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

MEMOIRE DE FIN D'ETUDES

Pour l'Obtention du Diplôme de Master en Informatique

Option : Ingénierie des Systèmes d'Information

THÈME :« Enrichissement d'un lexique de noms propres »

Etudiant(e) : « Elkaid Maroua »

Etudiant(e) : « Mokhtari Mohammed »

Encadrant(e) : « Kenniche Ahlem »

Résumé

Le dictionnaire électronique relationnel multilingue de noms propres, Prolexbase, issu de nombreux travaux de recherche sur le TAL, comporte à ce jour dix langues, parmi lesquelles trois sont bien couvertes : le français, l'anglais et le polonais. Pour le traitement automatique des langues (TAL), Les bases de données lexicales sont indispensables comme l'extraction d'information, la reconnaissance d'entités nommées et la traduction automatique des noms propres.

Ce rapport présente les différentes étapes de la conception d'un système d'enrichissement d'un lexique de nom propres. Ce système s'appuie sur la reconnaissance des entités nommées arabe à l'aide d'un dictionnaire électronique relationnel multilingue de noms propres.

Mots-clés

Nom propre, Prolexbase, Bases lexicales multilingues, Langue arabe, Wikipédia.

Abstract

The multilingual relational electronic dictionary of proper names, Prolexbase, resulting from numerous researches works on NLP, currently includes ten languages, of which three are well covered: French, English and Polish. For automatic language processing (TAL), lexical databases are essential such as information extraction, recognition of named entities and automatic translation of proper names. This report presents the different stages of the design of a system for enriching a lexicon of proper nouns. This system is based on the recognition of Arabic named entities using a multilingual relational electronic dictionary of proper names.

Key Words

Proper noun, Prolexbase, Multilingual lexical databases, Arabic language, Wikipedia.

DÉDICACES

Je dédie ce modeste travail en signe de respect, de
Reconnaissance et de remerciement à mes parents pour leur
Soutient et patience ainsi qu'à tous les membres de ma
Famille, mes amies : Hadjer, Nihel, Meriem.

Elkaid Maroua

Je dédie ce travail d'abord à mes parents
Sans oublie mes amies qui m'ont Beaucoup aidé et soutenu.

Mokhtari Mohammed

REMERCIEMENTS

Nous tenons à remercier tous ceux qui nous ont soutenus

A réaliser ce travail en particulier, notre encadreur

Mme Kenniche Ahlem qui nous a orientés, accompagnes

Tout le long de ce petit projet.

Et grâce à ses conseils et son accompagnement que nous avons pu abouti à ce résultat.

Nos vifs remerciements vont également aux membres du jury qui ont accepté de juger notre travail.

Merci !!

Liste figure

Liste des tableaux

Liste des abréviations

Abréviation	Expression complet
EN	Entité Nommée
ENA	Entité Nommée Arabe
REN	Reconnaissance des Entités Nommées
TAL	Traitement Automatique des Langues
CoNLL	Conference On Natural Language Learning
MUDC	Message Understanding Conference

Table des matières

Introduction Générale

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 et la traduction automatique des noms propres. Elles nécessitent un développement et un enrichissement permanents via l'exploitation des ressources libres et riches en textes du web sémantique, entre autres, l'encyclopédie universelle Wikipédia.

Prolexbase est une ressource libre et un dictionnaire relationnel multilingue de noms propres. Il comporte à ce jour dix langues, parmi lesquelles trois sont bien couvertes : le français, l'anglais et le polonais. Notre objectif principal consiste à effectuer l'enrichissement d'un lexique de noms propres dans l'objectif et d'améliorer la performance de Prolexbase par l'ajout des liens Wikipédia et des numéros de pivot, ce qui contribuera à préparer l'extension à d'autres langues. A cet effet, nous avons également choisi d'élargir notre étude à l'arabe, par ce que La langue arabe est très répandue dans le monde car elle est parlée par plus de 300 millions de locuteurs.

Dans le premier chapitre, nous avons défini La reconnaissance des entités nommées leurs catégorisations et les trois approches principales de REN : Approche symbolique, Approche statistique et l'Approche hybride. Ensuite, nous décrivons la REN arabe. Puis, nous présentons la langue arabe et ses principes.

Dans le deuxième chapitre, nous allons décrire prolexbase et ses niveaux (Le niveau des instances, Le niveau linguistique, Le niveau conceptuel, méta-conceptuel). Déplus, nous allons définir la structure de l'encyclopédie Wikipédia. Puis, nous allons présenter la Wikimédia, l'accès au Wikipédia (DBPEDIA, dumps). En fin nous allons discuter de la Wikipédia arabe.

Dans le troisième chapitre, d'abord nous allons définir la méthode pour extraire les données d'un nom propre à partir d'une url, ensuite nous avons présenté les détails de notre conception du système d'enrichissement en langue arabe.

Dans le quatrième chapitre, nous avons défini les outils et les langages utilisés dans notre thème et présenté les différentes interfaces sur enrichissement et les résultats obtenus.

Enfin nous terminons ce rapport par une conclusion générale.

Chapitre1 :

Reconnaissance des Entités nommées

I. Introduction :

Une entité nommée est une expression linguistique qui désigne un nom de lieu, un nom de personne ou un nom d'organisation. Les entités nommées sont utilisées dans le domaine du traitement automatique du langage ou dans l'analyse de corpus de textes. Dans le domaine des graphes de connaissance, les entités nommées sont utilisées par les moteurs de recherche pour constituer les éléments des bases de connaissance. On appelle les graphes de connaissance "knowledge graph" en anglais. Ces graphes sont utilisés pour les résultats de la recherche vocale dans les moteurs de recherche qui tentent de devenir des moteurs de réponses.

I.1 La reconnaissance des entités nommées

I.1.1 Définitions d'une entité nommée

Le terme EN (Entité Nommée) est apparu au cours de la sixième conférence MUC (Conférences sur la compréhension de messages, en anglais, Message Understanding Conference) en 1996. Une EN désigne les noms de tous les personnes, organisations et lieux dans un texte.

La REN est également appelée extraction d'entité ou identification d'entité. C'est une sous-tâche très importante de l'extraction de l'information qui vise à trouver et classifier le nom dans un texte Plusieurs chercheurs ont proposé des définitions différentes pour ce terme. Notez que la définition d'EN est relié au domaine intéressé.[13]

I.1.2 Catégorisations d'une entité nommée

I.1.2.1 Catégorisation des conférences MUC

La conférence MUC a été créée dans le but de promouvoir la recherche en invitant les chercheurs à venir participer avec leurs outils et leurs systèmes à une compétition annuelle d'extraction de l'information.

Les participants étaient alors invités à développer un système qui permet l'extraction du plus grand nombre d'informations possibles sur des entités bien précises.

À partir de la sixième édition de MUC, baptisée MUC-6, la tâche d'extraction des EN a été créée et par la même occasion la notion d'entités nommées a été introduite.

La conférence MUC-7 a distingué trois types d'entités à reconnaître et à catégoriser, soit ENAMEX, NUMEX et TIMEX. La Figure 1 montre ces trois grandes catégories et leur limite par rapport à la grande famille des EN. On remarque clairement qu'il y a une majeure partie d'EN qui n'est pas couverte par la classification MUC.[2]

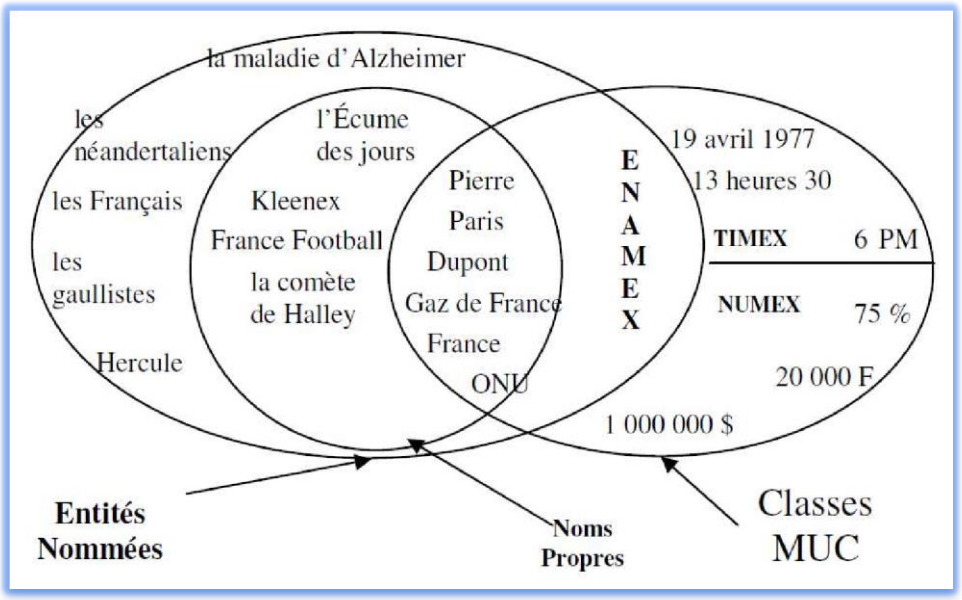

Figure 1 : Les entités nommées vs la classification MUC [9]

I.1.2.2 Catégorisation de la conférence CoNLL

Les conférences CoNLL 2002 et 2003 [20] procèdent aussi à quelques changements par rapport à la catégorisation de MUC. En effet, elles reprennent uniquement les catégories Personne, Lieu et Organisation et ajoutent la catégorie Miscellaneous pour annoter les entités n'appartenant pas aux classes de MUC.[1]

I.1.2.3 Catégorisation de la compagne ESTER

La campagne française ESTER a imposé la catégorisation des EN en huit catégories, qui sont : Personne, Organisation, Groupe geo-Socio-Politique (GSP), Lieu, Bâtiment, Produit, Temps et Quantité, auxquelles est ajoutée, en cas d'incertitude, la catégorie Inconnu ; l'héritage d'ACE se fait ici sentir au travers des catégories GSP (proche de GPE) et Bâtiment, celui de IREX au travers de la catégorie Produit.[1]

I.1.3 Approches de REN

I.1.3.1 Approchesymbolique

L'approchesymbolique(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 description d'enchaînements possibles de syntagmes nominaux. Ces patrons exploitent généralement des informations d'ordre morphosyntaxique telles que les mots déclencheurs (السيّدMrملعب , stade), ainsi que celles contenues dans des ressources (lexiques ou dictionnaires).

Dans ce cadre, [21]a défini l'architecture d'un système de reconnaissance des EN arabes. Ce système est fondé sur une approche linguistique utilisant les grammaires locales implémentées

dans la plateforme linguistique NooJ.[1] L'architecture de ce système est illustrée dans la Figure 2. [21]

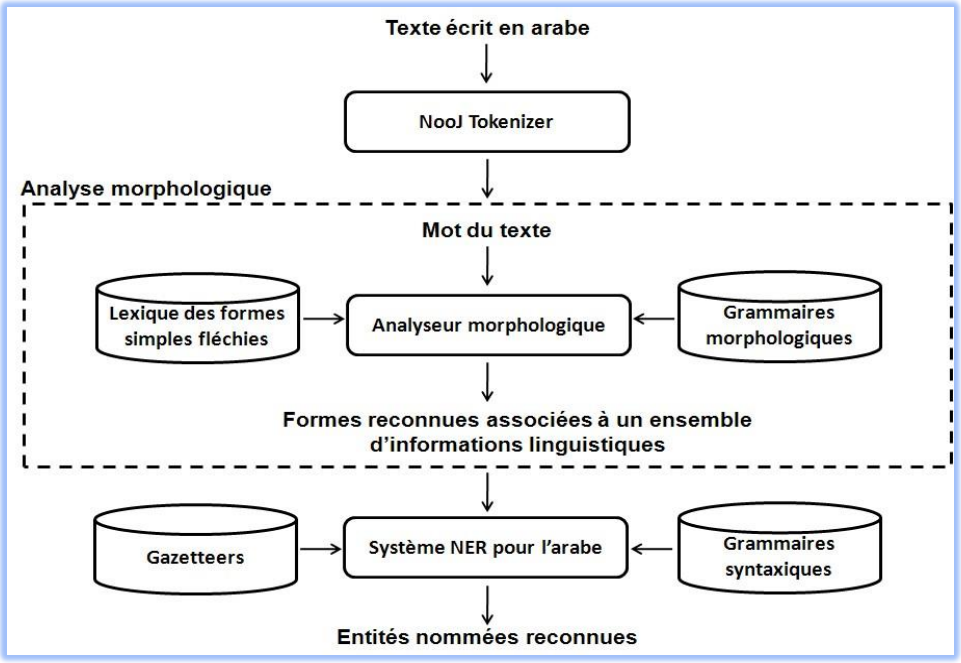

Figure 2. Architecture générale d'un système de reconnaissance des EN [21]

I.1.3.2 Approche statistique

Elle emploie des techniques mathématiques, ainsi elle utilise des corpus de textes pour le développement des modèles des phénomènes linguistiques. Un modèle fréquemment utilisé est le modèle caché de Markov (HMM). Le HMM est un automate d'états finis avec un ensemble d'états et des probabilités attachés aux transitions (automate flou). L'approche statistique a été typiquement utilisée dans des tâches telles que la reconnaissance vocale, l'acquisition de lexique, le POS tagging, la traduction statistique, ect.[3]

I.1.3.3 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, cette idée se traduit par l'association de la puissance descriptive des solutions linguistiques et des facilités statistiques d'apprentissage. Généralement, la partie essentielle de la méthode d'extraction est statistique. Quant à la partie linguistique, elle consiste à filtrer les termes en fonction de leur catégorie syntaxique [30]. Bien évidemment, l'utilisation d'information sémantique est envisagée bien qu'elle soit difficile à mettre en œuvre.[1]

Parmi les projets admettant cette approche hybride, nous citons le système d'extraction des EN arabes proposé par [22]. Les étapes linguistiques sont illustrées dans la **Figure 3**. [1]

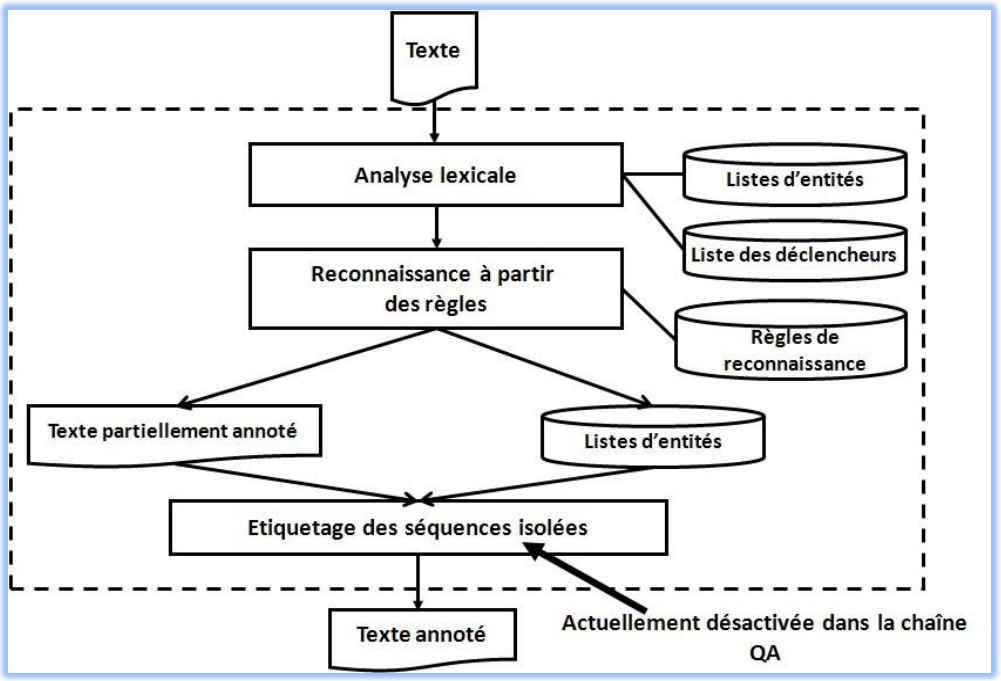

Figure 3. Démarche proposée [1]

