MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITÉ ABDELHAMID IBN BADIS - MOSTAGANEM

Faculté des Sciences Exactes et d'Informatique

Département de Mathématiques et informatique

Filière : Informatique

RAPPORT DE MINI-PROJET

Option : Ingénierie des Systèmes d'Information

THÈME :

« Recherche d'entité arabe sur Wikipédia »

Etudiant(e) : « Bensaber kaouter »

Etudiant(e) : « Dahmeche Hadjer »

Soutenu le : 18/09/2022

Devant le jury composé de :

MR.Moussa M	Garde	Université de Mostaganem	Examinateur
MR.Hanni F	Garde	Université de Mostaganem	Président
Mme.Kennich A	Garde	Université de Mostaganem	Encadreur

Année Universitaire 2021-2022

Résumé

Les bases de données lexicales jouent un rôle important dans plusieurs domaines du traitement automatique des langues (TAL), comme l'extraction d'information, la reconnaissance d'entités nommées. La reconnaissance des entités nommées (REN) consiste à identifier des entités nommées (EN) dans des ressources textuelles et les classer en catégories prédéfinies.

Ce rapport présente les différentes étapes de la conception d'un système de recherche d'entités nommées arabe sur Wikipédia. Ce système s'appuie sur la reconnaissance des entités nommées arabe à l'aide d'un dictionnaire électronique relationnel multilingue de nom propres qui s'appelle prolexbase.

Mots-clés/ entité nommée, Prolexbase, Wikipédia, Dictionnaire électronique, Reconnaissance d'entités nommées, langue Arabe.

Abstract

Lexical databases play an important role in several areas of automatic language processing (NLP), such as information extraction, named entity recognition. Named entity recognition (REN) consists in identifying named entities (EN) in textual resources and classifying them into predefined categories.

This thesis presents the different stages of the design of an Arabic named entity search system on Wikipedia. This system is based on the recognition of Arabic named entities, which is done using a multilingual relational electronic dictionary of proper names called prolexbase.

We have implemented our system to increase the Arabic volume in a Wikipedia corpus.

Key Words

Named entity, Prolexbase, Wikipedia, electronic dictionary, Named entity recognition, Arabic.

Dédicace

Je dédie ce travail à mes très chers parents, pour leurs présences, soutien moral Permanent et leur amour infini tout au long de ma vie, ce travail est le fruit de leurs Sacrifices qu'ils ont consentis pour mon éducation et ma formation.

À ma sœur Sabria.

À Tous ceux qui ont participé de près ou de loin à la réalisation de ce travail.

Bensaber Kaouter

Dédicace

Je dédie ce modeste travail : A mes grands-parents qu'allah nous les gardes.

A mon très cher père qui m'a vraiment soutenu dans mon chemin scolaire et qui était toujours présent pour me pousser et m'encourager. Sans lui le n'aurait pu être ce que je suis-je lui exprime mes sincères gratitude

A ma source d'Amour très chères mère qui n'a jamais cesser de m'éclairer par ses conseils forts précieux et parfois exigeants.

A mon marie qui m'a soutenu et encouragé durant ces années d'études

A mes frères: (Khair-Eddine, Yousef, Houssine)

A mes sœurs :(Hasina, Amina)

À tous mes amis qui m'ont toujours encouragé, surtout mon binôme avec lequel j'ai partagé des moments inoubliables, et à qui je souhaite plus de succès et de réussite dans leur vie.

Hadjer Dahmeche

Remerciements

Nous remercions ALLAH de nous avoir donné la force et la capacité qui nous ont Menées à ce niveau. Nous tenons à exprimer notre profonde gratitude et reconnaissance à notre

Encadrante, **Dr. Kenniche Ahlem**. Nous vous remercions d'avoir bien assuré la direction et l'encadrement de notre travail malgré vos nombreuses occupations, nous vous remercions pour vos qualités humaines, votre précieuse attention, implication et vos précieux conseils et orientations qui ont mené à l'aboutissement de ce travail.

Nous remercions aussi les membres de jury pour avoir bien voulu donner de leur temps pour lire et juger notre projet.

Merci aussi à tous nos amis, nos collègues, et à tous ceux qui nous ont aidé de près ou de loin. Nous leur exprimons notre profonde sympathie et leur souhaitons beaucoup de bien.

Nous n'oublierons pas non plus de remercier toutes les personnes que nous avons pu côtoyer pendant ces cinq ans à la faculté des sciences exactes et d'informatique pour leur soutien moral et amical.

Liste des figures

Liste des abréviations

EN : Entité Nommée

ENA : Entité Nommée Arabe

REN : Reconnaissance des Entités Nommées

RS : Relation Sémantique

TAL : Traitement Automatique des Langues

EI : D'Extraction d'Information

Table des matières

Introduction Générale

Avec l'évolution rapide des technologies de l'information et de la communication, le besoin s'est rapidement fait sentir de s'appuyer sur les techniques linguistiques. Parallèlement, la linguistique a pu profiter de la puissance des ordinateurs pour acquérir une nouvelle dimension et ouvrir la voie à de nouveaux domaines de recherche. Parmi ces domaines RNE arabe. Cette dernière, reste encore une piste de recherche intéressante.

Nous utiliserons comme corpus le volume arabe de l'encyclopédie Wikipédia qui est une ressource libre. La Wikipédia possède un volume arabe très riche en termes d'EN arabes (ENA) qui possèdent une fréquence d'apparition très importante et une présence dans divers contextes.

Après avoir effectué une étude linguistique profonde sur la reconnaissance des entités nommées arabe, le dictionnaire électronique multilingue relationnel et étudié le fonctionnement de la Wikipédia arabe et les travaux existant. Notre objectif principal consiste à effectuer une reconnaissance des entités nommées arabe à partir d'un corpus extrait de la Wikipédia arabe. Et construire un système de recherche d'entité nommée arabe sur Wikipédia arabe à base de dictionnaire « prolexbase » dans le but d'augmenter le volume arabe dans cette derniére qui est faible en comparent avec le volume français et le volume anglais.

Dans le premier chapitre, nous allons commencer par la présentation de la notion de la reconnaissance d'entité nommée arabe. Puis, nous allons décrire les catégorisations. Ensuite, nous allons aborder trois approches d'Extraction d'Information (EI) principales (symbolique, statistique et hybride) ainsi que les travaux associés. Enfin, nous clôturons ce chapitre par la définition du dictionnaire électronique multilingue relationnel spécifique aux noms propres et leur ontologie.

Dans le deuxième chapitre, nous allons rappeler la définition de la Wikipédia. Après, nous allons présenter en détail la ressource libre Wikipédia arabe. Puis, pour chaque approche nous allons présenter les systèmes de reconnaissance des entités

nommées arabe. Déplus, nous allons définir le Wikimédia. Finalement, nous allons citer les travaux qui exploitent la Wikipédia arabe c'est dernière temps.

Dans le troisième chapitre, nous allons citer les étapes de traitement automatique avec unitex et prolexbase. Puis nous allons étudierons la segmentation de clitique et en phrase pour la langue arabe.

Dans le quatrième chapitre, nous allons discuter sur les étapes que nous avons suivies pour l'implémentation de notre système.

Chapitre 1

Reconnaissance d’entité nommée arabe

Introduction

La reconnaissance des entités nommées arabe consiste à identifier des entités nommées arabe dans des ressources textuelles et à les classer en catégories prédéfinies. Le système de la reconnaissance des entités nommées (REN) est basé sur des règles linguistiques qui exploitent l'étiquetage syntaxique, des déclencheurs et des dictionnaires de noms propres. Parmi les dictionnaires utilisés par le système REN on trouve prolexbase. Il s'agit d'une base de données lexicale qui contient toutes les informations syntaxiques, morphologiques et sémantiques concernant les noms propres. Le prolexbase comporte à ce jour dix langues. Dans notre thèse on doit faire l'étude sur les entités nommées arabe.[1]

1. La reconnaissance des entités nommées

1.1.Définition

La reconnaissance d'entités nommées est une sous-tâche de l'activité d'extraction d'information dans des corpus documentaires. Elle consiste à rechercher des objets textuels catégorisables dans des classes telles que noms de personnes, noms d'organisations ou d'entreprises, noms de lieux, quantités, distances, valeurs, dates.[2]

1.2.Les catégories des entités nommées

1.2.1. Les noms propres

Le nom propre est une sous-catégorie de nom. Ainsi, Un nom propre appartient donc à un référent déterminé (une personne, un animal ou une chose), que ce référent, réel ou imaginaire, existe naturellement (un élément géographique par exemple) ou qu'il soit artificiellement créé par l'homme (une œuvre d'art, une œuvre littéraire).[3]

1.2.2. Les noms de lieux

Les noms de lieux désignent les villes, les pays, les villages, les montagnes et les fleuves. Par exemple, le mot Algérie الجزائرaljaZaa'ir désigne le pays ou la capitale.[3]

1.2.3. Les noms d’organisations

Les noms d’organisations sont assez nombreux et sont difficilement quantifiables puisque leur apparition et leur disparition dépendent de la situation dans le monde.[3]

1.2.4. Les entités numériques

Les entités numériques sont divisées en deux grandes formes d’ENA : les expressions de temps et les nombres. Les expressions de temps incluent les dates, la période et toute autre expression exprimant le temps et les nombres incluent principalement les systèmes de mesures (poids, distance, volume, vitesse), les pourcentages, ainsi que les devises.[3]

2. Reconnaissance d’entité nommée Arabe

2.1. Définition

La détection des entités nommées (EN) en langue arabe est un prétraitement potentiellement utile pour de nombreuses applications du traitement des langues, en particulier pour la traduction automatique. Cette tâche représente un sérieux défi, compte tenu des spécificités de l'arabe.[4]

2.2. Phénomènes linguistiques rencontrés

L’étude des différentes formes d’entités nommées nous a montré l’existence de plusieurs phénomènes linguistiques qui peuvent causer des problèmes dans le processus de reconnaissance des EN arabes. Parmi ces phénomènes, nous citons les suivants :[3]

2.3. L’agglutination

La langue arabe est une langue fortement agglutinante du fait que les clitiques se collent aux substantifs, verbes, adjectifs auxquels ils se rapportent. De ce fait, nous trouvons des particules qui se collent aux radicaux en empêchant leurs détections. Par exemple, la détermination peut s’exprimer par :

- L’agglutination de l’article AL avant le mot :

Le livre : الكتاب

- L’agglutination d’un clitique à la fin du mot :

Son livre : كتابه

Ce qui rend son analyse automatique une tâche pénible à réaliser. D’ailleurs, Ces phénomènes d’agglutination augmentent, considérablement, le taux d’ambiguïté en introduisant d’autres supplémentaires au niveau de la segmentation des mots.[5]

2.4. La détermination :

Certains constituants sont toujours déterminés (cas des adjectifs). D’autres peuvent être déterminés ou non sans qu’il y ait des règles qui régissent ces différentes situations Prenons le cas des toponymes. Cette situation peut être aussi rencontrée pour les anthroponymes. Et le traitement de ce problème nécessite l’ajout dans le dictionnaire d’un trait supplémentaire sur la possibilité ou non d’une détermination. [3]

2.5. Longueur des noms propres

Contrairement aux langues latines, les noms propres arabes ne commencent pas par des lettres majuscules (la majuscule n’existe pas dans la langue arabe). Les noms propres sont donc difficiles à identifier. Aussi, leur longueur n’est pas connue d’avance et peut dépendre des traditions de la région dans laquelle est née la personne. Ainsi, il n’est pas possible de mettre dans un dictionnaire tous les noms propres avec toutes leurs variantes d’écriture. [3]

2.6. La syntaxe

Dans la langue arabe, la grammaire de construction des EN est riche et très variée. En effet, la longueur des EN (ou le nombre de constituants) ne peut pas être connue à l’avance ; elle est variable. Pour compléter le sens et le rendre non ambigu, on a tendance à ajouter un adjectif supplémentaire. Ainsi, une EN peut contenir une partie essentielle et une autre facultative qui vient juste pour l’enrichir ou le rendre non ambiguë. Notons aussi, qu’un même type de constituant peut se trouver à des positions différentes. Ce changement de position s’accompagne d’un changement de la structure de l’EN notamment au niveau des conjonctions et de la forme de détermination de certains constituants. C’est principalement le cas de l’adjectif qui

ne suit pas toujours le nom auquel il se rapporte. A tous ces phénomènes, nous pouvons ajouter aussi les différentes formes d'écritures d'un même mot notamment les mots d'origine étrangère.

