Centre d'Etudes Doctorales : Sciences et Techniques de l'Ingénieur

N° d'ordre : 84 /2020

THESE DE DOCTORAT

Présentée par

Mr : KHALFI Mustapha

Spécialité : Informatique

Sujet de la thèse : Une nouvelle ressource lexicale riche pour la langue Arabe classique : Conversion digitale du dictionnaire ***al=qāmūs al=muḥīṭ***.

Thèse présentée et soutenue le Samedi 26 décembre 2020 devant le jury composé de :

Nom Prénom	Titre	Etablissement	
Mr. Mohammed EL MOHAJIR	PES	Faculté des Sciences Dhar El Mehraz Fès	Président
Mr Azzeddine MAZROUI	PES	Faculté des Sciences - Oujda	Rapporteur
Mr. Si Lhoussain AOURAGH	PES	Faculté des Sciences Juridiques Economiques et Sociales -Salé	Rapporteur
Mme. Ilham CHAKER	PH	Faculté des Sciences et Techniques – Fès	Rapporteur
Mr. Azeddine ZAHI	PH	Faculté des Sciences et Techniques – Fès	Examinateur
Mme. Ouafae NAHLI	Dr	Institut des Linguistiques Computationnelles – Pise (Italie)	Directeurs de thèse
Mr. Arsalane ZARGHILI	PES	Faculté des Sciences et Techniques – Fès	
Mr. Rachid BEN ABBOU	PH	Faculté des Sciences et Techniques – Fès	

Laboratoire d'accueil : Systèmes Intelligents et Applications

Une nouvelle ressource lexicale riche pour la langue Arabe classique: Conversion digitale du dictionnaire *al=qāmūs al=muḥīṭ*

Remerciements

Je remercie, tout d'abord, DIEU qui garde toujours un œil bien veillant sur nous.

Je tiens à remercier mes directeurs de thèse Mr ***Arsalane Zarghili*** *et Mr* ***Rachid BEN ABBOU*** *à la FST d'avoir rendu possible la réalisation de ce travail et de m'avoir accordé leurs confiances, pour leur encadrement, leur disponibilité et leurs conseils fructueux qu'ils m'ont prodigués le long de ma thèse.*

Un immense merci à ma Co-encadrante Mme ***Ouafae NAHLI*** *de ILC-Pise qui a clairement enrichie la définition du mot encadrer dans le contexte de cette thèse. Je salue notamment l'intérêt constant qu'elle a portée à ces travaux, sa disponibilitée et ses conseils méticuleux, et qui n'a jamais cessée de me guider et de me faire bénéficier de son grand savoir.*

Je remercie Mr ***Azzeddine MAZROUI****, Mr* ***Si Lhoussain AOURAGH*** *et Mme.* ***Ilham CHAKER*** *qui m'ont fait l'honneur d'être rapporteurs de cette thèse et qui m'ont fait part de leurs commentaires précieux.*
Merci aussi à Mr. ***Mohammed EL MOHAJIR*** *et Mr* ***Azeddine ZAHI*** *qui ont accepté de faire partie du Jury de cette soutenance.*

J'ai eu la chance durant mon stage de recherche à l'ILC-pise d'être entouré et encouragé, merci donc à mes collègues et notamment à ***Angelo Mario Del Grosso*** *qui m'a prêté une oreille attentive et aidé, à sa manière, quand j'en avais besoin.*

De manière plus générale, un grand merci pour ma chère famille. J'ai une chance incroyable de vous avoir et d'être si bien entourée. Je remercie enfin mes parents et mes frères, ma sœur et ma femme pour être ce qu'ils sont et pour leur soutien sans faille.

Finalement un immense merci à mon ***Cher père (que dieu bénisse ton âme),*** *Merci pour le père que tu as été. Merci pour ta gentillesse, ton soutien, ton humour et ton originalité qui ont marqué ma vie.*

Résumé

Actuellement, les grandes ressources lexicales acquièrent une pertinence potentielle élevée pour les systèmes d'information et le besoin de ressources lexicales dans les domaines du Traitement Automatique des Langues Naturelles (TALN) est primordial.

Pour contribuer à répondre à ces besoins pour la langue Arabe qui souffre d'une grave pénurie de ce genre de ressources numériques, nous construisons une ressource lexicale à base du célèbre dictionnaire *al=qāmūs al=muḥīṭ* (AQAM). En utilisant une approche basée sur des règles, nous avons conçu un système qui permet d'extraire des informations morpho-syntaxiques, sémantiques et lexicales du célèbre dictionnaire. Nous avons donc obtenu une version numérisée et structurée d'AQAM, enrichie d'informations explicites morpho-syntaxiques et lexicales. Dans cette thèse, nous décrivons les étapes suivies pour la conversion d'AQAM en un format lisible par machine, en commençant par la phase de segmentation du texte, puis l'extraction des lemmes et des sens, suivie de la phase d'identification de la catégorie grammaticale (Part Of Speech), l'extraction d'informations morpho-syntaxiques et lexicales. Le dictionnaire est très riche de noms propres, nous avons donc dû faire face à une méthodologie automatique permettant de les identifier et les extraire. La tâche d'extraction des noms propres est effectuée sur la base d'une liste de balises définie constituée de mots-clés utilisés par l'auteur d'AQAM.

De plus, la ressource obtenue est enrichie par des traductions anglaise du lemme et des sens simples qui l'accompagnent à l'aide de deux dictionnaires bilingues Arabe-Anglais. Ensuite, nous présentons un aperçu d'un alignement d'expérience de la section de la lettre *bā'* sur le WordNet de Princeton (PWN) et Suggested Upper Merged Ontology (SUMO). Cette expérience s'est avérée intéressante car elle a révélé que le mappage d'une ressource lexicale arabe sur une ressource anglaise montre des points communs entre les deux langues, mais elle permet surtout de souligner les non-équivalences entre elles.

Toutes les ressources obtenues sont représentées au format XML et distribuées sous licence gratuite sous la plate-forme CLARIN-IT.

Mots-Clés — Extraction d'informations, lexique arabe, Al Qamus Al Muhit, dictionnaire lisible par machine, ressource lexicale arabe.

Abstract

Currently, large lexical resources are getting a high potential relevance for information systems and a need of Lexical resources in Natural Language Processing (NLP) fields is paramount. To contribute to these needs for the Arabic language which suffers from a serious shortage of this kind of digital resources, we build a lexical resource from the famous dictionary *al=qāmūs al=muḥīṭ* (AQAM). Using a rule based approach, we have designed a system that allows extracting morpho-syntactical, semantics and lexical information from the famous dictionary. So, we obtained a digitized and structured version of AQAM, enriched by morpho-syntactical and lexical explicit information. In this thesis, we describe the followed steps of AQAM conversion to a machine readable format, starting by text segmentation phase, thereafter lemma and senses extraction, followed by POS (Part Of Speech) tagging phase, morpho-syntactical and lexical information extraction. The dictionary is also rich of proper names, so we had to face the task of identifying an automatic methodology which makes it possible to identify and extract them. Proper names extraction task is carried out based on a defined tags list consisting of keywords used by AQAM's author.

In addition, the obtained resource is enriched by English translations of lemma and accompanying senses using a bilingual English-Arabic dictionary. Then we present an overview of an experiment alignment of the section of the letter *bā'* on Princeton's WordNet (PWN) and Suggested Upper Merged Ontology (SUMO). This experience turned out to be interesting because it revealed that mapping an Arabic lexical resource on an English resource shows commonality between the two languages, but it allows especially to emphasize the non-equivalences between them.

All obtained resources are represented in XML format and distributed under free license in the *CLARIN-IT* platform.

Keywords — Information Extraction, Arabic Lexicon, Al Qamus Al Muhit, Machine-readable dictionary, Arabic Lexical Resource.

مُلخـص

معلوم أن الموارد المعجمية الرقمية الضخمة ذات صلة وطيدة بالأنظمة المعلوماتية، وعليه تعد الحاجة إلى توظيف هذه الموارد في معالجة اللغات الطبيعية أمرا لا غنى عنه.

وللإسهام في تلبية احتياجات اللغة العربية، التي تعاني من خصاص مهول في هذا النوع من الموارد الرقمية، نقوم ببناء مورد معجمي يعتمد في الأساس على واحد من أشهر قواميس اللغة العربية، وهو القاموس المحيط. باعتماد نمط قائم على مجموعة من القواعد، صممنا نظاما يسمح باستخراج المعلومات الصرفية والنحوية والدلالية والمعجمية من القاموس المحيط. وبالتالي حصلنا على نسخة رقمية موَسّمة ومنظّمة منه، غنية بالمعلومات الممثلة بشكل واضح وجَليّ. وفي هذه الأطروحة، نقوم بوصف الخطوات المتَّبعة لتحْويل هذا القاموس إلى نسق رقمي قابل للاستخدام الآلي. عَنِيَتْ المرحلة الأولى بتجزئة النص، ثُمّ تليها مرحلة استخلاص الكلمة الرئيسية ومعانيها، تليها مرحلة تحديد الفئة النحوية (نوع الكلمة)، بعد ذلك ننتهي إلى استخراج المعلومات الصرفية والنحوية. يعتبر هذا القاموس حافلا بأسماء الأعلام، لذا كان علينا التعامل معه بمنهجية تُتيح التّعرّف عليها واستخراجها، على أساس قائمة علامات محددة تتكون من كلمات مفاتيح يستخدمها صاحب القاموس.

إضافة إلى ذلك، تَمّ إثراء المورد الرقمي المحصل عليه بترجمات انجليزية للكلمات الرئيسية ومرادفاتها، باعتماد قاموسين مزدوجي اللغة عربي انجليزي. بعد ذلك نقدم نظرة عامة لتجربة مواءمة "باب حرف الباء" على شبكة الكلمات (Prinston Wordnet) وعلى اقتراح الأنطولوجيا المدمجة العليا (Suggested Upper Merged Ontology). تبين من خلال هذه التجربة أن المواءمة بين مورد عربي وآخر انجليزي، يظهر نقاطا مشتركة بين هاتين اللغتين، ولكنه قبل ذلك يبين ويوضح عدم التكافؤ الحاصل بينهما.

يتم تمثيل جميع الموارد التي تم الحصول عليها بتنسيق XML ويتم توزيعها بموجب ترخيص مجاني عبر منصة CLARIN-IT

كلمات المفاتيح: استخلاص المعلومات، المعجم العربي، القاموس المحيط، القاموس المقروء آليا، الموارد المعجمية العربية.

Table des matières

Introduction Générale **10**

Chapitre 1 Étude lexicographique de al=qāmūs al=muḥīṭ et préparation des données **15**

1.1 Étude lexicographique de al=qāmūs al=muḥīṭ 15

1.1.1 Introduction 15

1.1.2 Caractéristiques de la langue Arabe 16

1.1.3 Aperçu de la lexicographie Arabe 17

1.1.3.1 École d'arrangement de phonèmes 17

1.1.3.2 École d'arrangement des racines par ordre alphabétique . 18

1.1.3.3 École des rimes (*'al=ǧawharī*) 18

1.1.3.4 École d'arrangement par ordre alphabétique ordinaire . . . 19

1.1.4 ***al=qāmūs al=muḥīṭ*** 19

1.1.4.1 Macrostructures d'AQAM 20

1.1.4.2 Microstructure d'AQAM 22

1.1.5 Conclusion 24

1.2 Préparation de données d'al=qāmūs al=muḥīṭ 25

1.2.1 Introduction 25

1.2.2 Préparation des données et phases de segmentation du texte 25

1.2.2.1 Préparation du texte 25

1.2.2.2 Segmentation du texte suivant la macrostructure et la microstructure 26

1.2.2.3 Remplacement des mot-vedettes 27

1.2.2.4 Nettoyage du texte 28

1.2.2.5 Segmentation de la section *wāw & yā'* 28

1.2.2.6 Résultats de la segmentation du texte 29

1.2.3 Conclusion 30

Chapitre 2 Conversion digitale d'al=qāmūs al=muḥīṭ 32
2.1 Phases d'extraction d'informations . . . 32
2.1.1 Extraction et normalisation du lemme . . . 34
2.1.1.1 Cas des entrées lexicales avec mot-vedette . . . 34
2.1.1.1.1 Sélection du mot-vedette . . . 34
2.1.1.1.2 Normalisation du lemme . . . 36
2.1.1.2 Cas de paradigme / mot-moule . . . 38
2.1.1.3 Cas de changement de voyelle . . . 39
2.1.1.4 Cas des adjectifs . . . 40
2.1.1.5 Catégorie grammaticale (Part Of Speech / POS) . . . 41
2.1.1.5.1 Cas des adjectifs . . . 41
2.1.1.5.2 Cas des verbes . . . 41
2.1.1.5.3 Cas des noms . . . 42
2.1.2 Extraction des informations morphosyntaxiques et dérivatives . . . 43
2.1.2.1 Cas des verbes . . . 43
2.1.2.2 Cas des noms et des adjectifs . . . 45
2.1.3 Cas des entrées imbriquées . . . 45
2.1.4 Extraction des sens et des informations lexicales . . . 46
2.1.5 Exemples de conversion XML . . . 47
2.1.6 Version digitale de *al=qāmūs al=muḥīṭ* . . . 52
2.2 Conclusion . . . 57
Chapitre 3 Marquage sémantique automatique d'AQAM en utilisant PWN et SUMO 58
3.1 Introduction . . . 58
3.2 Ressources . . . 58
3.2.1 Lexiques bilingues . . . 58
3.2.1.1 An Advanced Learner's Arabic-English Dictionary . . . 59
3.2.1.2 Dictionary of Modern Written Arabic . . . 59
3.2.2 Princeton WordNet (PWN) . . . 62
3.2.3 Suggested Upper Merged Ontology (SUMO) . . . 62
3.3 Méthodologie . . . 64
3.4 Processus de traduction . . . 66
3.5 Processus d'alignement / mapping . . . 66

3.6 Discussion et analyse de quelques exemples 68
3.6.1 Exemple du lemme verbal *'abba* 68
3.6.2 Exemple du lemme *kitāb* 71
3.6.3 Exemple du lemme *ṯawāb* 77
3.7 Analyse des résultats statistiques 80
3.8 Conclusion 83
Chapitre 4 Traitement et extraction des noms propres cités dans *al=qāmūs al=muḥīṭ* **85**
4.1 Introduction 85
4.2 Défi de la reconnaissance des noms propres 86
4.2.1 Manque de capitalisation 86
4.2.2 Manque des ressources 87
4.3 Les approches de reconnaissance des noms propres 87
4.3.1 L'approche à base de règle 87
4.3.2 L'approche statistique 89
4.3.3 L'approche Hybride 89
4.4 Extraction automatique des noms propres à partir de *al=qāmūs al=muḥīṭ* . 89
4.5 Marquage sémantique des noms propres 91
4.6 Conclusion 95
Conclusion et Perspective **97**
Annexes **99**
.1 Listes des suffixes nominatives et transitives 99
.2 Liste des lettres solaires utilisées 100
.3 Exemple de certains tags/mots clés utilisés 101
Bibliographie **103**

Table des figures

1.1 Extrait d'AQAM 21
1.2 Microstructure d'entrée lexicale AQAM 22
1.3 Famille de la racine *(' t̲ ')* 23
1.4 La Macrostructure d'AQAM 27
1.5 Exemple des entrées lexicales de la section wāw & yā' 29
1.6 Phase 1 : Segmentation de texte selon la macrostructure 29

2.1 Phase 2 : Extraction d'information en fonction de la Microstructure 34

3.1 Structure du DMWA avant les procédures de normalisation. 60
3.2 Extrait du fichier restructuré des verbes. 61
3.3 Processus de Mapping. 65
3.4 Flux de travail 69

Liste des tableaux

1.1 statistiques d'AQAM TXT 31

2.1 Stemming vs stemming léger 36
2.2 Stemming vs stemming léger 37
2.3 Marqueurs de changement de voyelle 40
2.4 Exemples des paradigmes morphologiques verbales 42
2.5 Quelques statistiques de conversion XML d'AQAM 53
2.6 Statistiques des entrées verbales extraites d'AQAM 54
2.7 Statistiques d'extraction des entrées d'AQAM 54

3.1 Statistiques des types de lemmes d'AQAM 81
3.2 Statistiques des types de définition d'AQAM 81
3.3 Statistiques de la traduction des lemmes et des sens de l'AQAM 81
3.4 Résultat contrôlé manuellement 82
3.5 Statistiques des entrées lexicales verbales d'AQAM mappées à SUMO . . . 82

4.1 statistiques des noms propres extrait à partir de AQAM 92
4.2 Statistiques des noms propres extraits 93
3 Suffixes Nominatives et Transitives 99
4 lettres solaires 100
5 Exemple de tags/mots clés pour l'extraction des noms propres 101

Introduction Générale

La croissance instantanée de la taille du contenu arabe sur le World Wide Web et la nature non structurée de la majorité de ces données représentent deux problèmes principaux concernant les humains et les machines. D'une part, d'énormes quantités de données ne peuvent pas être traitées par des humains, et d'autre part, un format non structuré ne peut pas être compris par des machines Al-Zoghby et al. (2018). Il est donc nécessaire d'introduire des ressources lexicales et sémantiques pour les machines qui leur permettent de traiter et d'analyser le texte non structuré.

En outre, les ressources lexicales sont essentielles dans la plupart des domaines du Traitement Automatique des Langues Naturelles **TALN!** (**TALN!**), tels que l'extraction d'informations, la recherche d'informations, la classification et la traduction automatique. Généralement, le Traitement Automatique de la Langue Arabe (TALA) souffre d'un manque de ressources linguistiques numériques (Shaalan (2014)). De plus, la langue arabe est hautement flexionnelle et dérivée, avec une morphologie riche et une syntaxe complexe Alqrainy et al. (2012), ce qui place le TALA devant un grand défi. Différents efforts sont faits pour faire face à ces tâches difficiles et récemment certains travaux ont été développés dans le domaine des ressources en arabe.

On commence par le plus connu, le WordNet Arabe (AWN), c'est une base de données lexicale arabe construite selon les méthodes développées pour EuroWordNet [1] Elkateb et al. (2006). *"Un wordnet -un riche répertoire de connaissances sur les mots - est un élément clé de nombreuses applications de traitement de texte ou de traitement de la langue réussies"*Piasecki et al. (2009). Wordnet est un dictionnaire informatisé de synonymes, thésaurus, base de données lexicale, taxonomie de concepts. Bien qu'ils existent depuis plus de 20 ans, Wordnets signifient toujours des choses différentes pour différentes personnes Piasecki et al. (2009). Cependant, en tant que source de sens des mots, Wordnet

1. EuroWordNet est une base de données multilingue avec WordNets pour plusieurs langues européennes (néerlandais, italien, espagnol, allemand, français, tchèque et estonien). http ://projects.illc.uva.nl/EuroWordNet/

ressemble à un thésaurus et est souvent présenté en tant que thésaurus (Fellbaum (1998a), p. 210). Le premier Wordnet, dont les origines remontent à près d'un quart de siècle, est le WordNet anglais connu sous le nom Princeton WordNet Miller (1995). Étant le mieux développé, il est devenu au cœur de nombreux travaux en traitement automatique du langage naturel. Princeton WordNet (PWN) a commencé comme une expérience psychologique qui visait à expliquer comment la signification lexicale est stockée dans l'esprit et à éclaircir l'acquisition de la signification lexicale par les enfants, comme l'a défini Miller dans (Miller (1995)) comme étant un système de référence lexical en ligne dont la conception s'inspire des théories psycholinguistiques actuelles de la mémoire lexicale humaine.

Contrairement à un dictionnaire standard, WordNet ne prend pas le mot, ou le lexème, comme son élément de base. Au lieu de cela, WordNet ressemble à un thésaurus en ce que ses unités sont des concepts, lexicalisés sous forme de groupe de mots. Un groupe de mots qui peuvent tous faire référence au même concept est appelé ensemble de synonymes, ou synset. Les synsets sont liés à d'autres mots ou synsets à travers des relations conceptuelles-sémantiques et lexicales. (Fellbaum (1998a), p. 210).

Pour les applications TALN, WordNet est une ressource électronique qui rapproche la signification des unités lexicales, bien que cela se limite souvent à de simples utilisations de l'hypernymie (une relation super-classe/sous-classe) (Piasecki et al. (2009)). Les synsets (groupes de mots étroitement liés sémantiquement) fournissent souvent des alternatives ou des ajouts interchangeables à des ensembles de mots-clés ou de termes de recherche.

PWN est couramment utilisé comme référence pour d'autres wordnets et pour les travaux liés aux wordnets. Cependant Arabic WordNet (AWN) est basé sur PWN qui est construit pour la langue anglaise, c'est-à-dire que AWN est construit sur des listes de traductions arabes suggérées pour différents mots contenus dans des synsets anglais correspondant à un ensemble de concepts de base Rodríguez et al. (2008a). Par conséquence, il ne couvre pas certains concepts spécifiques de la langue arabe. Plusieurs tentatives pour enrichir et étendre le contenu AWN ont été faites AlKhalifa and Rodríguez (2009) ; Rodríguez et al. (2008b) ; Del Gratta and Nahli (2014) ; Regragui et al. (2016). Alkhatib et al. propose dans Alkhatib et al. (2017) une approche pour développer un WordNet basé sur *al=ḥadīṯ*[2] en établissant des connexions sémantiques entre les mots afin de fournir une meilleure compréhension des significations des mots *al=ḥadīṯ*. Cependant, l'écart entre

2. Il fait référence à un enregistrement des mots, les actions et l'approbation silencieuse du prophète Mohammed

l'AWN et la langue arabe ainsi que d'autres WordNet similaires est resté l'une des limites de son utilisation, car il ne prend pas certaines particularités de la langue arabe Regragui et al. (2016).

