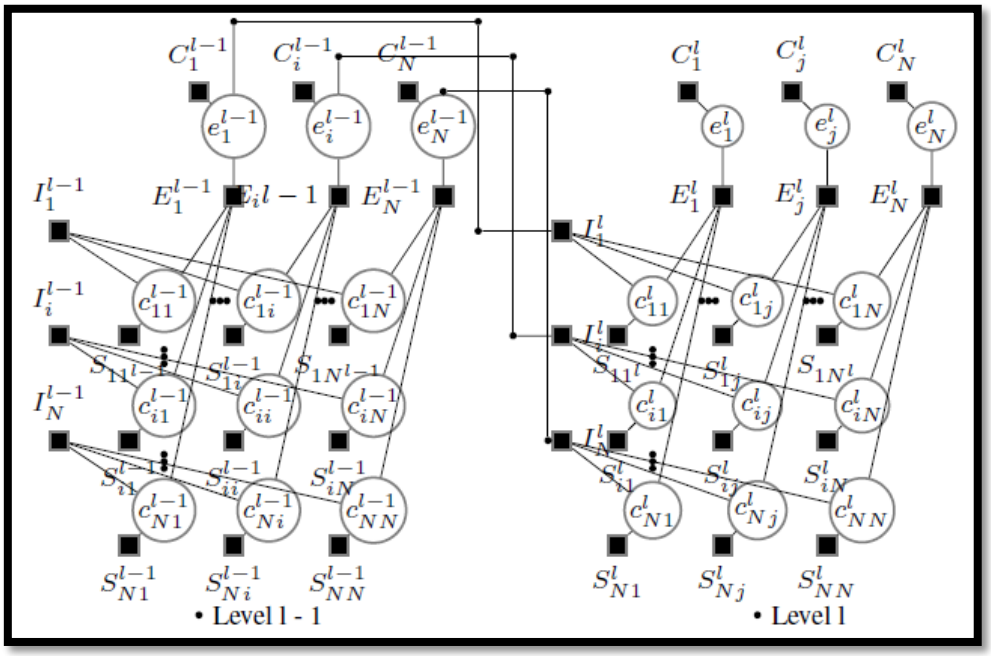


Figure II-7: Fragment d'un graphe de facteurs pour les deux niveaux l-1 et l.

II.6. Travaux relatifs en langue arabe

La plupart des systèmes de la segmentation thématique ou autrement dit étude des relations entre phrases traitent des textes des langues latines. Le peu de travaux enregistrés relativement en langue arabe concerne des segmentations formelles en paragraphes ou phrases. D'autre part, les travaux qui ont touché à la segmentation thématique se contentaient de faire des comparaisons de quelques algorithmes après application aux textes arabes. La détection des relations entre les segments dans ces travaux est généralement statistique et basée sur la répétition des occurrences des termes. Les travaux recensés dans ce cadre sont :

- [54] : 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.
- [55] et [56] 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).
- [57]: Propose une approche symbolique basée sur la relation entre segments noyaux et satellites en utilisant RST qui l'ont appliquée aux résumés automatiques.

- [58] : 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 segmentation.
- STAr [36] est un segmenteur de textes arabes basé sur l'approche de segmentation thématique. Il est réalisé avec le langage de programmation Perl. Il accepte en entrée un texte arabe en format .txt et génère en sortie un texte segmenté thématiquement en paragraphes et en phrases.
- Segmenter [41] effectue une segmentation linéaire basée sur les chaînes lexicales présentes dans le texte arabe. Ces chaines relient les occurrences des termes dans les phrases.
- (Belguith et el, 2005)[37] ont proposé une approche à base de règle pour segmenter des textes arabes non-voyelles en phrase. L'approche consiste en une analyse contextuelle des signes de ponctuation, des conjonctions des coordonner et une liste de particules qui sont considérer comme des critères de segmentation.
- Réalisation du système SEGA de SEGmentation de textes Arabes non voyellés permettant le découpage d'un texte (de type .txt) en paragraphes et en phrases. Ce système est réalisé dans le laboratoire LARIS. Il est basé sur l'étude des contextes gauche et droit des signes de ponctuation, de quelques conjonctions de coordination telle que le waw و ainsi que d'autres connecteurs.

II.7. Evaluation de la segmentation thématique

Évaluer un système de la segmentation thématique est délicat. De nombreux problèmes sont soulevés, et peuvent grosso modo être ramenés à deux questions :

- Quelle référence ? (b) Quel score d'évaluation ? [59]
- Pour évaluer un tel système, il faut le comparer à une segmentation de référence.

II.7.1. 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.

II.7.2. 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 calcul 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.

II.7.3. Evaluation automatique

Pour évaluer de tel système de façon automatique, d'autres scores ont été proposés, dont les plus usités sont les mesures Pk [60] et WindowDiff [61]. 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.

II.8. Conclusion

La segmentation thématique d'un texte a un rôle crucial dans plusieurs traitements automatiques, tel que la recherche d'information ou le résumé automatique, etc. L'amélioration de la manière de segmenter thématiquement un texte influe positivement sur les résultats rendus par les autres traitements automatiques des textes écrits.

Dans de nombreuses applications, la segmentation thématique est une première étape avant l'identification du sujet et dont l'objet est la classification ou le regroupement des thèmes réels discutés au sein de chaque segment.

En effet, nous pensons que la combinaison de plusieurs critères (par exemple, distributionnels et sémantiques) pour la détection les relations qui peuvent exister entre paragraphes améliorera majoritairement les résultats rendus par processus automatique de segmentation thématique.

La nature et le format des textes à segmenter ont un rôle très important dans le choix d'algorithme de segmentation. Les algorithmes de segmentation linéaire donnent de meilleurs

résultats dans le cas où les idées sont exposées séquentiellement dans le texte, cependant, les textes ne sont pas toujours ainsi organisés car les auteurs peuvent développer les mêmes idées dans plusieurs reprises et dans différentes positions dans le texte d'où une segmentation en blocs contigus n'est pas toujours parfaite. Les algorithmes de segmentation hiérarchique (non-linéaires) quant à eux, produisent des segments thématiques dépendants les uns des autres et ne fournissent pas des sous thèmes distincts d'où la segmentation est considérée problématique dans certaines applications évitant les redondances.

La partie qui suit sera consacrée à la conception proprement dite du système de segmentation, cette partie sera marquée par le choix d'un corpus de textes écrits en arabe standard moderne. La catégorie de textes étudiée est une collection d'articles économiques publiés dans une revue économique. Notre système sera basé sur une approche globale, non-linéaire et non-supervisée.

Chapitre III :

Conception

III. Conception

III.1. Introduction

La segmentation thématique constitue en effet une très bonne solution, Elle sert à identifier le maximum possible de relations entre différents segments d'un document, afin de regrouper les plus cohérentes les unes avec les autres. Ainsi des relations distributionnelles réitératives ou sémantiques sont utilisées.

Notre système de segmentation thématique, sera basé sur une approche de segmentation globale, non-linéaire et non-supervisée. La similarité entre segment est calculée par les méthodes distributionnelles de détection de cohésion lexicale. Les textes étudiés sont des articles économiques, une catégorie de textes écrits en arabe standard moderne non-voyellé.

La technique de détection de cohésion lexicale que nous utilisons est basée sur une combinaison de critères statistiques (distributionnels) et linguistiques (sémantiques). En effet, le peu de travaux existants relativement à la langue arabe, se basent essentiellement sur la répétition des termes lexicalement proche.

Nous pensons que l'ajout d'autres critères comme la synonymie, améliore un tel système de segmentation thématique.

Afin de réduire les redondances lexicales et sémantiques, nous utilisons une technique de clustering strict appliquée sur une matrice de similarité entre segments.

III.2. Description de l'approche adoptée

L'approche que nous proposons pour détecter automatiquement des segments reliés thématiquement dans d'un texte arabe est une approche globale, non-linéaire et non supervisée, 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 un terme avec toute les autres termes du texte, au contraire des approches locales qui comparent seulement les mots voisins.

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

L'approche est dite non-supervisée, en effet, le système de segmentation ne dispose d'aucune information ou référence sur la segmentation objective, il se base seulement sur un calcul statistique distributionnel et tente de trouver une meilleure segmentation.

La catégorie de textes que nous étudions est une collection d'articles économiques arabes publiés dans une revue moderne (الاقتصاد المعاصر) éditée par la faculté des sciences économiques et de gestion de l'université de Khemis Miliana

Les critères statistiques sont la distribution des occurrences des termes (fréquences) et la position des termes dans les paragraphes du texte. L'importance des termes (vis-à-vis la répétition et la synonymie) est quantifiée en utilisant la métrique OKAPI qui est basée sur le calcul des fréquences d'apparition et de la distribution des termes dans le texte.

Un score final est attribué à chaque terme du texte en fonction de son poids OKAPI et de son poids par rapport à sa position relative dans le paragraphe, Les termes sont classés en fonction de leurs scores finaux.

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 « Arabic Stemmer », un outil linguistique simple à intégrer et qui utilisent le célèbre algorithme de « Khoja Stemmer ».

Pour trouver les synonymes des termes, nous utilisons une base de données lexicale riche et accessible « Arabic WordNet ».

Pour mesurer la similarité (taux de cohésion) entres les segments, nous calculons deux score Cosinus qui mesure la similarité entre les pairs de segment vis-à-vis la répétition des termes et la distribution des synonymes des termes dans le texte.

Une matrice d'adjacence sera construite et dont les éléments sont les valeurs de similarité Cosinus déjà calculées entre les segments. Cette matrice peut être représentés finalement par un graphe G=(S,A), S étant les sommets qui représentent les segments et A les arêtes qui représentent les liens entre les segments.

Afin de construire les sous-groupes finaux qui constituent les sous thèmes, nous utilisons une méthode de clustering strict. Les différents segments thématiques seront finalement représentés graphiquement.

III.3. Présentation du Corpus

Nous avons choisi corpus d'étude une collection d'articles économiques publiés dans la revue économique internationale « Economie contemporaine », 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 en .TXT ou DOC est 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 économiques et comme tout autres type d'articles spécialisés utilisent généralement un même vocabulaire dans plusieurs reprises dans le texte ce qui 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'où nous avons choisi le racineur (Arabic Stemmer) comme analyseur morphologique.

D'une autre part, un même topic peut être discuté dans des segments ou paragraphes séparés, ce qui favorise l'utilisation d'une approche globale qui consiste à comparer un segment avec tous 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 paragraphes (segmentation formelle). Cette dernière étant le segment minimal représentant un sens complet. Nous

nous sommes basé sur le signe de ponctuation ‘le point (.)’ pour marquer les frontières des paragraphes. En effet, l’utilisation d’autres séparateurs comme la virgule ou les expressions langagière comme le و / waw peut générer des défauts de segmentation car ils sont considérés comme des frontières ambiguës et ne jouent pas toujours le rôle de séparateurs.

III.4. Les Outils Utilisés

III.4.1. L’analyseur Arabic Stemmer

Nous avons obtenu une version Java simplifiée de Shereen Khoja stemmer qui est Arabicstemmer, ce dernier 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 la liste de tous les caractères diacritiques, caractères de ponctuation, des articles précis, et 168 mots vides.

- Afin de calculer une racine, Arabic stemmer suit cette procédure suivante:

1. Retirer le diacritique représentant la vocalisation.

2. Retirer les stopwords, la ponctuation et les chiffres.

3. Retirer article défini "ال"

4. Retirer le conjointement inséparables "و".

5. Retirer les suffixes.

6. Retirer les préfixes.

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

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

9. Remplacer les lettres faibles (حروف العلة) " ي,و,ا " par "و"

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

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

Exemple : Comment l'analyseur de Khoja analyse le mot : يعلمكم ?

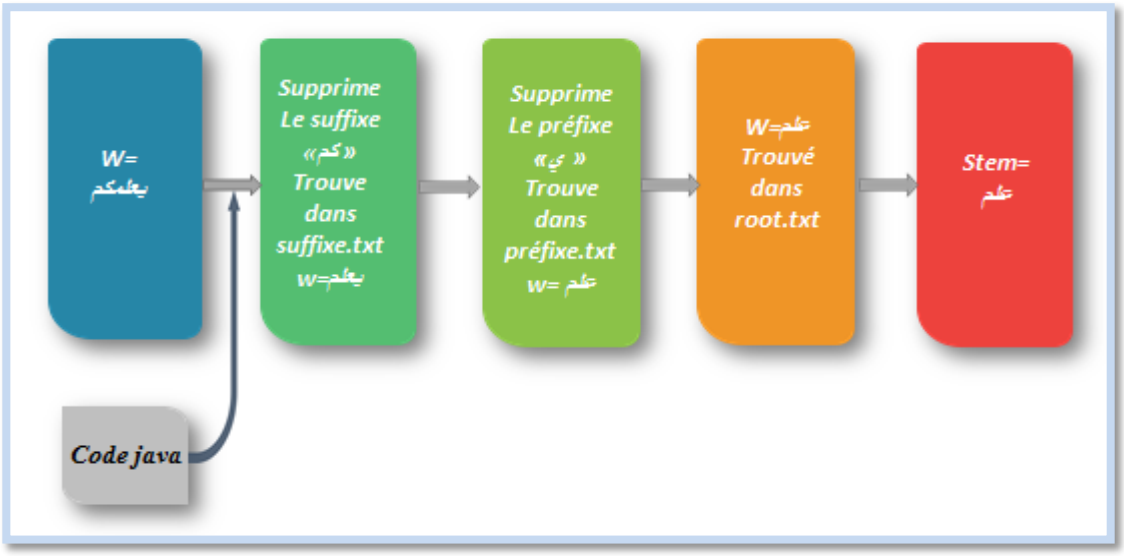


Figure III-1: Analyse du mot -يعلمكم– par Arabic Stemmer.