En outre, nous constatons d'une part, les ambiguïtés des déclencheurs. Par exemple, «الجزائر, aljaZaa'ir» qui peut désigner l'Algérie ou l'Alger.[3]

2.7. Approches de reconnaissance des EN

Les approches de reconnaissance d'EN ont connu un fort développement depuis la fin des années 80 sous l'impulsion des conférences MUC. Trois grandes approches sont généralement suivies : l'approche linguistique ou symbolique, l'approche statistique ou à base d'apprentissage et l'approche hybride. Ce qui distingue les approches citées, n'est pas la nature des informations prises en compte, mais plutôt leur acquisition et leur manipulation. Dans ce qui suit, nous donnons un aperçu de ces approches tout en s'appuyant sur quelques exemples représentatifs des systèmes de reconnaissance d'EN.[3]

2.7.1. Approche linguistique

L'approche linguistique repose sur l'intuition humaine, avec la construction manuelle des modèles d'analyse, le plus souvent sous la forme de règles contextuelles. C'est pourquoi, cette approche est appelée aussi approche à base de règles. Ces dernières, qui expriment l'information à reconnaitre, prennent la forme de patrons d'extraction permettant la descriptiond'enchaînements possibles de syntagmes nominaux. Dans ce cadre, (Mesfar, 2008) a défini l'architecture d'un système de reconnaissance des EN arabes. Ce système est fondé sur une approche

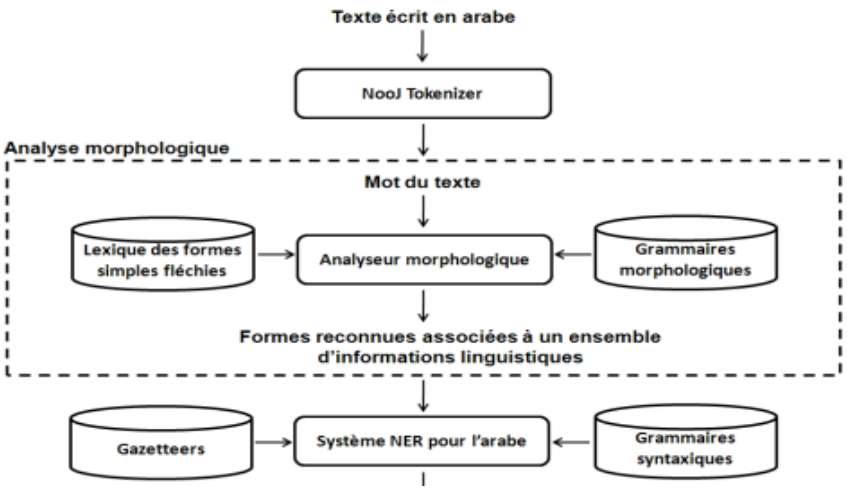

FIGURE 1: ARCHITECTURE GENERALE D'UN SYSTEME DE RECONNAISSANCE DES EN

linguistique utilisant les grammaires locales implémentées dans la plateforme linguistique NooJ.[3].L'architecture de ce système (Mesfar, 2008) est illustrée **figure 1**[3]

Comme indiqué dans la **Figure 1**, le système de (Mesfar, 2008) passe par une analyse morphologique suivie d'une analyse syntaxique. L'analyse morphologique, qui se sert des dictionnaires et des grammaires morphologiques, permet l'identification des caractéristiques de chaque mot du texte existant dans les dictionnaires construits et la résolution de certains problèmes liés à la langue arabe comme l'agglutination.[3]

2.7.2. Approche statistique

L'approche statistique a pour principe de base la mise au point automatique de modèles d'analyse à partir de volumes importants de données (ou corpus). Ces méthodes sont dites statistiques ou à base d'apprentissage car elles apprennent, à partir de corpus annotés, des modèles d'analyse de textes. Le système de reconnaissance d'EN arabes ANERsys développé par (Benajiba, 2009). Ce système possède deux versions différentes basées sur une approche d'entropie maximale. Ces systèmes à base d'apprentissage se sont considérablement multipliés ces dernières années, eu égard à leur facilité de mise en œuvre.

Nous commençons tout d'abord avec le système de reconnaissance d'EN arabes ANERsys développé par (Benajiba, 2009). Ce système possède deux versions différentes basées sur une approche d'entropie maximale. La première version passe par une seule étape comportant deux phases. Ces deux phases sont illustrées

dans **Figure 2**[3]

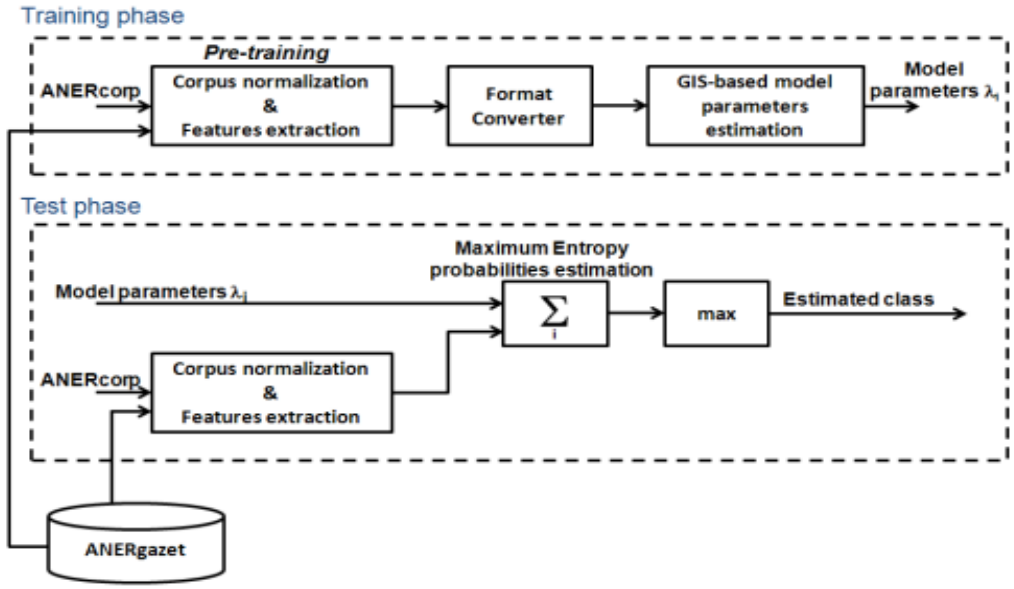

FIGURE 2 : ARCHITECTURE DE LA PREMIERE VERSION D'ANERSYS

2.7.3. APPROCHE HYBRIDE

L'approche hybride consiste à combiner l'approche linguistique et l'approche statistique afin de profiter des avantages des deux. En extraction d'informations. Parmi les projets admettant cette approche hybride, nous citons le système d'extraction des EN arabes proposé par (El Kateb-Gara, 2004).

Parmi les projets admettant cette approche hybride, nous citons le système d'extraction des EN arabes proposé par (El Kateb-Gara, 2004). Sa méthodologie mixte passe nécessairement par trois étapes linguistiques et une étape statistique :

- L'analyse lexicale du texte ou éventuellement d'un corpus.
- La reconnaissance des séquences pertinentes à travers des grammaires dédiées qui traduisent les différentes règles de reconnaissance.
- L'étiquetage des séquences isolées afin d'avoir un texte annoté. Les étapes linguistiques sont illustrées dans la **Figure 3** [3]

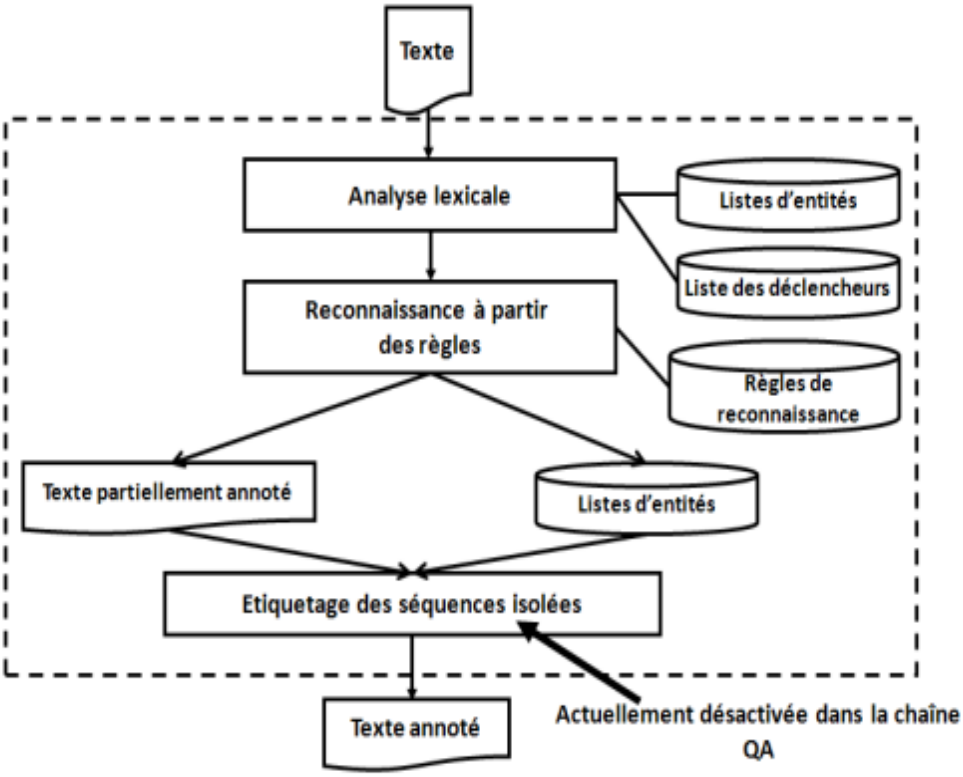

FIGURE 3 : DEMARCHE PROPOSEE

3. Prolexbase

3.1. Définition

Le Projet Prolex a été lancé en 1990 avec pour simple objectif de produire une base de données de noms d'habitants français et de toponymes avec des informations linguistiques. Aujourd'hui, Prolexbase est un dictionnaire électronique multilingue relationnel spécifique aux noms propres, disponible librement sur le site Web CNRTL [6]. La modélisation du domaine des noms propres définie dans le projet Prolex se base sur deux concepts fondamentaux : un nom propre conceptuel, le pivot, le référent de différents points de vue et la projection de ce pivot dans une langue donnée, appelé «prolexème» ; chaque prolexème est relié à un seul pivot inter langue.[7]

3.1.1. L'ontologie de Prolexbase

L'ontologie de Prolexbase vise à modéliser la classe linguistique des noms propres ; elle est structurée en deux parties : la partie supérieure commune à toutes les langues traitées, qui elle-même est constituée de deux niveaux : le niveau

conceptuel (les pivots numériques) et le niveau méta conceptuel (types et super types)[7]. La partie inferieure propre à une langue donnée est divisée en deux niveaux dépendants de la langue : le niveau des instances (les noms propres tels qu'ils apparaissent dans un texte écrit dans une langue donnée) et le niveau linguistique (les prolexèmes).[3] **La figure 4** [7] représente l'architecture générale de prolexbase.

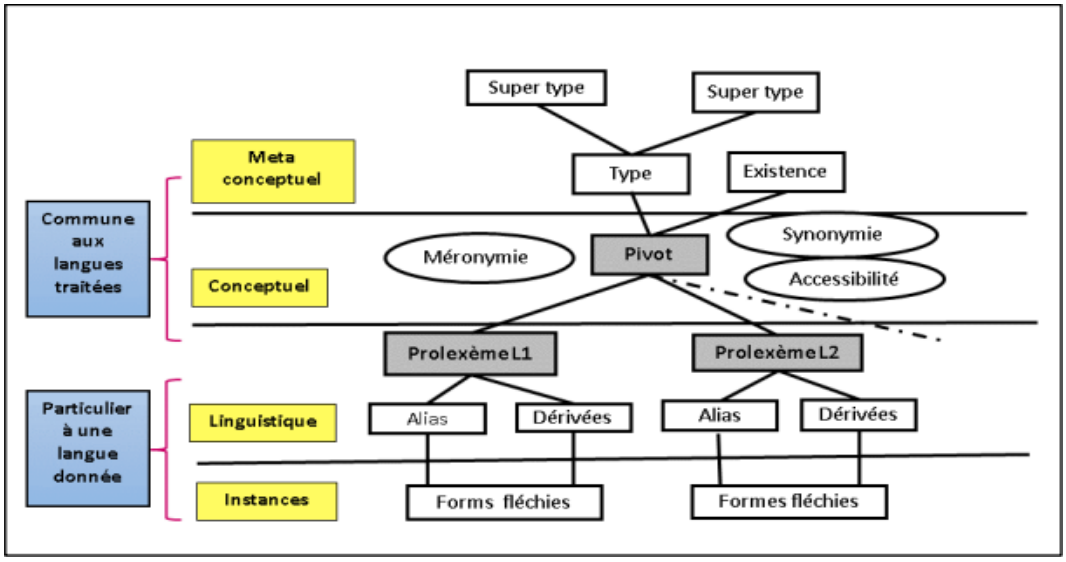

FIGURE 4 : L'ARCHITECTURE GENERALE DE L'ONTOLOGIE DES NOMS

Conclusion

La reconnaissance des entités nommées arabe (RENA) reste encore une piste de recherche intéressante vue la difficulté de la syntaxe et de la grammaire de la langue arabe.