Boudelaa et Marslen-Wilson ont présenté dans Sami and D (2010) une base de données lexicale pour l'arabe standard moderne nommée *Aralex*[3] et basée sur un corpus de textes contemporains de 40 millions de mots. Il fournit des informations sur les fréquences des jetons, des bigrammes et des trigrammes, c'est-à-dire la séquence de deux ou trois éléments adjacents à partir d'une chaîne de jetons respectivement. En outre, il fournit des informations sur les formes orthographiques, les racines et les motifs de mots.

Les auteurs de Attia et al. (2011) proposent une base de données lexicale basée sur un corpus représentative de l'arabe standard moderne. Ils ont utilisé un corpus arabe moderne standard de 1 089 111 204 mots, un outil de pré-annotation, des techniques d'apprentissage automatique et une correspondance de modèles basée sur les connaissances pour acquérir et filtrer automatiquement les connaissances lexicales sur les attributs morpho-syntaxiques et les paradigmes d'inflexion.

Namly et Bouzoubaa ont présenté dans Namly and Bouzoubaa (2015) la conversion LMF du dictionnaire de langue arabe contemporaine *mu'ğam al=luġah al=mu'āṣirah*[4] qui est un dictionnaire monolingue arabe réalisé par *Ahmed Mukhtar Abdul Hamid Omar* en 2008.

Zerrouki et Balla proposent dans Zerrouki and Balla (2017) un corpus de textes vocalisés arabes appelé Tashkeela qui contient 75 millions de mots entièrement vocalisés extraits de principalement 97 livres d'arabe classique et moderne. Tashkeela peut être utilisé comme outil de ressources linguistiques pour le TALN, comme les systèmes diacritiques automatiques, le mécanisme de désambiguïsation, les fonctionnalités et l'extraction de données.

Cependant, on note que la plupart des ressources lexicales numériques disponibles concernent l'arabe standard moderne (MSA : Modern Standard Arabic), ainsi il y a une pénurie de ressources dédiées à l'arabe classique (CA : Classical Arabic). Donc, pour minimiser le manque de ressources linguistiques lisibles par machine pour l'arabe classique, notre travail vise à créer une nouvelle ressource lexicale basée sur le dictionnaire arabe

3. https ://aralex.mrc-cbu.cam.ac.uk/aralex.online/login.jsp
4. http ://arabic.emi.ac.ma/alelm/ ?q=Resources

classique médiéval القَامُوس المُحيط *al=qāmūs al=muḥīṭ* (AQAM).

La ressource obtenue pourrait être utilisée dans d'autres projets de TALA, tels que la reconnaissance d'entités nommées, la recherche d'informations, l'analyse morphologique, la reconnaissance optique de caractères, et plus précisément dans le traitement du patrimoine manuscrit arabe.

Après cette section d'introduction, cette thèse se compose de quatre chapitres qui illustrent l'approche qui permet de créer une nouvelle ressource lexicale riche à partir du célèbre dictionnaire *al=qāmūs al=muḥīṭ* (AQAM). La version numérique et enrichie est publiée en open source aidera les études d'arabe classique à gagner sur plusieurs fronts (lexicographie, sémantique, philologie, etc.).

Cette thèse est complémentaire avec la thèse de Mme Nahli Ouafae (Nahli (2018b)), dont elle traite le sujet d'un point de vue lexicale et philologique, tandis que notre thèse le traite d'un point de vue informatique.

Pour donner un aperçu des étapes de conversion et d'enrichissement de la ressource que nous avons accomplies dans cette thèse, nous avons divisé cette dernière en quatre chapitres :

Le premier chapitre se compose de deux sections, la première illustre une étude lexicographique d'AQAM où on présente un aperçu des différentes écoles lexicographiques de la langue Arabe. Ensuite une étude de la macrostructure et la microstructure qui caractérisent AQAM, permet d'identifier les différentes parties dont il se compose. La deuxième section présente une phase de préparation particulière et détaillée du texte brut AQAM. Suivant la macrostructure, la microstructure de AQAM et les particularités stylistiques de son auteur, on a ajouté différents marqueurs permettant d'identifier et diviser les parties qui constituent le dictionnaire AQAM.

Le deuxième chapitre traite la conversion numérique d'AQAM. En suivant un mécanisme complètement automatique, nous illustrons les phases d'extraction et de normalisation des lemmes, d'extraction d'informations morphosyntaxiques, de sens et des informations lexicales, en plus de l'identification automatique des mots dérivés qui accompagnent une entrée principale. Nous sommes en mesure d'identifier de nombreuses expressions régulières et règles morphologiques et stylistiques qui permettent l'extraction d'informations lexicales, morphologiques et syntaxiques et permettent que la conversion numérique AQAM soit effectuée automatiquement. Ensuite, nous discutons quelques exemples d'entrées lexicales et de certains résultats statistiques obtenus.

Le troisième chapitre illustre la phase d'enrichissement de notre ressource avec le mar-

quage sémantique. D'une part, on a enrichi notre ressource par la traduction anglaise des lemmes en utilisant deux dictionnaires bilingues (Salmoné (1978), Wehr (1976)). D'autre part elle a été enrichie en liant, en tant que possible, les lemmes aux concepts correspondants en utilisant PWN et SUMO.

Les dictionnaires modernes ne contiennent pas les noms propres, par contre les dictionnaires médiévales en sont riches. C'est l'une des caractéristiques spécifiques distinguant AQAM, d'où vient la nécessité de les extraire et les classer. Le dernier chapitre traite le cas particulier des noms propres, leurs extractions, classification et marquage sémantique.

Une conclusion finale termine le travail conçu jusque là et trace les lignes pour le futur afin de rendre la ressource obtenue uniforme et systématique pour être utilisable par la communauté scientifique.

Chapitre 1

Étude lexicographique de al=qāmūs al=muḥīṭ et préparation des données

1.1. Étude lexicographique de al=qāmūs al=muḥīṭ

1.1.1. Introduction

Traditionnellement, la lexicographie est définie comme «la branche de la linguistique appliquée concernée par la conception et la construction de lexiques à usage pratique» (Gibbon (2000)). Comme Ide et Véronis l'expliquent dans Ide and Véronis (1995), les dictionnaires en général présentent une forte dualité entre leur structure de surface (le texte) et leur structure profonde (le contenu d'information), et une grande partie des informations de structure profonde n'est pas explicite dans la structure de surface. Pour que la reconnaissance des entrées lexicales et des composants se fasse automatiquement et correctement, il est nécessaire d'étudier la macrostructure et la microstructure et de connaître les conventions abrégées et la disposition du dictionnaire.

Dans cette étude précise, on a également dû faire face au fait que la structure du dictionnaire arabe médiéval est très variable et différente de celle moderne.

Cette section se consacre à l'étude du dictionnaire médiéval de l'arabe classique al=qāmūs al=muḥīṭ (AQAM), sa macrostructure et microstructure, les raisons de son choix, ainsi que les caractéristiques faisant de lui un très bon candidat pour la conversion en format lisible par machine (*Machine-Readable Format*).

1.1.2. Caractéristiques de la langue Arabe

La langue Arabe est une langue flexionnelle, avec une morphologie riche, ainsi qu'une syntaxe complexe, ce qui complique les différentes phases du traitement automatique de cette langue. La tâche reste difficile en raison des caractéristiques suivantes de la langue Arabe :

1. Langue utilisée : L'arabe est classée comme la 6ème langue parmi les principales langues dans le monde. De plus, c'est la langue officielle de 22 pays qui s'étendent du nord de l'Afrique jusqu'au Moyen-Orient. Dans la littérature, on parle de trois types de langue arabe :

   — L'Arabe Classique (CA, Classical Arabic) est la langue utilisée dans le Coran (livre sacré de l'Islam). Cette forme est utilisée comme langue de l'Islam depuis plus de 1500 ans. En outre, l'arabe classique a été largement utilisée jusqu'au 9ème siècle, et ce fut pendant les temps Abbasside et la plupart des textes historiques et religieux arabes sont manuscrits en Arabe Classique.

   — L'Arabe Standard Moderne (MSA, Modern Standard Arabic) est la langue officielle du monde arabe, utilisée aujourd'hui comme langue officielle pour les communications entre les gouvernements et les secteurs publics ou privés et les individus. En principe, MSA peut être traitée comme la version évoluée de l'arabe classique. MSA est la forme la plus connue de nos jours, et c'est la forme largement utilisée dans l'éducation, les journaux et les magazines. les Nations Unies considèrent le MSA comme l'une des principales langues officielles.

   — Les Dialectes Arabes Familiers (Colloquial Arabic Dialects) généralement utilisés pour une communication quotidienne informelle. Ils ne sont pas enseignés dans les écoles ou même standardisés bien qu'il existe une riche culture dialectale populaire des contes populaires, des chansons, des films et des émissions de télévision. Les dialectes sont principalement utilisés dans la communications orale et dans les communications à travers les médias sociaux. Cependant il n'y a pas d'uniformité dans l'écriture du script pour les dialectes. (Shaalan and Habash (2010)).

2. Script : L'une des caractéristiques de l'écriture arabe imposant des défis au reconnaissance des noms propres, est que les mots arabes sont écrits avec des scripts

connectés, ce qui n'est pas le cas de nombreuses autres langues telles que l'anglais.

3. Diacritisation : L'Arabe a des signes diacritiques qui représentent les voyelles (voyelles courtes). Les diacritiques jouent deux rôles principaux en fonction de leurs positions dans le mot. La voyelle finale dans un mot -qui apparaît généralement sur la dernière lettre- indique son rôle syntaxique ; tandis que les signes diacritiques sur le reste des lettres indiquent le choix lexical (Darwish et al. (2017)). Le changement de ces signes diacritiques d'un mot arabe implique le changement de signification du mot. De conséquence, le texte arabe non vocalisé crée une ambiguïté, mais il reste compréhensible par les arabophones.

1.1.3. Aperçu de la lexicographie Arabe

La lexicographie arabe est l'un des arts originaux et fondamentaux de la littérature arabe. Au cours des 12 derniers siècles, de nombreux types de lexiques de langue arabe ont été construits, ces lexiques sont différents dans l'ordre, la taille et le but de la construction. De nombreux linguistes et lexicographes de langue arabe ont étudié la construction, le développement et les différentes méthodologies utilisées pour construire ces lexiques.

La lexicographie arabe a connu une variété d'écoles, dont les plus importantes sont particulièrement distinctes selon l'ordre de présentation des entrées lexicales (ordre alphabétique ou phonétique - ordre par racines ou par lexèmes). Généralement, la structure de la langue arabe ne permet pas de classer les entrées lexicales par ordre alphabétique et les mots sont ramenés à la racine dont ils sont dérivés. Les lexicographes ont emprunté plusieurs voies pour organiser les entrées lexicales de leurs dictionnaires, de sorte qu'ils sont devenus des moyens connus pour ceux qui voulaient collecter et organiser les mots de la langue. Concernant la langue Arabe, il existe quatre méthodes de classifications les plus répandues (Abou-sikkīn (1981)).

1.1.3.1 École d'arrangement de phonèmes

C'est la première école connue dans l'histoire du lexique arabe. الخليل *al=ẖalīl* (718-786) est l'imam de cette école et des lexicologues arabes en général. Il a créé dans son

dictionnaire العين *al=‘ayn*, un moyen d'arranger les racines. Dans ce système *al=ẖalīl* a arrangé les lettres suivant leurs phonétiques. Il a commencé par la lettre عين *‘ayn* qui sort du milieu de la gorge puis هَاء *hā’*, et ainsi de suite jusqu'à ce qu'elle se termine par les lettres des lèvres [1].

1.1.3.2 École d'arrangement des racines par ordre alphabétique

C'est la deuxième école, dont son pionnier أبو بكر بن دريد *abū bakr ibnu darīd* a suivi dans son dictionnaire معجم الجمهرة *mu‘ğam al=ğamharat* le même système de classement adopté par *al=ẖalīl*, mais en suivant le classement alphabétique des lettres Arabe (أ ب ت ث…) qui est plus simple par rapport à l'ordre phonétique. Par exemple, les racines ضرب, برض، بضر، ربض، رضب sont cités dans le chapitre de la lettre بَاء *bā’*, parce que c'est la première lettre suivant l'ordre.

1.1.3.3 École des rimes (*’al=ğawharī*)

Le système initié par الجوهري *’al=ğawharī* se base sur l'arrangement des entrées lexicales selon le dernier radical du mot au lieu du premier. Dans ce système, les éléments lexicaux sont classés par ordre alphabétique sur la base des radicaux de leurs racines, en commençant par le dernier suivi du premier, puis les consonnes intermédiaires, c'est-à-dire que le dernier radical est considéré en premier, suivi du premier radical et d'autres radicaux sont disposés à l'avant (d'où 3ème + 1er + 2ème en trilitères et 4ème + 1er + 2ème + 3ème en quadrilitères). [2]

En fait ce système se base sur le dernier radical du mot, ce qui facilite la recherche par rime. Par exemple le mot دَرَسَ *darasa* est classé sous la section السّين *as=siīn* et le chapitre الدَّال *ad=dāl*.

Parmi les lexicographes les plus célèbres qui ont suivi ce système dans leurs dictionnaires, on trouve :

---

1. ع /ح /ه /خ /ق /ك /ج / ش / ض /ص /س / ز / ط / د / ت /ظ /ذ /ث /ر /ل /ن /ف /ب /م /و /ا /ي /أ

2. voir Ba‘labakkī (2014), pages 363 et 394.

— إبن منظور *ʻibnu manẓūr* dans "لسَان العرب *lisān al=ʻarab*"

— الفيروز آبَادي *al=fīrūz ʻābādī* dans son lexique "القَاموس المحيط *al=qāmūs al=muḥīṭ*"

— الزّبيدي *az=zabīdī* dans "تَاج العروس *tāğ al=ʻarūs*"

— أحمد فَارس الشّديَاق *ʻahmad fāris aŝ=ŝidyāq* dans "الجَاسوس علَى القَاموس *al=ğāsūs ʻalā al=qāmūs*".

1.1.3.4 École d'arrangement par ordre alphabétique ordinaire

Cette école est venue comme conclusion des écoles lexicographes, à travers laquelle la facilitation dans le lexique arabe a atteint le moyen le plus simple. Le mot est lu à partir de la première lettre puis la deuxième et ainsi de suite, et cette école a organisé les mots en tenant compte de l'ordre alphabétique des lettres constituant le mot, sans regarder la structure morphologique à laquelle se réfère. La classification selon cette méthode a commencé à une époque où certains lexicographes suivent la méthode des rimes. Parmi les plus anciens lexicographes qui ont utilisé cette méthode est *az=zamaḫŝarī* dans son dictionnaire *"ʻasās al=balāğah"*. Dernièrement la plupart des lexicographes ont continué à utiliser cette méthode, grâce à sa simplicité, et elle est devenue l'école moderne utilisée par plusieurs dictionnaires compilés dernièrement.

1.1.4. *al=qāmūs al=muḥīṭ*

AQAM a été compilé par le lexicographe perse الفيروز آبَادي *al=fīrūz ʼābādī* (1329-1414) qui a réduit le contenu original des dictionnaires sources, en éliminant les exemples, les citations coraniques, la poésie, etc. En fait, le style succinct utilisé par l'auteur facilite l'extraction d'informations. Ainsi, AQAM est bien structuré et contient de courts éléments lexicaux, ce qui est en fait un excellent candidat pour la conversion en un lexique computationnel (Nahli et al. (2016)).

De plus, le choix d'AQAM est dû à d'autres raisons, principalement en raison de son autorité lexicale et de son exhaustivité en termes de nombre d'entrées. En fait, AQAM

est l'un des dictionnaires arabes les plus utilisés à ce jour, en particulier pour l'étude des manuscrits, de la poésie et de la littérature arabe classique.

1.1.4.1 Macrostructures d'AQAM

AQAM appartient à l'école lexicographique ordonnant les entrées lexicales selon la racine dont elles dérivent et suivant le système de rimes. Comme ça était cité précédemment, les éléments lexicaux sont classés par ordre alphabétique sur la base des radicaux de leurs racines, en commençant par le dernier suivi du premier, puis les consonnes intermédiaires, c'est-à-dire que le dernier radical est considéré en premier, suivi du premier radical et d'autres radicaux sont disposés à l'avant (d'où 3ème + 1er + 2ème en trilitères et 4ème + 1er + 2ème + 3ème en quadrilitères). [3]

Le dictionnaire est divisé en sections (أَبوَاب). Chaque section (بَاب) est consacrée à une consonne alphabétique correspondant à la dernière consonne radicale.

Chaque section est divisée en chapitres (فُصُول) classés selon la première consonne radicale. Dans la même section, chaque chapitre (فَصل) contient les racines ayant les mêmes radicaux en première et troisième position.

Chaque chapitre est également divisé en différentes parties regroupant la famille de racine (root family), c'est-à-dire toutes les entrées lexicales qui ont la même racine. Dans le chapitre, les racines sont classées par ordre alphabétique selon la deuxième (et la troisième) consonne radicale. Enfin, les entrées lexicales sont regroupées sous la racine dont elles dérivent.

En résumé, AQAM se compose donc de 29 sections. Chaque section est divisée en chapitres ordonnés selon la première consonne du lemme. Ces chapitres sont également divisés en différentes parties rassemblant les entrées lexicales dérivant de la même racine.

La Figure 1.1 illustre un extrait de la section همزة *hamzah* et des deux premiers chapitres *hamzah et bā'*, d'une version imprimée [4] montrant les différentes parties de la macrostructure d'AQAM.

3. voir Ba'labakkī (2014), pages 363 et 394.
4. Al-fīrūz'ābādī (1998)

باب الهمزة

فصل الهمزة

• الأباءة، كعباءة[1]: القصبة، ج: أباء، هذا موضع ذكره كما حكاه ابن جني عن سيبويه، لا المعتل كما توهمه الجوهري وغيره. وأبأته بسهم: رميته به.

• أتأة، كحمزة: امرأة من بكر بن وائل: أم قيس بن ضرار، وجبل.

• الأثبة، كالأثبة: الجماعة. وأتأته بسهم: رميته به، هنا ذكره أبو عبيد، والصغاني في: ث و أ، ووهم الجوهري فذكره في ثأثأ. وأصبح مؤتئا، أي: لا يشتهي الطعام.

• أجأ: جبل لطيئ وبرزته[2]، وة بحضر، ويؤنث فيهما. وكجعل: هرب. وكسحابة: ع لبدر بن عقال، فيه بيوت ومنازل.

• أرأ الغنم، كمنع: أشبعها، و. عن الحاجة: جبن ونكص.

• الأشاء، كسحاب: صغار النخل. قال ابن القطاع[3]: همزته أصلية عن[4] سيبويه، فهذا موضعه، لا كما توهم[5] الجوهري.

• أكأ، كمنع، استوثق من غريمه بالشهود.

أبو زيد: أكأ إكاءة، كإجابة، وإكاء: إذا أراد أمراً، ففاجأته على نية[6] ذلك، فهابك، ورجع عنه.

فصل الباء

• بأباه، و. به: قال له: بأبي أنت، و. الصبي: قال: بابا. والبؤبؤ، كالهدهد: الأصل، والسيد الطريف، ورأس المكحلة، وبدن الجرادة، وإنسان العين، ووسط الشيء، وكسرسور وذحناح: العالم. وتبأبأ: عدا.

• بتأ بالمكان، كمنع: أقام.

• كبأ.

• بدأ به، كمنع: ابتدأ، و. الشيء: فعله ابتداء، كبأبدأه وابتدأه، و. من أرضه: خرج، و. الله الخلق: خلقهم، كبأبدأ فيهما. ولك البدء والبداءة والبداءة ويضمان، والبدية، أي: لك أن تبدأ، والبديئة: البديهة، كالبداءة. والفعلة بدءاً، وأول بدء، وبادي بدء، وبادي بدئ[7]، وبادي بدأة، وبدأة ذي بدء، وبدأة ذي بداءة (وبداءة ذي بداءة، وبدأة

(١) أي موازن له في حركاته وسكناته، وقد ضبط المؤلف في هذا الكتاب غالباً الألفاظ التي تشتبه عند العامة، وإن لم تشتبه عند الخاصة، بذكر مثال مشهور عنده، أو بالنص على حركات حروفه التي يحصل بها التمرين، حذراً من تحريف السماع وتصحيفهم. وإنما قل الانتفاع باللغة لعسر الترتيب، أو قلة الضبط بالموازين، والنص على الحركات اعتماداً على ضبطها بالشكل، وظهورها عند الخواص، وقد أجاد الجوهري الترتيب وأهمل الضبط الذي يطرق إليه التحريف والتبديل مما قريب، وغلبه ما مر. (مناوي).

