UNIVERSITÉ FRANÇOIS – RABELAIS DE TOURS

ÉCOLE DOCTORALE [MIPTIS]

[LI]

THÈSE présentée par :

[Mouna Elashter]

Soutenue le : **04 juillet 2017**

Pour obtenir le grade de : **Docteur de l'université François – Rabelais de Tours**

Discipline/ Spécialité : Informatique

Gestion et extension automatiques du dictionnaire relationnel multilingues de noms propres Prolexbase

Mise à jour multilingues et création d'un volume arabe via la Wikipedia

THÈSE dirigée par :

Denis Maurel	PR, université François-Rabelais de Tours

RAPPORTEURS :

Béatrice Daille	Professeur, université de Nantes
Kais Haddar	MCF HDR, université de Sfax, Tunisie

JURY :

Béatrice Daille	Professeur, université de Nantes
Kais Haddar	MCF HDR, université de Sfax, Tunisie
Béatrice Markhoff	MCF HDR, université François-Rabelais de Tours
Denis Maurel	PR, université François-Rabelais de Tours

A ma mère et à ma famille.

Remerciements

Tout d'abord, je souhaite adresser tous mes remerciements à mon directeur de thèse, Monsieur Denis Maurel, pour sa bienveillance, ses encouragements et sa disponibilité tout au long de ma recherche.

Je remercie également les membres du jury, Madame Béatrice Daille, Madame Béatrice Markhoff et Monsieur Kais Haddar d'avoir accepté de juger mon travail.

J'adresse un remerciement très sensible à ma mère Fatma qui m'a inspirée tout au long de ma vie et m'a encouragée et soutenue, depuis la Libye.

Je remercie chaleureusement toute ma famille, et en particulier mon mari Hassan YOUSSEF pour son soutien, mes adorables filles et mon cher fils, qui m'ont beaucoup aidés et encouragés tout au long de ce travail de recherche.

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 et la traduction automatique des noms propres. Toutefois, 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, DBpedia (Auer *et al*., 2007), Geonames et Yago2 (Hoffart *et al.*, 2012).

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. Il a été conçu manuellement et une première tentative semi-automatique a été réalisée par le projet ProlexFeeder (Savary *et al.* 2013). Notre travail avait pour objectif d'élaborer un outil de mise à jour et d'extension automatiques de ce lexique, et l'ajout de la langue arabe. Tout d'abord, une mise à jour multilingue de la base de données a été effectuée grâce à l'établissement d'un système automatique de consolidation des liens Wikipédia dans Prolexbase en nous servant du concept interlangue de Wikipédia. En conséquence, un nombre considérable de nouveaux liens Wikipédia a été ajouté dans toutes les langues constituant la base de données, et cet ajout a été précédé, le cas échéant, d'un traitement des redirections.

Un système entièrement automatique a également été mis en place qui permet de calculer, via l'encyclopédie Wikipédia, un indice de notoriété pour les entrées de Prolexbase ; cet indice dépend de la langue et participe, d'une part, à la construction d'un module de Prolexbase pour la langue arabe et, d'autre part, à la révision de la notoriété actuellement présente pour les autres langues de la base. Pour calculer la notoriété, une technique multicritères de l'aide à la décision a été utilisée : la méthode SAW incluant le calcul de l'entropie de Shannon, à partir de cinq valeurs numériques déduites de l'encyclopédie Wikipédia.

Finalement, l'utilisation des liens Wikipédia a été l'instrument fondamental pour la création d'un volume arabe dans Prolexbase par un processus d'extraction de noms propres arabes depuis leurs liens Wikipédia obtenus précédemment.

Mots Clés : Nom propre, Prolexbase, Bases lexicales multilingues, Notoriété, Langue arabe, Wikipédia.

Abstract

Lexical databases play a significant role in natural language processing (NLP), such as information retrieval, recognition of named entities, and automatic translation of proper names. However, they require permanent development and enrichment through the exploitation of free resources rich in texts from the semantic web, among others, the universal encyclopedia Wikipedia, DBpedia (Auer et al., 2007), Geonames and Yago2 (Hoffart et al. Al., 2012).

In particular, the multilingual relational electronic dictionary of proper names, Prolexbase, which issued of numerous studies on NLP, has ten languages, three of which are well covered: French, English and Polish. It was manually designed; the first semi-automatic attempt was made by the ProlexFeeder project (Savary et al., 2013).

The objective of our work was to create an automatic updating and extension tool for this lexicon. First, a multilingual update of the database was carried out by establishing an automatic system for consolidating Wikipedia links in Prolexbase using the interlanguage concept of Wikipedia. As a result, a considerable number of new Wikipedia links have been added in all the languages constructed the database, preceded by redirection processing if needed.

In addition, a fully automatic system has been implemented to calculate, via Wikipedia, the notoriety of the entries of Prolexbase. This notoriety is language dependent, is the first step in the construction of an Arabic module of Prolexbase, and it takes a part in the notoriety revision currently present for the other languages in the database. To calculate the notoriety, we present a multi criteria technique, the method SAW (preceded by the calculation of Shannon entropy), starting from five numerical values deduced from Wikipedia.

Finally, the use of Wikipedia links was the fundamental instrument for creating an Arabic volume in Prolexbase by a process of extracting Arabic proper names from their previous Wikipedia links.

Key Words: Proper noun, Prolexbase, Multilingual lexical databases, Notoriety, Arabic language, Wikipedia.

Table des matières

Liste des tableaux

Chapitre 1

Chapitre 2

Chapitre 4

Chapitre 5

Chapitre 6

Liste des figures

Chapitre 3

Chapitre 4

Chapitre 5

Chapitre 6

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. Toutefois, 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, DBpedia (Auer *et al*., 2007), Geonames et Yago2 (Hoffart *et al*., 2012).

Prolexbase (Tran & Maurel, 2006) est une ressource libre[1], un dictionnaire relationnel multilingue de noms propres, au format LMF (ISO 24613) depuis les travaux de Bouchou et Maurel (2008). Prolexbase, issue 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 ; ce dictionnaire a été conçu manuellement, mais une première tentative d'enrichissement semi-automatique a été réalisée par le projet Prolexfeeder (Savary *et al.* 2013) en augmentant fortement le nombre d'entrées polonaises dans la base de données. A cet égard, l'objectif fondamental de notre travail de thèse est d'automatiser d'une manière générale le traitement de l'ensemble des langues contenues dans cette base de données.

En effet, notre recherche vise à élaborer un ensemble d'outils pour la mise à jour et l'extension automatiques de ce lexique. Tout d'abord, nous allons effectuer une mise à jour multilingue des différentes tables de prolexèmes dans Prolexbase en développant un système automatique de consolidation des liens Wikipédia qui sont dans la base, en nous servant du concept interlangue relatif à l'encyclopédie Wikipédia. Dès lors, notre objectif est d'ajouter un nombre important de nouveaux liens Wikipédia dans les dix langues constituant la base de données. A cet égard ; il conviendra auparavant de détecter les redirections éventuelles afin d'y remédier, et ensuite de procéder à la complétion des liens manquants dans certaines langues de Prolexbase.

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

Dans un deuxième temps, une extension automatique du nombre de langues dans Prolexbase sera effectuée via ce même système en permettant tout d'abord l'introduction de la langue arabe. Dès lors, notre travail, par l'ajout des liens Wikipédia et des numéros de pivot, contribuera à préparer l'extension à d'autres langues. A cet effet, nous avons également choisi d'élargir notre étude à l'arabe égyptien et au wolof. (Par une collaboration avec un chercheur du Sénégal, Alla Lo).

Une autre appréciation de notre recherche est la génération automatique d'un dictionnaire de noms propres arabes pour le TAL. Cependant, celui-ci ne doit pas être trop volumineux pour limiter l'ambiguïté. Autrement dit, la constitution d'un dictionnaire commence par le choix des entrées à y placer. Cela est surtout vrai pour les dictionnaires édités sur papier, mais aussi pour les dictionnaires électroniques où certaines entrées peuvent augmenter inutilement l'ambiguïté. De plus, "il faut mentionner l'important problème que pose l'absence des noms propres dans les dictionnaires de langue" (Rey, 1977) et, pourrait-on ajouter aujourd'hui, dans les dictionnaires électroniques. Mikheev *et al*. (1999) ont constaté que, dans la reconnaissance des entités nommées, "without gazetteers […] locations come out badly", mais ils ont ajouté: "The collection of gazetteers need not be a bottleneck: […] relatively small gazetteers are sufficient to give good Precision and Recall". En effet, les noms propres et leurs dérivés sont très souvent ambigus et une trop grande taille du dictionnaire peut créer des difficultés.

La recherche du nom Paris sur Geonames[2] donne des centaines de résultats, score à peine amélioré en précisant le type *city* (210 résultats). Sur l'encyclopédie Wikipédia[3], la même recherche conduit directement à la capitale de la France (qui est le nom propre le plus connu désigné par le mot Paris) et indique qu'il existe un grand nombre de pages homonymes. Paris (France) a plus de notoriété que d'autres villes du même nom. Bien sûr, d'autres critères spécifiques au corpus traité peuvent influencer les choix, mais un dictionnaire générique doit s'appuyer sur cette notion. L'évaluation de cette notoriété n'est cependant pas évidente. Doit-elle juste reposer sur le choix d'un expert ? Plusieurs travaux suggèrent l'utilisation du Web sémantique, et plus précisément de l'encyclopédie en ligne Wikipédia, entre autres, (Chevalier *et al*., 2010 ; Eom et Shepelyansky, 2013 ; Yu *et al*., 2016).

[2] http://www.geonames.org/
[3] https://fr.wikipedia.org/

À ce stade, il est essentiel d'indiquer qu'au niveau multilingue de Prolexbase se trouvent des pivots qui représentent un point de vue sur les noms propres et se projettent sur chaque langue en des prolexèmes, ensembles de formes morphosémantiquement liées (alias et dérivés). A chaque prolexème est associé, éventuellement, un lien vers l'encyclopédie Wikipédia et, obligatoirement, un indice de notoriété basé sur trois valeurs, conformément à la norme ISO 12620 : l'indice le plus fort étant 1 et le plus faible 3. Les indices actuels ont été choisis manuellement, ce que nous souhaitons changer, d'abord pour permettre l'ajout automatique de nouvelles entrées et de nouvelles langues, mais aussi pour autoriser une réévaluation régulière de cette notoriété, qui évolue dans le temps.

Pour cela, un système entièrement automatique permettra de calculer, via l'encyclopédie Wikipédia, un indice de notoriété pour les entrées de Prolexbase ; cet indice de notoriété dépendra de la langue et participera, d'une part, à la construction d'un module de Prolexbase pour la langue arabe et, d'autre part, à la révision de la notoriété actuellement présente pour les autres langues de la base. Pour calculer la notoriété, nous avons choisi d'utiliser une technique multicritère de l'aide à la décision, à savoir : la méthode SAW incluant le calcul de l'entropie de Shannon, à partir de cinq valeurs numériques déduites de l'encyclopédie Wikipédia.

Un troisième outil, qui consiste à extraire les noms propres arabes ou lemmes de leurs liens Wikipédia produits lors de la première phase de notre travail, sera finalement établi. A cet égard, l'utilisation des liens Wikipédia représente un instrument fondamental pour la création d'un volume arabe dans Prolexbase.

Notre étude s'articulera en deux temps. Dans la première partie, nous présenterons l'état de l'art (chapitre 1), puis certains concepts de la langue arabe (chapitre 2), et enfin les fondements des ressources utilisées pour ce travail : Prolexbase et l'encyclopédie Wikipédia (chapitre 3). La deuxième partie sera consacrée à la réalisation pratique de notre projet. A cet égard, nous allons concevoir trois traitements multilingues automatiques dans l'objectif d'améliorer la performance de Prolexbase: la mise à jour (chapitre 4), le calcul de notoriété (chapitre 5) et l'ajout de la langue arabe (chapitre 6).

Première partie

Disposition du Travail

Chapitre 1 État de L'Art

Introduction

Construire une base de connaissance solide et riche est un point de départ essentiel pour atteindre les objectifs de notre recherche. Nous avons lu plus d'une cinquantaine d'articles et parmi eux nous avons choisi ceux qui ont des liens avec notre travail. Les travaux sont classés chronologiquement selon le sujet traité et selon les méthodes utilisées et le niveau des résultats obtenus.

Nous avons classé ces études dans les six points suivants :

1. L'analyse morphologique de textes arabes ;
2. La reconnaissance d'entités nommées dans les textes arabes ;
3. La gestion des bases de données lexicales multilingues ;
4. L'enrichissement de Prolexbase ;
5. L'exploitation de l'encyclopédie Wikipédia ;
6. Les mesures de la popularité des articles de l'encyclopédie Wikipédia.

1 L'analyse morphologique de textes arabes

Plusieurs travaux avaient pour objectif de construire des analyseurs morphologiques de textes arabes ; parmi eux, l'analyseur et le générateur morphologique à états finis pour la langue arabe Xerox (Beesley, 2001).

Xerox propose un analyseur-générateur morphologique pour l'arabe standard moderne qui a été rendu disponible pour les tests sur Internet[4]. Ce système est capable d'analyser les mots entièrement voyellés, partiellement voyellés ou non voyellés ; la présence de signes diacritiques limite automatiquement l'ambiguïté de la sortie.

[4] http://www.xrce.xerox.com/research/mltt/arabic.

L'analyseur–générateur est basé sur les dictionnaires d'un projet antérieur de la société ALPNET [5]; le système entier est reconstruit à l'aide de la technologie Xerox à états finis. L'implémentation vise à résoudre trois défis de l'arabe : les morphotactiques[6], les voyelles courtes et la recherche dans le lexique (lookup).

Le lexique consiste en 5000 racines et 400 modèles phonologiques distincts. Le calcul de Xerox à états finis a été utilisé pour insérer des racines dans leurs modèles et générer efficacement 85 000 radicaux valables ; le transducteur du lexique contient également des préfixes et suffixes appropriés qui sont ajoutés aux radicaux d'une manière concaténaive régulière. Le résultat de l'analyse inclut la racine de la forme de base suivie par les traits morphosyntaxiques pertinents du mot cible, (Beesley, 1996 ; Beesley 1998). Les avantages du système Xerox sont sa large couverture, la reconstruction des voyelles courtes et le glossaire anglais fourni pour chaque mot. Cependant, dans le système, il manque la spécification des expressions de mots composés (MWEs).

On peut également citer d'autres inconvénients de Xerox, qui sont :

- la surgénération dans la dérivation de mots en raison de la répartition inégale des modèles pour les racines ;
- le niveau élevé d'ambiguïté qui produit de nombreuses analyses pour la plupart des mots. (Attia, 2008).

Parmi les travaux, se trouve aussi BAMA (Buckwatter, 2004) qui est considéré comme l'un des meilleurs analyseurs de l'arabe et qui est par conséquent très répandu.

Le texte en entrée doit être translittéré en ASCII avant tout traitement, et le résultat doit être reconverti en arabe pour que cela soit compréhensible. Ce système ne permet pas l'analyse des textes contenant des chiffres 0...9.

---

[5] ALPNET est une société à Provo, Utah, États-Unis qui a écrit un certain nombre de systèmes de traitement en langage naturel.

[6] L'ensemble des règles qui définissent la manière dont les morphèmes (morpho) peuvent se toucher (tactiques) l'un l'autre. Selon le site : https://en.wikipedia.org/wiki/Morphotactics

BAMA utilise une approche concaténative du lexique, où les règles morphologiques et orthographiques sont intégrées directement dans le lexique. Le système comporte trois outils :

- un lexique qui contient trois listes : les préfixes, les racines et les suffixes. Chaque entrée du lexique consiste en sa catégorie morphologique, son étiquette grammaticale (pos) et sa traduction anglaise ;
- trois tables de catégories morphologiques compatibles ;
- un algorithme assez simple d'analyse qui consiste à :

1. découper le mot d'entrée en préfixes, infixes et suffixes pour extraire sa racine ;
2. déterminer toutes les segmentations possibles du mot puis chercher les résultats dans les listes des radicaux, des suffixes et des préfixes ;
3. vérifier si les morphologies de chacun des éléments sont compatibles entre elles en examinant les trois tables de correspondances : préfixe-radical, préfixe-suffixe et radical-suffixe.

En sortie, BAMA produit plusieurs analyses pour une entrée. Par exemple : pour le mot للكتب/llktb 'pour les livres', BAMA produit les analyses suivantes :

- sa diacritisation entière lilkutubi ;
- son identifiant de lemme kitAb-1 ;
- ses traits morphologiques et son étiquette grammaticale(POS) :

Li/PREP+ Al/DET+ kutub/NOUN+ i/CASE-DEF-GEN.

L'utilisation des lexèmes comme des formes de base dans ce système engendre un taux supplémentaire d'ambiguïté; par exemple, le lemme كتابِ / kitAb du lexème (1-كتاب/kitAb-1/livre) a pour catégorie (Nud), qui n'est pas compatible avec la catégorie du marqueur du féminin ة/ap (Nsuff-ap). Le même lemme, كتاب / kitAb, apparaît comme l'un des radicaux du

lexème **(**1-كتابة/kitAbap-1/écriture) avec une catégorie qui nécessite un marqueur du féminin comme suffixe (N. HABASH, 2004).

Quelques points faibles de cet analyseur ont été abordés par Attias (2006):

- absence de règles de génération : tous les lemmes sont listés manuellement et tous les lexèmes des formes fléchies associées sont énumérés, ce qui finit par augmenter le coût de maintenance du lexique.

- problème au niveau du traitement des proclitiques interrogatifs qui se localisent au début des verbes et des noms (exemple : « أأقول » «أمحمد »).
- l'intégration de certains termes de l'arabe classique qui ne sont plus utilisés, ce qui augmente l'ambiguïté. Par exemple : (حنيفة / haniyfa/orthodoxe) est un prénom en arabe standard moderne, alors que l'analyseur de Buckwater produit des traductions anglaises qui n'existent plus.

D'autres méthodes ont été proposées pour l'analyse morphologique de textes arabes non voyellés. Belguith *et al.* (2005) ont réalisé MORPH2, une méthode d'analyse et de désambiguïsation morphologique de textes arabes non voyellés. Elle se concentre sur l'ambiguïté morphologique due à l'absence des marques de diacritisation et la présence de formes nominales irrégulières. Il s'agit d'une méthode d'analyse morphologique approfondie qui permet de déterminer pour chaque mot sa racine suivie de la liste de toutes ses caractéristiques morphologiques possibles.

En tenant compte des différents types d'ambiguïtés (l'agglutination, l'affixation, la transformation, etc.), le système permet l'analyse des différents types de mots arabes (les noms, les verbes, les adjectifs, les particules, etc.).

L'approche repose sur cinq étapes, à savoir :

1. la segmentation du texte en mots qui est réalisée par le système STAr ;

2. le prétraitement qui consiste à supprimer les proclitiques et les enclitiques pouvant être agglutinés au mot, le mot restant sera filtré pour tester s'il s'agit d'une particule, d'un nom propre, d'un nombre ou d'une date ;

3. l'analyse affixale qui a pour objectif de reconnaître les éléments de base qui entrent dans la constitution d'un mot à savoir, la racine (R) ou la forme canonique et les affixes (préfixe (P), infixe (I) et suffixe(S)) ;
4. l'analyse morphologique qui consiste à déterminer, à partir de la forme (R, P, I, S) obtenue pour chaque mot, toutes ses caractéristiques morphosyntaxiques possibles (i.e. POS, genre, nombre, cas, personne, etc.) ;
5. le post-traitement qui a pour but de détecter les cas d'ambiguïté. Il s'agit de vérifier pour chaque couple de mots successifs ayant la catégorie NOM, s'il représente une agglutination.

Le système MORPH2 est un analyseur morphologique basé sur la méthode proposée et permettant l'analyse morphologique de textes arabes non voyellés. Ce système est réalisé avec le langage de programmation JAVA et utilise un ensemble de données nécessaires pour chaque étape d'analyse. Le résultat de l'analyse est enregistré dans un fichier XM (Belguith et al, 2004).

L'analyseur morphologique MORPH2 a été évalué sur deux corpus : le premier corpus est un livre scolaire tunisien composé de 81 textes arabes non voyellés contenant 29 188 mots. Le second corpus est un extrait du web qui représente un ensemble d'articles de journaux sur divers thèmes contenant 22 216 mots.

Les résultats de l'évaluation montrent que l'analyseur MORPH2 a pu analyser avec succès presque 74 % des noms propres du premier corpus et 62% des noms propres du deuxième corpus. Les cas d'échecs pour les noms sont dus essentiellement à l'absence de la forme canonique du nom dans le lexique et aux transformations des voyelles longues et de la lettre hamza. Pour le deuxième corpus (articles de journaux), les fautes d'orthographe, les omissions de «chadda » et l'absence du hamza sont les principales causes de l'écart avec la mesure du premier corpus.

Un autre système d'analyse robuste de textes arabes non vocalisés est celui de MASPAR[7] (Belguith *et al*., 2006), basé sur une approche multi-agent. Il comporte six agents qui sont

[7] *Multi-Agents Système for Parsing Arabic.*

représentés par les phases fonctionnelles suivantes : «Segmentation», «Lexique», «Morphologie», «Syntaxe», «Anaphore» et «Ellipse».

L'objectif principal de la phase "Morphologie" consiste à déterminer pour chaque mot décomposé en racine et affixes, la liste de ses caractéristiques morphosyntaxiques suivant une méthode d'étiquetage qui repose sur trois étapes :

1- Identifier la/les catégorie(s) grammaticale(s) du mot ;

2- Déterminer pour chaque catégorie grammaticale identifiée à l'étape 1, la liste de ses caractéristiques morphologiques candidates ;

3- Réaliser un filtrage des liste candidates résultants de l'étape 2 (un enrichissement du lexique de base par des Formes Canoniques (FC) non verbales, la forme de masculin singulier /ou celle de féminin singulier, associés aux mots ambigus à cause de l'absence de voyelles).

Un échec dans la reconnaissance des caractéristiques morphosyntaxiques de certains mots provoque une analyse partielle de 16,5% des phrases du corpus de test et un taux de 22,5% des phrases mal analysées.

Parmi leurs perspectives, Belguith *et al* envisagent une modélisation plus fine du système par l'attribution de plus d'un agent à chaque phase afin d'améliorer la qualité de solutions trouvées. L'enrichissement de base des connaissances lexicales et morphologiques de MASPAR est aussi envisagé.

Un autre analyseur morphologique fonctionnel a été proposé par Smrž (2007), ELIXIRFM qui représente une implémentation inspirée par la méthodologie de morphologie fonctionnelle arabe (Forsberg *et al*. 2004) ; il consiste en deux composants principaux, une bibliothèque de programmation HASKEL[8] et un lexique linguistique morphologique. Une interface en ligne[9] de ElixirFM est à la disposition de l'utilisateur.

[8] www.haskell.org.
[9] http://quest.ms.mff.cuni.cz/elixir/

Le lexique est dérivé de la source ouverte Buckwalter[10] (Buckwalter, 2002) et a été amélioré par des informations provenant des annotations syntaxiques de dépendance du Prague arabe Treebank (Hajic *et al*., 2004).

Ce système est fondé sur un ensemble de règles morpho-phonémiques (alternation de Ta-Marbuta). L'analyseur morphologique ELIXIRFM permet, étant donné une forme fléchie en ASM, d'en extraire le lemme et la racine. ELIXIRFM fournit à l'utilisateur quatre modes de fonctionnement différents :

1) «Resolve» qui offre une segmentation[11] et une analyse morphologique du texte inséré (éventuellement en Unicode, Buckwalter ou ArabTeX notations). Chaque item lexical reçoit sa propre étiquette (POS) et ses analyses distinguées ;
2) « Inflect» qui produit toutes les formes fléchies possibles pour un lexème donné ;
3) «Derive» qui transforme les mots en leurs homologues ayant un sens similaire et des catégories grammaticales différentes. Les formes des mots sont codées à l'aide des modèles morpho-phonémiques ;
4) «Lookup» qui permet de rechercher pour un lexème souhaité, ses lemmes, sa racine et ses dérivées, comme un dictionnaire en ligne.

Si on prend l'exemple du nom propre باريس /Paris, la fonction «Resolve» effectue une segmentation de trois tokens, et pour chaque token (item lexical), il produit son POS, son lemme, sa racine, son modèle morphémique (ses morphes) et ses traductions en anglais. Le tableau 1.1, ci-dessous, illustre les résultats obtenus avec cette fonction ; de même, la fonction «Derive» engendre les 3 dérivées indiquées dans le tableau 1.2, et la fonction «Inflect» génère leur liste de formes fléchies.

[10] Il contient environ 40.000 entrées qui sont regroupées dans environ 10.000 entrées imbriquées. (Buckwalter, 2002**)**

[11] La représentation extérieure d'ELEXIRFM comprend une décision principale de découpage qui suit les conventions de l'ARABE Treebank PENN et la DÉPENDANCE PRAGUE ARABE Treebank.

Token	POS	Lemme	Modèle morphemique	Traduction en anglais
bi / بِ	P	bi / بِ		"by" - "with"
'irrīs / إرِّيس	N	r s / أرس	FiCCiL	"peasant", "farmer"
arīs /أرِيس	N	r s / أرس	FaCiL	"peasant", "farmer"

Tableau 1. 1 : Les résultats obtenus avec la fonction «Resolve» pour le nom propre Paris (باريس)

Dérivée	Son étiquette (POS)
باريسي/ bārīsīy/"Parisian"	A (adjective)
باريسي/ bārīsīy/"Parisian"	N (nom masculin)
باريسية / bārīsīyat /"Parisian"	N (nom féminin)

Tableau 1. 2 : Les trois dérivées du nom propre Paris (باريس) engendrées par la fonction «Derive»

Une particularité de ELIXIRFM est qu'il représente ses items lexicaux dans une représentation phonémique, qui est ensuite convertie en une chaîne de caractères dans la notation ARABTX qui peut elle-même être convertie en transcription orthographique ou phonétique arabe.

De plus, l'utilisation de modèles morpho-phonémiques dans la conception de ELIXIRFM permet d'éviter la définition des règles de transformation comme en MAGEAD (Habash et al, 2005), et le listage de formes de surface pour chaque entrée lexicale comme en BAMA (Buckwalter, 2004). Cependant, la taille de ces modèles est de 4 290 entrées, ce qui introduit le problème de la couverture.

En outre, l'utilisation du lexique du code source ouvert Buckwalter hérite des inconvénients du système, tels que le manque de spécification des mots polylexicaux et la définition de règles d'orthographe qui sont inappropriées.

Ici, nous présentons MAGEAD (Altantawy et al, 2010) en tant qu'analyseur morphologique et générateur pour l'arabe standard moderne (ASM) et ses dialectes. MAGEAD a été introduit

dans des travaux pour l'ASM et des verbes arabes du Levant[12]. (Habash et al. 2005 ; Habash and Rambow, 2006). Le système est un analyseur morphologique et générateur fonctionnel, c'est-à-dire qu'il effectue une analyse morphologique profonde.

Partant d'une forme nominale de l'arabe, il en fait l'analyse sous la forme d'une racine, d'une classe et de traits morphologiques.

Ce travail nous apprend que la morphologie des noms arabes est plus complexe que celle des verbes à cause de la prévalence de formes fléchies irrégulières dans le système nominal arabe. Egalement, le même morphème de surface peut avoir différentes fonctions morphologiques selon le contexte : par exemple, le morphème Ta-marbuta/ة généralement associé au féminin singulier comme dans le mot شجرة / /šjrh/ 'arbre', peut apparaître dans le pluriel de certains noms masculins (أنظمة /ÂnDˇm_h / systèmes).

Le système comporte trois phases principales :

- la création manuelle de ressources linguistiques qui sont utilisées pour une instance spécifique de MEGEDA. Les ressources sont : La hiérarchie de la classe de comportement morphologique[13](MBCH), la grammaire hors contexte (CFG),[14] les règles[15] orthographiques/ et morpho phonémiques et le lexique ;

- la compilation de ces ressources linguistiques pour produire deux transducteurs à états finis (un pour la production et l'autre, son inverse, pour l'analyse) ;

- l'utilisation bidirectionnelle de MAGEAD pour l'analyse /et la génération morphologique.

---

[12] Le syro-libano-palestinien ou arabe levantin septentrional est une variété d'arabe dialectal parlé en Syrie, au Liban, et dans certaines régions urbaines de Palestine.

[13] Fonction qui associe des morphèmes à des traits linguistiques.

[14]«Context free gramme : En linguistique et en informatique, une grammaire non contextuelle, grammaire hors contexte ou grammaire algébrique est une grammaire formelle dans laquelle chaque règle de production est de la forme X\to w où X est un symbole non terminal et w est une chaîne composée de terminaux et/ou de non terminaux. », selon le site : https://fr.wikipedia.org/wiki/Grammaire_non_contextuelle

[15] MAGEAD a deux types de règles : les règles morpho-phonémiques /et phonologiques qui transforment des représentations morphématiques en représentations phonologiques et orthographiques, et les règles orthographiques identifiant la représentation orthographique.

La phase de génération morphologique se compose de 4 représentations :

1. Une représentation profonde sous la forme d'une racine, d'une classe, appelée MBC (pour Morphologic Behavioural Class) et de traits morphologiques ;

2. Associer des traits morphologiques aux morphèmes abstraits[16] en utilisant la classe comportement morphologique « MBC » ;

3. Représenter les morphèmes abstraits sous la forme de morphèmes concrets correspondants[17] ;

4. passer de la représentation en morphèmes concrets à la représentation de surface par la concaténation des affixes et l'application des règles morpho phonémiques.

Une évaluation détaillée de MAGEAD a été effectuée en le comparant à l'analyseur morphologique couramment utilisé SAMA (Graff *et al*., 2009) qui est la version 3.1 de BAMA (Buckwalter, 2004). La mise en œuvre montre une bonne couverture et la facilité d'utilisation avec une grande précision. Un travail sur la création d'un lexique amélioré et la mise en œuvre d'un système nominal de MAGEAD pour les dialectes arabes a été envisagé.

Enfin, cette étude concernant les analyseurs morphologiques des textes arabes (A. Hamadi, 2012) a montré que la prise en compte des diacritiques lors d'une analyse morphologique peut réduire le taux d'erreurs de cette analyse ; autrement dit, l'accroissement du taux des diacritiques dans le texte traité provoque une réduction du taux d'erreurs résultant de l'analyse du texte.

La méthode proposée consiste à représenter l'entrée fournie à MADA[18] sous un automate à état fini, ainsi que les solutions sortant de MADA elles-mêmes représentées par des automates à états finis. Cette représentation facilite les tests de compatibilité entre l'entrée et les sorties.

---

[16] Les morphèmes abstraits sont des morphèmes qui pourront se réaliser différemment dans les différentes variétés de l'arabe.

[17] Il s'agit de remplacer les symboles 1, 2 et 3 du schème par le premier, second et troisième symbole de la racine. Les symboles V sont quant à eux remplacés par les symboles qui constituent le vocalisme.

[18] MADA (*Morphological Analyzer and Disambiguator of Arabic*) (Habash, 2005), est un analyseur morphologique de l'arabe qui ne tient pas compte des diacritiques.

Deux automates sont compatibles si leur intersection est non nulle, c'est-à-dire, s'il existe un chemin commun entre eux de l'état initial à l'état final.

Une évaluation sur 10 corpus de test contient un pourcentage de diacritiques entre 10% et 100%. Les performances de l'analyse morphologique passent de 84,25% à 95,59%, si MADA prend en considération les diacritiques présents dans les textes entièrement vocalisés. Les résultats obtenus prouvent un apport considérable de la diacritisation[19] dans l'analyse morphosyntaxique de l'arabe.

2 La reconnaissance d'entités nommées dans les textes arabes

Il s'agit des travaux nous informant que le marquage et l'extraction des entités nommées, en particulier de noms propres, est une clé importante pour améliorer l'efficacité des différents systèmes de traitement automatique de langues comme le système question-réponse.

Par exemple, Abuleil (2004) a développé une nouvelle technique pour extraire les noms propres dans les textes arabes par la construction d'une base de données et des graphes pour représenter les mots dans ces phrases ainsi que les relations entre ces mots ; les mots sont représentés par les nœuds du graphe et les liens entre ces mots représentent leurs relations. Le système opère par :

- marquage des phrases nominales ;
- application des règles pour extraire les noms, les règles sont fondées sur les mots clés et certains verbes spéciaux ;
- classement des noms propres en fonction de deux types : grande classe (personne, lieu, organisation et événement) et sous-classes spécifiques (président, pays, journal, guerre, etc.).

Des dizaines de mots clés et de verbes spéciaux ont été recueillis dans un projet de recherche antérieure (Abuleil et Evens, 2002) et ont été classés dans les différentes catégories.

[19] C'est l'opération qui consiste à attribuer des diacritiques aux lettres des mots non vocalisés.

L'implémentation s'appuie sur deux types de relations [20] entre les mots (nœuds) et cinq types de classe de mots (nœuds) dans le graphe.

La relation entre deux mots est obtenue par le calcul du nombre de fois où ces deux mots apparaissent de manière consécutive dans le texte. La technique a été testée sur un corpus de 500 articles, 78,4% des noms propres sont successivement classés.

Une mauvaise classification de certains noms peut être générée pour diverses raisons :

- un nom de personne apparaît dans une phrase qui contient à la fois un titre et un nom d'organisation ou un nom de lieu ;
- de nombreux titres apparaissent avec le même nom de personne ;
- plusieurs organisations différentes (université, centre, banque, etc.) utilisent le même nom.

Dans ce contexte, MESFAR (2008) présente un système de reconnaissance automatique d'entités nommées dans les textes arabes qui a été construit en utilisant les informations utiles fournies par l'analyse morphologique (noms de professions (ex : طبيب /tabiyb/docteur) considérés comme des marqueurs lexicaux de noms de personnes).

Cet outil est basé sur des règles utilisant les preuves internes[21] et externes[22], ainsi que des dictionnaires de mots déclencheurs pour l'identification et la catégorisation des entités nommées. Les règles sont représentées par des grammaires locales de Nooj (écrites sous forme de RTN)[23] utilisant des informations morphosyntaxiques et sémantiques présentes dans le dictionnaire ElDicAr. [24]

---

[20] Relation forte : entre les mots qui forment un certain nom propre ; Relation faible : entre les mots et les noms communs.

[21] Ils sont fournis par les constituants de l'entité nommée. Ils se trouvent dans les listes de mots déclencheurs / noms propres.

[22] Ils sont fournis par le contexte dans lequel l'entité nommée se trouvait. Ils se basent sur les relations syntaxiques au sein d'une phrase contenant l'entité à traiter (résultant de l'analyse morphologique).

[23] Recursive Transition Network (réseau de transitions récursif).