I.2 Reconnaissance d'entité nommée Arabe

I.2.1 Définition

La définition que nous avons retenue est inspirée de plusieurs travaux (p.ex, [31],[32],[9], [33]et[34]) et des encyclopédies libres (p.ex., l'Atalapédie et le Wikipédia). Cette définition consiste à considérer l'EN comme étant un nom propre dans son sens élargi (i.e., prénom, syntagme nominal), une expression numérique, une fonction, un événement ou un terme. En effet, il s'agit d'une catégorisation plus large et qui peut regrouper les différentes catégories de la langue arabe. De plus, toutes les définitions proposées pour les EN exigent le principe de référence et la majorité d'elles n'ont pas imposé l'existence d'autres critères. Dans ce cas, une EN n'est pas nécessairement unique et peut être un nom propre ou une expression de temps ou de quantité bien qu'il existe d'autres définitions qui exigent l'unicité. Donc, nous considérons que par exemple ملعب الطيب المهيري بصفاقسmal`abelTayibelmhyrybiSafaaqus 28 (stade de taieb al mhiri à sfax) est une EN mais ملعب الطيب المهيريmal`abelTayibelmhyry (stade de taieb almhiri) est aussi une EN. En effet, les deux entités réfèrent à un nom de lieu et exactement à un nom de stade.[1]

I.2.2 Présentation de la langue arabe اللغة العربية

La langue arabe est l'une des langues les plus parlées et utilisées dans le monde. Elle est la langue officielle de plus de 22 pays parlé par plus de 320 millions de personnes et elle est utilisée comme vecteur de transmission religieux pour tous les croyants musulmans au nombre de 1 milliard et demi à travers les cinq continents du globe. Elle constitue ainsi un élément principal dans la culture et la pensée d'une partie importante de l'humanité et du patrimoine mondial. A l'origine, les peuples de la péninsule arabe tenaient le monopole de cette langue qui est sémitique (comme l'hébreu ou l'araméen), mais du fait qu'elle est la langue du coran elle s'est étendue au-delà du golfe arabo-persique, atteignant l'Afrique du nord et l'Asie mineur. De plus, l'expansion territoriale de l'empire musulman a fait de l'arabe une langue d'administration, de culture et de sciences à travers son utilisation dans la définition et la rédaction des contrats et des lois, la rédaction de manuscrits et de livres, la transmission et la formation, etc. Par ailleurs, la diversité des populations arabes et de leurs cultures ont fait émerger différentes variantes de l'arabe allant de l'arabe classique utilisé dans le coran, à l'arabe standard moderne (ASM). [4]

I.2.3 Système d'écriture de l'arabe

Comme mentionné dans la section précédente, l'arabe est classé sous le groupe des langues sémitiques contemporaines qui s'écrit de droite à gauche. Son système graphique se compose d'un alphabet arabe de type abjad constitué de 28 lettres. Cet alphabet contient 25 consonnes et 3 voyelles longues « ا« ,» و «et « ي .«L'écriture arabe comporte aussi des voyelles courtes qui sont généralement facultative mais essentielles dans les textes religieux (Coran, Hadith, etc.). Il existe de plus, une série d'autres diacritiques dont les plus courants comme l'indication de l'absence de voyelle (سكون -sukun) et la gémination des consonnes (شدة- shadda). En arabe les mots indéfinis, qui ne sont pas associé à des articles ou à des compléments du nom, prennent les désinences (nounatation ou tanwine) notées par des diacritiques spéciaux.

Nous signalons également que les notions de lettres majuscules et de lettres minuscules n'existent pas dans la langue arabe (l'écriture est donc monocamérale). Aussi, l'arabe est semi cursive dans le sens où son alphabet est unique mais la forme des lettres change en fonction de la position qu'elles occupent dans le mot. Chaque lettre possède une forme spécifique en fonction de sa position dans un mot (au début, au milieu ou à la fin) ou si elles sont utilisées de façon isolée.[5]

I.2.4 Lexique et grammaire

Dans cette section nous donnons une présentation sommaire du lexique et grammaire de la langue arabe, tout en mettant l'accent sur les éléments qui seront pris en charge en priorité dans notre étude. Nous trouvons différentes structurations du lexique de l'arabe, basées essentiellement sur les sous-ensembles : noms, verbes et particules, et augmentées avec d'autres sous-ensemble afin d'avoir suffisamment d'éléments pour un traitement automatique de la langue. Nous trouvons entre autres les classifications de [25]et[24]. Nous considérons dans étude une classification proche de celle de [24]ayant les éléments suivants :

I.2.4.1 Nom

Est une entité ou un élément qui exprime un sens indépendamment du temps pour désigner un objet ou un être. Nous pouvons répartir les noms en trois catégories selon le système morphologique comme suit :

A- Les primitifs : sont les noms qui constituent le glossaire fondamental de la langue arabe, et représentent les noms qui ne peuvent pas être rattachés à une racine verbale. Cette catégorie inclue aussi les noms propres, les noms communs et les racines bilitères. Par exemple, nous citons رأسraa's 'tête', مُحَمَّد' Mohammed' et فَمfam 'bouche'.[5]

B- Les dérivés : sont les noms formés à partir d'une racine verbale. Le statut de cette dernière détermine la nature et le nombre de ces formes. Nous trouvons dans cette catégorie les participes actifs (ضار بٌ َ – ِcelui qui frappe), les participes passif (– مضروبfrappé), les noms de lieux ou de temps (مَضْرِبُ_lieu de frappe), le nom d'instrument (مِضْرَبْ-ِraquette), le nom d'une fois (– ضربةune frappe), etc.[5]

C- Les nombres : ce sont les numéros simples représentant les unités (de '– 'صِفْرsifr : zéro- à '– 'تِسْعَةtis'at : neuf-), les dizaines ('– عِشْرُونishruwn : vingt-) et les centaines (' 'مئة : -cent-), etc ; et les numéros composés comme les cardinaux, par exemple ' ستة عشر. Seize ' – [5]

I.2.4.2 La morphologie verbale

Le système morphologique des verbes arabes est un système très régulier ; ce système compte un nombre limité de patrons (modèles) : dix modèles trilittéraux et deux modèles quadrilatéraux.[26]

Les verbes se fléchissent en aspect, mode, voix et sujet :

- L'aspect a trois valeurs : accompli (ماضيmADiy), non accompli (مضارعmuDAriÇ) et impératif (أمرAamr) ;
- Le mode a trois valeurs : indicatif (مرفوعmarfuw''), subjonctif (منصوبmanSuwb) et jussif (مجزومmajzuwm) ;
- La voix à deux valeurs : passif et actif ;
- Le sujet possède trois traits morphologiques :

  1.le trait personne qui a trois valeurs : (1er personne, متكلمmutakal~im),

  (2ème personne, مخاطبmuxATab) et (3ème personne,غائب ghaAyib) ;

  2. le genre ayant deux valeurs : masculin et féminin ;

  3. le nombre dispose de trois valeurs : singulier, duel et pluriel.[6]

I.2.5 Specificités de la langue arabe

L'arabe, comme toutes les langues naturelles, est caractérisée par un ensemble de phénomènes créant des difficultés et des problèmes qu'il faut prendre en considération. Cette caractéristique introduit, de fait, une forte ambiguïté avec laquelle il va falloir dans le cadre du traitement automatique de la langue.

I.2.5.1 voyellation

Contrairement à la langue française, les voyelles dans la langue arabe ne sont pas des lettres, mais des signes qui s'écrivent au-dessus ou au-dessous des lettres (consonnes) et qui remplissent la fonction de voyelle.

A titre d'exemple, prenons-le mot كتب/ktb et comptabilisons ses diverses voyellation [Debili, 2002] : « كَتَبَ/ kataba » (Il a écrit)

« كُتِب/ kutiba » (Il a été écrit)

« كُتُبَ/ kutub » (des livres)

« كَتْب/ katb » (un écrit)

[7]

I.2.5.2 L'agglutination

Contrairement aux langues latines, l'arabe est une langue agglutinante. Les articles, les prépositions et les pronoms collent aux adjectifs, noms, verbes ; ce qui nécessite de procéder au découpage des mots avant la tâche de lemmatisation. La plupart des mots arabes sont composés par l'agglutination d'éléments lexicaux élémentaires. 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

La forme agglutinée correspond à une suite de formes « collées ».[7]

I.2.5.3 L'écriture de droite à gauche

Traiter automatiquement la langue arabe suppose d'utiliser un éditeur de texte capable de soutenir l'écriture de droite à gauche et surtout, de combiner les différents sens d'écriture (par exemple, du français ou de l'anglais avec de l'arabe). Il est en effet courant de trouver dans les textes arabes, des noms écrits en langue latine avec leur acronyme et, à côté, de trouver la traduction du nom de la société en arabe.[7]

I.2.5.4 La détermination

Certains constituants du domaine du sport sont toujours déterminés (le 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 : nom de ville qui suit directement la catégorie comme par exemple l'EN ملعب صفاكس (stade Sfax) où le toponyme est non déterminé et (ملعب الرياض stade du Ryadh) où le toponyme est déterminé.[1]

I.2.5.5 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 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. Par exemple, le mot stade en arabe peut s'écrire stade ,ستادistaadاستاد

En outre, nous constatons d'une part, les ambiguïtés des déclencheurs. Par exemple, le mot stade peut être un déclencheur pour des noms de stades commeملعب الطيب المهيري (stade de taeib el mhiri).[1]

I.2.6 Catégorisation d'ENA

La catégorisation est une étape visant déterminer la catégorie adéquate qui décrit convenablement une ENA.

I.2.6.1 Catégorie Nom de personne

La catégorie Nom de personne est dédiée à représenter les différentes formes décrivant un nom de personne arabe. Un nom de personne arabe contient cinq parties ne suivant aucun ordre particulier : al-ism, al-kuniyah, al-nasab, allaqab[27]. La combinaison de ces quatre parties permet de construire un nom de personne quand elles sont regroupées au sein de la même ENA.

I. Al-ism

Al-ism est ce qu'on appelle le prénom. Ce prénom est le premier nom donné à une personne lors de sa naissance. Le prénom peut être masculin comme «عبد الله /Abdullah ; عادل/Adel ; حسين/Hussein » ou féminin comme «فاطمة /Fatma ».[10]

II. Al-kuniyah

Al-kuniyah est un élément utilisé comme une forme informelle pour s'adresser à quelqu'un par respect comme l'utilisation de « oncle » ou « tante ». Elle indique aussi que quelqu'un est le père ou la mère d'une personne particulière. Par exemple, «أم كلثوم /Om Kalthoum » signifie la mère de kalthoum.[10]

III. Al-nasab

Al-nasab est un élément décrivant un nom patronymique qui commence par un lien comme « بن/bin » ou «بنت/bint » signifiant respectivement « le fils de » ou « la fille de ». Cet élément suit directement Al-ism (prénom) comme « فهد بن عبد العزيز/ Fahad ibn Abdul Aziz » qui désigne « Fahd le fils de Abdul-Aziz ». Il faut mentionner que l'existence de lien n'est pas toujours obligatoire.[10]

IIII. Al-laqab

Al-laqab est définie comme une épithète qui est généralement religieuse ou descriptive. Prenons les deux exemples suivants : le mot «الرشيد / Al-Rashid » signifie « le bien guidé » et le mot « الفضل/ Al-fadl » signifie « le proéminent ».[10]

IIIII. Les formes de noms de personnes identifiées

En fait, nous identifions 24 formes décrivant les formes alternatives d'un nom de personnes. Les ENA suivantes sont des exemples illustrant quelques formes identifiées durant notre étude linguistique. Nous remarquons que l'ENA se compose au moins d'un parmi les quatre éléments déjà expliqués.

(18) الأستاذ فؤاد بك العادلي

Le Professeur Foued Bek Al-Adeli

(20) عبد الفتاح أبو غدة

Abd Al-fattah Abu Ghoda

Dans l'exemple (18), l'ENA ayant la catégorie nom de personne est précédé par un indicateur externe exprimant une profession. Cette ENA contient un autre indicateur interne « بك/ bek » situé entre le prénom et le nom de la famille. L'exemple dans (20), décrivant la catégorie nom de personne prend la forme ordinaire : un prénom suivi par un laqab ayant la forme de kuniyah.[10]

I.2.6.2 Catégorie Nom de lieu

Les noms de lieux en arabe, comme dans d'autres langues, désignent les villes, les pays, les villages, les montagnes et les fleuves. Il existe trois sous-catégories, appartenant à la catégorie Nom de lieu, qui apparaissent fréquemment qui sont les suivantes : Nom de lieu absolu, relatif et géographique. Dans notre corpus, cette catégorie inclut tout ce qui représente un lieu sportif tel que les noms de stade, de piscines, de salles de sport, de cité. La liste des noms de lieux connus et existants dans le monde est relativement stable dans la mesure où les noms de lieux ne changent pas souvent. Toutefois, à l'instar des noms de personnes, certains noms de lieux sont ambigus. Par exemple, le mot Tunisie en arabe تونسTunis ou Algérie الجزائر aljaZaa'ir désigne le pays ou la capitale. De plus, le mot Maroc en arabe المغرب désigne soit le pays qui est le Maroc soit la grande région du Maghreb située en Afrique du Nord. Parfois et afin de lever l'ambiguïté, on appelle le Maroc المغرب الأقصىalmaghrab al'aqsaa qui signifie le (Maroc lointain) et المغرب الكبير almaghrab alkabyr pour désigner la région entière du Maghreb. [1]

Tableau 1. Sous-catégories associées à un nom de lieu relatif [10]

Sous-catégorie	Exemple
Musée	المتحف المسيحي المبكر بقرطاج
	Musée Paléo-chrétien de Carthage
Palais	قلعة السلع
	Le palais de Sela
Mosquée	الجامع العمري
	La mosquée d'Omari