Ce chapitre, présente le concept de base de la reconnaissance des entités nommées arabe, les approches fondamentales de la reconnaissance des entités nommées et les différents phénomènes linguistiques liés au repérage des ENA et l'ontologie du dictionnaire électronique multilingue relationnel.

Le chapitre suivant sera consacré à la description de la Wikipédia arabe, et la Wikimédia et les systèmes de REN pour chaque approche citer dans le premier chapitre ainsi que les travaux réalisé sur Wikipédia pour l'extraction des entités nommées arabe.

Chapitre 2
Wikipédia arabe

Introduction

Wikipédia est un projet d'encyclopédie collective en ligne, universelle, multilingue et fonctionnant sur le principe du wiki. La richesse de la Wikipédia en termes d'EN et son aspect multilingue ont joué un rôle important pour la proposition des systèmes de reconnaissance des entités nommées.[8]

La Wikipédia est devenue une ressource riche et un outil important utilisé dans le système de recherche d'information, et dans ce cadre, notre recherche se base fondamentalement sur l'utilisation de la Wikipédia arabe. Nous allons présenter de nombreuses études qui ont été faites sur l'extraction des entités nommées.[7]

1. Wikipédia

1.1. Définition

Le terme Wikipédia est étymologiquement issu de la fusion de deux termes : wiki-, issu de l'hawaïen wiki, qui signifie rapide, se référant au fait que l'encyclopédie ait toujours vocation à s'améliorer rapidement et à être constamment active par son mode de fonctionnement, et -pédia, lui-même dérivé du mot grec paideia, instruction et éducation.[9]

2. Wikipédia arabe

2.1. Définition Wikipédia arabe

Aujourd'hui, l'encyclopédie Wikipédia se classe comme le cinquième site le plus visité au monde selon le classement Alexa[10].la Wikipédia possède un volume arabe très riche en termes d'EN arabes (ENA) mais il reste encore les articles arabe sur Wikipédia moins que les articles en français et anglais.

La Wikipédia arabe contient un système hiérarchique de catégorisation, dans lequel chaque article appartient selon son sujet à au moins une catégorie, et les catégories sont elles-mêmes classées dans d'autres catégories, thématiquement plus larges. [11]

2.2. La structure générale d'une page Wikipédia

Les articles de la Wikipédia sont constitués dans les différentes versions linguistiques avec généralement une structure quasi identique ; ils se composent de

textes écrits en langage naturel, d'images et aussi d'autres informations structurées et de plusieurs types de liens. [7]

Ci-dessous, nous détaillons certains de ces composants que nous considérons importants.

2.2.1. Les Info boxes

Elles représentent les caractéristiques d'une entité donnée, correspondent aux tableaux reprenant des informations factuelles et structurée. Le contenu de ces Info boxes est une base pour l'alimentation de la base de données DBpedia, cependant, leur présence est limitée ; dans le cas des articles biographiques, moins d'un article sur trois propose ainsi une Info box ; les biographies font pourtant partie d'un des types d'articles les plus fréquents sur la Wikipédia [12]. Les Info boxes ou les boîtes d'information affichent des informations pertinentes pour le sujet de l'article en utilisant la fonctionnalité du logiciel modèle considérant le type d'entrée ; ces informations peuvent être des clés pour les recherches d'informations.[7]

2.2.2. Les catégories

Elles indexent chaque page de la Wikipédia où un ensemble de catégories mères visibles et cliquables par l'utilisateur est placé en bas de chaque page.[7]

2.2.3. L'historique

Il désigne un lien nommé « Historique ». via ce lien on peut accéder à la page de l'historique conservant l'ensemble des modifications qui ont été effectuées à la page cible depuis sa création. La page de l'historique permet de connaître la date, l'auteur et la teneur exacte de chaque modification ; elle contient des outils externes et statiques relatifs à la page cible : Auteurs et statistiques, Recherche de l'auteur d'un passage de l'article, Statistiques de consultation, Contributeurs suivant cette page et Modifications par utilisateur.[7]

2.2.4. La discussion

Il existe un lien appelé « Discussion ». Qui conduit vers la page de discussion où se trouvent les différents points de vue des contributeurs et les résultats du système d'évaluation fourni par le projet Wikipédia sur le contenu de la page cible.[7]

2.2.5. Page liées

C’est un lien vers une page d’outil via lequel on peut connaître la liste des pages liées à la page cible ; cette page contient un outil externe pour le nombre de pages liées, les inclusions, les liens internes et les redirections contenus dans la page cible.[7]

2.2.6. Informations sur la page

C’est un lien vers une page contenant des informations de base sur la page cible comme le titre, la taille, le nombre de contributeurs, le nombre de redirections vers cette page et d’autres informations.[7]

2.2.7. Les liens inter langues

Ce sont des liens vers les articles correspondants dans les autres langues ; ces liens sont situés dans un cadre à gauche de la page. Ainsi, le lecteur ou le contributeur peut trouver l’article équivalent dans les autres langues.[7]

2.2.8. Les liens inter wiki

Ils sont appelés aussi liens inter-projet car ce sont des liens entre les différents projets de la fondation Wikimédia ; ce sont des liens intégrés dans le texte comme les liens internes ordinaires, à utiliser principalement dans les discussions, en dehors donc des articles.[13]

2.2.9. Liens externes

Ce sont des hyperliens qui mènent vers d’autres sites web que la Wikipédia. Dans les articles de la Wikipédia, on peut en trouver à deux endroits différents. Tout d'abord, dans la liste des sources permettant de vérifier ce qui est écrit dans l'article. Ce type de lien externe, aussi appelé source ou référence, est généralement regroupé dans une section intitulée Références ou bien Notes et références. Un deuxième endroit possible pour ces liens est une section tout simplement appelée Liens externes en fin d’article.[14]

2.2.10. Liens internes

Ce sont des liens internes à la Wikipédia ou wiki liens pointant vers d’autres articles de la Wikipédia ; ils se mettent dans le corps de l'article. Leur utilisation peut parfois pécher dans leur pertinence.[15]. Les liens internes connexes à un article sont regroupés en fin d’article dans une sous-rubrique Articles connexes de

la rubrique Voir aussi, un lien interne s'affiche par défaut en bleu et quand il pointe vers un article qui n'existe pas, il s'affiche en rouge.

2.2.11. Référence

Elles se trouvent à la fin d'un article Wikipédia et elles sont des sources qui sont insérées dans le texte d'un article en les précédant par «↑» pour les distinguer des autres types.[16] Pour terminer cette section, nous illustrons en images des exemples clarifiant certains composants qui sont mentionnés plus haut.

Les **figures 5** et **6**[24] montrent respectivement la forme de la page Wikipédia arabe et les références.

FIGURE 5 : UNE PARTIE DE LA PAGE EL AMIR ABD EL KADER (VERSION WIKIPEDIA ARABE) COMPRENANT INFO BOX, DISCUSSION, HISTORIQUE ET LIEN INTERNE

FIGURE 6 : LES REFERENCES DANS EL AMIRE ABD EL KADER DE LA VERSION WIKIPEDIA ARABE

3. Wikimédia

3.1. Définition

Wikipédia est un mouvement mondial. Elle a pour finalité d'encourager la croissance, le développement et la distribution de savoirs et contenus libre et multilingues. Pour cela, elle met à disposition du public, l'intégralité de ses projets, fondés sur des Wiki. Ainsi fait-elle fonctionner quelques-uns des plus importants projets de l'édition collaborative, en particulier Wikipédia.[17]Wikipédia est l'un des dix sites les plus populaires dans le monde. Son contenu ainsi que celui de tous les autres projets Wikimédia sont créés, améliorés et mis en ligne par des bénévoles.[18]

3.2. Les chapitres de Wikimédia

Les chapitres Wikimédia incitent les personnes du monde entier à rassembler et développer du contenu éducatif sous une licence libre ou dans le domaine public et de le disséminer effectivement et mondialement. Ces associations locales indépendantes permettent d'agir selon les priorités et spécificités d'une région géographique donnée.[17] La **figure 7** [17] présente chapitre en 2017 :

associations existantes en bleu foncé, approuvées mais pas encore fondées en turquoise foncé, planifiés en vert et chapitres en discussion en bleu clair.

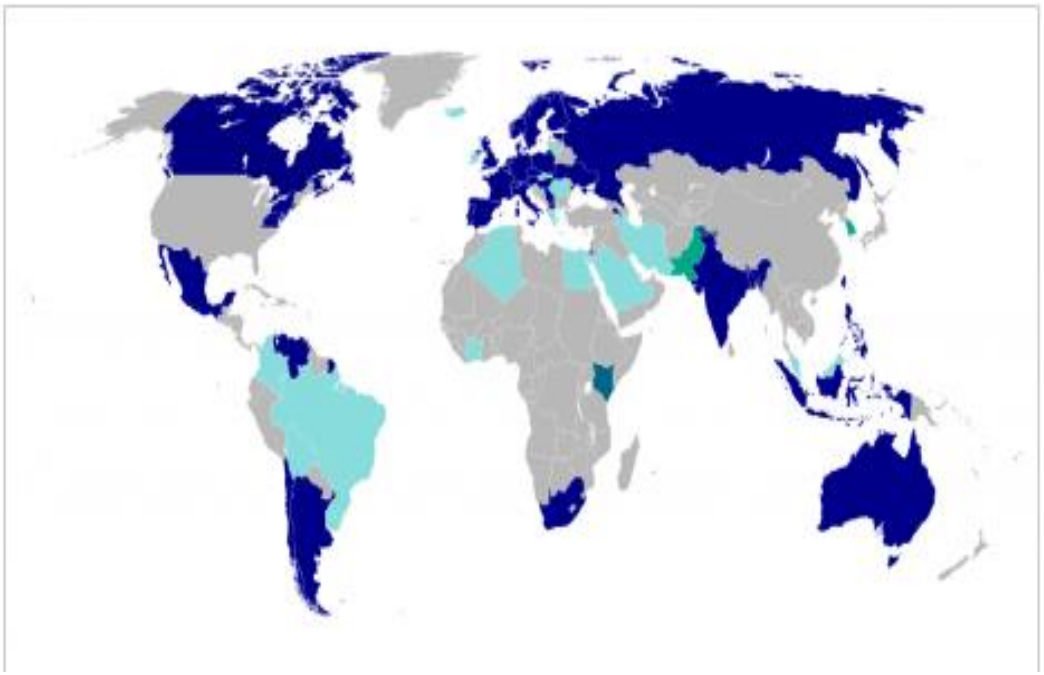

FIGURE 7 : CHAPITRE EN 2017

4. L'accès au contenu de l'encyclopédie Wikipédia arabe

La fondation Wikimédia fournit plusieurs outils de recherche et de traitement de données constituant les pages Wikipédia. Les principaux outils qui permettent l'accès au contenu de l'encyclopédie sont [7]:

4.1. Les dumps

La fondation Wikimédia publie des sauvegardes de la base de données qui peuvent être téléchargées.[19] Et utilisées pour consulter l'encyclopédie Wikipédia hors ligne après avoir installé localement le logiciel Media Wiki.[20] Les dumps permettent aussi de créer un site miroir qui conçoit une copie exacte d'un autre site web dans l'objectif de fournir plusieurs copies de la même information.[21]

En d'autres termes, les dumps sont les copies brutes de l'état de la mémoire informatique de tous les projets Wikimédia ; ils contiennent les publications, les historiques, les métadonnées, les liens inter wiki et les liens externes ; ce sont des fichiers de grande taille au format XML ou SQL ; il y a néanmoins des problèmes

associés à l'utilisation de cette solution d'accès au contenu de l'encyclopédie Wikipédia puisqu'elle est très gourmande en mémoire vive et n'est pas adaptée aux débutants.[22]

4.2. DBPEDIA

C'est un projet universitaire et communautaire d'exploration et d'extraction automatiques de données dérivées depuis l'encyclopédie Wikipédia. Son principe est de proposer une version structurée, sous forme de données normalisées au format du web sémantique, des contenus encyclopédiques de chaque fiche de la Wikipédia.[23]