[24] «Permet de rattacher l'ensemble des informations flexionnelles, morphologiques, syntactico-semantiques à la liste des lemmes. Les routines de flexion et dérivation automatique à partir de cette liste produisent plus de 3 millions de formes fléchies. Ce dictionnaire joue le rôle de moteur linguistique pour l'analyseur morphosyntaxique. » (S.Mesfar, 2008)

Pour l'identification de noms propres (personnes, lieux et organisations), une catégorisation est soigneusement établie (par exemple : pour les noms de lieux, Mesfar a repris la classification faite dans [Piton et Maurel, 2004] classant comme un nom de lieu/toponyme : les pays, villes, fleuves, montagnes, îles, états, etc. Pour chaque classe, il a appliqué les étapes suivantes :

- élaboration de dictionnaires (par exemple : un dictionnaire de noms de villes comme : وهران- wahraān - Wahrân) ;
- recensement des listes des marqueurs lexicaux (par exemple : **مدِينة** (madiyna - ville)) ;
- utilisation de ces dictionnaires et listes pour la description des règles de reconnaissance au sein des grammaires locales.

On a abouti à une couverture lexicale de plus que 95% sur un corpus du journal « Le Monde Diplomatique » dans sa traduction arabe.

Dans leurs travaux, Shihadeh *et al*. (2010) ont implémenté ARNE[25], un outil pour la reconnaissance d'entités nommées de texte arabe. Ce travail a pour objectif de présenter un logiciel pipeline[26] de reconnaissance de l'entité nommée arabe (ARNE).

Le système comprend une segmentation, une analyse morphologique, la translitération de Buck Walter et le POS de l'entité nommée (personne, lieu, organisation). On distingue deux méthodes: la première est une méthode simple, rapide et multilingue[27] qui utilise une liste d'entités nommées connues[28]. Si un mot est un élément dans cette liste, alors il est étiqueté comme une entité nommée, autrement non. La deuxième approche est le système ARNE qui consiste en un prétraitement en quatre étapes avant de passer à l'étape de REN. Ces quatre étapes sont :

1. La segmentation et l'analyse morphologique du texte d'entrée en utilisant le système ElixirFM développé par Smrz (2007). L'entrée de cette étape est un fichier de texte arabe.

[25] « *Arabic named entity recognition* ».
[26] L'avantage d'un tel modèle est que la sortie d'un élément est l'entrée du suivant, ce qui permet d'utiliser différentes ressources et différentes informations pour la reconnaissance des entités nommées (Shihadeh *et al,* 2010).
[27] La possibilité d'utiliser le même Système de REN pour toute autre langue par l'échange de la liste utilisée.
[28] ANERgazet : Trois listes de noms propres : 1. liste de 2 309 de noms de personnes ; 2. liste de 1 950 noms de lieux ; 3. liste de 262 noms d'organisations ; ces listes sont développées par Benajiba *et al.*

Ce fichier passe à ElixirFM qui délivre un texte contenant six colonnes qui comportent les informations suivantes :

- ✓ les unités lexicales ;
- ✓ ses notations ArabTEX qui indiquent la prononciation et l'orthographe ;
- ✓ ses translitérations de Buckwalter en fonction de sa prononciation en colonne 2 ;
- ✓ ses traits morphologiques ;
- ✓ ses positions dans le dictionnaire ElixirFM ;
- ✓ ses traductions anglaises.

2. La translitération du texte arabe[29] en utilisant le logiciel Encode arabe développé par Buckwater. L'entrée de ARNE dans cette étape est la sortie de la première étape ; la sortie est un texte en 4 colonnes :

- ✓ les positions des unités lexicales dans ses phrases ;
- ✓ les unités lexicales ;
- ✓ les translitérations Buckwater ;
- ✓ Les numéros de blocs ExilirFM.

3. L'étiquetage[30] consiste à attribuer son POS à chacune des unités lexicales identifiées. L'entrée de cette étape est le texte transcrit résultant de la deuxième étape et la sortie est ce même texte avec une colonne de plus qui contient les POS des unités lexicales.

4. La reconnaissance d'entité nommée (REN) consiste à rechercher (look up) les unités lexicales dans ANERgazet et à étiqueter celles qui sont des entités nommées en utilisant la méthode BIO-étiquetage. L'entrée de ARNE dans cette étape est le texte sorti de l'étape

[29] Pour les lecteurs qui n'ont pas la capacité de lire le script arabe, ils peuvent le lire en latin.
[30] Cette tâche n'a pas encore été mise en œuvre. Shihadeh *et al* vont intégrer leur propre tagger base-SVM qui est basé sur (Gimenez et Marquez, 2004) ; dans cet article, la valeur de POS par défaut est "NULL''.

précédente et la sortie est ce même texte avec une colonne de plus qui contient les entités nommées étiquetées par le BIO-étiquetage.

L'évaluation du système ARNE a retourné une f-mesure de 32,5 %, les petites dimensions des listes utilisées sont la cause d'un tel résultat. Beaucoup d'entités nommées pourraient ne pas être reconnues par ARNE parce qu'elles ne font pas partie de listes ANERGazet. Cependant, l'utilisation d'informations morphologiques sortant de l'analyseur ExilirFM a amélioré le taux de f-mesure de 32,5% à 33,7%. Par exemple, ARNE ne peut pas reconnaître le groupe nominal وسوريا[31] /wswrya/'' et Syrie'' à cause de l'agglutination, avant d'utiliser l'information donnée par ElixirFM qui peut trouver si un groupe nominal inclut une agglutination ou non.

D'autres approches ont été étudiées au début de notre recherche afin d'approfondir nos connaissances en ce domaine comme Zribiet *et al*. (2010) qui proposent une méthode d'apprentissage[32] hybride pour l'extraction des entités nommées dans les textes arabes.

La méthode est fondée sur l'utilisation d'un ensemble de règles pour extraire et classer les entités nommées. L'approche consiste en trois phases principales, chaque phase est précédée par une étape de préparation qui contient deux modules :

1. Une analyse morphologique vise à déterminer les caractéristiques morphosyntaxiques (genre, nombre, type, etc.) de chaque mot du corpus et à remédier aux erreurs engendrées par la nature agglutinative de la langue arabe ;

2. Un module d'extraction des attributs se base sur des listes de noms propres et sur l'ensemble des caractéristiques morphologiques afin de définir les valeurs des attributs associées à tout mot dans le corpus.

Les trois phases principales de cette approche sont :

[31] Le groupe nominal'' وسوريا /wswrya/et Syrie '' n'est pas dans le lexique utilisé.
[32] Ils ont exploité le corpus ANERcorp2 qui est formé de 150.000 mots extraits d'un ensemble d'articles de presse de types variés. Il contient 10.519 EN de type lieu, organisation et personne. L'apprentissage a été réalisé sur 76.608 mots contenant 5.192 EN. Pour la validation des règles, ils ont utilisé 38.488 mots contenant 2.778 EN et le reste du corpus a été utilisé pour tester la performance système.

1) La génération automatique d'un ensemble de règles à l'aide d'un algorithme d'apprentissage[33] ; ce sont des règles capables de détecter les mots qui composent l'EN selon leur type ; elles permettent la classification des mots selon trois types

- B-TYPE (Begin) : si un mot constitue le premier syntagme d'une EN ;
- I-TYPE (Inside) : si un mot appartient à une EN ;
- O (Out) : si un mot n'appartient pas à une EN.

2) L'extraction manuelle d'un autre ensemble de règles pour la validation de celles générées automatiquement dans la première phase ;
3) L'extraction des entités nommées par l'application des règles générées automatiquement et validées issues de la deuxième étape, et la correction des erreurs survenues.

Par exemple : l'application des règles n'a pas pu reconnaitre le mot « العمال » (ouvriers) /AlEmAl/ comme une partie de l'EN «حزبِ العمال الكردستاني» (Parti des ouvriers Kurdistan) /Hzb AlEmAl AlkrdstAny/, on obtient l'annotation (B-ORG O I-ORG) au lieu de (B-ORG I-ORG I-ORG). Pour corriger l'erreur dans l'annotation (B-ORG O I-OR), il faut appliquer la règle : « classe mot= O and classe_mot-1=B-ORG and classe_mot+1= I-ORG => classe mot=I-ORG » ;

Cette règle signifie que si un mot porte l'étiquette O et s'il est précédé par un mot qui porte l'étiquette B-ORG et suivi par un mot qui porte l'étiquette I-ORG, alors, ce mot appartient à une EN de type organisation et l'étiquette O sera remplacée par l'étiquette I-ORG.

Les résultats de l'évaluation de la méthode proposée sont encourageants ; un taux global de F-mesure[34]égal à 79,24% a été obtenu. Ils ont identifié plusieurs erreurs dont la présence a provoqué l'échec de l'identification et de la classification de quelques ENs :

- des erreurs dues à la segmentation des mots avec l'analyse morphologique (Certains noms propres étrangers ressemblent à des mots agglutinés lorsqu'ils sont translittérés en arabe). Par exemple, le nom propre « Virginie » sera translittéré en «فيرجيني » (et il espère de moi) qui désigne un verbe auquel s'attache une conjonction de coordination « ف » ;

[33] L'algorithme d'apprentissage des règles RIPPER (Cohen, 1995).

[34] Mesure populaire qui combine la précision et le rappel par leur pondération, nommée F-mesure (soit Fmeasure en anglais) ou F-score. Selon le site : https://fr.wikipedia.org/wiki/Pr%C3%A9cision_et_rappel#F-mesure

- des erreurs dues à des insuffisances dans le corpus d'apprentissage[35] .

Citons aussi Saddane (2013) qui a proposé un nouveau système d'extraction de connaissances[36] à partir d'un texte arabe. Ce système est capable de repérer les entités nommées et les relations sémantiques qui les relient en se basant sur une ontologie métier[37].

Le processus d'extraction s'appuyait sur une analyse morphosyntaxique profonde qui est fondée sur la technologie des automates d'états finis. Le résultat de cette analyse est exploité par des règles d'extraction sémantique pour produire des informations concernant un même événement.

Le système de REN est basé sur des règles linguistiques qui exploitent l'étiquetage syntaxique, des déclencheurs et des dictionnaires de noms propres.

Pour mettre en avant l'efficacité de son approche, Saddane a comparé les performances de la phase de segmentation avec l'outil de Stanford[38] en apportant des modifications[39] aux résultats pour pouvoir les comparer avec son système ; elle a obtenu une précision de 0,98% contre 0,96% avec l'outil de Stanford.

Une évaluation du système d'extraction d'entités nommées a été réalisée sur le corpus ANER [40](Benajiba et al., 2007) qui est composé de 150 000 occurrences de mots (personne, lieu, organisation, divers). Une précision de 89,05% pour la détection des entités nommées de type personne, 91% pour les lieux et 83,41% pour les organisations.

Dans ce même contexte, nous présentons un module arabe libre sous Unitex[41]. Afin d'implémenter ce module, Doumi *et al.* (2013) ont ajoutés à Unitex les composants suivants :

---

[35] La génération des règles de REN se base sur la structure et le contexte de l'EN. Si une EN apparaît dans un contexte différent de celui du corpus d'apprentissage, alors elle ne sera pas reconnue.
[36] Ce système reconnaît les mots simples, les mots composés, les expressions idiomatiques, les entités nommées et les relations syntaxiques dans les phrases.
[37] Un domaine particulier modélisé dans une ontologie.
[38] http://nlp.standford.edu/projects/arabic.shtml
[39] Par exemple, dans l'outil de Stanford, l'article défini (ال/al/le ou la) fait partie du mot, par contre, le segmenteur de Saddane le considère comme un token indépendant.
[40] http://users.dsic.upv.es/~ybenajba/downloads.html
[41] http://www-igm.univ-mlv.fr/~unitex

- ✓ l'alphabet arabe d'Unicode[42] ;
- ✓ un corpus libre de droits (32 pages) composé de 18 261 formes simples et partiellement voyellé[43] ;
- ✓ un jeu de 17 catégories et 21 traits morphosyntaxiques a été choisi après une étude approfondie comparative des jeux d'étiquettes des projets sur le TAL arabe.

Dans ce travail, l'objectif principal est la construction du lexique arabe représenté par des dictionnaires DELAS et DELAF[44] et divisé en deux grandes catégories :

1) les catégories syntaxiques fermées contenant les verbes d'état[45], les adverbes de lieu/temps, les pronoms et les particules ;
2) les catégories syntaxiques ouvertes contenant les noms communs, les adjectifs, les verbes.

Deux algorithmes de génération automatique de graphes de flexion ont été conçus, le premier produit les paradigmes flexionnels de verbes ; le second a pour but de générer ceux des noms (un paradigme est représenté par un graphe calculé).

Les deux algorithmes ont le même principe ; l'algorithme général de flexion des noms consiste à faire parcourir le corpus de test par un expert linguiste à travers une interface graphique et pour chaque mot, il s'agit de :

1. lemmatiser et introduire le lemme au programme ;
2. calculer son patron flexionnel ;
3. rechercher ce patron dans la liste des patrons de noms qui existent, s'il est trouvé, on l'ajoute au DELAS, sinon le linguiste détermine sa classe et son pluriel et son féminin en cas d'irrégularité ;
4. utiliser sa classe choisie, un graphe modèle et des affixes pour générer son graphe de flexion, ce graphe sera appliqué sur le lemme pour produire ses formes fléchies. Le graphe

[42] http:/www.unicode.org/charts/PDF/U0600.pdf
[43] Il contient seulement les diacritiques de nounation (tanwine) et le signe de germination (chadda).
[44] Dictionnaires DELA de mots simples de formes fléchies.
[45] Les lemmes de verbes d'état (verbes particuliers) sont au nombre de 13.

calculé représente le paradigme d'un ensemble de mots, un petit algorithme a été conçu pour détecter les mots ayant le même paradigme.

Le système de génération automatique de graphes a donné des résultats encourageants. Les graphes obtenus par l'algorithme de flexion de verbes engendrent 264 formes verbales fléchies en utilisant 5 thèmes ; ceux de noms communs/adjectifs ont produit 63 formes fléchies nominales avec un seul thème. Le module couvre plus de 95% des catégories fermées ; les dictionnaires graphes établis seront développés par la construction de dictionnaires de polylexicaux et des dictionnaires de noms propres afin d'utiliser Unitex pour des applications de haut niveau, telles que la reconnaissance des entités nommées dans les textes arabes.

Pour terminer cette classe, Ben Mesmia et al (2017) ont proposés l'outil CasANER pour reconnaître et annoter les entités nommées arabes.

Pour réaliser CasANER, les auteurs ont tout d'abord établi une hiérarchie de cinq catégories principales : date, personne, lieu, organisation et évènement (chacune d'entre elles est divisée en sous catégories), en s'appuyant sur deux corpus[46] de la Wikipédia arabe.

Puis, ils ont implémenté une méthode qui vise à la création d'un ensemble de dictionnaires, à partir d'une liste des règles d'extraction qui dépend essentiellement des mots déclencheurs et d'un ensemble de transducteurs typiques permettant la reconnaissance de différentes catégories d'entités nommées arabes précédemment définies.

En fait, le système CasANER est composé de deux types de cascades de transducteurs : l'analyse et la synthèse, qui sont mises en place via la plateforme linguistique libre Unitex en utilisant l'outil CasSys. La cascade d'analyse regroupe les transducteurs qui assurent les phases d'analyse, de filtrage et de marquage générique. La phase de filtrage est destinée à rectifier les chemins des transducteurs d'analyse afin d'avoir des ENA reconnues structurées (et annotées). Cependant, la phase de marquage générique aide à améliorer les performances du système[47].

---

[46] Ils ont expérimenté le système CasANER à travers deux corpus : un corpus de test qui contient des fichiers texte de 95 378 tokens, et un corpus d'étude contenant des fichiers texte pour un ensemble de 146 000 tokens en permettant la création de nouveaux dictionnaires et ainsi que la mise à jour des dictionnaires disponibles sous la plate-forme Unitex. Les deux corpus sont collectés à partir de la Wikipédia arabe avec l'outil Kiwix: http://wiki.kiwix.org/wiki/Main_Page.

[47] Selon (Ben mesmia et al, 2017), l'utilisation d'un transducteur de marquage générique permet d'éviter des problèmes d'ambiguïté, par exemple, pour la catégorie «personne », il faut deviner si un mot est un prénom ou un nom de famille et non un adjectif ou un nom commun.

La cascade de synthèse regroupe les transducteurs qui permettent de transformer l'annotation des ENA reconnues en une annotation TEI, ce qui permet de générer un corpus de sortie structuré, utile pour plusieurs applications de TAL.

Enfin, l'efficacité de CasANER est évaluée[48] sur le corpus annoté ANERcorp en démontrant que les resultats fournis par leur système sont plus efficaces que ceux de ANERsys dans la reconnaissance et l'annotation d'entités nommées des catégories personne et organisation. Cependant, les ANERsys peuvent reconnaître et déclarer les ENA possédant la catégorie lieu avec de meilleurs résultats que le système. CasANER. Dans des travaux futurs, Ben Mesmia et al (2017) exploiteront la sortie de système CasANER pour extraire les relations pertinentes entre les ENA reconnues et annotées afin de créer un dictionnaire électronique des entités nommées arabes.

3 La gestion des bases de données lexicales multilingues

Les bases de données lexicales multilingues sont devenues le noyau central de différents types d'applications du TAL, notamment la recherche d'informations et la traduction automatique. Cependant, la gestion de ces ressources linguistiques soulève toujours des problèmes et nécessite une amélioration permanente pour assurer un bon fonctionnement et une crédibilité des programmes qui les utilisent.

Plusieurs études ont démontré l'importance de la gestion des bases de données lexicales multilingues, par exemple, les travaux de la thèse de Ying ZHANG(2016) qui commence par une étude détaillée de l'évolution des idées en lexicographie computationnelle de 1980 à 2012. Nous citons ici un exemple de l'évolution vers des bases lexicales à partir des années 1991, à savoir, la construction des bases de données lexicales symétriques grâce aux notions de lexie (sens de mot dans un dictionnaire) et d'acception (sens de mot en usage)[49] ; puis la création de

[48] L'évaluation comprend trois étapes. La première consiste à supprimer les tags existants dans ANERcorp pour récupérer le corpus initial. La deuxième étape est l'application de cascade de l'analyse sur l'ANERcorp initial pour fournir un corpus annoté. Finalement, les ENA reconnues dans le nouveau corpus sont annotées à l'aide de {}. Par exemple, à l'issue de la première étape l'ENA "فرانكفورت / Francfort" a été annoté comme [فرانكفورت B-LOC] et il est transformé en <LOC> فرانكفورت </ LOC>.

[49] « le mot français bleu correspond à plusieurs lexies : bleu_nm#couleur, bleu_nm#fromage, bleu_nm#contusion, bleu_adj#couleur, bleu_adj#cuisson, etc. »,

PARAX[50], dans laquelle, il y a 5 volumes monolingues (français, japonais, chinois, espagnol et russe), chacun d'entre eux contant des entrées (mots). Chaque mot est relié à une ou plusieurs lexies (sens), et chaque lexie est reliée à un UW (Universal Word)[51].

L'objectif initial de travaux de Ying ZHANG (2016) a été l'amélioration de la base de données lexicale PIVAX[52]. Tout d'abord, JIBIKI, la plate-forme sous-jacente pour la construction et la gestion de bases de données lexicales, a été rénové en adaptant Pivax-1 en Pivax-2 ; il s'agissait d'augmenter la vitesse, la sécurité, et de transformer Pivax en un serveur lexical opérationnel. Par conséquence, PIVAX-2 a été utilisé pour mettre à disposition sur le Web toutes les ressources lexicales mises dans PIVAX-2 par le projet ANR TRAOUIERO[53].
Egalement, Ying ZHANG(2016) a traité les problèmes de gestion d'acronymes en constituant Pivax-3, un prototype de mettre en correspondance ces termes en employant la notion de liens riches relative à JIBIKI-2, en utilisant la notion de PROLEXÈME (Tran, Maurel, 2006) et en créant des nouvelles notions appelées PROAXIE.
De plus, le manque de services génériques lexicaux peut poser plusieurs difficultés dans les bases de données multilingues comme le délai de temps de réponse pour un grand nombre de requêtes. En réalité, la présence de ces outils est important pour effectuer une bonne gestion de certains travaux, entre autres, la lemmatisation et la création de mini-dictionnaires ; pour cela, Ying ZHANG(2016) a mis en place le service générique LEXTOH (Intergiciel de lemmatisation), sans lequel, il fallait programmer un service de lemmatisation pour chaque système et chaque langue. Un outil de création de mini-dictionnaire (CREATDIC) a été aussi implémenté lors de ce travail de thèse. Parmi les perspectives envisagées, on peut souligner l'amélioration de LEXTOH et de CREATDICO et l'importation de plusieurs ressources à la

[50] Le but de PARAX était la création d'une base de données lexicales multilingues à acceptions interlignes basée sur HYPERCARD (un programme et un environnement de programmation développé par Apple qui ne fonctionne que sous Mac OS versions 9 et précédentes) et qui a été convertie en REVOLUTION, puis en LIVECODE où REVOLUTION est un descendant d'HYPERCARD, porté sur les plates-formes usuelles (MacOs, Unix, Windows), et s'appelle désormais LIVECODE.
[51] « terme dénotant les unités lexicales d'UNL (Universal Networking Language), qui sont des "lexèmes interlingues ».
[52] C'est une base lexicale à pivot par acceptions monolingues et interlingues pour la mise en commun de ressources lexicales ouvertes et propriétaires pour la TA, réalisée par Hong Thai Nguyen dans le cadre de sa thèse [Nguyen, H.-T., 2009].
[53] TRAOUIERO : TRAduction : Outils Unifiés, Intégrables, Embarquables, et Ressources Opérationnelles (Projet ANR-emergence).

base de données du prototype PIVAX-3 comme la ressource complète de PROLEXBASE, la grosse base CJK, les listes d'abréviations en multilingue de la WIKPÉDIA1 (abréviations en informatique, en médecine, abréviations militaires, etc.).

Dans le même contexte, il convient d'évoquer la recherche d'Andon Tchechmedjiev (2016) qui est consacrée à l'étude de l'interopérabilité sémantique de ressources lexico-sémantiques[54] en s'appuyant sur l'alignement des ressources entre elles au niveau des sens de mots. D'abord, il a présenté l'alignement de deux ressources dans la même langue en soulignant deux problèmes fondamentaux :

- il faut déterminer quels éléments sont équivalents entre les deux ressources.
- il faut trouver une équivalence partielle hiérarchique entre les sens des deux ressources lorsque la granularité des distinctions sémantiques est différente.

Ensuite, il a montré que les problèmes de granularité et d'équivalence sémantique partielle augmentent quand les deux ressources sont dans deux langues différentes, et il faut projeter les informations sémantiques des deux langues dans un espace commun où elles sont comparables (traduction, système translingue, équivalences structurelles…).

Dans le cas des alignements de trois ressources ou plus, un paradigme d'alignement doit être défini pour gérer l'ensemble des ressources alignées et garder une solidité entre tous les alignements en présentant deux architectures possibles :

- une architecture par transfert (alignements bilingues entre toutes les paires de ressources), par exemple, les corpus parallèles et les dictionnaires bilingues.
- une architecture par pivot interlingue[55] qui demande à trouver une représentation interlingue commune pour représenter les alignements des sens entre toutes les ressources en même temps, par exemple Babel Net, EuroWordNet ont fait le choix d'utiliser l'anglais comme pivot interlingue.

[54]« Une ressource lexico-sémantique est une ressource langagière à base d'entrées. Chaque entrée correspond à un mot du lexique et est associée à un ensemble de sens. Les sens dénotent les différents usages du mot ayant des significations différentes. L'ensemble des mots et des sens forment un graphe où différents mots et sens peuvent être reliés par des relations lexicales ou sémantiques (synonymie, antonymie, hyperonymie, traduction, etc.). »

[55] Il existe deux types de pivots, le pivot naturel, qui utilise une ressource dans une langue particulière comme pivot (par exemple le projet WordNet) et un pivot interlingue (artificiel) qui ne dépend d'aucune des langues alignées(par exemple le projet Papillon).

De plus, à travers cette thèse, Andon Tchechmedjiev a effectué une analyse détaillée des architectures d'interopérabilité utilisées pour les ressources multilingues en les classifiant en deux catégories principales : ressources construites manuellement et collaborativement, et ressources construites automatiquement.

Dans l'objectif d'assurer l'interopérabilité au niveau sémantique de ressources lexico-sémantiques multilingues par l'utilisation d'un pivot à acceptions interlingues (pivot artificiel), l'auteur a proposé deux algorithmes de construction initiale d'acceptation sur la base du calcul des cliques[56] dans le graphe des alignements bilingues, et des techniques de mise à jour de sens (ajout, suppression, correction) aux hiérarchies d'acceptations.

L'évaluation de ces algorithmes a été réalisée via un cas d'étude pratique, celui de DBNary[57] qui admet l'extraction des différentes éditions de langue du dictionnaire Wiktionary dans un format de données lexicales liées (lemon). En effet, DBNary est interopérable avec les nombreuses autres ressources et outils qui gèrent ce format et permet aussi de réaliser des alignements avec des ontologies et d'autres ressources liées ouvertes.

Parmi les futures travaux considérés, Andon Tchechmedjiev a cité l'ajout d'une représentation de travail de la ressource sans passer par des requêtes SPARQL afin d'optimiser la performance d'accès et l'élaboration des algorithmes pour convertir les ressources existantes (telles que BabelNet) dans une architecture d'alignement par pivot interlingues.

4 L'enrichissement de Prolexbase

Okinina *et al*. (2011) présentent une méthode hybride combinant un ensemble de règles d'extraction et un apprentissage supervisé. L'objectif essentiel de ce travail est de réaliser un enrichissement semi-automatique de Prolexbase depuis la Wikipédia.

[56]« Une clique est un sous ensemble de sommets d'un graphe, qui forment un graphe où tous les sommets sont adjacents deux à deux (un sous-graphe complet) ».

[57] Le projet DBnary vise à extraire des données lexicales structurées en RDF, à partir des différentes éditions de wiktionnaires (www.wiktionary.org). Ces données sont disponibles à http://kaiko.getalp.org/dbnary/.

Il s'agit d'alimenter la base de données par l'ajout de nouveaux noms propres français en grande quantité et avec une grande fiabilité (bruit limité)[58].

La méthode générale consiste à :

1. Un apprentissage binaire par type ; c'est-à-dire, pour chaque type dans Prolexbase (Célébrités, Œuvres, Entreprises, Fêtes, etc.) est constitué un corpus d'apprentissage contenant deux parties : les exemples positifs[59] et ceux appelés négatifs[60] ;

2. Un algorithme séquentiel qui est conçu pour détecter l'appartenance d'un article du corpus d'apprentissage à un type Prolexbase. L'algorithme s'exécute en 4 étapes représentant 4 types de règles de classification :

   - règles sur des infoboxes : ces règles sont performantes pour classifier tous les types d'entités nommées grâce à la précision des infoboxes Wikipédia, cependant, plus de deux tiers des articles Wikipédia ne contiennent pas d'infoboxes ;

   - SVM[61] : si l'article n'a pas été classé à l'aide des règles précédentes, il faut passer à une classification par SVM de trois sortes. Cette méthode utilise les conceptions suivantes :

   1) SVM linéaire qui donne de meilleurs résultats pour les types célébrités et entreprises.

   2) SVM avec noyau RBF (Radial basis function) est utilisé pour les autres types ;

   3) SVM avec d'autres paramètres[62] ;

---

[58] Une forte dégradation des performances a été constatée lorsque les corpus sont bruités. La sélection des éléments pour constituer le corpus d'apprentissage est déterminante et doit être réalisée avec soin. La présence d'homonymie dans Prolexbase nécessite un filtrage pour constituer les exemples positifs du corpus.
[59] Les exemples positifs sont sélectionnés manuellement au sein de l'encyclopédie Wikipédia, ou récupérés par la mise en correspondances EN existantes et non ambigües de Prolexbase et Wikipédia.
[60] Les exemples négatifs sont sélectionnés manuellement au sein de l'encyclopédie Wikipédia (pour les types peu représentés dans Wikipédia) ou par tirage au sort.
[61] « Les machines à vecteurs de support ou séparateurs à vaste marge (en anglais Support Vector Machine, SVM) sont un ensemble de techniques d'apprentissage supervisé destinées à résoudre des problèmes de discrimination et de régression »,
selon le site : https://fr.wikipedia.org/wiki/Machine_%C3%A0_vecteurs_de_support.
[62] Paramètre gamma de contrôle de la forme de l'hyperplan séparateur des classes (RBF kernel) Prise en compte ou non des titres –- Prise en compte ou non du premier paragraphe des articles et des catégories.

- règles sur les titres : ces règles sont utilisées si l'article n'a pas été classé par le SVM ; elles sont moins robustes que celles sur les infoboxes. Il était important que ces règles arrivent en fin de chaîne de traitement[63] à cause de l'homonymie entre les types Prolexbase.

- règles sur les catégories : en dernier, si l'article n'a toujours pas été classé, des règles sur les catégories sont appliquées. Ces règles sont peu robustes car les catégories ont déjà été considérées par le SVM pour certains types.

Le résultat a montré que la méthode la plus facile à utiliser et la plus robuste est celle des règles sur les infoboxes. L'augmentation de la base d'apprentissage améliore les performances du classifieur SVM. Les résultats montrent qu'il est possible d'atteindre un enrichissement important de Prolexbase avec un bruit limité.

Savary *et al.* (2013) ont construit ProlexFeeder, un module intéressant que nous avons étudié au début de notre travail et nous le considérons comme une ressource importante dans ce travail. ProlexFeeder présente une méthode semi-automatique pour l'enrichissement de Prolexbase par l'ajout de nouveaux noms propres en trois langues (polonais, anglais et français) qui sont extraits de ressources libre comme l'encyclopédie Wikipédia et GeoNames.

Le processus pour enrichir Prolexbase comprend trois sources de données : les éditions Wikipédia polonaise, anglaise et française, les noms propres polonais dans GeoNames et Translatica[64].

D'après les auteurs, la méthode principale de l'enrichissement consiste dans les principaux points suivants :

- une sélection automatique de classes dans la Wikipédia et GeoNames ;

[63] Par exemple, le SVM classifie les séries télévisées parmi les œuvres, alors que ce sont des produits dans la classification Prolexbase. L'erreur du filtre est liée au fait que les films sont classés comme œuvres dans Prolexbase. Ils procèdent à un post-traitement en regardant les mots-clefs comme « série télévisée » parfois précisées entre parenthèses dans le titre.

[64] Une machine finie pour générer les formes fléchies de noms polonais.

- une mise en correspondance manuelle (mapping) entre les classes sélectionnées et la typologie de Prolexbase ;

- l'extraction automatique[65] de classes et l'estimation[66] de leur popularité (fréquence) ;

- la génération automatique de formes fléchies de noms propres polonais (simples et composés) par Translatica ;

- la procédure d'intégration qui s'opère en deux aspects, à savoir :

  1. une vérification automatique de l'entrée pour éviter sa duplication dans Prolexbase ;

  2. une validation manuelle de l'entrée candidate et ses informations morphosyntaxiques.

  Ce processus est fait par un expert, via une interface graphique, une interface de validation pour valider les données à ajouter avant de les intégrer dans la base.

Une des particularités de ProlexFeeder est qu'il exploite des critères de sélection appropriés permettant de conserver les noms propres les plus pertinents, populaires et stables. Ces critères sont basés sur la popularité des articles de l'encyclopédie Wikipédia correspondants, des listes systématiques de certaines grandes catégories trouvées dans les GeoNames, et des techniques approfondies d'intégration de données pour éviter la duplication des données au cours d'un processus d'enrichissement (par opposition à une extraction à partir de zéro).

Une évaluation est effectuée pour estimer à la fois la performance du processus d'intégration de données dans Prolexbase et l'utilisation de l'interface de validation humaine. A cet effet, des échantillons, de trois types différents, issus de l'encyclopédie Wikipédia ont été sélectionnés : célébrité, œuvre et ville comptant 500 entrées chacun. Une précision de 97,4% pour le type célébrité, 96,2 % pour le type œuvre et de 98% pour le type ville a été obtenue.

---

[65] ProlexFeeder utilise le Template d'infobox comme facteur principal pour l'extraction et la classification de noms propres ; les liens redirigés dans l'encyclopédie Wikipédia peuvent être de précieuses sources d'alias et de synonymes pour les entrées extraites.

[66] Chaque prolexème en Prolexbase reçoit l'une des trois valeurs standards de fréquence : (1 : Fréquemment utilisé, 2 : peu utilisé et 3 : rarement utilisé).

5 L’exploitation de l’encyclopédie Wikipédia

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. Dans cette section, nous allons présenter de nombreuses études qui ont été faites sur l’exploitation de la Wikipédia en tant que ressource intéressante d’informations.

Sellami *et al.* (2012) ont présenté des méthodes simples et efficaces pour l’extraction d’un corpus comparable [67] et d’un lexique bilingue arabe-français à partir de la Wikipédia. Il s’agit d’une exploitation à grande échelle de la structure d’articles de la Wikipédia pour construire deux corpus qui seront très utiles dans le domaine de traitement automatique de langues naturelles.

La méthode d'extraction du corpus comparable arabe-français depuis la Wikipédia comprend les quatre étapes suivantes :

- la première étape consiste à télécharger la base de données (Janvier /Février 2012) des éditions Wikipédia arabe et française au format XML ;
- deuxièmement, tous les sujets arabes qui ont des traductions françaises sont extraits d’articles de la Wikipédia ; par exemple :
  Le titre arabe ’’ لاعبو كرةِ مضربِ المان’’ mène à traduction française ‘’joueurs de tennis allemands’’ en suivant la syntaxe de liens interlangues.
- troisièmement, pour chaque paire de sujets, il faut extraire tous les articles arabes et français qui ont un lien texte au thème choisi ;
- la quatrième étape consiste à nettoyer l’extrait d’articles en supprimant les marges Wikipédia.

[67] Les Corpus comparables sont les « ensembles des textes en différentes langues qui ne sont pas des traductions de l'autre» (Bowker et Pearson, 2002), mais ils contiennent des textes du même domaine. » [Sellami *et al*. 2010].

La sortie de ce processus est un corpus comparable de textes arabes et français partageant les mêmes sujets. Le corpus compte 20 533 sujets arabes et leurs traductions en langue française.