Eglise	الكنيسة المارونية
	L’église maronite
Bain publique	حمامات عفرا المعدنية
	Les chutes minérales d’Ofra
Salle	قاعة المؤتمرات
	Salle de conférences
Hôtel	فندق الرويال
	L’hôtel royal
Marché	سيتي مول
	Centre commercial
Stade	ملعب إستاد السّلام الرياضي
	Le stade sportif de Al-salam
Aéroport	مطار عمان المدني
	Aéroport civil d’Amman
Théâtre	المسرح الدولي بالجزائر
	Le Théâtre international algérien
Hôpital	مستشفى الملكة علياء
	L’hôpital de la reine Alia
Jardin	حديقة العامرات الطبيعية
	Le jardin naturel Amrat
Monument	سارية العلم الأردني
	La hampe de drapeau Jordanien
Tombe	كعب بن عمير الغفاري الصحابي ضريح
	Le tomb de sahabi Kaab bin Amir Alghafari
Rue	شارع الثقافة
	Rue de la culture

I.2.6.3 Catégorie Organisation

Dans notre corpus d’étude, nous rencontrons de nombreuses formes d’ENA décrivant la catégorie Organisation. Les noms d’organisation identifiés possèdent les natures suivantes : Entreprise, Administration, Média, Organisation culturelle ou politique ou éducative. Il est vrai qu’il existe des organisations qui peuvent apparaître sous la forme d’acronymes. Néanmoins, l’utilisation des acronymes est relativement rare dans notre corpus d’étude.[10]

Figure 4. Sous-catégories associées à un nom d'organisation [10]

I.2.6.4 Catégorie Evènement

Un événement est une composition nominale qui peut avoir différentes formes. Cette catégorie peut également être composée de sous-catégories. En se basant sur notre corpus d'étude, nous constatons qu'un événement peut être imbriqué dans un nom de lieu, une date ou un nom de personne. En fait, nous avons identifié 3 sous-catégories d'un évènement qui sont : événement politique Prenons l'exemple de l'ENA «سليم أبو سجن مجزرة/ le massacre de la prison d'Abu Salim », celle-ci décrit un événement politique contenant un nom de lieu relatif «سجن أبو سليم/la prison d'Abu Salim » dont il inclut à son tour un nom de personne «أبو سليم/ AbuSalim»,Évènement culturel par exemple Dans notre corpus, un événement culturel peut être avoir dans sa composition un événement religieux tel que «مهرجان عيد الفطر/ festival d'Eid al-Fitr » sachant que l'ENA «عيد الفطر/ Eid al-Fitr » est une fête religieuse en l'islam et évènement religieux comme «المولد النبوي الشريف/le Mawlid».[10]

I.2.6.5 Catégorie Date

La catégorie Date qui décrit une ENA fait partie des expressions numériques. Les expressions de temps incluent les dates, la période et toute autre expression exprimant le temps. La plupart des entités temporelles dans la langue arabe sont identifiables grâce à une liste de marqueurs lexicaux comme par exemple *jour*, *mois*, *année*, etc. Concernant les dates, il existe une différence dans l'usage des calendriers qui varient d'un pays arabe à un autre tels que le calendrier grégorien (ex., *01 نوفمبر 2007*), le calendrier syriaque (ex., 02 تشرين الثاني2007) et le calendrier musulman (ex., 21 شوال من 1428 هـ).

I.2.7 Les problèmes d'analyse du traitement automatique de la langue arabe

L'arabe, comme toutes les langues naturelles, est caractérisée par un ensemble de phénomènes créant des difficultés et des problèmes qu'il faut prendre en considération lors d'un traitement

automatique. Dans la présente section, nous présentons les phénomènes que nous considérons les plus importants pour l'arabe.

I.2.7.1 L'absence de voyelle – voyellation –

L'absence des voyelles courtes, appelées aussi les diacritiques, dans les textes en arabe. Cette absence génère plusieurs cas d'ambiguïté compliquant ainsi le traitement automatique. Ces ambiguïtés lexicales sont dues essentiellement au fait que chaque consonne peut prendre l'une des sept voyelles de l'arabe, ce qui crée des combinaisons de mots dont le nombre diffère d'un mot non voyelles à un autre en fonction de l'existence de la combinaison obtenue dans le vocabulaire ou pas. Selon [28], l'absence de diacritiques en arabe entraîne une complexité de calcul d'un ordre de grandeur plus grand que la manipulation de ses homologues langues latines. Ce problème est d'autant plus complexe qu'un mot en arabe peut avoir différentes prononciations sans aucun effet orthographique en l'absence de diacritiques comme dans l'exemple suivant :[5]

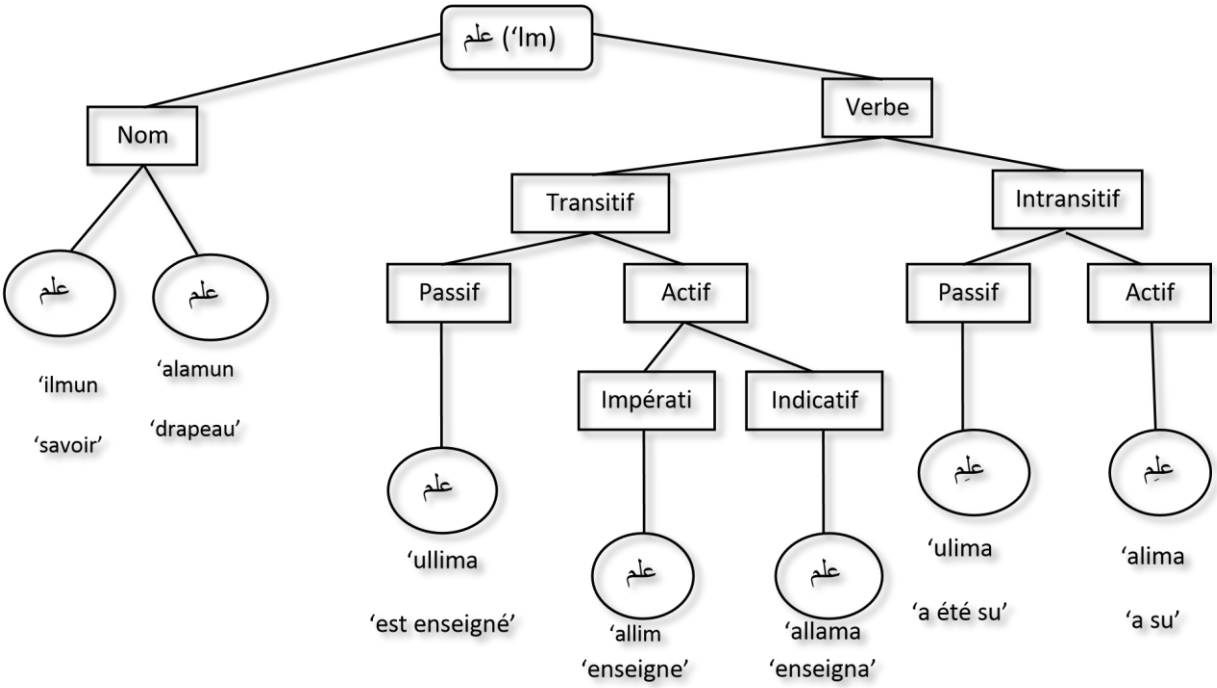

Figure 5. Ambiguïté causée par le manque de diacritiques [23]

I.2.7.2 Les proclitiques

Les proclitiques permettent de donner des traits syntaxiques (coordonnant, déterminant …) pouvant accompagner un mot arabe. Leur nombre est fini et peuvent se combiner entre eux pour être utilisés comme préfixes rattachés au 'mot minimal' ou détachés comme c'est le cas des conjonctions de coordination. Lorsqu'un proclitique est rattaché à un verbe il dépend exclusivement de son aspect verbal, ainsi ils prennent tous les pronoms et par conséquent ils sont compatibles avec tous les préfixes pris par l'aspect. Nous pouvons répartir les proclitiques dans les catégories suivantes :

- Les proclitiques réservés aux noms et adjectifs :
  - ❖ L'article défini 'ال'(al- le)
  - ❖ L a préposition ب (bi - avec), ل(li - pour), ك(ka - comme)
- Les proclitiques réservés aux verbes :
  - ❖ La particule du subjonctif : nasbل (li - pour)
  - ❖ La particule du futur س (sa)
- Les proclitiques généraux utilisés indépendamment de la catégorie des mots auxquels ils s'attachent :
  - ❖ Les conjonctions de coordination ف(fa - et), et و(Wa - et)
  - ❖ L'article d'interrogation أَ(a - est ce que)

Cette classification n'omet pas le fait qu'il existe certaines exceptions de proclitiques qui peuvent jouer différents rôles, comme pour le proclitique 'و' (wa) utilisé généralement comme particule de liaison (conjonction de subordination et de coordination), mais également peut être utilisé comme particule d'accompagnement (واو المعية) ou de serment (واو القسم).[5]

I.2.7.3 Les enclitiques

Les enclitiques présentent les pronoms suffixes qui s'attachent toujours à la fin du mot graphique, leur liste est constituée des 17 éléments suivants : ـِ هِما هِم هِنَّ ني ي نا كُمَا كُمْ هُ هَا هُما هُمْ هُنَ هِUn mot graphique ne contient qu'un seul enclitique à la fois. Ils s'attachent aux verbes comme étant un complément-objet et aux noms et aux prépositions comme un complément du nom ou complément d'objet indirect. Leurs utilisations est régie par certaines restrictions.

Par ailleurs, certains mots nécessitent des transformations morphologiques avant de leur rattacher des enclitiques, c'est le cas des noms se terminant par une hamza, une "ى" ou une "ي" (y). Par exemple la forme مَلْهى (malha - un manège), nécessite une transformation de celle-ci en "ا "avant sa suffixation pour produire la forme agglutinée, ملهاهُ (malhAhu - son manège).[5]

I.2.7.4 La variation de l'ordre des mots dans la phrase arabe

Notons aussi qu'un mot déclencheur d'une relation peut se trouver à des positions différentes par rapport aux EN dans la phrase. Ceci est dû au fait que l'ordre des mots dans une phrase arabe est relativement libre (VSO, SVO et OVS). Si nous prenons par exemple le cas d'une phrase composée d'un verbe, sujet et complément, nous constatons que les positions relatives de ces constituants peuvent varier sans trop affecter le sens global :

❖ VSO :
سافر أحمد إلى مصر\saAfara ÂH. Madu Ălý miS.ra\Ahmed s'est rendu en Egypte.

❖ SVO :
أحمد سافر إلى مصر \ÂH. Madu saAfara Ălý miS.ra\Ahmed s'est rendu en Egypte.

❖ OVS :
أحمد سافر إلىمصر\ Ălý miS.ra saAfara ÂH. Madu\ en Egypte, Ahmed s'est rendu.
Dans ces exemples le mot déclencheur سافر\saAfara\se rendre est situé avant la première EN dans le premier exemple, entre les deux EN dans les deux autres exemples.[2]

I.2.7.5 Le manque de ponctuation dans les textes arabes

Un degré supplémentaire de difficulté est imposé à l'extraction des relations en raison d'un manque de ponctuations régulières associé à des phrases longues (i.e., renfermant plusieurs propositions reliées). Cela génère une ambigüité au niveau de l'identification des relations entre EN éloignées comme le montre l'exemple ci-dessous extrait de [29]. Ce problème peut être atténué par un traitement préalable de segmentation de la phrase en prépositions.

وقف ما ذكرته وكالة الأناضول للأنباء أن السيد أحمد داود اغلو وزير الخارجية الجمهورية التركية التقى مع السيد فلاديمير ماكاي وزير الخارجية روسيا البيضاء في استنبول\ mA ðkrthwkAlĥAlÂnAĎwlllÂnbA' ÂnAlsydÂHmddAwd Âγlwzyr xArjyĥAljmhwryĥAltrkyĥAltqýmçAlsydflAdymyrmAkAywzyrxArjyĥrwsyAAlbyDA' fy AsTnbwl\. [2]

I. Conclusion

La reconnaissance des entités nommées reste encore un intérêt de recherche car elle peut se baser divers critères d'identification d'une EN en cherchant leur portée sémantique.

Dans ce chapitre, nous avons présenté la définition de l'EN que nous avons retenue. Ensuite, nous avons effectué une étude sur l'EN arabe, les Catégorisation d'une entité nommée et les trois approches de REN. Nous avons mis en avant les différents phénomènes linguistiques liés au repérage des EN et à leur traduction. Le chapitre suivant sera consacré à la description des deux ressources principales de ce travail de recherche : la base de données lexicale multilangues Prolexbase et l'encyclopédie Wikipédia et la Wikimédia.

Chapitre2

Les ressources utilisées

II. Introduction

Le traitement automatique des langues (désormais TAL) est fortement dépendant des ressources linguistiques adaptées aux applications envisagées et aux méthodes utilisées. Certains composants logiciels n'utilisent que des corpus d'apprentissage pour permettre ensuite la mise en œuvre de traitements statistiques. Mais un grand nombre d'applications possèdent une composante linguistique qui nécessite une liste de mots avec des codes morphologiques, syntaxiques ou autres, autrement dit un dictionnaire électronique. Prolexbase est un dictionnaire électronique multilingue relationnel spécifique aux noms propres (constituant 10 % des textes journalistiques), disponible librement sur le site Web CNRTL, au format LMF (ISO 24613) depuis les travaux de [35].

II.1 Prolexbase

II.1.1 Définition de Prolexbase

Prolexbase est un dictionnaire multilingue. Pour une langue donnée et une entrée linguistique de Prolexbase est un prolexème. Chaque prolexème d'une langue est relié à un et un seul pivot interlangue qui est un identificateur unique. C'est par ce pivot que passe la traduction d'une langue à l'autre. Le prolexème, qui est une famille structurée de lexèmes. Autour d'eux, sont définis d'autres concepts et des relations (synonymie, méronymie, accessibilité, éponymie, etc.). Chaque pivot est en relation d'hyperonymie avec un type et un paradigme d'existence.