(٢) أي وهو على وزنه، وكأنه أشار به إلى ضبطه، وهو اصطلاح له، ويدل لذلك ما سيأتي له في: ق ب ل. (ش) فانظره.

(٣) هو علي بن جعفر بن علي السعدي. (ش).

(٤) في الشرح «عند».

(٥) في الشرح «توهمه».

(٦) أي حينه ووقته. (ش).

(٧) بسكون الياء وإن كانت في محل نصب هكذا يتكلمون به وربما تركوا همزه لكثرة الاستعمال اهـ مناوي عن الصحاح لكن الشارح مرتضى ضبط بادي بفتح الياء وبادي بد، وبادي بداء الأول كشج والثاني كسماء والياء ساكنة في بادي كما في الشارح اهـ (مصححه).

FIGURE 1.1 – Extrait d'AQAM

La section commence par le titre de la lettre concernée, suivie du titre du chapitre. Chaque famille de racine est citée dans un paragraphe distinct regroupant les différentes entrées lexicales dérivant de la même racine.

1.1.4.2 Microstructure d'AQAM

En analysant les entrées lexicales dont se compose AQAM, on remarque que généralement, l'entrée contient le mot à définir et des informations morphologiques et lexicales.

Par exemple, pour la microstructure de l'entrée illustrée dans la Figure 1.2, on remarque que la séquence avant les deux points concerne le mot-à-définir أَبَّ *'abba*, puis on trouve certaines informations dérivatives et syntaxiques concernant le lemme, (les formes imparfaites et les noms verbaux). On note que le mot à définir est une forme qui se trouve toujours au début de l'entrée lexicale. Pour ceci, on l'appelle ***mot-vedette (headword)***. Par contre, la séquence suivante représente un sens qui est dans ce cas sous forme d'un synonyme, et peut être aussi sous forme d'une phrase explicative ou bien d'un ensemble de sens séparés par des virgules.

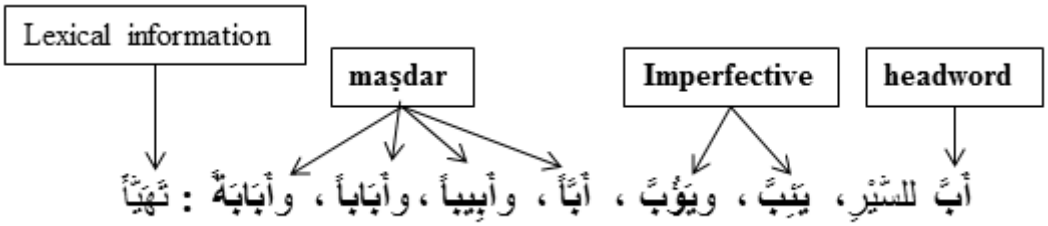


FIGURE 1.2 – Microstructure d'entrée lexicale AQAM

En résumé, la segmentation des différentes parties de l'entrée lexicale se fait à base des signes de ponctuations. Les **virgules** et les **deux points** servent à délimiter les informations lexicales d'une part, et les informations morphologiques d'autre part. Souvent, les informations lexicales sont placées après le mot-vedette, suivies d'autres informations pertinentes.

La Figure 1.3 illustre un extrait des entrées lexicales appartenant à la famille de racine ء ث ء *(' t̲ ')*.

Cette famille se compose de trois entrées lexicales : i) la première entrée nominale commence par la forme orthographique الأُثْئِيَّةُ *al='ut̲'iyyat-u*, suivie du mot كَالأُثْفِيَّة "*comme al='ut̲fiyyat-i*" qui sert comme schème/pattern pour contrôler la vocalisation du

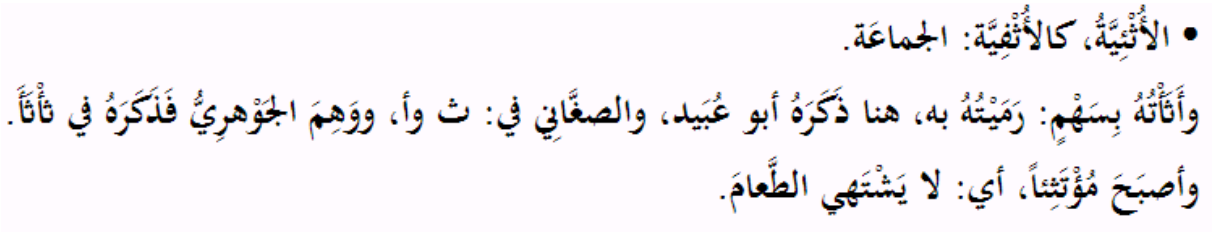

• الأُثْنِيَّةُ، كالأُثْفِيَّة: الجماعَة.
وأَثَأْتُهُ بِسَهْمٍ: رَمَيْتُهُ به، هنا ذَكَرَهُ أبو عُبَيد، والصغَّانِي في: ث وأ، ووَهِمَ الجَوْهرِيُّ فَذَكَرَهُ في ثأْثَأَ.
وأصبَحَ مُؤْتَثِئاً، أي: لا يَشْتَهي الطَّعامَ.

FIGURE 1.3 – Famille de la racine *(' ṯ ')*

mot à définir. L'entrée se termine avec la définition الجَمَاعَة *al=ǧamā'at* qui signifie la communauté, le groupe ou la cohorte et se trouve après deux points.

ii) La deuxième entrée verbale commence par وَأَثَأْتُهُ بِسَهْمٍ *wa='aṯa'-tu=hu bi=sahm-in*[5] "et *'aṯa'-tu=hu* avec une flèche", après deux points, on trouve le sens رَمَيْتُهُ بِهِ *ramay-tu=hu bihi* qui peut être traduit par "Je l'ai jeté avec elle".

iii) La dernière entrée adjectivale concerne أَصْبَحَ مُؤْتَثِئاً *'aṣbaḥa mu'taṯi'-an* "*il est devenu mu'taṯi'-an*", après on trouve la conjonction qui annonce une explication, أَيْ *'ay* "c'est-à-dire". Enfin l'entrée lexicale se termine par لَا يَشْتَهِي الطَّعَامَ *lā yaštahī aṭ=ṭa'ām-a* qui veut dire "il n'a pas envie de manger".

Dans les dictionnaires classiques le *mot-vedette* peut être une forme fléchie du lemme, c-à-d n'est pas forcement une forme canonique (masculin singulier, infinitif...). On remarque dans la première entrée nominale de la Figure 1.3 illustrant le mot-vedette *al='uṯ'iyyat-u* que ce dernier est mentionné avec l'article défini *al=* et le suffixe nominatif *-u* qui doivent être supprimés afin d'avoir une forme nominale canonique suivant les conventions lexicales modernes. La deuxième entrée lexicale traitant le mot-vedette *'aṯa'-tu=hu* qui est une forme fléchie composée par le verbe *'aṯa'-a*, le suffixe nominatif *=tu* et le pronom accusatif *=hu*. Ce mot-vedette nécessite un traitement supplémentaire qui consiste à la suppression des suffixes et préfixes. La normalisation de l'adjective *mu'taṯi'-an* figurant dans la dernière entrée consiste à la suppression du suffixe *-an*. Le traitement de chaque cas sera détaillé dans la deuxième section du 3ème chapitre (3.2.1).

À partir du ***mot-vedette***, on peut extraire le lemme suivant le standard lexicographique moderne, cependant étant une forme fléchie et complexe, il a besoin de traitements

5. Pour la syntaxe et la sémantique des gloses interlinéaires, On utilise "The Leipzig Glossing Rules : Conventions for Interlinear Morpheme-by-morpheme Glosses " (Max Planck Institute for Evolutionary Anthropology, 2015. Disponible sur le lien : http ://www.eva.mpg.de/lingua/resources/glossing-rules.php). Les gloses inter-linéaires morphème par morphème fournissent des informations sur les significations et les propriétés grammaticales de mots individuels et de parties de mots. Ainsi, les morphèmes segmentables sont séparés par des tirets et les frontières clitiques sont marquées par un signe égal. En fait, dans l'exemple *wa='aṯa'-tu=hu*, la conjonction, le verbe et le pronom accusatif sont séparés par un signe égal.

spécifiques.

En résumé, Pour révéler un mot spécifique dans AQAM, on suit les étapes suivantes :

1. Rendre le mot à son origine, en supprimant les pronoms associés au mot (suffixes et préfixes),
2. Extraire la racine de laquelle il dérive ;
3. Rechercher la dernière lettre de la racine dans la section correspondante. Par exemple pour دَعَبَ *da'aba* il sera donc dans la section *bā'*, دَمِيَ *damiya* dans la section *wāw et yā'*, et خَيْر *ẖayr* dans la section *rā'*, et ainsi de suite.
4. Rechercher dans le chapitre correspondant à la première lettre radicale. Par exemple, كَتَبَ *kataba* est cité dans le chapitre *kāf* de la section *bā'*, دَعَبَ *da'aba* dans le chapitre *dāl* de la section *bā'*, et زَهْرَة *zahrat* appartient au chapitre *zāy* de la section *rā'*, et ainsi de suite.
5. Parcourir les regroupements des racines selon l'ordre alphabétique de la deuxième consonne radicale.

1.1.5. Conclusion

On a présenté un aperçu de la lexicographie Arabe et ses écoles. Ensuite, on a discuté le dictionnaire médiéval AQAM, ses caractéristiques et les raisons de son choix, qui se résument dans son autorité lexicale, l'exhaustivité en terme de nombre d'entrées, la fusion des contenus originaux des dictionnaires sources [6], ainsi que sa bonne structuration au niveau de la marcostructure et la microstructure, faisant de lui un excellent candidat pour la conversion en un lexique computationnel de l'Arabe Classique.

L'étude lexicographique permet d'identifier les caractéristiques et les composants du lexique dans le but de le convertir en format digitale. Le processus est complexe et commence d'abord par une préparation rigoureuse et soignée du contenu textuel pour reconnaître les différentes parties qui le composent. La section suivante illustre les phases

6. Al-fīrūz'ābādī (1998), page 5

exécutées pour rendre la macrostructure et la microstructure lisibles par machine et permettant la segmentation du contenu.

1.2. Préparation de données d'al=qāmūs al=muḥīṭ

1.2.1. Introduction

Dans le but de convertir AQAM en un format lisible par machine, on a utilisé une copie sous format Unicode (.txt) disponible en accès libre sur : "https ://ar.wikisource.org", cependant, on utilise comme référence une copie éditée (Al-fīrūz'ābādī (1998)).

Afin de permettre la segmentation du texte AQAM, le processus commence par la phase de préparation en identifiant et en marquant les parties essentielles qui caractérisent la macrostructure et la microstructure d'AQAM. Dans ce chapitre on traite et on décrit les signes de surface utilisés pour la préparation du texte.

1.2.2. Préparation des données et phases de segmentation du texte

1.2.2.1 Préparation du texte

Le fichier source téléchargé contenant le texte brut d'AQAM, se compose de 28 sections, commençant par la section *hamza* (بَاب الهَمزَة), et finissant par la section 'alif (بَاب الأَلِف اللَّيِّنَة), par contre la section *wāw & yā'* (بَاب الوَاو وَاليَاء) rassemble les deux lettres *wāw* et *yā'* dans une seule section.

Bien que *al=fīrūz 'ābādī* a vocalisé tout le dictionnaire, le texte que nous avons n'est pas entièrement vocalisé, ce qui nécessite une étape de normalisation, dans laquelle on a remplacé tous les noms des lettres (titre des sections et chapitres...) avec les mêmes variantes, car parfois on trouve dans la version Unicode utilisée quelques noms de chapitres écrits avec différentes manières. Par exemple les variantes du nom du chapitre *ǧīm* utilisées dans les différentes sections :

الجيْم ، الجيم

On remarque que dans le premier cas, la *yā'* est vocalisée, tandis que pour la deuxième, tout le mot n'est pas vocalisé. Pour remédier à ce genre d'ambiguïté, nous avons remplacé toutes les occurrences de chaque chapitre par la variante utilisée dans le titre de la section. Donc pour l'exemple donné précédemment de la lettre *ǧīm*, la variante utilisé est « الجيْم ».

1.2.2.2 Segmentation du texte suivant la macrostructure et la microstructure

Après la phase de normalisation du texte, et afin de délimiter et identifier les différents segments d'AQAM, on a préparé le texte brut en ajoutant quelques indicateurs et signes, comme le montre le fragment d'AQAM illustré dans la Figure 1.4, et comme décrit ci-dessous :

— Reconnaissance des sections et chapitres : les symboles ***1*** et ***2*** indiquent respectivement la section et le chapitre. Par conséquent, ils indiquent respectivement les valeurs de la troisième et de la première consonne radicale. Par exemple, dans la Figure 1.4, le symbole *1* indique la section de la lettre *bā'*, par conséquent la dernière consonne radical est *bā'*. Par contre, le symbole *2* indiquant le chapitre *dāl* donne également la première valeur radicale qui est *dāl*.

— Reconnaissance de la famille de racine : le symbole **@** indique le début d'une nouvelle famille de racine et par conséquent il indique le changement du deuxième radical. Cependant, il ne permet pas son identification, qui se fait à base d'ordre alphabétique. Par conséquent, la famille de racine commence par le symbole **@**. Ainsi, entre deux symboles **@** successifs, on trouve des éléments lexicaux dérivés de la même racine. Dans la Figure 1.4, il est mis en évidence la famille de racine caractérisée par C1 = *dāl*, C2 = *'ayn* et C3 = *bā'*.

— Reconnaissance des éléments lexicaux : le symbole **@** indique également la première entrée lexicale. Les entrées lexicales successives commencent par le symbole **$**. L'entrée lexicale est donc identifiée par **@** ou **$** et se termine par le caractère "retour chariot".



FIGURE 1.4 – La Macrostructure d'AQAM

1.2.2.3 Remplacement des mot-vedettes

L'exemple 1 suivant illustre les entrées lexicales appartenant à la famille de la racine ء ز ء (*'z'*) .

Exemple 1.

@أَزَأَ الغَنَمَ، كَمَنَعَ : أَشْبَعَهَا

$وَ - عَنِ الحَاجَةِ : جَبُنَ وَنَكَصَ

@'aza'a al=ġanama, kamana'a : 'ašba'ahā

$wa - 'ani al=ḥāğati : ğabuna, wa=nakuṣa

Le signe "-" au début de la deuxième entrée réfère au mot-vedette *'aza'a*. Donc la deuxième entrée devrait être :

أَزَأَ عَنِ الحَاجَةِ : جَبُنَ وَنَكَصَ

wa'aza'a 'ani al=ḥāǧati : *"être lâche, avoir peur, se retirer, se détourner."*

Dans la deuxième étape de la phase de préparation du texte d'AQAM, on a développé un script qui remplace le signe "-" au début de chaque entrée par le dernier mot vedette de l'entrée lexicale qui précède.

1.2.2.4 Nettoyage du texte

Comme illustré dans la Figure 1.4, à part la première entrée de chaque famille de racine c-à-d celles qui commencent par le symbole @, nous notons que presque toutes les autres entrées lexicales commencent par la conjonction *wa* "et". Au début de l'entrée, la conjonction *wa* est considérée comme un mot de remplissage sans contenu lexical, donc elle peut être supprimée sans perdre aucune information. Par conséquent, cette phase de nettoyage du texte (text cleaning) consiste à supprimer la conjonction *wa* "et" qui regorge les entrées lexicales. L'opération se produit après avoir vérifié que la première consonne radicale n'est pas *wāw*. Dans ce dernier cas, le traitement est fait avec des contrôles manuels (c-à-d les entrées appartenant au chapitre *wāw* de chaque section).

1.2.2.5 Segmentation de la section *wāw* & *yā'*

Rappelant que la section *wāw* et *yā'* contient les entrées lexicales des deux lettres, ce qui provoque une ambiguïté au niveau de la reconnaissance du dernier radical de la racine, d'où la nécessité de la segmentation de cette section en deux sous-sections. Cependant, l'auteur a tenu à indiquer la valeur de la dernière consonne radicale comme indiqué dans la figure 1.5. On remarque qu'il s'agit d'une root-family finissant par la consonne *wāw* « :و@», par contre la deuxième entrée lexicale dérive d'une racine finissant par la consonne *yā'* « :ي@». Grâce aux caractéristiques stylistiques de l'auteur, la division de la section est faite d'une manière automatique, en se basant sur la valeur de la consonne indiquée par l'auteur d'AQAM au début de chaque root-family. En effet, chacune d'elles est affectée à sa sous-section correspondante, après la suppression du signe désignant la consonne finale

de la racine et les deux point qui la suit (:ي ou bien :و).

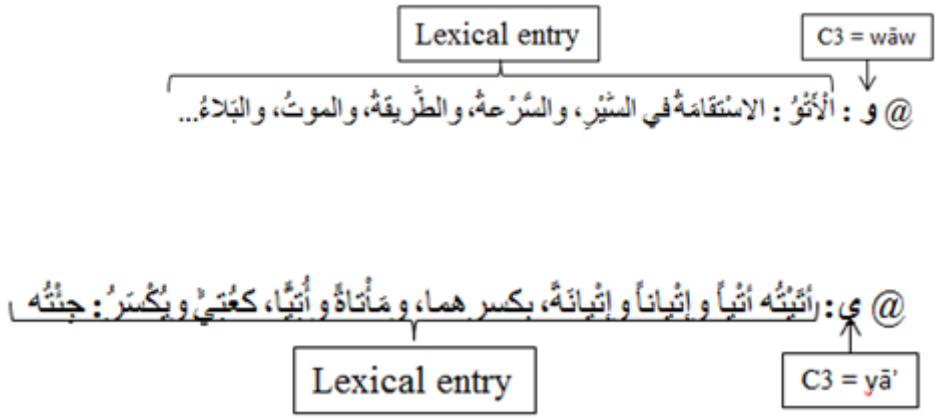


FIGURE 1.5 – Exemple des entrées lexicales de la section wāw & yā'

1.2.2.6 Résultats de la segmentation du texte

La Figure 1.6 illustre les phases de la segmentation du texte.

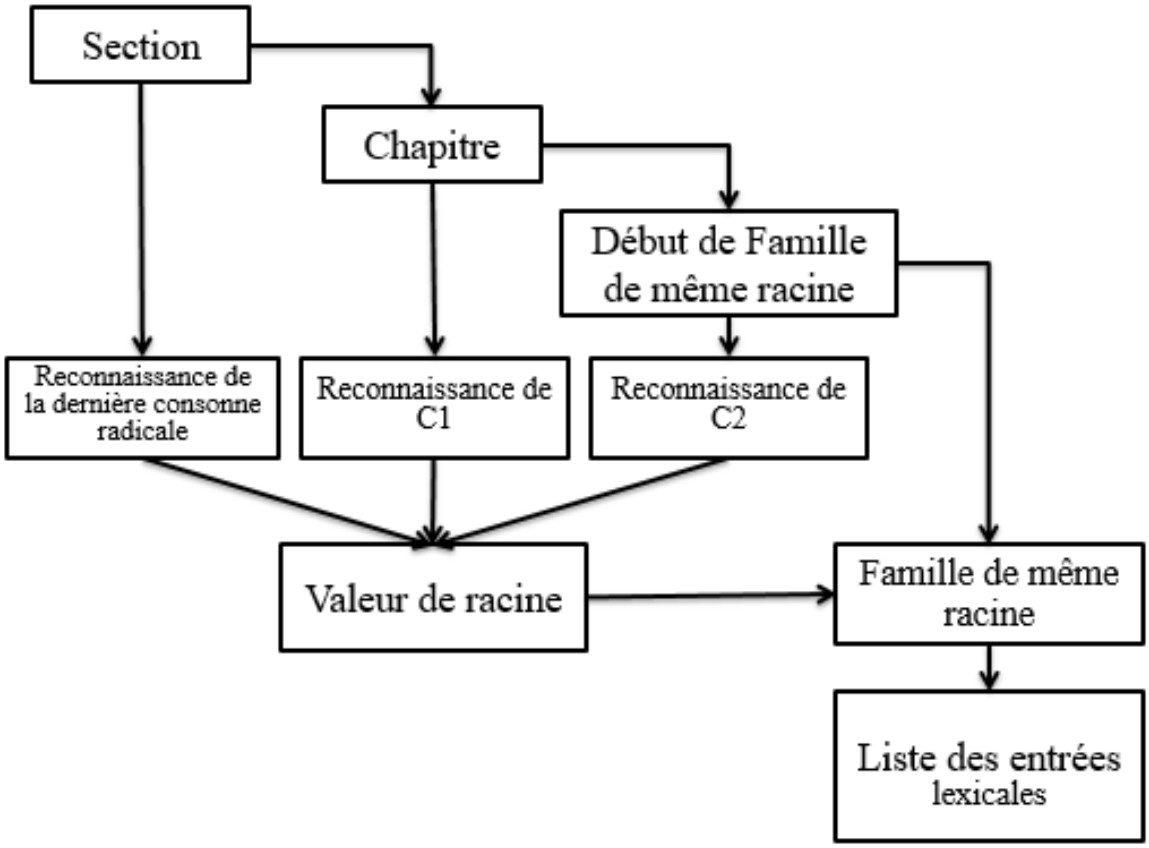


FIGURE 1.6 – Phase 1 : Segmentation de texte selon la macrostructure

La subdivision en sections dont le titre nous permet de reconnaître la valeur de la dernière consonne radicale. Chaque section est divisée en chapitres dont le titre nous

permet de reconnaître la valeur de la première consonne radicale. À l'intérieur de chaque chapitre, le début de chaque famille de racine (et la première entrée lexicale aussi) nous permet d'identifier la deuxième (et éventuellement la troisième) consonne radicale. Enfin, on peut identifier la valeur des racines et les listes d'éléments lexicaux qui en dérivent. Le texte est subdivisé en entrées lexicales autonomes don le début est marqué par l'indicateur @ et la fin est marquée par le retour à la ligne.