III.4.2. ArabicWordNet (AWN)

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 (Suggested Upper Merged Ontology). ArabicWordNet a été développé par DOI / REFLEX (2005-2007).

La base de données AWN est librement et publiquement disponible. L'ontologie Arabic WordNet contient 9228 concepts (6252 nominals et 2260 verbals, 606 adjectival, et 106 adverbals), contient 18,957 expressions et 1155 concepts nommés. Le fichier base de l'AWN est sous format **XML** et il contient les quatre balises suivantes :

• **Item** : Contient les concepts, 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.

III.4.2.1 Utilisation d'AWN

A partir des deux balises <form> et <word>, la recherche des liens sémantiques se fait comme suit :

- **<form value="شري" wordid="$Ariy_1" type="root" authorshipid="11270" />**
- **<word wordid="$Ariy_1" value="شارِي" synsetid="mu$otarK_n1AR" frequency="" corpus="" authorshipid="11270" />**

Après avoir entré la racine du mot à chercher :

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"

- La figure ce dessus résume l'utilisation de la base de données lexicale AWN :

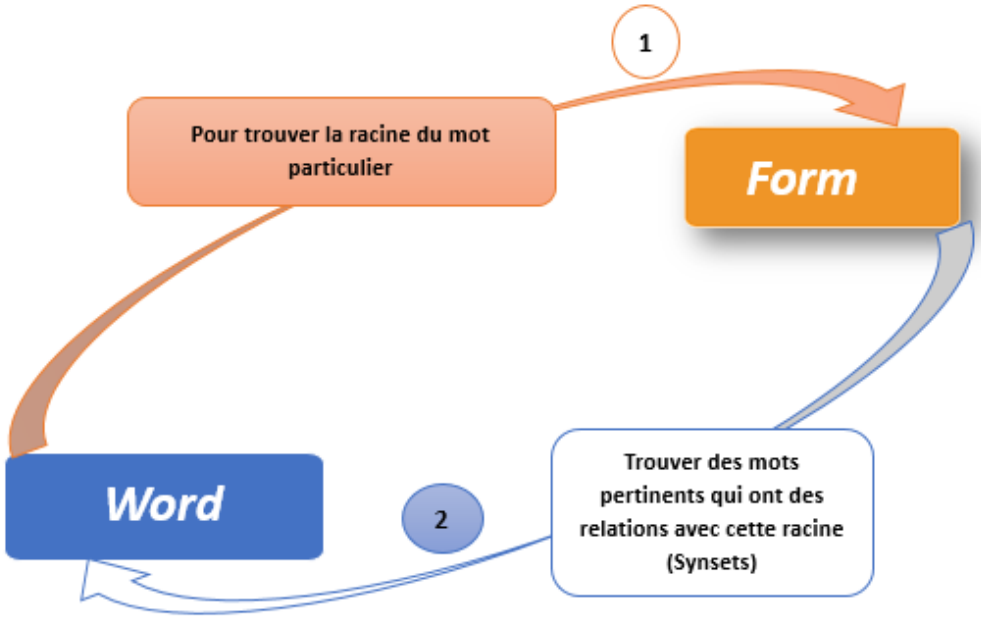


Figure III-2: L'utilisation de la base de donné lexical AWN.

Dans la figure suivante, nous avons montré comment utilisé arabic stemmer pour extraire les racines des mots ainsi que les racines des synsets :

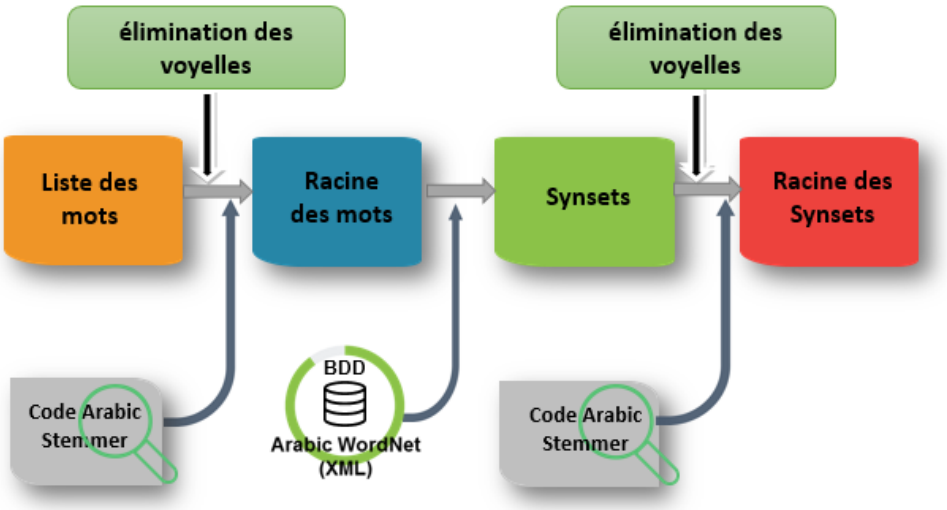


Figure III-3: L'utilisation d'Arabic stemmer pour extraire les racines des mots et les racines des synsets.

III.5. Architecture du système

Notre système de segmentation est basé principalement sur un calcul distributionnel (OKAPI) et une mesure de similarité (COSINUS) entre pairs de segment textuels. La mise en œuvre fonctionnelle du système est représentée par la figure III.4

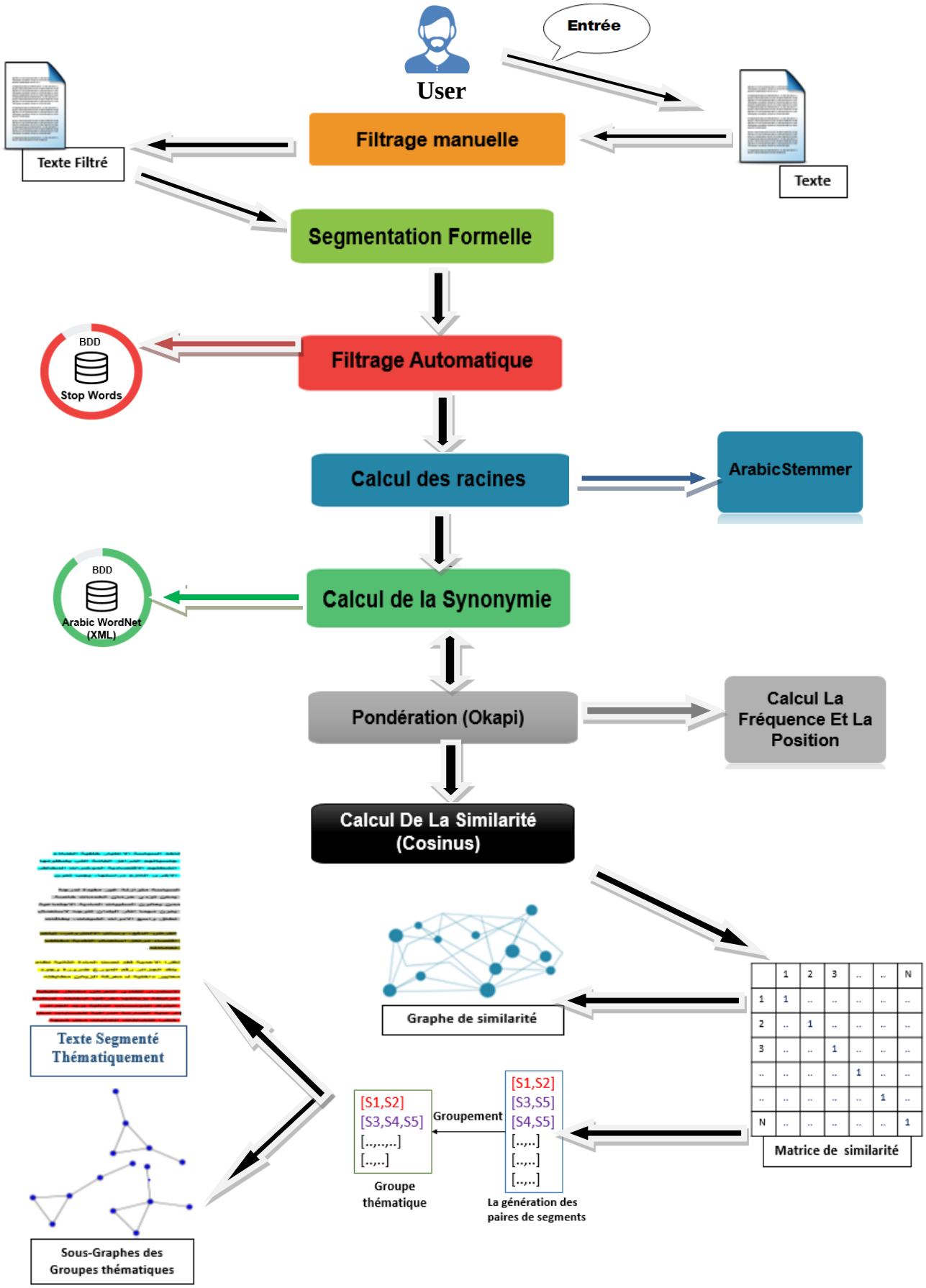


Figure III-4: Architecture du système.

III.6. Description détaillée des différents modules du système

III.6.1Pré-traitements

III.6.1.1 Filtrage manuel

Le filtrage manuel des articles commence par la suppression des figures et des schémas qui sont inutiles car nous nous intéressons au contenu textuel, Le fichier d'entrée peut être un (.txt) ou Word (.doc). Nous aurons en sortie un fichier filtré.

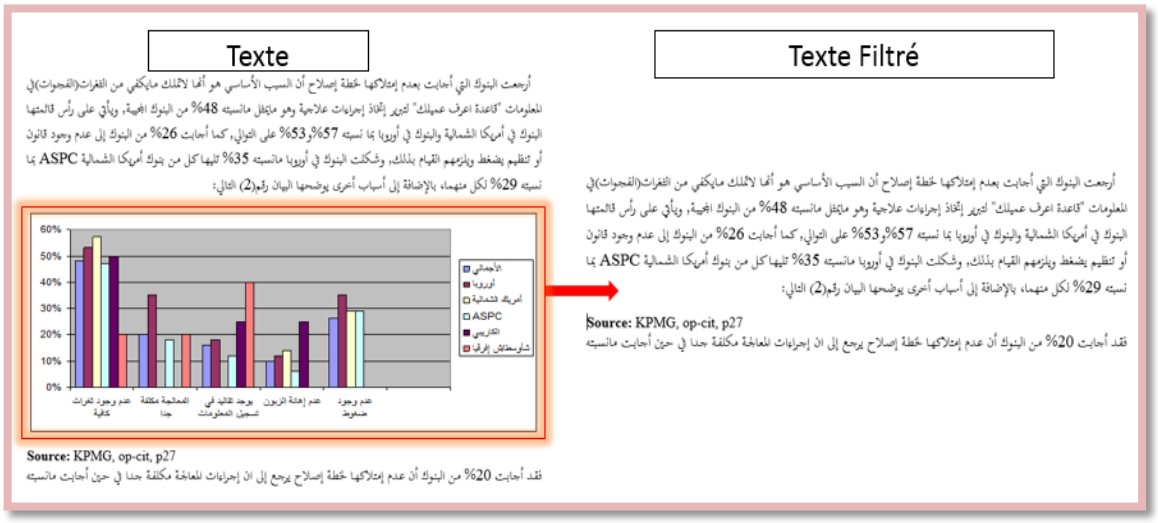


Figure III-5: Exemple de filtrage manuel.

III.6.1.2. codage des textes

L'encodage unique des textes en format standard, permet de représenter les textes sans aucune déformation au niveau de caractère lors de lecture. Tous les textes de notre corpus sont représentés avec un encodage UTF-8.

III.6.1. La segmentation formelle

La segmentation est une étape nécessaire et signifiante dans tout traitement de la langue naturelle. La fonction d'un segmenteur est de couper un texte courant en segments.

Dans notre approche nous avons opté à une segmentation en paragraphe. Le paragraphe étant le segment qui garantit la couverture d'une unité thématique a sens complet. Une frontière d'un paragraphe est facilement détectée par un point et/ou un saut de ligne.