5. Les Systèmes de REN existants

5.1. Système de Fehri

En se basant sur les transducteurs, dans [Fehri, 2012], l'auteur a proposé un module de REN pour la langue arabe dont le domaine choisi est le sport. Le module de REN proposé se compose de deux étapes dont la première est dédiée à identifier un ensemble de dictionnaires et des patrons syntaxiques à partir du corpus d'étude.[1]

5.2. Système d'Aboaoga et Aziz

Dans [Aboaoga et Aziz, 2013], les auteurs ont développé un système reconnaissant les EN ayant seulement la catégorie nom de personne dans des textes arabes. Le système proposé couvre trois domaines qui sont : le sport, la politique et l'économie et repose sur trois étapes dont la première étape est dédiée à effectuer un prétraitement.[1]

5.3. Système de Hkiri et al

L'ANERcorp est utilisé également pour tester un système de REN pour la langue arabe élaboré par [Hkiri et al. 2016]. Le système reconnaît trois catégories d'ENA qui sont nom de personne, nom de lieu et nom d'organisation. Cependant, le nombre de catégories reste insuffisant vis-à-vis le nombre de corpus exploités qui sont le ANERcorp, le corpus News Commentary et le corpus United Nations.[1]

6. Travaux exploitant la Wikipédia

La ressource Wikipédia est très utilisée pour construire des ressources linguistiques et les enrichir également. C'est dans ce cadre que s'inscrit le travail de [Sellami et al. 2012] consistant à construire un lexique bilingue à partir de la Wikipédia en profitant de son aspect multilingue. Et Parmi les domaines qui ont choisi la ressource libre Wikipédia pour élaborer des systèmes puissants, nous citons celui de la traduction automatique. Dans ce contexte, [Sellamiatal, 2013] ont proposé un système de traduction automatique statistique à partir de corpus comparables. Dans ce système, les auteurs ont choisi la paire de langues arabo-française. Le travail effectué consiste à exploiter les liens inter-langues qui relient les articles en arabe à ceux en français. Ce type de lien facilite l'extraction entre les termes (simples ou composés) arabes et leurs traductions en français et vice versa.

La richesse de la Wikipédia en termes d'EN et son aspect multilingue ont joué un rôle important pour la proposition des systèmes de REN. Parmi ces travaux, nous citons le travail de [Biltawi et al., 2016] qui se concentre sur la création des dictionnaires en exploitant cette ressource libre. Les dictionnaires créés sont classés en trois principales catégories : nom de personne, nom de lieu et nom d'organisation. Cette création de différents types de dictionnaires est une initiation à un processus de REN.[1]

Conclusion

Dans ce chapitre, Nous avons définie Wikipédia arabe .Puis, nous avons exposé certains composants de la structure d'une page Wikipédia arabe, en particulier les liens internes, les liens externes, le lien nommé et les liens inter langues. Ensuite, nous avons exploré des systèmes de la reconnaissance d'entités nommées. La présentation de ces systèmes nous a montré les résultats obtenus par chaque approche utilisée selon la langue traitée et la nature de corpus. Enfin, Nous avons conclu ce chapitre par Travaux exploitant la Wikipédia.

Chapitre 3

Conception de notre approche et traitements automatiques avec unitex

Introduction

Dans le présent chapitre, nous allons présenter les différentes étapes de traitement automatique avec Unitex et Prolexbase. Notre tâche étant d'entamé l'approche et les méthodes utilisées pour mettre en œuvre la détection et l'extraction d'entités nommées arabes. Cela veut dire parcourir le texte brut qui représente le corpus à traiter et appliquer toutes les règles pour récupérer le fichier nécessaire au traitement Pour l'utilisé dans Prolexbase.

1. Les Traitements automatiques avec Unitex

1.1Les étapes de traitement avec Unitex

Le corpus utilisé dans notre travail de détection des entités nommées et des Relations est seulement une partie des articles de Wikipédia arabe. Ce corpus est connu sur le nom d'article Wikipédia. **La figure 8** montre les étapes de traitement avec Unitex.

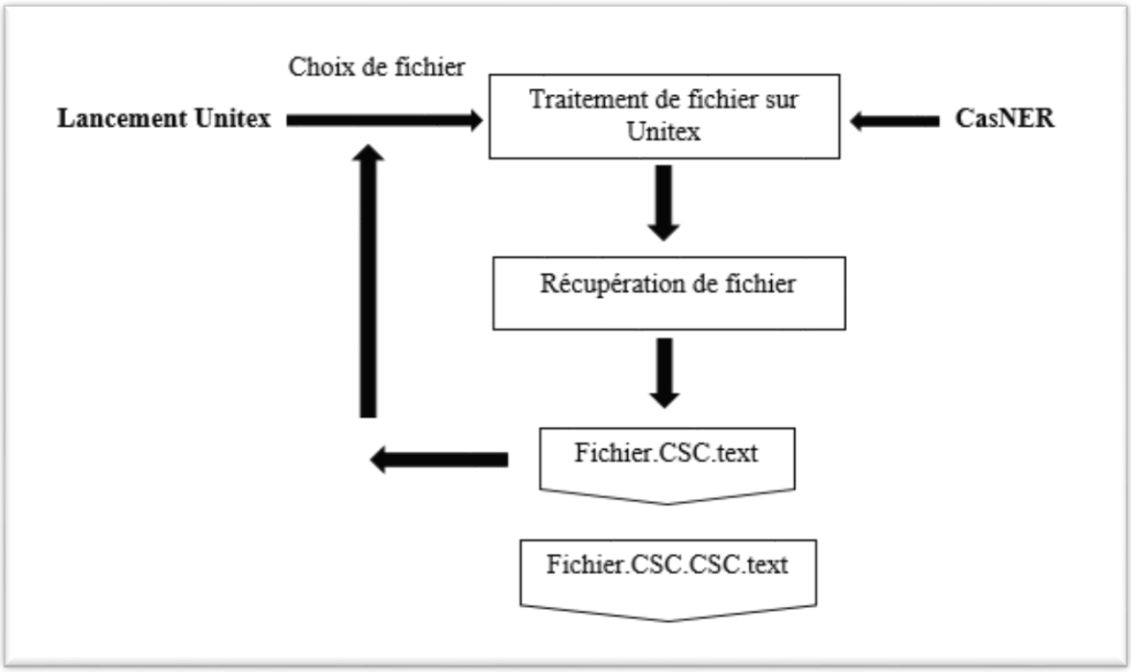

FIGURE 8 : LES ETAPES DE TRAITEMENT AVEC UNITEX

Première étape : L'ouverture de logiciel et le choix de texte quand veut traiter et après le traitement effectué par Unitex on récupéré le fichier.

CSC.text. Ce fichier contient les entités nommées arabe reconnues par le système CasANER.

CasANER : Le système CasANER est une cascade de transducteurs comportant cinq modules principaux. Ces modules sont organisés selon un ordre précis fixé selon plusieurs tests. Chaque module décrit une catégorie principale faisant partie de la typologie d'ENA. Dans la **figure 9** nous décrivons l'architecture de système CasANER et les modules qui le composent**.[1]**

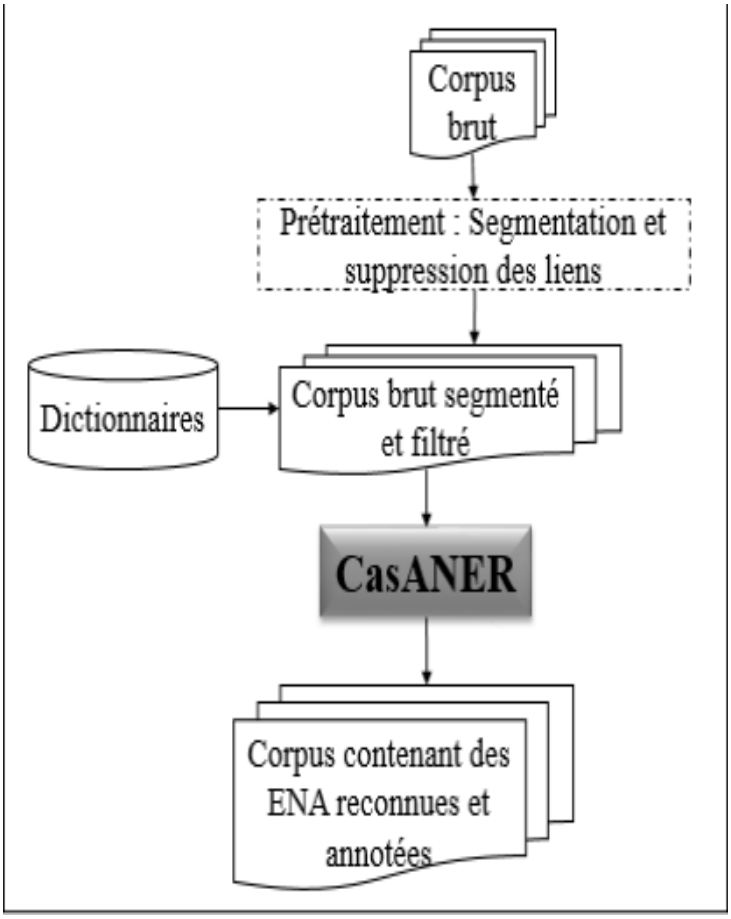

FIGURE 9 : ENTREES /SORTIE DU SYSTEME CASANER

Deuxième étape : En refait la première étape mais cette fois-ci avec le fichier. CSC.text et après le traitement en récupérer le fichier.CSC.CSC.text. Ce fichier contient les entités nommées arabe mais plus précis que le premier fichier. **.[1]**

1.2Les étapes de Traitement avec Prolexbase

Prolexbase est un dictionnaire électronique multilingue relationnel spécifique aux noms propres. La modélisation du domaine des noms propres définie dans le projet Prolex se base sur deux concepts fondamentaux : un nom propre conceptuel, le pivot**.[1]**. La **figure 10** montre les étapes de traitement avec Polexbase**..[1]**

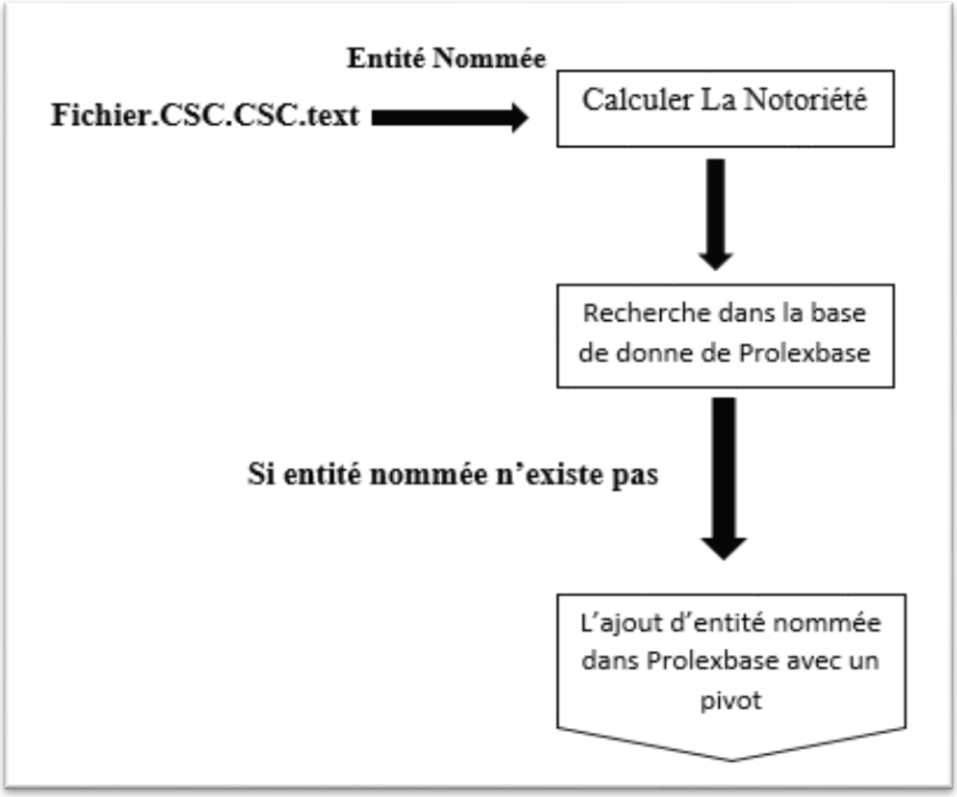

FIGURE 10 : LES ETAPES DE TRAITEMENT AVEC PROLEXBASE

Calculer la notoriété : Dans la première étape de traitement avec Prolexbase nous avons calculé la notoriété pour chaque entité nommée et nous avons suivi la méthode de « Mouna-elshter » qui est bien défini dans ca mémoire. **[7] .**

La recherche dans Prolexbase : Après avoir calculer la notoriété nous avons effectué une recherche dans prolexbase de l'entité nommée arabe.