En conséquence, on a obtenu 29 fichiers publiés au format Unicode (TXT) en tant que ressources téléchargeables à travers la plate-forme CLARIN-IT [7] :

— 26 sections correspondant aux consonnes à partir du *hamzah* jusqu'au *hā'* ;

— une section contenant les deux consonnes *wāw* et *yā'* divisée en deux sous-sections (fichiers), l'une pour *wāw* et l'autre pour *yā'* ;

— une section pour *al='alif al=layyinah*, c'est-à-dire le *'alif* qui marque l'allongement de la voyelle /a/ à /ā/. Cette section ne contient que 32 entrées de particules, par exemple, إِلَى *'ilā* qui peut être une préposition (à, vers) ou une conjonction (jusqu'au).

Le tableau 1.1 illustre le contenu statistique d'AQAM dans son format Unicode, subdivisé en 29 fichiers, où chaque fichier correspondant à une section du dictionnaire. On note, également que le dictionnaire est composé de 687 chapitres et 10373 racines, faisant 70992 entrées lexicales en total. Le tableau 1.1 détaille pour chaque section le nombre de chapitres, le nombre de racines et le nombre des entrées lexicales qu'elle contient.

1.2.3. Conclusion

La phase du pré-traitement concernant la préparation du texte source (plain text) d'AQAM, consiste dans la normalisation du script et le nettoyage. Afin de convertir AQAM en un format plus utilisable et exploitable par machine, nous avons préparé le texte brut du dictionnaire en y ajoutant des signes de surface, permettant la segmentation suivant la macrostructure et la microstructure du dictionnaire et l'extraction automatique des différentes parties (sections, chapitres, famille de même racine et finalement les entrées lexicales). Nous avons également suivi un contrôle minutieux du texte, de la vocalisation et de la ponctuation. Bien que cette phase a pris beaucoup de temps, elle a été nécessaire pour une conversion digitale et automatique du dictionnaire AQAM dont le processus et ses phases sont expliqués dans le chapitre suivant.

7. "http ://hdl.handle.net/20.500.11752/ILC-97"

TABLE 1.1 – statistiques d'AQAM TXT

Section (bāb)		chapitre (faṣl)	Racine	Entrées Lexicales
hamza		25	272	1499
bā'		28	650	5065
tā'		28	239	987
ṯā'		25	203	894
ǧim		28	461	1921
ḥā'		23	302	2191
ḫā'		26	209	883
dāl		27	458	3544
ḏāl		24	126	490
rā'		28	858	7942
zāy		24	284	1450
sīn		24	522	2564
šīn		25	284	1276
ṣād		21	224	1289
ḍād		21	114	1056
ṭā'		27	345	1902
ẓā'		18	88	337
'ayn		26	426	4025
ġayn		23	137	668
fā'		27	436	3746
qāf		26	468	3617
kāf		26	252	1501
lām		28	833	6400
mīm		28	740	5586
nūn		28	551	3619
hā'		25	154	784
wāw & yā'	wāw	28	374	2877
	yā'	28	331	2778
'alif		-	32	-
Total				
29		687	10373	70992

Chapitre 2

Conversion digitale d'al=qāmūs al=muḥīṭ

2.1. Phases d'extraction d'informations

Les exemples 2 et 3 illustrent respectivement une entrée lexicale verbale et une entrée lexicale nominale.

Exemple 2.

وَآبَهُ اللَّهُ : أَبعَدَهُ

wa='āb-a=hu Allah : 'ab'ad-a=hu

Dans l'exemple 2, l'entrée lexicale commence par le mot-vedette وَآبَهُ *wa='āba=hu* qui est composée par : i) la conjonction *wa= "et"* comme proclitique ; ii) le verbe *'āba* composé de la tige (stem) verbale *'āb* et le suffixe flexionnel *-a* indiquant la troisième personne du singulier et iii) le pronom accusatif *=hu* comme enclitique.

Selon les conventions lexicales arabes classiques, il est important de prendre en considération la présence du pronom accusatif car cela indique que le verbe est transitif. Après les deux points, la forme orthographique أَبعَدَهُ *'ab'ad-a=hu*, composée du verbe *'ab'ad-a* et du pronom accusatif *=hu*, correspond au synonyme du verbe *'āba=hu* : *"Il l'a éloigné"*.

Exemple 3.

وَالأَبَابُ : المَاءُ، وَالسَّرَابُ

wa=al='abāb-u : Al=mā'-u, wa=s=sarāb-u

L'exemple 3 commence par la forme orthographique وَالأَبَابُ *wa=al='abāb-u*, composée par la conjonction *wa= "et"*, l'article défini *al=* ; le nom *'abāb* et le suffixe défini nominatif *-u*. Dans ce cas, il est important de prendre en considération la présence de l'article défini *al=* car il indique que le mot à définir est un substantif. Après les deux points, on trouve deux sens *al=mā'-u "eau"* et *wa=s=sarāb-u "mirage"* séparés par une virgule.

Comme le montre la figure. 2.1, nous avons identifié différentes étapes d'extraction d'informations en fonction des caractéristiques de la microstructure :

— Extraction du mot-vedette qui doit subir un traitement pour obtenir le lemme correspondant selon les conventions lexicales modernes. [1]

— Les informations grammaticales regroupent les informations morpho-syntaxiques de l'unité lexicale :

  i. POS indique la partie du discours (nom, verbe, adjectif, etc.)

  ii. Information syntaxique, par exemple transitivité / intransitivité d'un lemme verbal.

  iii. Informations morphologiques qui indiquent par exemple la catégorie de flexion associée à une unité lexicale, et la catégorie dérivée (trilitère, quadrilitère ou verbe dérivé)

— L'information sémantique regroupe différents sens, de plus, quand il est possible d'identifier la relation entre le lemme et le sens, comme la synonymie, l'antonymie.

— Extraire les mots dérivés associés, comme *maṣdar* lorsqu'il est cité avec un lemme verbal.

1. La forme canonique du lemme verbal dans les dictionnaires arabe moderne est la troisième personne du masculin singulier au passé. Par contre, la forme canonique du lemme des noms est le singulier et pour les adjectifs c'est le masculin singulier.

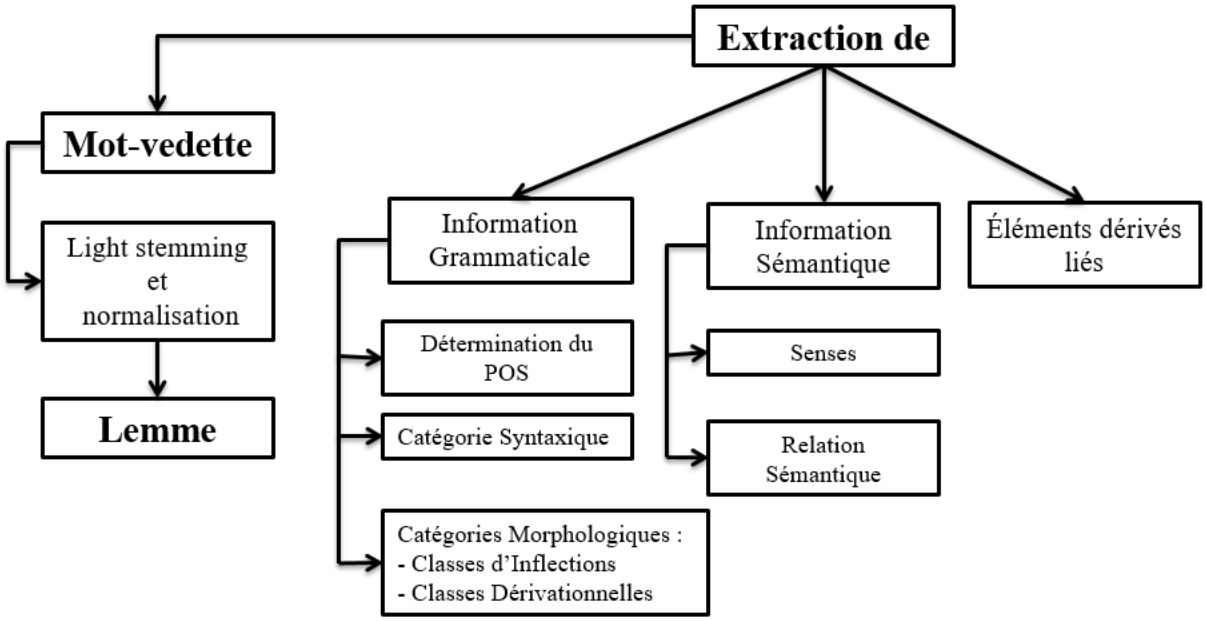


FIGURE 2.1 – Phase 2 : Extraction d'information en fonction de la Microstructure

2.1.1. Extraction et normalisation du lemme

2.1.1.1 Cas des entrées lexicales avec mot-vedette

2.1.1.1.1 Sélection du mot-vedette

La sélection du mot-vedette candidat est basée sur les radicaux de la racine (c'est-à-dire qu'il doit contenir tous les radicaux). Ensuite, il est nécessaire de préciser la nature grammaticale du mot-vedette afin de déterminer quel traitement doit-il subir afin d'obtenir le lemme.

L'algorithme 1 illustre la procédure adoptée. Dans une première phase, il est plus facile de définir des motifs morphologiques de verbes car leur nombre est limité. Premièrement, nous avons défini la liste des suffixes de flexion verbale et la liste des suffixes de transitivité (voir le tableau 3 dans l'annexe). Deuxièmement, nous avons défini une liste regroupant 34 expressions régulières qui présentent d'éventuels motifs verbaux fléchis et dérivés (quelques exemples sont présentés dans le corps de l'algorithme 1). Le mot-vedette est une forme verbale s'il correspond à l'une des motifs verbaux définis.

Dans le cas où il ne correspond à aucun motif verbal, nous testons s'il commence par

l'article défini *al=*, soit se termine par *tanwīn -un* , ou sa dernière lettre est égale à ً, ce qui indique qu'il s'agit d'un cas d'entrée nominale. Finalement, nous devons faire subir le mot-vedette à un processus de normalisation afin d'obtenir le lemme selon les conventions lexicales modernes.

Algorithme 1 : Headword Surface Form Classification

input : Headword W, List NominativeSuffixes, List TransitivitySuffixe, List Verbal_derived_Form, List Nominal_Form, List Adjectival_Form
output : R :Headword Surface Form

```
/* Examples of Verbal_derived_Form                                    */
verbI_a = C + a +"{0,1}?"+ C + a + C + NominativeSuffixe+TransitivitySuffixe+"{0,1}?";
/* fa‘ala                                  فَعَلَ                      */
verbI_cAc= C + a +"{0,1}?"+ "ا" C + NominativeSuffixe+TransitivitySuffixe+"{0,1}?";
/* qaāla                                   قَالَ                       */
verbI_A = C + a +"{0,1}?"+ C + a + "{0,1}?" + "ا";
/* danā                                    دَنَا                       */
verbI_Y = C + a +"{0,1}?"+ C + a + "{0,1}?" + "ى";
/* maŝaY                                   مَشَى                       */
verbI_CC = C + a +"{0,1}?"+ C + C + a;
/* ’abba                                   أَبَّ                       */
verbe_Q = C + a + C + o +C + a + C + NominativeSuffixe+TransitivitySuffixe+"{0,1}?";
/* fa‘lala                                 دَحرَجَ                     */
/* ****************************************                           */
/* Examples of Nominal_Surface_Form                                   */
noun_un = Arabic_Word + "ً" + vowel + "{0,1}?";
/* wardat                                  وَردَة                      */
noun_al = "ال" + Arabic_Word + vowel + "{0,1}?";
/* al=kitāb-u                              الكِتَابُ                   */
noun_add = Arabic_Word + u + Arabic_Word + i ;
/* kitāb-u ar=rajul-i                      كِتَابُ الرَّجُلِ           */
/* ****************************************                           */
/* Examples of Adjectival_Surface_Form                                */
adj_un = Arabic_Word + un + Arabic_Word + un ;
/* hallāb-un                               هَلَّابٌ                    */
adj_Pronoun = pronoun + Arabic_Word + un ;
/* huwa mu’taššib-un                       هُوَ مُؤَتَشِبٌ             */
/* ****************************************                           */
for v in Verbal_derived_Form do
    if W.Matches(v) then
        R ← Verb;
end
for n in Nominal_Surface_Form do
    if W.Matches(n) then
        R ← Noun;
end
for a in Adjectival_Surface_Form do
    if W.Matches(a) then
        R ← ADJ;
end
return R;
```

2.1.1.1.2 Normalisation du lemme

L'exemple 4 illustre l'entrée lexicale verbale dont le mot-vedette تَصَاخَبُوا *taṣāḫab=ū* est constitué par le radical تَصَاخَب *taṣāḫab* et le pronom nominatif *ū* qui indique qu'il s'agit d'un verbe.

De plus, il est accompagné de deux sens تَصَايَحُوا *tašāyaḥ=ū* et تَضَارَبُوا *wa=taḍārab=ū* l'accompagner avec les significations *"ils se crient"* et *"ils se bagarrent ou ils se frappent"* respectivement.

Exemple 4.

تَصَاخَبُوا : تَصَايَحُوا، وَتَضَارَبُوا

taṣāḫab=ū : taŝāyaḥ=ū, wa=taḍārab=ū

Le stemming (radicalisation) est le processus de réduction des mots fléchis (ou dérivés) à leur racine. Par contre le light stemming est le processus de les normaliser à leur forme de base ou canonique.

Le tableau 2.1 illustre les résultats du processus du stemming et du processus de stemming léger (Light-Stemming) du mot-vedette *taṣāḫab=ū* et des deux sens *taŝāyaḥ=ū* et *wa=taḍārab=ū*.

TABLE 2.1 – Stemming vs stemming léger

Mot	Stemming	Stemming léger
taṣāḫab=ū	*ṣḫb*	*taṣāḫab*
taṣāyaḥ=ū	*ṣyḥ*	*taṣāyaḥ*
wa=taḍārab=ū	*ḍrb*	*taḍārab*
Résultat	Racine	Stem

On note que le processus de stemming est agressif dans le sens qu'il réduit les mots à leurs racines, en supprimant les clitiques, les affixes dérivatifs et flexionnels. Par contre, le processus de stemming léger ne supprime que les clitiques et les préfixes et suffixes flexionnels. Par conséquent, cela n'affecte pas la valeur sémantique des mots Duwairi et al. (2007). En fait, cela n'affecte pas les affixes de dérivation, comme le préfixe *ta* et la voyelle longue médiane *ā* qui caractérisent la forme verbale dérivée *taC1āC2aC3-a* dans les verbes de Exemple 4.

Donc, pour ces raisons, on a adopté le stemming léger pour traiter les mots-vedettes et leurs sens, afin d'obtenir une forme canonique standard qui respecte les conventions lexicales modernes.

Dans le cas d'une entrée lexicale nominale, illustrée dans l'exemple 5, le mot-vedette *at̲=t̲awb-u* est accompagné du sens *al=libāsu* avec le sens *"vêtements"*.

Exemple 5.

الثَّوبُ : اللِّبَاسُ

at̲=t̲awb-u : al=libās-u

Dans ce cas aussi, comme le montre le tableau 2.2, le stemming léger respecte les caractéristiques sémantiques des mots et ne supprime que l'article défini (et les éventuelles conjonctions) et la voyelle finale qui marque le cas grammatical.

TABLE 2.2 – Stemming vs stemming léger

Mot	Stemming	Stemming léger
at̲=t̲awb-u	t̲wb	t̲awb
al=libās-u	lbs	libās
Résultat	Racine	Stem

La procédure du stemming léger appliqué dépend des résultats de la nature grammaticale des mots-vedettes, retournée par l'algorithme 1. En fait, nous utilisons un traitement spécial pour chaque catégorie et le processus se fait automatiquement suivant les instructions de l'algorithme 2 décrit ci-dessous.

Algorithme 2 : Headword Light-Stemming

```
input  : Lemma X, POS P, List NominativeSuffixes, List TransitivitySuffixes
output : R : Light stemmed word
if X.getRoot().getC1() <> "و" then
    X ← eliminatePrefixe("و");
if X.getPOS() == "Verb" then
    for ts in TransitivitySuffixes do
        if X.endsWith(ts) then
            X ← eliminateSuffixes(X, ts);
    end
    for ns in NominativeSuffixes do
        if X.endsWith(ns) then
            X ← eliminateSuffixes(X, ns);
    end
    R ← X;
else if X.getPOS() == "Noun" then
    if X.endsWithVowel() then
        X ← X − (Last_Vowel);
    if X.startsWith("ال") then
        X ← X − ("ال");
    if X.charAt(1) ==ṣadda then
        X ← eliminateGemination(X);
    R ← X;
else if X.getPOS() == "ADJ" then
    if X.endsWithVowel() then
        X ← X − (Last_Vowel);
    R ← X;
return R;
```

Dans le cas des verbes, on supprime les suffixes, en utilisant les deux listes prédéfinies contenant les suffixes de transitivités et nominatifs possibles. En conséquence, l'algorithme 2 présenté renvoie le mot-vedette traité. Par exemple, à partir de *taṣāḫab=ū* on obtient *taṣāḫab* qui nécessite une étape de plus en y ajoutant la voyelle *-a* afin d'obtenir la forme canonique verbale du lemme *taṣāḫab-a*.

Dans le cas des noms, on supprime l'article défini et la voyelle finale. De plus, si le premier radical est l'une des lettres solaires, il assimile la lettre *lām* de l'article défini précédent *al=* ce qui donne une consonne doublée. Dans ce cas, l'assimilation peut être exprimée en mettant une marque de gémination *šaddah* sur le premier radical. Après avoir spécifié la liste des lettres solaires, (voir table 4), l'algorithme du stemming léger permet de supprimer les marques de gémination. Par exemple, à partir du mot-vedette الثَّوبُ *aṯ=ṯawb-u* on obtient le mot tronqué (light stemmed word) ثَوب *ṯawb* qui correspond au lemme nominal selon les conventions lexicales modernes.

Lorsque les mots-vedettes ne contiennent pas les consonnes radicales, les entrées ne peuvent pas être sélectionnées en suivant cette procédure. Elles sont exclues et regroupées dans un fichier séparé, ce qui nous permet de les étudier et d'identifier d'autres procédures d'identification, sélection et extraction.

2.1.1.2 Cas de paradigme / mot-moule

Parfois, l'auteur ne mentionne pas le mot-vedette, mais il utilise des mots très connus qui agissent comme motifs pour construire le lemme, comme illustré dans l'exemple 6.

Le mot-vedette commence par la conjonction كَ "*like*" suivie d'un mot populaire, *kitāb-in "book"*, qui peut être utilisé comme moule pour produire le lemme. Après avoir identifié la famille de racine (ح ب ب) (*ḥ b b*), on produit automatiquement le lemme en remplaçant les consonnes du mot-moule donné par les consonnes radicales de la famille de racine. Par conséquent, le lemme construit est حِبَاب *ḥibāb*, ayant le paradigme *C1iC2āC3* "*comme kitāb*".

De plus, le lemme obtenu (*ḥibāb*) est classé comme nom sur la base de la voyelle finale

du mot-moule, *tanwīn al=kasra (-in)* qui correspond à l'une des voyelles déclencheur des cas des noms indéfinis.

Exemple 6.

أُمُّ حُبَابٍ : الدُّنْيَا
كَكِتَابٍ : الـمُحَابَبَةُ

Par contre, l'entrée de Exemple 7 appartient à la famille de racine ع ش ب (*ʿ š b*). Mais, le mot-vedette commence par la conjonction كَـ(*ka "comme"*) suivie du verbe *fariḥa "Il était heureux"*. Dans ce cas, le lemme construit عَشِبَ *ʿašiba* a le paradigme *C1aC2iC3a* et il est reconnu comme verbe trilitère.

Exemple 7.

كَفَرِحَ : يَبِسَ

2.1.1.3 Cas de changement de voyelle

Dans d'autres cas, sans évoquer clairement le lemme, l'auteur mentionne des motifs stylistiques pour indiquer sa vocalisation en s'appuyant sur le lemme précédent.

L'exemple 8 illustre deux entrées lexicales, dont la première a le mot-vedette *al=ǧabāb-u*. Par contre, le second commence par بِالكَسْرِ *bi=al=kasr-i* "*avec la voyelle* /i/".