Par contre Il y a des cas particuliers des points (.) à éviter pour segmenter le texte. Avant de considérer le point comme frontière d'un paragraphe, on le compare avec des éléments dans la base de données contenant tous les cas de figure du point comme la montre la figure suivante :

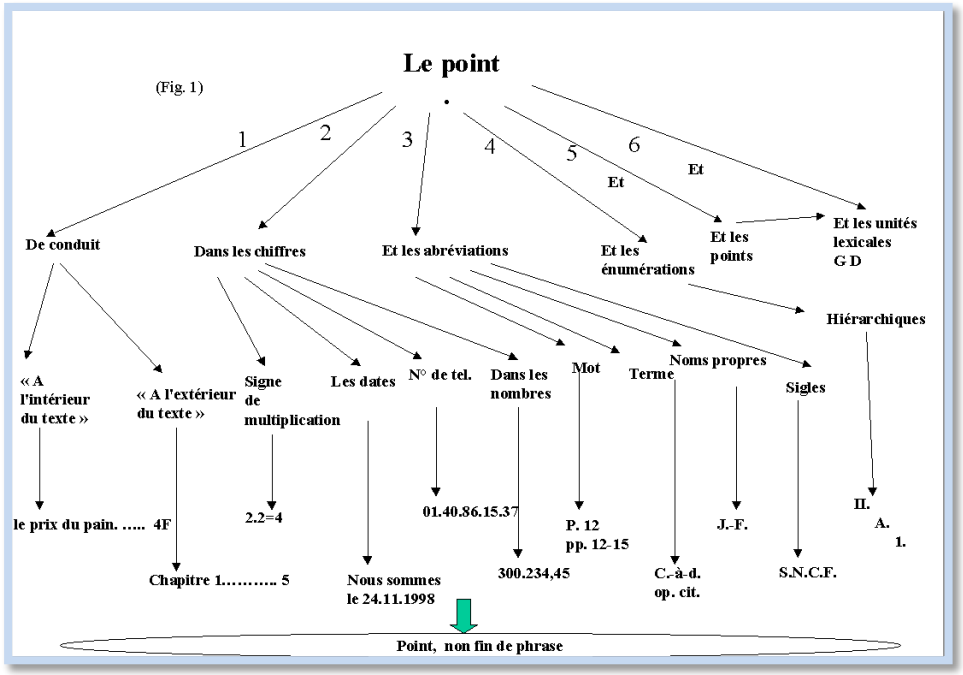


Figure III-6: Les cas de figure du caractère point (.)**.**

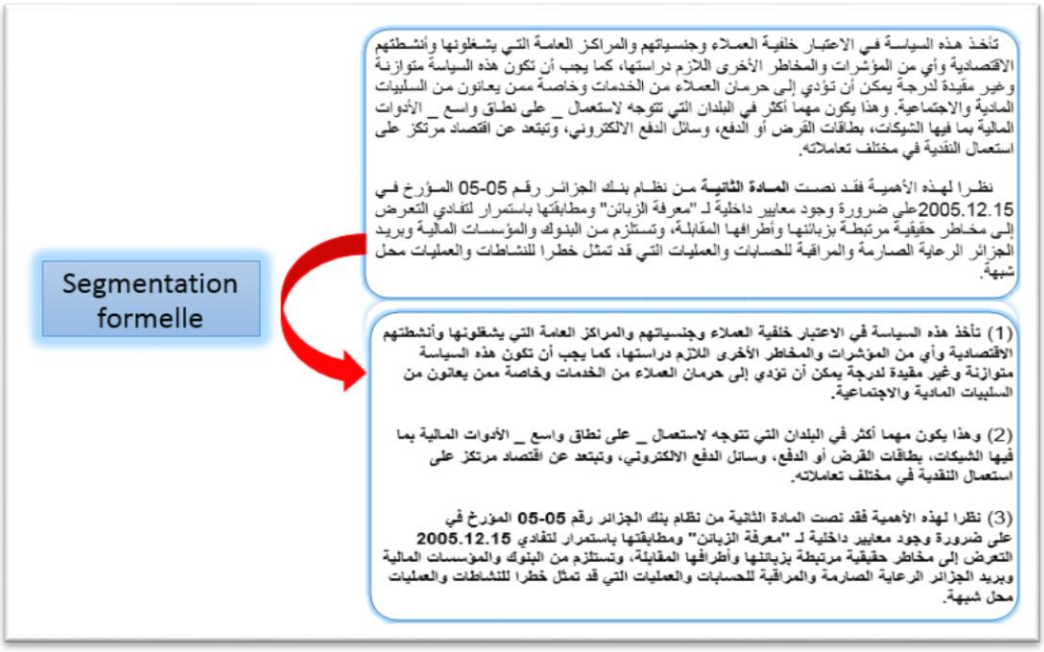


Figure III .7 : Exemple d'une segmentation formelle

III.6.2. Le filtrage automatique

Le filtrage automatique consiste à prendre chaque segment de l'étape précédente (le paragraphe) et éliminer tous les mots non significatifs. Pour chaque mot reconnu, on le compare avec un des éléments dans la base de données qui contient tous les mots non-significatifs. Si un mot en fait partie, il ne sera pas pris en considération pour le calcul de sa fréquence et on le supprime directement. La base de données regroupe tous les particules et/ou les mots vides (stopwords).

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

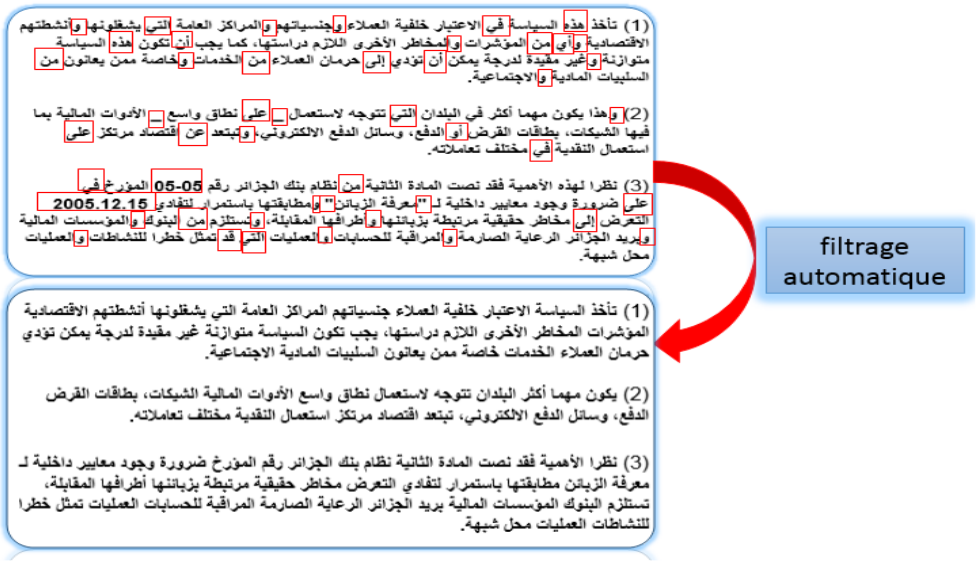


Figure III-7: Exemple de filtrage automatique.

III.6.3. Traitement linguistique

Afin de manipuler les variations du texte qui peuvent être représentées en arabe, on applique plusieurs genres de normalisation sur le texte. Par exemple, dans l'arabe écrit, les voyelles sont souvent omises dans les textes, néanmoins, on peut parfois trouver quelques voyelles présentes avec les mots. Alors, l'élimination de ces voyelles est nécessaire pour fin de normalisation. Certaines lettres subissent une simple modification dans l'écriture qui n'influe pas considérablement sur le sens du mot. Mais l'encodage de ces lettres change d'un mot à un autre. Une autre raison pour ce prétraitement est que l'on a tendance fréquemment à mal écrire ces différentes formes de hamza. Ce genre d'erreurs est très répandu dans les textes arabes.

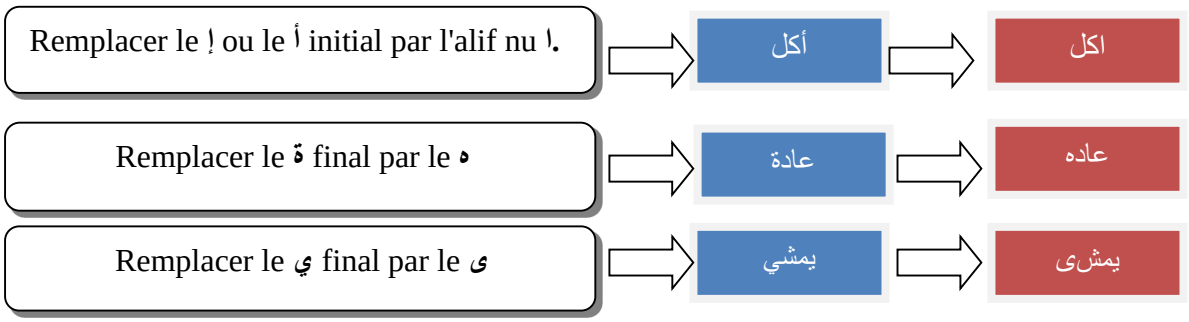


Figure III-8: Normalisation des lettres.

La liste des signes de ponctuation, des signes diacritiques, et des non-lettres est celle utilisée par analyseur *Arabic Stemmer.*

III.6.3.1. Analyse morphologique (Stemming)

Un des traitements les plus importants pour la langue arabe, en vue de la recherche d'information et d'indexation est la lemmatisation des mots. L'objectif de la lemmatisation est de trouver la forme représentative d'index d'un mot à partir de sa forme représentée dans le document et réduire son espace caractéristique par l'application de la troncature des affixes.

Pour un mot significatif normalisé, on applique une lemmatisation profonde en utilisant l'analyseur Arabic Stemmer 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 termes.

إن تطوير روح الإبداع المحاسبي في المؤسسات مرهون بمدى توفر مجمو عة من الشروط والعوامل لعل أهمها المناخ الملائم ومدى التحفيزات التي تقدمها المؤسسة لمبدعيها بالخصوص المحاسبين المبدعين. كما أن تسيير نشاطات الإبداع قد يختلف من مؤسسة لأخرى إلا أنه وماهو متفق عليه، فإن احترام العلاقات الإنسانية في تسيير الإبداعات المحاسبية وترقيتها.

كما تظهر أهمية الإبداع في كونه يعطي للمؤسسة القدرة التنافسية على مواجهة منافسيها وهذا على المستوى الجزائي بينما على المستوى الكلي يعتبر الإبداع أداة رئيسية لتحقيق التنمية الاقتصادية.

Figure III-9: Calcul de distribution des termes à partir de leurs racines.

Comme on remarque dans le paragraphe le mot « ***الإبداع*** » il a plusieurs schèmes mais ils ont le même contexte, dans ce cas il est mieux de prendre la racine du mot que prendre le lemme.

III.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. Afin de mesurer l'importance d'un terme dans le texte, nous avons utilisé la métrique Okapi (voir chapitre02), une nouvelle mesure qui vient de remplacer la célèbre tf-idf.

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 le métrique *okapi.* Le score de synonymie est quantifié en utilisant encore une fois *okapi* adaptée au calcul de distribution des synonymes d'un terme.

Pour calculer l'importance d'un terme dans un document, Okapi considère un document par rapport une collection de document (voir Chapitre 2). Afin de considérer un segment par rapport à un ensemble de segment dans un texte, Okapi peut être calculée ainsi :

$$\text{Okapi}(\text{terme}) = TF_{BM25}(\text{terme}) \times \log\left(\frac{N - DF(\text{terme}) + 0{,}5}{DF(\text{terme}) + 0{,}5}\right) \quad (\mathbf{III.1})$$

$$\boldsymbol{TF\ BM25} = \frac{\boldsymbol{TF(terme)} \times (\boldsymbol{k_1} + \mathbf{1})}{\boldsymbol{TF(terme)} + \boldsymbol{k_1} \times \left(\mathbf{1} - \boldsymbol{b} + \boldsymbol{b} \times \frac{\boldsymbol{DL}}{\boldsymbol{DL_{moyenne}}}\right)} \quad (\boldsymbol{III.2})$$

- Tel que :
- **N** : Nombre de segment dans le texte.
- **TF** : représente nombre d'occurrences d'un terme dans le segment courant (nombre d'apparition du terme lui-même ou l'un de ses synonymes).
- **DF** : représente le nombre de segments dans lequel il est présent (nombre d'apparition du terme lui-même ou l'un de ses synonymes).
- **DL** : La langueur du segment analysé.

- **DL moyen** : La langueur moyenne des segments du texte.
- **K1 et b** : Constants « 2 », « 0,75 » respectivement.

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

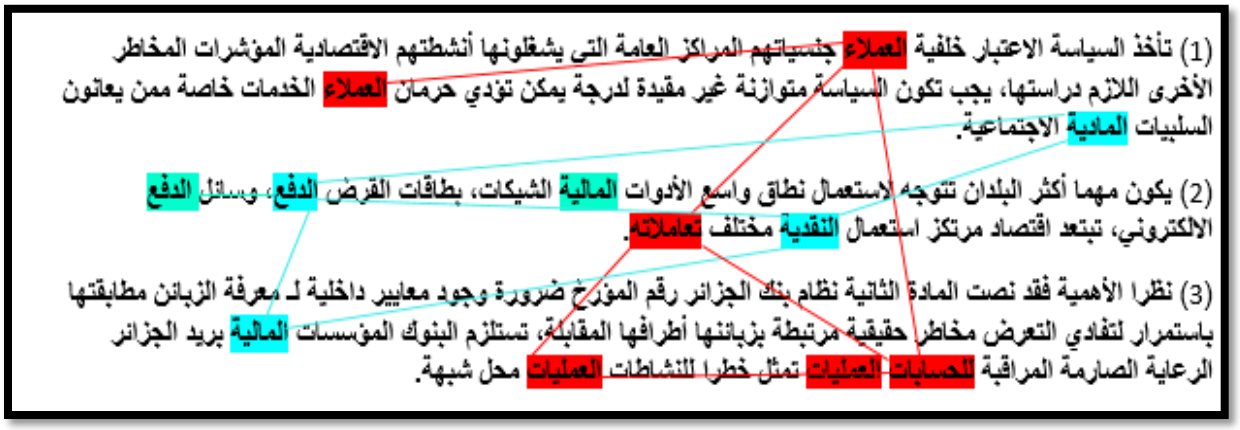


Figure III-10: Exemple des fréquences pondérées.