Prolexbase est une base de données lexicale qui contient toutes les informations syntaxiques, morphologiques et sémantiques concernant les noms propres. Ce type de ressource est très utile et efficace dans plusieurs applications de TAL, telles que la recherche d'informations interlangues, la traduction automatique, l'alignement des textes multilingues, ect. [6]

II.1.2 L'ontologie de Prolexbase

La gestion cohérente d'une base de données passe par la définition d'un modèle d'organisation des noms propres, autrement dit d'une ontologie. A partir d'une étude approfondie dans le domaine des noms propres, notamment à travers les différents mécanismes morphologiques et dérivationnels qui peuvent s'appliquer aux noms propres dans différentes langues et sans oublier les relations qui lient les noms propres entre eux, nous avons établi une ontologie hiérarchisée en quatre niveaux: le niveau des instances, le niveau linguistique, le niveau conceptuel (pivot), et le niveau méta conceptuel[8].La figure 1.1 représente l'architecture générale de Prolexbase méta conceptuel[8].La figure 1.1 représente l'architecture générale de Prolexbase

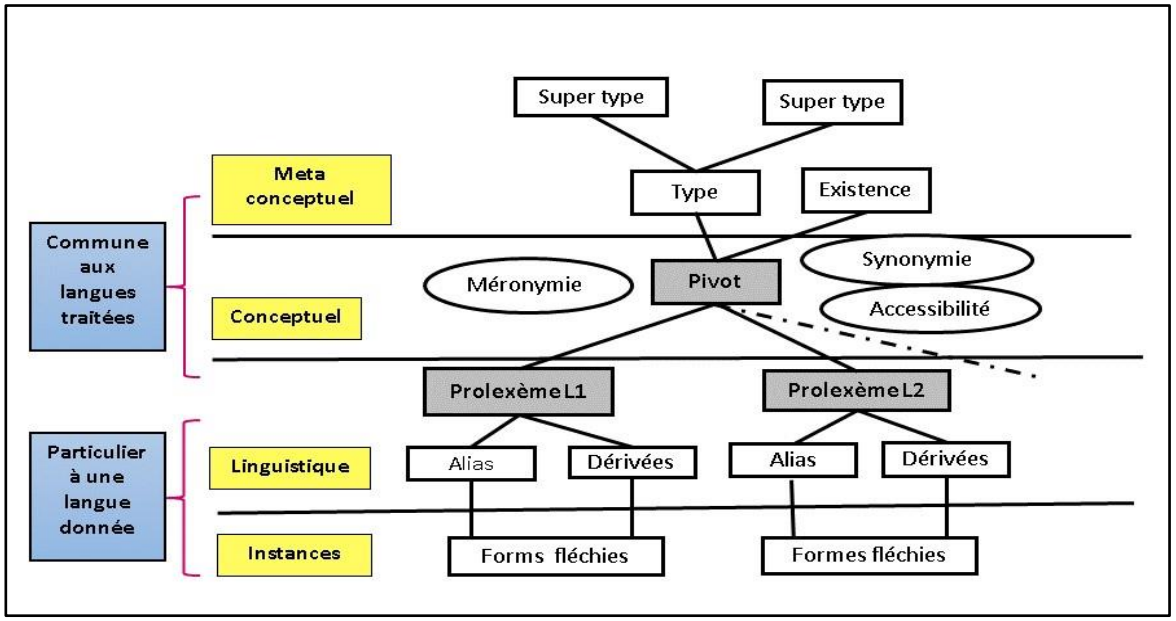

Figure 6 : L'architecture générale de l'ontologie des noms propres (Prolexbase) [8]

II.1.2.1 Le niveau des instances

Les instances correspondent à l'ensemble des formes fléchies qu'un prolexème peut générer dans une langue traitée ; le niveau d'instances contient des ensembles de toutes les formes réelles de prolexèmes et de leurs alias et dérivés.

Dans ce niveau, le nombre de formes fléchies dépend complètement de la morphologie de la langue cible ; en particulier, la langue anglaise et la langue française sont moins importantes au niveau morphologique, en comparaison avec les langues slaves qui sont morphologiquement plus riches comme le polonais et le serbe.

Pour illustrer quelques différences entre les langues, considérons le pivot représentant Italie qui en anglais possède 5 instances (Italy, Italian, Italians, Italian, Italo), en français 10 : (Italie, Italien, Italiens, Italienne, Italiennes, italien, italiens, italienne, italiennes, italo) et en polonais 70 instances. [6]

II.1.2.2 Le niveau linguistique

Le niveau linguistique se décompose en trois sous-niveaux :

Le premier niveau dépendant de la langue contient les prolexèmes qui sont les formes canoniques (lemmes) représentant les noms propres lexicaux dans une langue donnée. Plus précisément, un prolexème peut être considéré comme un lemme de l'ensemble de différentes formes d'apparition d'un nom propre dans un texte.

Il s'agit de la projection du nom propre conceptuel (le pivot) dans une langue donnée où chaque prolexème est relié à un seul pivot ; notons que cette relation entre ces deux concepts (pivot

prolexème) est utilisée pour la traduction d'un prolexème d'une langue vers une autre langue dans Prolexbase. [6]

Par exemple, le prolexème français Platon, le prolexème anglais Plato, le prolexème polonais Platon et le prolexème arabe أفلاطون seront reliés au même nom propre conceptuel (le pivot (52307). La Figure 7 montre le pivot et les prolexèmes de nom propre Platon dans les quatre langues.

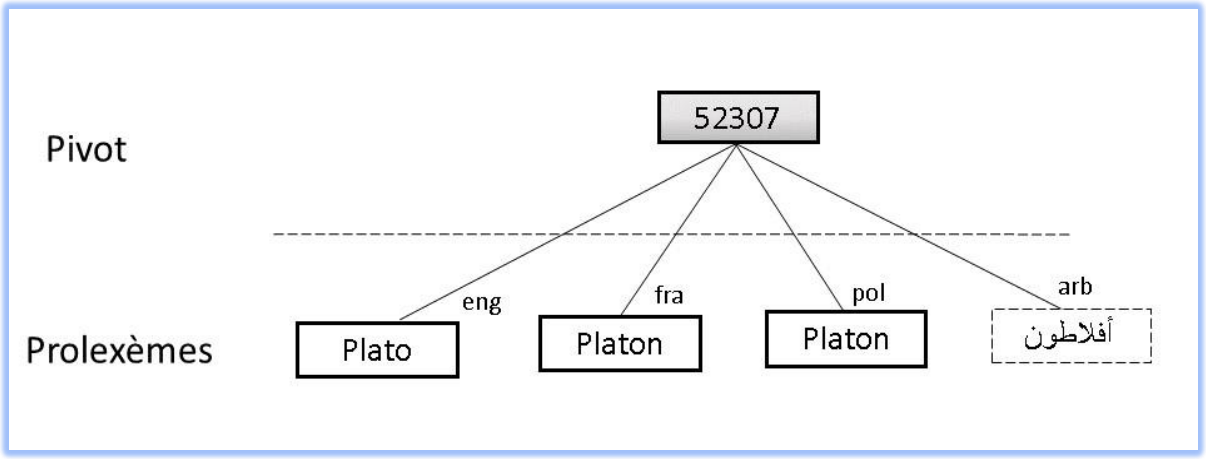

Figure 7 : Pivot et prolexèmes du nom propre Platon [6]

Dans ce même niveau, les prolexèmes peuvent avoir des dérivés et des alias dépendants de la langue ;[36] ont défini les alias comme des synonymes qui dépendent de la langue regroupant d'une part des synonymes exacts, les variantes d´écriture (caractères, abréviations, acronymes et sigles, transcriptions) et d'autre part des synonymes approximatifs, diatopiques ou diastratiques.), les dérivés sont obtenus par dérivation morphosémantique ; ils comprennent les adjectifs relationnels et les noms relationnels ; par exemple, Parisien et Parigot sont les dérivés du nom propre Paris. [6]

II.1.2.3 Le niveau conceptuel

Le cœur de l'ontologie reste le niveau conceptuel à travers la notion de pivot. Les pivots correspondent à des numéros d'identité uniques (ID) qui sont associés à chaque prolexème de chaque langue, tels qu'ils sont définis au niveau linguistique. Le niveau conceptuel est un niveau interlangue. Ainsi, chaque nom propre représentant le même concept possèdera le même pivot.

Un nom propre n'est la traduction d'un autre nom propre dans une autre langue que si tous deux partagent le même pivot. Il existe trois relations sémantiques qui ne dépendent pas de la langue et sont associées aux pivots dans ce niveau : Synonymie, Méronymie et Accessibilité. Illustrons cela par un exemple, le nom propre Paris qui possède le pivot unique 38558 est en relation de synonymie avec le pivot 55120 dont le prolexème français est Ville de Lumière ; parallèlement, il est en relation de méronymie avec le pivot 5, qui réfère au nom propre Île de France (Paris fait partie de la région Île de France) ; Encore, Paris est la capitale

de la France, alors, il est en relation d'accessibilité avec le nom propre France portant le pivot 27. La figure8 reprend ces trois relations entourant le pivot 38558. [6]

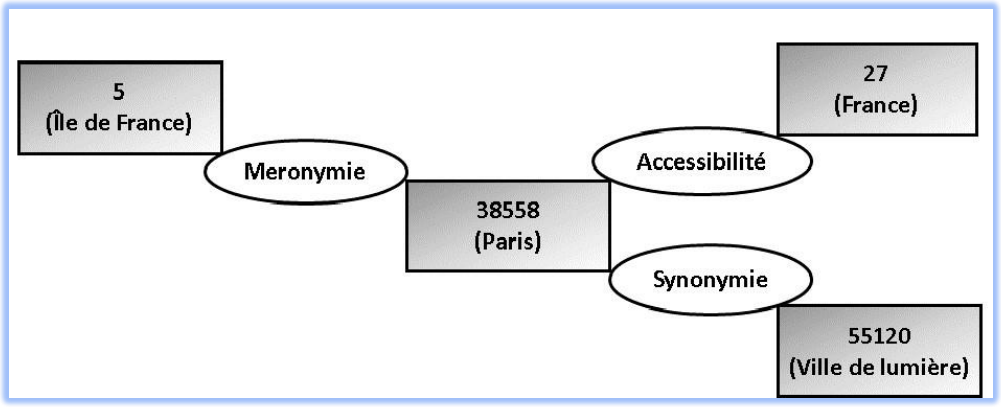

Figure 8 : Les relations Synonymie, Accessibilité, Méronymie entourant le pivot 38558 (Paris) [6]

II.1.2.4 Le niveau méta-conceptuel

Ce niveau permet d'avoir une classification homogène des noms propres sur la base des super types et des types qui sont associés à chaque nom propre.

Nous distinguons, ci-dessous, quatre super types qui se réfèrent aux caractéristiques sémantiques primaires telles que l'humain, le lieu, le concret et l'événement. Trente types ont été définis pour l'organisation de la structure de Prolexbase. Les super types nous fournissent des informations de base sur les noms propres, tandis que les types sont plus précis et fournissent une classification plus fine.

1. Les toponymes (trait locatif) comprennent tous les noms de lieux au sens général en les rassemblant (pays, région, supranational. ..ect), En particulier, le type région est indiqué pour une subdivision d'un pays, comme les régions, les provinces, les départements et les voïvodies (*ex : Cambridgeshire*).
2. Les anthroponymes (trait humain) sont partagés en deux autres super types : les anthroponymes individuels (célébrité, prénom, patronyme, pseudo-anthroponyme), et les anthroponymes collectifs (dynastie, ethnonyme, association, ensemble, entreprise, institution et organisation) ;
3. Les ergonymes (trait inanimé) conçoivent l'objet, le produit, la pensée (*ex : catholicisme, marxisme*), le vaisseau (*Titanic*) et les œuvres ;
4. Les pragmonymes (trait événement) incluent les types désastre, manifestation, fête, histoire et météorologie.

Un super type est hyperonyme de plusieurs types, chaque type est lié à un seul super type. Tout nom propre est associé à un seul type, sinon, ils sont considérés comme homonymes et attribués à des pivots différents. Par exemple, le nom propre Washington est considéré comme un

toponyme (ville), un anthroponyme (célébrité) et à nouveau, comme un toponyme (région), ces trois homonymes obtiennent trois pivots différents. [6]

II.2- L'encyclopédie Wikipédia

Wikipédia est une encyclopédie multilingue créée par Jimmy Wales et Larry Sanger le 15 janvier 2001. Cette encyclopédie apporte des liens explicatifs et offre un contenu librement réutilisable, objectif et vérifiable. Elle fournit aussi l'accessibilité et la reconnaissance automatique des sujets mentionnés dans des textes non structurés à travers des liens.[10]

«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 »[12], et elle fonctionne sur le principe de wiki c'est-à-dire une application web permettant la modification des pages web écrites en utilisant un langage de balisage par ses visiteurs via un navigateur web. [6]

II.2.1- La structure générale d'une page Wikipédia

La Wikipédia se compose de textes écrits en langage naturel, d'images et aussi d'autres informations structurées et de plusieurs types de liens, elle contient des différentes versions linguistiques avec généralement une structure quasi identique.

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

❖ LES Infoboxes

Une **infobox** (ou infoboîte) est une table de données présentant sommairement des informations importantes sur un sujet. Elle prend la forme d'un cartouche ou d'un encadré, placé en général en haut à droite de l'article. Une infobox doit rester synthétique quant à son contenu. Elle ne doit pas remplacer l'article mais le compléter, fournir une vue globale, par l'apport d'informations générales.[11] la Figure 9 représenteune partie de la pagel'émir Abdelkader dans l'édition Wikipédia Arabe en entourant les liens Historique,

Discussion, Pages liées, Informations sur la page, l'Infobox et un lien interne.

FIGURE 9 : Une partie de la page l'émir Abdelkader (version Wikipédia arabe) comprenant info box, discussion, historique et lien interne

❖ L 'historique

Il désigne un lien nommé « Historique » dans la version française, placé en haut à droite, près du moteur de recherche ; 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 : statistiques, Recherche, Statistiques de consultation, Contributeurs suivant et Modifications par utilisateur. [6] elle est représentée dans la figure 9.

❖ 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. [6]

❖ La discussion

Il existe un lien appelé « Discussion » (en français) en haut à gauche de la page, 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. [6] comme l'exemple de la figure 9.

❖ Pages 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. [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. [6]

❖ Les liens interlangues

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. [6]

❖ Liens internes

Les liens internes à Wikipédia ou wikiliens sont des liens qui pointent vers un autre article de Wikipédia. Leur utilisation peut parfois pécher dans leur pertinence, leur efficacité ou leur esthétique. Les liens internes connexes à un article sont regroupés en fin d’article dans une sous-rubrique Articles connexes de la rubrique Voir aussi. Dans le cas où la rubrique Voir aussi ne présenterait pas de liens externes, on admettrait qu’elle soit utilisée pour les articles connexes. Un même lien répété plusieurs fois est inutile puisque, pour appréhender l’ensemble du sujet, un lecteur lira la section recherchée ou l’article en entier plutôt qu’une phrase précise et aura ainsi autant de chances de trouver les liens adéquats qui lui permettront d'approfondir certains points précis.[4] regarde la figure 9.

❖ *Liens externes*

Ce sont des hyperliens qui mènent vers d’autres 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]la figure 10 représente la forme des liens externes se trouvant dans la page « l’émir Abdelkader ».

Figure 10 : Les liens externes de la page l'émir Abdelkader

❖ Références

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.[15]
la figure 11 représente la référence de la page El Amire Abd Elkader.