L'auteur explique dans son introduction que, dans ce cas, il faut changer la voyelle de la première consonne du lemme précédent (*al=ǧ**a**bāb-u*) avec la voyelle /i/. Par conséquence, le mot-vedette de la deuxième entrée est *al=ǧ**i**bāb-u*.

Exemple 8.

الحَبَابُ : القَحطُ الشَّديدُ $
بِالكَسرِ : المُغَالَبَةُ $

La liste des motifs de vocalisation est expliquée dans l'introduction d'AQAM (Al-fīrūz'ābādī (1998), page 20) et résumée dans le tableau 2.3. Le changement de voyelle par

rapport au lemme précédent ne concerne que la première consonne radicale C1. Dans ce cas, l'auteur commence l'entrée lexicale par la préposition بِـ *"avec"* attachée au nom de la nouvelle voyelle qui peut être *fatḥah* /a/, *ḍammah* /u/ ou bien *kasrah* /i/.

Un seul cas, signalé par le motif بِالتَّحْرِيكِ *bi=at=taḥrīk-i*, où *taḥrīk* signifie aussi la voyelle /a/ mais, dans ce cas, elle prendra la place de la voyelle du premier et du deuxième radical (C1 et C2).

TABLE 2.3 – Marqueurs de changement de voyelle

Motif	Translittération	Exemple	changer la voyelle de
بِالفَتْحِ	*avec=la=fatḥah*	C1a...	C1 en /a/
بِالضَّمِّ	*avec=la=ḍammah*	C1u...	C1 en /u/
بِالكَسْرِ	*avec=la=kasrah*	C1i...	C1 en /i/
بِالتَّحْرِيكِ	*avec=la=taḥrīk*	C1aC2a...	C1 et C2 en /a/

2.1.1.4 Cas des adjectifs

Les exemples 9 et 10 expliquent la microstructure de certaines entrées lexicales des adjectifs. Selon les conventions lexicales arabes classiques, l'adjectif peut être introduit comme prédicat dans une phrase nominale.

Dans l'exemple 9, on remarque qu'il y a un sujet indéfini رَجُلٌ *rağul-un* suivi de l'adjectif *'alūb-un*. La phrase signifie "un homme [est] *'alūb-un*".

Exemple 9.

رَجُلٌ **أَلُوبٌ** : سَرِيعُ إِخرَاجِ الدَّلوِ، أَو نَشِيطٌ

*Un homme **'alūb-un** : rapide pour sortir le seau du puits, ou actif*

Dans d'autres cas, le sujet de la phrase nominale peut être un pronom nominatif. Dans l'exemple 10, l'entrée lexicale commence par le pronom nominatif, هُوَ *huwa* qui signifie "Il [est]" suivi de l'adjectif **مُؤَتَشِبٌ** *mu'tašib-un*.

Exemple 10.

وَهُوَ مُؤْتَشِبٌ : غَيْرُ صَرِيحٍ فِي نَسَبِهِ

Il [est] ***mu'taš**ib-un* *: il n'est pas clair de déclarer sa dynastie*

Dans ces cas, le mot-vedette se trouve à la deuxième place et se termine par le suffixe nominatif indéfini *un*.

Dans l'exemple 9, on vérifie si la séquence avant les deux-points correspond à la règle modélisée (1), où on teste si le premier mot W (supposé sujet indéfini) et le mot suivant (supposé un adjectif candidat) se termine par le suffixe nominatif indéfini*un*.

- Règle (1) : W-*un* + ADJ-*un*.

Dans l'exemple 10, le pronom nominatif *huwa* est considéré comme un signe de surface pour capturer les adjectifs. Dans ce cas, on vérifie que la séquence d'entrée correspond à la règle modélisée (2), c-à-d le premier mot est l'un des pronoms nominatifs [2] et le deuxième mot se termine par le suffixe nominatif indéfini *un* .

- Règle (2) : Pronom Nominative + ADJ-*un*.

Ensuite, il faut s'assurer que le deuxième mot contient tous les radicaux avant de décider qu'il s'agit du mot-vedette candidat. Après avoir supprimé le suffixe nominatif *un*, le lemme obtenu est classé comme adjectif.

2.1.1.5 Catégorie grammaticale (Part Of Speech / POS)

2.1.1.5.1 Cas des adjectifs

Pour définir la catégorie grammaticale (POS) du lemme, on utilise des techniques spécifiques à chaque classe grammaticale. Comme déjà expliqué dans la section précédente, l'adjectif est automatiquement étiqueté lorsqu'il est introduit par un sujet non défini ou un pronom nominatif.

2.1.1.5.2 Cas des verbes

Concernant les verbes, on cherche si le lemme obtenu correspond à l'un des paradigmes morphologiques listés dans Table 2.4. [3] Comme on sait, le texte arabe peut ne pas être entièrement vocalisé. Dans notre travail, nous avons vérifié pendant la phase de préparation que le texte était bien vocalisé. Cependant, pour reconnaître qu'un lemme est un

2. Les pronoms nominatifs sont : هُوَ *huwa* - هِيَ *hiya* - هُمَا *humā* - هُم *hum* et هُنَّ *hunna*.

3. Dans le tableau, nous utilisons la translittération de Buckwalter car c'est une translittération caractère par caractère, permettant de représenter tous les caractères du mot, y compris : *sukūn* (o) ; *šaddah* (∼) *hamza-sur-'alif* (>). Pour plus de détails, voir : http ://www.qamus.org/transliteration.htm

TABLE 2.4 – Exemples des paradigmes morphologiques verbales

Verbe	Paradigme Translitéré	Paradigme Arabe
I	C1aC2(a|u|i)C3-a	فَعَلَ ـ فَعُلَ ـ فَعِلَ
II	C1aC2∼aC3-a	فَعَّلَ
III	C1AC2aC3-a	فَاعَلَ
IV	>aC1oC2aC3-a	أَفْعَلَ
V	taC1aC2∼aC3-a	تَفَعَّلَ
VI	taC1AC2aC3-a	تَفَاعَلَ
VII	AinoC1aC2aC3-a	انْفَعَلَ
VIII	AiC1taC2aC3-a	افْتَعَلَ
IX	AiC1oC2aC3∼-a	اِفْعَلَّ
X	isotaC1oC2aC3-a	اسْتَفْعَلَ
Q_I	C1aC2oC3aC4-a	فَعلَلَ
Q_II	taC1aC2oC3aC4-a	تَفَعلَلَ
Q_III	AiC1oC2anoC3aC4-a	اِفعَنلَلَ
Q_IV	AiC1oC2aC3aC4∼-a	اِفعَلَلَّ

verbe, il est nécessaire que toutes les voyelles soient présentes, sinon un contrôle manuel est nécessaire.

2.1.1.5.3 Cas des noms

Le mot-vedette X sélectionné est étiqueté nom s'il correspond à l'une des règles modélisées :

— X se termine par la lettre ة *tā' fermée* qui est une forme orthographique spécifique aux noms.

— (**al=X-u**) : X commence par l'article défini "ال" *al=* et se termine par la voyelle définie nominative *ḍamma* /u/.

— (**X-u**) (**al=Y-i**) : X constitue le premier mot de la construction génitive où X ne commence pas par l'article défini mais se termine par la voyelle définie nominative

/u/ à condition que Y commence par l'article défini, comme illustre l'exemple 11.

Exemple 11.

مَثَاعِبُ الـمَدينَةِ : مَسَايِلُ مَائِهَا

maṯā'ib-u la ville : Le canal de son eau.

Cependant, la qualité du travail nécessite un contrôle manuel en particulier du fichier des noms car il y a des cas, comme dans l'exemple 12, où le lemme est reconnu comme nom mais en fait c'est un adjectif.

Exemple 12.

المُغَلَّبُ : المَغلُوبُ مِرَارًا، وَالمَحكُومُ له بِالغَلَبَةِ

*al=**mugallab**-u : vaincu plusieurs fois et voué à l'échec*

2.1.2. Extraction des informations morphosyntaxiques et dérivatives

2.1.2.1 Cas des verbes

Bien qu'il existe quelques règles phonologiques et sémantiques, les formes imparfaites du verbe trilitère restent ambiguës et varient selon la voyelle médiane. De même, les maṣdars ou noms verbaux des verbes trilitères sont nombreux et variés et non prévisibles. Pour cette raison, les formes imparfaites et les noms verbaux sont indiqués dans les dictionnaires et il est utile de les extraire.

L'exemple 13 illustre la manière dont ces informations sont rapportées dans AQAM. on note qu'il y a deux formes imparfaites *ya'ibb-u* et *ya'ubb-u* qui suivent le mot-vedette verbal *'abb-a*. Après les formes imparfaites, on trouve quatre verbes infinitifs qui se terminent par le suffixe accusatif indéfini *-an* : *'abb-an*, *'abīb-an*, *'abāb-an* et *'abābat-an*.

Exemple 13.

أَبَّ لِلسَّيْرِ، يَئِبّ، وَيَؤُبُّ، أَبًّا، وَأَبِيبًا، وَأَبَابًا، وَأَبَابَةً : تَهَيَّأَ

'abb-a *for walking, ya'ibb-u, and ya'ubb-u, 'abb-an, and 'abīb-an, and 'abāb-an, and 'abābat-an : préparer*

Pour extraire les formes imparfaites, on suit les instructions présentées dans l'algorithme 3. On parcourt la séquence après le mot-vedette, en recherchant un mot qui commence par le préfixe imparfait يَـ *ya*, finissant par *-u* et contenant toutes les consonnes radicales.

Algorithme 3 : Extraction de la forme Imparfait

```
input  : String PlainText, char[] root
output : R : Vector of imperfect Form
seq ← getSeqBeforeColon(PlainText);
tokens ← seq.tokenizer();
for token in tokens do
    if token.startsWith("يَـ") && token.endsWith(u) && token.contains(root)
     then
        R add(token);
end
return R;
```

L'extraction de Maṣdar se fait également automatiquement en vérifiant que le jeton candidat se termine par le suffixe accusatif indéfini *-an* et contient évidemment tous les radicaux de la racine. Avant d'affecter les maṣdars extraits à l'entrée verbale correspondante, on leur applique le processus du stemming léger afin d'éliminer les éventuelles particules et la voyelle finale.

Notons également que les verbes dérivés, en particulier les verbes de cause, sont fréquemment indiqués dans la même entrée lexicale du verbe trilitère correspondante, comme l'éclaircit l'exemple 14 :

Exemple 14.

تَوَهَّمَ : ظَنَّ، **وَأَوْهَمَهُ وَوَهَّمَهُ** غَيْرُهُ

tawahham-a *: to believe, and someone* ***'awham-a=hu****, and* ***wahham-a=hu***

Le lemme verbal *tawahhama* est intransitif et accompagné du sens *ẓanna* avec le sens "penser, croire, supposer". De plus on note les verbes imbriqués au niveau de la dernière séquence, *'awhama=hu* et *wahhama=hu* avec la signification "quelqu'un d'autre le fait croire". La présence du pronom accusatif *=hu "lui"* indique qu'ils sont transitifs. On note également qu'ils sont dérivés de la même racine que le lemme verbal et correspondent à

un motif de causalité >aC1oC2aC3a (أَفْعَلَ) et C1aC2C2aC3a (فَعَّلَ). Après avoir éliminé la conjonction *wa "et"* et le pronom accusatif, et en vérifiant l'existence de tous les radicaux de racines, les verbes *'awham-a* et *wahham-a* sont marqués comme verbes de *cause*.

2.1.2.2 Cas des noms et des adjectifs

Le pluriel brisé ou interne est une caractéristique des noms et des adjectifs. Il est impossible de deviner combien de pluriels peut avoir un nom, si ce n'est en recourant à des dictionnaires comme AQAM. La détection des pluriels brisés est très simple grâce aux motifs de surfaces spécifiques qui les précèdent, et ils sont validés et extraits en fonction des radicaux des racines. Étudions l'exemple 15 suivant :

Exemple 15.

جَذَبَتِ النَّاقَةُ : قَلَّ لَبَنُهَا، **فَهِيَ** جَاذِبٌ وَجَاذِبَةٌ وَجَذُوبٌ، ج. جَوَاذِبُ وَجِذَابٌ

ǧaḏaba-ti la chamelle : son lait a diminué, et elle [est] ǧāḏib-un, et ǧāḏibat-un et ǧaḏūb-un

Le mot-vedette *ǧaḏaba-ti* est une forme fléchie du lemme *ǧaḏaba* et répertorié sous la famille de racine (ج ذ ب) (*ǧ ḏ b*). La séquence citée après les deux-points se compose de trois parties :

1. le sens *qalla laban-u=hā* "son lait a diminué",
2. trois adjectifs dérivés de la même racine *ǧ ḏ b* : *ǧāḏib-un*, *ǧāḏibat-un* et *ǧaḏūb-un* et introduit par la forme de surface d'adjectif *fa=hiya "alors elle [est]"*,
3. deux pluriels brisés d'adjectif : *ǧawāḏib-u* et *wa=ǧiḏāb-un* précédés du motif de surface *ǧ.* (ج) qui est l'acronyme de *ǧam' "pluriel"* (جَمع).

2.1.3. Cas des entrées imbriquées

On peut trouver le cas contraire du dernier cas présenté, c'est-à-dire que le mot-vedette est un adjectif et le verbe trilitère est imbriqué dans l'entrée lexicale. Prenons l'exemple

16 qui présente ce cas.

Exemple 16.

البَاتُّ : المَهزُولُ، **وَقَد** بَتَّ يَبِتُّ بُتُوتاً، وَالأَحمَقُ، وَالسَّكرَانُ.

*al=**bātt**-u* : *al=**mahzūl**-u*, ***wa=qad** batt-a ya-bitt-u butūt-an*, *wa=l=**'aḥmaq**-u*, *wa=s=**sakrān**-u*.

Après le processus du stemming léger, le lemme extrait *bātt* a trois sens :

1. *mahzūl "maigre ; faible"* ;
2. *'aḥmaq* "*stupid*" ;
3. *sakrān* "*ivre*".

Après le premier sens, on trouve également le verbe parfait *batt-a* dérivé de la même racine *b t t* que le mot-vedette. On note qu'il est précédé de la particule spécifique qui ne précède que les verbes وَقَدْ *wa=qad*, correspondant parfois à l'adverbe *"déjà"*.

Dans AQAM, le style d'auteur est très précis, la particule *wa=qad* n'est utilisée que pour introduire le verbe dans l'entrée adjectivale. De plus, le verbe ne concerne que le sens qui le précède.

Par conséquent, dans l'exemple 16, le verbe *batt-a* ne concerne que le premier sens et signifie "*il est devenu décharné*". On note également qu'il est suivi de la forme imparfaite *ya-bitt-u* et du maṣdar *butūt-an*.

Ainsi, la particule *wa=qad* est considérée comme le signe de surface qui déclenche les contrôles automatiques. En conséquence, le système vérifie si le mot suivant (qui est sûrement un verbe) est formé des mêmes radicaux que le mot principal. Ensuite, il crée automatiquement l'entrée lexicale spécifique pour le verbe.

2.1.4. Extraction des sens et des informations lexicales

Après avoir reconnu et extrait le lemme et les différentes caractéristiques morphosyntaxiques, telles que la catégorie grammaticale (POS), l'imperfectif, les maṣdars, le pluriel,

etc. On passe au niveau lexical pour définir les sens, ainsi que pour établir les différentes relations entre les lemmes et leurs sens, telles que la synonymie et l'antonymie.

1. Relation de synonymie : le style de l'auteur est très concis, c'est-à-dire que le sens qui est synonyme se compose d'un seul mot (c'est l'une des raisons pour lesquelles nous avons choisi AQAM). Par exemple, dans l'exemple 16, les sens *mahzūl "décharné ; maigre"* ; *'aḥmaq* "*stupide*" et *sakrān* "*ivre*" sont automatiquement marqués comme synonymes du lemme *bātt*.

2. Relation d'antonymie : dans l'exemple 17, la définition est composée d'une phrase *ḍiddu ǧamada*, dans laquelle le premier mot *ḍiddu ǧamada* signifie *"contraire"*, ainsi que le second mot *ǧ***amada** *"geler"* est marqué comme antonyme du lemme ***d̲āba*** *"fondre, liquéfier"*. Il existe deux motifs de surface ضِدُّ *ḍidd-u* et عَكْسَ *ʿaks-a* dont la signification est *"opposé de"* et permettant de détecter la relation d'antonymie entre le lemme et les sens.

   Exemple 17.

   ذَابَ ذَوبًا وَذَوَبَانًا : **ضِدُّ جَمَدَ**

   ***d̲āba**, d̲awb-an, wa=d̲awabā-an : opposite [to] freeze*

3. Autrement, la définition peut être une phrase explicative, comme dans Exemple 18 lorsque le lemme *sabbābah "l'index"* est défini par la phrase *"suit le pouce"* !

Exemple 18.

السَّبَّابَةُ : تَلِي الاِبهَامَ.

*as=**sabbābat**-u : "suis le pouce"*

2.1.5. Exemples de conversion XML

Après l'extraction de diverses informations morpho-syntaxiques et sémantiques, nous présentons dans cette partie un aperçu des résultats obtenus, qui est décrit dans la conversion XML de quelques exemples riches.

L'exemple XML 2.1 montre la conversion XML de l'entrée discutée dans l'exemple 2 et nous trouvons les balises XML suivantes :

- *root_family* représente la famille de racine à laquelle appartient l'entrée. Dans cet exemple, la famille de racine est C1 = *"hamza"* ; C2 = *"wāw"* ; C3 = *"bā'"*
- *lexical_entry* : délimite chaque entrée.
- *plain_text* : contient la partie originale du texte AQAM dont on extrait les informations.
- *lemma* : contient le lemme extrait du texte brut. Dans cet exemple, le lemme est *'āba*.
- *pos* : pour déclarer la catégorie grammaticale du lemme.
- *morphological_information* : pour désigner le motif morphologique du lemme. Le lemme *'āba* est un verbe trilitère de forme I.
- *morphosyntactic_information* : pour marquer les informations de transitivité du verbe. Le lemme *'āba* est transitif.
- *sense* : tient dans la signification et la définition du lemme concerné. Le verbe *'āba* signifie *'ab'ad-a*.
- *root* : les radicaux de la racine à partir de laquelle le lemme est dérivé (*' w b*), présentée sous forme de script de lettre de la *root_family*.

XML Example 2.1 – Conversion de l'entrée lexicale de *'āba*

```
<root_family c1="الهَمزَة" c2="الوَاو" c3="البَاء">
 <lexical_entry>
  <plain_text>آبَهُ اللَّهُ : أَبْعَدَهُ</plain_text>
  <lemma>آبَ</lemma>
  <pos>Verb</pos>
  <morphological_information>I</morphological_information>
  <morphosyntactic_information>Transitif</morphosyntactic_information>
  <sense att="أَبْعَدَ">
   <synonym att="أَبْعَدَ" />
  </sense>
  <root>ء و ب</root>
 </lexical_entry>
```

```
</root_family>
```

L'exemple XML 2.2 présente les résultats de conversion d'une entrée verbale illustrée dans l'exemple 14. Le système extrait le lemme *tawahhama "croire"* et le reconnaît comme verbe à l'intérieur de la balise <pos> ayant la forme V comme motif morphologique. Les verbes *'awhama* et *wahhama "faire croire"* sont extraits et classés comme action de causalité du lemme. Enfin, le sens *ẓanna* est marqué comme synonyme du lemme *tawahhama*.

XML Example 2.2 – Conversion de l'entrée lexicale de *tawahhama*

```
<root_family c1="الوَاو" c2="الهَاء" c3="المِيم">
 <lexical_entry>
  <plain_text> .ظَنَّ، وَأَوْهَمَهُ وَوَهَّمَهُ غَيْرُهُ :تَوَهَّمَ</plain_text>
  <lemma>تَوَهَّمَ</lemma>
  <pos>Verb</pos>
  <morphological_information>V</morphological_information>
  <cause>أَوْهَمَ</cause>
  <cause>وَهَّمَ</cause>
  <sense att="ظَنَّ">
   <synonym att="ظَنَّ" />
  </sense>
  <root>و ه م</root>
 </lexical_entry>
</root_family>
```

L'extrait XML 2.3 montre la conversion XML de l'entrée dans l'Exemple 15. Le lemme *ǧaḏab-a* est marqué comme verbe de forme I. Le système reconnaît trois adjectifs dérivés du verbe marqué par la balise <adjective> et extrait deux pluriels brisés d'adjectifs marqués par la balise <plural>. Enfin, la balise <sense> contient la phrase *"son lait a diminué"* qui explique le lemme *ǧaḏab-a*.