III.6.5. Calcul de la similarité

Cette étape consiste à calculer la similarité entre les paires de segments en utilisant la mesure Cosinus (voir chapitre 2). La version standard de Cosinus calcul la similarité sur la base des fréquences d'apparition des termes dans une paire de segments. Nous proposons une nouvelle version de Cosinus qui prend en considération les fréquences pondérées au lieu de simples fréquence *f*. Cette nouvelle version mesure non seulement le nombre de relations répétitives entre segments mais aussi l'importance de ces relations. Pour un système qui calcul la pertinence des segments vis-à-vis les relations, une telle mise à jour de Cosinus améliore fortement les résultats.

L'équation du Cosinus mise à jour qui calcul la similarité entre deux segments (S1,S2) est calculée par l'équation suivante :

$$Cosinus(S1, S2) = \frac{\sum_{j \epsilon S1 \cap S2} OKAPI_{S1,j} \times OKAPI_{S2,j}}{\sqrt{\left(\sum_{j \epsilon S1} {OKAPI_{S1,j}}^2\right) \times \left(\sum_{j \epsilon S2} {OKAPI_{S2,j}}^2\right)}} \quad (\mathbf{III.3})$$

$OKAPI_{Si,j}$ dénote le score Okapi du terme *j* dans le segment *Si*

Après avoir calculé Cosinus entre les paires de segments, une matrice de similarité SIM est construite. Cette dernière est une matrice d'adjacence carrée et dont les lignes et les colonnes sont les segments S. Un élément SIM [i,j] représente la valeur de similarité Cosinus entre les deux segments i et j.

	S1	S2	S3	……	Sn
S1	**SIM [1,1]**	**SIM [1,2]**	**SIM [1,3]**	**……**	**SIM [1, n]**
S2	**SIM [2,1]**	**SIM [2,2]**	**SIM [2,3]**	**……**	**SIM [2, n]**
S3	**SIM [3,1]**	**SIM [3,2]**	**SIM [3,3]**	**……**	**SIM [3, n]**
…	**……..**	**…….**	**……**	**……**	**……**
Sn	**SIM [n, 1]**	**SIM [n, 2]**	**SIM [n, 3]**	**……**	**SIM [n, n]**

Table III-1: Matrice de similarité.

Cette représentation vectorielle de la matrice d'adjacence peut être présentée par un graphe ***G=(S, A)***, S étant les sommets qui représentent les segments et A les arêtes qui représentent les liens de cohésion entre les segments. Et on finira par un texte segmenté en sous thèmes à partir de regroupement thématique.

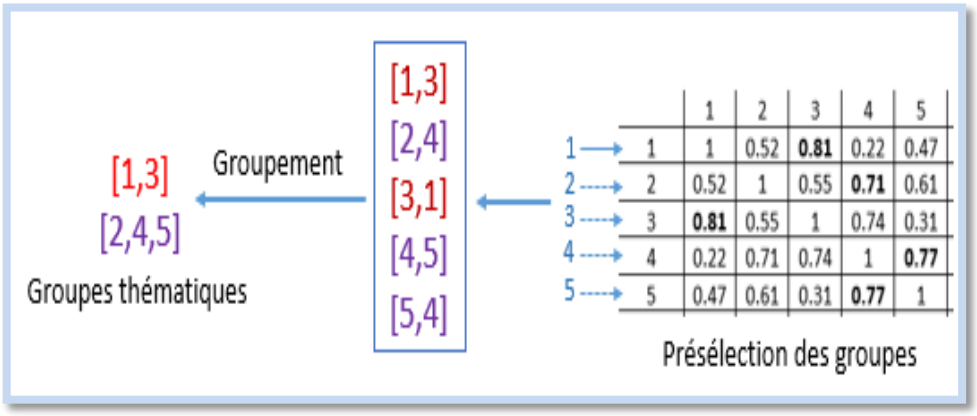
III.6.5.1. Sélection des paires de segments

Afin d'évaluer le degré de la cohésion entre un segment i et les autres segments j et d'après le calcul de la valeur de similarité SIM [i, j] on fait la présélection des groupes en comparant chaque valeur de similarité du segment i avec les autres segments, et on

prend la valeur maximale du SIM [i, j], ensuite on fait le groupement pour avoir des paires de segment cohérents.

Nous pouvons réduire le nombre d'arrêtes dans le graphe ***G(S, A)*** en ne gardant que les relations de cohésion les plus importantes.

La génération des paires de segments et le groupement avec clustering strict à partir de la matrice de similarité est schématisé par la figure suivante :



	1	2	3	4	5
1	1	0.52	**0.81**	0.22	0.47
2	0.52	1	0.55	**0.71**	0.61
3	**0.81**	0.55	1	0.74	0.31
4	0.22	0.71	0.74	1	**0.77**
5	0.47	0.61	0.31	**0.77**	1

Figure III-11: Calcul des groupes thématiques.

Dans cet exemple on prend le premier segment (1) on choisit dans la ligne 1 de la matrice la valeur de cohésion la plus élevée 0.81 qui représente le degré de cohésion avec le segment 3.), ainsi la première paire est formée du segment 1 et 3. Nous répétons cette procédure avec toute la ligne de la matrice.

III.6.5.2 Groupement avec Clustering strict

D'après la sélection des paires de segments, nous avons obligé de former les groupes thématiques avec le clustering strict afin d'éliminer les segments redondants en passant par les étapes suivantes :

1) Sélectionner une paire de segment et la Comparer avec les autres paires afin de former des groupes ayant des segments en communs.
2) Eliminer les segments redondants dans les groupes.

3) Comparer le groupe obtenu (cluster) à nouveaux avec les paires de segments restants.
4) Passer à la paire de segments suivante.
5) Refaire les étapes à partir de 1.

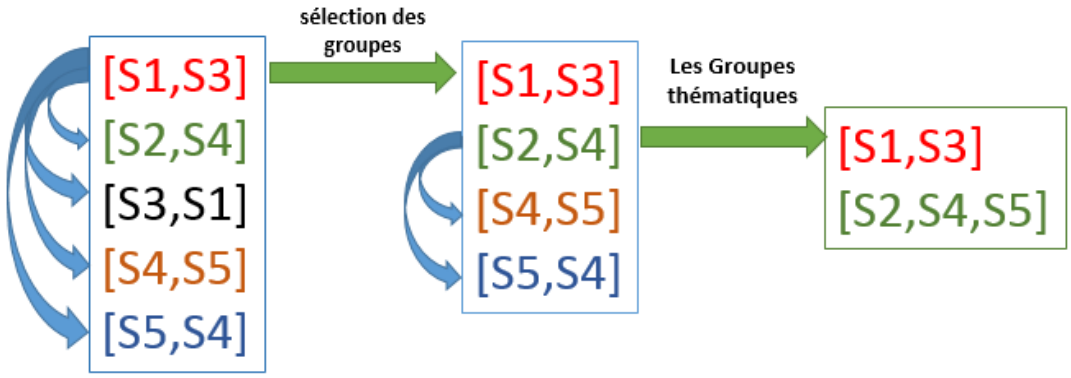


Figure III-12: Clustering des paires.

Finalement, nous obtiendrons des groupes de segments distincts qui formeront les segments thématiques.

III.7. Conclusion

Dans ce chapitre, nous avons présenté l'approche adoptée pour le développement de notre système d'automatisation de la segmentation thématique des textes arabes et plus précisément les articles économiques, en décrivant les outils utilisés et l'architecture du notre système avec une description des différents modules de traitement, Cette conception sera mise en fonction dans le chapitre qui suit.

Chapitre IV :

Implémentation Et Evaluation

IV. Implémentation et évaluation

IV.1. Introduction

Il est évident que les méthodes et les outils choisis pour concevoir et développer une application doivent être en fonction de l'environnement et du domaine d'application de celle-ci. Cela est bien expliqué par le génie logiciel.

Ce chapitre nous permet de montrer le déroulement de notre application dans sa phase de test et d'essai, ce qui nous permettra d'envisager les améliorations possibles. Les difficultés rencontrées nous ont permis de dégager un ensemble de pistes susceptibles de nous aider.

IV.2. L'environnement de travail

Pour que notre travail atteigne l'objectif qu'on visait, on a pris l'initiative d'exploiter et d'implémenter notre programme sur la version: Windows 7 Ultimate Edition SP1 64 BITS et afin de mener à bien ce projet, nous avons utilisé un ensemble de matériels dont les principales caractéristiques sont les suivantes :

- **Machine numéro 01:**
  - CPU: Intel® Core™ i3-2120 CPU @ 3.30 GHz.
  - RAM: 4 GO.
  - DISQUE DUR: Capacité 1 TO.

- **Machine numéro 02:**
  - CPU: Intel® Core™ i7-4500 CPU @ 3.30 GHz.
  - RAM: 8 GO.
  - DISQUE DUR: capacité 1 TO.

Figure IV-1: Matériel informatique utilisé**.**

IV.3. Outils de développement

On présente dans cette section, le langage de programmation Java utilisé, et l'environnement NetBeans ainsi que la base de données MySql.

IV.3.1. Environnement de développement

IV.3.1.1Java

Notre application a été codée en sa globalité par le langage JAVA, c'est un langage de programmation récent (les premières versions datent le 1995) développé par Sun MicroSystems. Il est fortement inspiré des langages C et C++.

Comme C++, Java fait partie de la grande famille des langages orientés objets. IL répond donc aux trois principes fondamentaux de l'approche orienté objet (POO): l'encapsulation, le polymorphisme et l'héritage.

- Java a rapidement intéressé les développeurs pour les raisons principales:

- C'est un langage orienté objet dérivé du C, mais plus simple à utiliser et plus « pur » que le C++. On entend par le « pur » le fait qu'en Java, on ne peut faire que la programmation orienté objet contrairement au C++ qui reste un langage hybride, c'est- à-dire autorise plusieurs styles de programmation. C++ est hybride pour assurer une compatibilité avec le C.

- Il est doté, en standard, de bibliothèques de classe très riches comprenant la gestion des interfaces graphiques (fenêtres, boites de dialogue, contrôles, menus, graphisme), la programmation multithreads (multitâches), la gestion des exceptions, les accès aux fichiers et au réseau…L'utilisation de ces bibliothèques facilitent grandement la tâche du programmeur lors de la construction d'applications complexes.

- Il est doté, en standard, d'un mécanisme de gestions des erreurs (les exceptions) très utile et très performant.

- Il est multi plates-formes : les programmes tournent sans modification sur tous les environnements où Java existe (Windows, Unix et Mac).

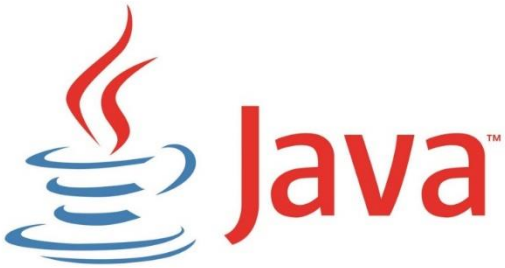


Figure IV-2: Le langage de programmation java.

IV.3.1.1. IDE: NetBeans

C'est un environnement de développement intégré (IDE) pour Java, placé en open source par Sun en juin 2000 sous licence CDDL (Common Development and Distribution License). En plus de Java, NetBeans permet également de supporter différents autres langages, comme Python, C, C++, XML 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).

NetBeans est disponible sous Windows, Linux, Solaris (sur x86 et SPARC), Mac OS X et Open.

- L'environnement de base comprend les fonctions générales suivantes :

- configuration et gestion de l'interface graphique des utilisateurs.
- support de différents langages de programmation.
- traitement du code source (édition, navigation, formatage, inspection..).
- fonctions d'import/export depuis et vers d'autres IDE, tels qu'Eclipse ou JBuilder.
- accès et gestion de bases de données, serveurs Web, ressources partagées.
- gestion de tâches (à faire, suivi …).
- documentation intégrée.

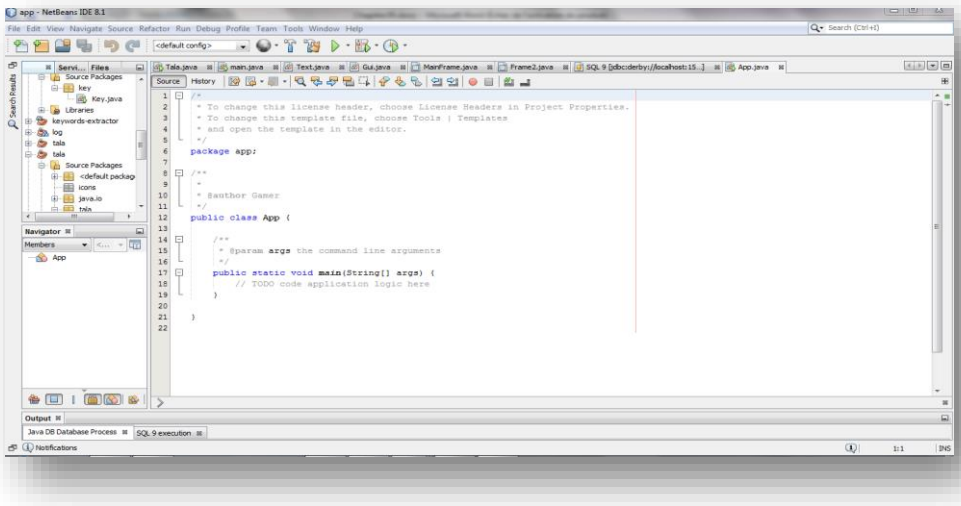


Figure IV-3: Environnement NetBeans.