FIGURE 11 : les référence dans l'émir Abdelkader de la version wikipidia arabe

II.2.2-Wikimédia

Wikimédia est un mouvement mondial dont la mission est d'apporter du contenu éducatif gratuit au monde. À travers divers projets, chapitres et la structure de soutien de la Fondation Wikimédia à but non lucratif, Wikimédia s'efforce de créer un monde dans lequel chaque être humain peut partager librement la somme de toutes les connaissances.[16]

II.2.2.1- Les chapitres de Wikimédia

Les chapitres de Wikimédia sont des organisations indépendantes fondées pour soutenir et promouvoir les projets Wikimédia dans une région géographique spécifique (dans la plupart des cas, un pays). Comme la Wikimédia Fondation, ils visent à "habiliter et engager les gens du monde entier à collecter et développer du contenu éducatif sous licence libre ou dans le domaine public, et à le diffuser efficacement et dans le monde entier". Il existe actuellement 38 chapitres, dont au moins un sur chaque continent habité.[17]

II.2.3-L'accès au contenu de l'encyclopédie Wikipédia

La fondation Wikimédia fournit plusieurs outils de recherche et de traitement de données constituant les pages Wikipédia. Trois approches peuvent être exploitées pour accéder au contenu de l'encyclopédie, nous avons situé deux :

II.2.3.1. DBPEDIA

DBpedia est un effort communautaire qui a démarré en 2007. Il vise à extraire des informations structurées de Wikipédia et à les rendre disponibles sur Internet. Le contenu extrait depuis l'encyclopédie est converti dans le format RDF. Plusieurs mécanismes d'accès sont proposés pour explorer DBpedia : l'accès aux données RDF directement par URI (*Universale Resource Identifier*), l'utilisation d'agents Web (exemple : navigateurs pour le Web sémantique) et les points d'accès SPARQL permettant l'interrogation de DBpedia au moyen d'un langage évoquant le SQL utilisé pour les bases de données relationnelles.[18]

II.2.3.2. Les dumps

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.[6]

En d'autres termes, Wikimédia fournit des vidages publics du contenu de nos wikis et des données connexes telles que les index de recherche et les mappages d'URL courts. Les dumps sont utilisés par les chercheurs et dans les projets de lecture hors ligne, pour l'archivage, pour l'édition de bot des wikis et pour la mise à disposition des données dans un format facilement interrogeable, entre autres. Les dumps peuvent être téléchargés et réutilisés gratuitement. [19]

II.2.4 Wikipédia en Arabe

Wikipédia en arabe (ويكيبيديا العربية Wīkībīdyā al-ʿArabiyya ou ويكيبيديا الموسوعة الحرة Wīkībīdyā, al-Mawsūʿa al-Ḥurra) est l'édition de Wikipédia en langue arabe, langue sémitique parlée dans le monde arabe. L'édition est lancée le 9 juillet 2003. Son code est : ar.

La Wikipédia possède un volume arabe très riche en termes d'EN arabes (ENA) mais il reste encore les articles arabes 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.

II. Conclusion

Dans ce chapitre Nous avons présenté les deux ressources principales de ce travail de recherche : la base de données lexicale multilangues Prolexbase et l'encyclopédie Wikipédia. Nous avons défini

L'ontologie de Prolexbase en discriminant les deux concepts essentiels : le concept multilingue (pivot) et le concept lexical (prolexème). Pour la Wikipédia, nous avons décri la structure d'une page donnée, en particulier les liens internes, les liens externes, le lien nommé et les liens Interlangues, Nous avons expliqué les différentes méthodes d'accès au contenu d'un article de la Wikipédia.

Finalement, Nous avons définie Wikipédia arabe.

Chapitre3

conception

III. Introduction

Ce chapitre vise à mettre en évidence la conception de notre travail, qui est l'ajout de la langue arabe dans Prolexbase. Nous présentons par la suite l'enrichissement de la base pour les deux autres langues les plus couverte sur prolexbase qui sont l'anglais et le français.

Premièrement nous procéderons à la récupération des entités nommées de la wikipedia pour cela on extrait les URLs du corpus ; ensuite nous calculons la notoriété pour les entrées du dictionnaire relationnel multilingue de noms propres Prolexbase. Cette notoriété servira au choix de la pertinence de l'ajout de l'entité en langue arabe sur prolexbase. Apres l'ajout de l'entité nous recalculons la notoriété pour les noms propre déjà existant dans les deux autres langues étudiées. Ses modifications nécessite des traitements des ressources textuelles sur wikipedia pour cela nous utilisons le logiciel Unitex.

III.1extraire les données d'un nom propre

Nous avons extrait les données d'un nom propre sous forme d'url dans un fichier.txt Dans la figure 12 nous avons scraper les données de la page html d'un nom propre spécifique (عبد القادر الجزائري) de l'encyclopédie wikipedia est l'avons enregistré dans un fichier.txt.

L'url va être traité en utilisant Unitex, pour ne sauvegarder que la partie concernant le nom propre lui-même.

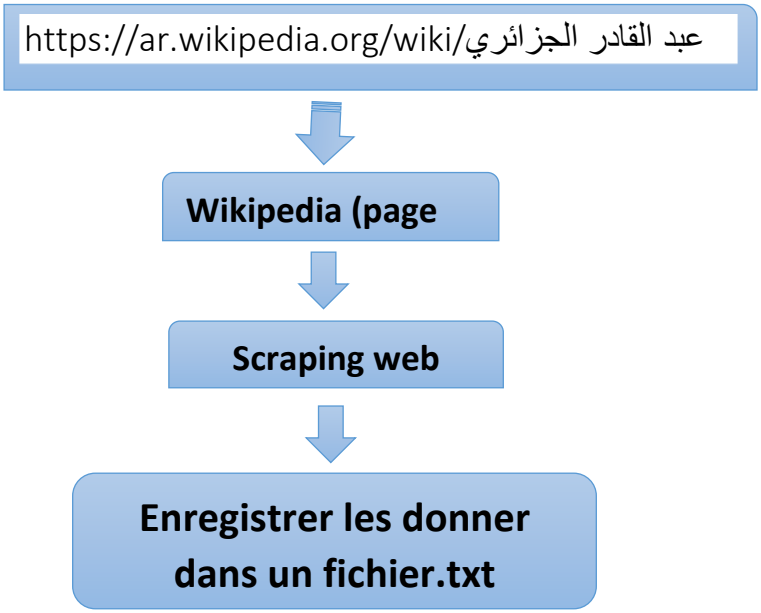

Figure 12 : les étapes pour extraire les données de nom propre

III.2 le Calcul de la notoriété

Afin de pouvoir enrichir prolexbase, nous devons choisir que les noms propres pertinents ; pour mesurer cette pertinence nous calculons une mesure de notoriété [6] de chaque nom propre sélectionnés à partir de la wikipedia. Les étapes pour calculer la notoriété de chaque nom propre , il existe deux partie pour calculer la notorité :

1. Le calcul des cinq indices :
   Les cinq indices sont :

- Le nombre de consultations de l'article ⇨ Pour calculer cet indice nous avons suivi la méthode de Mouna [6], nous avons calculé la moyenne de coefficient d'oubli du chaque mois dans les années [2016,2022].

- Le nombre de contributeurs à l'article
- La taille de l'article
- Le nombre de liens internes à la Wikipédia pointant vers l'article
- Le nombre de liens externes à la Wikipédia contenus dans l'article

Pour calcule la notoriété de ces indices Il existe une url pour chacun afin d'extraire les données.

Url de chaque critère d'abord connecter dans une api ensuit Engistrer les donner format. Json enfin non affichons le résultat.

2. Le calcul de la notoriété :
   Nous avons donc obtenu pour chaque nom propre cinq indices de notoriété. Nous allons maintenant calculer une valeur finale, égale à 1, 2 ou 3. Pour cela, nous avons utilisé un calcul multicritère, **la méthode SAW** (simple additive weighting) [6], qui nécessite d'attribuer un poids à chaque critère.[6]
   - **La méthode SAW :** Cette méthode représente une technique multicritère qui consiste à calculer pour chaque entrée (nom propre) la somme de toutes les valeurs normalisées de ses critères correspondants, chacune d'entre elles est multipliée par le poids d'importance qui lui a été associé.[6]

➢ La figure suivante représente les étapes de calcule la notoriété.

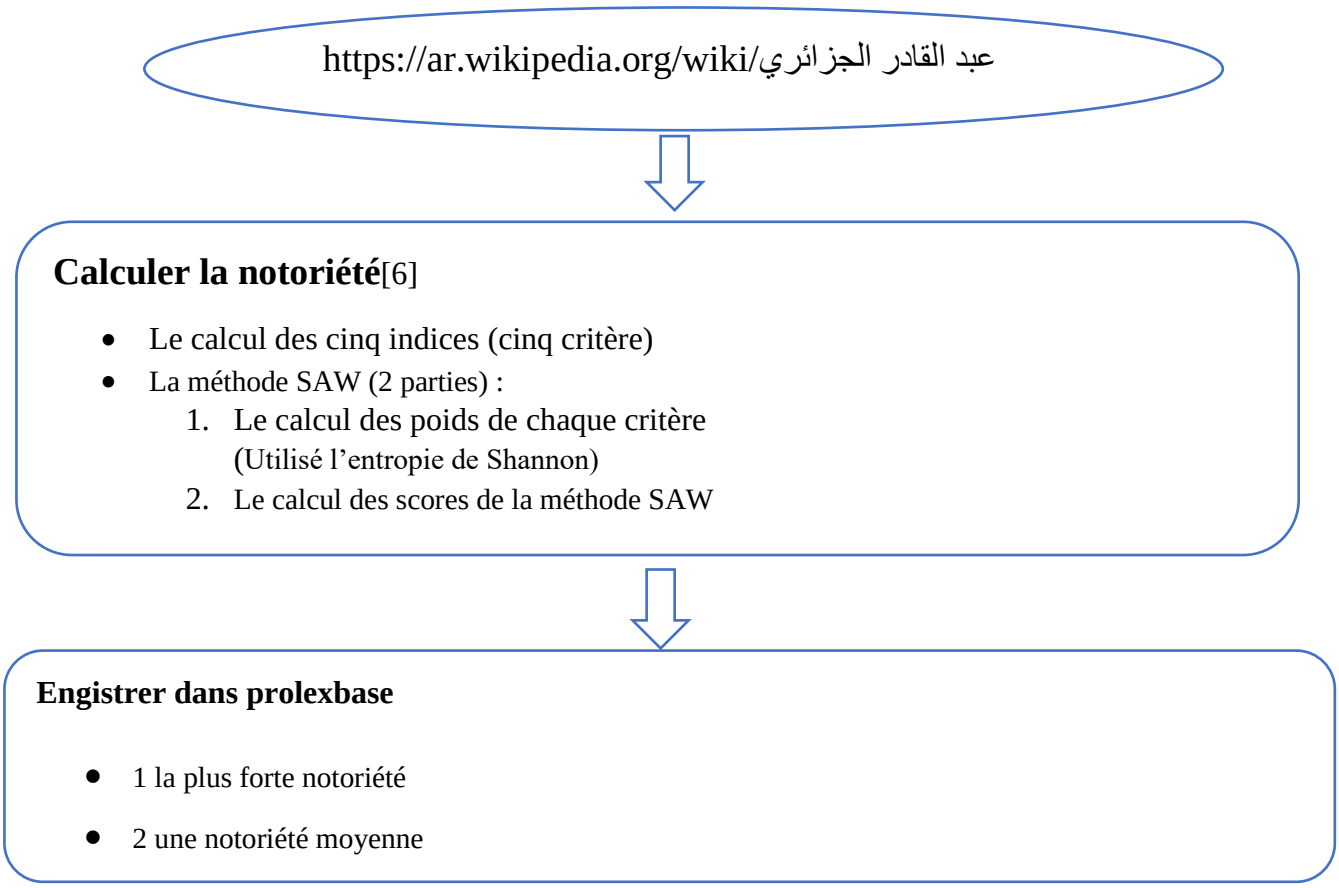

Figure 13 : les étapes de calcul de la notoriété d'un nom propre

III.3 choix des entités nommées

Nous avons choisi les entités nommées à partir de la notoriété, ce schéma représente les étapes de choix d'EN :

1. Calculons la notorité si la notorité =1 ou notorité=2
2. Ajoutée EN dans prolexbase
3. Si non (notorité=3) nous engriserons juste dans la table de notoriété

➢ Ce schéma représente les étapes de choix d'entité nommée

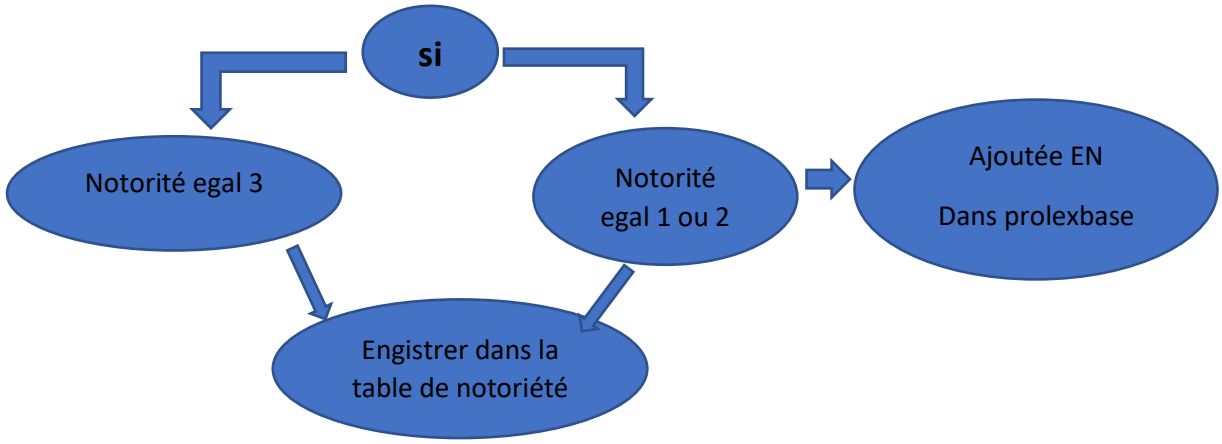

Figure 14 : les étapes de choix d'entité nommée

III.4 Ajout des noms propres arabes

L'arabe possède un vocabulaire riche et étendu, avec de nombreuses nuances de sens. Apprendre et mémoriser de nouveaux mots peut représenter un défi pour les non-locuteurs natifs, en particulier en raison des racines trilitères et quadrilitères qui nécessitent une compréhension de la structure des entités nommées arabes. Par exemple l'entité nommée **عبد القادر بن محي الدين**peut varié a**عبد القادر الجزائري**les deux entités nommées représenté la même personne.

III.4.1 Les Traitements automatiques avec Unitex

Unitex est une plateforme linguistique libre, qui consiste à traiter des ressources textuelles en des langues naturelles via un environnement de travail manipulable graphiquement ou à travers des lignes de commande [10]. Le corpus utilisé dans notre travail de détection des entités nommées et des Relations est seulement une partie de la wikipedia qui est représenté par l'infobox. Mais vue que tous les articles de la wikipedia ne contiennent pas d'infobox dans ce cas nous utilisons le premier paragraphe des articles de Wikipédia arabe qui représente un bon résumé de l'article. Nous sauvegardons ses données dans un fichier texte afin de pouvoir le traiter et le simplifier sous Unitex[41].

Il existe des grammaires il faut installer dans Unitex pour simplifie les textes un de ces grammaires est le CasANER.