L'évaluation a montré l'existence d'une différence importante en termes de taille des catégories et de définition entre l'édition Wikipédia arabe et l'édition Wikipédia française. Par exemple, 41 articles arabes partagent la catégorie "بحيرات / Lac» contre seulement 9 articles français qui partagent la même catégorie. Cependant, la différence entre le nombre d'articles arabes et le nombre d'articles français est moins importante que prévue puisque l'extraction est réalisée à partir de l'édition Wikipédia arabe.
La méthode d'extraction d'un lexique bilingue à partir de la Wikipédia comprend :

1. le téléchargement de la base de données de la Wikipédia (Janvier 2012) au format XML et permettant ainsi d'extraire 104 - 104 (arabe – français) liens interlangues, chaque lien correspondant à une paire de titres arabe-français ;
2. l'exploitation des liens interlangues[68] entre les articles de l'encyclopédie Wikipédia afin d'extraire les titres arabes (simples et composés) et leur traduction en français ;
3. l'utilisation d'une approche statistique[69] pour aligner[70] les mots composés ; avant d'aligner ces titres, une étape de prétraitement consiste à enlever les mots d'arrêt arabes et français ;
4 un filtrage basé sur une méthode linguistique qui utilise l'étiqueteur Tagger2 développé à l'Université de Stanford[71].

Par l'application de ce système, 235 938 titres arabes français sont tenus comme des candidats pertinents.

---

[68] Ces liens sont créés par les auteurs des articles dans Wikipédia ; Par ailleurs, un article paru dans la langue source est lié à un seul article dans la langue cible. (Sellami *et al*., 2012)
[69] L'alignement des mots de chaque titre est basé sur les modèles IBM [5.1] (Brown *et al*., 1993) combiné avec un modèle de Markov caché (Vogel *et al*., 1996). Ces modèles standard ont déjà prouvé leur efficacité dans de nombreuses recherches. (Sellami *et al*., 2012)
[70] L'objectif de l'étape d'alignement est d'avoir un lexique constitué de mots simples.
[71] http://nlp.stanford.edu/software/tagger.shtml

Certaines erreurs de traduction des lexiques arabe–français sont dues au fait que certains titres d'articles sont introduits dans une autre langue que l'arabe. D'autres erreurs sont dues au fait que des paires de titres ne sont pas d'exactes traductions mais se rapportent à la même notion ; par exemple, la paire de titres Noël /عيد où le mot « عيد » (ide) a été traduit par Noël, or, la traduction exacte du mot « عيد» en français est « fête » et non pas « Noël ».

L'encyclopédie Wikipédia représente aussi une ressource de noms propres. Robert Viseur (2013) propose une recherche portant sur un cas pratique de l'extraction de données biographiques relatives à des personnalités originaires de Belgique. Ce travail est organisé en trois sections :

Dans la première section, l'auteur présente un état de l'art sur l'exploitation de la Wikipédia. Parmi les méthodes qui sont adoptées dans ce domaine, il a cité :

- ✓ l'utilisation des copies XML de la Wikipédia disponibles en ligne (Biadsy *et al.*, 2008);
- ✓ l'utilisation de DBpédia qui a démarré en 2007 (Auer *et al.* 2007) ;
- ✓ l'utilisation des infobox associées aux articles de l'encyclopédie Wikipédia (Hellmann *et al.* 2009).

Dans la deuxième section, il présente un cas pratique d'extraction de données biographiques de personnalités belges dans le but d'alimenter une base de données biographique. Cette approche est basée sur :

- l'identification des articles anglophones et francophones pertinents de la Wikipédia. Les concepts appliqués sont l'interrogation de DBpédia via une requête SPARQL et la technique de crawling sur le site de l'encyclopédie ;
- l'extraction des données dans le texte visant à une analyse de texte brut [72] en deux opérations :

[72] «Une copie des articles est sauvegardée en local. En pratique, nous travaillerons sur la version des articles dans le format propre à l'encyclopédie Wikipédia. Cette version est accessible via des URLs de la forme : http://fr.wikipedia.org/w/index.php?action=raw&title=xxxxx et permet d'obtenir un texte brut (texte + syntaxe Mediawiki) », (Robert Viseur, 2013).

1) l'extraction de l'infobox lorsqu'elle existe ;

2) l'identification des phrases de la biographie contenant des informations biographiques importantes, telles que la date de naissance, la date de décès (le cas échéant) et la profession, par la mise en œuvre d'un jeu d'expressions régulières.

Ces données structurées sont stockées dans un fichier CSV. Ce fichier contient 10 610 entrées, avec les champs suivants : nom, date de naissance, date de décès, profession, URL de la catégorie et URL de l'article au format HTML

Dans la troisième section, il a abordé les difficultés rencontrées, selon trois éléments fondamentaux :

- les articles ne sont accompagnés d'une infobox que dans moins d'un cas sur trois, les propriétés des infoboxes ne sont pas totalement standardisées[73] ;
- les formats de dates ne sont pas homogènes[74] ;
- la sélection préalable de phrases candidates pour l'extraction de données nécessite une mise en œuvre plus fine que celle utilisée dans cette recherche (un jeu d'expressions régulières).

L'évaluation a été réalisée sur un ensemble de 2 980 entrées contenant des infoboxes, une comparaison a été faite entre les dates de naissance extraites dans le texte des articles de la Wikipédia et les dates de naissances extraites dans les infoboxes.

Sur un ensemble de 1 644 dates de naissances comparables : 90,4 % sont des dates identiques, 9,6 % sont des dates différentes, un taux d'erreur d'extraction de 1,9%.

Suivant le même principe, Sadat *et al.* (2010) ont fait une étude visant à extraire automatiquement une terminologie bilingue ou multilingue des articles de la Wikipédia pour la construction et l'enrichissement de ressources linguistiques multilingues telles que les

[73] Par exemple, selon Robert Viseur (2013) : « La date de naissance pourra ainsi être annoncée par date_naissance, date naissance, date de naissance, date_de_naissance ou encore naissance.».

[74] « Les dates peuvent être écrites avec des chiffres uniquement, avec le mois écrit en lettres ou encore être complétées par d'autres informations comme le lieu de naissance ou le type d'activité pour laquelle la personne s'est faite remarquer. » (Robert Viseur, 2013).

dictionnaires et ontologies. Le processus adopté consiste en deux parties : la première est celle de la construction des corpus comparables (c.-à-d. dans deux langues au moins) à partir des articles de la Wikipédia et via une requête composée de n mots dans une langue source S. Ensuite, on utilise des liens interlangues pour la même requête afin de créer un corpus comparable dans la langue source S et la langue cible T. La deuxième partie est l'extraction d'une terminologie bilingue de ce corpus.

Ils ont évoqué les fondements du processus de l'extraction :

- les repérages des termes sources et cibles des deux corpus comparables ;
- la construction des vecteurs contextes dans les deux langues ;
- le transfert du contenu des vecteurs contextes de la langue source vers la langue cible en utilisant les traducteurs interlangues de la Wikipédia ;
- la construction des vecteurs de similarité en utilisant des mesures de similarité telles que le cosinus, la distance de Jaccard et le coefficient Dice ;
- Les informations linguistiques des mots sources et cibles sont utilisées afin de réordonner les termes et leurs traductions candidates et ainsi éliminer les termes cibles inutiles.

Les résultats obtenus sont favorables, et les évaluations préalables qui utilisent les paires de langues français-anglais, japonais-français et japonais-anglais ont montré une bonne qualité des paires de termes extraites.

Parmi leurs perspectives considérées, ils envisagent d'augmenter la taille des corpus comparables afin d'obtenir une meilleure terminologie ; de définir des relations sémantiques entre les différents termes en langues source et cible ; et aussi d'utiliser plusieurs requêtes et des évaluations pour déterminer la qualité des traductions.

Pour leur part, Chevalier *et al*. (2010) ont implémenté WikipediaViz[75], un ensemble de visualisations basé sur un mécanisme d'agrégation de données d'édition Wikipédia pour aider les utilisateurs occasionnels[76] à estimer l'impact sur la qualité et la fiabilité des articles de la Wikipédia.

Ce travail a commencé par l'organisation de deux sessions de conception participative avec 4 administrateurs Wikipédia, 2 contributeurs majeurs et 2 sociologues, dans l'objectif d'identifier les besoins des utilisateurs, en particulier lorsqu'ils naviguent comme simples lecteurs, et lister les indices sur lesquels ces experts se basent pour estimer la qualité d'un article.

Les sessions ont abouti à deux méthodes : la méthode objective qui expose des données factuelles et la méthode subjective qui tente d'agréger des scores pour fournir un niveau de qualité. Les auteurs ont choisi la méthode objective et ils ont listé les cinq indices de qualité suivants :

1) Nombre de mots ;
2) Nombre de contributeurs et taux de contributions ;
3) Nombre et taille des éditions ;
4) Nombre de références et de liens internes (wiki links) ;
5) Taille et activités de la discussion.

Enfin, ils ont conçu WikipediaViz : cinq visualisations représentent les cinq indices de qualité, chaque visualisation correspond à un objet dynamique HTML ou une image, insérée dans le panel gauche, sous le logo de la Wikipédia.

[75] «WikipediaViz » est implémenté en PHP et intégré comme un plugin dans le système Mediawiki. Il utilise la base de données standard Wikipédia avec des tables additionnelles que nous calculons pour visualiser rapidement la frise et les contributions des auteurs. » (Chevalier *et al*., 2010).

[76] « Qui ne maîtrisent pas les moyens de récolter ces indices, qui parfois ne sont pas conscients des problèmes de fiabilité, et qui ne sont familiers ni des statistiques, ni des techniques de visualisation. » (Chevalier *et al.*, 2010)

Attia *et al.* (2010) ont réussi à enrichir le lexique multilingue des entités nommées MINELEX[77] (*Multilingual, Interoperable Named Entity Lexicon*) avec la langue arabe via l'exploitation de deux ressources lexicales, le Word Net Arabe (*AWN*) et l'édition Wikipédia arabe (*AWK).* La méthode est fondée sur l'extraction des entités nommées du Word Net Arabe et l'identification de leurs correspondants dans les catégories et sous-catégories d'hyponymie dans la Wikipédia arabe puis leur insertion dans le lexique arabe.

Pour augmenter la taille de ce lexique, ils ont appliqué la recherche par mot-clé sur les résumés des articles de la Wikipédia arabe qui n'ont pas de correspondance dans les autres langues[78], en utilisant le site AWK[79] qui fournit un fichier XML contenant tous les titres des entrées suivies par une brève description de l'entrée.

En outre, une étape de post-traitement a été effectuée afin de récupérer un nombre supplémentaire d'entités nommées de l'édition Wikipédia arabe qui ne sont pas accessibles par le Word Net Arabe.

Finalement, ils ont fait une étude de la diacritisation via la base de données Géonames et l'outil MADA+TOKAN (Habash *et al*., 2005), pour restaurer les marques de voyelles des entités arabes.

Avec cette méthode, ils ont construit un lexique arabe bien structuré contenant 45 000 entités nommées arabes et une version servant la norme ISO LMF (Lexical Markup Framework) a été créée. Ils ont testé leurs résultats par une évaluation quantitative et qualitative du lexique par rapport à un corpus annoté manuellement « Gold Standard » et obtenu des scores d'une précision de 95,83% (avec rappel de 66,13%) à 99,31% (avec rappel de 61,45%) selon les différentes valeurs d'un seuil.

[77] http://www.ilc.cnr.it/ne-repository
[78] Catalan, néerlandais, anglais, français, italien, norvégien, portugais, roumain, espagnol, suédois.
[79] http://download.wikimedia.org/arwiki/

6 Les mesures de la popularité de certains articles de l'encyclopédie Wikipédia

Plusieurs travaux suggèrent l'utilisation du web sémantique, et plus précisément de l'encyclopédie en ligne Wikipédia pour estimer la notoriété ou la popularité de noms propres. Eom Y-H *et al*. (2015), en particulier, ont fait une étude qui vise à analyser les liens Wikipédia en 24[80] langues et de classer leurs articles en utilisant différentes méthodes dans le but d'identifier l'interaction entre les 24 cultures différentes.

Dans cette recherche, ils ont considéré chaque édition de la Wikipédia comme un réseau d'articles. Chaque article correspond à un nœud du réseau et des liens hypertextes entre les articles correspondent aux liens du réseau. Notons que l'analyse des liens a été faite indépendamment pour chaque édition Wikipédia donnée. En effet, il s'agit de trois phases principales :

- L'extraction automatique des articles biographiques[81] à partir de l'édition anglaise de la Wikipédia. Ils ont extrait 1,1 million d'articles biographiques. Ils ont identifié le lieu de naissance, la date de naissance et le genre de chaque figure historique sélectionnée de DBpedia.

- L'extraction des titres correspondant dans les 23 autres éditions linguistiques, en utilisant les liens interlangues fournis par Wikidata[82] ;

- L'application de deux algorithmes de classement afin d'obtenir une liste des 100 meilleurs personnages historiques, pour chaque édition et pour chaque algorithme. Les deux algorithmes utilisés sont :

[80] « We consider 24 different language editions of Wikipedia: English (EN), Dutch (NL), German (DE), French (FR), Spanish (ES), Italian (IT), Portuguese (PT), Greek (EL), Danish (DA), Swedish (SV), Polish (PL), Hungarian (HU), Russian (RU), Hebrew (HE), Turkish (TR), Arabic(AR), Persian (FA), Hindi (HI), Malaysian (MS), Thai (TH), Vietnamese (VI), Chinese (ZH), Korean (KO), and Japanese (JA). »[Eom Y-H et al., 2014]

[81] «To identify biographical articles, we considered all articles belonging to "Category:living people", or to "Category:Deaths by year" or "Category:Birth by year" or their subcategories in the English Wikipedia» (Eom *et al*, 2014).

[82] «The transcriptions of names from English to the other 23 selected languages are harvested from WikiData (http://dumps.wikimedia.org/wikidata) and not directly from the text of articles. » (Eom *et al*, 2014).

1) Le page Rank Google (en considérant les liens entrants à un article donné) ;

2) 2DRank [83] (en considérant les liens entrants et les liens sortants pour un article donné).

Par exemple, leur analyse en appliquant la notion PageRank a conçu 3 groupes de personnages historiques ; notons que NA est le nombre d'apparitions dans différentes éditions de la Wikipédia pour une personne donnée, et HKI est le classement moyen d'une personne donnée au fil des éditions Wikipédia pour chaque algorithme de classement :

1) des personnages historiques mondiaux apparaissent dans la plupart des éditions Wikipédia (NA ≥ 18) et ils sont bien classés (HKI ≤50) pour chaque édition de la Wikipédia, comme : Charles Linné, Platon, Jésus, et Napoléon ;

2) des personnages historiques locaux apparaissent dans quelques éditions de Wikipédia (NA <18), mais ils sont bien classés (HKI ≤ 50) dans leurs éditions de Wikipédia tels que : Tycho Brahe, Sejong le Grand, et de Sun Yat-sen ;

3) des personnages historiques locaux apparaissent dans plusieurs éditions Wikipédia (NA <18) et ils sont mal classés (HKI > 50).

En résumé, ils ont identifié 2 400 meilleures figures historiques pour chaque algorithme de classement. Cependant, certains personnages historiques tels que Jésus, Aristote ou Napoléon apparaissent dans plusieurs éditions de la Wikipédia. En conséquence, 1 045 figures historiques ont un PageRank élevé et 1 616 figures historiques sont classées supérieures par l'algorithme 2Drank.

Également, le Macro Connections group au MIT Media Lab qui travaille sur la quantification, l'analyse et la visualisation de la culture mondiale a développé le projet Panthéon (Yu *et al*., 2016) qui s'intéresse au calcul d'une notoriété universelle pour les personnes célèbres à travers l'histoire. Dans ce projet, on prend en compte le nombre des différentes versions linguistiques de la Wikipédia ayant un article sur cette personne (en appliquant un filtre : ce nombre doit être supérieur à 25) et on calcule un *indice de popularité historique* (*historical popularity index*).

[83] «Briefly speaking, nodes with both high PageRank and CheiRank get high 2DRank ranking. » (Eom *et al*, 2014)

Cette approche a été testée de plusieurs manières, entre autres par une comparaison avec le jeu de données de l'accomplissement humain (HA) ; HA est une compilation de 3 869 personnes notables dans les arts et les sciences, basée sur des encyclopédies imprimées. Contrairement à Panthéon, HA peut classer un individu dans plusieurs domaines, avec des scores différents. Par exemple, Galilée est classé comme un astronome (avec un score de 100 qui le place comme l'astronome le plus influent du monde) et aussi un physicien (cinquième rang avec un score de 83). Toutefois, Panthéon contient 40 % des entrées disponibles dans HA, avec une corrélation significative entre les mesures d'impact historique dans Panthéon (*L, HPI*) et dans HA. On a constaté une corrélation entre l'indice de HA et le nombre d'éditions linguistiques sur la Wikipédia (*L*) avec une variance de 18% et une corrélation avec *HPI* où la variance est 12%.

Conclusion

Nous avons lu un grand nombre d'articles qui ont un lien direct ou indirect avec notre travail de recherche, nous les avons classés selon le sujet traité en six catégories : l'analyse morphologique des textes arabes, la reconnaissance d'entités nommées arabes, l'enrichissement de Prolexbase, l'amélioration des bases de données lexicales multilingues, l'exploitation des articles de l'encyclopédie Wikipédia et les mesures de popularité.

En fin de ce chapitre, il conviendra de souligner les quatre travaux principaux que nous allons utiliser dans la réalisation de ce travail de recherche : le projet Prolexfeeder (Savary et al. 2013) pour l'enrichissement de Prolexbase ; le mécanisme WikipediaViz (Chevalier et al. 2010) pour le choix de critères de notoriété ; l'étude sur l'analyse des liens Wikipédia en 24 langues (Eom et al. 2015) pour l'utilisation de concept interlangue relatif à cette encyclopédie. Finalement nous allons comparer les resultats de notre système d'estimation de notoriété d'un nom propre via la Wikipédia avec ceux du projet Panthéon (Yu et al. 2016). Nous aurions souhaité le faire aussi pour les résultats de (Eom et al. 2015), mais ceux-ci ne sont pas disponibles librement.

Chapitre 2 La Langue Arabe اللغة العربية

Introduction

La langue arabe est très répandue dans le monde car elle est parlée par plus de 300 millions de locuteurs. L'arabe est la langue officielle de 26 pays de l'Afrique du Nord et du Moyen Orient, ainsi que de la Ligue Arabe. Il est aussi une des langues officielles de l'Union Africaine, de l'OTAN et des Nations Unies et surtout, elle est la langue intellectuelle et liturgique de l'Islam.

Il s'agit d'une langue sémitique, qui s'écrit et se lit de droite à gauche et dont l'alphabet est un abjad[84]. La langue arabe se caractérise par l'absence de voyelles [85] courtes dans la plupart des textes et aussi l'absence de capitalisation à l'inverse d'autres langues comme le français ou l'anglais ; le manque de capitalisation rend évidement difficile la détection des entités nommées qui est une partie essentielle de la recherche d'information (IR) (Darwish, 2013). Comme les autres langues sémitiques, la morphologie de l'arabe est très riche et complexe avec deux types, dérivationnel et flexionnel (Attia, 2008).

La langue arabe est une collection de dialectes et une langue écrite standard ; l'arabe dialectal diffère de l'arabe standard moderne (MSA) morphologiquement, lexicalement[86] et phonologiquement. En outre, il n'existe pas d'orthographe standard et aucune académie de langue pour l'arabe dialectal. Par conséquent, les outils et les ressources conçus pour l'arabe standard moderne ne fonctionnent pas avec l'arabe dialectal (Habash *et al*., 2013).

La langue écrite standard est la même dans le monde entier : l'arabe standard moderne, (MSA), basé sur l'arabe classique qui est la langue de la littérature historique ; elle est enseignée dans

[84] Un abjad désigne un alphabet qui ne note que des consonnes (ou note principalement les consonnes), comme en arabe ou hébreu (Mesfar, 2008).
[85] Les voyelles jouent un rôle proche des accents en français pour un mot comme peche qui peut être interprété comme pêche, pèche et péché. Par contre, en arabe chaque lettre de chaque mot devrait posséder sa voyelle ce qui n'est en général pas le cas (Douzidia, Fouad Soufiane, 2005).
[86] Par exemple : ARZ طربيزة Tarabiza « Table » correspond à MSA طاولةTawila «Table » (Habash *et al*, 2013).

les écoles contemporaines et également utilisée dans certains écrits de communication comme les éditions d'informations et les débats parlementaires. Toutefois, aujourd'hui, les dialectes arabes sont en train d'émerger comme des langues de communication informelle en ligne, dans les courriels, les blogs, les forums de discussions, etc. On peut identifier cinq grandes régions dialectales en arabe : Egypte, Golfe, Maghreb, Levant et Irak ; récemment, le traitement automatique du dialecte arabe a généré une quantité considérable de recherches en TAL (Shoufan et Alameri, 2015), (Habash, 2010).

Dans la suite de ce chapitre, nous allons présenter certains concepts fondements de la langue arabe.

1 L'alphabet arabe

L'alphabet arabe est constitué principalement de 28 consonnes, sur et sous lesquelles se placent des voyelles et des signes diacritiques qui jouent un rôle primordial dans la langue arabe ; en effet, il suffit d'un simple changement de voyelle, pour qu'un mot prenne un tout autre sens. Les 28 consonnes sont divisées en deux groupes :

1) 14 consonnes appelées « solaires » se confondent avec le « L » de l'article défini « AL» au début du mot et ainsi le « L » n'est pas prononcé devant ces lettres : ت ث د ذ ر ز س ش ص ض ط ظ ل ن;
2) 14 consonnes appelées « lunaires » ne se confondent pas avec le « L » de l'article défini « AL» au début du mot et ainsi le « L » est bien prononcé devant ces lettres : ا ب ج ح خ ع غ ف ق ك م ه و ي;

Parmi les consonnes dites lunaires, 3 sont des voyelles longues (ا و ي) ; parallèlement, il existe d'autres formes particulières de lettres :

1. la lettre Hamza : il en existe 6 formes (ؤ-أ-ئ-إ-آ-ء) ;
2. la lettre T-marbuta : c'est le marqueur du féminin (ة) ;
3. la lettre Alif-maqsura : c'est le marqueur dérivationnel qui ne s'utilise qu'en fin de mot ; c'est une lettre de prolongement pour le phonème /a/ (ى).

Lettre arabe	Correspondant français	Prononciation		Lettre arabe	Correspondant français	Prononciation
ا	a	Alef		ض	d	Dad
ب	b	Ba'		ط	t	Tah
ت	t	Ta'		ظ	z	Zah
ث	th	Tha'		ع	‘ ‘	Ayn
ج	j	Jim		غ	gh	Ghayn
ح	h	Hha'		ف	f	Fa
خ	kh	Kha'		ق	q	Qaf
د	d	Dal		ك	k	Kaf
ذ	d	Thal		ل	l	Lam
ر	r	Ra		م	m	Mim
ز	z	Zayn		ن	n	Nun
س	s	Sin		ه	h	Ha
ش	sh	Shin		و	w	Waw
ص	s	Sad		ي	y	Ya

Tableau 2. 1: Les 28 lettres arabes, (Leclerc, 2000) extrait de (Douzidia et al, 2005)

2 La diacritisation

Le mot arabe s'écrit avec des consonnes et des voyelles ; les consonnes changent de forme de présentation selon leur position dans le mot (au début, au milieu ou à la fin) ; les voyelles sont de deux types : les voyelles courtes (diacritiques) et les voyelles longues.

Contrairement au français, les voyelles arabes ne sont pas des lettres de l'alphabet mais ce sont des signes diacritiques qui se rajoutent aux lettres et qui jouent le même rôle que les voyelles dans les autres langues.

La diacritisation est le processus qui consiste à recouvrir un mot avec des signes diacritiques ; même si l'utilisation des diacritiques dans les textes arabes est optionnelle, la présence de ces signes enlève l'ambigüité et facilite la compréhension de texte. En général, les textes arabes sont non vocalisés et c'est au lecteur de deviner les diacritiques des textes lors de la lecture. En revanche, les textes religieux, les livres éducationnels pour enfants et les textes littéraires sont entièrement vocalisés. On distingue aussi la vocalisation partielle qui répond dans les éditions soignées à la levée de certaines ambigüités de première lecture.

Les diacritiques (voyelles) principaux sont :

- les voyelles courtes (brèves) : (Fatha ـَـ /a/, Damma ـُـ /ou/, Kasra ـِـ - /i/) ;
- les voyelles longues : ا (A), و (U), ي (I) ;
- le tanwine : le doublement de voyelles courtes (ـًـ (an), ـٌـ (un), ـٍـ (in)) ;
- le Chadda : le marqueur de la gémination de consonne (ـّـ) ;
- le Soukoon : le marqueur de l'absence de voyelle courte (ـْـ).

Nom de la voyelle	Appliquée à la lettre (ra)	Prononcé comme :
Fatha	رَ	ra
Damma	رُ	rou
Kasra	رِ	ri
Chadda	رّ	rr
Soukoon	رْ	r

Tableau 2. 2: L'application de diacritiques sur la lettre ra (ر)

Bien que les diacritiques soient destinés à enlever l'ambiguïté des textes arabes, la majorité des analyseurs morphosyntaxiques de l'arabe, entre autres, Buckwalter (Buckwalter, 2004), Xerox (Beesley, 2005) ou MADA (Habashe, 2005) ne traitent que les textes non vocalisés car la majorité des textes arabes ne sont pas vocalisés. De plus, si l'entrée est partiellement voyellée, ces analyseurs éliminent tous les diacritiques et analysent l'entrée comme si elle était non voyellée. Or, prendre en considération la vocalisation partielle dans le texte arabe joue un rôle important pour réduire son ambiguïté (Hamadi, 2012).

3 La morphologie

La langue arabe se caractérise par sa morphologie fortement flexionnelle, dérivationnelle et agglutinante. Les mots arabes peuvent être classés en trois catégories morphosyntaxiques : Noms, Verbes et Particules.

Généralement, la plupart des mots arabes sont formés par l'application de schèmes (modèles) aux racines ; une racine est constituée de trois consonnes (trilitères) ou quatre consonnes (quadrilatère) ; un schème est une suite de voyelles et d'autres consonnes (affixes) qui se placent avant, après ou entre les consonnes de la racine. Par conséquent, le schème d'un mot permet de détecter les lettres qui constituent sa racine (Habash, 2010).

Par exemple : le participe actif كاتب kAtib « écrivain » est un nom dérivé qui est formé par l'association de la racine 'k t b' sur le schème 1A2i3 (les nombres représentent les consonnes de la racine).

3.1 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 (Habash, 2010).

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

- l'aspect a trois valeurs : accompli (ماضي mADiy), non accompli (مضارع muDAriÇ) et impératif (أمرِ Aamr) ;
- le mode a trois valeurs : indicatif (مرفوع marfuw''), subjonctif (منصوب manSuwb) et jussif (مجزوم majzuwm) ;
- le voix à deux valeurs : passif et actif ;
- le sujet possède trois traits morphologiques :
    1. le trait personne qui a trois valeurs : (1er personne, متكلم mutakal~im),

       (2ème personne, مخاطب muxATab) 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.

3.2 La morphologie nominale

En comparaison avec le système verbal, le système morphologique nominal est plus complexe et idiosyncratique (Habash, 2010). Les noms arabes (noms, adjectifs et noms propres) se fléchissent en genre, nombre, état et cas :

- genre : masculin, féminin ;
- nombre : singulier, duel, pluriel ;
- cas : nominatif, accusatif, génitif ;
- état : défini, indéfini, construit[87].

Dans ce système, on peut distinguer six classes, chaque classe rassemble les noms disposant des traits morphologiques identiques :

3.2.1 Les triptotes

Il s'agit d'une classe de noms qui ont une flexion casuelle à trois cas, nominatif, accusatif et génitif, déclinés respectivement par (Damma ـُ /u/, Fatha ـَ /a/ et Kasra ـِ /i/) avec l'état défini ou construit. De même, cette classe comprend les noms à trois cas, nominatif, accusatif et génitif, déclinés par les trois voyelles courtes de tanwine (ـٌ(un), ـً (an), ـٍ (in)) avec l'état indéfini .

Notons que les noms qui ne se terminent pas par Ta-marbuta (ة)/ou Alif Hamza ((اء reçoivent un trait particulier en leur ajoutant un extra Alif[88](ا) dans le cas accusatif indéfini.

Pour comprendre comment fonctionnent les triptotes, prenons comme exemple caractéristique les paradigmes flexionnels d'un nom commun triptote : kitab (livre), il se présente comme dans le tableau 2.3.

[87] L'état construit indique que le nom est la tête d'une construction d'Idafa, c'est-à-dire le premier mot (مضاف muDAf) qui est possédé par le syntagme nominal qui le suit. Par exemple : le mot كتاب kitAbu 'livre' dans la phrase nominal (اكتاب الطالب kitAbu AlTAlibi ' le livre de l'étudiant').

[88] C'est l'alif (ا)/a/ avec le diacritique (ً (an)).

état / cas	I. Indéfini (un livre)	II. Défini par l'article (le livre)	III. Défini par l'annexion (Le livre de l'enfant)
Nominatif	kitab-u-n	al-kitab-u	kitab-u al-walad-i
Accusatif	kitab-a-n	al-kitab-a	kitab-a al-walad-i
Génitif	kitab-i-n	al-kitab-i	kitab-i al-walad-i

Tableau 2. 3: Les trois paradigmes flexionnels du nom commun triptote (livre).

3.2.2 Les diptotes

Il s'agit d'une classe lexicale spécifique qui ne connaît qu'une flexion à deux cas seulement. Contrairement aux triptotes, les noms indéfinis sont privés ou « interdits » de tanwine ; d'autre part, le génitif ne possède pas de marque spécifique et il a la même forme morphologique que l'accusatif, c'est-à-dire que les diptotes sont déclinés par le suffixe diacritique Fatha ـَـ /a/ dans les deux cas accusatif et génitif. Par exemple, dans la phrase : le livre d'Ahmed (كتاب أحمد), le nom Ahmed (أحمد) est un diptote au cas génitif. Il est bien au cas génitif à cause de l'annexion et en tant que diptote, son signe caractérisant le cas génitif est la *Fatha (et pas une Kasra).*
Les noms communs qui sont sur le schème (مفاعيلmf"ail m1a2i3) comme : mafatih (clés) sont considérés diptotes, leurs paradigmes flexionnels sont illustrés dans le tableau 2.4 ci-dessous.

état / cas	I. Indéfini (des clés)	II. Défini par l'article (les clés)	III. Défini par l'annexion (Les clés de l'enfant)
Nominatif	mafatih-u	al-mafatih-u	mafatih-u al-walad-i
Accusatif	mafatih-a	al-mafatih-a	mafatih-a al-walad-i
Génitif	mafatih-a	al-mafatih-i	mafatih-i al-walad-i

Tableau 2. 4: Les trois paradigmes flexionnels du nom commun diptote (clés)

3.2.3 Les indéclinables, les défectifs et les invariables

- **Les indéclinables**

Les noms non déclinables en arabe ne changent pas de voyelle finale en fonction de leur cas et leur état comme vu précédemment De même, ils ne prennent pas de *tanwine* qui est propre aux

noms déclinables. Les outils interrogatifs, les pronoms démonstratifs, les prépositions et les pronoms sont des mots non déclinables.

Cette classe comprend aussi les noms masculins singuliers en leur appliquant le suffixe Fataha ـَ/a/ pour les trois cas : nominatif, accusatif et génitif et le suffixe de tanwine (ـً /an/) pour les trois cas ayant l'état indéfini ; par exemple : signification (معنى maÇnaÝã).

- **Les défectifs**

Ils sont dérivés d'une racine avec un radical final faible (y), ils reçoivent deux traits morphologiques : le suffixe du génitif (Kasra - /i/) pour les cas nominatif et génitif et le suffixe de l'accusatif pour le cas accusatif. Également, ils autorisent le tanwine quand l'état du nom est indéfini, mais toujours avec deux suffixes : le suffixe de tanwine (- (an)) pour le cas accusatif et le suffixe de tanwine (- (in)) pour les cas nominatif et génitif en supprimant le glide final, par exemple : Le nom'un juge' قاضٍ qADin (nominatif, génitif) contre قاضياً qADiyan (accusatif).

- **Les invariables**

Ce sont les noms singuliers dont la voyelle finale ne change pas pour tous les cas (nominatif, accusatif et génitif) ; par exemple : les noms qui se terminent par une voyelle longue comme : France (فرنسا), Libye (ليبيا).

3.2.4 Le duel

Les noms duels ne se déclinent pas par des diacritiques comme les noms singuliers ; ils sont déclinés par des lettres comme suit :

- par le suffixe littéral (اِ+A) au cas nominatif, genre masculin, et état construit ; ex : (كتابا الطالب kitAbaA AlTAlibi ' les deux livres d'étudiant') ;
- par le suffixe littéral (ان +A+ni) au cas nominatif, genre masculin et état défini ou indéfini; ex :(كتابان kitAbaAni 'deux livres'/ou alkitAbaAni الكتابان ' les deux livres') ;
- par le suffixe littéral (ي +ya) au cas accusatif ou génitif, genre masculin et état construit ; ex : (كتابي الطالب kitAbaay AlTAlibi ' les deux livres de l'étudiant') ;
- par le suffixe littéral (ين +ay +ni) au cas accusatif ou génitif, genre masculin et état défini ou indéfini ; ex :(كتابين kitAbaayni 'deux livres' /ou al kitAbaayni

الكتابين‘les deux livres ’) ;

- par le suffixe littéral (ا +ta+A)[89] au cas nominatif, genre féminin et état construit ; ex : (نافذتا البيت nAfidatA Albyati ’ les deux fenêtres de la maison’) ;
- par le suffixe littéral (ان ta +A+ni) au cas nominatif, genre féminin et état défini ou indéfini ; ex :(نافذتا ن nAfidatAni ‘deux fenêtres’ /ou al nAfidatAni النافذتان’ les deux fenêtres ’) ;

- par le suffixe littéral (تي ta +ya) au cas accusatif ou génitif, genre féminin et état construit ; ex : (نافذتي البيت nAfidatay Albyati ’ les deux fenêtres de la maison’) ;

- par le suffixe littéral (ين ta + ya + ni) au cas accusatif ou génitif, genre féminin et état défini ou indéfini ; ex :(نافذتين nAfidatayni ‘deux fenêtres’/ou al nAfidatayni النافذتين ا‘les deux fenêtres ’) ;

3.2.5 Le Pluriel

En arabe, il existe trois types de pluriel : pluriel masculin, pluriel féminin et pluriel brisé :

1. le pluriel masculin externe

C’est une classe régulière dont la déclinaison est effectuée par des suffixes littéraux :

- le suffixe littéral (و +uw) au cas nominatif, genre masculin et état construit, ex : (كاتبوا النص kAtibuw al nasse ’ les écrivains de texte’) ;
- le suffixe littéral (ون+uw+na) au cas nominatif et état défini ou indéfini, ex : (كاتبون kAtibuwna ‘écrivains’/ou al kAtibuwna الكاتبون ‘les écrivains’) ;
- le suffixe littéral (ي +iy) au cas accusatif ou génitif et état construit, ex : (كاتبي النص kAtibiy al nasse ’les écrivains de texte’) ;
- le suffixe littéral (ين +iy +na) au cas accusatif ou génitif et état défini ou indéfini ; ex : (كاتبين kAtibiyna ‘écrivains /ou الكاتبين al kAtibiyna ‘ les écrivains’) ;

[89] Pour produire le duel féminin, nominatif ayant l’état construit, on convertit le marqueur de féminin Tamarbouta (ة) en Ta-maftouha (ت) ; on le décline par la voyelle courte (Fatha - /a/) et on ajoute le suffixe Aا .