L'ajout dans Prolexbase : si l'entité n'a pas été trouvé dans Prolexbase. Nous rajoutons l'entité avec un pivot.

3. Les étapes de traitement des entités arabes

3.1. La segmentation des clitique

Le phénomène de l'agglutination dans la morphologie arabe augmente le degré d'ambiguïté lexical et accroit la complexité de l'analyse d'un mot en ses composants. Ces derniers composants des affixes et des clitique arabe qui peuvent s'agglutiner en plusieurs niveaux. **La figure 11** montre les composants d'un mot arabe. **La figure 12** donne un exemple de l'analyse d'un mot arabe en affixe et en clitique.**[25]**

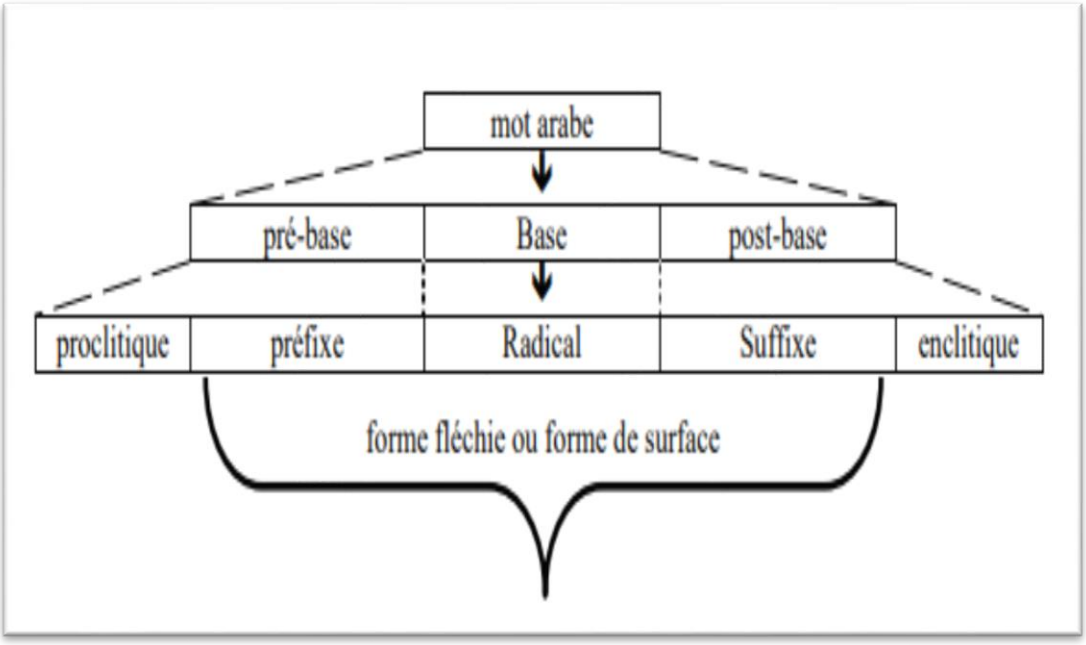

FIGURE 11 : MORPHOLOGIE D'UN MOT ARABE

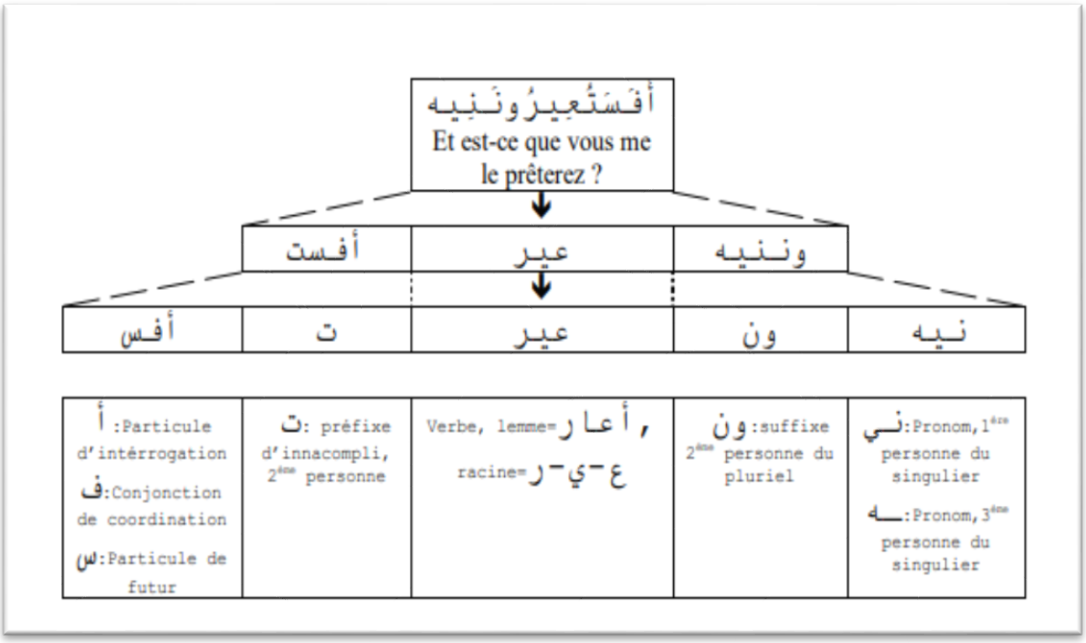

FIGURE 12 : EXEMPLE D'AGGLUTINATION/ ANALYSE DES CLITIQUES ET DES AFFIXES DANS UN MOT ARABE

3.2. La segmentation en phrase

La segmentation d'un texte en arabe en phrase révèle une grande importance dans le TAL arabe, car tout traitement qui repose sur la structure syntaxique nécessaire une segmentation du texte en petite structure syntaxique que nous l'appelons ici phrase **[25]**. Dans notre tokenizer de phrase de la **figure 13** on se base sur les signes de ponctuation pour effectuer par la segmentation.**[25]**

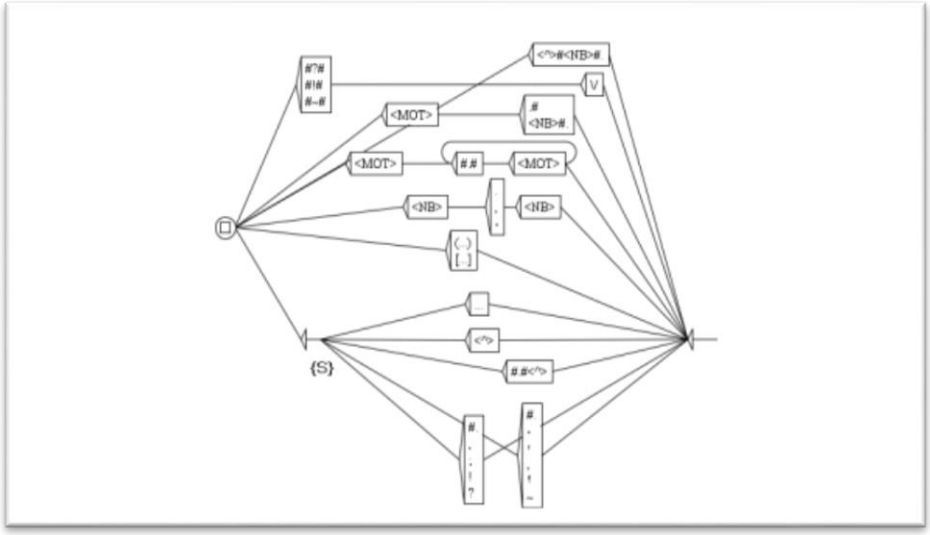

FIGURE 13 : TRANSDUCTEUR DE SEGMENTATION DE CORPUS EN PHRASES

4. La détection des entités nommées arabes

Pour détecter les entités nommées arabes on applique une cascade de transducteurs. Chaque élément dans cette cascade apporte un changement servira par la suite aux autres transducteurs qui viennent après celui-ci. Pendant l'annotation on utilise la typologie Quaero qui respect la norme TEI.**[25]**

4.1. TEI

La TEI (TextEncoding Initiative) est un projet universitaire pluridisciplinaire visant à uniformiser autant que possible le codage de documents en vue de leur échange et de leur publication en ligne ou hors ligne. Il s'agit d'un format de codage de documents dit structuré : il a besoin d'un langage, XML, pour aider à la saisie d'un texte en lui donnant une structure compatible à la fois avec les exigences des différentes communautés qui l'utilisent et avec les possibilités des outils de consultation.**[25]**

5. L'annotation des entités nommées arabes

5.1. Les noms de personnes

Les noms de personnes peuvent se trouver dans un texte arabe sous différentes formes. Les noms de personnes dans un texte de l'arabe classique se caractérisent par l'absence de la structure moderne alors que les noms de personnes de l'arabe standard moderne varient entre les régions du monde arabe : les noms de la région du Maghreb diffèrent des noms de l'Egypte.**[25]**

5.2. Le dictionnaire des prénoms

Pour notre travail nous avons opté pour la collection des prénoms arabes à partir de Wikipédia arabe proposant des prénoms pour les nouveau-nés dans le monde arabe. **La figure 14** montre un extrait de la liste des prénoms arabes compilée sous forme d'un dictionnaire LADL sous la plateforme Unitex/GramLab.

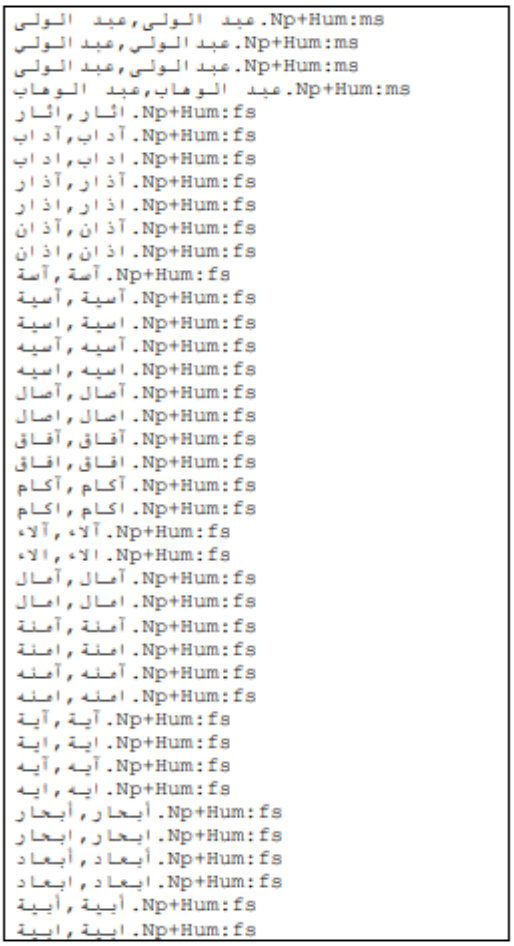

FIGURE 14 : EXTRAIT DU DICTIONNAIRE DES PRENOMS UTILISE DANS LES TRAVAUX DE RECONNAISSANCE DES EN DE TYPE NOM DE PERSONNE SOUS LA PLATEFORME UNITEX/GRAMLAB

5.3. Les dates

L'entité date est diviser en deux classe : date et heure. Ces deux classes peuvent être absolues ou relative **[25]**. La **figure 15**montre les classes et les sous classes de la catégorie temps. **[25]**

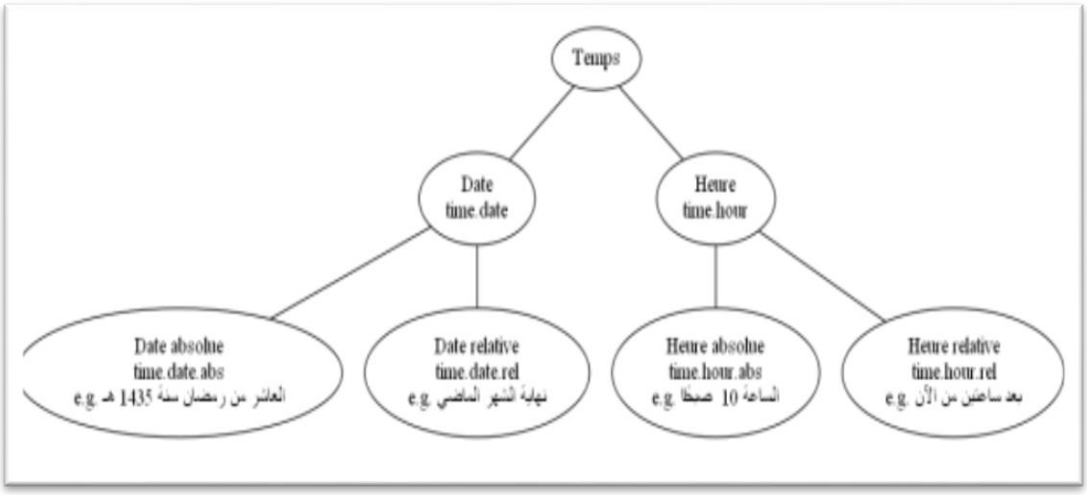

FIGURE 15 : LES CLASSES ET LES SOUS CLASSES DE CATEGORIE TEMPS.