XML Example 2.3 – Conversion de *ǧaḏaba*

```
<root_family c1="الجيم" c2="الذَّال" c3="البَاء">
 <lexical_entry>
```

<plain_text> جَذَبَتِ النَّاقَةُ :قَلَّ لَبَنُهَا، فَهِيَ جَاذِبٌ وَجَاذِبَةٌ وَجَذُوبٌ، ج. جَوَاذِبُ وَجِذَابٌ، كَنِيَام</plain_text>
<lemma>جَذَبَ</lemma>
<pos>Verb</pos>
<morphological_information>I</morphological_information>
<adjective>جَاذِبٌ</adjective>
<adjective>جَاذِبَةٌ</adjective>
<adjective>جَذُوبٌ</adjective>
<plural>جَوَاذِبُ</plural>
<plural>جِذَابٌ</plural>
<sense att="قَلَّ لَبَنُهَا" />
<root>ج ذ ب</root>
</lexical_entry>
</root_family>

L'exemple XML 2.4 illustre les informations extraites de l'objet d'entrée lexicale de l'Exemple 16. La première information récupérée est le lemme *ḏāba* et sa catégorie grammaticale comme Verbe dans <pos> et sa classe morphologique comme forme I dans <morphologique_information>. Les noms verbaux *ḏawb et ḏawabān* sont extraits et marqués comme maṣdars. Ensuite, les verbes *'aḏāba* et *ḏāwwaba* "*dissoudre quelque chose*" sont marqués par la balise cause. Enfin le verbe *ǧamada* est marqué comme antonyme du lemme *ḏāba*.

XML Example 2.4 – Conversion de *ḏāba*

<root_family c1="الذَّال" c2="الوَاو" c3="البَاء">
<lexical_entry>
<plain_text> ذَابَ ذَوبًا وَذَوَبَانًا، مُحَرَّكَةً :ضِدُّ جَمَدَ، وَأَذَابَهُ غَيرُهُ وَذَوَّبَهُ.</plain_text>
<lemma>ذَابَ</lemma>
<pos>Verb</pos>
<morphological_information>I</morphological_information>
<masdar>ذَوب</masdar>
<masdar>ذَوَبَان</masdar>
<cause>أَذَابَ</cause>

```
  <cause>ذَوَّبَ</cause>
  <sense att="ضِدُّ جَمَدَ">
   <antonym att="جَمَدَ" />
  </sense>
  <root>ذ و ب</root>
 </lexical_entry>
</root_family>
```

Le résultat obtenu après la phase de conversion de l'entrée mentionnée dans l'exemple 17, est présenté en deux entrées lexicales séparées. Le premier, comme indiqué dans l'exemple XML 2.5, concerne le lemme *bātt* accompagné de définitions marquées comme synonymes : *mahzūl "émacié", 'aḥmaq "insensé" et sakrān "ivre".*

XML Example 2.5 – Conversion XML de *bātt*

```
<root_family c1="البَاء" c2="التَّاء" c3="التَّاء">
 <lexical_entry>
  <plain_text> .البَاتُّ: المَهزُولُ، وَقَد بَتَّ يَبِتُّ بُتُوتاً، وَالأَحمَقُ، وَالسَّكرَانُ</plain_text>
  <lemma>بَاتّ</lemma>
  <pos>Noun</pos>
  <sense att="مَهزُول">
   <synonym att="مَهزُول" />
  </sense>
  <sense att="أَحمَق">
   <synonym att="أَحمَق" />
  </sense>
  <sense att="سَكرَان">
   <synonym att="سَكرَان" />
  </sense>
  <root>ب ت ت</root>
 </lexical_entry>
</root_family>
```

Une autre entrée lexicale est créée séparément pour le verbe imbriqué *batt-a* et montrée dans l'exemple XML 2.6. Ayant le même texte brut que l'entrée incubateur, le lemme

batta est suivi de sa forme imparfaite *ya-bitt-u*. Les deux étiquettes <pos> et <morphological_information> indiquent qu'il s'agit d'un verbe de forme I. Enfin, on trouve la forme maṣdar *butūt* et enfin la racine *(b t t)*.

XML Example 2.6 – Conversion XML de *batta*

```
<root_family c1="البَاء" c2="التَّاء" c3="التَّاء">
 <lexical_entry>
  <plain_text> .المَهزُولُ، وَقَد بَتَّ يَبِتُّ بُتُوتاً، وَالأَحمَقُ، وَالسَّكرَانُ :البَاتُّ</plain_text>
  <lemma>بَتَّ</lemma>
  <imperfective>يَبِتُّ</imperfective>
  <pos>Verb</pos>
  <morphological_information>I</morphological_information>
  <masdar>بُتُوت</masdar>
  <root>ب ت ت</root>
 </lexical_entry>
</root_family>
```

2.1.6. Version digitale de *al=qāmūs al=muḥīṭ*

Le système conçu est capable de convertir automatiquement toutes les sections, identifier les entrées lexicales, les lemmes en plus des informations lexicales, grammaticale et dérivationnelles qui les concernent. Le tableau 2.5 illustre le nombre de verbes, noms, adjectifs contenus dans chaque section et le tableau 2.6 illustre le nombre de verbes trilitères, quadrilitères et dérivés contenus dans chaque section. En analysant les résultats obtenus, notre système est capable de reconnaître un très grand nombre d'entrées à partir de AQAM. En raison de l'énorme quantité de données, nous avons contrôlé les résultats obtenus pour la section *d̲āl*, qui contient 95 verbes, 270 noms et 27 adjectifs. Nous avons pris en considération :

- TP (True Positives), où nous avons correctement prédit la catégorie grammaticale des lemme. Par exemple, le système prédit que le lemme est un Verbe et qu'il s'agit en réalité d'une entrée verbale.
- FP ou (False Positives), où nous avons mal prédit la catégorie grammaticale du lemme. Donc, le résultat du système est faux, quand il prédit que le mot d'entrée est un Verbe, mais en fait il s'agit d'une entrée non verbale.

TABLE 2.5 – Quelques statistiques de conversion XML d'AQAM

Section (bāb)	Verbes	Noms	Adjs
hamza	637	569	19
bā'	1244	2961	385
tā'	285	487	48
t̲ā'	227	483	31
ǧim	507	1024	98
ḥā'	661	1078	138
ḫā'	322	435	63
dāl	855	1716	163
d̲āl	92	268	23
rā'	2052	4231	353
zāy	453	738	68
sīn	575	1356	104
šīn	447	612	73
ṣād	448	623	79
ḍād	376	388	53
ṭā'	649	866	77
ẓā'	121	148	15
'ayn	1306	1846	227
ġayn	278	289	32
fā'	1135	1767	209
qāf	1071	1728	201
kāf	468	732	70
lām	1834	3191	319
mīm	1450	2685	232
nūn	835	1756	142
hā'	271	307	43
wāw	638	1182	47
yā'	440	1156	60
Total			
28	**19677**	**34233**	**3204**

– FN (False Negatives), lorsque le système exclut incorrectement un lemme. Par exemple, le lemme n'est pas considéré comme un verbe, mais en fait il l'est. Donc le résultat du système est erroné.

Enfin, le tableau 2.7 présente des résultats statistiques concernant la section *d̲āl*, respectivement :

– Précision : Nombre total d'entrées extraites et classées correctement / Nombre total d'entrées extraites.

$P = \frac{TP}{TP+FP}$

– Rappel : Nombre total d'entrées correctement extraites et classées / Nombre total d'entrées pertinentes dans AQAM.

TABLE 2.6 – Statistiques des entrées verbales extraites d'AQAM

Section	I	II	III	IV	V	VI	VII	VIII	IX	X	Q_I	Q_II	Q_III	Q_IV	Total
hamza	397	25	25	66	24	6	8	5	-	11	56	10	4	-	637
bā'	626	109	36	192	77	17	27	19	1	36	71	17	-	16	1244
tā'	180	17	9	31	11	1	11	9	1	5	9	-	-	1	285
ṯā'	126	18	1	15	19	3	9	6	1	11	13	4	-	1	227
ǧim	234	36	7	85	44	4	14	18	3	13	38	6	3	2	507
ḥā'	324	54	23	106	45	14	12	14	2	17	37	8	5	-	661
ḫā'	166	37	6	36	23	5	10	9	6	3	18	3	-	-	322
dāl	358	84	25	165	74	15	27	12	4	42	27	5	1	16	855
ḏāl	36	6	5	18	1	2	5	-	1	5	7	6	-	-	92
rā'	744	220	61	321	174	43	91	41	12	82	173	40	6	44	2052
zāy	216	37	14	55	35	11	23	5	-	14	25	9	3	6	453
sīn	255	32	21	68	45	25	21	12	5	5	63	20	3	-	575
šīn	196	37	8	52	37	9	34	9	1	10	32	15	2	5	447
ṣād	212	47	9	56	25	7	23	16	4	4	37	8	-	-	448
ḍād	171	40	10	72	25	3	19	5	5	13	9	4	-	-	376
ṭā'	312	42	20	75	50	10	35	21	2	14	45	12	3	8	649
ẓā'	72	7	4	15	6	-	5	1	1	4	6	-	-	-	121
'ayn	637	103	28	203	109	14	39	51	2	36	57	20	7	-	1306
ḡayn	133	20	6	45	29	2	3	13	-	5	18	4	-	-	278
fā'	487	100	32	242	69	16	41	28	3	49	48	9	1	10	1135
qāf	487	100	35	183	69	7	33	53	9	19	56	15	5	-	1071
kāf	297	11	4	27	25	13	37	5	9	10	19	8	3	-	468
lām	715	159	57	286	153	31	79	29	7	70	181	30	1	36	1834
mīm	666	129	40	241	120	13	57	17	6	29	82	29	8	13	1450
nūn	394	77	31	141	53	15	31	3	4	29	28	4	-	25	835
hā'	132	35	14	31	24	6	1	2	-	3	18	5	-	-	271
wāw	588	3	10	21	2	3	7	-	1	-	3	-	-	-	638
yā'	375	19	12	16	3	2	7	-	-	5	1	-	-	-	440
Total															
28	**9536**	**1604**	**553**	**2864**	**1371**	**297**	**709**	**403**	**90**	**544**	**1177**	**291**	**55**	**183**	**19677**

$R = \frac{TP}{TP+FN}$,

– F-Mesure : la moyenne harmonique pondérée de Précision et Rappel.

$F = \frac{2 \cdot P \cdot R}{P+R}$.

Notre système a une précision de 100%, 95% pour le rappel et 97% pour la F-mesure pour les verbes. Pour les noms, nous avons 99% pour la précision, 97% pour le rappel et 98% pour la F-mesure. Enfin pour les adjectifs, nous avons obtenu 100% pour la précision, 85% pour le rappel et 92% pour la F-mesure.

TABLE 2.7 – Statistiques d'extraction des entrées d'AQAM

Classe	**Précision**	**Rappel**	**F-Mesure**
Verbes	100%	95%	97%
Noms	99%	97%	98%
adjectifs	100%	85%	92%

Certains cas doivent être résolus manuellement, par exemple, les entrées lexicales indiquant que le système n'a pas pu reconnaître leurs POS. Dans l'exemple 19, le lemme عَلبَى *'albā* dérivé de la racine trilitère (ع ل ب) *('l b)* a un modèle verbal annexé au modèle verbal quadrilatitère. Autrement, pour l'Exemple 20, l'absence de la voyelle médiane /**a**/

dans le texte AQAM utilisé du lemme قَأَبَ *qa'**a**ba*, crée une ambiguïté dans la classification du lemme comme verbe, donc pour résoudre ces cas, nous devons ajouter cette voyelle dans le texte brut AQAM.

Le système regroupe les cas non reconnus dans un fichier séparé, dont on extrait toutes les parties des éléments lexicaux reconnus sans spécifier la catégorie grammaticale. Le résultat extrait à partir de l'exemple 19 et 20 sont affichés dans l'exemple XML 2.7 et 2.8 respectivement.

Exemple 19.

عَلبَى الرَّجُلُ : ظَهَرَتْ عَلَابِيُّهُ كِبَرَاً

'albā *l'homme : ẓaharat 'alābiyyuh-u kibar-an*

Exemple 20.

قَأَبَ الطَّعَامَ : أَكَلَهُ

qa'ba *la nourriture : il l'a mangée*

XML Example 2.7 – Conversion XML de *'albaā*

```
<root_family c1="العين" c2="اللَّام" c3="البَاء">
 <lexical_entry>
 <plain_text> عَلبَى الرَّجُلُ: ظَهَرَتْ عَلَابِيُّهُ كِبَرَاً</plain_text>
  <lemma>عَلبَى</lemma>
  <pos>-</pos>
  <sense att="ظَهَرَتْ عَلَابِيُّهُ كِبَرَاً">
  <root>ع ل ب</root>
 </lexical_entry>
</root_family>
```

XML Example 2.8 – Conversion XML de *qa'ba*

```
<root_family c1="القَاف" c2="الهَمزَة" c3="البَاء">
 <lexical_entry>
  <plain_text> قَأَبَ الطَّعَامَ: أَكَلَهُ</plain_text>
  <lemma>قَأَبَ</lemma>
  <pos>-</pos>
  <sense att="أَكَلَهُ">
  <root>ق ء ب</root>
 </lexical_entry>
</root_family>
```

Un autre exemple apparaissant dans le fichier des cas non reconnus est illustré dans l'Exemple 21 qui est une expression idiomatique, et qui devrait être résolu manuellement.

Exemple 21.

جَازِئُكَ مِن رَجُلٍ : نَاهيكَ

ğāzi'uk-a min rağul-in : "cet homme te suffira"

Enfin, les exemples 22 et 23 illustrent deux lemmes, le premier est *hā'i*, c'est une humeur impérative de *hā'a* qui signifie *"Donnez ! Apportez !"*. Le second est le nom *ğabādi*. Les deux ont un motif inhabituel se terminant par *kasra*. De plus, ils ont différents POS. Donc, la catégorie grammaticale de ces cas devraient être étiquetés manuellement.

Exemple 22.

هَاءِ، بِالكَسرِ : هَاتِ

hā'i, bi=al=kasri : "Donner ! Apporter !"

Exemple 23.

جَبَاذِ : المَنِيَّةُ، أو النِّيَّةُ الجَابِذَةُ

ğabādi : "Sort ; destin ; mort"", ou "idée attrayante"

2.2. Conclusion

Dans cette section, on a expliqué les méthodologies suivies dans chaque phase pour obtenir une version complète et numérique d'AQAM. Les résultats dépendent de la connaissance de la structure profonde d'AQAM, du style de son auteur et du contrôle systématique de la ponctuation à l'intérieur des entrées lexicales.

Les fichiers XML qui résultent de cette phase de travail sont également publiés en tant que sources libres au moyen de l'infrastructure *CLARIN-IT*[4]. Ainsi, dans chaque dossier correspondant à une section, en plus du fichier texte ".TXT" enrichi d'indicateurs et signes de surface qui marquent le début des chapitres ; les familles de racines et les entrées lexicales, on trouve également un dossier contenant les fichiers XML extraits et répartis selon les catégories grammaticales d'entrées lexicales (Verbe ; Noms ; Adjectifs), en plus d'un fichier où sont regroupées les formes non identifiées par le système.

Pour construire une ressource digitale plus complète et exhaustive, le troisième chapitre propose une méthode pour l'alimenter grâce à l'intégration des traductions anglaises des lemmes et à l'insertion des liens qui permettent de les connecter aux concepts déjà définis dans des ressources anglaises préexistantes telles que Princeton WordNet (PWN) et Suggested Upper Merged Ontology (SUMO).

4. http ://hdl.handle.net/20.500.11752/ILC-97. Nous pouvons également trouver les dossiers enregistrés en cherchant le mot *"qāmūs"* sur l'adresse : https ://vlo.clarin.eu/

Chapitre 3

Marquage sémantique automatique d'AQAM en utilisant PWN et SUMO

3.1. Introduction

Actuellement, les lexiques sémantiques et les ontologies lexicales sont des ressources importantes dans le traitement sémantique du langage naturel qui devient de plus en plus important à mesure que la technologie progresse. Pour ces raisons, nous avons pensé à enrichir la nouvelle version numérique obtenue de AQAM avec des informations sémantiques et ontologiques. Notre motivation d'augmenter la capacité de traitement de la langue arabe a été réalisée en deux phases :

1. en enrichissant la ressource digitale de AQAM avec les traductions anglaises des lemmes et cette phase a été effectuée sur toutes les sections de AQAM ;
2. en liant la version numérique de AQAM à des ressources ontologiques préexistantes telles que Princeton WordNet (PWN) et Suggested Upper Merged Ontology (SUMO). Cette phase pilote de mapping entre AQAM et PWN et SUMO a été effectuée sur la section de la lettre *bā'*.

3.2. Ressources

3.2.1. Lexiques bilingues

Afin de relier, autant que possible, les entrées lexicales de l'AQAM aux concepts SUMO, nous utilisons comme "pont" deux lexiques arabe-anglais, "Dictionary of Mo-

dern Written Arabic" et "An Advanced Learner's Arabic-English Dictionary" que nous décrivons ci-dessous.

3.2.1.1 An Advanced Learner's Arabic-English Dictionary

Le dictionnaire bilingue "An Advanced Learner's Arabic-English Dictionary" (ALAED), Compilé par H. Anthony Salmoné a été publié en 1889. Dans son introduction, l'auteur a précisé qu'il utilisait, entre autres, "An Arabic English Lexicon" d'Edward William Lane, lui-même basé sur *tāğ al=ʻarūs min ğawāhir al=qāmūs* pour az=zabīdī (2001) qui est le dictionnaire étendu d'AQAM. Heureusement, nous avons une version lisible par machine au format TEI (Text Encoding Initiative) disponible en ligne. [1]

3.2.1.2 Dictionary of Modern Written Arabic

Dictionnaire de l'arabe écrit moderne (DMWA) est un dictionnaire bilingue arabe-anglais compilé par l'arabe-allemand Hans Wehr (1909-1981). Le dictionnaire présente le vocabulaire de l'arabe moderne, il est basé sur la forme de la langue utilisée à travers le monde arabe de l'Irak au Maroc. Ainsi selon l'auteur de ce dictionnaire, les index et collections disponibles de la terminologie arabe dans les domaines de la religion (islam et de l'Église orientale), de la jurisprudence, de la philosophie, de la grammaire arabe, de la botanique et autres ont été utilisés (Wehr (1976)). Le dictionnaire organise ses entrées selon l'ordre traditionnel des racines arabes.

(*) préparation du DMWA :

(a) Normalisation : Avant de démarrer la phase de traduction, nous avons effectué un pré traitement sur le lexique bilingue DMWA. Le format numérique du DMWA a été obtenu à l'aide d'un système de reconnaissance optique de caractères OCR (voir la figure. 3.1) avec quelques erreurs de reconnaissance. Parfois, il confond les lettres arabes avec une calligraphie presque similaire, et la seule différence entre elles est la position du point, comme :

(i) ت "t" & ث "$\underline{t}$"

1. Fondé par la National Science Foundation, publié par les administrateurs de l'Université Tufts, dans le cadre du projet Perseus : http ://www.perseus.tufts.edu.

(ii) د "d" & ذ "$\underline{d}$"

(iii) ص "$ṣ$" & ض "$ḍ$"

(iv) ح "$ḥ$" & ج "$ǧ$" & خ "$\underline{ḥ}$"

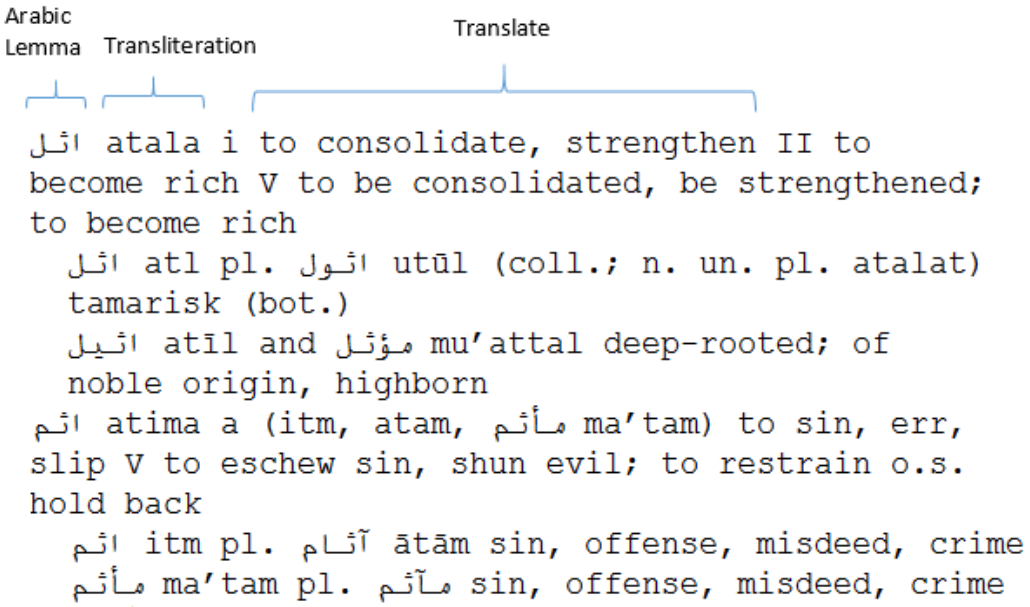


FIGURE 3.1 – Structure du DMWA avant les procédures de normalisation.