2. le pluriel féminin externe

En général, tous les noms féminins pluriels sont engendrés par l'ajout du suffixe littéral (اتAt) à ses lemmes[90] , cependant, ces noms sont déclinés par des signes diacritiques comme suit :

- ✓ par la voyelle (Damma ـُ+At+u) au cas nominatif et état construit ou défini,

  ex : (كاتبا تُ النص kAtibAtu al nasse ' les écrivains du texte' /ou alkAtibAtu تُ الكاتبا 'les écrivaines') ;

- ✓ par la voyelle du tanwine (ـٌ+At+un) au cas nominatif et état indéfini, ex :(كاتبا تٌ kAtibAtun 'écrivaines ') ;

- ✓ par la voyelle de génitif (Kasra ـِ + At+i) au cas accusatif ou génitif et état construit ou défini, ex : (كاتبا تِ النص kAtibAti al nasse 'les écrivaines du texte' /ou al kAtibAti الكاتباتِ 'les écrivaines' ;

- ✓ par la voyelle du tanwine (ـٍ + At+in) au cas accusatif ou génitif et état indéfini ; ex :(كاتبا تٍ kAtibAtin ' écrivaines').

3. le pluriel brisé / ou interne

Cette classe assure un défi réel pour la morphologie de la langue arabe ; sa complexité est due à l'irrégularité du système de flexion, c'est-à-dire qu'il n'existe pas de règles figées d'affixation comme dans les deux autres types (pluriel masculin et pluriel féminin). Environ plus de la moitié des pluriels arabes sont des pluriels brisés. Par exemple : le pluriel du nom bureau (مكتب maktab) est مكاتب makAtib 'bureaux' et n'est pas مكتبونِ* *maktabuwn (non correct).

De plus, il existe des noms pluriels qui sont sur des schèmes des autres noms singuliers, ce qui pose un problème morphologique ; par exemple : les noms كتاب kitAb 'livre' (singulier) et رجال rijAl 'hommes' (pluriel interne) partagent le schème 1i2A3 (Habash, 2010).

[90] Le lemme d'un nom en arabe est sa forme au masculin singulier en cas d'existence ; sinon, son lemme est sa forme au féminin singulier. Le lemme d'un verbe en arabe est la forme de la troisième personne au masculin singulier. (Habash, 2010)

En arabe, il y a des noms qui ont deux types de pluriel (pluriel interne et externe) ou deux pluriels internes, comme le nom طريقِ Tariyqe ‘ chemin’ qui peut être considéré comme féminin singulier ou masculin singulier, ce mot possède deux pluriels :طرق turque ‘chemins’ (pluriel interne) et طرقات turuqueAte ‘ chemins’ (Pluriel féminin).

Certains noms commençant par *mim* ont un pluriel en *(مفاعيل mafa il)*, par exemple, le pluriel du nom مصباح Misbah est مصابيحMasabih sur le schème de *ma1A2i3* ; Cette distribution n'est nullement aléatoire, mais le choix du schème du pluriel dépend de la qualité de la dernière voyelle du nom singulier. Ici encore, il ne suffit pas de croiser une racine et un schème, il faut, crucialement, regarder un troisième terme : le singulier du nom (Boha, 2014).

Pour terminer, la langue arabe présente un vaste domaine approprié à la recherche et à l’exploitation d’applications du traitement automatique des langues naturelles.

Conclusion

Nous avons introduit les caractéristiques générales de la langue arabe en tant que langue riche morphologiquement, flexionnelle et qui est constituée de plusieurs dialectes régionaux ; parmi lesquels, nous distinguons l’arabe égyptien pour être ajouté dans Prolexbase, car il est largement répandu avec plus de 78 millions de locuteurs et parce qu’il possède une édition sur la Wikipédia. À travers ce chapitre nous avons détaillés certains concepts de la morphologie nominale arabe en les illustrant par quelques exemples. À cet égard, il faut indiquer que nous allons effectuer l’ajout d’un module arabe via le corpus de la Wikipédia arabe où les textes sont non voyellés.

Chapitre 3 Ressources

1 Prolexbase

1.1 Introduction

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 pour le TAL.

Aujourd'hui, Prolexbase est un dictionnaire électronique multilingue relationnel spécifique aux noms propres (constituant 10 % des textes journalistiques), disponible librement sur le site Web CNRTL[91], au format LMF (ISO 24613) depuis les travaux de Bouchou et Maurel (2008).

Il s'agit d'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, etc.

En outre, la motivation derrière Prolexbase n'est pas de représenter autant de noms disponibles que possible, comme dans le cas des autres grandes ontologies construites automatiquement, entre autres, YAGO (Suchanek *et al*., 2007) ou DBpedia (Mendes *et al*., 2012) ; Prolexbase se caractérise comme un module qualifié et supervisé de noms propres (Savary *et al*., 2013).

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 interlangue.

Ici, on s'intéresse à une présentation détaillée de l'ontologie à quatre niveaux de Prolexbase. Cette présentation est inspirée notamment par la référence (Maurel *et al*., 2014).

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

1.2 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). 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). La figure 3.1 représente l'architecture générale de Prolexbase.

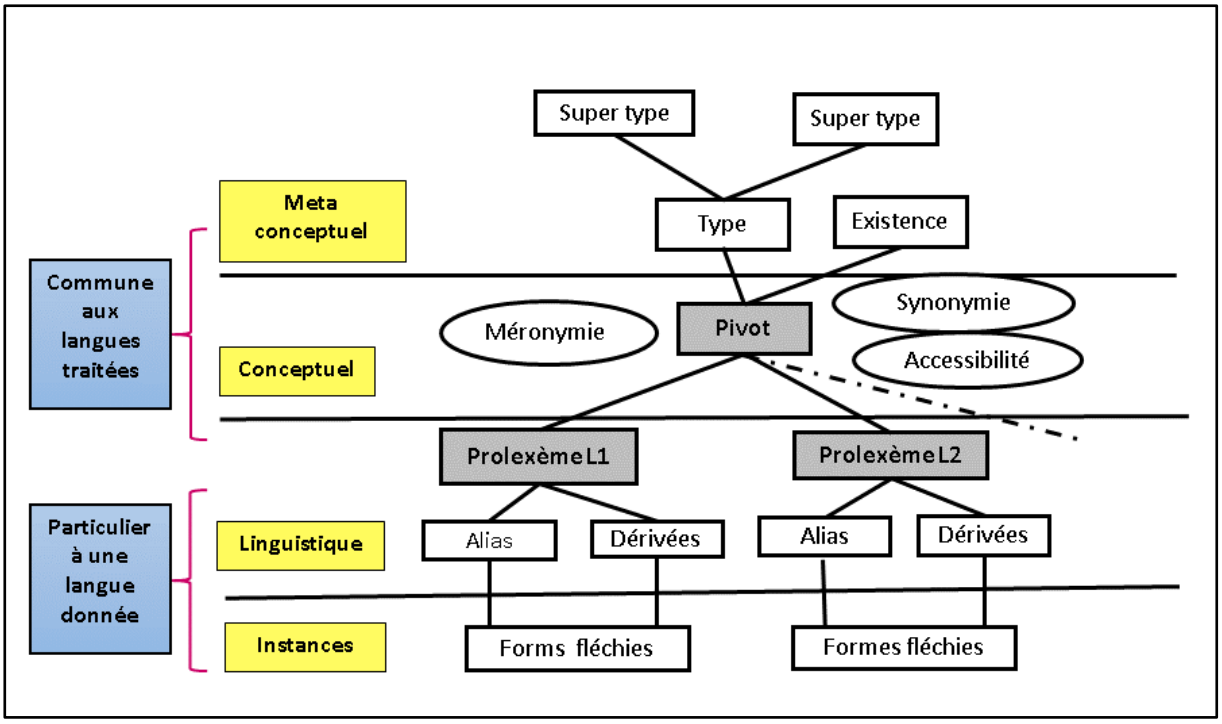


Figure 3. 1 : L'architecture générale de l'ontologie des noms propres (Prolexbase)

1.2.1 La partie commune aux langues traitées

1.2.1.1 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 anthroponymes (trait humain) sont partagés en deux autres super types : les anthroponymes individuels (célébrité[92], prénom, patronyme[93], pseudo-anthroponyme[94]), et les anthroponymes collectifs (dynastie, ethnonyme[95], association, ensemble[96], entreprise, institution et organisation) ;
2. Les toponymes (trait locatif) comprennent tous les noms de lieux au sens général en les rassemblant (pays, région, supranational), et d'autres emplacements comme astronyme, édifice[97], ville, géonyme[98], hydronyme et voie. 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*). Les îles (*Maui : la deuxième plus grande des îles hawaïennes*) appartiendront également à ce type, car elles ne sont pas un Etat indépendant ; or, les Bahamas qui constituent un Etat indépendant, seront considérées comme un pays et non pas comme une région.

---

[92] Les pseudonymes appartiennent également à ce type (Stephen King pour Richard Bachman).
[93] Noms de famille (Sarkozy, Hollande).
[94] Noms d'animaux, de robots, de machines, etc. (Laika - chien spatial soviétique).
[95] Noms de personnes (Angles). Si un ethnonyme est un nom relationnel qui est associé à un toponyme, il ne constituera pas un prolexème distinct (par exemple, Polonais ne sera pas une entrée de nom propre, mais plutôt sous le prolexème Pologne). (Maurel *et al*., 2014).
[96] Noms d'ensemble artistique ou de club sportif, y compris des équipes de football (Manchester United) et des groupes de rock (The Smashing Pumpkins).
[97] Noms de bâtiments, y compris les noms de parcs, jardins, monuments, ponts et théâtres (Jardin du Luxembourg, Tour Eiffel).
[98] Noms de sites naturels (désert du Sahara, Mont Blanc), définis comme des formes de paysage naturel.

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[99] et météorologie.

En outre, alors qu'un super type est hyperonyme de plusieurs types, chaque type est lié à un seul super type. De même, 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.
En plus de cette typologie, le niveau métalinguistique possède une relation nommée existence ; chaque pivot est lié à une seule valeur pour l'existence ; cette fonctionnalité est souvent importante pour la traduction de noms propres. Trois types d'existence sont distingués :

- Historique : ce sont les noms de personnes, d'événements qui existent ou ont existé (Mozart, Paris) ;
- Fictionnel : ce sont les noms propres inventés par des auteurs de romans, d'histoires, pièces de théâtre, films, etc. (Tintin, Atlantis) ;
- Religieux : ce troisième trait dépend de la foi des gens. Selon les auteurs, par exemple, tandis que Jésus et Mahomet sont des noms propres historiques, ce n'est pas le rôle du linguiste de dire si l'archange Gabriel a réellement existé.

1.2.1.2 Le niveau conceptuel

Le deuxième niveau indépendant de la langue est le niveau conceptuel[100], organisé autour du pivot, qui est représenté par un numéro d'identification unique. Le pivot joue le rôle d'un identificateur interlangue, permettant la connexion de noms propres qui représentent les mêmes concepts dans les différentes langues traitées.

Selon Denis Maurel *et al.* (2014), cette représentation par pivot est commune dans de nombreuses bases de données lexicales, y compris EuroWordnet (Vossen, 1998), Balkanet

[99] Noms d'événements historiques ou politiques (Révolution française).
[100] Dans ce niveau, les noms propres qui portent le même pivot sont des traductions interlangues ; par exemple, les noms propres' Libye' (fra),'Libya' (eng) et 'Libia'(pol), ont le même pivot (43156).

(Tufis *et al*., 2004) et Papillon (Mangeot-Lerebours *et al*., 2003). Les pivots, aussi appelés noms propres conceptuels, représentent les différents points de vue du référent d'un nom propre de sorte qu'ils ne correspondent pas directement au référent de la langue.

Par exemple, bien que Jorge Mario Bergoglio et le pape François réfèrent à la même personne, ils possèdent deux pivots différents parce qu'ils représentent deux points de vue différents.

Egalement, il existe trois relations sémantiques qui ne dépendent pas de la langue et sont associées aux pivots dans ce niveau : Synonymie, Méronymie[101] 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 figure 3.2 reprend ces trois relations entourant le pivot 38558.

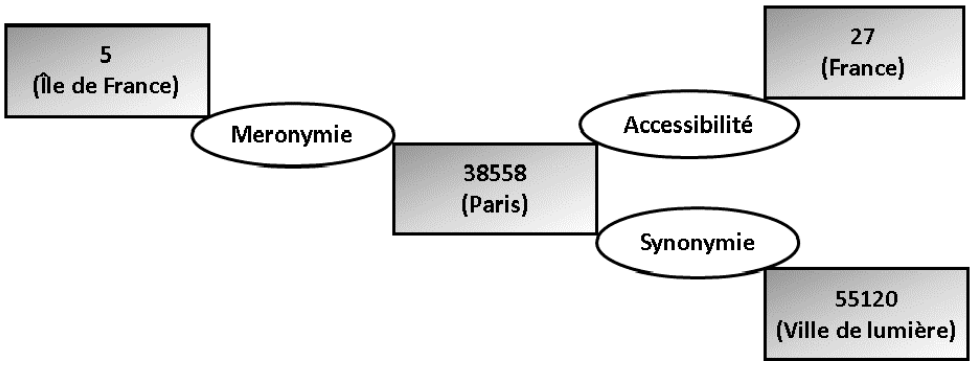


Figure 3. 2: Les relations Synonymie, Accessibilité, Méronymie entourant le pivot 38558 (Paris)

[101] La méronymie est une relation d'inclusion, une relation partie-tout qui concerne tous les types de noms propres par exemple : Tours -> Indre-et-Loire ->Région Centre -> France.

1.2.2 La partie propre à une langue donnée

1.2.2.1 Le niveau linguistique

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.

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 3.3 montre le pivot et les prolexèmes de nom propre Platon dans les quatre langues.

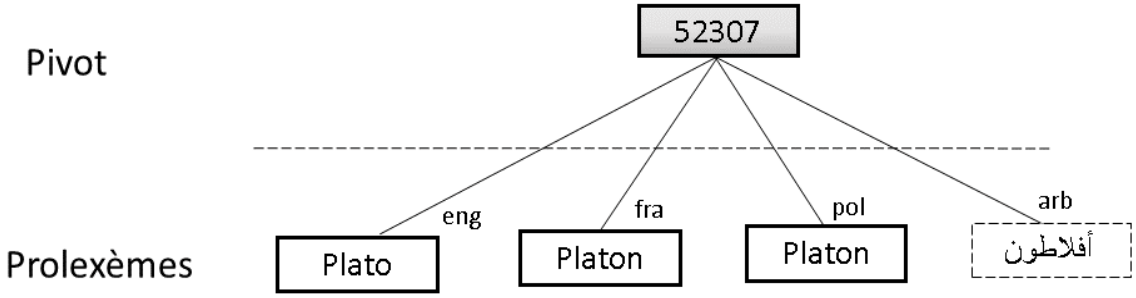


Figure 3. 3: Pivot et prolexèmes du nom propre Platon

Dans ce même niveau, les prolexèmes peuvent avoir des dérivés et des alias dépendant de la langue ; Maurel *et al.* (2006) 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.

Par exemple, Nations unies et ONU sont les alias du nom propre Organisation des Nations Unies alors qu'Onusien est un dérivé ; de même, George W. Bush, George Bush et Bush sont les alias du nom propre George Walker Bush.

Toutefois, selon Maurel *et al*. (2006), 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.

Comme nous l'avons déjà mentionné, dans Prolexbase, il existe des relations indépendantes de la langue entre les pivots, telles que la synonymie, la méronymie et l'accessibilité. De même, nous pouvons distinguer des relations qui dépendent de la langue concernant les prolexèmes : Collocation[102], Contexte d'accessibilité[103], Eponymie[104]. Chaque prolexème reçoit une fréquence qui indique la popularité du nom propre ; trois valeurs standards (ISO 12620) possibles pour cet indicateur de notoriété : 1 fréquemment utilisé, 2 peu utilisé et 3 rarement utilisé. De plus, chaque nom propre lexical est lié à un tri (par exemple Jacques Chirac est trié par rapport aux 2 mots, une langue (ex : fra pour le français, eng pour l'anglais et pol pour le polonais) ; enfin, un lien Wikipédia a été ajouté afin de permettre une population automatique de Prolexbase avec la Wikipédia.

1.2.2.2 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.

---

[102] Elle indique le lien qui peut être établi entre un nom propre et un mot de fonction, tels que des déterminants et des prépositions. Par exemple, en français, presque tous les noms de pays exigent un article (la France, le Portugal) (Maurel *et al*, 2014).

[103] Par exemple, le pivot 38 558 est dans une relation d'accessibilité avec le pivot 27, le pivot 38 558 fait référence à Paris, alors que le pivot 27 fait référence à la France. Le contexte d'accessibilité est la capitale de. Ceci produit ce qui suit : Paris, la capitale de la France. La capitale de la France est en apposition à Paris et, en même temps, «explique» le nom propre. (Maurel *et al*, 2014).

[104] Contrairement aux autres relations, l'Eponymie nous informe que la traduction ne se réfère pas à un nom propre mais à un nom commun, une terminologie ou un idiome. (Maurel *et al*, 2014).

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.

2 L'encyclopédie Wikipédia

2.1 Introduction

«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 »[105]. La Wikipédia est une encyclopédie collaborative universelle et multilingue en ligne[106] qui 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.

Le premier wiki a été créé en 1995 par Ward Cunningham pour réaliser la section d'un site sur la programmation informatique. Dans ce contexte, on peut noter que la modification d'une page wiki se fait en cliquant sur le composant «modifier» (dans la version française ou ses équivalents dans les autres versions linguistiques), il se trouve en haut à droite de la page ; cette modification peut être effectuée en respectant les règles de contribution indiquées par le projet Wikipédia.

[105] http://igm.univ-mlv.fr/~dr/XPOSE2011/Wikipedia/presentation_wikipedia.html)
[106] http://wikipedia.org/

« Créée en 2001, elle est alimentée chaque jour par plus de cent mille contributeurs à travers le monde. Elle est visitée chaque mois par près de 500 millions de visiteurs et propose plus de 30 millions d'articles dans plus de 280 langues. Plus de 25 000 articles sont créés par jour sur les différentes versions linguistiques de la Wikipédia et on compte plus de 10 millions de modifications par mois [107] » ; à l'heure actuelle, les éditions de la Wikipédia les plus importantes en nombre d'articles sont la Wikipédia en anglais, en allemand, en français, en néerlandais et en suédois[108]. L'encyclopédie Wikipédia est hébergée par la fondation Wikimédia[109] aux États-Unis et entièrement financée par des dons ; publiés sous une licence libre (CC-BY-SA)[110], beaucoup de contenus existent déjà sur la Wikipédia, mais certains sujets ne sont pas encore approfondis ou n'ont pas d'article du tout. La Wikipédia compte sur des contributeurs bénévoles pour créer de nouveaux articles ou enrichir des ébauches[111].

Certains considèrent l'encyclopédie Wikipédia comme peu fiable, en se basant notamment sur le fait qu'elle n'est pas assez exacte et ne fait pas autorité, car n'importe qui peut écrire n'importe quoi sur la Wikipédia, et ce sans aucune forme de contrôle[112] ; pourtant, aujourd'hui, l'encyclopédie Wikipédia se classe comme le cinquième site le plus visité au monde selon le classement Alexa[113]. La Wikipédia constitue une ressource pour les chercheurs et les développeurs travaillant sur des problèmes de bases de données, d'indexation ou de classification de documents. Ce site peut être utilisé pour extraire des connaissances dans un domaine précis comme l'extraction des entités nommées ; ces données sont semi-structurées ayant un certain degré de formalité et de structure ; il est constamment actualisé ou mis à jour, ce qui est un point très important pour la maintenance des dictionnaires spécifiques (Bouamor, 2009).

En outre, sur la Wikipédia, chaque article ne décrit qu'un seul sujet, le titre de chaque article ressemble à un terme d'ontologie ; les concepts équivalents sont considérés comme homonymes

---

[107] https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Statistiques#cite_note-1
[108] https://fr.wikipedia.org/wiki/Wikipédia : À propos
[109] https://fr.wikipedia.org/wiki/Wikimedia_Foundation
[110] https://fr.wikipedia.org/wiki/Licence_Creative_Commons
[111] «Une ébauche est un article qui donne trop peu d'informations sur le sujet pour être suffisamment informatif et vérifiable. Tout contributeur peut créer une ébauche, ou en compléter une. Le bandeau ébauche apposé en tête d'article invite les contributeurs à en enrichir le contenu. », selon le site :
[112] http://www.liberation.fr/ecrans/2008/01/17/ceux-qui-disent-non-a-wikipedia_959345?page=article
[113] http://www.alexa.com/topsites

et regroupés par des liens redirigés. En effet, la Wikipédia 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.[114]. Toutes ces caractéristiques font de la Wikipédia une ontologie potentielle qui peut être exploitée pour enrichir et ainsi améliorer le contenu des autres ontologies, comme les ontologies de programmes scolaires de base (Gueffaz *et al*. 2014) ou encore les dictionnaires électroniques typiques, pour lesquels la Wikipédia représente une ressource sémantique essentielle (Savary *et al*. 2013).

En résumé, l'encyclopédie Wikipédia, avec son contenu extrêmement vaste et la quantité des liens inter-articles, est une formidable ressource multilingue d'informations et un défi pour l'exploitation.

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.

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

A. Les Infoboxes

Elles représentent les caractéristiques d'une entité donnée, correspondent aux tableaux reprenant des informations factuelles et structurées, et sont placées en général en haut à droite de certains articles. Le contenu de ces Infoboxes 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 Infobox ; les biographies font pourtant partie d'un des types d'articles les plus fréquents sur la Wikipédia[115]. Les Infoboxes ou les boîtes d'information affichent des informations pertinentes pour le sujet de l'article en utilisant la

[114] http://fr.wikipedia.org/wiki/Wikipédia:Catégories
[115] https://www.cetic.be/Exploiter-le-contenu-de-Wikipedia

fonctionnalité du logiciel modèle considérant le type d'entrée ; ces informations peuvent être des clés pour les recherches d'informations.

B. 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.

C. 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 : Auteurs et statistiques, Recherche de l'auteur d'un passage de l'article, Statistiques de consultation, Contributeurs suivant cette page et Modifications par utilisateur.

D. 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.

Le système d'évaluation fourni par l'encyclopédie Wikipédia

Ce système permet à un projet wiki de surveiller la qualité des articles dans ses domaines d'étude et ainsi de prioriser le travail sur ces articles.

Les évaluations de qualité sont effectuées principalement par des membres de Wiki Projets, qui balisent des pages de discussion d'articles. Ces balises sont ensuite collectées par un robot, qui génère une sortie telle qu'une table, un journal et des statistiques[116]

[116] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_assessment

Généralement, l'évaluation des articles des différentes éditions de la Wikipédia se fait par la détermination de l'état des articles en se basant sur leur état d'avancement qui désigne le stade quantitatif et qualitatif du contenu de l'article. En principe, chaque article reçoit un stade ou un label d'évaluation dans une version linguistique donnée et doit répondre à certains critères le concernant :

- le contenu évaluant comment le sujet de l'article a été couvert et traité ;
- les sources dépendant de la qualité et de la quantité des références qui sont citées dans l'article ;
- la mise en page en considérant le plan général et les illustrations, leur choix et leur place dans l'article ;
- le style comprend le niveau de langage, l'orthographe, la syntaxe, la typographie, etc. ;
- la wikification déterminant l'usage correct du code wiki pour créer des titres, du gras, de l'Italique et aussi l'usage des liens internes dans l'article.

Ici, nous nous intéressons à l'évaluation des articles dans l'édition Wikipédia française où l'avancement est défini par quatre stades :

1. « ébauche», qui désigne les articles bien existants, mais à peine commencés ;
2. « BD» ou bon début qui désigne les articles peu développés, représentant le plan sommaire et des informations essentielles ;
3. « B » ou bien construit qui désigne les articles bien développés ;
4. «A » ou article avancé.

De plus, un système de labellisation qui est désigné par l'avis de la communauté Wikipédia avec deux valeurs possibles :

1) «BA » pour bon article ;
2) «ADQ » pour un article de qualité.

E. 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.

F. 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.

G. 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.

H. 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 [117].

I. 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[118].

J. Liens internes

Ce sont des liens internes à la Wikipédia ou wikiliens pointant vers d'autres articles de la Wikipédia ; ils se mettent dans le corps de l'article. Leur utilisation peut parfois pécher dans

[117] https://fr.wikipedia.org/wiki/Aide:Lien_interwiki

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

leur pertinence (le lien doit apporter une information utile)[119], leur efficacité (l'article correspondant à ce lien doit exister et le lien ne doit pas être répété) ou leur esthétique[120]. 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.

K. 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[121].

Pour terminer cette section, nous illustrons en images des exemples clarifiant certains composants qui sont mentionnés plus haut ; la Figure 3.4 représente une partie de la page Platon dans l'édition Wikipédia française en entourant les liens Historique, Discussion, Pages liées, Informations sur la page, l'Infobox et un lien interne. La Figure 3.5 indique le résultat de l'évaluation de cette page dans la page de Discussion ; la page Wikipédia française de Platon reçoit le stade **A** pour l'avancement ce qui signifie « Article Avancé » ; finalement, les Figures 3.6 et 3.7 montrent respectivement la forme des références et liens externes se trouvant dans la page «Platon».

[119] Un lien vers carbone dans l'article diamant apporte une information au sujet, mais dans une phrase telle que « Pierre est mort dans un accident de voiture », faire un lien vers « voiture » n'apporte aucune information à l'article. D'après le site : http://fr.howtopedia.org/wiki/Aide:Liens_internes

[120] Par exemple, il n'est pas nécessaire de mettre en majuscule la première lettre d'un lien si les règles de français n'y obligent pas car, pour cette première lettre, et elle seule, le logiciel convertit la minuscule en majuscule. Selon le site : https://fr.wikiquote.org/wiki/Aide:Liens_internes

[121] https://fr.wikipedia.org/wiki/Aide:Insérer une référence

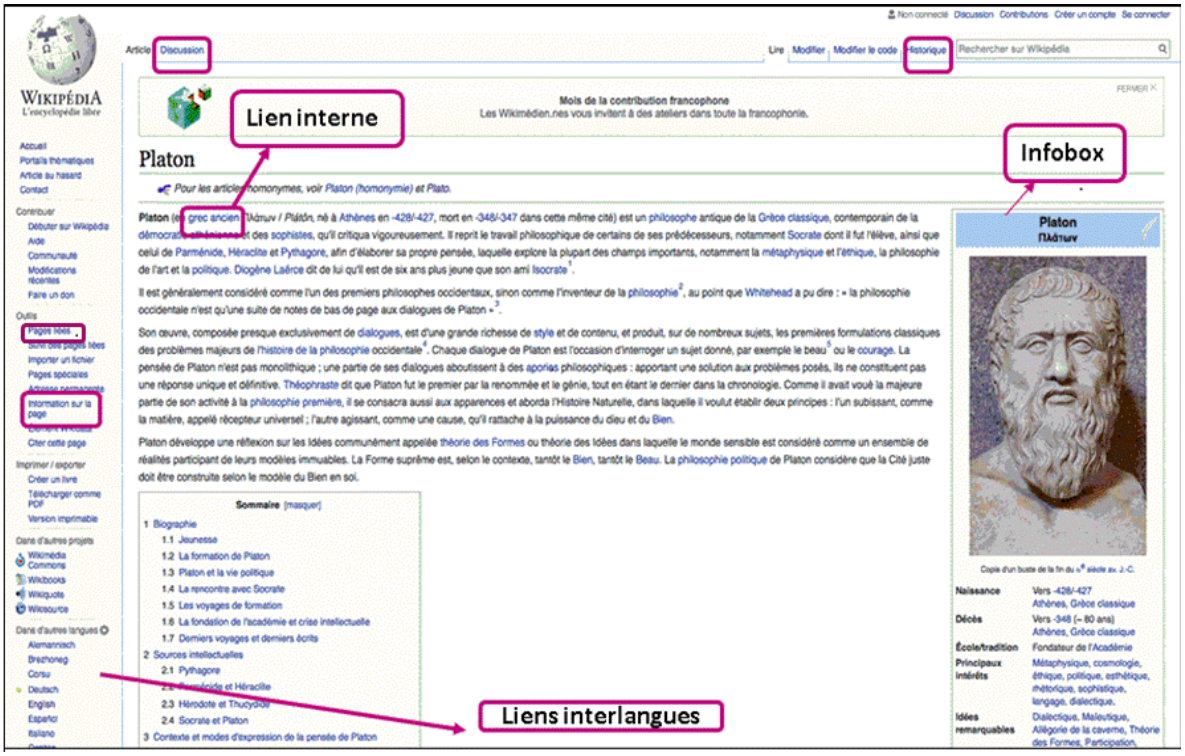


Figure 3. 4: Une partie de la page Platon (version Wikipédia française) comprenant infobox, Discussion, Historique, lien interne, Pages liées et Information sur la page et liens interlangues.

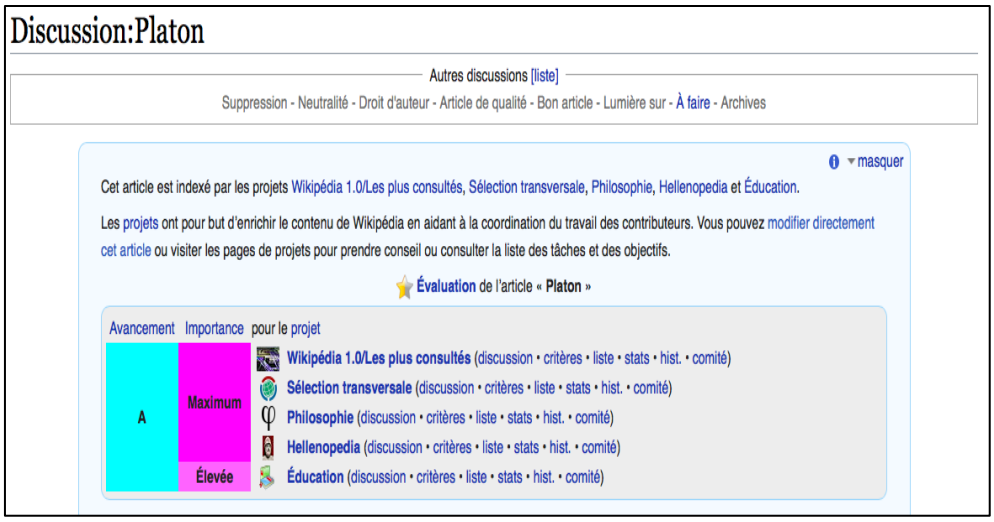
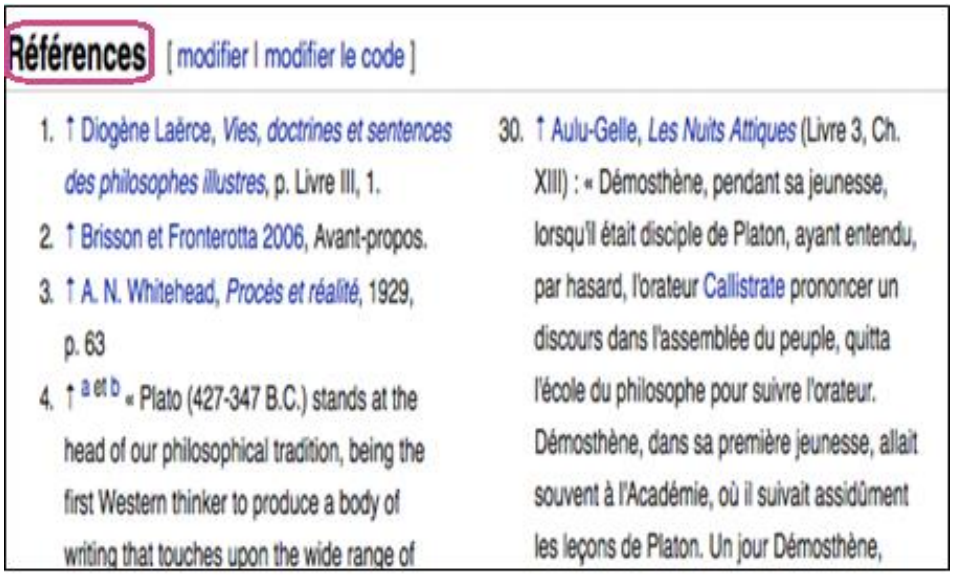


Figure 3. 5: L'évaluation de la page Platon dans la page de Discussion



Références [modifier | modifier le code]

1. ↑ Diogène Laërce, *Vies, doctrines et sentences des philosophes illustres*, p. Livre III, 1.
2. ↑ Brisson et Fronterotta 2006, Avant-propos.
3. ↑ A. N. Whitehead, *Procès et réalité*, 1929, p. 63
4. ↑ [a et b] « Plato (427-347 B.C.) stands at the head of our philosophical tradition, being the first Western thinker to produce a body of writing that touches upon the wide range of

30. ↑ Aulu-Gelle, *Les Nuits Attiques* (Livre 3, Ch. XIII) : « Démosthène, pendant sa jeunesse, lorsqu'il était disciple de Platon, ayant entendu, par hasard, l'orateur Callistrate prononcer un discours dans l'assemblée du peuple, quitta l'école du philosophe pour suivre l'orateur. Démosthène, dans sa première jeunesse, allait souvent à l'Académie, où il suivait assidûment les leçons de Platon. Un jour Démosthène,

Figure 3. 6: Les références dans la page Platon de la version Wikipédia française

Liens externes [modifier | modifier le code]

Papyri [modifier | modifier le code]

- (en) Bibliothèque d'Oxyrhinchus [archive]

Éditions et traductions en ligne [modifier | modifier le code]

- (fr) Traductions, sur Wikisource

Bibliographie [modifier | modifier le code]

- (fr) Bibliographie platonicienne [archive] par Luc Brisson

Figure 3. 7: Les liens externes de la page Platon de la version Wikipédia française

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. Les principaux outils qui permettent l'accès au contenu de l'encyclopédie sont :

2.3.1 Les dumps

La fondation Wikimédia publie des sauvegardes de la base de données qui peuvent être téléchargées[122] et utilisées pour consulter l'encyclopédie Wikipédia hors ligne après avoir installé localement le logiciel MediaWiki[123]. 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[124], et aussi, de faire tout type de traitement automatique sur le contenu. Tout le texte contenu est réutilisable selon les termes de la licence Creative Commons paternité partagée à l'identique[125].