5.4. Les lieux

Les entités nommées de type lieux se divisent en cinq sous-classes : administratif, géographique, voies, bâtiments et adresses. Une sixième sous-classe (autre) est ajoutée pour regrouper le reste des lieux qui ne peuvent pas être annotés par les cinq classes.**[25]**

Conclusion

Unitex est un système multiplateforme capable de fonctionner aussi bien sous Windows que sous Linux ou OS X. Unitex est un ensemble de logiciels permettant de traiter des textes en langues naturelles en utilisant des ressources linguistiques. Ces ressources se présentent sous la forme de dictionnaires électroniques, de grammaires et de tables de lexique-grammaire.

Ce chapitre, décrit la segmentation des cliques et la segmentation en phrase, définit la détection des entités nommées arabe et leur annotation sur Unitex et montre les étapes avec lesquelles nous avons fait le traitement automatique avec logiciel Unitex et le dictionnaire Prolexbase.

Chapitre 4

Implémentation

Introduction

Dans le présent chapitre, nous allons présenter le processus de création et de construction du module arabe de la plateforme Unitex /GramLab. Cette opération est nécessaire pour nous permettre d'atteindre notre premier objectif qui est l'extraction des entités nommées à partir du texte arabe.

Notre tâche étant de traiter un corpus textuel, nous allons à cet effet utiliser la plateforme Unitex/GramLab. Cette plateforme est un logiciel de traitement automatique de corpus qui regroupe un ensemble de programme réalisant les différents taches dont l'utilisateur a besoin.

1. L'utilisation de logiciel Unitex

Lorsque l'on utilise Unitex pour la première fois, le logiciel demande à l'utilisateur de choisir un répertoire de l'ordinateur où il veut stocker ses données. Dans ce répertoire, le logiciel installe six dossiers : Cassys, Corpus, Dela, Elag, Graphs, Inflection. L'utilisateur place dans le dossier Corpus les fichiers à soumettre à l'analyse : il sera ainsi possible de les sélectionner depuis le menu Text. **La figure 16** montre comment sélectionner un texte pour le traiter.

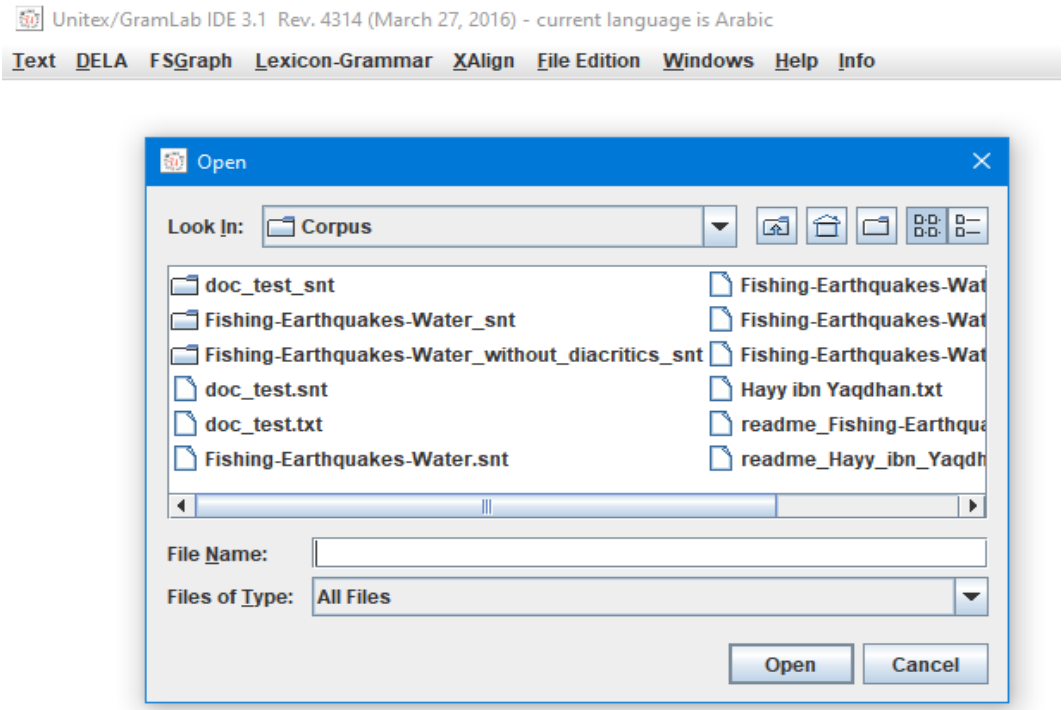

FIGURE 16 : SELECTION UN TEXTE A TRAITER

Une fois ouvert le texte, le logiciel ouvre une interface qui demande si l'on veut prétraiter le texte. Par défaut, le logiciel applique au texte les dictionnaires disponibles pour la langue choisie. Si c'est nécessaire à de l'analyse, on peut demander de produire aussi l'automate du texte à la fin de l'étape de prétraitement, en cochant la case ConstructTextAutomaton (en bas à gauche).**La figure 17**définir l'interface qui nous donne l'accès au prétraitement.

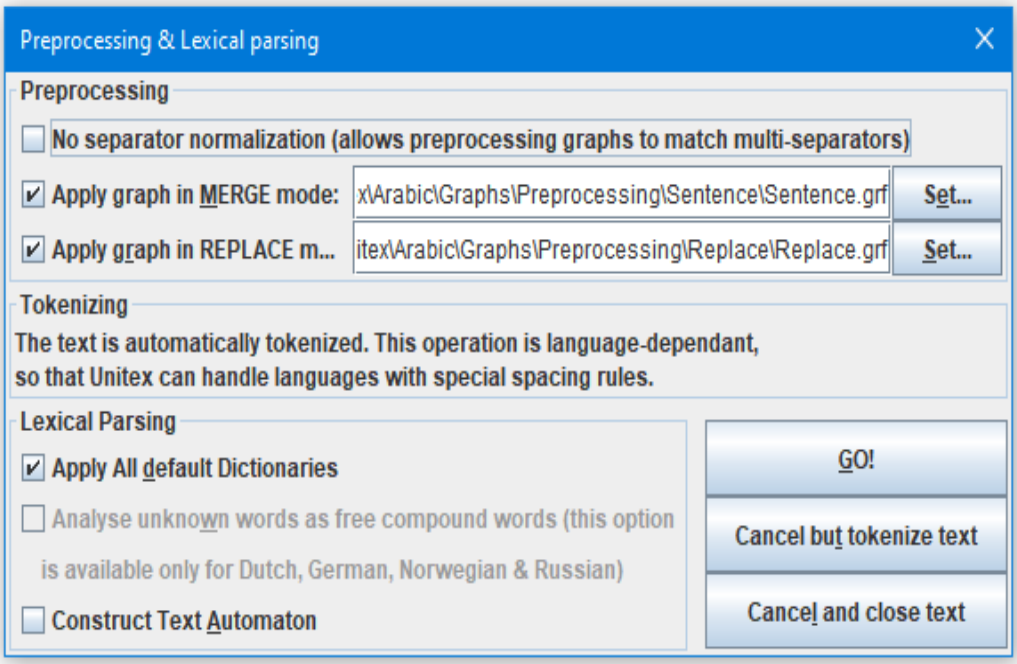

FIGURE 17 : L'INTERFACE PERMETTANT L'ACCES AU PRETRAITEMENT

Une fois que nous avons cliqué sur GO, le logiciel commence à faire le chargement du texte qu'on n'a sélectionnés ensuite unitex ouvre une interface dans les quelles en trouve le texte avec toutes les informations.

La **figure 18 est l'**interface du texte que nous avons choisir.

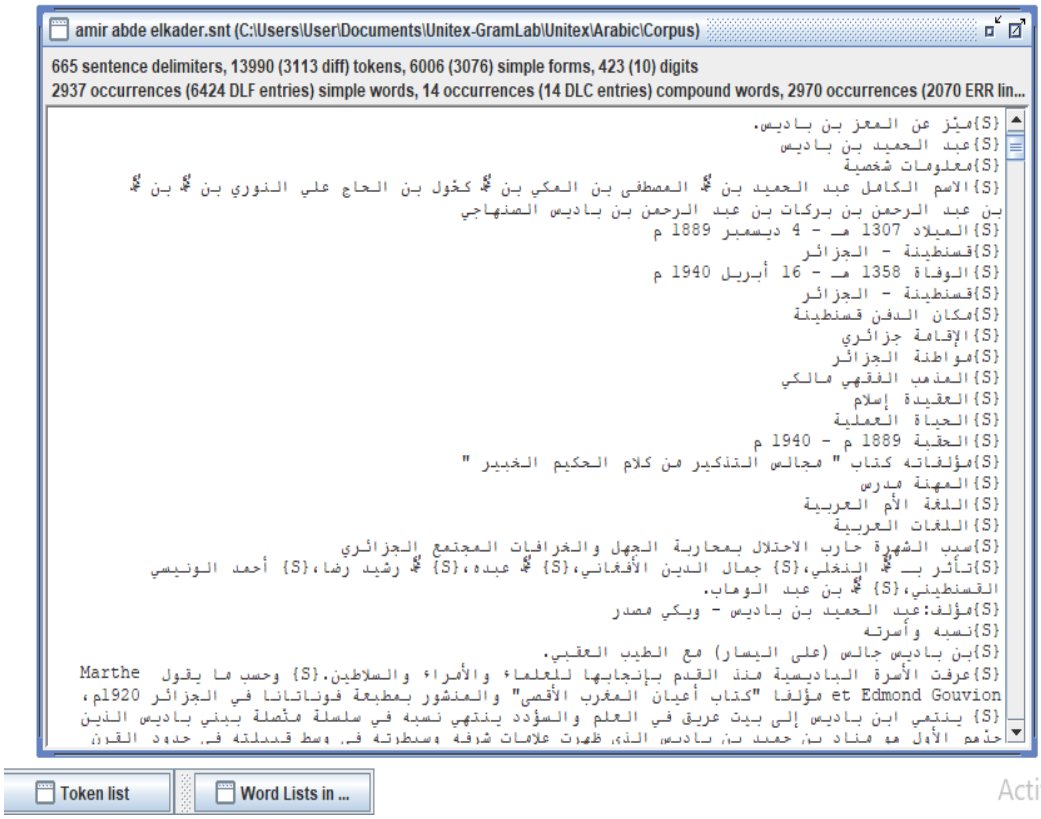

FIGURE 18 : LE TEXTE QUE NOUS AVONS CHOISI

Trois opérations fondamentales sont exécutées pendant la phase de prétraitement : le comptage des formes du texte, l'étiquetage de ces formes, la segmentation du texte en phrases. Les résultats de ces opérations sont affichés dans trois fenêtres différentes. Ainsi, dans la première fenêtre (Token List) sont données toutes les formes présentes dans le texte (signes diacritiques inclus) avec le nombre d'occurrences. Il est possible d'afficher la liste par fréquence (ordre décroissant) ou par ordre alphabétique.

La figure 19 montre Token List par fréquence.

La figure 20 montre Token List avec le nombre d'occurrences.