IV.3.2. Bases de données utilisées

IV .3.2.1 MySql-Wamp

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 sous MySql. Des requêtes comme les créations de table de données, insertions, mises à jour, suppressions et modifications de structure de la base de données, ainsi que l'attribution et la révocation de droits et l'import/export. Ce système permet de sauvegarder commodément une base de données sous forme de fichier .sql et d'y transférer ses données, même sans connaître SQL.

Figure IV-4: MySql-phpMyAdmin.

IV.3.2.1. La base de données lexicale Arabic WordNet : Awn.Xml

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

```
<form value="وشك" wordid="&gt;awo$aka_1" type="root"
authorshipid="13685" />
<form value="وضح" wordid="&gt;awoDaHa_1" type="root"
authorshipid="13686" />
<form value="وضح" wordid="&gt;awoDaHa_2" type="root"
authorshipid="13687" />
<form value="وضح" wordid="&gt;awoDaHa_3" type="root"
authorshipid="13688" />
<form value="وصي" wordid="&gt;awoSaY_1" type="root"
authorshipid="13689" />
<form value="وصي" wordid="&gt;awoSaY_b_1" type="root"
authorshipid="13690" />
<form value="وصي" wordid="&gt;awoSaY_b__1" type="root"
authorshipid="13691" />
<form value="وصل" wordid="&gt;awoSala_1" type="root"
authorshipid="13692" />
<form value="وصل" wordid="&gt;awoSala_2" type="root"
authorshipid="13693" />
<form value="وصل" wordid="&gt;awoSala_3" type="root"
```

Figure IV-5: Extrait de Awn.Xml.

IV.4. Description de notre application

Dans cette partie nous allons présenter l'interface principale de notre application avec son fonctionnement.

IV.4.1. Architecture de fonctionnement

Nous avons résumé le fonctionnement de notre application dans le schéma de la figure ci-dessous (**Figure IV.6**):

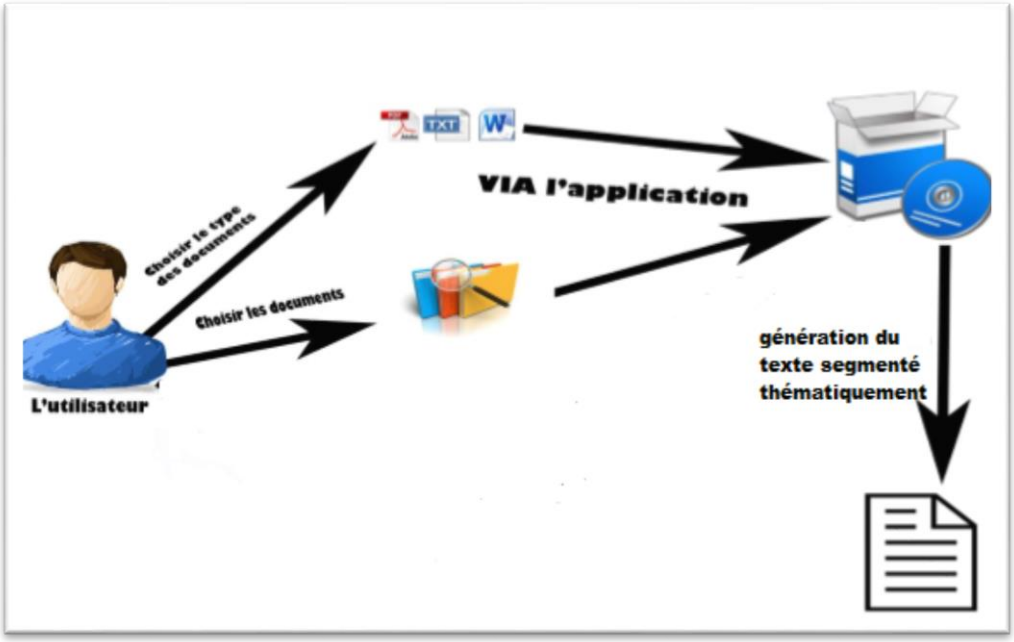


Figure IV-6: Architecture de fonctionnement du système topic segmentation.

IV.4.2. Les bibiothèques utilisées

- 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 :
  - MySql-connector-java-5.1.42-bin.jar
- Générer les fichers word,txt,pdf :
  - Pdfbox-app-2.0.jar
  - Poi-3.9.jar
  - Poi-ooxml-3.8.jar
- Design amilioré :
  - Weblaf-complete-1.29 .jar

- Générer le spalch-screen :

  - Jspachscreen.jar

- Gestion des graphes :

  - Gs-core-1.2.jar
  - Gs-ui-1.3.jar

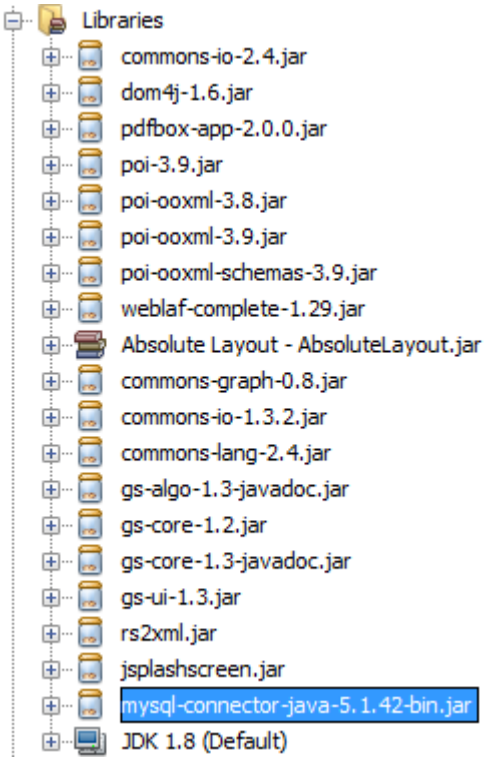


Figure IV-7: Jars utilisés.

IV.4.3. Interface d'utilisateur

Dans cette partie nous expliquons les différentes interfaces de notre application, comme les fonctionnalités des boutons et le rôle de chaque outil dès la sélection de textes jusqu'à la génération des sous graphes des segments thématiques, en passant bien sûre par les étapes intermédiaires: prétraitement, traitement syntaxique, linguistique à l'aide de la figure ci-dessous **(figure 4.8)**

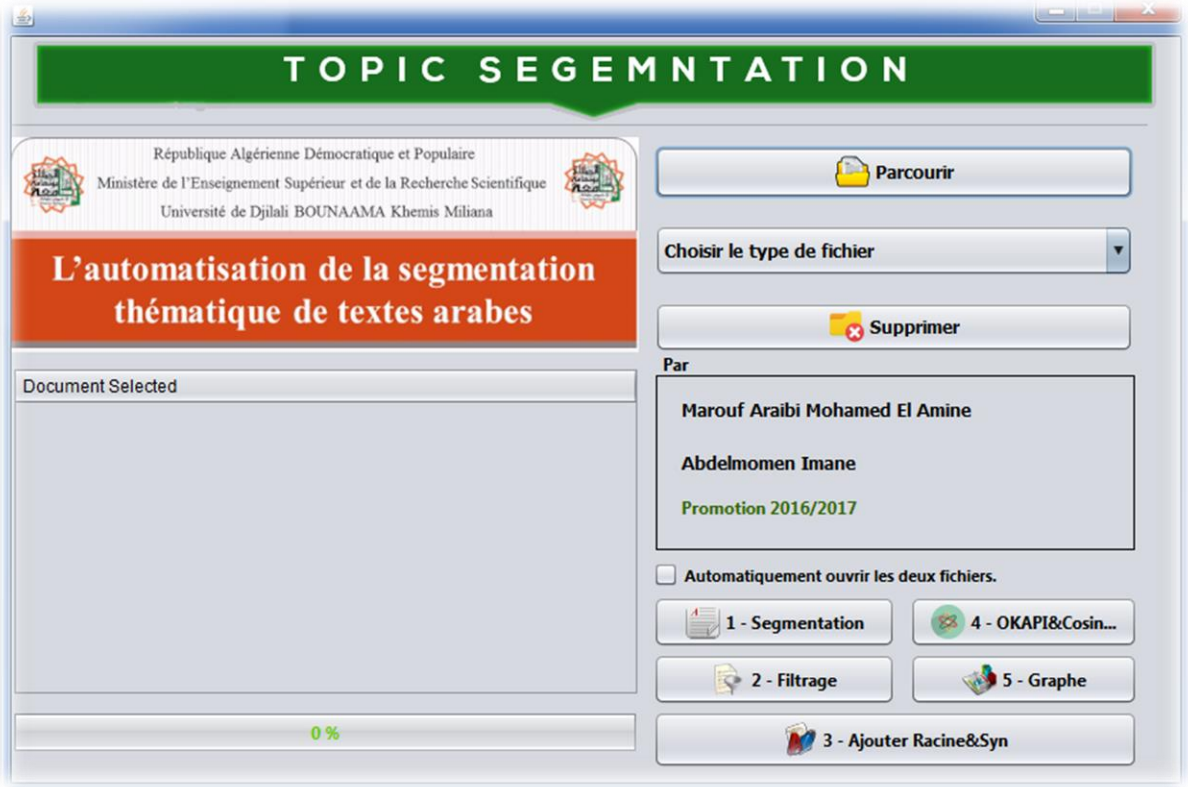


Figure IV-8 : Interface principale de l'application.

- **Liste des documents :** Un tableau qui affiche les documents sélectionnés par l'utilisateur.
- **Liste des types :** L'utilisateur doit sélectionner le type des documents avant de choisir ces derniers (TXT, WORD).
- **Bouton parcourir :** Ouvre une interface dialogue où l'utilisateur peut parcourir le répertoire de son système et choisir le document à segmenter.
- **Supprimer :** L'utilisateur peut supprimer ou annuler la sélection des documents en cas d'erreur de sélection.
- **Automatiquement ouvrir les deux fichiers :** le fichier segmenté et filtré seront ouvert automatiquement quand on coche cette fonctionnalité.
- **Segmentation :** ce bouton permet de segmenté le document sélectionné et le montrer dans un fichier TXT.
- **Filtrage :** ce bouton permet de filtrer le fichier segmenté et l'afficher dans un fichier TXT.
- **Ajouter racine-syn** : ce bouton permet de calculer et d'ajouter les racines des mots, leur Synsets ainsi que les racines de ces Synsets dans la table de la base de données.

- **Okapi** : ce bouton permet de calculer le score okapi pour chaque mot ainsi que le score **Cosinus** entre les segments et faire le clustering strict en sous-groupes thématiques.
- **Graphe** : permet de générer le graphe des segments thématique.

IV.4.3.1. Sélection des textes

Les textes que nous avons utilisés sont préparés à l'avance et ils sont enregistrés en format « txt ou word». Comme nous l'avons déjà mentionné dans la description de notre approche, ces textes sont collectés à partir d'un ensemble d'articles arabes discutant le domaine d'économie contemporaine. Le traitement commence par la sélection d'un fichier texte à traiter en actionnant le bouton « parcourir » d'après la sélection de type de document comme le montre la figure ci-dessous:

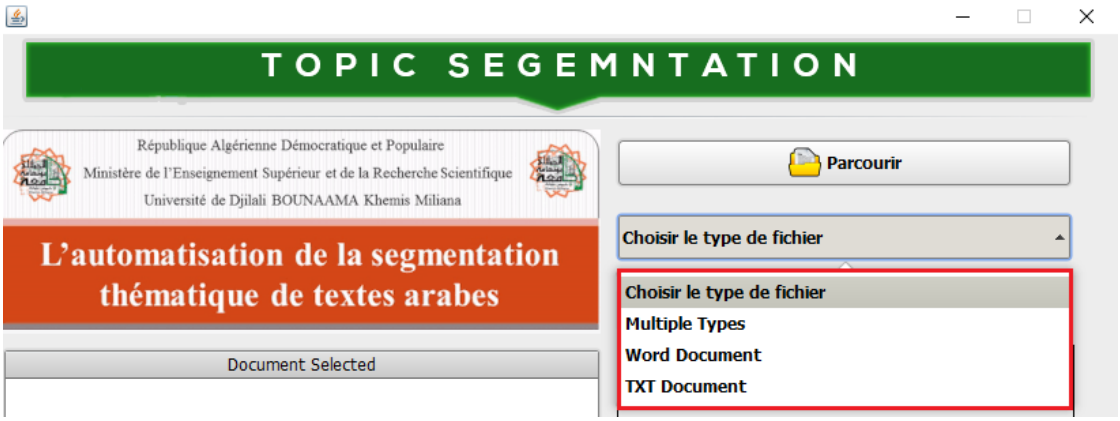


Figure IV-9: Sélection le type de document.