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[126].

Enfin, plusieurs travaux ont utilisé des dumps ou des parties des versions locales de la Wikipédia en tant que source d'information externe entre autres (Leva *et al.*, 2013 ; Savary *et al*., 2013 ; Robert Viseur, 2013).

---

[122] http://dumps.wikimedia.org/
[123] https://fr.wikipedia.org/wiki/Aide:MediaWiki
[124] https://fr.wikipedia.org/wiki/Site_miroir
[125] https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Citation_et_r%C3%A9utilisation_du_contenu_de_Wikip%C3%A9dia
[126] https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Wikip%C3%A9dia_hors-connexion#1re_.C3.A9tape_:_Installer_MediaWiki

2.3.2 L'action API Médiawiki

C’est un service web qui permet d’accéder très facilement et efficacement au contenu du wiki, ses bases de données, ses métadonnées via des requêtes HTTP. La requête HTTP doit commencer par une URL principale ; par exemple, l’URL de base pour l’API Wikipédia anglaise est https://en.wikipedia.org/w/api.php; tous les wikis Wikimédia ont des URL de base qui suivent ce schéma. Ainsi, pour obtenir l’URL de base d’un autre wiki, il suffit de remplacer le code de langue « en » dans l’adresse précédente par celui du wiki cible[127].

Via ces requêtes, les clients demandent des "actions" particulières en définissant un paramètre action. En général, ce paramètre est associé à la valeur query (action = query) qui est considérée comme une des actions de l’API la plus importante[128] ; ce type d’action permet de récupérer les différentes informations sémantiques concernant l’historique, les modifications récentes, la liste des contributeurs, la taille d’une page, les pages appartenant à une catégorie, etc. Les réponses à ces requêtes peuvent être au format JSON[129] ou XML en spécifiant un paramètre format dans la requête. Dans les nouvelles installations MediaWiki, le service web est activé par défaut, mais un administrateur peut le désactiver.

Pour terminer cette section, la figure 3.8 montre un exemple d’une requête API Wikipédia française, elle se compose d’un schéma de base concaténé avec certains paramètres en demandant à ce service web de récupérer la taille de la dernière version de la page wiki « Platon » et de retourner le résultat au format JSON.

---

[127] https://nl.wikipedia.org/w/api.php est API Wikipédia Néerlandaise et https://fr.wikipedia.org/w/api.php est API Wikipédia Française

[128] https://www.mediawiki.org/wiki/API:Query

[129] JavaScript Object Notation, est un format de données textuelles dérivé de la notation des objets du langage JavaScript. Il permet de représenter de l’information structurée selon le site : https://fr.wikipedia.org/wiki/JavaScript

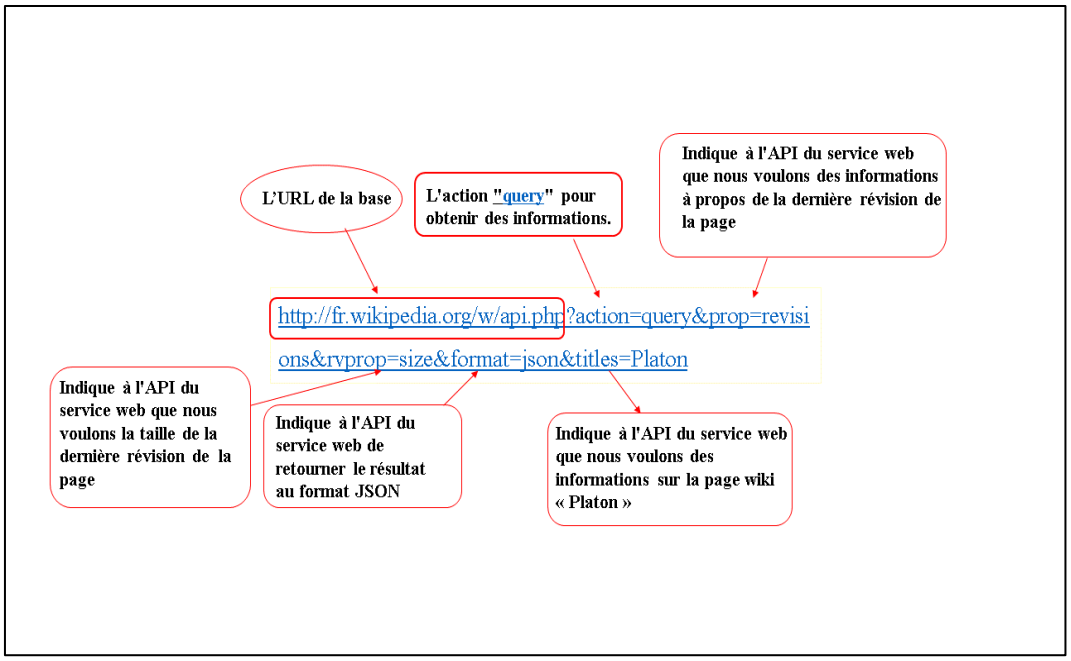


Figure 3. 8: La requête API Wikipédia française pour récupérer la taille de la dernière version de la page wiki Platon au format JSON

Finalement, en utilisant n'importe quel langage de programmation pour envoyer cette requête HTTP à cette URL, ou en consultant simplement ce lien dans un navigateur, nous obtiendrons comme ci-dessous un document au format JSON qui contient la taille de la page «Platon» qui est égale à 153 345 octets :

{"batchcomplete":"","query":{"pages":{"7201":{"pageid"
:7201,"ns":0,"title":"Platon","revisions":[{"size":153345}]}
}}}

2.3.3 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[130].

Il est possible d'accéder au contenu des articles de l'encyclopédie Wikipédia via l'utilisation de DBpedia. DBpedia est un effort communautaire qui a été lancé en 2007 par Auer *et al.* (2007) ; il consiste à extraire des informations sémantiques structurées à partir de la Wikipédia, surtout les informations contenues dans les infoboxes associées aux articles, et à les rendre accessibles.

Le contenu extrait depuis la Wikipédia est converti dans le format RDF[131]. Selon Robert Viseur (2013), on peut accéder au contenu de DBpedia (un extrait du contenu de la Wikipédia converti au format RDF) via l'une des trois solutions suivantes : la première est l'accès direct aux données RDF par URI (*Universale Resource Identifier*) ; la deuxième est l'utilisation d'agents Web comme les navigateurs pour le Web sémantique ; et la dernière solution est celle de l'interrogation de DBpedia depuis les points d'accès SPARQL[132].

Enfin, en nous inspirant de Robert Viseur (2013), nous finissons cette section en citant deux raisons pour lesquelles on ne peut considérer DBpedia que comme un outil partiel et non complet d'accès au contenu de l'encyclopédie Wikipédia. Premièrement, la couverture linguistique de DBpedia est actuellement limitée : au moment de la rédaction de cette thèse, il y a 18 langues (chapitre localisé) seulement possédant une version DBpedia[133] ; en particulier, s'agissant de notre travail, on trouve du français, de l'anglais et du polonais, mais pas d'arabe.

---

[130] https://fr.wikipedia.org/wiki/DBpedia

[131] « *Resource Description Framework* » (RDF) est un modèle de graphe destiné à décrire les ressources Web et leurs métadonnées, de façon à permettre le traitement automatique de telles descriptions. Il est développé par le W3C , RDF est le langage de base du Web sémantique.D'apres le site : https://fr.wikipedia.org/wiki/Resource_Description_Framework.

[132] « *SPARQL Protocol and RDF Query Language* » est un langage de requête et un protocole qui permet de rechercher, d'ajouter, de modifier ou de supprimer des données RDF disponibles à travers Internet. », D'après le site : https://fr.wikipedia.org/wiki/SPARQL

[133] http://oldwiki.dbpedia.org/Internationalization/Chapters

Deuxièmement, le processus d'extraction est fondé notamment sur le contenu des Infoboxes (Auer *et al.* 2007 ; Hellmann *et al.* 2009), mais, toutes les pages de l'encyclopédie Wikipédia ne proposent pas d'Infoboxes et ces dernières ne sont pas toujours complètes.

2.4 Exemples de grandes ontologies issues de l'encyclopédie Wikipédia

Hormis DBpedia, plusieurs études (Medelyan *et al.*, 2009 ; Hovy *et al*., 2013 ; Flati *et al*., 2014) indiquent que la création de grandes base de connaissances est rendue possible grâce à la disponibilité de ressources en ligne, collaboratives et multilingues comme l'encyclopédie Wikipédia. Ces ressources, bien qu'elles soient partiellement structurées, fournissent des connaissances précieuses qui peuvent être récoltées et transformées sous forme entièrement structurée.

Citons ici des exemples de ces grandes ontologies et de réseaux sémantiques qui sont principalement issus de la Wikipédia et d'autres ressources comme WordNet :

- **BabelNet** (Navigli et Ponzetto, 2012) est à la fois un dictionnaire encyclopédique multilingue qui fournit des entrées lexicalisées dans la majorité des langues et un réseau sémantique qui relie ces concepts et ces entités nommées dans un très grand réseau de relations sémantiques, composé d'environ 14 millions d'entrées, appelées synchrones de Babel ; il regroupe les mots de différentes langues par groupes de synonymes appelés Babel synsets. Pour chaque Babel synset, BabelNet fournit des définitions textuelles appelées gloses en plusieurs langues, qui sont obtenues à partir de WordNet et de la Wikipédia.

- **YAGO** (Hoffart *et al*. 2013) (Yet Another Great Ontology) est une énorme base de connaissances sémantiques, créée par l'Institut Max-Planck d'informatique à Sarrebruck. Elle est constituée à partir d'informations extraites de la Wikipédia (catégories, redirections, infoboxes), de WordNet (synsets, hyponymie) et de GeoNames. Actuellement, YAGO a une connaissance de plus de 10 millions d'entités nommées (personnes, organisations, villes, etc.) et contient plus de 120 millions de faits sur ces

entités. YAGO est une ontologie ancrée dans le temps et dans l'espace, caractérisée par les points suivants [134]:

1) la précision qui a été évaluée manuellement est confirmée à 95% ;
2) la combinaison de la taxonomie propre de WordNet avec la richesse du système de catégorie de la Wikipédia ;
3) En plus d'une taxonomie, YAGO intègre les domaines thématiques (tels que «musique» ou «science») de WordNet Domains[135] ;
4) YAGO extrait et combine des entités et des faits de 10 éditions linguistiques de l'encyclopédie Wikipédia ;
5) Enfin, YAGO est liée à des ontologies comme DBpedia, et ainsi, le type et la sous-classe des faits de YAGO sont importés dans un espace de noms propres dans DBpédia.

Conclusion

Nous avons présenté dans ce chapitre les deux ressources principales de ce travail de recherche : la base de données lexicale multilangues Prolexbase et l'encyclopédie Wikipédia. Nous nous sommes intéressés à l'ontologie de Prolexbase en distinguant les deux concepts essentiels : le concept multilingue (pivot) et le concept lexical (prolexème). Pour la Wikipédia, nous avons exposé certains composants de la structure d'une page donnée, en particulier les liens internes, les liens externes, le lien nommé (information sur la page) et les liens interlangues. Nous avons expliqué également les différentes méthodes d'accès au contenu d'un article de la Wikipédia en soulignant l'action API mediawiki que nous allons utiliser afin d'extraire les informations concernées.

[134] Selon le site :http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/
[135] http://wndomains.fbk.eu/

Deuxième partie

Réalisation du Travail

Chapitre 4 La Consolidation des Liens Wikipédia dans Prolexbase

Introduction

Ce chapitre est consacré à la présentation de la première phase de notre travail étroitement lié à Prolexbase ; comme nous l'avons déjà évoqué, notre point de départ dans ce travail est Prolexbase qui, en tant que base de données lexicale, nécessite une révision ponctuelle et régulière ainsi qu'un enrichissement de ses entrées. Ce processus est une étape fondamentale afin de rendre ce lexique plus efficace et fonctionnel pour différents types d'applications de traitement automatique de langues (TAL).

La méthode utilisée fournit une consolidation entièrement automatique comprenant deux principes :

- premièrement, une validation des liens Wikipédia qui se trouvent dans les tables des prolexèmes de différentes langues dans Prolexbase ; si des liens ne conduisent pas à la même page, ils doivent être envoyés à un expert pour les valider ;
- deuxièmement, une complétion des liens ; si des liens n'existent pas, ils doivent être complétés ; en particulier, cette complétion est une mise en correspondance automatique qui consiste, dans un premier temps, à remplir les liens manquants des entrées déjà existantes dans la base de donnée et, dans un deuxième temps, à ajouter de nouvelles entrées (numéros de pivots et liens Wikipédia) à Prolexbase si ces entrées n'existent pas dans une langue donnée mais existent dans une autre.

En résumé, c'est un mappage automatique de différentes langues dans Prolexbase depuis les liens Wikipédia.

L’approche proposée

Dans cette approche, nous avons travaillé sur la table « pivot » et les différentes tables de prolexèmes dans Prolexbase ; en effet, Prolexbase comporte actuellement 10 langues[136], dont 3 sont bien couvertes avec toutes leurs informations morphologiques et flexionnelles, à savoir : le français, l’anglais et le polonais. Pour chaque langue donnée, il y a une table nommée « prolexeme_iso » où « iso » est remplacé par le code de cette langue à trois lettres, par exemple, « prolexeme_fra » pour le français, « prolexeme_eng » pour l’anglais et « prolexeme_pol » pour le polonais. Chaque table de prolexèmes consiste en plusieurs attributs. Ici, on s’intéresse à deux d’entre eux : en premier, le numéro de pivot qui est obligatoirement associé à chaque entrée donnée et son rôle en tant qu’identificateur numérique unique est de relier la même entrée dans toutes les langues ; en deuxième, le lien Wikipédia qui est éventuellement associé à une entrée donnée.

La méthode utilisée vise d’abord à relier les mêmes entrées de tables de prolexèmes, et ainsi à compter le nombre de liens Wikipédia qui sont non nuls, associés à chacune de ces entrées dans les différentes langues constituant Prolexbase. Nous avons utilisé le concept multilingue (le numéro de pivot) en parcourant la table « pivot », et pour chaque numéro de pivot (un attribut de la table pivot), il fallait le chercher dans toutes les tables de prolexèmes en les interrogeant via une requête SQL afin de récupérer l’entrée dans les différentes langues possédant ce même numéro de pivot. En conséquence, la réponse obtenue contient le numéro de pivot recherché, l’ensemble des liens Wikipédia qui lui sont associés dans toutes les langues et le nombre de ceux qui sont non nuls.

À ce stade, pour chaque réponse obtenue en haut (numéro de pivot, ensemble de liens), un traitement automatique de l’ensemble des liens Wikipédia a été effectué incluant les trois phases suivantes :

[136] Ce sont le français, l’anglais, le polonais, le serbe, l’italien, le portugais, l’allemand, le néerlandais, le coréen et l’espagnol.

1 Le traitement des redirections

Dans le but d'assurer qu'un lien Wikipédia dans Prolexbase ne conduit pas à une page de redirection, nous avons envoyé chaque lien de l'ensemble résultant de la réponse vers une fonction appelée «update_ifredirect » afin de traiter ce problème. Le principe consiste à exploiter le contenu de la page de l'outil « Information sur la page » de ce lien, puis à vérifier s'il contient le terme «redirectsto»/ «Rediriger vers », qui signifie que ce lien est une redirection vers une autre page ; ainsi, il suffit d'extraire le lien attribué qui correspond au lien exact de la page cible. En conséquence, la fonction retourne ce lien exact si c'est le cas, sinon, elle retourne le même lien traité ; finalement, un fichier SQL nommé « update_redirect» est produit à cette étape contenant les liens exacts qui sont retournés par cette fonction.

Pour mettre au clair cette phase, la figure 4.1 montre les fondements du traitement des redirections représentant un lien de prolexeme_fra nommé «Limousin » et qui conduit à une page de redirection. La fonction a extrait le lien exact «Limousin (ancienne région administrative) » qui est ensuite stocké dans un fichier SQL de type (mise à jour/ou Update). Finalement, au niveau des résultats, 903 redirections ont été modifiées par ce même programme dans la table « prolexeme_eng », 731 redirections dans le prolexème français, 1 521 dans le polonais et 21 redirections parmi les entrées du prolexème serbe.

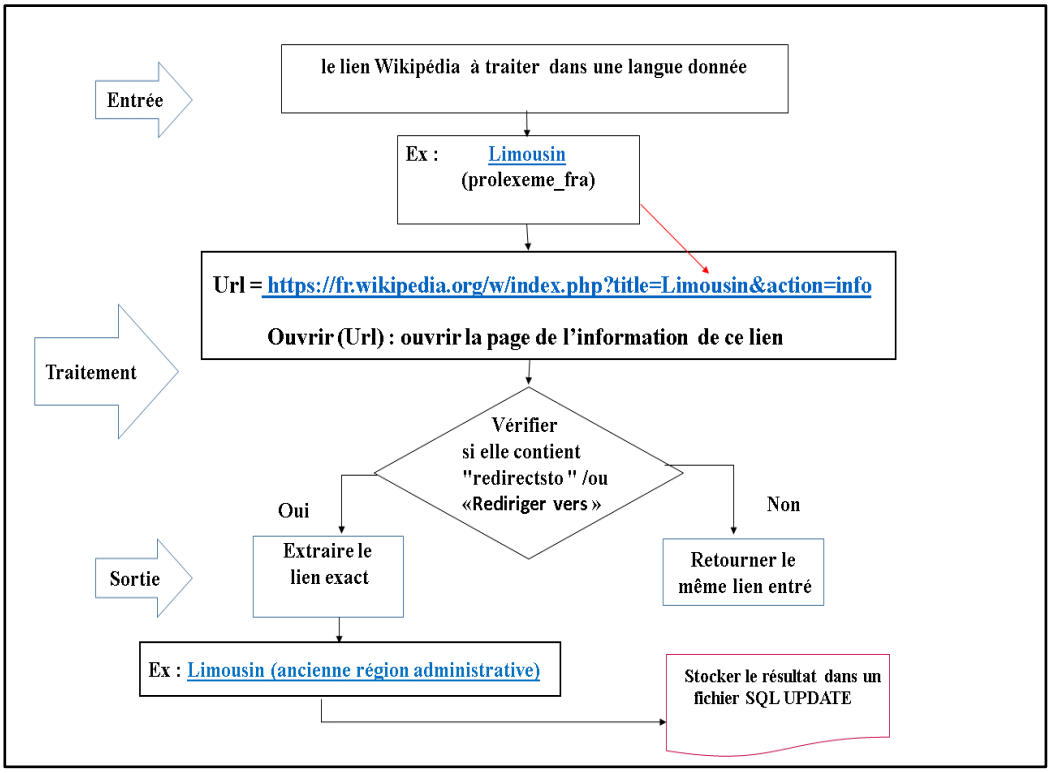


Figure 4. 1: Algorithme général du traitement des redirections avec un exemple d'une redirection de prolexeme_fra nommé «Limousin »

2 La validation de liens

Le processus de validation a été réalisé selon le nombre de liens Wikipédia non nuls qui sont reliés à chaque entrée :

1. si ce nombre est égal au nombre de langues traitées, c'est-à-dire, s'il n'y a pas de liens manquants, dans ce cas, il faut comparer tous les liens ;
2. si ce nombre est supérieur à un, c'est-à-dire, s'il n'y a pas plus d'un lien non nul, mais s'il y a des liens manquants, dans ce cas, il faut comparer les liens existants ;
3. s'il n'y a qu'un seul lien, dans ce cas, on ne compare pas.

La technique élaborée s'appuie sur la comparaison du lien cible (lien à valider) dans une langue donnée de Prolexbase aux liens de même langue qui sont précédemment générés à partir de leurs équivalents dans d'autres langues via le concept interlangue de la Wikipédia. En effet, s'ils ne sont pas identiques, ils nécessitent d'être stockés dans un fichier de contrôle afin de les

faire examiner par un expert ; pour faciliter la tâche de l'expert, le programme génère un fichier texte pour chaque langue donnée.

En particulier, si une entrée multilingue donnée (par exemple : le numéro de pivot 52307) «Platon en prolexeme_fra» relie 3 liens Wikipédia non nuls de 3 langues traitées (l'anglais, le français et le polonais) dans Prolexbase. Or, ils demandent une validation qui consiste avant tout à extraire les liens interlangues qui lui correspondent dans les trois langues depuis chacun d'entre eux, c'est-à-dire, pour chaque lien, le programme produit 2 liens interlangues, un ensemble de 9 liens de trois langues. Puis, on regroupe les liens possédant le même code linguistique ; donc, l'ensemble est divisé en 3 groupes de 3 liens ; ensuite, il faut comparer les liens de même langue constituant chaque groupe ; s'ils ne sont pas identiques, alors, comme nous l'avons évoqué ci-dessus, ce groupe de liens d'une langue donnée doit être conservé dans un fichier de contrôle typique à cette langue et par la suite envoyé à l'expert pour qu'il les vérifie et les corrige.

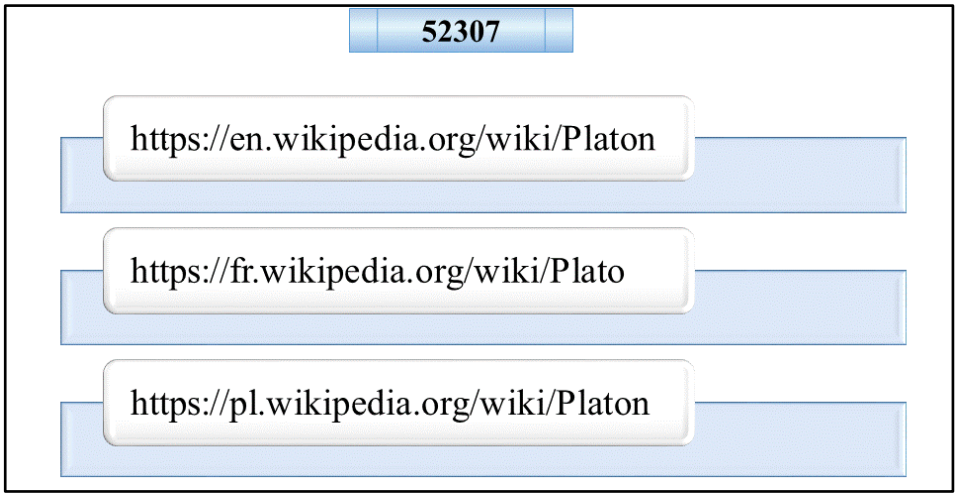


Figure 4. 2: Ensemble de 3 liens Wikipédia qui sont associés au numéro de pivot «52307 » (Platon en prolexeme_fra)

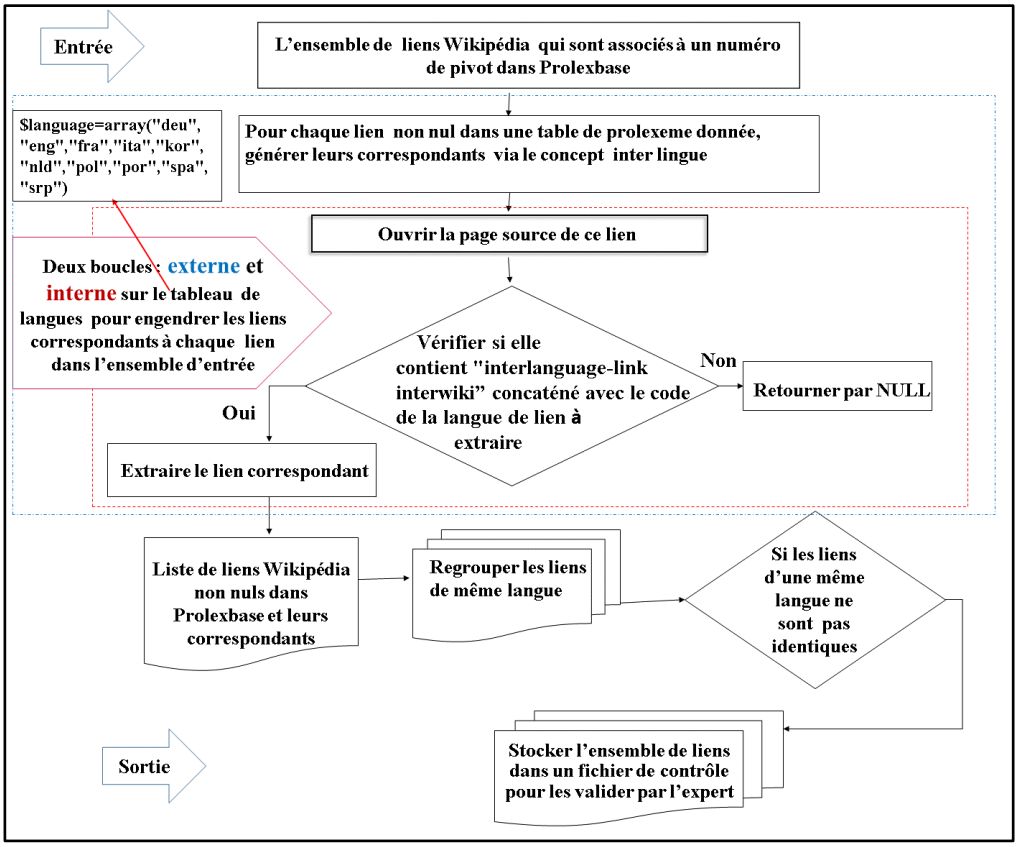


Figure 4. 3: Algorithme général de la phase « Validation de liens »

La figure 4.4 illustre un exemple du processus de validation avec un ensemble de 9 liens Wikipédia. Le numéro de pivot 19324 (entrée multilingue), reliant trois liens en trois langues (anglais, français et polonais), ainsi que leurs correspondants qui sont précédemment générés ; d'abord, il s'agit de regrouper les liens avec le même code linguistique, ensuite, de les comparer et enfin de stocker ceux qui ne sont pas identiques dans un fichier de contrôle pour les faire vérifier et aussi corriger par un expert.

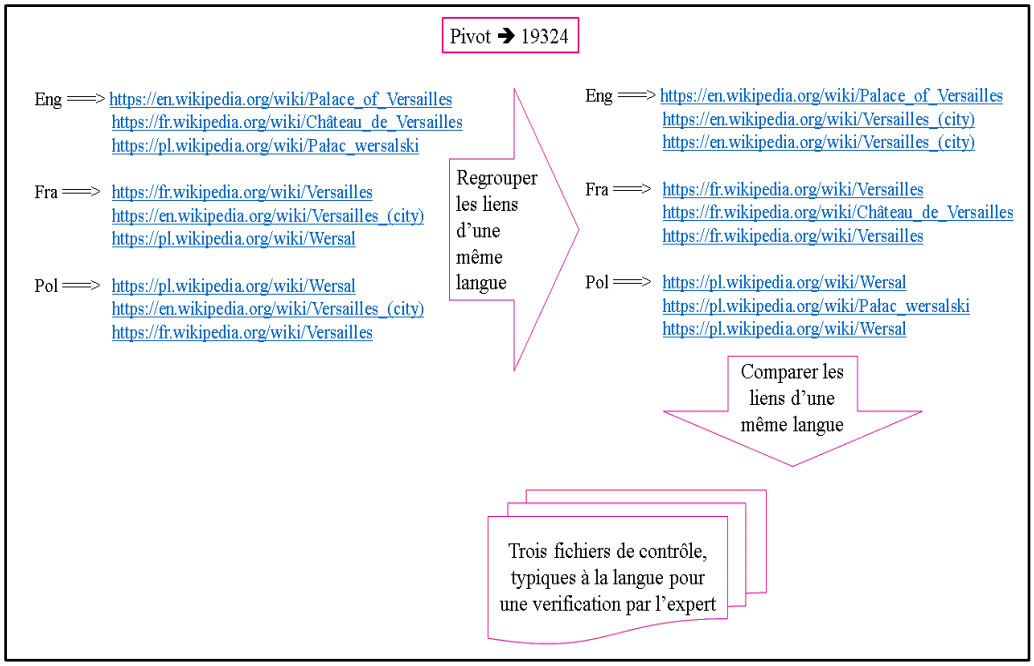


Figure 4. 4 : Exemple des liens considérés non valides par le processus« Validation de liens »

Pour terminer cette section, il est important d'indiquer que l'expertise a été réalisée par un étudiant en Master de linguistique (Alexandre Herry-Cailler) ; nous lui avons fourni un fichier de contrôle composé de 130 entrées multilingues (numéros de pivots), chacune d'entre elles étant reliée à un ensemble de 9 liens en trois langues (l'anglais, le français et le polonais).

Tous les liens qui ont été repérés comme pouvant être « non valides » par notre procédure ont été validés ou corrigés par l'expert ; en conséquence, un fichier résultant de liens valides a été importé dans Prolexbase pour une mise à jour grâce à ce travail.

3 La complétion de liens

Après la validation des liens Wikipédia se trouvant dans les tables de prolexèmes de Prolexbase, nous avons établi un processus d'enrichissement de ses entrées, toujours dans l'objectif de l'améliorer et de le rendre plus pertinent et fonctionnel. Ce processus est fondé sur deux principes :

- remplir les liens manquants dans des entrées d'une table de prolexèmes d'une langue donnée lorsqu'ils sont fournis dans les tables de prolexèmes d'une autre langue ;
- ajouter des nouvelles entrées (numéros de pivot et liens Wikipédia) à une table de prolexèmes d'une langue donnée lorsqu'elles sont fournies dans les autres tables des autres langues.

En d'autres termes, une mise en correspondance entre les différentes tables de prolexèmes des différentes langues constituant Prolexbase au niveau des liens Wikipédia, a été réalisée à cette phase. Dans ce contexte, il est nécessaire de citer quelques chiffres relatifs aux tables prolexeme_iso dans Prolexbase, par exemple, la table «prolexeme_pol » comporte le plus grand nombre de liens Wikipédia avec 18 528 liens sur 27 266 entrées, suivie par celle « prolexeme_eng » contenant 17 806 liens sur 19 397 entrées. On trouve ensuite le français avec 11 907 liens sur 70 973 entrées ; le serbe qui arrive en quatrième place, dispose seulement de 584 entrées avec tous leurs liens associés. Pour leur part, les tables prolexeme_deu, prolexeme_ita, prolexeme_spa, prolexeme_por prolexeme_nld et prolexeme_kor comprennent zéro lien avec, respectivement, 807, 754, 744, 689,525 et 228 entrées

L'algorithme de complétion a été établi en partant du nombre de liens Wikipédia non nuls qui a été déterminé au départ de notre programme ; en effet, si le nombre de liens non nuls était égal à un, c'est-à-dire, qu'il n'y avait qu'un seul lien dans une langue donnée, il a fallu compléter ses équivalents dans les autres langues dans la base de données. Également, si ce nombre était supérieur à un et inférieur au nombre de langues traitées, alors, il a fallu compléter les liens manquants.

La figure 4.5 montre en détail les fondements de l'algorithme appliqué dans les deux cas ; la complétion est basée sur le concept interlangue caractérisant la structure des pages de la Wikipédia ; les liens manquants correspondants sont générés à partir d'un lien dans une table de prolexèmes d'une langue donnée à condition que celui-ci soit confirmé comme valide par la phase de validation. Notons que, quand il n'existe qu'un lien non nul, nous utilisons le premier code linguistique enregistré dans un tableau contenant les codes des langues de liens non nuls, ce tableau a été constitué au départ de notre programme ; un autre tableau rassemble les codes

de langues des entrées qui n'existent pas dans une langue donnée, et servira à décider dans quel fichier de résultats on doit mettre un nouveau lien engendré.

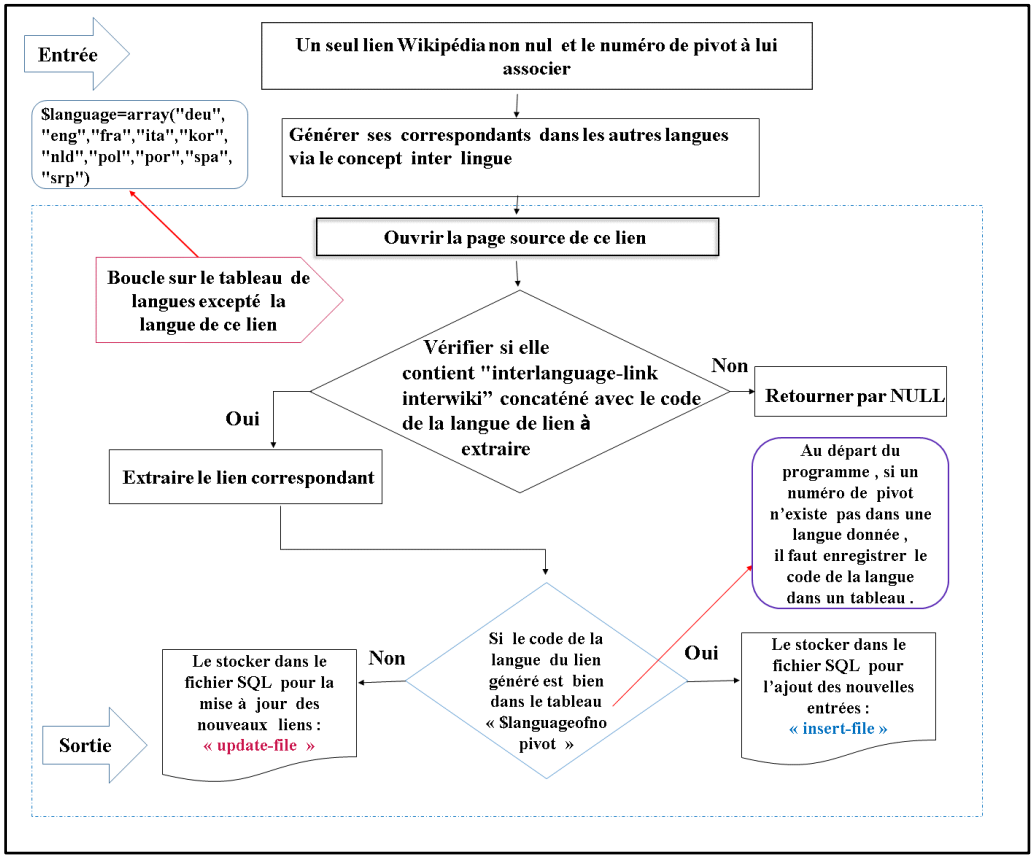


Figure 4. 5: Algorithme général de la phase « complétion de liens »

Au niveau des résultats, la procédure produit deux fichiers SQL : le premier pour mettre à jour les nouveaux liens à remplir «update_file » et le second pour ajouter les nouvelles entrées «insert_file» ; le tableau 4.1 décrit une comparaison du nombre de liens Wikipédia dans les différentes tables de prolexèmes avant et après l'importation de ces deux fichiers résultant de la complétion.