Dans les deux exemples suivants d'entrées lexicales DMWA, درب *dariba* et ذرب *dariba* avaient la même translittération (*dariba*) :

- درب *dariba* a (*darab*, درب *daruba*) to be accustomed, be used (ب to), be practiced, trained, skilled
- ذرب *dariba [dariba]* a (*darab*) to be sharp, cutting

Pour éviter ces confusions, on a généré une nouvelle version de DMWA dont on a vérifié et corrigé tous ces cas ambiguës, en générant la nouvelle translittération correcte, et en gardant l'ancienne entre crochets. Après, on a fait une normalisation pour les deux ressources AQAM et DMWA en utilisant le même système de translittération.

(b) Génération des formes dérivées : On a divisé le fichier DMWA en deux fichiers, un pour les verbes et le second pour les noms, afin d'économiser les ressources lors de la phase de recherche des traductions. Par exemple, lors de la traduction de verbes, on n'utilise que le fichier des verbes DMWA (comme le montre l'extrait de fichier dans la Figure 3.2). Ainsi que la nouvelle version DMWA a été bien structurée en y ajoutant la racine pour chaque entrée lexicale de la même famille de racine.

```
أثـل
أثـل ʾaṯala i to consolidate, strengthen
أثـل II [ʾaṯṯala] to become rich
أثـل V [taʾaṯṯala] to be consolidated, be
strengthened; to become rich
أثـم
أثـم ʾaṯima a (itm, atam, مـأثـم maʾtam) to
sin, err, slip
أثـم V [taʾaṯṯama] to eschew sin, shun
evil; to restrain o.s. hold back
```

FIGURE 3.2 – Extrait du fichier restructuré des verbes.

Pour les entrées verbales, on note qu'il mentionne les verbes dérivés avec le nombre romain correspondant à la classe des formes, nous avons donc créé une nouvelle entrée lexicale pour chaque forme dérivée (voir Tableau. 2.4). Tel que dans l'exemple suivant :

- اثل *aṯala* i to consolidate, strengthen II to become rich V to be consolidated, be strengthened ; to become rich

Après avoir divisé l'entrée en entrées lexicales séparées, on génère la translittération spécifique pour chaque forme dérivée de chaque nouvelle entrée lexicale créée en utilisant les règles mentionnées dans l'annexe a, de sorte que l'entrée précédente est transformée en :

- اثل *ʾaṯala* i to consolidate, strengthen
- اثل II [*ʾaṯṯala*] to become rich

– اثل V [*ta'a<u>tt</u>ala*] to be consolidated, be strengthened ; to become rich

(c) Translittération : La translittération est le processus de conversion de mots d'un alphabet d'une langue source donnée en un alphabet d'une langue cible, de manière à préserver les aspects phonétiques et orthographiques des mots translittérés (Ameur et al. (2017)). Malheureusement, il n'y a pas de système conventionnel unifié disponible que les savants du monde entier puissent utiliser à cette fin(Alginahi et al. (2017)). Nous avons donc utilisé deux systèmes de translittération différents, chaque système étant spécifique à chaque dictionnaire.

Par exemple, comme illustré dans l'entrée DMWA suivante :

– اثل ***a<u>t</u>ala*** i to consolidate, strengthen

En fait, si le lemme arabe commence par *hamza*, il n'est pas translittéré, donc nous ajoutons la translittération *hamza* "'" à tous ces cas afin de normaliser la translittération. Par conséquent, dans l'exemple précédent, nous remplaçons la translittération précédente par ***'a<u>t</u>ala***.

3.2.2. Princeton WordNet (PWN)

WordNet (PWN3.0) est une grande base de données lexicale pour l'anglais Fellbaum (1998b). [2] Elle regroupe les mots en fonction de leur signification dans des ensembles de synonymes. L'ensemble de synonymes est appelé ***synset***. Chaque synset représente un *concept* sous-jacent, c'est-à-dire qu'un ensemble de synonymes partage la même signification dans un contexte donné (Miller (1995)). De plus, les synsets sont inter-connectés au moyen de relations lexicales et conceptuelles-sémantiques, telles que : synonymie, antonymie, hyponymie, hypernymie, méronymie, troponymie, etc.

3.2.3. Suggested Upper Merged Ontology (SUMO)

En science de l'information, une ontologie supérieure (Upper Ontology), (également connue sous le nom de Top level ou Foundation ou Universal Ontology) est définie comme

2. http ://wordnet.princeton.edu/

«une ontologie qui décrit des concepts très généraux qui sont les mêmes dans tous les domaines de connaissances» (Ladislav et al. (2012)). L'ontologie suggérée supérieure fusionnée (SUMO) [3] a été créée en fusionnant un certain nombre d'ontologies de niveau supérieur existantes. De plus, c'est la seule ontologie formelle qui a été mappée à tout le lexique WordNet [4] (Pease et al. (2002) ; Niles and Pease (2003)). La base de données utilisée contient le mapping de la base de données lexicale WordNet vers SUMO. Le dernier ensemble de mapping concerne WordNet 3.0 et s'appelle WordNetMappings30-xxx.txt, où «xxx» est l'un des [nom, verbe, adj, adv]. Ci dessous un exemple de la base de données utilisée, illustrant une entrée WordNet enrichie par les concepts SUMO :

```
01828405 37 v 03 hanker 0 long 2 yearn 0 007 @ 01825237 v 0000 + 10270878
n 0303 + 07486628 n 0302 + 10270878 n 0201 + 07486628 n 0201 + 07486922 n
0101 ~ 01805684 v 0000 01 + 22 00 | desire strongly or persistently
&%IntentionalPsychologicalProcess+
```

La première partie de l'enregistrement indique que le numéro 01828405 est l'identifiant unique du verbe synset {hanker, long, yearn}. Il est relié à d'autres synsets par des relations lexicales et hiérarchiques grâce à des pointeurs étiquetés. Le symbole "@" indique que ce synset est subsumé par le synset {desire, want } dont l'identifiant est 01825237. Chaque synset après le symbole "+" regroupe des noms liés par dérivation, par exemple, le synset 10270878 {longer, thirster, yearner}. Enfin le symbole $\sim$ indique la relation de troponymie avec le synset 01805684 {ache, yearn, yen, pine, languish}. Après le signe "|" on retrouve la glosse définissant le sens du synset (*desire strongly or persistently*) [5] et enfin en dernière position le concept SUMO correspondant (*IntentionalPsychologicalProcess+*) qui est désigné par le symbole "&%". Le concept se termine par le symbole "+" qui indique (en position finale) que le synset 01828405 {hanker, long, yearn} est subsumé par le concept SUMO. Il y a trois symboles, "=", "+" ou "@", qui indiquent la relation précise entre le concept SUMO et le synset WordNet. Ils signifient, respectivement, que le synset WordNet est équivalent en sens, subsumé ou une instance du concept SUMO.

3. http ://www.adampease.org/OP/

4. Les ressources SUMO et PWN3.0 utilisées dans cette thèse sont disponibles en téléchargement sur le lien :
https ://github.com/ontologyportal/sumo/blob/master/WordNetMappings/

5. Pour plus de détails sur tous les composants constituant la ligne vous pouvez consulter le lien suivant : https ://wordnet.princeton.edu/documentation/wndb5wn

3.3. Méthodologie

La première étape consiste à traduire le lemme à l'aide du dictionnaire arabe-anglais. Pourtant, la polysémie est l'un des problèmes majeurs de la traduction automatique. Cependant, le mot polysémique n'a qu'une seule signification dans un contexte spécifique.

Nous avons noté (généralement) que l'auteur d'AQAM a défini la signification du lemme en le présentant dans un exemple / contexte spécifique et le sens est un seul mot considéré comme synonyme du lemme qui spécifie sa signification. Par exemple, le lemme *'abba* est plus polysémique dans un contexte isolé et peut être traduit par : *"prepare, get ready to, yearn, long"*.

Cependant dans le contexte " *'abba* sa patrie" de l'exemple 24, le lemme *'abba* a le sens *ištāqa* traduit par : *"desire,* ***yearn, long****"*.

Par contre dans le contexte " *'abba* pour la marche" de l'exemple 25, cela signifie *tahayya'a* qui se traduit : *"be ready for ; be possible ; feasible for ;* ***prepare for****"*.

Exemple 24.

أَبَّ إِلَى وَطَنِهِ أَبًّا وَ إِبَابَةً وَأَبَابَةٌ : **اشْتَاقَ**

'abba *sa patrie ..... :* ***iŝtāqa***

Exemple 25.

أَبَّ لِلسَّيْرِ يَئِبُّ وَيَؤُبُّ أَبًّا وَأَبِيبًا وَأَبَابًا وَأَبَابَةً : **تَهَيَّأَ**

'abba *pour la marche ..... :* ***tahayya'a***

Ainsi pour traiter le problème de polysémie, en plus de la traduction des lemmes, nous avons également eu recours à la traduction des sens qui l'accompagnent lorsqu'il s'agit d'un verbe ou d'un adjectif ou d'un nom commun.

Pour lever l'ambiguïté sur la signification du lemme, nous considérons un algorithme qui prend en compte les mots traduits en anglais qui résultent de l'intersection entre les traductions des lemmes et les traductions des sens. Dans Exemple 24, le lemme *'abba* est traduit par *"****yearn, long****"*. Par contre dans l'exemple 25, il se traduit par *"****prepare for****"*.

Dans le cas spécifique des noms propres, on ne traduit pas justement le lemme mais on le classe grâce au sens qui l'accompagne. Dans l'exemple 26, le lemme *ğal'ib-un* est un nom propre de *"**une montagne** dans la Médina"* et donc c'est un toponyme c'est-à-dire un nom de lieu.

Exemple 26.

جَلعِبٌ : جَبَلٌ بِالمَدِينَةِ

ğal'ib-un : "une montagne dans la Médina"

Figure. 3.3 présente un résumé et les étapes de la procédure adoptée. Après le processus de traduction, les étapes suivantes consistent à rechercher d'éventuels équivalents de traduction dans Princeton WordNet et SUMO afin de trouver des synsets et des concepts qui peuvent correspondre au lemme arabe.

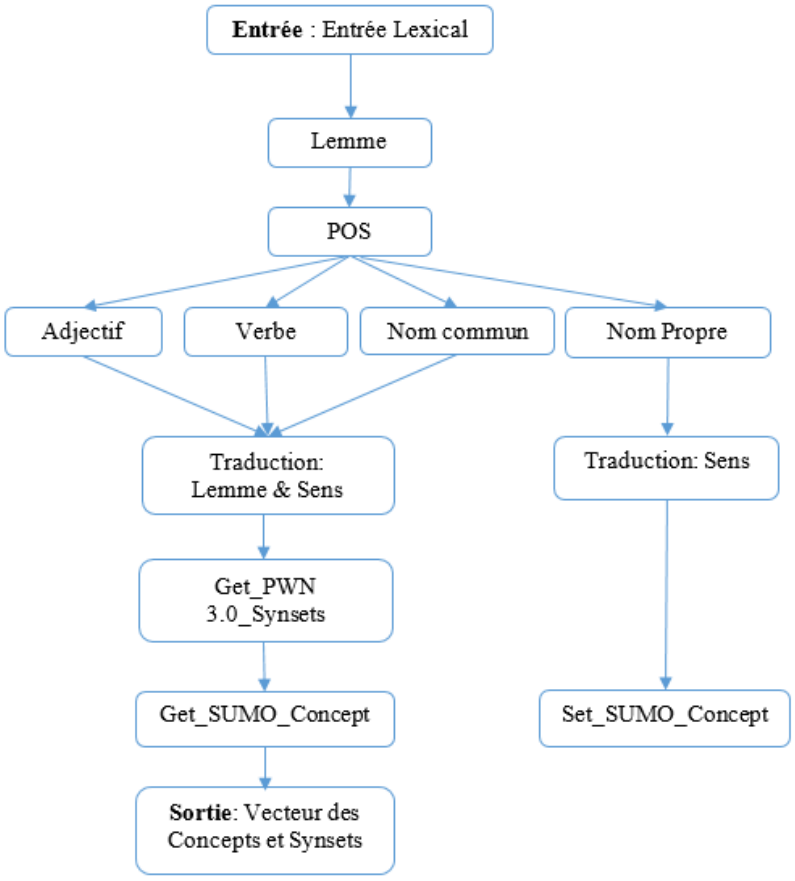


FIGURE 3.3 – Processus de Mapping.

3.4. Processus de traduction

L'algorithme 4 illustre les étapes suivies. Les ressources utilisées diffèrent par de nombreux détails, la première étape donc était de normaliser leurs entrées.

Les lemmes arabes sont présentés dans la version numérique TEI du dictionnaire bilingue en utilisant une translittération non standard. La première étape consiste donc à translittérer les mots candidats en utilisant le même système de translittération.

Ensuite, on cherche la forme translittérée dans le lexique arabe-anglais pour obtenir sa traduction anglaise. Cependant, le texte AQAM que nous avons, n'est pas entièrement vocalisé, par contre le lexique bilingue est entièrement vocalisé. Avant d'entamer le processus d'appariement/alignement, on définit pour chaque voyelle trouvée dans un mot candidat d'AQAM qu'elle doit être obligatoire. En fait, comme présenté dans l'algorithme 4, on affecte aux voyelles de la forme translittérée recherchée une priorité égale à 1. Après avoir eu l'expression régulière, nous récupérons donc toutes les traductions correspondantes.

En outre, Salmonè, l'auteur du dictionnaire bilingue utilisé (Salmoné (1978)), présente les traductions de verbes en mode passé simple. Afin de relier AQAM à l'anglais WordNet (PWN), nous devons avoir la forme verbale infinitive. Donc on utilise l'API JAWS [6] (Java API for WordNet Searching), qui permet d'obtenir la forme de base, basée sur la recherche du résultat obtenu dans PWN verbs. Par exemple, dans les exemples 24 et 25, le lemme *'abba* est traduit par *"prepared, got ready to, longed, yearned"*, après les avoir converti en leurs formes de base : *"prepare, get ready to, yearn, long"*.

3.5. Processus d'alignement / mapping

L'algorithme 5 illustre le processus de alignement automatique d'AQAM avec PWN et SUMO qui est résumé dans les étapes suivantes :

1. Recherchez chaque élément E de la liste de traduction anglaise du lemme L(X) dans PWN.
2. Récupérez la liste des Synsets $L_S(E)$ pour chaque élément E.
3. Pour chaque élément p d'un synset donné de $L_S(E)$, récupérez le concept SUMO correspondant.
4. Classez le résultat par concept dans List_Lemma_Concepts $L_{LC}(X)$.

6. https ://github.com/jaytaylor/jaws

Algorithme 4 : Traduction des lemmes et sens simples d' AQAM

```
input  : String qAmusLemma, List ArabicVowels, Lexicon ALAED
output : Translation : Vector of matched words Translation
qAmusLemmaTransliterated ← Transliteration(qAmusLemma);
/* Transliteration of أَبَّ 'abba = "A^ab~a"                                  */
Char [ ] splitedLemma ← splitToChar(qAmusLemmaTransliterated);
for c in splitedLemma do
    if c matches (ArabicVowels) then
        s ← s + c + "{1}?";
    else
        s ← s + c + "{1}?.{0,1}?";
    end
end
s ← clean(s) // delete additional .{0,1} ?
/* s= A{1}?^{1}?a{1}?b{1}?~{1}?a{1}?                                          */
for ALAEDLemma in ALAED do
    if s matches(ALAEDLemma) then
        addToTranslation(ALAEDLemma.getTranslation());
end
/* Translation={Prepared, got ready to, for; Yearned for, longed to see}     */
if qAmusLemma.getPOS() == "Verb" then
    for t in Translation do
        get_basic_form(t);
    end
/* Translation={Prepare, get ready to, for; Yearn, long}                     */
return Translation;
```

5. Pour chaque sens s, répétez de 1 à 4 et obtenez les concepts du sens de la liste $L_{SC}(s)$.

6. Comparez les éléments de $L_{LC}(X)$ et $L_{SC}(s)$.

   - Si $L_{LC}(X) = L_{SC}(s)$: la liste de concept du sens et la liste de concept du lemme sont égales, cela signifie que le sens et le lemme sont des Synonymes parfaits "Perfect Synonym".

   - $R = L_{LC}(X) \cap L_{SC}(s)$, Si $R \neq \varnothing$: la liste de concept du sens et la liste de concept du lemme partagent certains concepts, cela signifie que le sens et le lemme sont "synonymes".

Algorithme 5 : Mapping d'une entrée lexicale AQAM à PWN et SUMO

input : Lemma X, Sens S, Lemma_Translation L1, Sens_Translation L2

output : R : Vector of corresponding SUMO concepts and PWN synsets

for *e1 in L1* **do**
    $List_Lemma_Syn \leftarrow getSynsets(e1)$;
    for *p1 in List_Lemma_Syn* **do**
        $List_Lemma_Concepts \leftarrow getSUMOConcept(p1)$;
    end
    $LLC \leftarrow sortByConcept(List_Lemma_Concepts)$;
end
for *e2 in L2* **do**
    $List_Sens_Syn \leftarrow getSynsets(e2)$;
    for *p2 in List_Sens_Syn* **do**
        $List_Sens_Concepts \leftarrow getSUMOConcept(p2)$;
    end
    $LSC \leftarrow sortByConcept(List_Sens_Concepts)$;
end
// Compare LLC && LSC
$R \leftarrow LLC \cap LSC$;
if $R == LSC$ && $R == LLC$ **then**
    X and S are perfect Synonyms;
else if $R <> \varnothing$ **then**
    X and S are Synonym;
else
    $R \leftarrow LLC$;
end
return R;

3.6. Discussion et analyse de quelques exemples

3.6.1. Exemple du lemme verbal *'abba*

Figure. 3.4 présente un diagramme qui résume le flux en prenant comme exemple l'entrée AQAM d'origine indiquée dans l'exemple 25. Le résultat final contient un niveau

morphosyntaxique et lexical qui concerne l'extraction du lemme accompagné de sa catégorie grammaticale, enrichi de différentes formes dérivées associées, et se termine par un enrichissement de composants sémantiques concernant l'extraction des sens, les traductions et le mapping avec PWN et SUMO.

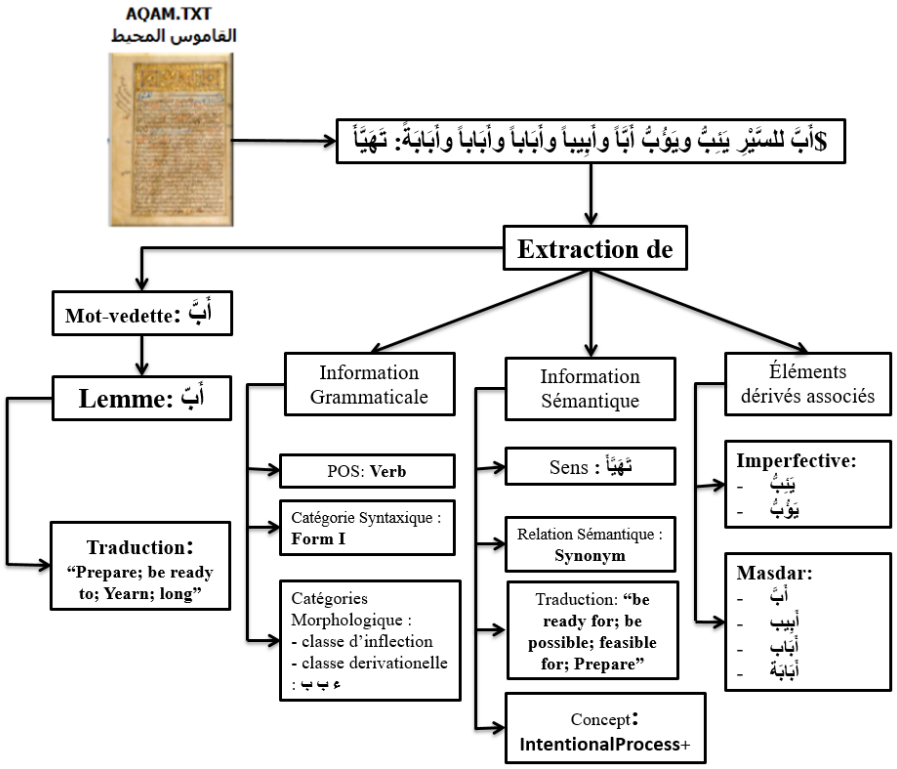


FIGURE 3.4 – Flux de travail

L'exemple XML 3.1 illustre l'entrée verbale indiquée dans l'exemple 24 concernant la conversion du lemme *'abba*. En plus des informations lexicales et dérivatives, le système reconnaît automatiquement :

1. Traduction du lemme *'abba* : *"Prepare, get ready to, long, yearn"*.
2. Traduction du sens *ištāqa* : *"desire, yearn, long"*.
3. On ne garde que les équivalents anglais qui sont communs entre les traductions du lemme et du sens : ***"long, yearn"***.
4. On recherche tous les synsets PWN3.0 correspondants pour chaque traduction équivalente en anglais commun :

   Le verbe *long* fait partie d'un synset :
   - {hanker, long, yearn} : "desire strongly or persistently".

Le verbe *yearn* fait partie de trois synsets :

- {yearn} : "have affection for ; feel tenderness for"

- {ache, languish, pine, yearn, yen} : "have a desire for something or someone who is not present"

- {hanker, long, yearn} : "desire strongly or persistently".