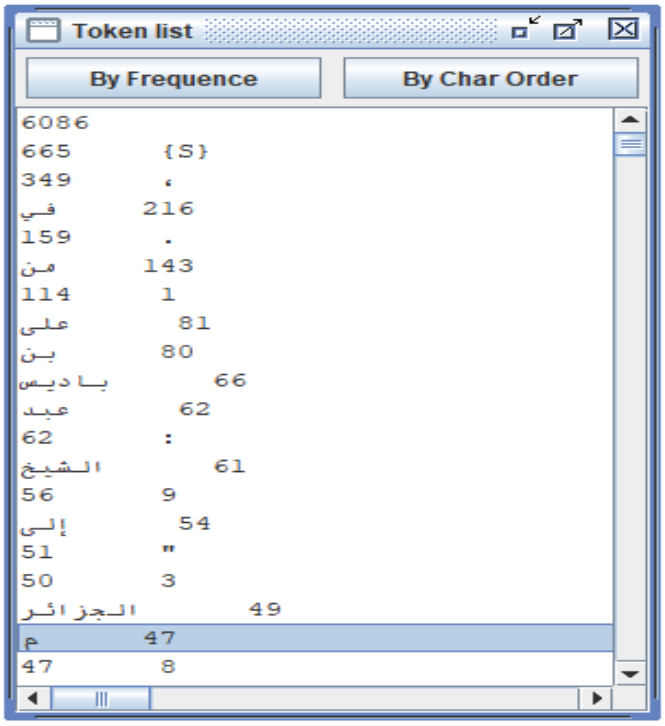

FIGURE 19 : LA FENETRE TOKEN LISTE PAR FREQUENCE

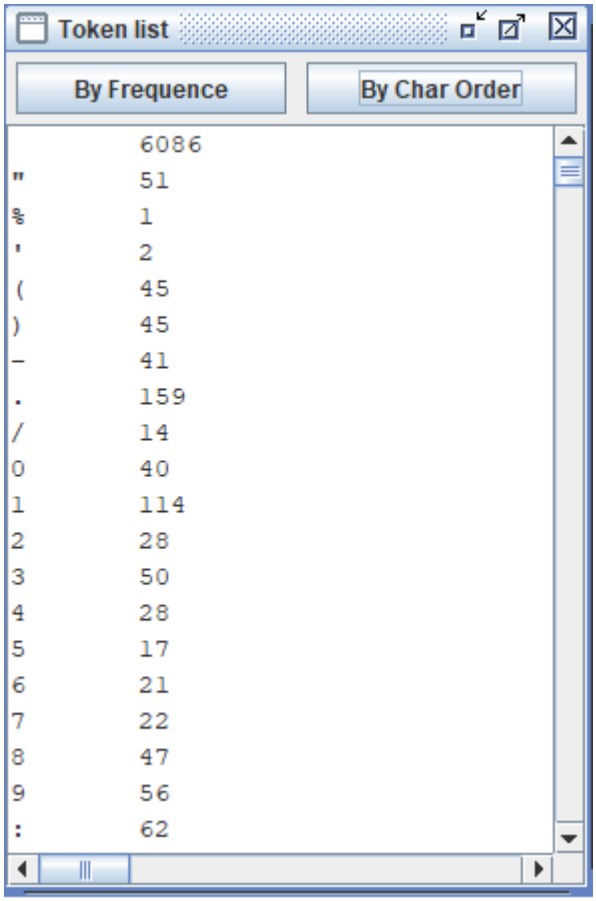

FIGURE 20 : LA FENETRE TOKEN LISTE AVEC NOMBRE D'OCCURRENCES

La deuxième fenêtre, Word Lists, est divisée en trois sous-fenêtres : une contenant les mots simples, une autre listant les formes composées (dans ces deux premiers cas, il s'agit des formes reconnues par les dictionnaires appliqués) et une dernière dans laquelle sont listées toutes les formes non reconnues par les dictionnaires**[26].**

La figure 21 : montre Word Liste et les trois fenêtres

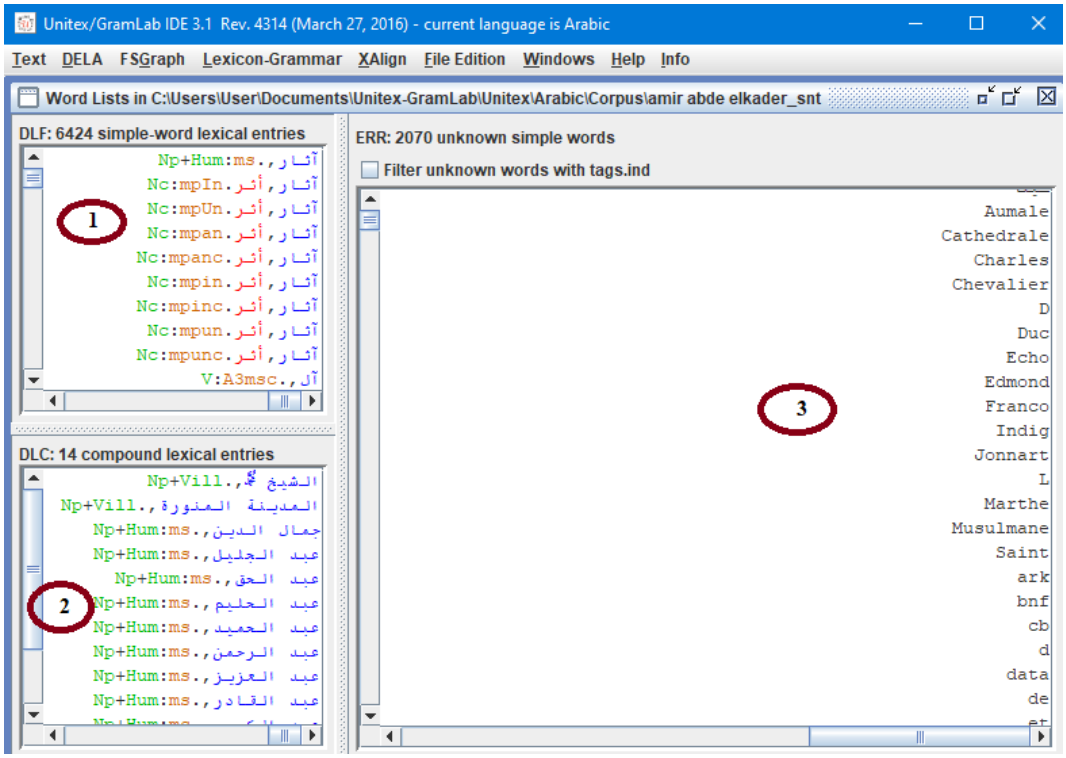

FIGURE 21 : LA FENETRE WORLD LISTS

1_ Dictionnaire des formes simple.

2_Dictionnaire des formes composées.

3_ Liste des noms non reconnus par le dictionnaire.

Après, terminer les étapes précédentes nous avons récupéré le fichier.snt. Ce fichier contient Tout les données de la fenêtre World Lists (dictionnaire des formes simple et dictionnaire des formes composées).

La figure 22 montres le fichier.SNT que nous avons ouvert avec bloc note.

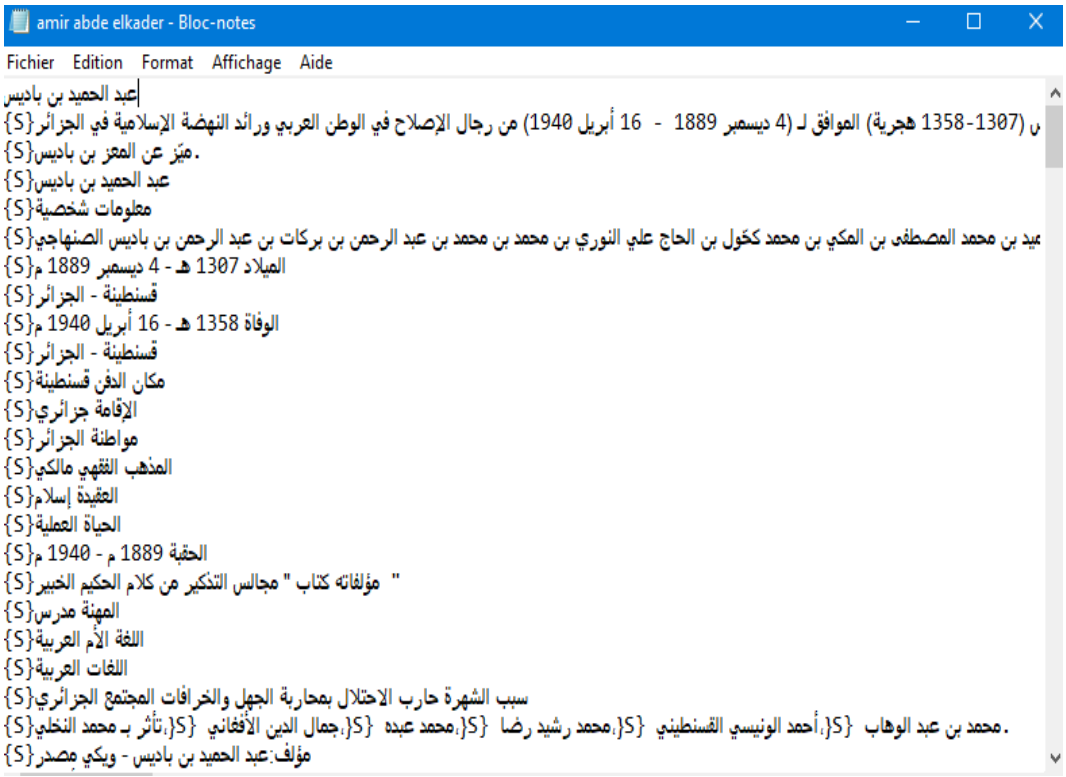

FIGURE 22 : LE FICHIER.SNT

2. L’utilisation de dictionnaire Prolexbase

Pour l’utilisation de dictionnaire de Prolexbase nous avons téléchargé la base de données de dictionnaire sur XamppServeur.

2.1. XamppServeur

XAMPP est une distribution Apache entièrement gratuite et facile à installer contenant MySQL, PHP et Perl. Le paquetage open source XAMPP a été mis au point pour être incroyablement facile à installer et à utiliser.**[28]**

1. Les étapes pour télécharger la base de données

Pour la première partie nous avons commencé par Xampp contrôle. **La figure 23** montre la démarche que nous avons suivie.

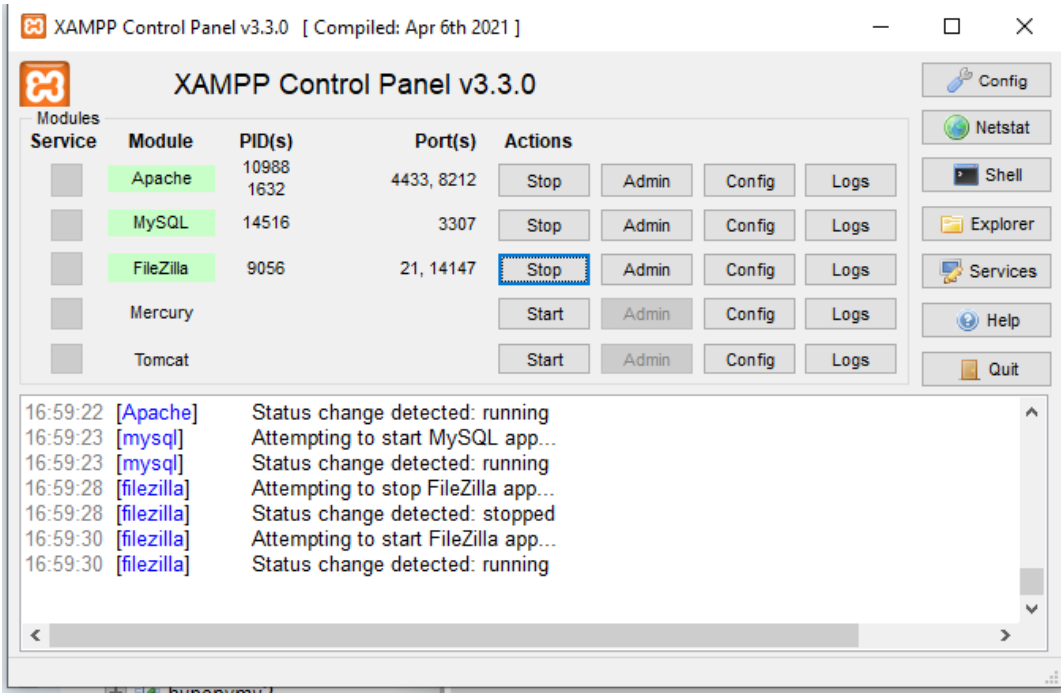

FIGURE 23 : LA FENETRE XAMPP CONTROL

La deuxième étape est télécharger la base de données de dictionnaire Prolexbase dans PHmyAdmine. **La figure 24** montre les tables de base de données de dictionnaire Prolexbase.