Le tableau 4.1 montre clairement que des résultats importants ont été obtenus concernant le nombre de liens Wikipédia qui ont été remplis ou ajoutés dans les tables : prolexeme_deu, prolexeme_por, prolexeme_spa, prolexeme_srp, prolexeme_ita, prolexeme_kor, et prolexeme_nld; mais des résultats moins importants pour les tables prolexeme_eng, prolexeme_fra et prolexeme_pol.

prolexeme _iso	Nombre de liens Wikipédia dans Prolexbase avant la complétion de liens	Nombre de liens Wikipédia dans Prolexbase après la complétion de liens	Nombre de nouveaux liens	Nombre de nouvelles entrées
prolexeme_eng	17 806	18 383	123	454
prolexeme _fra	11 907	14 963	2 767	289
prolexeme_pol	18 528	19 024	111	385
prolexeme_srp	584	8 530	1	7 945
prolexeme_por	0	11 495	434	11 061
prolexeme _spa	0	10 649	600	10 049
prolexeme _deu	0	13 700	662	13 038
prolexeme _ita	0	11 958	608	11 350
prolexeme _kor	0	8 743	217	8 526
prolexeme_nld	0	13 309	563	12 746

Tableau 4. 1: Comparaison du nombre de liens Wikipédia dans les tables de prolexeme_iso avant et après la complétion de liens.

Conclusion

Dans ce chapitre, nous avons présenté les trois phases principales d'un système multilingue entièrement automatique que nous avons établi pour réaliser et maintenir une consolidation efficace des liens Wikipédia dans Prolexbase. Nous avons détaillé le traitement des redirections, la validation des liens et en particulier, la complétion de liens qui a offert des résultats importants via le remplissage des liens manquants et l'ajout de nouvelles entrées dans les différentes tables de prolexèmes dans la base de données. Par la suite, nous allons démontrer dans le chapitre 6 la fonctionnalité et l'efficience de ce système rendu capable d'ajouter de nouvelles langues et de nouveaux dialectes.

Chapitre 5 Estimation de la Notoriété d'un Nom Propre via Wikipédia

Introduction

Ce chapitre propose un système automatique qui permet de calculer, via la Wikipédia, un indice de notoriété pour les entrées du dictionnaire relationnel multilingue de noms propres Prolexbase. Cet indice de notoriété dépend de la langue et participera, d'une part, à la construction d'un module de Prolexbase pour la langue arabe et, d'autre part, à la révision de la notoriété actuellement présente pour les autres langues de la base. En effet, comme nous l'avons précédemment mentionné, dans cette base, à chaque entrée lexicale (prolexème) est associé, éventuellement, un lien vers la Wikipédia et, obligatoirement, un indice de notoriété basé sur trois valeurs, conformément à la norme ISO 12620 : l'indice le plus fort étant 1 et le plus faible 3. Les indices actuels ont été choisis manuellement, ce que nous avons voulu changer, d'abord pour permettre l'ajout automatique de nouvelles entrées et de nouvelles langues, mais aussi pour autoriser une réévaluation régulière de cette notoriété, qui évolue dans le temps.

Le point de départ de ce travail était l'ajout de la langue arabe ; notons qu'un premier travail d'ajout de langue a été réalisé par Savary *et al*. (2013) qui a fortement augmenté le nombre d'entrées polonaises si l'on se base sur le total des consultations de l'encyclopédie Wikipédia dans l'année précédant leurs travaux. Actuellement, Prolexbase contient dix langues, mais c'est principalement le français, l'anglais, le polonais qui sont bien couvertes et sont ainsi les seules considérées dans cette partie du travail de thèse.

La recherche de la notoriété est basée sur les liens vers la Wikipédia présents dans Prolexbase ; plus précisément, en préliminaire à notre calcul de notoriété, nous avons dû choisir un ensemble de critères (cinq valeurs numériques) déduites de la Wikipédia, et ensuite, établir

un système de calcul qui a été fondé sur deux techniques multicritères de décision : la méthode SAW et l'entropie de Shannon.

Finalement, nous avons mené une application des résultats issus de notre système de calcul de la notoriété comprenant des exemples dans trois langues principales de type célébrité dans Prolexbase ; celle-ci est suivie par une discussion du choix de la notion « coefficient d'oubli » qui a été appliquée pour l'obtention du nombre de pages vues, l'un des cinq critères. Finalement, nous terminerons ce chapitre par une évaluation des résultats provenant de ce travail via une comparaison avec ceux du projet Panthéon (Yu *et al*., 2016), voir notre Chapitre 1- section 1.5.

1 L'approche proposée

L'approche utilisée pour l'estimation de la notoriété a été menée respectivement à travers les étapes suivantes.

1.1 Le choix des critères

Nous avons commencé notre estimation de la notoriété en choisissant cinq critères : les trois premiers critères concernent l'article lui-même et les deux autres les liens de cet article avec les autres articles et l'ensemble de la Toile. Précisons que nous avons fait ce calcul à l'intérieur d'une même édition linguistique, considérant que la notoriété d'un nom propre dépend de la langue. Comme il nous semble peu intuitif de comparer la notoriété d'une célébrité à celle d'un lieu, nos calculs se font type par type. Pour cette partie du travail, nous nous focalisons sur le type célébrité.

Voici les cinq critères :

1) le nombre de consultations de l'article ;
2) le nombre de contributeurs à l'article ;
3) la taille de l'article ;
4) le nombre de liens internes à la Wikipédia pointant vers l'article ;
5) le nombre de liens externes à la Wikipédia contenus dans l'article.

Le premier indice ne se limite pas à une année, mais à l'ensemble des données disponibles sur le nombre de consultations mensuelles de l'article depuis 2008. Les deux indices suivants permettent d'estimer la fiabilité de l'article consulté et les deux derniers d'estimer son intégration dans la Wikipédia et dans la Toile. En préliminaire, la sélection de ces cinq critères a été inspirée

des précédents travaux dans ce domaine, entre autres, (Chevalier *et al*. 2010 ; Savary *et al*., 2013 ; Eom, Shepelyansky, 2013 ; Yu *et al*. 2015) ; ensuite, nous avons testé manuellement le lien entre les statistiques de ces indices et la qualité de l'article Wikipédia concerné sur une trentaine d'articles. Dès lors, nous avons constaté une corrélation entre les valeurs de ces indices et la qualité d'un article Wikipédia ; en outre, un article ayant un grand nombre de contributeurs, une grande taille et/ou un grand nombre de consultations peut intuitivement être considéré comme un article de qualité.

Autrement dit, un article d'une taille importante réfère à un sujet important qui a attiré une contribution considérable et *vice versa*, un sujet important doit certainement inciter un grand nombre de participants à enrichir ou à modifier son article, etc.

De même, la taille des contributions montre s'il y a eu d'importants changements ou simplement des corrections mineures ; aussi, le nombre de contributeurs différents traduit la diversité d'un article, « les articles de qualité ont substantiellement plus de contributeurs impliqués »[137].

En considération de l'opportunité de l'indice de liens pointant vers une page donnée de la Wikipédia, nous avons choisi de calculer le nombre de ses liens entrants ; de même, nous avons considéré l'intégration d'un article dans les autres sites web en optant pour le nombre de liens externes sortant de cet article vers ceux-ci comme l'un des critères de notoriété.

Finalement, c'est la combinaison de ces cinq indices qui nous permettra de calculer la notoriété que nous attribuons au nom propre (section 1.3).

1.2 Le calcul des cinq indices

Comme nous l'avons déjà illustré dans le chapitre 4, chacune des trois tables de prolexèmes anglais, français et polonais est alimentée par un grand nombre de liens Wikipédia ; ici, nous avons employé ces liens afin de récupérer des informations de la Wikipédia.

En effet, pour chaque nom propre possédant un lien vers la Wikipédia dans une table de prolexèmes donnée, nous avons calculé alternativement les cinq critères de notoriété sélectionnés en section 1.1 ; le calcul de chacun d'entre eux est expliqué en détail comme suit.

[137] Selon Wilkinson et Huberman (2007), cités par Chevalier *et al*. (2010) dans l'article intitulé «Visualisation de mesures agrégées pour l'estimation de la qualité des articles Wikipédia ».

1.2.1 Le calcul du nombre de consultations

Comme indiqué ci-dessus, nous disposons pour le calcul du premier indice, via des services web, du nombre de consultations mensuelles de l'article depuis 2008. Un premier service[138] permet de connaître ce nombre à partir de décembre 2007 pour l'anglais et de février 2008 pour les autres langues et ce, jusqu'à fin 2015. Pour les consultations postérieures, nous utilisons un service web de Wikimédia[139] qui commence sa compilation en 2016.

Nous avons choisi de prendre comme point de départ de notre compilation janvier 2008 et comme fin le dernier mois complet avant la date du jour. Dans cette section, nous considérerons la période 2008-2015 qui couvre 96 mois. Tout d'abord, l'agrégation des statistiques de consultations pour un article cible a été effectuée mois par mois à partir de 2008 et jusqu'à 2015 via l'outil «Statistiques de consultation » de la Wikipédia. Rappelons ici que cette partie du travail a été réalisée langue par langue et type par type. A la fin de cette étape, nous avons créé une table SQL pour chaque type dans une langue donnée. Ces tables contiennent les consultations mensuelles concernées ; la figure 5.1 montre en détail l'algorithme appliqué afin d'obtenir cet ensemble de fichiers et de les importer dans Prolexbase.

[138] La concaténation http://stats.grok.se/fr/201501/Paris nous permet de connaître le nombre de consultations de l'article Paris dans l'édition française de Wikipédia en janvier 2015.

[139] De même, via https://wikimedia.org/api/rest_v1/metrics/pageviews/per-article/fr.wikipedia/all-access/all-agents/Paris/daily/2016010100/2016013100, nous obtenons le nombre de consultations de l'article Paris dans l'édition française de Wikipédia en janvier 2016.

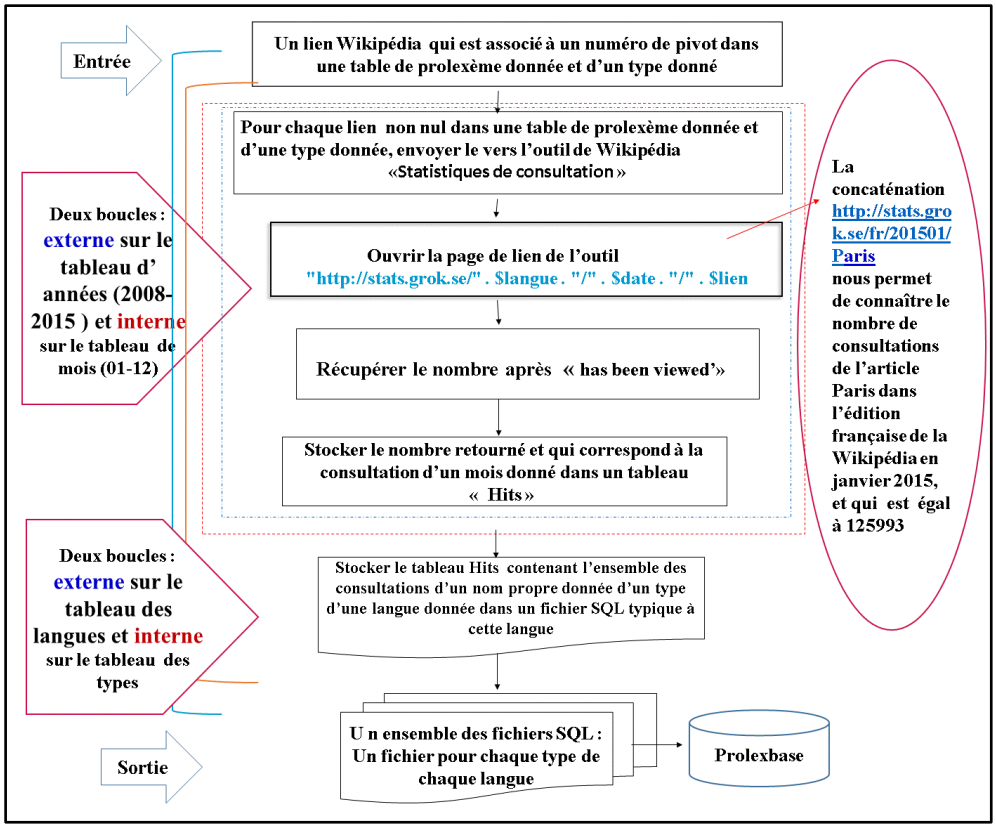


Figure 5. 1: Algorithme appliqué pour construire l'ensemble des fichiers SQL contenant les consultations mensuelles des types de chaque langue dans Prolexbase.

Ensuite, et comme la notoriété d'un nom propre peut varier dans le temps, nous n'attribuons pas la même valeur à chaque mois, mais nous associons à un mois ce que nous avons appelé son coefficient d'oubli. Ainsi, janvier 2008 sera associé au coefficient 1/96, février 2008 au coefficient 2/96, mars 2008 au coefficient 3/96, etc ; en nous servant des tables de consultations obtenues précédemment, nous calculons donc pour chaque nom propre et pour chaque mois, le produit du nombre de consultations de son article par le coefficient d'oubli du mois en question. Finalement, nous faisons la somme de ces valeurs. C'est cette somme qui constituera notre premier indice (Figure 5.2).

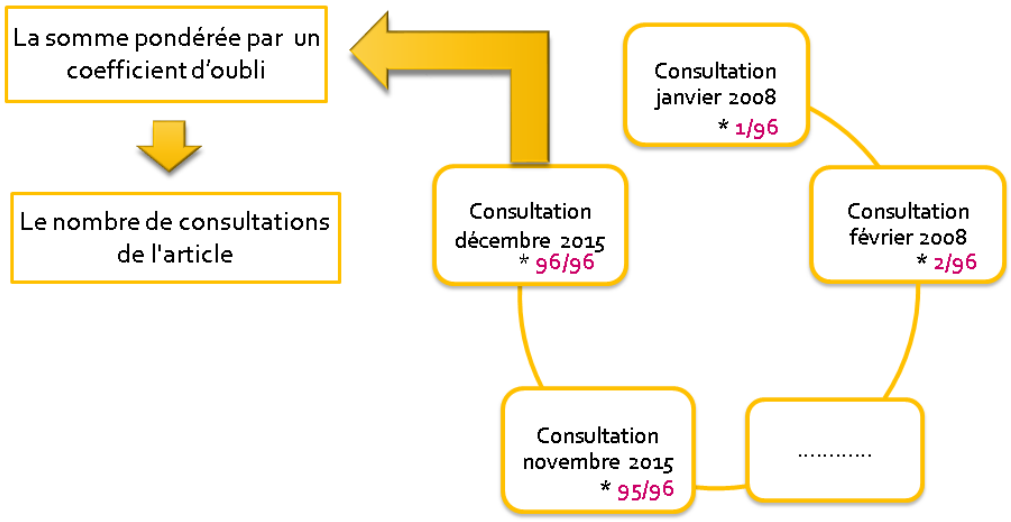


Figure 5. 2 : Le nombre de consultations d'un article donné est égal à la somme pondérée par un coefficient d'oubli

Avant de terminer cette section, il convient de souligner que le choix de cette conception pour construire le premier critère de notoriété « nombre de consultations d'un article » a été fait suite à plusieurs expérimentations avec d'autres notions statistiques. Par exemple, dans un premier temps, nous avons décomposé l'ensemble des statistiques en 8 années (2008-2015) et dans un deuxième temps, nous avons calculé la moyenne des moyennes annuelles comme le premier indice. Cette idée a été rapidement abandonnée à cause des valeurs extrêmes ou aberrantes « outliers » dans les statistiques mensuelles de chaque année.

Une autre idée consistait à se baser sur l'utilisation de ce premier critère comme un filtre pour choisir les entrées à conserver ou à intégrer dans Prolexbase avant qu'il ne soit utilisé pour le calcul de la notoriété avec les 4 autres critères. Plus précisément, nous avons au départ calculé la moyenne de l'ensemble des consultations mensuelles et l'écart type pour chaque table de prolexème, ensuite, en utilisant le z score[140], nous avons éliminé les noms propres qui possèdent un z score inférieur ou égal à zéro. Cependant, nous avons abandonné ce concept et finalement

[140] Z-score = $(\boldsymbol{x}-\boldsymbol{\mu})/\ \sigma$ tel que x est le nombre de pages vues en un mois donné, μ est la moyenne des statistiques et σ est l'écart type.

décidé que le filtrage dépendrait de l'indice final de la notoriété. En d'autres termes, une fois l'estimation terminée, on pourra éliminer les entrées ayant l'indice de notoriété le plus faible.

1.2.2 Le calcul des autres indices

Pour un article donné, la Wikipédia fournit un service web[141] permettant d'obtenir la liste de tous ses contributeurs, et donc, leur nombre. Le même service permet d'obtenir la taille de l'article, le nombre de liens entrants et le nombre de liens externes. En effet, le principe général effectué pour récupérer ces informations d'une page de la Wikipédia est l'action API de Media Wiki (chapitre 3), l'un des outils d'exploitation du contenu de la Wikipédia qui se base sur une requête soumise via une URL ; elle permet d'interroger la Wikipédia, et de fournir une réponse dépendant des paramètres utilisés et de leurs valeurs associées.

La figure 5.3 décrit les étapes suivies afin de calculer ces indices pour chaque type de chaque langue donnée ; ensuite, la figure 5.4 présente les adresses URL de l'action API pour récupérer les quatre valeurs numériques relatives au nombre de contributeurs, à la taille, au nombre de liens internes et au nombre de liens externes de l'article Platon de l'édition française Wikipédia.

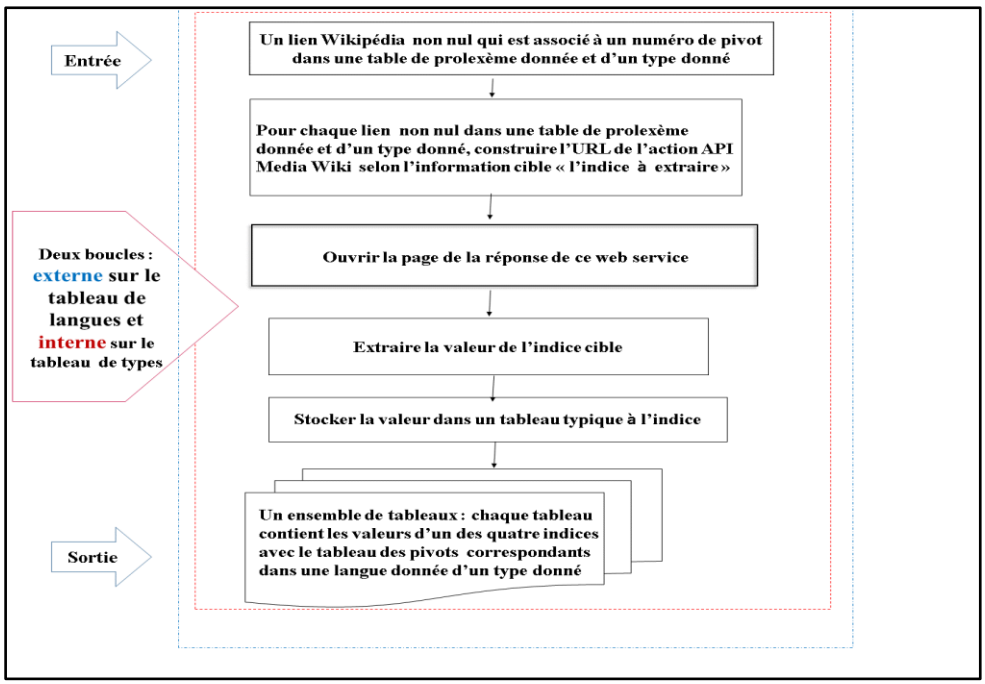


Figure 5. 3 Algorithme du calcul des quatre indices : le nombre de contributeurs, la taille de l'article, le nombre de liens internes et le nombre de liens externes

[141] https://www.mediawiki.org/wiki/API:Main_page/fr

Figure 5. 4: Adresses API pour récupérer les quatre indices de notoriété indiqués de l'article Platon de l'édition française Wikipédia.

Nous terminons cette partie par l'illustration de trois exemples de la version Wikipédia française avec les valeurs de ces cinq indices récupérés par notre programme. Cette illustration a pour objectif de mettre au clair les différentes valeurs possibles qu'un article de la Wikipédia peut avoir selon la réputation et la célébrité de son sujet.

Autrement dit, les valeurs des cinq critères sélectionnés peuvent varier selon l'importance du sujet de l'article en question. Par exemple, en observant les chiffres dans le Tableau 5.1, nous constatons qu'un article qui comporte le sujet « Platon » en tant que l'un des premiers philosophes mondiaux est associé à un grand nombre de consultations (1 481 827), et de contributeurs (513), à une taille importante (1 546 77 Octets/Bytes), un nombre élevé de liens internes (3 120) et aussi un certain nombre de liens externes (43). Egalement, un article est consacré à David Beckham, footballeur international, dans l'édition Wikipédia française avec des valeurs intéressantes qui sont bien montrées dans le tableau. Au contraire, l'article consacré à Stefan Niesiolowski, politicien polonais non connu en France, est une ébauche dans l'édition française de la Wikipédia qui ne contient à la date actuelle [06/05/2017] aucun lien externe, et seulement 4 liens internes, 11 contributeurs et donc, une taille quasi nulle de 487 octets et la somme pondérée du nombre de consultations (1 298) est très petite en comparaison avec celle des deux autres articles.

Nom propre	Nombre de consultations	Nombre de contributeurs	Taille de l'article	Nombre de liens internes	Nombre de liens externes
Platon	1 481 827	513	1 546 77	3 120	43
David Beckham	1 398 370	499	82 251	518	63
Stefan Niesiolowski	1 298	11	487	4	0

Tableau 5. 1 : Ensemble de trois noms propres et leurs valeurs de cinq indices de notoriété récupérées via notre programme dans l'édition française de la Wikipédia.

1.3 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*) (Afshari, 2010), qui nécessite d'attribuer un poids à chaque critère. Ce poids est parfois défini arbitrairement par l'utilisateur. Nous avons préféré le déduire du calcul de l'entropie de Shannon (1948), comme, entre autres, Safari *et al*. (2012) et Karami et Johansson (2014).

1.3.1 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.

En effet, comme nous l'avons déjà indiqué, l'attribution des poids d'importance peut être réalisée par estimation arbitraire, cependant, nous avons utilisé la méthode de l'entropie de Shannon qui est capable de calculer ces poids automatiquement via l'application de certaines formules comme suit :

1.3.1.1 Le calcul des poids de chaque critère

Pour commencer, pour chaque nom propre *i* et chaque critère *j*, nous normalisons les valeurs x_{ij} obtenues précédemment en une valeur c_{ij} comprise entre 0 et 1 ; si *m* est le nombre total de prolexèmes considérés dans une langue donnée :

$$c_{ij=}\frac{x_{ij}}{\sum_{i=1}^{i=m} x_{ij}} \text{ pour i} = 1..\text{m}, \text{j} = 1..5$$

Puis, nous calculons l'entropie E_j (comprise entre 0 et 1) :

$$E_j = \left(\frac{-1}{\ln(m)}\right) \sum_{i=1}^{i=m} \left[c_{ij} \ln(c_{ij})\right] \text{ pour j} = 1..5$$

$$\text{avec, par convention, } c_{ij}\ln(c_{ij}) = 0 \text{ pour } c_{ij} = 0$$

Et le poids W_j de chaque critère qui est calculé via la formule suivante :

$$W_j = \left(\frac{1-E_j}{\sum_{j=1}^{j=5}(1-E_j)}\right) \text{pour j} = 1..5$$

1.3.1.2 Le calcul des scores de la méthode SAW

Nous commençons par multiplier chaque valeur normalisée c_{ij} par le poids W_j du critère correspondant et qui a été calculé juste avant et nous obtenons le score S_i d'un nom propre en additionnant ces cinq valeurs :

$$S_i = \sum_{j=1}^{j=5} c_{ij} * W_j \text{ pour i} = 1..\text{m}$$

1.3.2 La répartition des prolexèmes entre les trois valeurs de notoriété : 1, 2 ou 3

Ici, nous nous intéressons aux résultats finaux de la notoriété en visant à obtenir 3 valeurs de notoriété (ou fréquence) : 1 pour la plus forte notoriété, 2 pour une notoriété moyenne et 3 pour une notoriété faible.

La répartition des prolexèmes entre les trois valeurs de notoriété n'est pas uniforme. Nous considérons qu'il y a un petit nombre de prolexèmes de notoriété 1, un nombre plus important de notoriété 2 et un grand nombre de notoriété 3 (la plus faible). Pour cela, nous attribuons tout d'abord la notoriété 1 aux prolexèmes de scores supérieurs à la moyenne plus l'écart-type de l'ensemble des scores :

$$\begin{cases} \bar{M} = Moyenne(S_i) \text{ pour i} = 1..\text{m} \\ \bar{E} = Ecart\_type(S_i) \text{ pour i} = 1..\text{m} \\ \quad Si\ S_i > \bar{M} + \bar{E}, N_i = 1 \end{cases}$$

Ensuite, nous attribuons la notoriété 2 aux prolexèmes de scores supérieurs à la moyenne plus la moitié de l'écart-type de l'ensemble des scores restants et la notoriété 3 aux autres prolexèmes.

$$\begin{cases} \bar{\bar{M}} = Moyenne(S_i) \text{ pour } \bar{S}_i \leq \bar{M} + \bar{E} \\ \bar{\bar{E}} = Ecart\_type(S_i) \text{ pour } \bar{S}_i \leq \bar{M} + \bar{E} \\ \quad Si\ \bar{M} + \bar{E} \geq S_i > \bar{\bar{M}} + \frac{1}{2}\bar{\bar{E}}, N_i = 2 \\ \quad\quad Si\ S_i \leq \bar{\bar{M}} + \frac{1}{2}\bar{\bar{E}}, N_i = 3 \end{cases}$$

1.4 Application

Nous avons testé notre calcul de notoriété comme annoncé. Le Tableau 5.2 donne les poids respectifs de chaque critère dans chaque langue, ces poids sont obtenus automatiquement via la technique de l'entropie de Shannon précédemment évoquée. Le Tableau 5.3 désigne la répartition des prolexèmes suivant leur notoriété.

Critères	(1)	(2)	(3)	(4)	(5)
Français	0,296	0,157	0,216	0,276	0,055
Anglais	0,289	0,202	0,160	0,284	0,066
Polonais	0,219	0,111	0,284	0,290	0,097

Tableau 5. 2 : Les poids respectifs de chaque critère dans chaque langue

Notoriété	1	2	3	Total
Français	10,22 %	23,48%	66,3%	3 825
Anglais	11,97%	26,68%	61,35%	4 385
Polonais	8,75%	25,23%	66,02%	4 400

Tableau 5. 3: La répartition des prolexèmes suivant leur notoriété

Comme illustré dans le Tableau 5.4, Platon et Louis XIV sont parfaitement connus dans ces trois langues, alors que Napoléon 1er est plus célèbre en français et en polonais qu'en anglais. Un autre exemple est celui d'Antoine de Saint-Exupéry qui est très connu en français, l'est moins en anglais et encore moins en polonais ; quant à Aldo Moro, il a une légère notoriété française et une très faible en anglais et en polonais ; *a contrario*, Czeslaw Niemen et Czeslaw

Kiszczak, respectivement très connu et connu en polonais, ne le sont guère en français et en anglais.

Nom propre	Notoriété		
	Français	Anglais	Polonais
Platon, Louis XIV	1	1	1
Napoléon 1er	1	2	1
Antoine de Saint-Exupéry	1	2	3
Aldo Moro	2	3	3
Czeslaw Niemen	3	3	1
Czeslaw Kiszczak	3	3	2

Tableau 5. 4 : Comparaison des notoriétés entre des noms propres de trois langues

2 Discussion du choix du coefficient d'oubli

Dans cette partie, nous avons voulu valoriser le choix de l'utilisation d'un coefficient d'oubli pour estimer le premier critère de la notoriété, celui du nombre de consultations d'une page de la Wikipédia. Pour cela, une répartition par année, de la période de 2008 à 2015, du nombre de consultations a été réalisée en appliquant le même principe évoqué précédemment, c'est-à-dire en multipliant chaque valeur mensuelle du nombre de pages vues pour un nom propre donné par son coefficient d'oubli en nous basant sur l'idée que la notoriété d'un nom propre varie dans le temps et qu'il y a certainement un oubli à un moment donné à travers cette période indiquée. Ci-dessous, le tableau 5.5 est constitué de trois noms propres et d'un ensemble de huit valeurs (Hits_2008, Hits_2009,…, Hits_2015) représentant les nombres de consultations annuelles leur correspondant dans la Wikipédia française. Chaque valeur annuelle est la somme des consultations mensuelles multipliées par leurs coefficients d'oubli respectifs ; par exemple, la valeur « HITS 2008 » est égale à la somme des consultations de janvier 2008 multipliée par 1/96 et jusqu'à celle de décembre 2008 multipliée par 12/96 ; de même, Hits_2015 est la somme des consultations de janvier 2015 multipliée par 85/96 à la consultation du mois de décembre multipliée par 96/96. Les valeurs sont normalisées et un écart moyen a été calculé afin d'identifier les entrées ayant un grand pic à une année donnée. Nous avons montré ces

différences via des graphiques illustrant les données associées à chaque nom propre et la position périodique de la différence.

Nom Propre	HITS 2008	HITS 2009	HITS 2010	HITS 2011	HITS 2012	HITS 2013	HITS 2014	HITS 2015	Ecart moyen
Michael Jackson	350	1 634	461	323	314	270	280	247	287
Christiane Taubira	248	173	137	167	602	543	110	149	153
Nelson Mandela	88	201	242	348	339	285	163	151	76

Tableau 5. 5: Ensemble de trois noms propres et leurs consultations annuelles normalisées et les écarts moyens correspondants

Nous avons choisi de transformer les données dans le Tableau 5.5 en des graphiques pour clarifier la position d'un pic important pendant la période indiquée.

Commençons par Michael Jackson, le graphique ci-après montre un pic conjoncturel dans le nombre de consultations en 2009, ce qui peut être nettement justifié par son décès cette année-là.

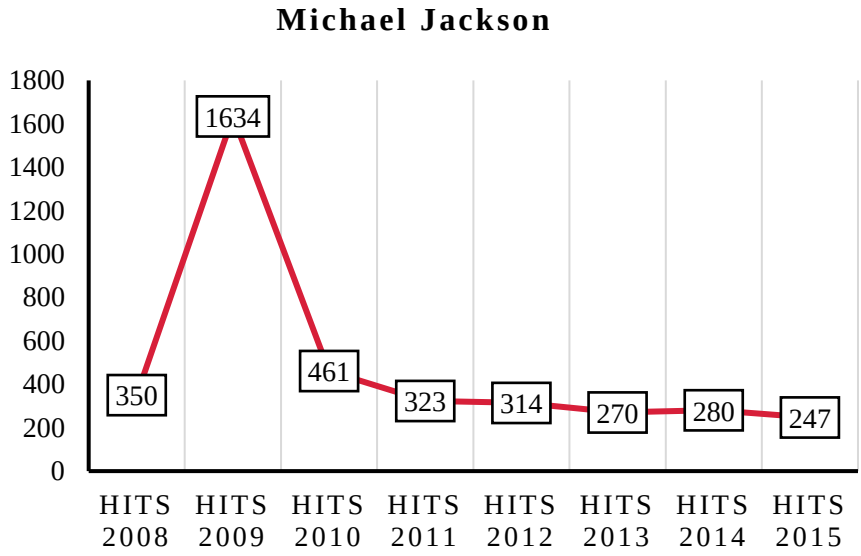


Figure 5. 5: Graphique illustrant les nombres de consultations annuelles normalisées de (2008-2015) de Michael Jackson

Ensuite, le graphique suivant montre un pic important de consultations de l'article de Christiane Taubira en 2012, alors qu'elle était nommée ministre de la justice.

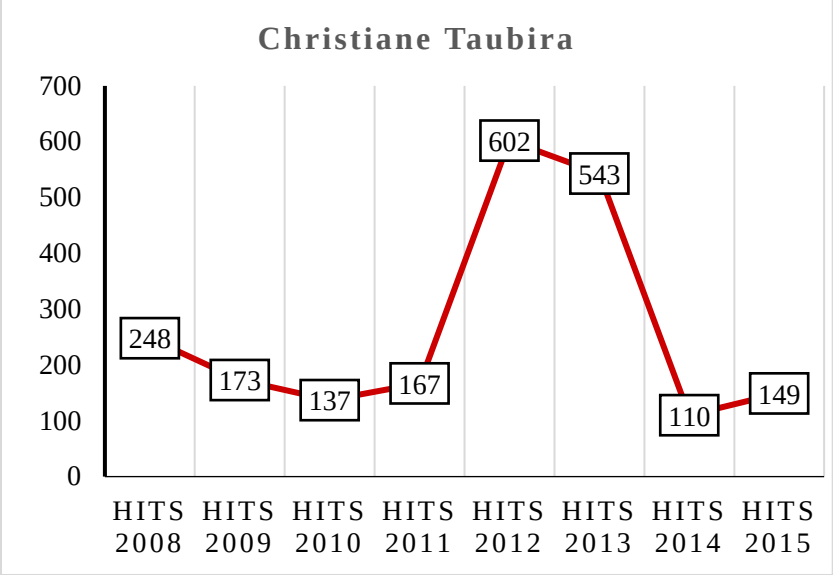


Figure 5. 6: Graphique illustrant les nombres de consultations annuelles normalisées de (2008-2015) de Christiane Taubira

Nous terminons cette partie par la figure 5.7 qui montre une augmentation continue du nombre de consultations à partir de 2011 et jusqu'à 2013, période durant laquelle Nelson Mandela a été malade et hospitalisé plusieurs fois avant de mourir en 2013.