CasANER [10] : 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 15 nous décrivons l'architecture de système CasANER et les modules qui le composent.[10]

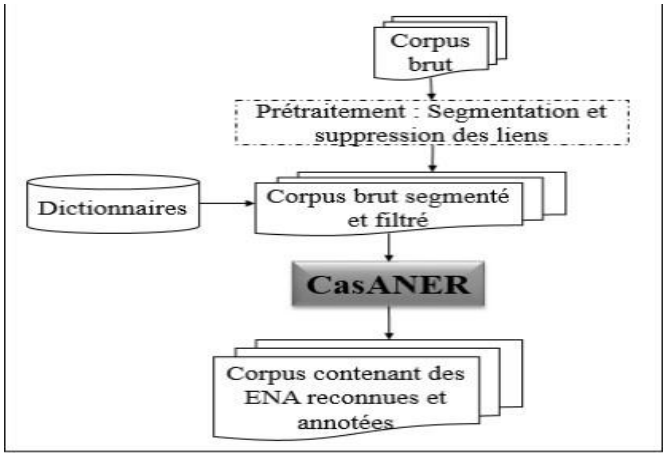

Figure 15 : l'architecture générale de CasANER[10]

[41]lien pour installe Unitex : http://unitexgramlab.org/fr

➢ **La figure**16montre les étapes de traitement avec Unitex.

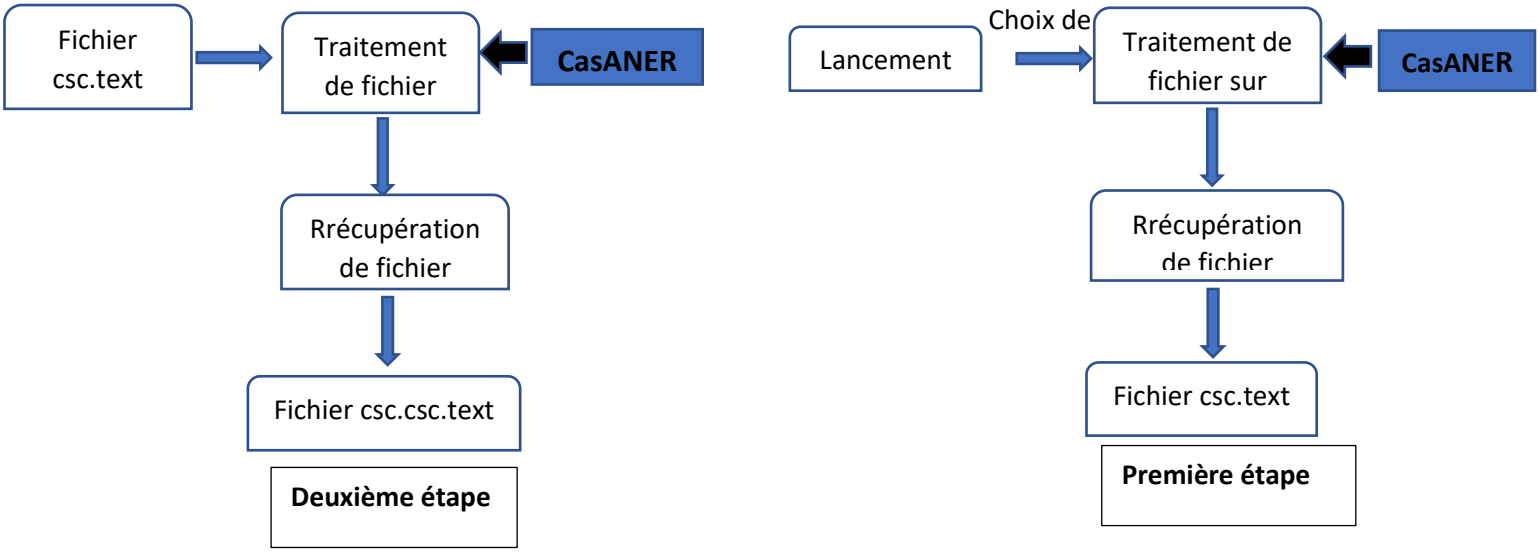

FIGURE 16: LES ETAPES DE TRAITEMENT AVEC UNITEX[10]

A. **Première étape** : Apres avoir choisie le texte à traiter par Unitex ; une série de traitement en langage naturel est effectué sur celui-ci (élimination des mots vides, lemmatisation, étiquetage, détection d'entité) on obtient après ce traitement un fichier Csc.text. Ce fichier contient les entités nommées arabe reconnues par le système CasANER [10].
B. **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.[10]

III.4.2L'enrichissement de Prolexbase

L'objectif essentiel de ce travail est de réaliser un enrichissement automatique de Prolexbase depuis la Wikipédia. Il s'agit d'alimenter la base de données par l'ajout de nouveaux noms propres arabe en grande quantité et avec une grande fiabilité (bruit limité).

III.4.2.2 l'ajout de l'Arabe et choisir le pivot

Notre thème basé sur l'ajout des entités nommées arabe et enrichir prolexbase. Pour l'ajout de la langue arabe nous avons suivi les étapes suivantes :

1. Extraire l'url d'un nom propre à partir d'une page wikipedia
2. Affecter une catégorie aux noms propres
3. Calculée la notoriété du nom propre
   Si la notoriété =1 ou la notoriété =2 en ajoute l'entité nomme dans prolex base

Pour ajouter le nom propre il faut parcourir la table qui contient les ENA pour avoir si EN existe ou non

4. Si l'entité existe en fait la mise à jour pour les autres langues en recalculons la notoriété dans la langue Français et anglais.
5. Sinon en vérifie si le nom propre existe dans les autres langues (FR, en) en parcourons les tables Français et anglais de prolexbase.
6. S'il existe en garde le même pivot et en crée un nouveau prolexème dans la langue arabe
7. Sinon on crée un pivot.

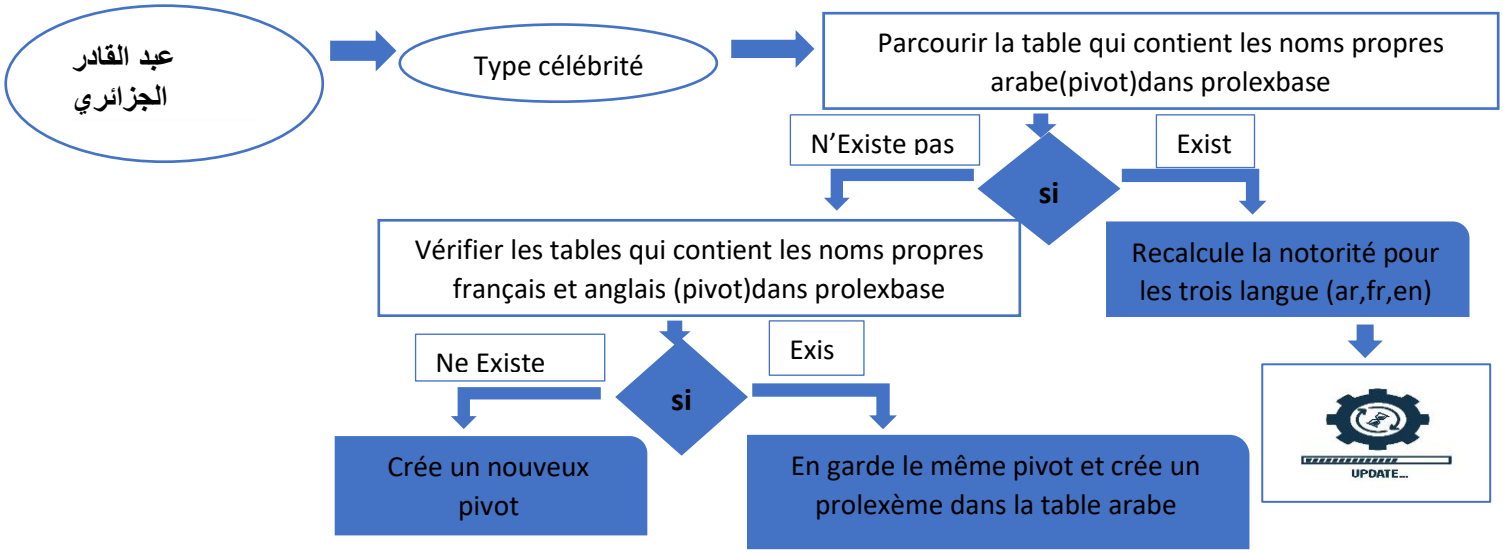

Figure 17: les étape pour ajoute la langue arabe dans prolexbase

- il existe trois relations sémantiques qui ne dépendent pas de la langue et sont associées aux pivots dans ce niveau : Synonymie, Méronymie et Accessibilité[6]. **La figure 18**montre un exemple, le nom propre **عبد القادر بن محي الدين** qui a une relation de synonyme avec **الأمير عبد القادر** et nous montrons les prolexèmes et les alias .

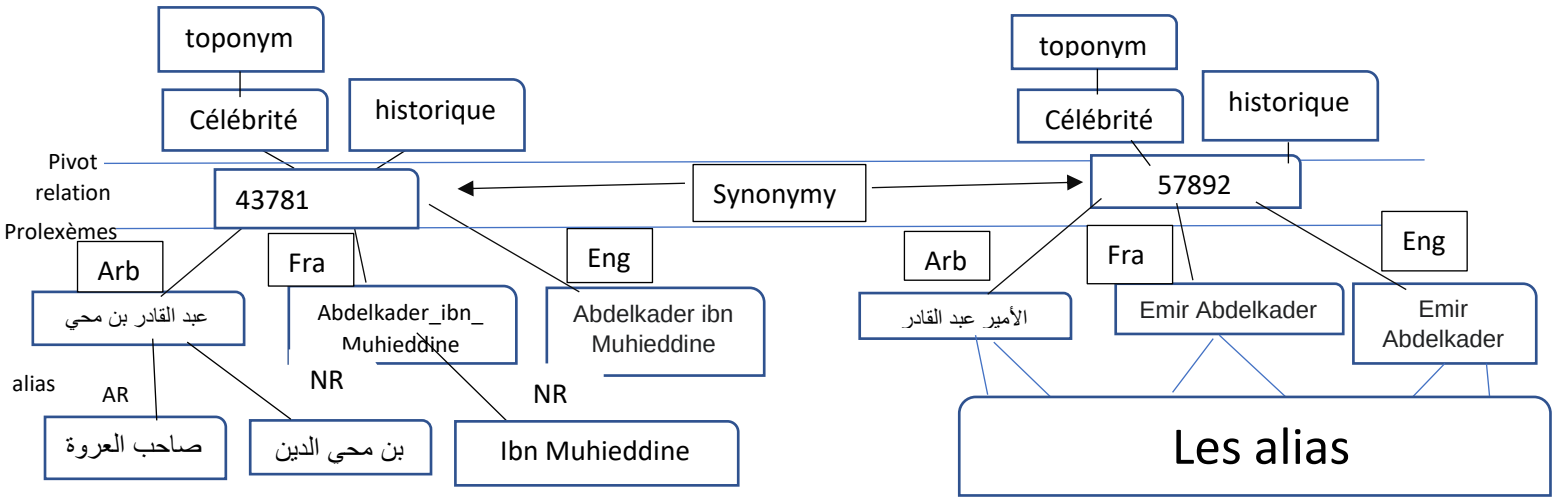

Figure **18:** exemple d'ajout d'un nom propre avec trois niveaux et deux prolexèmes

AR : adjectif relationnel NR : nom relationnel

Dans cette figure nous avons prié comme exemple deux pivot avec une relation de synonymie. Le premier pivot représente le nom propre عبد **القادر بن محي الدين** qui est le synonyme de **الأمير عبد القادر (deuxième pivot)** et nous avons présenté les prolexèmes des deux noms propres en arabe en Français et en anglais et les alias de chaque prolexèmes.

III.5 Les diagrammes UML

III.5.1 Diagramme de cas d'utilisation

Les diagrammes de cas d'utilisation capturent le comportement d'un système, d'un sous-système, d'une classe ou d'un composant tel qu'il est vu par les utilisateurs externes. La figure 19 montre un diagramme de cas d'utilisation de notre système.

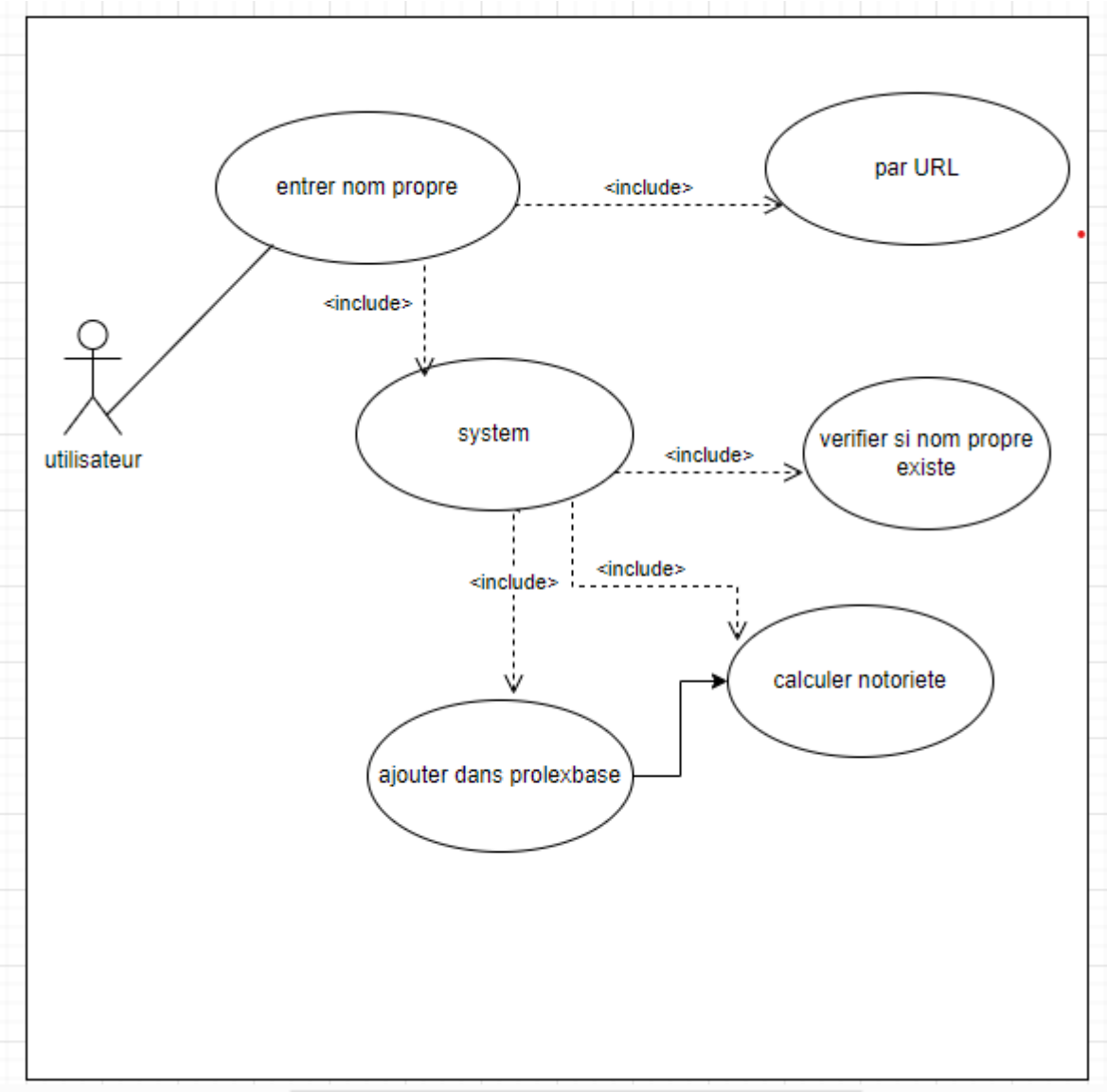

Figure 19 : diagramme de cas d'utilisation

III.5.2 Diagramme de classe

Les diagrammes de classes fournissent une vue holistique d'un système en montrant ses classes, ses interfaces et ses collaborations, ainsi que les relations entre elles. Les diagrammes de classes sont statiques : ils montrent en quoi consiste l'interaction, mais pas ce qui se passe pendant l'interaction.[41]