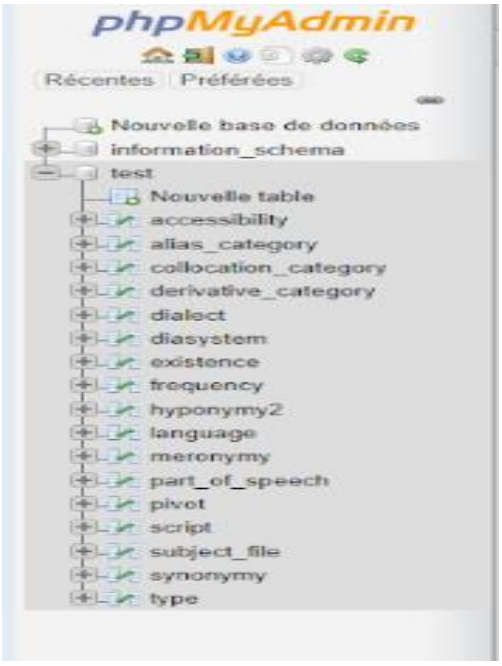

FIGURE 24 : LES TABLES DE PROLEXBASE

La figure 25 montre la table de pivot.

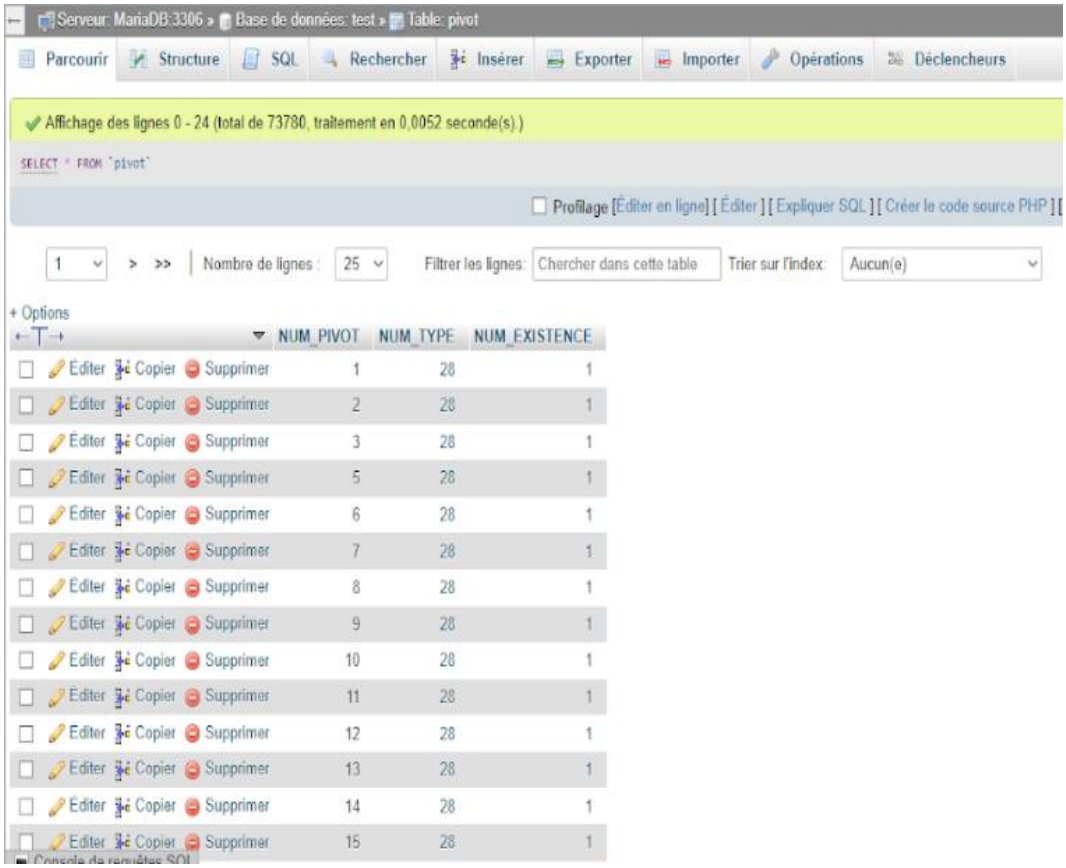

FIGURE 25 : TABLE DE PIVOT

3. L'utilisation de Notepad2

Notepad2 est un éditeur de texte libre. Nous avons utilisé le code de calculer notoriété de « mouna-elshter » et dans le code nous avons ajouté la fonction ajouter pour l'ajout des entités nommée dans Prolexbase.

La figure 26 montre une partie de code.

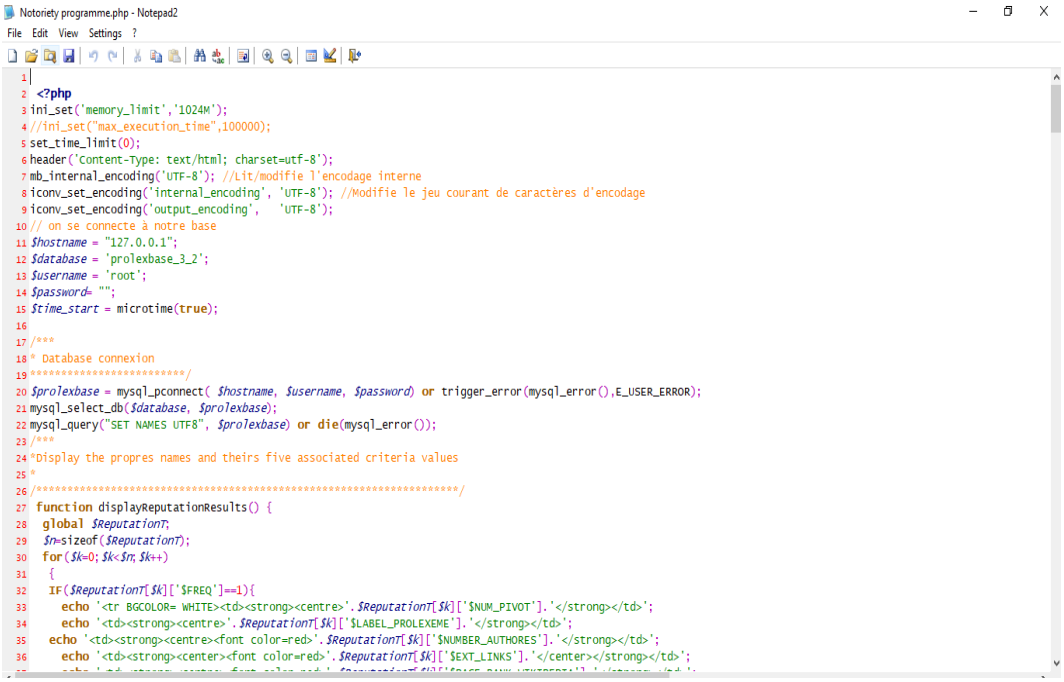

FIGURE 26 : PARTIE DE CODE

Conclusion

Unitex présente un potentiel d'exploitation intéressant pour satisfaire les besoins des terminologues (notamment, la recherche de termes, contextes et traductions).

Dans ce chapitre, nous avons fourni quelques données génériques à propos du logiciel. Après, nous avons illustré l'utilisation de logiciel Unitex, nous avons montré aussi l'utilisation de la base donnée de prolexbase et nous avons fini ce chapitre par l'utilisation de notepad2.

Conclusion générale

Dans le présent travail, nous avons pu réaliser un système de reconnaissance d'entité nommée. Ce système permet de reconnaître les entités nommées arabe. Le système agit sur un corpus Wikipédia arabe qui est le fruit d'un processus de reconnaissance d'entité nommée effectué par ce système. Le système se base sur de dictionnaire de haute couverture qui est participé à leur tour de guider le processus de reconnaissance. Pour ce faire, nous avons effectué un état de l'art qui se compose de deux parties.

Dans la première partie, nous avons étudié la reconnaissance d'entité nommée arabe. Cette étude nous a permis de découvrir les trois approches de la RENA. De plus, nous avons distingué les catégorisations des EN. Nous avons effectué aussi une étude linguistique sur de plusieurs phénomènes rencontrés dans ENA. Ensuite, nous avons définie dictionnaire électronique multilingue relationnel.

Dans la deuxième partie de l'état de l'art, nous avons compris les ressources libres plus précisément la Wikipédia arabe. Nous avons effectué aussi une étude sur La structure générale d'une page Wikipédia. Ensuite, nous avons discuté les systèmes des REN arabe. De plus, nous avons cités les travaux exploités sur Wikipédia spécifié à la langue arabe.

Dans la troisième partie, nous avons cite les traitements automatique avec logiciel Unitex et le dictionnaire Prolexbase et nous avons défini la segmentation des clitique et en phrase.

Dans le quatrième chapitre, nous avons discuté sur l'implémentation de système. Nous avons cité les étapes que nous avons suivi avec logiciel Unitex, dictionnaire Prolexbase et notepad2.

Finalement, toutes les tâches déjà mentionnées vont participer à un projet de création d'un système de la reconnaissance d'entité nommée arabe dans un corpus Wikipédia arabe en utilisant un dictionnaire électronique multilingue relationnel « Prolexbase » ce dictionnaire électronique comporte des entités

nommées arabe normalisées et bien définies via les relations entre elles et vers des ressources libres.

Bibliographie

[1] Fatma Ben Mesmia, «Reconnaissance des entités nommées à partir de Wikipédia arabe » Traitement du texte et du document. Université de Tunis El Manar, (2019). Français.

[3] HelaFehri, « Reconnaissance automatique des entités nommées arabes et leur traduction vers le français » Université de Franche-Comté ; Université de Sfax. Faculté des sciences, (2012). Français

[4] SouhirGahbiche-Braham— Hélène Bonneau-Maynard —François Yvon, « Traitement automatique des entités nommées en arabe : détection et traduction »Université Paris Sud & LIMSI-CNRS, le 26/09/2013

[5] HoudaSaadane, « Le traitement automatique de l'arabe dialectalisé : aspects méthodologiques et algorithmiques »Linguistique. Université Grenoble Alpes, (2015)

[7] MounaElashter, « Gestion et extension automatiques du dictionnaire relationnel multilingues de noms propres Prolexbase», thèse doctorat 'université François – Rabelais de Tours (2017).

[25] NourddineDoumi, « extraction de connaissances a partir du texte », thèse doctorat 'université Sidi Bel abbes (2017).

[26]Rosa Cetro.,« Lexique-grammaire et Unitex». Université Paris-Est, 2013. Français.

Webographie

[2]https://fr.wikipedia.org/wiki/Reconnaissance_d%27entit%C3%A9s_nomm%C3%A9es Consulte le 2 février 2021

[6] http://www.cnrtl.fr/lexiques/prolex/

[8]https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Accueil_principal Consulte le 8 novembre 2019

[9] https://igm.univ-mlv.fr//~dr/XPOSE2011/Wikipedia/presentation_wikipedia.html

[10]http://www.alexa.com/topsitesConsulte le 2022

[11]https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Cat%C3%A9goriesConsulte le 10 janvier 2022

[12]https://www.cetic.be/Exploiter-le-contenu-de-Wikipedia

[13]https://fr.wikipedia.org/wiki/Aide:Lien_interwikiConsulte le 19 juillet 2021

[14]https://fr.wikipedia.org/wiki/Wikipedia:Liens_externes

[15]http://fr.howtopedia.org/wiki/Aide:Liens_internes

[16]https://fr.wikipedia.org/wiki/Aide:Insérer_une_référenceConsulte le 8 février 2021

[17]https://www.wikimedia.fr/le-mouvement-wikimedia/Consulte 2019

[18]https://www.wikimedia.fr/

[19]https://dumps.wikimedia.org/

[20]https://fr.wikipedia.org/wiki/Aide:MediaWikiConsulte le 18 mai 2019

[21]https://fr.wikipedia.org/wiki/Site_miroirConsulte le 24 juillet 2021

[22]https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Wikip%C3%A9dia_hors-connexion#1re_.C3.A9tape_:_Installer_MediaWikiConsulte le 13 février 2022

[23]https://fr.wikipedia.org/wiki/DBpediaConsulte le23 février 2022

[24]https://ar.wikipedia.org/wiki

[15]http://fr.howtopedia.org/wiki/Aide:Liens_internes

[16]https://fr.wikipedia.org/wiki/Aide:Insérer_une_référenceConsulte le 8 février 2021

[17]https://www.wikimedia.fr/le-mouvement-wikimedia/Consulte 2019

[18]https://www.wikimedia.fr/

[19]https://dumps.wikimedia.org/

[20]https://fr.wikipedia.org/wiki/Aide:MediaWikiConsulte le 18 mai 2019

[21]https://fr.wikipedia.org/wiki/Site_miroirConsulte le24 juillet 2021

[22]https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Wikip%C3%A9dia_hors-connexion#1re_.C3.A9tape_:_Installer_MediaWikiConsulte le 13 février 2022

[23]https://fr.wikipedia.org/wiki/DBpediaConsulte le23 février 2022

[24]https://ar.wikipedia.org/wiki

[27]https://fr.wikipedia.org/wiki/Notepad2

[28]https://www.clubic.com/telecharger-fiche27009-wampserver.html