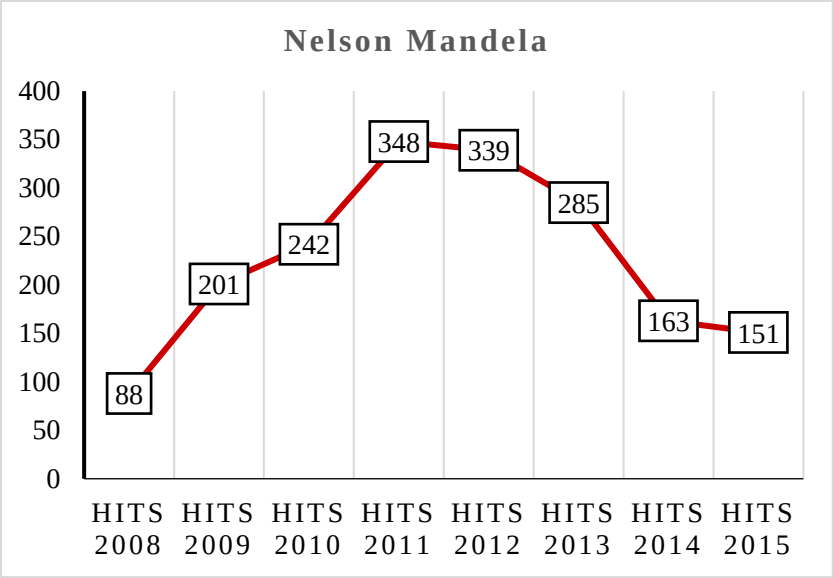


Figure 5. 7: Graphique illustrant les nombres de consultations annuelles normalisées de (2008-2015) de Nelson Mandela

3 Comparaison entre les anciennes et les nouvelles fréquences

Nous proposons dans cette partie une comparaison entre les anciennes et les nouvelles valeurs de notoriété des noms propres dans Prolexbase. Une simple évaluation nous a paru intéressante pour voir la différence entre ces deux estimations.

En effet, la comparaison consiste à calculer le nombre de noms propres qui ont changé d'1 point, c'est-à-dire, combien d'entre eux sont passés de la notoriété 1 à 2, 2 à 3 ou l'inverse (écart 1) ; également, il nous faut compter ceux qui ont changé de 2 points (écart 2), en passant de la notoriété 1 à 3 ou de 3 à 1 ; l'écart 0 est pour ceux qui possèdent deux valeurs identiques de notoriété. Nous avons considéré les noms propres de type célébrité dans les trois langues principales (français, anglais et polonais) ; le tableau 5.6 montre les résultats de cette comparaison en constatant qu'un grand nombre de noms propres ont conservé leurs indices de notoriété, un nombre moyen a eu un changement de 1 point et un petit nombre a eu un changement de 2 points.

Prolexeme_iso	Nombre de noms propres	Écart 0	Écart 1	Écart 2
FRA	3 825	59,03%	35,21%	5,61%
ENG	4 385	67,84%	28,38%	3,79%
POL	4 400	46,28%	42,05%	11,67%

Tableau 5. 6: Comparaison entre les anciennes et les nouvelles fréquences de noms propres de type célébrité

Rappelons que le changement d'indice de notoriété peut être du a une évolution de celle-ci à travers le temps.

4 La comparaison avec le projet Panthéon

Dans le but de valoriser les résultats obtenus via notre calcul de notoriété, nous avons réalisé une comparaison avec le classement «People » de Panthéon (Yu *et al*., 2016). L'idée principale a été basée sur un alignement entre la liste des noms propres de type célébrité de prolexeme_fra et celle de prolexeme_eng et la liste des cent premiers noms (Top 100) [142] de Panthéon.

Comme nous l'avons déjà mentionné, notre calcul de notoriété dépend de la langue, aussi, il est important de noter que, par la répartition des résultats que nous avons effectuée pour obtenir les trois valeurs de notoriété (1, 2 ou 3), nous avons obtenu un petit nombre de noms propres de notoriété 1, un nombre moyen de notoriété 2 et un grand nombre de notoriété 3 (section 1.3.2). Dès lors, il n'était pas évident de les comparer avec Panthéon en tant que projet mondial. Par ailleurs, il convient d'indiquer ici, que nous avions envisagé d'effectuer une comparaison avec les travaux de Eom et Shepelyansky (2013), mais les résultats n'étaient pas disponibles.

Pour mettre en œuvre la comparaison avec Panthéon, nous avons dans un premier temps, établi le pourcentage de noms propres qui avaient une forte notoriété (1) dans les deux tables de prolexèmes considérées et qui dans le même temps figuraient dans le top 100 de Panthéon à

[142] La liste provient du site : http://pantheon.media.mit.edu/rankings/people/all/all/-4000/2010/H15

savoir : 48 %. Ensuite, nous avons compté ceux qui ont obtenu la notoriété 2 dans les deux tables et qui figuraient parmi le top 100, et nous avons capté 21% de noms de célébrités.
Puis, nous avons constaté que 16% de célébrités avaient une notoriété forte (1) dans une langue et une notoriété moyenne (2) dans l'autre ; citons ici : Pythagore, Hannibal Barca et Cléopâtre VII qui ont obtenu la notoriété 1 en français, et 2 en anglais ; au contraire, Bouddha, Gengis Khan et David ont obtenu la notoriété 1 en anglais mais 2 en français. Nous n'avons retenu que 4% de célébrités qui ont obtenu la notoriété 2 ou la notoriété 3, citons ceux du prolexème français : Sappho et Johannes Gensfleisch Gutenberg de notoriété 2, et Zarathushtra et Ésope de notoriété 3. Finalement, 3% de noms propres ont eu la notoriété la plus faible (3) dans les deux langues : Lao Tseu (ou Lao Zi), Héraclite et Léonidas ; 8 % ne figuraient pas dans Prolexbase. Ces résultats sont montrés dans le tableau 5.7 ci-dessous.

Dans les prolexeme_fra et prolexeme_eng	Notoriété 1	Notoriété 2	Notoriété 1 ou Notoriété 2	Notoriété 2 ou Notoriété 3	Notoriété 3	Ne sont pas dans Prolexbase
Dans le top 100 de Panthéon	48%	21%	16%	4%	3%	8%

Tableau 5. 7: Résultats de la comparaison avec le top-100 de Panthéon (Type célébrité)

Conclusion

En utilisant principalement la Wikipédia, nous avons pu associer automatiquement un critère de notoriété aux prolexèmes de Prolexbase. Comme attendu, la valeur de ce critère diffère suivant les langues. La prochaine étape pour l'ajout de l'arabe dans Prolexbase sera la création des prolexèmes correspondant à des liens Wikipédia entre les éditions française ou anglaise et l'édition arabe ; il s'agira d'extraire de l'article le nom propre correspondant.

Chapitre 6 L'ajout de la langue Arabe dans Prolexbase

Introduction

Ce chapitre vise à mettre en évidence la dernière phase de notre travail de thèse, celle de l'ajout de la langue arabe dans Prolexbase. La méthode utilisée consiste dans un premier temps à la création d'un volume arabe comprenant deux étapes, à savoir :

l'ajout des URL de l'édition arabe de la Wikipédia, grâce au système automatique que nous avons établi dans la première phase et dont le détail est expliqué au chapitre 4 ;

le calcul de la notoriété des articles correspondant à ces liens en utilisant notre système de calcul qui a été présenté dans le chapitre 5.

Dans un deuxième temps, nous allons extraire les labels des prolexèmes arabes (lemmes) à partir de leurs liens Wikipédia correspondants obtenus précédemment.

Enfin, nous devons créer les tables des instances, des dérivées et des alias ; à cet égard, nous allons présenter nos premières avancées en citant certaines règles de flexion morphologique des noms propres arabes que nous avons illustrées par des graphes de flexion dans Unitex.

A cet effet, il conviendra de noter que nous n'avons pas encore développé le niveau morphologique dans ce module, étant donné que nous travaillons sur les articles de l'encyclopédie Wikipédia où les textes ne sont pas voyellés. Or, l'attribution des fonctions morphologiques en arabe dépend largement de la diacritisation du texte. L'approche proposée dans cette phase de notre travail vise les trois principes suivants.

1 La création d'un volume arabe dans Prolexbase

1.1 L'extraction des liens Wikipédia arabes de leurs équivalents dans les autres langues de Prolexbase

A cette étape, nous avons produit les liens Wikipédia arabes à partir de leurs équivalents dans les autres langues de Prolexbase. La procédure appliquée vise à parcourir la table pivot, pour chaque numéro de pivot dans Prolexbase, après le traitement des redirections et la validation des liens Wikipédia qui lui sont associés effectuée par le système de consolidation (Chapitre 4), puis à parcourir le tableau des langues contenues dans Prolexbase. Depuis un lien dans une langue donnée, et via le concept interlangue relatif à la Wikipédia, nous récupérons son équivalent dans la Wikipédia arabe s'il existe. Cependant, il se peut qu'une langue ne possède pas de lien Wikipédia, mais qu'un tel lien existe dans une autre langue et conduise à une page en arabe. S'il n'existe pas, nous devons l'extraire de la langue suivante et le stocker dans un fichier SQL d'ajout (Insert).

A l'issue de cette étape, nous avons un fichier SQL d'ajout (Insert) qui contient 8 409 entrées arabes (numéros de pivots et liens Wikipédia). Il est important de noter que cette même technique nous a permis de produire un fichier SQL de 1 632 entrées de l'arabe égyptien[143], et un autre de 601 entrées de la langue wolof[144]. Il suffit de parcourir un tableau des codes des nouvelles langues incluant ceux de ces deux dernières (arz et wo) et de récupérer les liens Wikipédia dans la langue cible en appliquant la procédure précédente. Les trois fichiers sont importés dans Prolexbase. Ci-dessous, la figure 6.1 présente l'algorithme utilisé pour obtenir ces résultats.

---

[143] «L'arabe égyptien est un dialecte parlé en Égypte. Il s'agit de la variété d'arabe dialectal ayant le plus de locuteurs : plus de 78 millions en Égypte et dans d'autres pays où des communautés égyptiennes habitent. De plus, en raison du rayonnement culturel de l'Égypte dans le monde arabophone, spécialement par le cinéma et la musique, il n'est pas rare qu'il soit compris par des personnes parlant d'autres variétés d'arabe. ». D'après le site : https://fr.wikipedia.org/wiki/Arabe_%C3%A9gyptien

[144] « Parfois écrit ouolof) est une langue parlée au Sénégal, en Gambie et en Mauritanie. », selon le site : https://fr.wikipedia.org/wiki/Wolof_(langue)

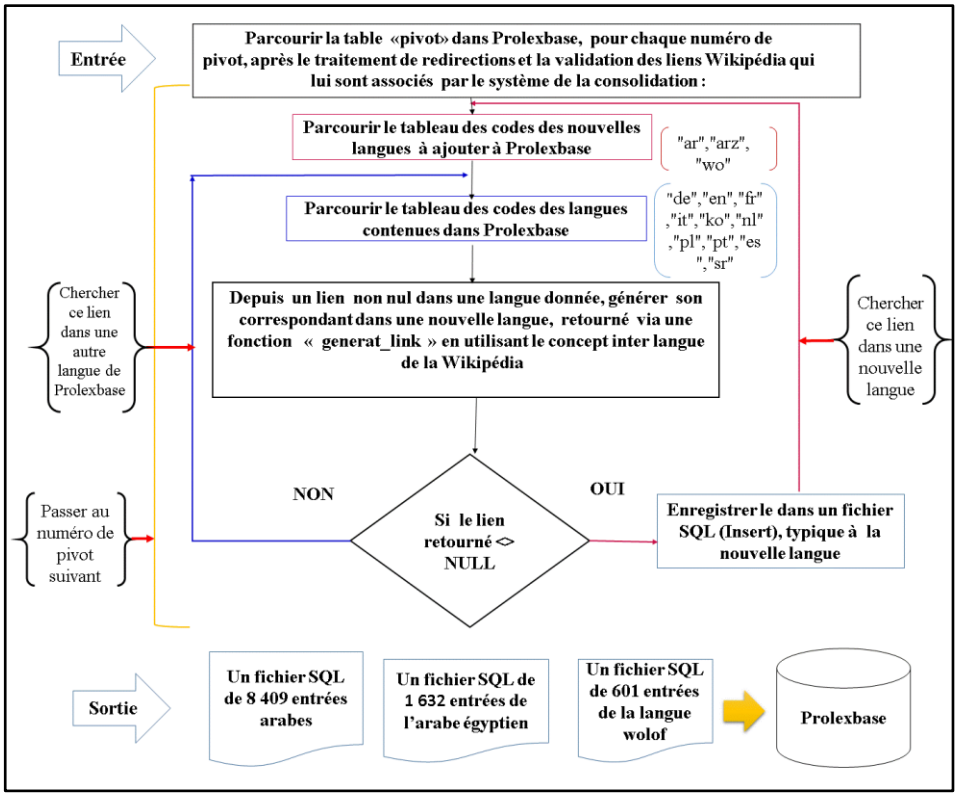


Figure 6. 1: L'algorithme de l'ajout des liens Wikipédia des nouvelles langues (l'arabe, l'arabe égyptien et le wolof) à Prolexbase

1.2 Le calcul de notoriété des noms propres arabes

A cette étape, nous avons rempli l'attribut «NUM_FREQUENCY» dans la table prolexème arabe, qui représente la valeur estimée de la notoriété d'un nom propre donné. Ce calcul est garanti grâce au système d'estimation de notoriété que nous avons établi dans le chapitre 5 ; il a suffi ici de faire tourner le programme de notoriété sur les entrées de la table prolexème arabe ; ce qui a produit un fichier SQL de mise à jour comportant 8 409 valeurs estimées de notoriété qui sont associées aux entrées traitées. Dans ce contexte, il nous paraît important de préciser que la répartition des noms propres arabes selon leur notoriété est comme suit : 6,43% de notoriété 1, 20,32% de notoriété 2 et 73,25% de notoriété 3.

Pour terminer cette partie, les tableaux 6.1, 6.2 et 6.3 illustrent des exemples pour comparer les résultats des trois types (célébrité, ville et organisation) de quatre langues (arabe, anglais, français et polonais).

En particulier, le tableau 6.1 ci-dessous montre que Platon et Saddam Hussein ont obtenu la plus grande notoriété dans ces quatre langues, d'autre part, Jacques Chirac est parmi les personnes les plus célèbres en français, alors qu'il est moins connu en arabe, en anglais et en polonais, Yasser Arafat et Zinedine Zidane sont moins célèbres en polonais, mais ils sont largement connus en arabe, en anglais et en français. En outre, Hannibal est plus populaire en arabe et en français qu'en anglais ou en polonais et Avicenne est très apprécié en arabe, mais il a obtenu la plus faible notoriété en polonais.

Noms Propres (Célébrité)		Notoriété			
		ARB	ENG	FRA	POL
Saddam Hussein (صدام حسين), Platon (أفلاطون)		1	1	1	1
Yasser Arafat (ياسر عرفات), Zinédine Zidane (زين الدين زيدان)		1	1	1	2
Hannibal Barca	حنبعل	1	2	1	2
Avicenne	ابن سينا	1	2	2	3
Angela Merkel	أنغيلا ميركل	2	1	1	1
Mohamed Ali	محمد علي	2	1	2	2
Jacques Chirac	جاك شيراك	2	2	1	2
Aldo Moro	الدرو مورو	3	3	2	3
Michelle Obama	ميشيل أوباما	3	1	3	3

Tableau 6. 1: Comparaison de notoriété entre des noms propres arabes et leurs traductions dans les prolexèmes anglais, français et polonais (Type célébrité).

Nom Propre (Ville)		Notoriété			
		ARB	ENG	FRA	POL
Oran	وهران	1	3	1	3
Monastir	المنستير	1	3	2	3
Alep	حلب	1	2	2	2
Beyrouth	بيروت	1	1	2	2
Naples	نابولي	2	1	1	1
Cordoue	قرطبة	2	3	1	2
Valence	فالنسيا	2	2	2	1
Cambridge	كامبريدج	3	2	3	3
Mascate	مسقط	3	3	3	3

Tableau 6. 2 : Comparaison de notoriété entre des noms propres arabes et leurs traductions dans les prolexèmes anglais, français et polonais (Type ville).

Nom Propre (Organisation)		Notoriété			
		ARB	ENG	FRA	POL
Organisation des Nations unies	الأمم المتحدة	1	1	1	1
Ligue arabe	جامعة الدول العربية	1	2	2	2
Organisation mondiale du commerce	منظمة التجارة العالمية	1	1	1	2
Organisation des Nations unies pour l'éducation, la science et la culture	يونسكو	2	1	1	2
Banque centrale européenne	البنك المركزي الأوروبي	3	3	2	3

Tableau 6. 3: Comparaison de notoriété entre des noms propres arabes et leurs traductions dans les prolexèmes anglais, français et polonais (Type organisation).

2 L'extraction des noms propres arabes correspondant aux liens Wikipédia arabes qui sont importés dans Prolexbase

Après l'introduction des liens Wikipédia arabes issus de la première étape dans Prolexbase, nous devons ajouter leurs noms propres correspondants. La méthode que nous avons conçue est basée sur l'exploitation des liens Wikipédia arabes.

2.1 La méthode proposée

Nous allons extraire les titres arabes représentant les noms propres arabes qui sont inclus dans ces liens. Par exemple, les noms propres Louis XIV(لويس الرابع عشر) Robert Schuman (روبير

شومان) et Zinédine Zidane (زين الدين زيدان) ont été respectivement extraits de leurs propres liens Wikipédia arabes illustrés dans le tableau ci-dessous.

Lien Wikipédia arabe	Nom propre arabe	Nom propre français
https://ar.wikipedia.org/wiki/لويس_الرابع_عشر_ملك_فرنسا	لويس الرابع عشر	Louis XIV
https://ar.wikipedia.org/wiki/روبير_شومان_(سياسي_فرنسي)	روبير شومان	Robert Schuman
https://ar.wikipedia.org/wiki/زين_الدين_زيدان	زين الدين زيدان	Zinédine Zidane

Tableau 6. 4: les liens Wikipédia arabes et les noms propres arabes extraits de ces liens

Plus précisément, un traitement du texte des URL a été effectué comprenant un processus d'extraction qui consiste à éliminer les parenthèses, les contenus entre les parenthèses, les titres nobiliaires (pour les noms propres de type célébrité), les marqueurs de lieu (pour les noms propres de type ville, pays et région), les caractères spéciaux, les soulignés, etc.

Après la création des listes de marqueurs de lieu, de titres de noblesse et autres, nous effaçons ces derniers dans les textes des URL en remplaçant tous les caractères soulignés par des espaces. Trois listes (tableaux) de mots déclencheurs ont été définies :

1) un tableau des titres de noblesse définis comme : ("الملك","الملكة","الأميرة","الأمير","الأمبراطور","الأمبراطورة","السير","القديس") qui sont respectivement traduits en français par :
   (Le saint, le sire, l'impératrice, l'empereur, le prince, la princesse, la reine, le roi) ;
2) un même tableau avec des titres nobiliaires indéfinis comme : ("ملك","ملكة","أميرة","أمير","إمبراطور","إمبراطورة","سير","قديس"), qui sont respectivement traduits par : (saint, sire, impératrice, empereur, prince, princesse, reine, roi) ;
3) un tableau des marqueurs de lieu qu'on peut voir dans un lien Wikipédia arabe comme : ("ولاية","مقاطعة","دولة","مدينة"),qui équivalent respectivement à :(ville, pays ou état, comté, état).

Par ailleurs, nous avons gardé les marqueurs de lieux dans certains cas comme les noms propres des hydronymes (fleuves, îles) ; en effet, il y a des noms propres qui sont des noms composés où le marqueur de lieu fait partie du nom propre comme جزر القمر (Comores), qui signifie en

français «îles de la lune » et donc il faut conserver le marqueur de lieu جزر(îles) en tant que partie-tout de ce nom propre.
Un fichier SQL d'ajout contient 8 409 «LABEL_PROLEXEME» ou lemmes, et a été importé dans Prolexbase après une évaluation des résultats obtenus à l'issue de cette étape.

2.2 Evaluation des résultats de la méthode précédente

2.2.1 Evaluation intrinsèque

Le tableau 6.5 montre une évaluation manuelle des noms propres arabes obtenus lors de la procédure précédente sur trois types (Célébrité, Ville et Organisation) ; les taux de précision sont importants ; des erreurs détectées dans les entrées du type Ville pourraient être facilement évitées en ajoutant des marqueurs de lieux manquants dans la liste déjà construite dans la procédure d'extraction. Par exemple, nous avons obtenu "district de Yaren" (ضاحية يارين) alors que le nom propre approprié est Yaren (يارين). Ainsi, nous devons ajouter le marqueur de lieu «district -ضاحية » à la liste cible. Les deux erreurs qui ont été captées dans les noms de célébrités sont : le nom propre « Diana Princess of Wales ديانا أميرة ويلز » qui devrait être Diana Spencer (ديانا سبنسر) et le nom propre George H. W. Bush (جورج بوش الأب) qui a été extrait, or, le nom propre correct est celui de George Herbert Walker Bush.(جورج هربرت واكر بوش) . Toutefois, la correction de ce type d'erreurs exige l'extraction du nom propre du premier paragraphe de son article de l'encyclopédie Wikipédia arabe. Une des perspectives de notre recherche constituera à analyser ce premier paragraphe.

Type	Nombre d'entrées	Nombre d'erreurs	Précision
Célébrité	2 830	2	99,92%
Ville	2 303	5	99,7%
Organisation	60	0	100%

Tableau 6. 5: Evaluation des résultats des entrées arabes sur trois types

(Célébrité, Ville et Organisation)

2.2.2 Évaluation extrinsèque

Dans le but de faire une évaluation extrinsèque de notre base de données lexicale arabe que nous avons construite à l'étape précédente, nous avons d'abord créé des dictionnaires de noms propres arabes sur la plate-forme libre Unitex. Nous avons réalisé l'évaluation considérée avec le corpus annoté ANERCorp, les Gazetteers ANERGazet[145] et le corpus Harry bin Yaqdhan[146]. Avant de comparer nos résultats avec le corpus ANERCorp, un prétraitement a été nécessaire en supprimant les étiquettes[147] qui s'y trouvaient.

Puis, nous avons comparé avec les mêmes mots en français sur le corpus 80jours distribué avec Unitex. Finalement, les résultats sur les corpus arabes sont identiques à ceux du français car le nombre de noms propres est petit. Les résultats sont présentés dans les deux tableaux ci-dessous.

Type	Nombre de noms propres dans le corpus Harry ibn Yqdhan	Nombre de noms propres dans le corpus ARNER	Nombre de noms propres dans ANERGazet	Taille du dictionnaire de noms propres arabes
Célébrité	1	134	15	2 830
Ville	4	365	811	2 302
Pays	2	252	303	629
Organisation	0	10	0	60

Tableau 6. 6: L'évaluation avec le corpus ARNER, les ANERGazet1 et le corpus Harry ibn Yqdhan

[145]Le corpus a été déjà cité dans le chapitre 1, page 21 et nous l'avons téléchargé depuis le site : http://users.dsic.upv.es/grupos/nle/?file=kop4.php.

[146] Un corpus libre de droits (32 pages) composé de 18 261 formes simples et partiellement voyellées et qui a été utilisé dans les travaux de (Doumi *et al.*, 2013) (chapitre de l'état de l'art, page 22).

[147] En fait, quatre types d'entités nommées arabes ont été définies (personne, lieu, organisation et autre) et chaque entité nommée est annotée par deux étiquettes selon sa catégorie, par exemple, la catégorie personne est structurée par B-PERS et I-PERS et la catégorie lieu est annotée par B-LOC et I-LOC (B pour Begin et I pour Inside).

Type	Nombre de noms propres dans le corpus 80 jours	Taille du dictionnaire de noms propres français
Célébrité	7	2 826
Ville	46	2 299
Pays	15	621
Organisation	0	60

Tableau 6. 7: L'évaluation avec le corpus 80 jours

3 Perspectives

3.1 Les flexions des noms propres arabes dans Prolexbase

Pour la flexion, Prolexbase contient des formes mais les flexions en arabe sont des modifications vocaliques non marquées dans les textes non voyellisés. Donc, nous avons rempli la table des instances avec trois entrées identiques. A cet égard, il convient de rappeler que le *corpus* de la Wikipédia ne comporte pas de voyelles. Par conséquent, dans les travaux futurs que nous envisageons d'effectuer, nous travaillerons sur l'ajout des flexions voyellisées. En prévision, en nous inspirant des travaux de Habash (2010)[148], nous avons déjà établi des règles de flexion, qui consistent à introduire les flexions des lemmes arabes qui se fléchissent en genre, nombre, cas et état comme le montre le tableau 6.8. Les flexions portent uniquement sur les voyelles. Nous présentons ci-dessous les premières ébauches de ce travail.

[148] Nisar Y. Habash, Introduction to Arabic Natural Language Processing, New York, 2010, pp. 52-58.

Genre	Nombre	Cas	État
Masculin	Singulier	Nominatif	Défini/Construit
Féminin		Accusatif	Indéfini
		Génitif	

Tableau 6. 8: Les flexions des noms propres arabes

3.1.1 Quelques règles de flexion des noms propres arabes

Dans le chapitre 2, paragraphes 3.2.1, 3.2.2, 3.2.3, nous avons expliqué les classes principales des noms propres arabes : chaque classe rassemble les noms disposant de traits morphologiques identiques ; de plus, nous avons détaillé avec des exemples les différents paradigmes de flexion qui peuvent exister dans ce domaine. Par la suite, nous avons cité quelques règles de flexion concernant les noms propres arabes et nous avons illustré ces règles par des graphes de flexion Unitex :

Règle 1 : Si le nom propre est un triptote, il aura six flexions :

genre : masculin/féminin, nombre : singulier, cas : nominatif avec l'état : défini/construit, et l'état indéfini, décliné respectivement par (Damma ـُ /u/) et la voyelle de tanwin (ـٌ(un) ;

genre : masculin/féminin, nombre : singulier, cas : accusatif avec l'état : défini/construit et de même avec l'état indéfini, décliné respectivement par (Fatha ـَ /a/) et la voyelle de tanwin (ـً (an) ;

genre : masculin/féminin, nombre : singulier, cas : génitif avec l'état : défini/construit et de même avec l'état indéfini, décliné respectivement par (Kasra ـِ /i/) et la voyelle de tanwin (ـٍ (in) ;

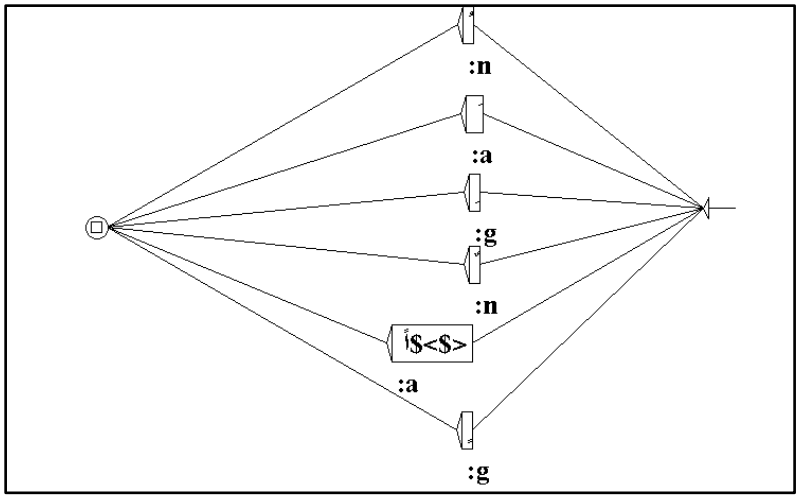


Figure 6. 2: Graphe de flexion montrant les six flexions possibles pour un nom propre triptote

Règle 2 : Si le nom propre arabe est un diptote, il aura six flexions :

genre : masculin/féminin, nombre : singulier, cas : nominatif avec l'état : défini/construit/indéfini décliné par (Damma ـُ /u/) ;

genre : masculin/féminin, nombre : singulier, cas : accusatif avec l'état : défini/construit/indéfini décliné par (Fatha ـَا /a/) ;

genre : masculin/féminin, nombre : singulier, cas : génitif avec l'état : défini/construit et de même avec l'état indéfini décliné par (Kasra ـِ /i/) et (Fatha ـَا /a/) ;

Nous avons défini certaines catégories de noms propres du type diptote qui se fléchissent suivant la règle précédente :

- les noms propres féminins comme : Mecca (مكة), zynabe (زينب) ;
- les noms propres masculins se terminant par le marqueur du féminin, par exemple Mou'awia (معاوية), Hamza (حمزة) ;
- les noms propres masculins se terminant par alif- noon (ان) selon le schème fa 'lan comme : Marwan (مروان), Othman (عثمان) ;
- les noms propres qui sont sur le schème (أفعِل af'alu a123) comme : Ahmed ((أحمد, Anwar (أنور) ;

- les noms propres “étrangers” qui ne sont pas d’origine arabe, par exemple : Paris (باريس), William (وليم) ;
- les noms propres masculins ayant pour modèle (يفعلُ yafalu ya123u)**,** par exemple : Yazid (يزيد).

Ci-dessous, la figure 6.3 montre deux exemples de deux types de noms propres qui sont classés diptotes et leurs flexions morphologiques associées. Il s’agit des noms propres étrangers et féminins Paris et La Mecque.

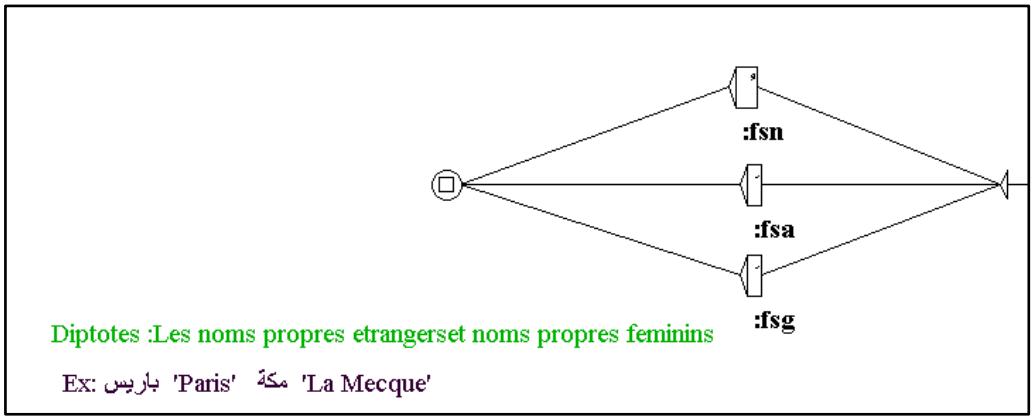


Figure 6. 3: Graphe de flexion montrant les trois traits possibles pour un nom propre diptote étranger ou féminin

Règle 3 : Si le nom propre est invariable, il aura la flexion suivante :

genre : masculin, nombre : singulier, cas : nominatif/accusatif/génitif avec l’état : défini/construit/ indéfini décliné par (Fatha ـَ /a/).

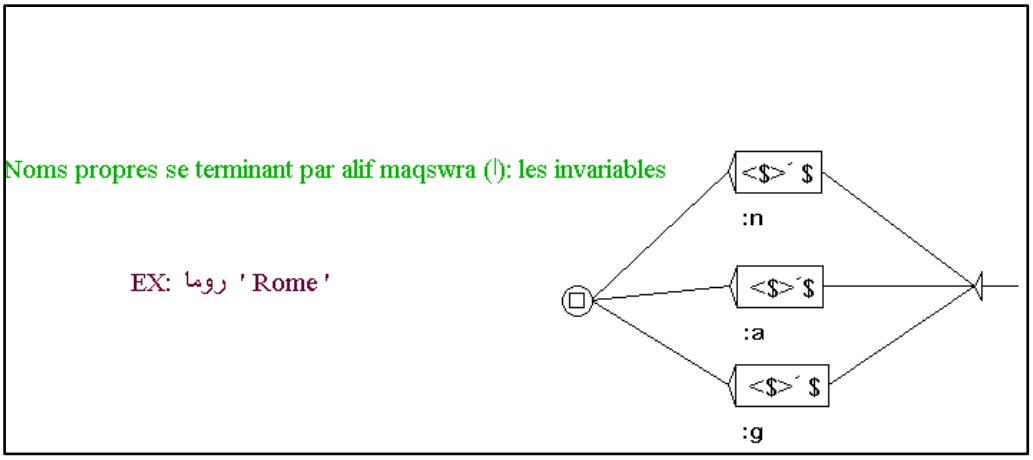


Figure 6. 4: Graphe de flexion montrant les trois cas possibles pour un nom propre se terminant par alif maqswra (ا)

3.1.2 Les dérivations

Les mots dérivés de noms propres du type lieu en arabe sont construits à partir d'une racine (stem) à laquelle on ajoute des affixes. Comme dans le cas de la langue française, ce sont des adjectifs et des noms rationnels. Par exemple, pour les noms féminins singuliers, on ajoute le suffixe " ية YTU", pour les noms féminins pluriels, on ajoute le suffixe (اتAt) aux cas accusatif ou génitif.

Pour les noms masculins singuliers, on ajoute le suffixe (ي +ya) et pour les noms masculins pluriels, on ajoute le suffixe littéral (ون+uw+na) au cas nominatif et le suffixe littéral (ين +ay +ni) au cas accusatif ou génitif. Aussi, ils se fléchissent au duel en ajoutant des suffixes littéraux (voir chapitre 2, paragraphe 3.2.4). La figure 6.5 montre les différentes dérivées des noms propres (ville ou pays) qui se terminent par un alif (A) comme le nom propre France. Pour générer leurs dérivées, on transforme l'alif en YA et ensuite on ajoute les suffixes appropriés. De même, le tableau 6.9 cite les différents noms dérivés du nom propre Paris qui sont obtenus par l'ajout de YA à la fin du nom propre Paris pour obtenir son dérivé au genre masculin et nombre singulier (parisien). Ensuite, à partir de ce nom dérivé, on peut produire le reste des dérivées par l'ajout des suffixes appropriés.