- La figure ci-dessous montre le répertoire pour choisir le document a traité:

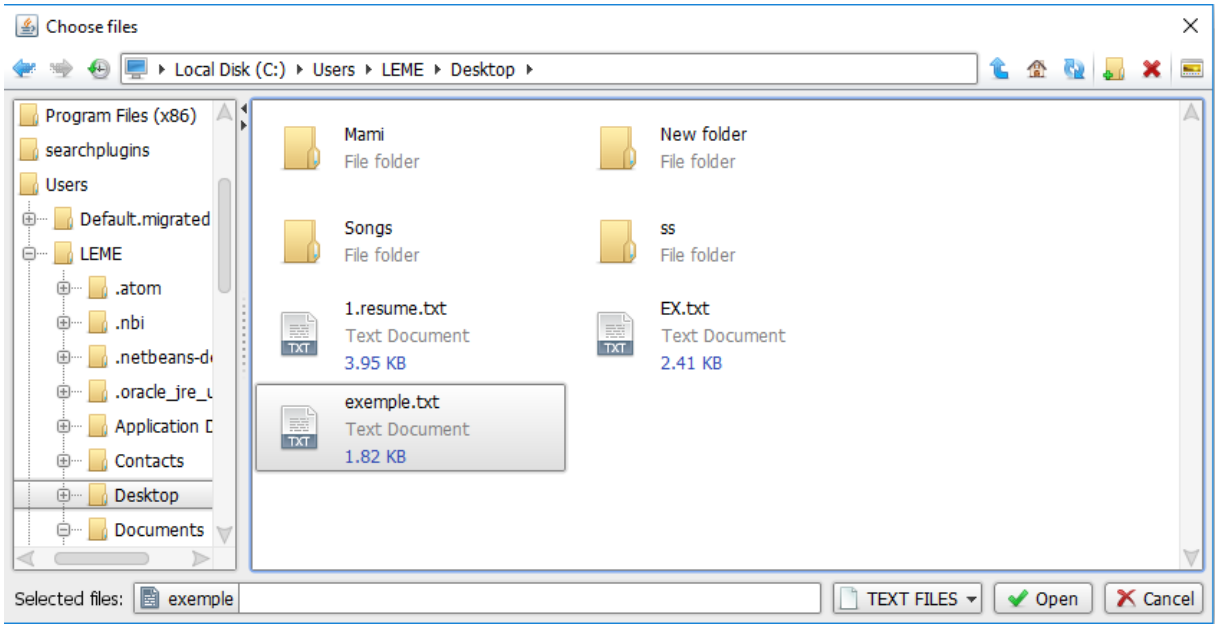


Figure IV-10: Sélection de document.

- Le tableau suivant montre le document sélectionné:

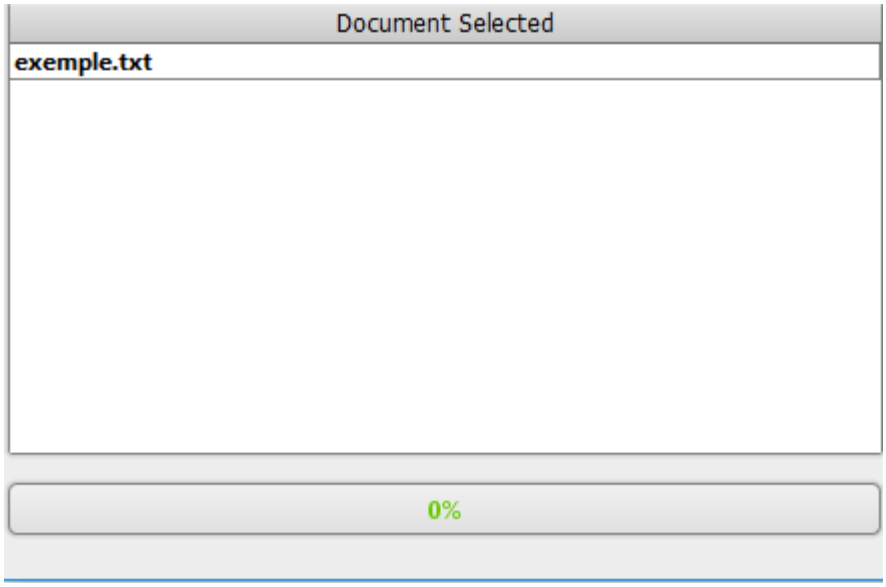


Figure IV-11: Document sélectionné.

IV.4.3.2. La segmentation formelle

Notre système offre la possibilité d'effectuer des traitements sur le texte comme la segmentation formelle à l'aide du bouton « segmentation » et l'affichage de texte segmenté en paragraphes comme le montre la figure suivante :

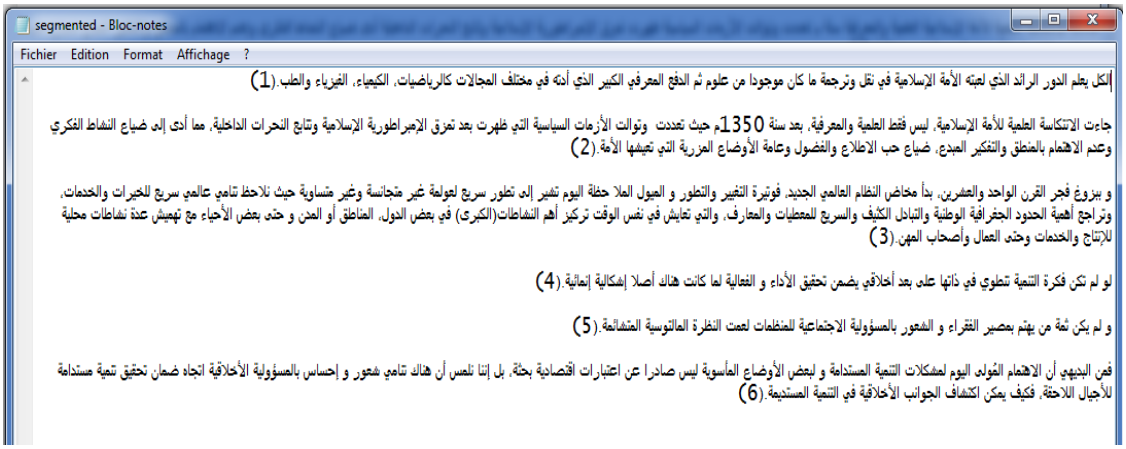


Figure IV-12: Fichier segmenté.

IV.4.3.3. Le filtrage automatique

Notre système offre la possibilité de filtrer le fichier déjà segmenté en éliminant les mots vides et les caractères spéciaux, ainsi que les chiffres et les lettres non-arabes, la figure suivante montre le résultat obtenu:

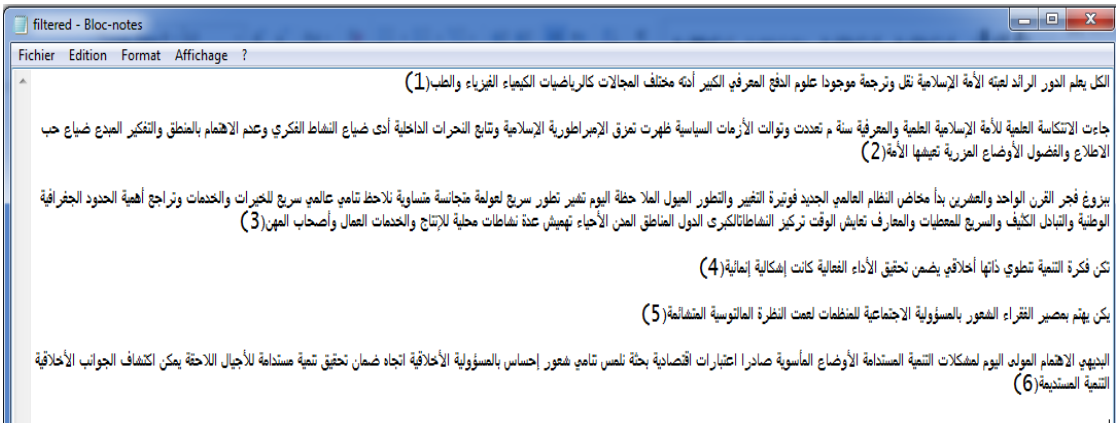


Figure IV-13: Fichier filtré.

IV.4.3.4. La lemmatisation et le calcul de la synonymie

Dans ces traitements nous avons fait l'extraction des racines des mots du fichier filtré ainsi que les racines de leurs synsets comme le montre la figure ci-dessous:

WORD	SEGMENT	POSITION	RACINE	SYNSETS
الكل	1	1	كل	ن نشأ قوم أسس ...
يعلم	1	2	علم	سل خطب خبر طل...
الدور	1	3	دور	س دبر روس سور ع...
الرائد	1	4	رود	شاء رضي رغب ب...
لعبته	1	5	لعب	ل وجه غش برا ش...
الأمة	1	6	لأم	سب واءم وفق حق...
الإسلامية	1	7	سلم	ب قبض ربح لقي ول...
نقل	1	8	نقل	حول عبر مرر حرك ه...
وترجمة	1	9	ترجم	ول نقل فسر سور ...
موجودا	1	10	وجد	م جود أسس جدد ...
علوم	1	11	علم	سل خطب خبر طل...
الدفع	1	12	دفع	

Figure IV-14: Calcul des racines et synsets.

Et Voice un exemple sur les racines trouvées pour les synonyme du mot « **الدور** » **:**

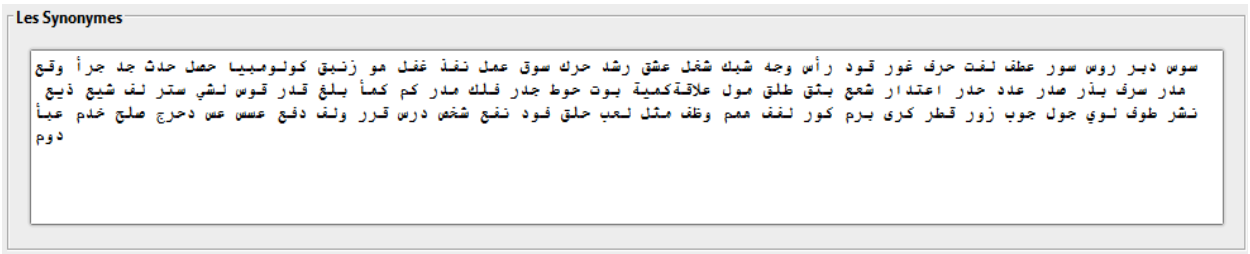


Figure IV-15: Liste des racines des synonymes du mot « الدور».

IV.4.3.5. La pondération

Dans cette partie nous montrons le résultat du calcul du score **okapi** pour chaque mot de document filtré ainsi pour leurs synsets en actionnant le bouton okapi qui montre le résultat final des traitements précédents :

WORD	SEGMENT	POSITION	RACINE	SYNSETS	TF	DF	DL	OKAPI
الكل	1	1	كل	ن نشأ قوم أسس ...	1	1	21	1.48489483900...
يعلم	1	2	علم	سل خطب خبر طل...	1	1	21	1.48489483900...
الدور	1	3	دور	رح صلح أخدم عبأ دوم	1	1	21	1.48489483900...
الرائد	1	4	رود	شاء رضي رغب ب...	1	2	21	0.67175618845...
لعبته	1	5	لعب	ل وجه عش برا ش...	1	1	21	1.48489483900...
الأمة	1	6	لأم	سب واءم وفق حق...	1	2	21	0.50640081899...
الإسلامية	1	7	سلم	ل قبض ربح لقي ول...	1	2	21	0.78683792592...
نقل	1	8	نقل	تحول عبر مرر حرك ه...	1	1	21	1.48489483900...
وترجمة	1	9	ترجم	ول نقل فسر سور ...	2	1	21	2.15053735304...
موجودا	1	10	وجد	م جود أسس جدد ...	1	3	21	0.0
علوم	1	11	علم	سل خطب خبر طل...	1	1	21	1.48489483900...
الدفع	1	12	دفع		1	1	21	1.48489483900

Figure IV-16: Résultat final des scores Okapi.

IV.4.3.6. Le groupement thématique

Dans cette partie nous sommes arrivés à notre dernier traitement qui concerne le calcul de la similarité entre les segments, et la présenté dans une matrice ainsi que la sélection de ces groupes et on finira par le groupement thématique pour ces derniers.

IV.4.3.6.1. Le calcul de similarité

Nous présentons dans cette partie le résultat de calcul de la similarité affiché dans une matrice d'adjacence et dont les lignes et les colonnes sont le nombre de segments obtenu d'après la segmentation formelle, comme il est indiqué dans la figure suivante :

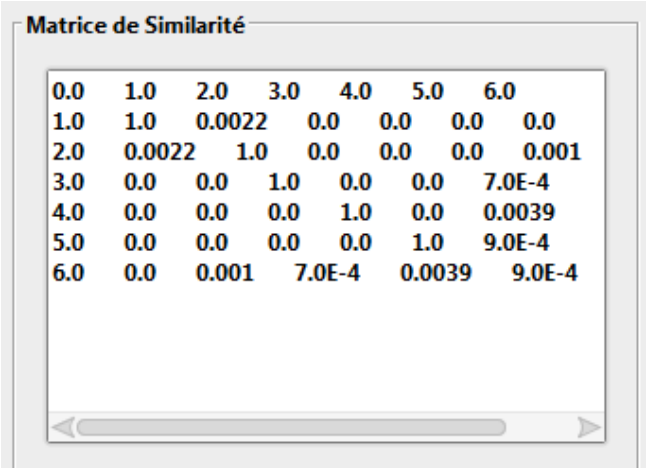


Figure IV-17: Matrice de similarité.