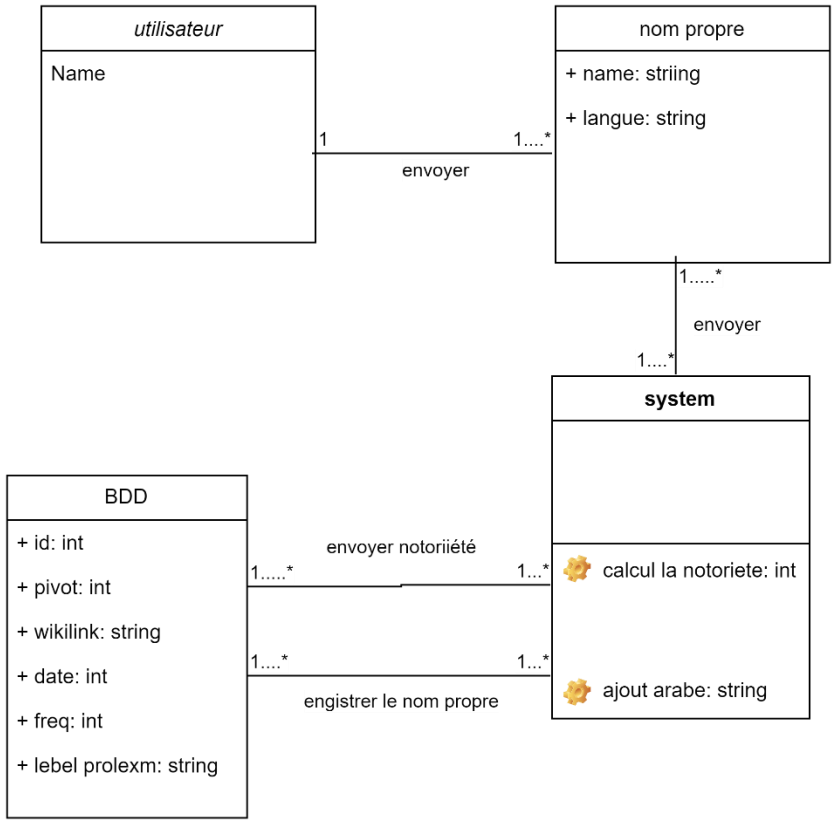

Figure 20 : diagramme de classe

III. Conclusion

Dans ce chapitre nous avons présenté notre algorithme pour l'enrichissement de prolexbase. Nous avons commencé par utiliser les liens Wikipédia (les url) dans la langue Arabe pour ajouter des noms propres arabe dans prolexbase. Ensuite, nous calculons la notoriété de chaque nom propre qui nous permet de décider de la pertinence de celui-ci dans la base. Vue les difficultés liées aux traitements de la langue arabe ; nous avons utilisé Unitex et le système CasANER [37] pour simplifier la détection des relations entre entités (les alias, les dérivé, les instances et les relation …ect) Finalement nous obtenons une base de données (prolexbase) qui contient des entités nommées arabe avec une mise à jour pour les langues français et anglais.

Chapitre4

Implémentation

IV Introduction

Dans ce dernier chapitre, nous présentons d'abord l'ensemble d'outils et d'environnements de programmation utilisés pour l'implémentation de notre programme. Ensuite nous présentons les résultats de notre enrichissement. Enfin nous terminons des démonstrations des interfaces de nôtres enrichissement.

IV.1Environnement de développement

IV.1.1Visual Studio Code

Un éditeur de code open-source développé par Microsoft supportant un très grand nombre de langages grâce à des extensions. Il supporte l'autocomplétions, la coloration syntaxique, le débogage, et les commandes git.[37]

IV.1.2 Draw.io

Une application gratuite en ligne, accessible via son navigateur (protocole https) qui permet de dessiner des diagrammes ou des organigrammes. Cet outil vous propose de concevoir toutes sortes de diagrammes, de dessins vectoriels, de les enregistrer au format XML puis de les exporter. [38]

IV.1.3XAMPP

un ensemble de logiciels permettant de mettre en place un serveur Web local, un serveur FTP et un serveur de messagerie électronique. Il s'agit d'une distribution de logiciels libres (X (cross) Apache Maria DB Perl PHP) offrant une bonne souplesse d'utilisation, réputée pour son installation simple et rapide. Ainsi, il est à la portée d'un grand nombre de personnes puisqu'il ne requiert pas de connaissances particulières et fonctionne, de plus, sur les systèmes d'exploitation les plus répandus.[39]

IV.1.4MongoDB

un système de gestion de base de données orienté documents, répartissable sur un nombre quelconque d'ordinateurs et ne nécessitant pas de schéma prédéfini des données. Il est écrit en C++. Le serveur et les outils sont distribués sous licence SSPL, les pilotes sous licence Apache et la documentation sous licence Creative Commons[2]. Il fait partie de la mouvance NoSQL.[40]

IV.2 Langages de programmation

IV.2.1 JavaScript

JavaScript est le principal langage de script des navigateurs Web et est essentiel aux applications Web modernes. Les programmeurs ont commencé à l'utiliser pour écrire des applications complexes, mais il y a encore peu de support d'outils pendant le développement.

VI.2.2 Html

Le langage de balisage hypertexte, souvent abrégé en HTML, est un langage de balisage destiné à représenter des pages Web. C'est un langage d'écriture hypertexte, D'où le nom. HTML peut également être sémantiquement et logiquement structuré, et mettre en forme le contenu de la page, y comprendre les ressources multimédias, y comprendre les images, les formulaires de saisie et les programmes informatiques.

VI.2.3 CSS

Les feuilles de style en cascade peuvent être traduites par "feuille de style en cascade". CSS est Langage informatique utilisé sur le Web pour formater des documents HTML ou XML. Les fichiers CSS contiennent le code qui gère la conception des pages HTML

IV.3 Résultats de l'enrichissement et l'ajout de la langue arabe

Nous avons présenté les résultats de l'enrichissement dans les tableaux suivants :

Le tableau 2 représente le nombre de pivot qui existe déjà, le nombre de pivot après l'enrichissement et le nombre des entités nommées arabe ajouté.

Nombre de pivot avant l'ajout dans toutes les langues	Nombre de pivot après l'ajout dans toutes les langues	Nombre des URL ajoutée (comme pivot et sans pivot)	Nombre d'entités nommées arabe ajoutée (comme pivot et sans pivot)	Nombre de prolexèmes ajoutée après l'enrichissement (dans la langue arabe)
36184	**36284**	**100**	**100**	**100**

Tableau 2 : le nombre de pivot avant, après l'enrichissement et le nombre ENA ajoutée

Le tableau 3 représente un exemple des pivots en arabe avec les prolexèmes

Numéro d'un pivot en arabe	Prolexème arabe	Prolexème français	Prolexème anglais
38558	**باريس**	**Paris**	**Paris**
52307	أفلاطون	**Platon**	**Platon**
45530	الجزائر	**Algérie**	**Algeria**

Tableau 3 : exemple des pivots en arabe avec les prolexèmes

Le tableau 4 représente un exemple de notoriété avant et après l'enrichissement

Le pivot	Label prolexème	Notoriété avant l'enrichissement	Notoriété après l'enrichissement
514	ستراسبورغ	3	2

Tableau 4: exemple de notoriété avant et après l'enrichissement

Le tableau 5 représente le nombre de prolexèmes dans chaque langue :

Nombre de prolexèmes avant l'ajout et l'enrichissement dans la langue arabe.	Nombre de prolexèmes avant l'ajout et l'enrichissement dans la langue française.	Nombre de prolexèmes avant l'ajout et l'enrichissement dans la langue anglais.
8403	7933	19848

Tableau 5: le nombre de prolexèmes dans chaque langue

IV.4 Présentation de travail

Dans notre travail nous avons utilisée deux bases de données dans prolexbase .la base de données 2_prolexbase_3_1_other_data(SQL) et la base de donnéechanonepro (no SQL).

La figure 22 représente l'ajout des noms propres arabe dans la base de données chanonepro.

Dans la base de données 2_prolexbase_3_1_other_data il existe une table prolexeme_arb qui contient les noms propres arabe avant l'enrichissement et l'ajout dans la langue arabe. Nous avons utilisé cet table pour ajoutée des entités nommées en vérifions si le nom propre existe déjà si oui nous gardons le même pivot et nous enregistrons le nom propre dans notre base de données chanonepro , un exemple est présenté dans la figure ci-dessous :

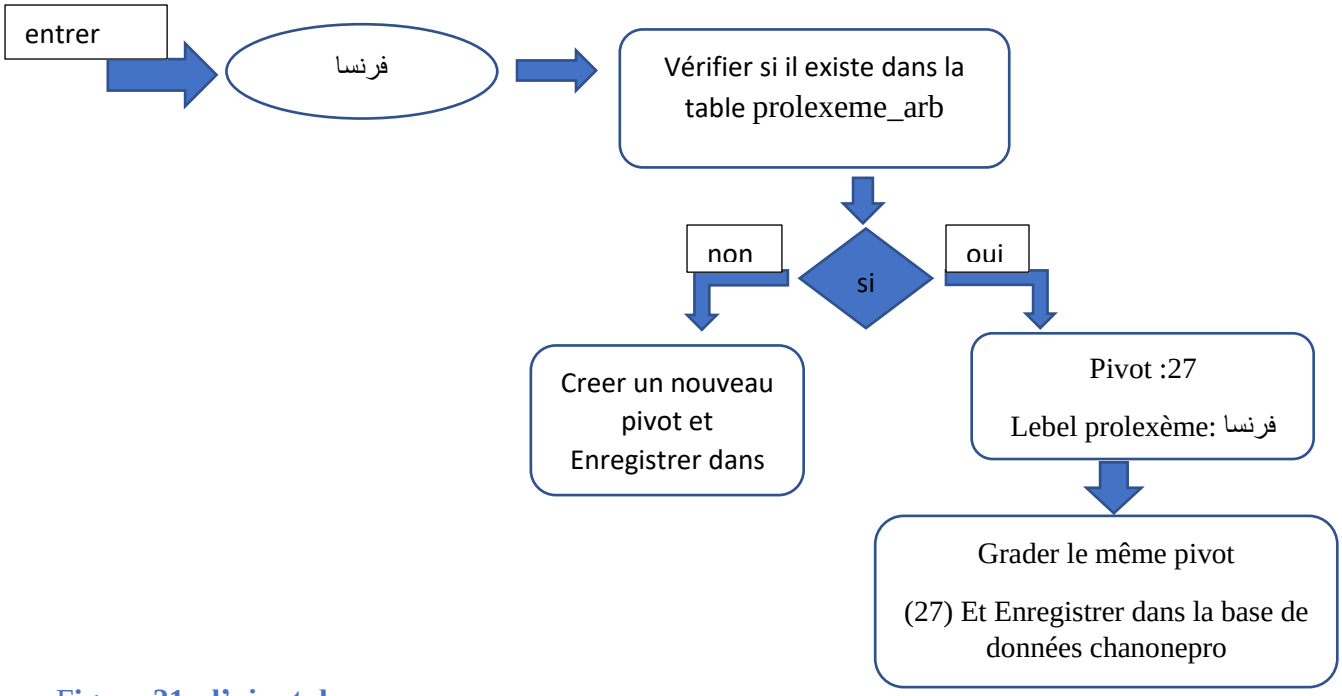

Figure 21 : l'ajout des noms propres

- **La base de données2_prolexbase_3_1_other_data**

La figure 22 représente une capture d'écran de la base de données 2_prolexbase_3_1_other_data et la table prolexeme_arb(SQL). Cette base contient une table prolexeme_arb dans la quelle ou le pivot sera vérifié. Dans cette table il existe les label prolexèmes (les noms propres) avec numéro de pivot, url de wikipedia (wikepedia link) et la notoriété (num-frequency) et num-prolexeme.

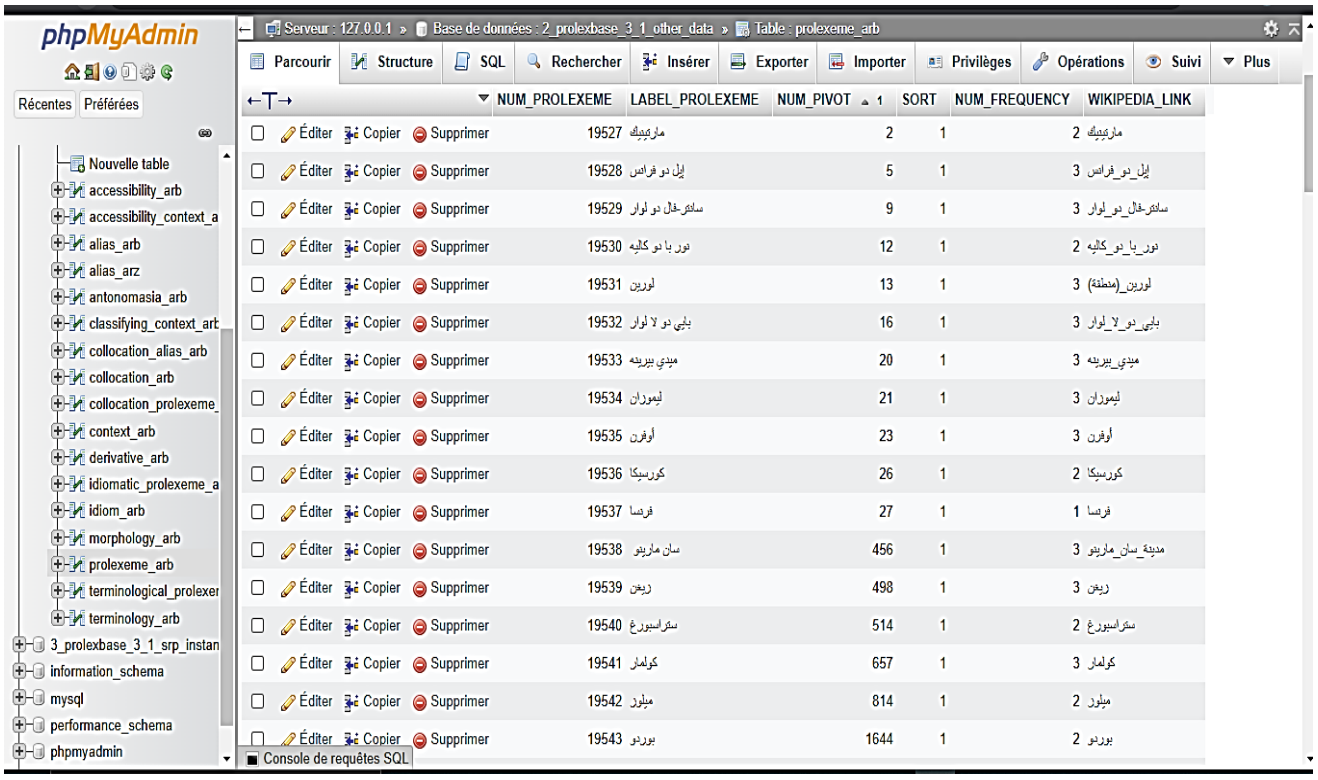

Figure 22 : la base de données 2_prolexbase_3_1_other_data et la table prolexeme_arb

- **La base de donnéechanonepro**
  Nous avons créé une base de données chanoneproc qui est une base de données no SQL pour enregistrer et ajouter des noms propres arabe dans la Table chanonepros.La figure 24 représente la base de données chanonepro. Dans l'exemple de la figure 23, le nom propre فرنسا est engistrer dans la base avec le num pivot 27, lebel prolexme : فرنسا et le fréq :1 c'est la notoriété et les 5 critères pour calculer la notoriété plus la date de l'enregistrement et les url de la wikipedia (wiki Link).