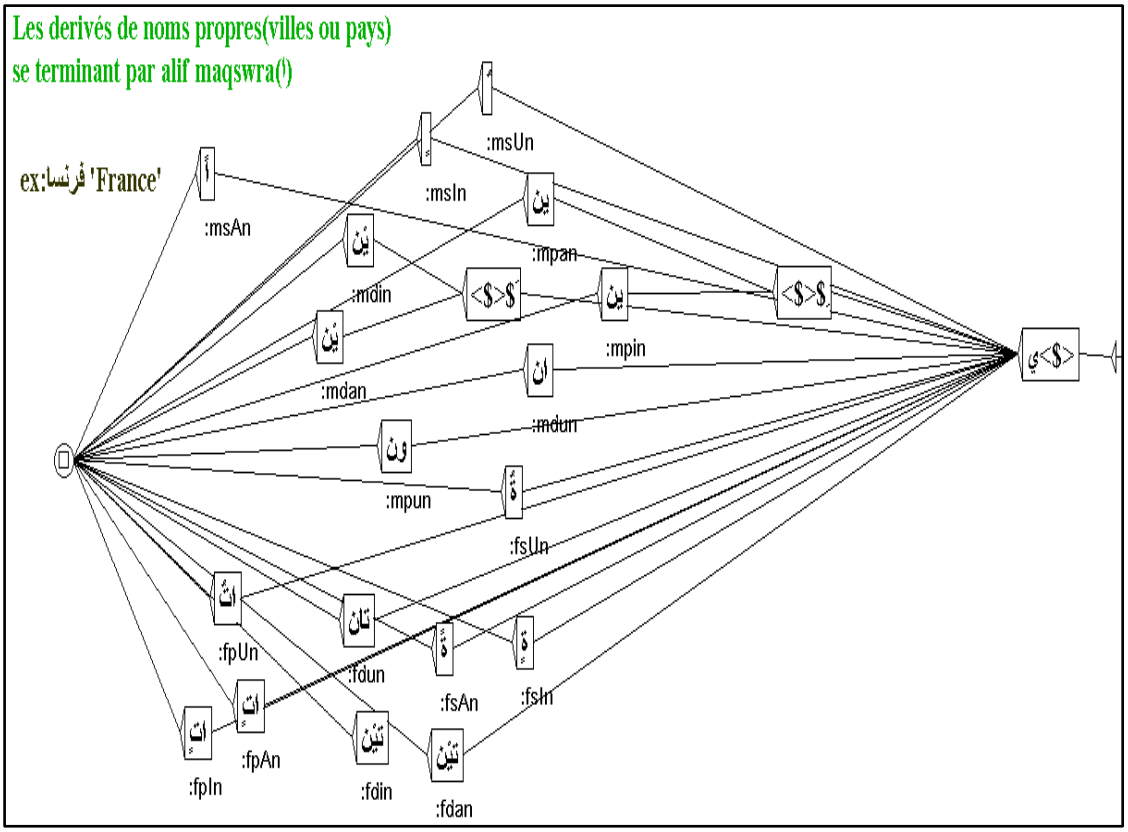


Figure 6. 5: Graphe de flexion montrant les dérivées des noms propres arabes (villes ou pays)

Dérivé		Flexion
باريسي	Parisien	MS
باريسية	Parisienne	FS
باريسيان	Deux Parisiens	MDN
باريسيين	Deux Parisiens	MDAG
باريسيتان	Deux Parisiennes	FDN
باريسيتين	Deux Parisiennes	FDAG
باريسيون	Parisiens	MPN
باريسيين	Parisiens	MPAG
باريسياتُ باريسياتِ	Parisiennes Parisiennes	FPN FPAG

Tableau 6. 9: Les noms dérivés du nom propre Paris (باريس)

Conclusion

Nous avons développé une approche entièrement automatique pour étendre la couverture de Prolexbase à la langue arabe. Nous avons commencé par utiliser les liens Wikipédia dans les trois langues principales de Prolexbase (l'anglais, le français et le polonais) pour produire leurs équivalents en arabe. Ensuite, nous avons extrait les noms propres arabes des liens qui ont été générés auparavant. Enfin, nous avons pu associer chaque entrée arabe (prolexème) à son index de notoriété via l'encyclopédie Wikipédia. Nous avons terminé notre méthode par le remplissage de la table morphologie de la langue arabe en décrivant quelques règles de flexion, une étude des noms propres arabes composés est aussi envisagée afin de définir les règles de flexion pour cette catégorie. Finalement, nous envisageons d'effectuer un traitement approfondi de l'analyse morphologique dans des travaux ultérieurs.

Conclusion Générale

En utilisant principalement l'encyclopédie universelle Wikipédia, nous avons pu réaliser un certain nombre de traitements multilingues automatiques. Ils nous ont permis, dans un premier temps, la mise à jour de l'ensemble des langues contenues dans Prolexbase. Un traitement de redirection éventuelle a réussi à détecter 904 redirections en prolexeme_eng, 731 redirections parmi les entrées de la table prolexeme_fra, 1 521 redirections en prolexeme_pol et 21 redirections dans la table de prolexèmes serbes. Toutes ces redirections ont été corrigées par notre système de mise à jour.

Un alignement entièrement automatique a également été effectué au début de ce travail de thèse, grâce auquel nous avons pu compléter et ajouter un grand nombre de liens manquants et de nouvelles entrées dans les dix langues de la base de données. A cet égard, il convient de préciser que pour certaines langues il n'existait aucun lien Wikipédia, à savoir : le portugais, l'espagnol, l'allemand, l'italien, le coréen et le néerlandais. Pour ces six langues, nous sommes passés respectivement de 0 à 434, 600, 662, 608, 217 et 563 nouveaux liens, et à 11 061, 10 049, 13 038, 11 350, 8 526 et 12 746 nouvelles entrées.

De plus, le nombre de liens Wikipédia a augmenté, incluant de nouvelles entrées, dans les prolexèmes français, anglais, polonais et serbe avec respectivement 3 056, 577, 496 et 7 946. Aussi, dans le but d'élargir Prolexbase à d'autres langues, nous avons réalisé, via notre système de consolidation, l'ajout de l'arabe égyptien et du wolof avec respectivement 1 632 et 601 numéros de pivots et liens Wikipédia.

Dans un deuxième temps, nous sommes parvenus à associer automatiquement un critère de notoriété aux prolexèmes de Prolexbase en nous basant sur la correspondance interlangue fournie par l'encyclopédie Wikipédia. Nous avons élaboré un outil de calcul fondé sur l'exploitation de l'encyclopédie Wikipédia et l'agrégation d'un ensemble de cinq mesures numériques relatives aux articles correspondant aux noms propres de Prolexbase. Nous avons choisi le nombre de consultations de l'article, le nombre de contributeurs, la taille de l'article, le nombre de liens internes pointant vers l'article et le nombre de liens externes contenus dans

l'article ; à cet effet, nous avons souhaité évaluer notre choix de ces cinq critères au début de notre recherche. Pour cela, nous avons testé manuellement sur une trentaine d'articles le lien entre les statistiques de ces critères et la qualité de l'article en question, et nous avons constaté une forte corrélation entre la popularité du sujet de l'article et les cinq valeurs sélectionnées pour tous les articles que nous avons testés.

Comme attendu, la valeur de ce critère diffère suivant les langues. De même, il nous semblait peu pertinent de comparer la notoriété d'une célébrité à celle d'un lieu, donc, nos calculs ont été effectués langue par langue et type par type. Nous notons surtout qu'une révision régulière de cette notion est assurée à travers notre système qui permet en outre de rendre l'utilisation de ce dictionnaire plus efficace et plus performante pour les applications de TAL.

Comme nous l'avons évoqué, la notoriété varie dans le temps et nous avons choisi d'attribuer un coefficient d'oubli pour le calcul du nombre de consultations d'un article Wikipédia cible. Dans le but de valoriser notre choix, nous avons produit un fichier CSV contenant les noms propres de type célébrité du prolexème français dans Prolexbase. À chacun d'entre eux, nous avons associé les nombres de consultations annuelles de la période 2008-2015 et nous avons calculé un écart moyen afin de différencier les noms propres ayant un grand pic à une année donnée. Finalement, nous avons justifié cette idée via des graphiques en reliant un tel pic à une année donnée avec un événement survenu dans cette même année.

Les résultats obtenus dans cette phase sont fortement satisfaisants. Nous avons effectué une application des résultats sur trois échantillons comportant 3 825, 4 385 et 4 400 noms propres de type célébrité dans les prolexèmes français, anglais et polonais ; cette application a été basée sur une comparaison du poids d'importance de nos cinq critères de notoriété qui ont été calculés via l'entropie de Shannon. Nous avons constaté que le nombre de consultations a eu le poids le plus important dans les prolexèmes français et anglais avec 0,296 et 0,289 ; il était de 0,219 dans le prolexème polonais ; cependant, le nombre de liens externes a présenté un poids faible dans les trois langues avec 0,055, 0,066 et 0,097 respectivement. Ensuite, par notre calcul, nous avons obtenu pour l'échantillon français un taux de 10,2% de noms propres de type célébrité de forte notoriété (1) ; 23,48 % de notoriété moyenne (2) et 66,30 % de notoriété faible (3). Pour l'anglais, il y a eu 11,97% de célébrités de notoriété 1, 26,68 % de notoriété 2 et 61,35%

de notoriété 3. En ce concerne le polonais, nous avons constaté un taux de 8,75% de noms propres de notoriété 1, 25,23% de notoriété 2 et 66,02 % de notoriété 3.

Ultérieurement, nous avons valorisé nos résultats obtenus ci-dessus, par une comparaison avec le projet mondial Panthéon. Il s'agit de compter les noms propres de type célébrité ayant une notoriété 1 ou 2 dans deux langues (le français et l'anglais) ou dans l'une de celles-ci, et qui figurent dans les cent premiers du classement «People » dans ce projet. Nous avons obtenu un bon résultat indiquant la validation de notre calcul de notoriété : 85 % de noms propres sont dans la liste des cent premiers de Panthéon, seulement 3% sont de notoriété 3 dans les deux langues considérées, 4% sont de notoriété 2 dans une langue et de notoriété 3 dans l'autre et 8 % ne figurent pas dans Prolexbase.

A la fin de cette phase, nous avons mené une comparaison de nos valeurs de notoriété avec les anciennes valeurs (fréquences) précédemment fournies dans Prolexbase depuis les travaux de Savaray *et al*. (2013). Nous avons fondé la comparaison sur la mise en correspondance entre les anciennes et les nouvelles fréquences parmi un nombre de 3 825, 4 385 et 4 400 prolexèmes français, anglais et polonais respectivement. Nous avons obtenu 59,03%, 67,84% et 46,28% de célébrités qui ont conservé leur ancien indice de notoriété ; toutefois, un nombre moyen a eu une transition de notoriété de 1 point (1 à 2, 2 à 1, 2 à 3 ou de 3 à 2) dans les trois langues avec 35,21%, 28,38% et 42,05%, et un petit nombre, soit 5,61%, 3,79% et 11,67%, a connu un changement de 2 points (de 1 à 3 ou de 3 à 1).

Pour répondre à l'objectif initial de notre travail de thèse, nous avons réalisé une nouvelle extension pour Prolexbase via l'ajout de l'arabe. Au départ, nous avons employé notre système de consolidation de liens, établi précédemment pour générer les liens Wikipédia arabes ; lors de cette dernière phase, nous avons utilisé ces liens pour extraire les noms propres correspondants, et éventuellement leur associer leurs instances (ou formes fléchies). Les résultats obtenus sont satisfaisants avec un volume de 8 409 entrées arabes (numéros de pivots, lemmes, fréquences /indices de notoriété et liens Wikipédia) qui ont été ajoutées dans le prolexème arabe de Prolexbase. Nous avons pu réaliser une évaluation manuelle des noms propres arabes extraits depuis la procédure implémentée à cette phase ; l'évaluation a été faite sur trois types : célébrité, ville et organisation et les taux de précision sont importants : 99,92% pour un ensemble de 2 830 noms de célébrités, 99,78% pour un ensemble de 2 303 villes et une

précision de 100% pour les 60 noms propres de type organisation. Il convient de noter ici que seulement deux erreurs ont été trouvées dans la liste de célébrités. Celles-ci sont dues au fait que dans certains cas les noms propres valides se trouvent dans le premier paragraphe de leurs articles Wikipédia et que notre méthode d'extraction de noms propres ne traite pas le premier paragraphe de l'article cible. Afin de remédier à ce type d'erreurs, nous envisageons de compléter prochainement la méthode utilisée par l'analyse du premier paragraphe de l'article cible.

Également, nous avons réalisé une évaluation extrinsèque de notre base de données lexicale arabe avec le corpus annoté ANERCorp et les Gazetteers ANERGazet ; les résultats de notre recherche nous encouragent à entreprendre de futurs travaux. Ceux-ci auront pour objectif d'augmenter le nombre d'entrées arabes via un processus d'enrichissement de Prolexbase à partir de l'édition arabe de l'encyclopédie Wikipédia. En effet, la base de données actuelle ne contient pas les noms de nombreuses célébrités arabes, telles qu'Ahmed Chawqi (أحمد شوقي), Taha Hussein (طه حسين), Naguib Mahfouz (نجيب محفوظ), Mahmoud Darwich (محمود درويش), Mustafa Mahmoud (مصطفي محمود) et Ahmed Zewail (أحمد زويل), etc.

De même, un traitement linguistique spécifique de la morphologie des noms propres arabes nous semble nécessaire afin d'assurer un module arabe qualifié et efficace par la suite dans les différentes applications de TAL sur les noms propres arabes.

Une autre perspective intéressante est celle de compléter l'ajout de l'arabe égyptien en tant que dialecte lié à la langue arabe, qui a déjà été préparé par la création de la table de prolexeme_arz alimentée par les numéros de pivots et liens Wikipédia. Par ailleurs, il est possible d'envisager une collaboration avec un chercheur sénégalais afin de compléter le module wolof dans Prolexbase.

Bibliographie

Abuleil, S. (2004, April). Extracting names from Arabic text for question-answering systems. In *Coupling approaches, coupling media and coupling languages for information retrieval* (pp. 638-647).

Abuleil, S., & Evens, M. (2002). Extracting an Arabic lexicon from Arabic newspaper text. *Computers and the Humanities*, *36*(2), 191-221.

Afshari A., Mojahed M., Mohd Y. R. (2010). Simple Additive Weighting approach to Personnel Selection problem. *International Journal of Innovation, Management and Technology*. 1:5, 511-515.

Attia M., Toral A, Tounsi L, et al. *(*2010). an automatically built named entity lexicon for Arabic[149].

Attia, M., & Somers, H. (2008). *Handling Arabic morphological and syntactic ambiguity within the LFG framework with a view to machine translation*. Manchester: University of Manchester.pp 279.[150].

Attia, M. (2006, October). An ambiguity-controlled morphological analyzer for modern standard arabic modelling finite state networks. In *Challenges of Arabic for NLP/MT Conference, The British Computer Society, London, UK* (Vol. 200610, No. 1.72).

Auer S., Bizer C., Kobilarov G, et al. (2007). DBpedia : *A Nucleus for a Web of Open Data*. In: *Aberer K. et al.* (eds) *The Semantic Web*, Berlin (2007). Heidelberg, Springer, 722-735.

Batra, M., & Sharma, S. (2013). Comparative Study of Page rank Algorithm with Different Ranking Algorithms Adopted by Search Engine for Website Ranking. *International Journal of Computer Technology and Applications*, *4*(1), 8. [151].

Beesley K R. (2001). Finite-state morphological analysis and generation of Arabic at Xerox Research: Status and plans in 2001. In *ACL Workshop on Arabic Language Processing: Status and Perspective.* Vol. 1, 1-8.

[149] http://doras.dcu.ie/15979/1/An_automatically_built_Named_Entity_lexicon_for_Arabic.pdf
[150] http://attiaspace.com/Publications/Attia-PhD-Thesis.pdf
[151] www.ijcta.com

Beesley, K. R. (1996, August). Arabic finite-state morphological analysis and generation. In *Proceedings of the 16th conference on Computational linguistics-Volume 1* (pp. 89-94). Association for Computational Linguistics.[152] .

Beesley, K. R. (2001, July). Finite-state morphological analysis and generation of Arabic at Xerox Research: Status and plans in 2001. In *ACL Workshop on Arabic Language Processing: Status and Perspective* (Vol. 1, pp. 1-8).[153].

Belguith L., Baccour L., Mourad G. (2005). Segmentation de textes arabes basée sur l'analyse contextuelle des signes de ponctuations et de certaines particules. In *Actes de la 12éme Conférence annuelle sur le Traitement Automatique des Langues Naturelles*, 451-456.

Belguith, L., Chaaben N. (2006). Analyse et désambiguïsation morphologiques de textes arabes non voyellés. *Actes de la 13ème confrence sur le Traitement Automatique des Langues Naturelles*, 493-501. [154]

Belguith, L. H., Aloulou, C., & Hamadou, A. B. (2007). MASPAR: De la segmentation à l'analyse syntaxique de textes arabes. *CÉPADUÈS-Editions, editeur, Revue Information Interaction Intelligence I*, *3*, 9-36.[155].

Ben Mesmia F., Haddar K., Friburger N., Maurel D. (2017). CasANER: Arabic Named Entity Recognition Tool. Studies in Big Data Series. Springer. To appear.

Benajiba, Y., & Rosso, P. (2007, December). ANERsys 2.0: Conquering the NER Task for the Arabic Language by Combining the Maximum Entropy with POS-tag Information. In *IICAI* (pp. 1814-1823). [156].

Bohas, G. (Ed.). (2014). *Développements récents en linguistique arabe et sémitique*. Presses de l'Ifpo.[157].

Bouamor, H. (2009). Extraction des connaissances à partir du Web pour la recherche des images géoréférencées. In *CORIA* (pp. 519-526).

[152] http://dl.acm.org/citation.cfm?doid=992628.992647
[153] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.28.5046&rep=rep1&type=pdf
[154] https://tel.archives-ouvertes.fr/tel-00756111v1/document
[155] https://pdfs.semanticscholar.org/4a2a/74b0078432b5bd09d334095e322287da27c2.pdf
[156] https://pdfs.semanticscholar.org/ee46/47bf7b0f82a7c0b985e42dc60f34c8340612.pdf
[157] http://books.openedition.org/ifpo/4916

Bouchou B., Maurel D. (2008). Prolexbase et LMF : vers un standard pour les ressources lexicales sur les noms propres, *Traitement automatique des langues*, 49(1) :61-88[158].

Brown, J. L., North, S., & Bussey, H. (1993). SKN7, a yeast multicopy suppressor of a mutation affecting cell wall beta-glucan assembly, encodes a product with domains homologous to prokaryotic two-component regulators and to heat shock transcription factors. *Journal of bacteriology*, *175*(21), 6908-6915. [159].

Buckwatter T. (2004). Arabic Morphological Analyzer Version 2.0 LDC2004L02. Web Download. Philadelphia : Linguistic Data Consortium.

Chevalier, F., Huot, S., & Fekete, J. D. (2010). Visualisation de mesures agrégées pour l'estimation de la qualité des articles Wikipedia. In *EGC 2010: Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances* (pp. Cépaduès-Éditions).

Cohen, W. W. (1995, July). Fast effective rule induction. In *Proceedings of the twelfth international conference on machine learning* (pp. 115-123).

Doumi, N., Lehireche, A., Maurel, D., & Ali Cherif, M. (2013, May). La conception d'un jeu de ressources libres pour le TAL arabe sous Unitex. In *TRADETAL2013, Colloque international en Traductologie et TAL, Oran-Algeria* (pp. 5-6).

Douzidia, F., & Lapalme, G. (2005). Un système de résumé de textes en arabe, 2ème Congrès International sur l'Ingénierie de l'Arabe et l'Ingénierie de la langue.

Eom Y-H, Shepelyansky DL. (2013) .Highlighting Entanglement of Cultures via Ranking of Multilingual Wikipedia Articles. PLoS ONE 8(10): e74554.[160]

Flati, T., Vannella, D., Pasini, T., & Navigli, R. (2014). Two Is Bigger (and Better) Than One: the Wikipedia Bitaxonomy Project. In *ACL (1)* (pp. 945-955).

Graff, D., Maamouri, M., Bouziri, B., Krouna, S., Kulick, S., & Buckwalter, T. (2009). Standard Arabic morphological analyzer (SAMA) version 3.1. *Linguistic Data Consortium LDC2009E73*.

Gueffaz, M., Deslis, J., & Moissinac, J. C. Peuplement automatisé d'ontologies par analyse des programmes scolaires. *Sixième Atelier Recherche d'Information SEmantique RISE, Nancy 18 mars 2014*, 42.

[158] http://www.atala.org/-Varia,55-
[159] http://jb.asm.org/content/175/21/6908.short
[160] https://doi.org/10.1371/journal.pone.0074554

Habash, N. (2004, April). Large scale lexeme based arabic morphological generation. In *Proceedings of Traitement Automatique du Langage Naturel (TALN-04).*

Habash, N. Y. (2010). Introduction to Arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, *3*(1), 1-187..[161].

Habash, N., & Rambow, O. (2006, July). MAGEAD: a morphological analyzer and generator for the Arabic dialects. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* (pp. 681-688). Association for Computational Linguistics.

Habash, N., Rambow, O., & Kiraz, G. (2005, June). Morphological analysis and generation for Arabic dialects. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages* (pp. 17-24). Association for Computational Linguistics.

Hamdi, A. (2012). Apport de la diacritisation dans l'analyse morphosyntaxique de l'arabe. *JEP-TALN-RECITAL 2012*, 247.[162].

Hellmann, S., Stadler, C., Lehmann, J., & Auer, S. (2009). DBpedia live extraction. *On the Move to Meaningful Internet Systems: OTM 2009*, 1209-1223.[163].

Hoffart, J., Suchanek, F. M., Berberich, K., & Weikum, G. (2013). YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence*, *194*, 28-61. [164].

Hovy, E., Navigli, R., & Ponzetto, S. P. (2013). Collaboratively built semi-structured content and Artificial Intelligence: The story so far. *Artificial Intelligence*, *194*, 2-27.

Iazzi, S., Yousfi, A., & Bellafkih, M. (2011). Analyseur morphologique des mots arabe en utilise le dérivé et schème de surface.[165].

ISO 12620:1999. *Computer applications in terminology - Data categories.*

ISO 24613:2008. *Language resource management - Lexical markup framework* (*LMF*).

Karami A., Johansson R. (2014). Utilization of Multi Attribute Decision Making Techniques to Integrate Automatic and Manual Ranking of Options. Journal of Information Science and Engineering. 30:519-534. [166].

[161] http://www.morganclaypool.com/doi/abs/10.2200/s00277ed1v01y201008hlt010
[162] http://www.aclweb.org/anthology/F12-3019
[163]https://www.researchgate.net/publication/220830431_DBpedia_Live_Extraction
[164] https://doi.org/10.1016/j.artint.2012.06.001
[165] event.ircam.ma/docs/TICAM2012/37.doc.
[166]http://www.iis.sinica.edu.tw/page/jise/2014/201403_14.pdf

Maurel, D., Spędzia-Baron, M., Bouchou-Markhoff, B., & Vitas, D. (2014). Prolexbase. A Multilingual Relational Database of Proper Names. *Cahiers de Linguistique*, *40*(2), 49-71.[167].

Medelyan, O., Milne, D., Legg, C., & Witten, I. H. (2009). Mining meaning from Wikipedia. *International Journal of Human-Computer Studies*, *67*(9), 716-754.

Mendes, P. N., Jakob, M., & Bizer, C. (2012, May). DBpedia: A Multilingual Cross-domain Knowledge Base. In *LREC* (pp. 1813-1817). [168].

Mesfar, S. (2008). *Analyse morpho-syntaxique automatique et reconnaissance des entités nommées en arabe standard* (Doctoral dissertation, Besançon).

Mikheev, A., Moens, M., & Grover, C. (1999, June). Named entity recognition without gazetteers. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics* (pp. 1-8). Association for Computational Linguistics.

Navigli, R., & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, *193*, 217-250. [169]

Okinina N, Nouvel D, Friburger N, Antoin J. (2013). Apprentissage supervisé sur ressources encyclopédiques pour l'enrichissement d'un lexique de noms propres destiné à la reconnaissance des entités nommées. *ATALA. TALN'2013, 20e conférence sur le Traitement Automatique des Langues Naturelles, Jun* 2011, Les Sable d'Olonne, France. 667-674. [170]

Rey, A. (1977). *Le lexique: images et modèles du dictionnaire à la lexicologie* (p. 16). Paris: Armand Colin.

Saadane, H. (2013). A linguistic approach for knowledge extraction from an Arabic text (Une approche linguistique pour l'extraction des connaissances dans un texte arabe)[in French]. *Proceedings of RECITAL 2013*, 124-137.[171].

Saaty, T. L., & Vargas, L. G. (2006). *Decision making with the analytic network process*. ». Springer Science+ Business Media, LLC, p 282.

[167] https://hal.archives-ouvertes.fr/hal-01119318/
[168] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.680.5157&rep=rep1&type=pdf
[169] http://wwwusers.di.uniroma1.it/~ponzetto/pubs/navigli12b.pdf
[170] <hal-01016545>
[171] http://www.aclweb.org/anthology/F13-5010

Saaty, T. L., & Vargas, L. G. (2013). Sensitivity analysis in the analytic hierarchy process. In *Decision making with the analytic network process* (pp. 345-360). Springer US.

Sadat, F., & Terrasa, A. (2010). Exploitation de wikipédia pour l'enrichissement et la construction des ressources linguistiques. In *Proceedings of TALN*. [172].

Safari H., Sadat Fagheyi M., Sadat Ahangari S., Reza Fathi M. (2012). Applying PROMETHEE Method based on Entropy Weight for Supplier Selection. *Business Management & Strategy*. Vol. 3:1, 97-106. [173]

Sajjad, H., Darwish, K., & Belinkov, Y. (2013, August). Translating Dialectal Arabic to English. In *ACL (2)* (pp. 1-6).

Savary Agata, Manicki Leszek, Baron Małgorzata. (2013). Populating a Multilingual Ontology of Proper Names from Open Sources. Journal of Language Modelling. 1-2:189-225.

Sellami R; Sadat F; Belguith L. (2012). Extraction de lexiques bilingues à partir de Wikipédia (Bilingual lexicon extraction from Wikipedia). In : *JEP-TALN-RECITAL 2012, Workshop TALAf 2012 : Traitement Automatique des Langues Africaines (TALAf 2012 : African Language Processing)*. S. 107-117.

Shannon C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, vol. 27. 379-423 et 623-656. [174]

Singh, A. K., & Kumar, P. R. (2009). A comparative study of page ranking algorithms for information retrieval. *International journal of electrical and computer engineering*, *4*(7), 469-480. [175].

Smrž, O. (2007, June). Elixirfm: implementation of functional arabic morphology. In *Proceedings of the 2007 workshop on computational approaches to Semitic languages: common issues and resources* (pp. 1-8). Association for Computational Linguistics. [176].

[172] https://www.semanticscholar.org/paper/Exploitation-de-Wikipe-dia-pour-l-Enrichissement-e-Sadat-Terrasa/c611a0690db3a913a641741346b1b2941aba0922

[173] https://doi.org/10.5296/bms.v3i1.1656

[174] http://www.science.oregonstate.edu/~hetheriw/whiki/ph415_s15/tasks/dsp/files/nyquist/shannon1948.pdf

[175] http://waset.org/publications/3153/a-comparative-study-of-page-ranking-algorithms-for-information-retrieval

[176] http://dl.acm.org/citation.cfm?id=1654576.1654578

Tchechmedjiev A. (2016). Interopérabilité sémantique multilingue des ressources lexicales en données lexicales liées ouvertes. Intelligence artificielle [cs.AI]. Université Grenoble Alpes. Français. <tel-01425123>

Tchechmedjiev, A. (2016). *Interopérabilité sémantique multilingue des ressources lexicales en données lexicales liées ouvertes* (Doctoral dissertation, Université Grenoble Alpes).

Tran M., Maurel D. (2006), Prolexbase : Un dictionnaire relationnel multilingue de noms propres, *Traitement automatique des langues*, Vol. 47(3) :115-139[177].

Viseur, R. (2013). Extraction of Biographical Data from Wikipedia. In *DATA* (pp. 248-252).[178] .

Viseur, R. (2014, August). Reliability of user-generated data: The case of biographical data in Wikipedia. In *Proceedings of The International Symposium on Open Collaboration* (p. 31). ACM.

Wilkinson, D. M., & Huberman, B. A. (2007, October). Cooperation and quality in wikipedia. In *Proceedings of the 2007 international symposium on Wikis* (pp. 157-164). ACM. [179].

Ying Z. (2016). Modèles et outils pour des bases lexicales "métier" multilingues et contributives de grande taille, utilisables tant en traduction automatique et automatisée que pour des services dictionnairiques variés. Informatique et langage [cs.CL]. Université Grenoble Alpes. Français.

Yu A. Z., et al. (2016). Pantheon 1.0, a manually verified dataset of globally famous biographies. *Scientific Data* 2:150075.doi: 10.1038/sdata.2015.75. [180].

[177] http://www.atala.org/-Varia,47-
[178] < DOI: 10.5220/0004595302480252>.
[179] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.88.4591&rep=rep1&type=pdf
[180] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4700860/

Sites web consultés

1. Open Xerox. [Consulté le 11 mars 2014],
   URL: http://www.xrce.xerox.com/research/mltt/arabi.
2. Haskell. [Consulté le 24 mars 2014],
   URL : http://www.haskell.org
3. Quest.ms.mff.cuni.cz. [Consulté le 24 mars 2014],
   URL: http://quest.ms.mff.cuni.cz/elixir/
4. Wikipédia, Grammaire non contextuelle. [Consulté le 12 mai 2014],
   URL: https://fr.wikipedia.org/wiki/Grammaire_non_contextuelle
5. nlp.standford. [Consulté le 14 mai 2014],
   URL: http://nlp.standford.edu/projects/arabic.shtml
6. users.dsic.upv.es. [Consulté le 17 mai 2014],
   URL: http://users.dsic.upv.es/~ybenajba/downloads.html
7. Unitex/GramLab. [Consulté le 20 janvier 2015],
   URL: http://www-igm.univ-mlv.fr/~unitex
8. Unicode. [Consulté le 20 janvier 2015],
   URL: http://www.unicode.org/charts/PDF/U0600.pdf
9. CNRTL : Centre National de Ressources Textuelles et Lexicales. [Consulté le 05 janvier 2014], URL : http://www.cnrtl.fr/lexiques/prolex/
10. The Stanford Natural Language Processing Group. [Consulté le 10 février 2014],
    URL: http://nlp.stanford.edu/software/tagger.shtml
11. ilc.cnr.it. [Consulté le 11 mars 2014],
    URL: http://www.ilc.cnr.it/ne-repository
12. Wikimédia, Index of /arwiki. [Consulté le 10 juin 2014],
    URL: http://download.wikimedia.org/arwiki
13. Wikipedia, Statistiques. [Consulté le 12 mars 2015],
    URL: https://fr.wikipedia.org/wiki/Wikipédia:Statistiques
14. Wikipedia, Wikimedia Foundation. [Consulté le 20 janvier 2015],
    URL: https://fr.wikipedia.org/wiki/Wikimedia_Foundation

15. Wikipedia, Licence Creative Commons. [Consulté le 11 janvier 2015],
URL : https://fr.wikipedia.org/wiki/Licence_Creative_Commons

16. Libération.fr. [Consulté le 15 mars 2015],

URL :http://www.liberation.fr/ecrans/2008/01/17/ceux-qui-disent-non-a-wikipedia_959345?page=article

17. Alexa. [Consulté le 15 mars 2015],
URL : http://www.alexa.com/topsites
18. Wikipedia, *Catégorie*. [Consulté le 16 mars 2015],
URL: http://fr.wikipedia.org/wiki/Wikipédia:Catégories
19. Cetic. [Consulté le 30 mars 2015],
URL: https://www.cetic.be/Exploiter-le-contenu-de-Wikipedia
20. Wikipédia, WikiProject assessment. [Consulté le 02 mai 2015],
URL: https://en.wikipedia.org/wiki/Wikipedia:WikiProject_assessment
21. Wikipédia, *Lien interwiki*. [Consulté le 03 mai 2015],
URL: https://fr.wikipedia.org/wiki/Aide:Lien_interwiki
22. Wikipédia, *Liens externes*. [Consulté le 03 mai 2015],
URL: https://fr.wikipedia.org/wiki/Wikipedia:Liens_externes
23. Wikipédia, *Insérer une référence*. [Consulté le 05 mai 2015],
URL :https://fr.wikipedia.org/wiki/Aide:Insérer une référence
24. Wikimedia, dumps. [Consulté le 12 mai 2015],
URL: http://dumps.wikimedia.org/
25. Wikipédia, *MediaWiki*. [Consulté le 12 mai 2015],
URL :https://fr.wikipedia.org/wiki/Aide:MediaWiki
26. Wikipédia, *Site miroir*. [Consulté le 14 mai 2015],
URL :https://fr.wikipedia.org/wiki/Site_miroir
27. Wikipédia, *Citation et réutilisation du contenu de Wikipédia*. [Consulté le 12 mai 2015],
URL :https://fr.wikipedia.org/wiki/Wikipédia:Citation_et_réutilisation_du_contenu_de_Wikipédia
28. Wikipédia, *Wikipédia hors-connexion*. [Consulté le 15 mai 2015],

URL:https://fr.wikipedia.org/wiki/Wikipédia:Wikipédia_hors-connexion-1re_.C3.A9tape_:_Installer_MediaWiki

29. Wikipédia, *DBpedia*. [Consulté le 20 juin 2015],
URL: https://fr.wikipedia.org/wiki/DBpedia
30. WordNet Domains. [Consulté le 26 juin 2015],
URL: http://wndomains.fbk.eu/
31. *Max-Planck-Institut für Informatik*. [Consulté le 26 juin 2015],
URL :http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/
32. Wikipédia, *ApiHelp*. [Consulté le 29 juin 2015],
URL :https://nl.wikipedia.org/w/api.php
33. Mediawiki, *API:Query*. [Consulté le 29 juin 2015],
URL: https://www.mediawiki.org/wiki/API:Query

Mouna ELASHTER

Gestion et extension automatiques du dictionnaire relationnel multilingues de noms propres Prolexbase Mise à jour multilingues et création d'un volume arabe via la Wikipedia

Résumé

Les bases de données lexicales jouent un grand rôle dans le TAL, mais, elles nécessitent un développement et un enrichissement permanents via l'exploitation des ressources libres du web sémantique, entre autres, l'encyclopédie Wikipédia, DBpedia, Geonames et Yago2. Prolexbase, comporte à ce jour dix langues, trois parmi elles sont bien couvertes : le francais, l'anglais et le polonais. Il a été conçu manuellement et une première tentative semi-automatique a été réalisée par le projet ProlexFeeder (Savary et al. 2013). L'objectif de notre travail était d'élaborer un outil de mise à jour et d'extension automatiques de ce lexique, et l'ajout de la langue arabe. Un système automatique a également été mis en place pour calculer via la Wikipédia l'indice de notoriété des entrées de Prolexbase ; cet indice dépend de la langue et participe, d'une part, à la construction d'un module de Prolexbase pour la langue arabe et, d'autre part, à la révision de la notoriété présente pour les autres langues de la base.

Mots Clés : Nom propre, Prolexbase, Bases lexicales multilingues, Notoriété, Langue arabe, Wikipédia

Abstract

Lexical databases play a significant role in natural language processing (NLP), however, they require permanent development and enrichment through the exploitation of free resources from the semantic web, among others, Wikipedia, DBpedia, Geonames and Yago2. Prolexbase, which issued of numerous studies on NLP, has ten languages, three of which are well covered: French, English and Polish. It was manually designed; the first semiautomatic attempt was made by the ProlexFeeder project (Savary et al., 2013). The objective of our work was to create an automatic updating and extension tool for Prolexbase, and to introduce the Arabic language. In addition, a fully automatic system has been implemented to calculate, via Wikipedia, the notoriety of the entries of Prolexbase. This notoriety is language dependent, is the first step in the construction of an Arabic module of Prolexbase, and it takes a part in the notoriety revision currently present for the other languages in the database.

Key words: Proper noun, Prolexbase, Multilingual lexical databases, Notoriety, Arabic language, Wikipedia