IV.4.3.6.2. Le clustering strict

A partir de la matrice de similarité, nous associons à chaque segment i le segment le plus lié. Ce dernier est le segment ayant la valeur de similarité la plus élevée dans la ligne du segment i. Ainsi un ensemble de paires de segments est formé.

Apres avoir formé les paires de segments**,** nous avons utilisés une technique de clustering strict pour former les sous-groupes thématiques, La figure ci-dessous montre le résultat d'exécution :

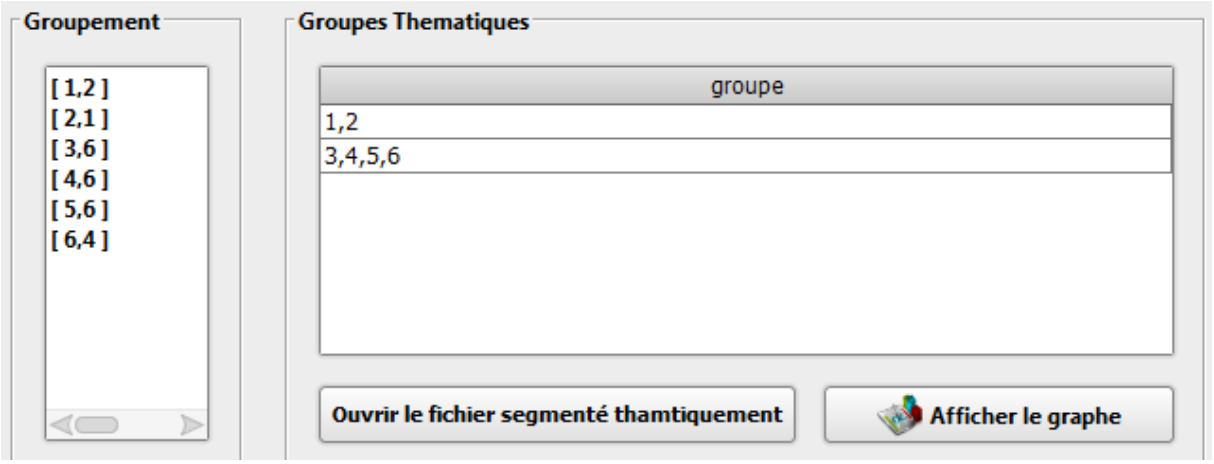


Figure IV-18: Exemple d'execution d'un clustering strict.

IV.4.3.6.3. La génération de graphe de similarité et les sous-graphes des groupes thématiques

A partir de la matrice de similarité nous avons la possibilité de générer un graphe de similarité, et à partir de la matrice de similarité et les groupes thématiques en peut aussi génère les sous-graphes des groupes thématiques, comme dans les figures ci-dessous :

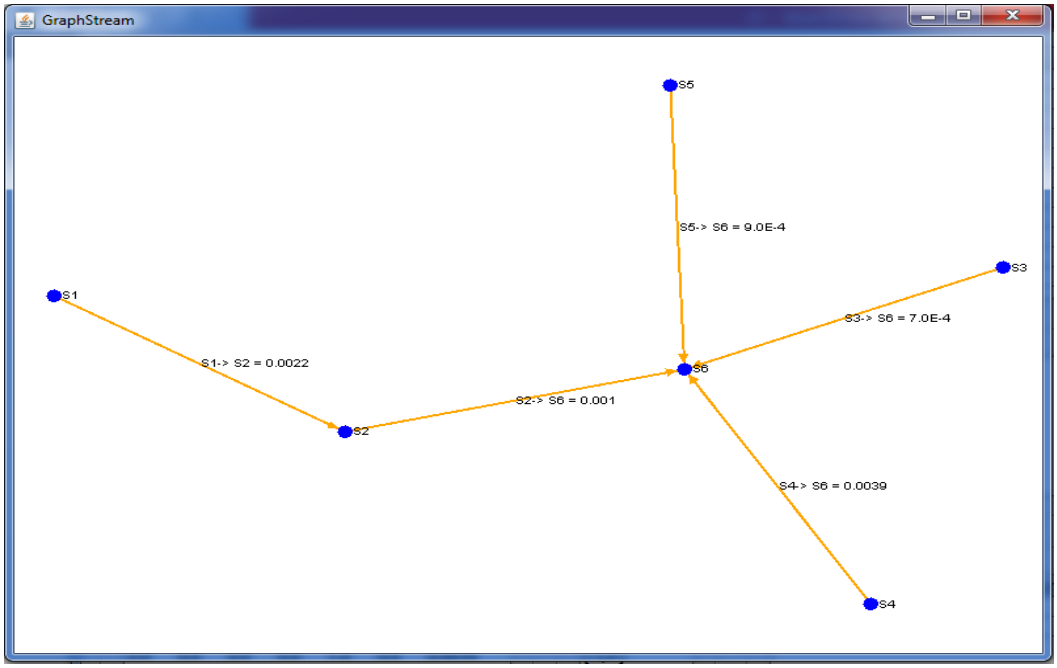


Figure IV-19: Graphe de similarité.

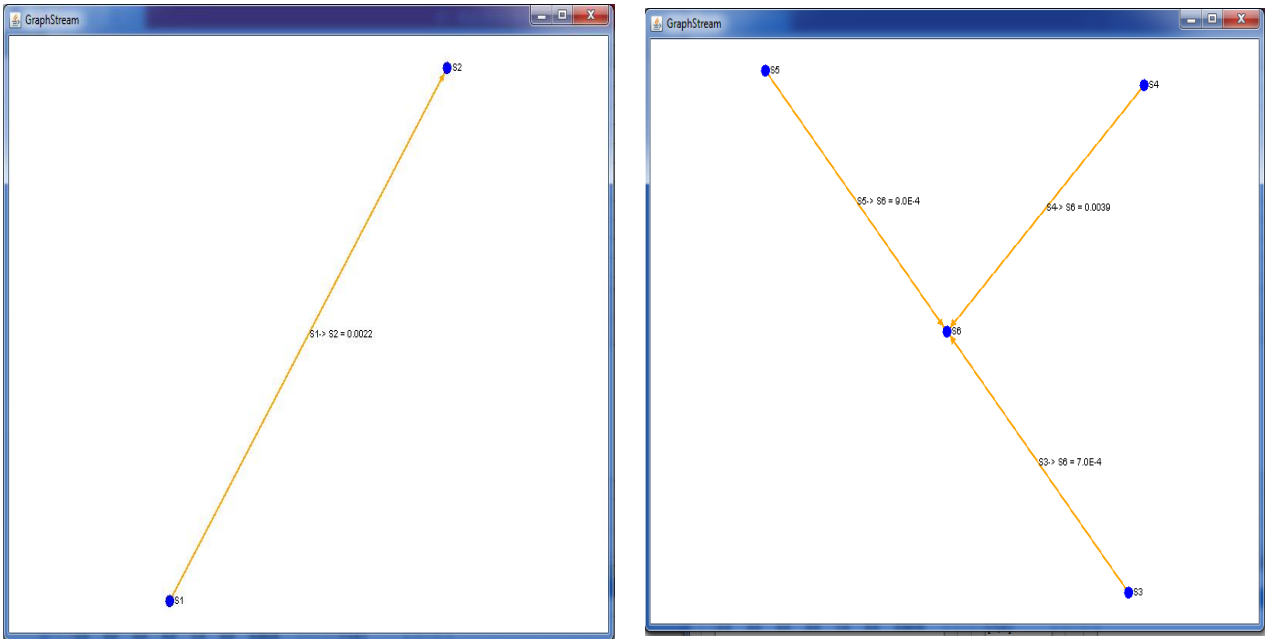


Figure IV-20 : Les sous-graphes des groupes thématiques.

IV.4.3.6.4. La génération du fichier segmenté thématiquement

A partir des groupe thématique déjà extrait, nous avons la possibilité d'affiché les groupes thématiques dans un fichier TXT comme dans la figure ci-dessous :

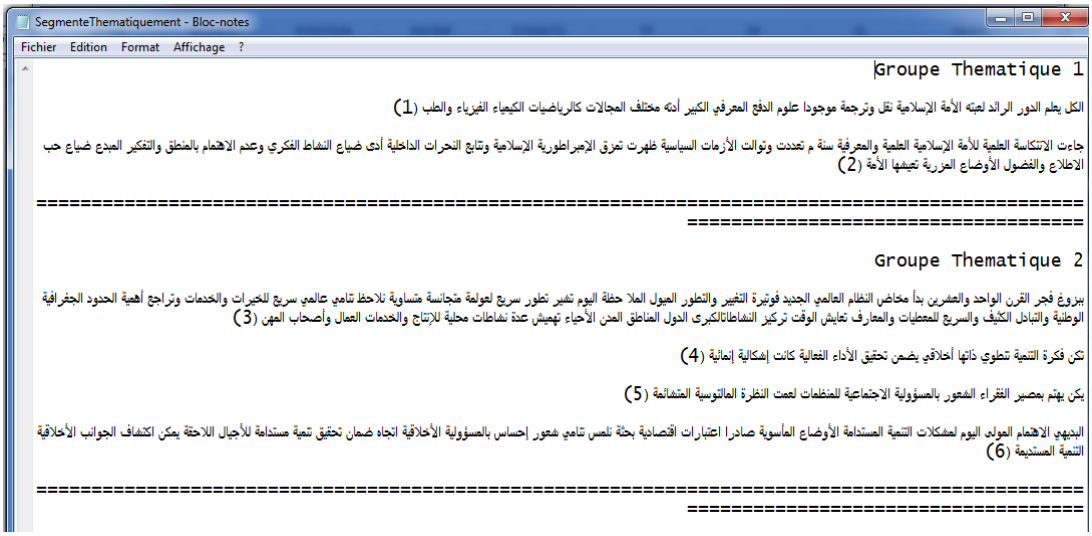


Figure IV-21: Fichier segmenté thématiquement.

IV.5. 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.

Afin d'évaluer la performance de notre système, nous avons construit un échantillon de 5 articles contenant des extraits hétérogènes.

Dans chaque article nous avons concaténés deux extraits provenant de deux articles discutant des sujets différents comme le montre la figure suivante:

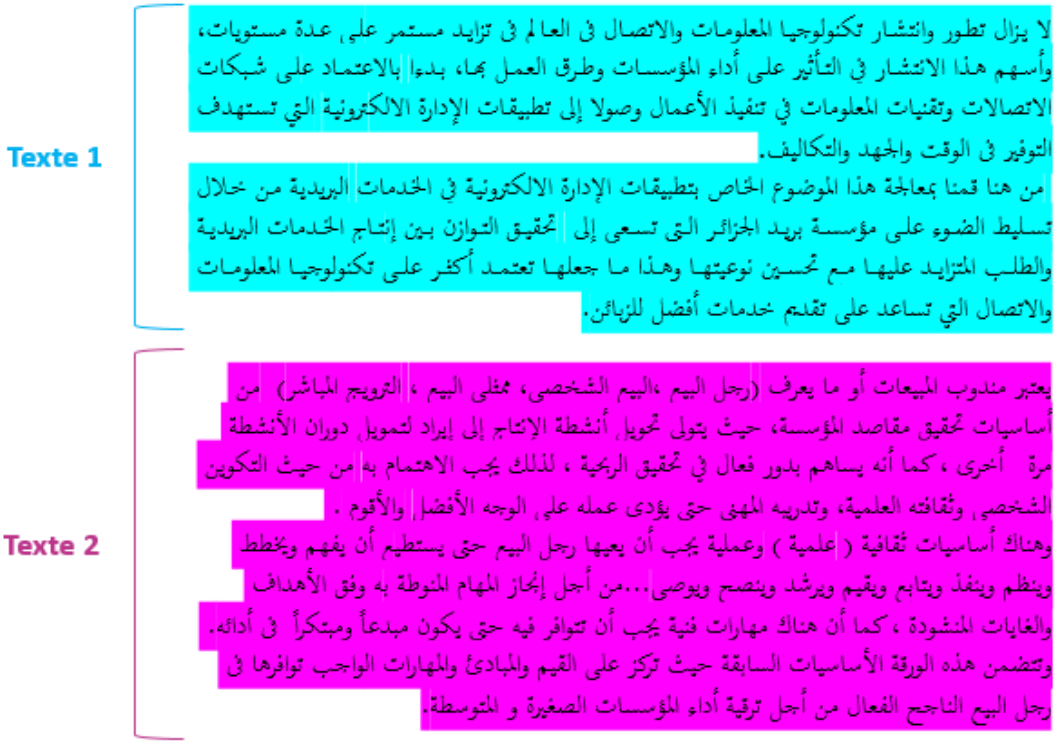


Figure IV-22: Concaténation des articles différents.