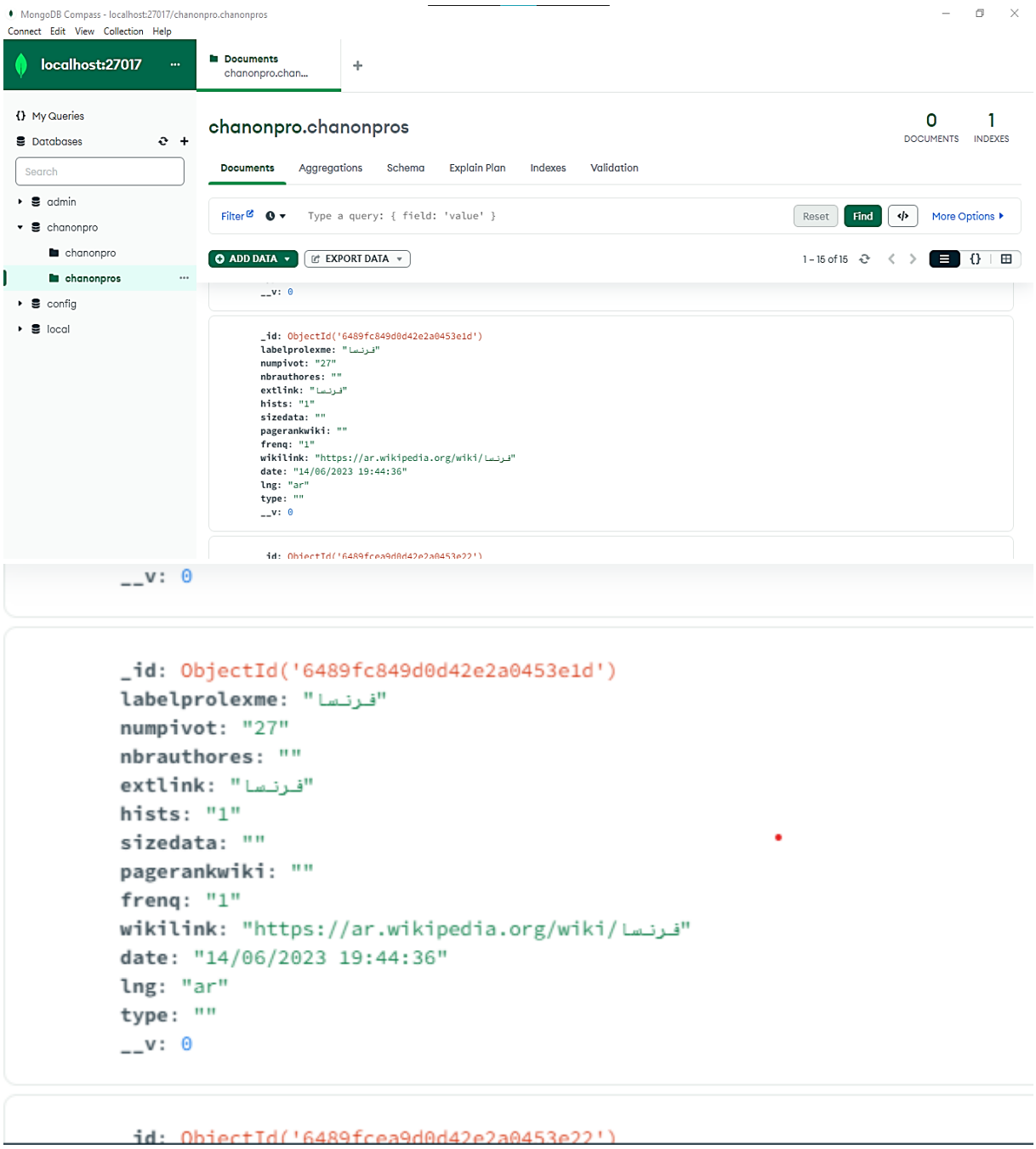

Figure 23 : base de données chanonepro

- **L’interface Enrichissement prolexbase Project**

Sur cette interface, l’utilisateur saisie le nom propre dans la barre de recherche. Si le nom propre est trouvé, l’info box de celui-ci sera affiché. Le nom propre sera conservé dans historique. La figure 24 représente l’interface Enrichissement prolexbase Project.

Figure 24 : Interface Enrichissement prolexbase Project

- Dans certain cas, le calcul de la notoriété nous signale une erreur car il manque des liens des contributeurs de la page (figure 25). Dans ce cas nous avons décidé de ne pas ajouter le nom propre a prolexbase mais les sauvegardés comme même, car ce lien n’existe pas aujourd’hui mais pourra exister dans les années à venir.

Figure 25: erreur de lien contrubitors

IV. Conclusion

Dans ce chapitre, nous avons présenté les environnements de développement et les différents langages de programmation et logiciels utilisés dans l'implémentation de notre système.

Nous avons aussi présenté le résultat obtenu est des exemples sur les tables et les données que nous avons ajoutées a prolexbase afin d'ajoutés les noms propres arabes.

Conclusion générale

L'enrichissement de Prolexbase a permis d'améliorer sa couverture lexicale dans différentes langues, offrant ainsi une plus grande diversité et une meilleure représentation de la langue dans différents contextes. Les informations supplémentaires ont permis d'enrichir les relations sémantiques entre les mots, ce qui facilite la recherche d'informations et l'analyse linguistique. De plus, l'ajout de synonymes, d'antonymes et de traductions a considérablement élargi les possibilités d'utilisation de Prolexbase. Les utilisateurs peuvent désormais trouver des mots équivalents, des termes opposés et des traductions dans différentes langues, ce qui facilite la compréhension et la communication inter linguistique.

Dans le présent travail, nous avons pu réaliser un système d'enrichissement d'un lexique de noms propres. Ce système permet d'enrichir la base de données prolexbase et ajouté des entités nommées arabe. Nous avons suivi plusieurs étapes pour réalisées ce travail.

Dans la première partie, nous avons défini la reconnaissance d'entité nommée. Nous avons présenté les Catégorisations et les trois approches d'une entité nommée, Nous avons effectué aussi une étude de Reconnaissance d'entité nommée Arabe en présentons la langue arabe et ces principes, les catégorisations d'EN arabe et Les problèmes d'analyse dans la langue arabe.

Dans la deuxième partie, nous avons décrit le dictionnaire prolexbase est ses niveaux. Nous avons effectué aussi une étude sur La structure générale d'une page Wikipédia et L'accès au contenu de l'encyclopédie. Ensuite nous avons parlé de la Wikipédia arabe.

La troisième partie, est consacré pour la conception et la modélisation du notre système. Dans Le quatrième chapitre, nous avons présenté les outils de l'implémentation et les résultats de l'enrichissement et de l'ajout de la langue arabe.

Finalement, toutes les tâches déjà mentionnées ont participé à un projet d'enrichissement de prolexbase et l'ajout d'entité nommée arabe en utilisant le corpus Wikipédia arabe.

Bien sûre ce travail n'est pas complètement achevé, car nous avons plusieurs perspectives, nous citons comme exemple :

1 -L'ajout des alias, dérivés, acronyme…ect, en détectant les relations sémantiques entre les noms propres, cela peut se faire à partir de l'info box ou du premier paragraphe de la wikipedia.

2-Le calcul de la notoriété dans les différentes langues du nom propres, pour voire l'évolution de celle-ci au fil des années et au niveau géographique. Ce qui permettra de rendre prolexbase comme un système de classification des noms propres.

Bibliographie

Héla FEHRI

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

Noureddine DOUMI

[2] Extraction de connaissances à partir du texte Université DJILLALI LIABES FACULTE DES SCIENCES EXACTES Sidi Bel Abbes Année universitaire 2016/2017.

Elizabeth D. Liddy

[3]Natural Language ProcessingSyracuse University2001. Recommended Citation Liddy, E.D. 2001. Natural Language Processing. In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc.

Houda Saadane

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

Mouna Elashter

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

Sylvie GUILLEMIN-LANNE, Fathi DEBILI, Zied Ben TAHAR, Chafik GACI

[7]Reconnaissancedes entités nommées en arabes (*) TEMIS, Tour Gamma B, 193-197 rue de Bercy, 75012 PARIS, France (**) LLACAN, INALCO, CNRS, 7, rue Guy Môquet, 94801 Villejuif, France.

Thierry Grass, Denis Maurel & Mickaël Tran

[8]Prolexbase : une ontologie pour le traitement multilingue desnoms propres.

Béatrice Daille, Nordine Fourour & Emmanuel Morin

[9] Catégorisation des noms propres : une étude en corpus. IRIN (Institut de recherche en informatique de Nantes), Université de Nantes,(2000).

Fatma Ben Mesmia Chaabouni

[10] Reconnaissance des entités nommées à partir de Wikipédia arabe Traitement du texte et du document. Université de Tunis El Manar, (2019). Français.

LAKEL Kheira

[13]Les annotations sémantiques dans les documents Web : application aux textes psychologiques en langue arabe, universitéOranUSTO-MB,Année Universitaire : 2017-2018.

Erik F. Tjong Kim Sang et Fien De Meulder

[20] Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition.University of Antwerp , (2003).

Slim Mesfar, 2008

[21]Analyse morphosyntaxique automatique et reconnaissance des entités nommées en arabe standard. Université de Franche-Comté, (2008).

Nasredine Semmar, Faıza Elkateb-Gara et Christian Fluhr.

[22] En utilisant un Stemmer dans un système de traitement du langage naturel pour traiter l'arabe Recherche d'informations multilingues,Proceedings of the Fifth Conference on Language Engineering, (2005).

Attia, M. (2008)

[23] Traitement de l'ambiguïté morphologique et syntaxique de l'arabe dans le cadre LFG en vue de la traduction automatique. University of Manchester, Faculté of Humanités.(2008).

(khoja et al., 2001) S.khoja , R.Garside, G.Knowles

[24] a tagset for the morpho-syntactic tagging of Arabic actes de la conferance internatinnel corpuslinguals 2001, lancaster,(2001).

D.E.Kouloughli

[25] Lexique fundamental de l'arabe standard moderne,paris,(1991).

HABASH, N

[26] Introduction au traitement du langage naturel arabe. Éditeurs Morgan &Claypool,(2010).

Shaalan K. and Oudah M

[27] Une approche hybride de Reconnaissance des entités nommées en arabe.Journal of Information Science,(2014).

Chalabi Achraf

[28] Arabisation transparente basée sur MT de l'internet TARJIM. COM. Dans Envisager la traduction automatique dans le futur de l'information,Springer Berlin Heidelberg, (2000).

Boujelbane Rahma, Ellouze Mariem, Béchet Frédéric, & Belguith Lamia

[29] De l'arabe standard vers l'arabe dialectal: projection de corpus et ressources linguistiques en vue du traitement automatique de l'oral dans les médias tunisiens,Traitement automatique des langues (TAL),(2015).

Meilland, J.-C. & Bellot, P

[30]Extraction automatique de terminologie à partir de libellés textuels courts. La Linguistique de corpus,(2005).

Friburger, N

[31]Reconnaissance automatique des noms propres ; application à la classification automatique de textes journalistiques. Thèse de doctorat. Université François Rabelais Tours, (2002).

Tran, M

[32]Prolexbase. Un dictionnaire relationnel multilingue de noms propres : conception implantation et gestion en ligne. Thèse de doctorat. Université François Rabelais Tours,(2006).

Sekine, S., Sudo, K. &Nobata, C.

[33]Extended Named Entity Hierarchy. In The Third International Conference on Language Resources and Evaluation. Canary Island, Spain, (2002).

Fourour, N. & Morin, E

[34]Apport du Web dans la reconnaissance des entités nommées. Revue Québécoise de Linguistique (RQL),(2003).

Bouchou B., Maurel D

[35]Prolexbase et LMF : vers un standard pour les ressources lexicales sur les noms propres. Traitement automatique des langues, (2008).

Maurel D., Tran M., Friburger N.

[36] Projet Technologue Noms Propres : Constitution et exploitation d'un dictionnaire relationnel multilingue de noms propres.TALN 2006, Cahiers du Cental, Louvain, Belgique,(2006).

Webographie

[4]https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Liens_internesconsulté le 5 février 2023

[11]https://fr.wikipedia.org/wiki/Aide:Infoboxconsulté le 5 février 2023

[12]http://igm.univ-mlv.fr/~dr/XPOSE2011/Wikipedia/presentation_wikipedia.html)consulté le 5 février 2023

[14]https://fr.wikipedia.org/wiki/Wikipedia:Liens_externesconsulté le 5 février 2023

[15] https://fr.wikipedia.org/wiki/Aide:Insérer une référenceconsultée le 5 février 2023

[16]https://www.wikimedia.org/consulté le27 décembre 2022

[17]https://meta.wikimedia.org/wiki/Wikimedia_chaptersconsulté le27 décembre 2022

[18]https://www.cetic.be/Exploiter-le-contenu-de-Wikipediaconsulté le 30 décembre 2022

[19]https://meta.wikimedia.org/wiki/Data_dumpsconsulté le27 décembre 2022

[37]https://framalibre.org/content/visualstudio-code consulte le 9 juin 2023

[38]https://www.tice-education.fr/tous-les-articles-er-ressources/articles-internet/819-draw-io-un-outil-pour-dessiner-des-diagrammes-en-ligne consulte le 9juin2023

[39]https://fr.wikipedia.org/wiki/XAMPP consulte 9juin2023

[40]https://fr.wikipedia.org/wiki/MongoDB consulte le 9juin2023

[41]http://docwiki.embarcadero.com/RADStudio/Rio/fr/D%C3%A9finition_des_diagrammes_de_classes_UML_1.5 consulte le 15juin2023