Nous comparons les segments thématiques produits par le système avec le groupement référence des segments dans les deux articles comme dans l'exemple suivant:

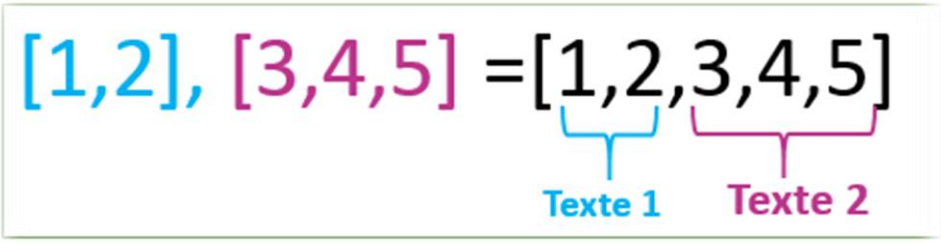


Figure IV-23: Comparaison entre les segments thématique et le groupement référence des segments.

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

$$\text{Rappel} = \frac{\text{Correctes}}{\text{Correctes} + \text{Oubliées}} \qquad \textbf{IV.1}$$

$$\text{Précision} = \frac{\text{Correctes}}{\text{Correctes} + \text{Incorrectes}} \qquad \textbf{IV.2}$$

- **Correctes**: Nombre de segments correctement groupées par le système
- **Incorrectes:** Nombre de segments incorrectement groupées par le système (les segments appartenant a article 1 groupés dans article 2).
- **Oubliées :** Nombre de segments groupés dans des groupes thématiques isolés.

Afin de donner une valeur d'évaluation globale De la segmentation du système, nous avons calculé la moyenne des valeurs Rappel et Précision pour les cinq articles, F-Mesure est calculée par l'équation suivante :

$$\text{F} - \text{mesure} = \frac{2.(\text{Rappel}.\text{Précision})}{\text{Rappel} + \text{Précision}} \qquad \textbf{IV.3}$$

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

- Les résultats sont montrés dans le tableau suivant :

	Texte1	Texte2	groupes thématique	rappel	précision
Test 1	[1,2]	[3,4,5]	[1,2],[3,4,5]	1	1
Test 2	[1,2]	[3,4,5]	[1,2,4],[3,5]	1	0,8
Test 3	[1,2,3,4,5,6]	[7,8,9]	[1,2,4],[3,5,6],[7,8,9]	0,66	1
Test 4	[1,2,3]	[4,5,6]	[1,2], [3,4,5,6]	1	0,85
Test 5	[1,2,3]	[4,5]	[1,2,3], [4,5]	1	1

Table IV-1: Résultats de tests pour 5 extraits hétérogènes.

Et le résultat des calculs de moyenne rappel, précision et F-mesure sont affiché dans le tableau ci-dessous :

Moyenne des rappels	0,93
Moyenne des précisions	0,96
F-mesure	0,94

Table IV-2: Résultat final d'évaluation.

IV.6. Conclusion

Dans ce chapitre, nous avons décrit brièvement le processus de réalisation de notre application en spécifiant l'environnement de développement, les bibliothèques utilisées et la démarche qu'on a adoptée pour réaliser notre projet. Ensuite pour évaluer notre système, on a choisi une comparaison semi-automatique Cela nous a permis de calculer les trois métriques (rappel, précision et mesure) afin d'obtenir des mesures plus précise de la qualité de notre résumé. Toutes les parties sont fonctionnelles, malgré qu'il y ait toujours quelques fonctionnalités à optimiser.

Conclusion Générale

Conclusion générale

L'automatisation de la segmentation thématique se base principalement sur les répétitions des termes ayant des structures morphologiques proches. Les nouveaux processus de segmentation visent à être plus intelligents en prenant en considération le sens des termes au lieu de considérer de simples chaines de caractères.

Dans ce travail notre objectif était de générer automatiquement une segmentation thématique d'un texte arabe qui regroupe les sous thèmes discutés dans des groupes distincts afin de faciliter l'accès aux informations et faciliter la tâche à d'autres traitements automatisés tel que l'indexation thématique ou le résumé automatique des textes arabes. 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 segments mais aussi de valoriser l'importance de ces relations et en se basant sur l'importance des termes qu'ils contiennent. Nous avons pu implémenter une technique de clustering strict en passant par un calcul de similarité entre segments textuels.

Par ailleurs, comme amélioration de ce travail, nous proposons l'étude d'autres relations telle que la collocation par exemple ainsi que l'utilisation d'une technique plus optimale pour le calcul des synonymes. 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.

Nous prévoyons compléter notre travail pour faire une évaluation totalement automatique en utilisant les formules Pk [60] ou WindowDiff [61].

Ce projet a été très bénéfique pour nous car il nous a permis de renforcer et enrichir nos connaissances théoriques dans le domaine de la conception, et de mettre en application nos connaissances acquises le long de nos études. Nous avons pu faire le lien entre tous les modules que nous avons étudié : L'intelligence artificielle, génie logiciel, reconnaissance des formes, Data Mining et la programmation orientée objet (POO), où il nous a donné l'occasion de maîtriser le langage de programmation Java, de découvrir de plus en plus le domaine vaste du TALN, et encore de nous familiariser avec la conduite des projets informatiques.

En perspective, on vise à améliorer la qualité des résultats obtenus, Nous espérons finalement par le modeste travail réalisé, apporter une contribution significative aux travaux liés au traitement de la langue arabe.

Références bibliographiques

Bibliographie

[1] : THÈSE de Souhir Gahbiche-Braham. Amélioration des systèmes de traduction par analyse linguistique et thématique .Application à la traduction depuis l'arabe. UNIVERSITÉ DE PARIS SUD, octobre 2013.

[2] :http://www.firdaous.com/wp-content/uploads/2014/03/monde_arabe.jpg. | Consulté le 18/06/2017.

[3] : Mémoire de Tahar DILEKH. Implémentation d'un outil d'indexation et de recherche des textes en arabe. Université Hadj Lakhdar – Batna, septembre 2011.

[4] : Mohand BEDDAR.vers un prototype de traduction automatique controlee français/arabe appliquee aux domaines a securite critique. Universite de franche-comte, avril 2013.

[5] : Mémoire de Fouad Soufiane Douzidia.Résumé automatique de texte arabe. Université de Montréal, Septembre 2004.

[6] : THÈSE de Souhir Gahbiche-Braham. Amélioration des systèmes de traduction par analyse linguistique et thématique. Application à la traduction depuis l'arabe. UNIVERSITÉ DE PARIS SUD, Septembre 2013.

[7] : Mémoire de Fouad Soufiane Douzidia.Résumé automatique de texte arabe. Université de Montréal, Septembre 2004.

[8] : T H È S E de Mohamed Hedi Maaloul. Approche hybride pour le résume automatique de Textes. Application à la langue arabe. Université Aix-Marseille. Université de Sfax, décembre 2012.

[9] : THÈSE de 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. Université de Nantes, Octobre 2008.

[10] : Mémoire de Tahar DILEKH. Implémentation d'un outil d'indexation et de recherche des textes en arabe. Université Hadj Lakhdar – Batna, septembre 2011.

[11] : Zoubeir Mouelhi. AraSeg : un segmenteur semi-automatique des textes arabes. S. Heiden, B.Pincemin JADT 2008 , Mar 2008, Rome, Italie. Presses Universitaires de Lyon, Actes JADT

2008 9es Journ´ees internationales d'Analyse statistique des Donn´ees Textuelles, pp.867-877, 2008.

[12] : Mémoire de Tahar DILEKH. Implémentation d'un outil d'indexation et de recherche des textes en arabe. Université Hadj Lakhdar – Batna, septembre 2011.

[13] : THÈSE de Baloul Sofiane Développement d'un système automatique de synthèse de la parole à partir du texte arabe standard voyellé. Université de Maine, France, le 27 mai 2003.

[14] : Mémoire de Filtrage sémantique de textes en arabe en vue d'un prototype de résumé automatique Motasem ALRAHABI, Ghassan MOURAD, Brahim DJIOUA. LaLICC (Langage, Logique, Informatique, Cognition et Communication) UMR 8139, Université Paris – Sorbonne, CNRS 96, Bd Raspail 75006 Paris – France, avril 2004.

[15] : De l'´évaluation en Traitement Automatique des Langues mémoire d'Habilitation à Diriger des Recherches École Doctorale d'Informatique Université Paris-Sud Patrick Paroubek Ingénieur de Recherche au LIMSI-CNRS (groupe ILES), avril 2013.

[16] : http://la-madrassa-du-mowa7id.blogspot.com/2011/11/lecon-1.html | Consulté le 18/06/2017.

[17] : http://www.musulmanspetitsetgrands.com/arabe/guide_enseignants_ch11.php | Consulté le 18/06/2017.

[18] : traitement automatique de la langue arabe, unité de recherche riadi faculté des sciences tunisie 2007.

[19] : http://al-dirassa.com/le%C3%A7on/la-phrase-nominale-et-la-phrase-verbale-en-arabe/ | Consulté le 18/06/2017.

[20]: A. Chalabi : Sakhr Web-based Arabic<>English MT engine, ACL/EACL 2001 Workshop on Arabic Language Processing, Toulouse July2001(http://www.elsnet.org/arabic2001/chalabi.pdf).

[21] : THÈSE de Mohamed Hedi Maaloul. Approche hybride pour le r_esum_e automatique de textes. Application _a la langue arabe.. Traitement du texte et du document. Universit_e de Provence - Aix-Marseille I, 2012. Fran_cais. <tel-00756111v1>.

[22]: Journal of Computational Information Systems4:3(2008) 1169-1180 Available at http://www.JofCI.org 1553-9105/ Copyright © 2008 Binary Information Press March, 2008 Automatic Keyword Extraction from Documents Using Conditional Random Fields.
[23]: http://www.larousse.fr/dictionnaires/francais/mot-cl%C3%A9_mots-cl%C3%A9s/52770 | Consulté le 18/06/2017.

[24] : http://www.zahhr.com/ | Consulté le 18/06/2017.

[25] : Mémoire de boubekeur yassmina.Identification automatique de mot clés dans les textes arabes, univercité de khmiss miliana 2015/2016.

[29] : Zoubeir Mouelhi. AraSeg : un segmenteur semi-automatique des textes arabes. S. Heiden, B.Pincemin JADT 2008 , Mar 2008, Rome, Italie. Presses Universitaires de Lyon, Actes JADT 2008 9es Journ´ees internationales d'Analyse statistique des Donn´ees Textuelles, pp.867-877, 2008.

[30] : Mémoire de Tahar DILEKH. Implémentation d'un outil d'indexation et de recherche des textes en arabe. Université Hadj Lakhdar – Batna, septembre 2011.

[31] : T h è s e de Alexandre LABADIE. Segmentation thématique de texte linéaire et non-supervisée : Détection active et passive des frontières thématiques en Français. Université Montpellier II, 2009.

[32] : Stéphanie Audrit - Détection automatique des changements de sujet au sein d'un texte

[33] : Evaluation de méthodes de segmentation thématique linéaire non supervisées après adaptation au français Laurianne Sitbon, Patrice Bellot Laboratoire d'Informatique d'Avignon - Université d'Avignon, avril 2004.

[34] : http://www.larousse.fr/dictionnaires/francais/segmentation/71825 | Consulté le 18/06/2017.

[35] : http://dictionnaire.reverso.net/francais-definition/segmentation | Consulté le 18/06/2017.

[36] : Leïla BACCOUR1, Lamia HADRICH BELGUITH1, Ghassan MOURAD2. Segmentation de textes arabes en phrases basée sur les signes de ponctuation et les mots connecteurs .Faculté des Sciences Economiques et de Gestion de Sfax. Mars 2003.

[37]: Lamia Hadrich Belguith, Nouha Chaâben. Analyse et désambiguïsation morphologiques de textes arabes non voyellés. Faculté des Sciences Économiques et de Gestion de Sfax

[38] : http://tadbur.com/ms/t6508.html | Consulté le 18/06/2017.

[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.

[40] : Analyse thématique du discours : segmentation, structuration, description et représentation Nicola Hernandez, Brigitte Grau LIMSI/CNRS, BP 13, F-91403 ORSAY (France).

[41] : T h è s e de Alexandre LABADIE. Segmentation thématique de texte linéaire et non-supervisée : Détection active et passive des frontières thématiques en Français. Université Montpellier II, 2009.

[42] : Alexandre Labadié, Jacques Chauché. Segmentation thématique par calcul de distance thématique. EGC'07: Extraction et Gestion des Connaissances, Namur, Belgique, pp.355-366, 2007. <lirmm-00161992>, le 12 Jul 2007.

[43] : Mémoire de CHOUCHAOUI Maiss et BRAHIMIA Yamna Affaf. Détection Automatique De La Cohésion Lexicale Entre Phrases Dans Les Textes Arabes, université de khemiss miliana 2015/2016.

[44] : Adrien Bougouin, État de l'art des méthodes d'extraction automatique de termes clés, TALN-RÉCITAL 2013, 17-21 Juin, Les Sables d'Olonne. Université de Nantes, France, Juin2013.

[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, le 24 novembre 2008.

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

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

[48]: 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.

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

[50] : 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

[51] : 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.

[52]: 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.

[53] : La segmentation thématique Application à la campagne DEFT'06,grenoble cadex 9-France Date : 21 Juin 2006.

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

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

[56] : 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.

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

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

[59] : 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.

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

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