5. On ne conserve que ceux qui sont communs entre différents ensembles de synsets :

   - {hanker, long, yearn} : "desire strongly or persistently" qui a l'ID = 01828405.

6. On récupére le concept SUMO correspondant sur la base du Synset-ID obtenu : **IntetionalPsychologicalProcess**+.

   Le suffixe '+' indique que le synset ID = 01828405 {hanker, long, yearn } est subsumé par le concept SUMO ***IntetionalPsychologicalProcess***.

7. Il y a quelques concepts communs entre le lemme et son sens, donc la valeur de mapping "Synonym" est ajoutée automatiquement à la balise ExternalLinks.

XML Example 3.1 – Exemple du lemme verbal *'abba*

```
<section value="البَاء">
 <chapter value="الهَمزَة">
  <root_familly c1="الهَمزَة" c2="البَاء" c3="البَاء">
   <lexical_entry>
    <plain_text> اشْتَاقَ: أَبَّ إِلَى وَطَنِهِ أَبّاً وَ إِبَابَةً وَأَبَابَةٌ </plain_text>
    <lemma>أَبَّ</lemma>
    <pos>Verb</pos>
    <morphological_information>I</morphological_information>
    <masdar>أَبّ</masdar>
    <masdar>إِبَابَة</masdar>
    <masdar>أَبَابَة</masdar>
    <lemma_translate att="أَبَّ" value="prepare, get ready to; yearn; long"/>
    <sense att="اشْتَاقَ">
     <synonym att="اشْتَاقَ">
      <translate att="اشْتَاقَ" value="Desire, yearn, long" />
      <external_links att="mapping" value="Synonym" />
       <concept att="SUMO" value="IntentionalPsychologicalProcess+">
```

```
        <synset ref="PWN" ver="3.0" id="01828405" syn="hanker, long, yearn"
          Gloss="desire strongly or persistently" />
       </concept>
      </external_links>
     </synonym>
    </sense>
   </lexical_entry>
  </root_familly>
 </chapter>
</section>
```

3.6.2. Exemple du lemme *kitāb*

L'exemple du lemme كِتَاب *kitāb* montre une entrée lexicale compliquée qui nécessite une étude plus approfondie.

الكِتَابُ : مَا يُكتَبُ فيه، وَالدَّوَاةُ، وَالتَّورَاةُ، وَالصَّحِيفَةُ، وَالفَرضُ، وَالحُكمُ، وَالقَدَرُ

En utilisant le dictionnaire bilingue, la traduction du lemme *kitāb* prouve également qu'il est très polysémique :

```
<lemma_translate att="كِتَاب" value="writing; book, volume; letter; record,
register; contract; Sacred Writings, Scriptures: the Pentateuch,
Old Testament; the Gospel; the Kur'ān; Decree, ordinance, statute,
prescript; prescription; sentence; Receptacle for ink" />
```

Dans AQAM, on a sept sens du lemme *kitāb* qui prouvent que le lemme est très polysémique :

Sens 1

Il se compose d'une phrase entière "*mā yuktabu fih*" qui signifie *"[matériel] sur lequel on écrit"*. Dans ce cas, l'intervention manuelle de l'expert est nécessaire. Après

élimination de l'ambiguïté, la phrase explicative est mappée sur le PWN3.0 au synset *{book, volume}*, ID = 02870092, défini par la Gloss *"physical objects consisting of a number of pages bound together"* et subsumé par le concept SUMO ***Book***.

```
<sense att="مَا يُكتَبُ فيه">
 <translate lang="eng" value="null" />
 <external_links att="mapping" value="null">
  <concept ref="SUMO" value="Book+">
    <synset ref="PWN" ver="3.0" id="02870092" syn="book, volume"
          Gloss="physical objects consisting of a number of pages
bound together"
 />
  </concept>
 </external_links>
</sense>
```

Sens 2

Initialement, le système ne trouve aucun équivalent anglais commun entre le lemme et le second sens "*dawāt*" traduit par *"Inkbottle, inkstand"*. Cependant, après une étude détaillée, nous constatons que le lemme a été traduit, dans le dictionnaire bilingue, par *"Receptacle for ink"*, phrase qui peut expliquer *"Inkbottle, inkstand"* *"Bouteille d'encre, encrier"*. Ainsi, après révision manuelle, le deuxième sens est étiqueté comme "synonyme", lié au synset *{encrier, encrier}*, ID = 03572321 avec la Gloss= *"a small well holding writing ink into which a pen can be dipped"* et subsumé par le concept SUMO ***Device***.

```
<sense att="دَوَاة">
 <synonym att="دَوَاة">
  <translate lang="eng" value="Inkbottle, inkstand" />
  <external_links att="mapping" value="synonym">
   <concept ref="SUMO" value="Device+">
    <synset ref="PWN" ver="3.0" id="03572321" syn="inkstand, inkwell"
      Gloss="a small well holding writing ink into which a pen can
be dipped" />
   </concept>
```

```
  </external_links>
 </synonym>
</sense>
```

Sens 3

Le troisième sens "*tawrāt*" est traduit par *"Torah, Pentateuch, Old Testament"*. Dans PWN, le mot anglais *"Torah"* est mappé à trois synsets qui sont liés au même concept SUMO **Text**. Les deux synsets ID = 06451891 et ID = 06452223 sont liés en instance grâce au symbole @. Par contre, le synset ID = 06408442 est subsumé par le même concept avec le symbole +. Cela démontre que le PWN est très détaillé et possède une grille très fine, ce qui ne facilite pas les procédures du mapping.

```
<sense att="تَورَاة">
 <synonym att="تَورَاة">
  <translate lang="eng" value="Torah, Pentateuch; Old Testament" />
  <external_links att="mapping" value="synonym">
   <concept ref="SUMO" value="Text@">
    <synset ref="PWN" ver="3.0" id="06451891" syn="Laws, Pentateuch, Torah"
    Gloss="the firs of three divisions of the Hebrew, Scriptures comprising
the first five books of the Hebrew Bible considered as a unit" />
    <synset ref="PWN" ver="3.0" id="06452223" syn="Torah" Gloss="the whole
body of the Jewish sacred writings and tradition including the oral tradition"
   </concept>
   <concept ref="SUMO" value="Text+">
    <synset ref="PWN" ver="3.0" id="06408442" syn="Torah" Gloss="(Judaism)
the scroll of parchment on which the first five books of the Hebrew Scripture
is written; is used in a synagogue during services" />
   </concept>
  </external_links>
 </synonym>
</sense>
```

Sens 4

"*ṣaḥīfah*" est traduit par "*leaf, paper, page*" et chaque traduction est plus polysémique :

- *leaf* fait partie de trois synsets dont le synset {leaf, fogliage} ID = 06256229 avec la gloss *"a sheet of any written or printed material (especially in a manuscript or book)"* et subsumé au concept SUMO ***Text*** peut correspondre aux significations du lemme.

- *paper* fait partie de huit synsets dont seul le synset {paper} ID = 14974264 défini par la gloss *"a material made of cellulose pulp derived mainly from wood or rags or certain grasses "* peut correspondre aux significations du lemme, et équivaut au concept SUMO ***Papier*** via le symbole "=".
- *page* fait partie de cinq synsets dont seul le synset {page} avec ID = 06256697 défini par *"one side of one leaf"* correspond au lemme, et il est équivalent au concept SUMO ***Page*** par le symbole "=".

```
<sense att="صَحِيفَة">
 <synonym att="صَحِيفَة">
  <translate lang="eng" value="leaf, page, paper" />
  <external_links att="mapping" value="synonym">
   <concept ref="SUMO" value="Text+">
    <synset ref="PWN" ver="3.0" id="06256229" syn="folio, leaf" Gloss="a
sheet of any written or printed material (especially in a manuscript ord
book)" />
   </concept>
   <concept ref="SUMO" value="Page=">
    <synset ref="PWN" ver="3.0" id="06256697" syn="page" Gloss="one side
of one leaf (of a book or magazine or newspaper or letter etc.) or the
written or pictorial matter it contains" />
   </concept>
   <concept ref="SUMO" value="Paper=">
    <synset ref="PWN" ver="3.0" id="14974264" syn="paper" Gloss="a
material made of cellulose pulp derived mainly from wood or rags or
certain grasses" />
   </concept>
  </external_links>
 </synonym>
```

```
</sense>
```

Sens 5 et 6

Parmi les traductions des deux sens polysémiques *farḍ* et *ḥukm*, on trouve "*statute, law, ordinance*" qui sont également listés comme traduction du lemme *kitāb*. Par conséquent, le système les a liés comme synonymes au lemme. Cependant, la relation n'est pas vraiment exacte, car le mot *kitāb* est utilisé dans un sens figuré dans ces cas, et désigne ces sens par extension.

Pour expliquer ces deux significations, az=zabīdī, dans son lexique *tāǧ al=ʿarūs* (az=zabīdī (2001), Volume 4, page 101) donne les exemples qui se réfèrent uniquement à Allah. [7]

De plus, Wehr dans son lexique (Wehr (1976)), marque *farḍ* et *ḥukm* par l'abréviation "Isl. Law" (loi islamique : *fiqh*) c'est-à-dire qu'ils font partie de la terminologie islamique traditionnelle :

farḍ : religious duty (Isl. Law). (c-à-d, un devoir religieux)

ḥukm : legal consequence of the facts of a case (Isl. Law). (c-à-d, la conséquence juridique des faits d'une affaire)

Dans PWN et SUMO, aucun concept ne considère la terminologie de la loi islamique qui représente un exemple de non-équivalence ontologique. Par conséquent, les liens externes restent des lacunes.

```
<sense att="فَرض">
 <synonym att="فَرض">
  <translate lang="eng" value="religious duty" />
  <external_links att="mapping" value="null">
   <concept ref="SUMO" value="null" />
    <synset ref="PWN" ver="3.0" id="null" syn="null" Gloss="null" />
   </concept>
  </external_links>
 </synonym>
</sense>
<sense att="حُكم">
 <synonym att="حُكم">
```

7. Notez que les entrées lexicales dans *tāǧ al=ʿarūs* sont très enrichis dans lesquels az=zabīdī explique les sens très stricts d'AQAM.

```
  <translate lang="eng" value="legal consequence of the facts of a case" />
  <external_links att="mapping" value="null">
   <concept ref="SUMO" value="null" />
    <synset ref="PWN" ver="3.0" id="null" syn="null" Gloss="null" />
   </concept>
  </external_links>
 </synonym>
</sense>
```

Sens 7

Le sens *qadar* est traduit par *fate, destiny ; divine decree ; providence*. Cependant, az=zabīdī, dans son lexique *tāǧ al=ʿarūs* (az=zabīdī (2001)), précise que le sens *qadar* concerne le Qur'ān :

وَكِتَابُ اللّهِ : قَدَرُهُ

Le Livre d'Allah : [est] "son ***qadar***".

Dans ce contexte précis, nous lions le lemme *kitāb* dans PWN 3.0 au synset ID = 06461609 qui identifie le synset {Book, Koran, Quran, al-Qur'an} subsumé à Concept SUMO ***Livre***.

```
<sense att="قَدَر">
 <synonym att="قَدَر">
  <translate lang="eng" value="statute, law, ordinance" />
  <external_links att="mapping" value="null">
   <concept ref="SUMO" value="Book+" />
    <synset ref="PWN" ver="3.0" id="06461609" syn="Book, Koran, Quran,
al-Qur'an" Gloss="the sacred writings of Islam revealed by God to the
prophet Muhammad during his life at Mecca and Medina" />
   </concept>
  </external_links>
```

```
    </synonym>
  </sense>
```

3.6.3. Exemple du lemme *ṯawāb*

الثَّوَابُ : العَسَلُ، وَالنَّحلُ، وَالجَزَاءُ، كَالمَثُوبَةِ، وَالمَثوَبَةِ

Bien que le "Dictionary of Modern Written Arabic" se limite à ne donner que les significations utilisées en arabe moderne, il est intéressant d'en tenir compte car il met l'accent sur le domaine sémantique, comme, "Islamic Law" c-à-d, la loi islamique. Par exemple, le lemme **ṯawāb** :

'requital, recompense, reward (for good deeds) ; ***(Isl. Law)*** *merit, credit (arising from a pious deed)'.* [8]

Par contre, les traductions dans "An Advanced Learner's Arabic-English Dictionary" Salmoné (1978) sont basées sur les significations trouvées dans le lexique *tāğ al=ʻarūs* az=zabīdī (2001) qui est à son tour basé sur AQAM. D'où le lemme *ṯawāb* se traduit par :

(a) reward, recompense. (b) Price. (c) Honey. [9]

On note que les significations sont classées selon leurs fréquences et les traductions anglaises sont accompagnées du symbole (a), (b) ou (c) (etc.) qui représentent la première, deuxième, troisième, (etc.) significations du lemme. Par exemple, le sens *ʻasal* "miel" qui n'est plus utilisé en arabe moderne est classé comme le troisième sens (c).

Sens 1 et 2

Des études approfondies d'AQAM nous permettent de noter qu'il existe des définitions récurrentes, par exemple, les significations *ʻasal "honey"* "miel" et naḥl "bee" "abeille" réapparaît respectivement 42 et 31 fois. Pour des raisons d'optimisation, nous avons identifié ces "mots clés" qui peuvent accélérer le mapping et relier automatiquement le lemme au concept correspondant. Les "Mots clés" identifiés dans la

8. Wehr (1976), page 108.
9. Salmoné (1978), page 81.

section *bā'* sont listés dans le tableau 5 dans l'annexe. Ainsi, les deux premiers sens *'asal* "miel" et *naḥl* "abeille" sont automatiquement liés à leurs concepts PWN et SUMO correspondants, respectivement à ***Honey***= et ***bee***=. De plus, le système les sépare en différents fichiers, respectivement File_Food et File_Animal.

```
<sense att="عَسَل">
 <synonym att="عَسَل">
  <translate lang="eng" value="honey" />
  <external_links att="mapping" value="synonym">
   <concept att="SUMO" value="Honey=">
    <synset ref="PWN" ver="3.0" id="15305917" syn="honey"
        Gloss="An edible, sweet substance produced by some species
of Bees" />
   </concept>
  </external_links>
 </synonym>
</sense>
<sense att="نَحل">
 <synonym att="نَحل">
  <translate lang="eng" value="bee" />
  <external_links att="mapping" value="synonym">
   <concept att="SUMO" value="Bee=">
    <synset ref="PWN" ver="3.0" id="15330350" syn="bee"
        Gloss="A hairy insect, some species of which produce honey
and/or sting" />
   </concept>
  </external_links>
 </synonym>
</sense>
```

Sens 3

Le troisième sens *ǧazā'* est traduit par "*repayment ; remuneration, recompense, requital, compensation*". Ayant des traductions communes avec le lemme, il est automatiquement considéré par le système comme synonyme du lemme **ṯawāb**. Cependant, ce dernier est spécifiquement lié au domaine de la loi islamique et ses sens

"*repayment ; remuneration, recompense, requital, compensation*" sont liés au bonnes ou mauvaises actions *"good or bad deeds"* qui ne sont pas prévues dans PWN et SUMO. Ainsi, comme cité par Nahli (2018a)), relier ces concepts qui ont une connotation religieuse à PWN et SUMO manquerait leur identité musulmane.

```
<sense att="جَزَاء">
 <synonym att="جَزَاء">
  <translate lang="eng" value="repayment; remuneration, recompense,
requital, compensation" />
  <external_links att="mapping" value="synonym">
   <concept ref="SUMO" value="null" />
    <synset ref="PWN" ver="3.0" id="null" syn="null" Gloss="null" />
   </concept>
  </external_links>
 </synonym>
</sense>
```

Sens 4 et 5

Les significations des deux derniers sens sont des cas spécifiques, cités selon le style de l'auteur. Introduits par la conjonction *ka* "comme", ils sont dérivés de la même racine du lemme et héritent du dernier sens qui précède. Après avoir vérifié que مَثُوبَة *maṯūbah* et مَثْوَبَة *maṯwabah* contiennent les consonnes radicales du lemme, ils héritent de la traduction de *ǧazā'*.

```
<sense att="مَثُوبَة">
 <synonym att="مَثُوبَة">
  <translate lang="eng" value="repayment; remuneration, recompense,
requital, compensation" />
  <external_links att="mapping" value="synonym">
  <concept ref="SUMO" value="null" >
```

```
      <synset ref="PWN" ver="3.0" id="null" syn="null" Gloss="null" />
     </concept>
    </external_links>
   </synonym>
  </sense>
  <sense att="مَثْوَبَة">
   <synonym att="مَثْوَبَة">
    <translate lang="eng" value="repayment; remuneration, recompense,
requital, compensation" />
    <external_links att="mapping" value="synonym">
     <concept ref="SUMO" value="null" />
      <synset ref="PWN" ver="3.0" id="null" syn="null" Gloss="null" />
     </concept>
    </external_links>
   </synonym>
  </sense>
```

3.7. Analyse des résultats statistiques

Les exemples présentés dans la section précédente montrent qu'il existe des cas complexes qui nécessitent une étude majeure et une révision manuelle. Le processus d'examen est donc compliqué et prend du temps. Comme point de départ, cette partie de notre travail a été expérimentée sur la section de *bā'*. Les tableaux suivants et les analyses statistiques qui en résultent donnent une idée générale de l'organisation d'AQAM et de la complexité du travail.

D'après le Tableau 3.1, le chapitre *bā'* est composé de 1244 lemmes verbaux, 2455 lemmes nominaux et 216 adjectifs. Les entrées lexicales peuvent être sans définition généralement extraites d'entrées imbriquées, comme expliqué dans les exemples 15 et 16. D'autres entrées lexicales peuvent avoir un lemme monosémique avec une seule définition ; ou lemme polysémique ayant plusieurs définitions. Notez que les lemmes polysémiques (en particulier les lemmes des noms) sont plus nombreux que ceux monosémiques. La table 3.2 illustre les types de définitions qui peuvent constituer une phrase explicative ou un mot simple et on note que :

— plus de la moitié des lemmes sont définis par des phrases explicatives qui ne peuvent pas être traitées automatiquement et qui nécessitent une révision et une étude manuels.

— les autres éléments sont définis par un simple mot qui est considéré comme des synonymes du lemme et peut être automatiquement traduit et mappé.

TABLE 3.1 – Statistiques des types de lemmes d'AQAM

Classe	**Entrées Lexicales**			
	Total	**Sans Définition**	**monosémique**	**polysémique**
Verbes	1244	122	386	276 (518 sens)
Noms	2455	-	515	774 (1651 sens)
Adjs	216	8	71	21 (102 sens)

TABLE 3.2 – Statistiques des types de définition d'AQAM

Classe	**Définition**	
	Mot Simple	**Phrase explicative**
Verbes	904	748
Noms	2166	2918
Adjs	102	152

La tableau 3.3 illustre le taux de traduction des lemmes verbaux, nominaux et adjectifs et des définitions simples. Notez que la couverture des dictionnaires bilingues n'est pas totale et qu'en général plus de la moitié des mots candidats (lemmes et définitions simples) ne sont pas traduits. La vérification de la traduction automatique donne de faibles résultats car elle est basée sur la comparaison entre la traduction de lemmes et la traduction de définitions uniques seulement.

TABLE 3.3 – Statistiques de la traduction des lemmes et des sens de l'AQAM

Classe	**Lemme**		**Sens**		**Synonyme**
	Lemme	**Lemme Traduit**	**Mot Simple**	**Sens Traduit**	
Verbes	1244	941	904	477	109
Noms	2455	1171	2166	1100	200
Adjs	216	59	102	32	3

Tableau 3.4 illustre les résultats du contrôle manuel et le pourcentage d'éléments sans correspondance sémantique dans SUMO et PWN. De nombreux cas de non-correspondance peuvent être expliqués car le sens ou la traduction est une phrase explicative qui empêche la mise en correspondance. D'autres cas concernent l'organisation différente des lexiques arabe et anglais Nahli (2018a).

TABLE 3.4 – Résultat contrôlé manuellement

POS	**Entrée Contrôlée**	**non correspondance**	
		Total	**Taux**
Verbes	511	204	40%
Noms	479	41	8.50%
Adjs	200	56	29.50%

Par exemple, les statistiques de mapping de verbes illustrées dans le tableau 3.5 montrent qu'un nombre important d'entrées verbales ne sont pas mappées. Concernant les verbes trilitères, nous avons obtenu un nombre important d'entrées sans mapping, en raison de la quantité importante de traductions qui suivaient la forme "*to be + participe*" ou "*be + adjective*". Alors qu'ils sont exprimés avec un verbe d'état en arabe, ils ne sont pas présents dans le lexique anglais. Par exemple, le verbe "*karuma*" qui signifie "*to be noble, magnanimous, generous*" ("*être noble, magnanime, généreux*") n'est pas mappable dans PWN et SUMO. Les verbes dérivés manquent également de mapping, car ils expriment des concepts dérivatifs arabes qui sont exprimés en anglais par une phrase entière. Par exemple, "*ta'assaba*" qui signifie "*to wind the turban round one's head*" "*enrouler le turban autour de la tête*".

TABLE 3.5 – Statistiques des entrées lexicales verbales d'AQAM mappées à SUMO

Forme	**Lemme**	**Lemme Traduit**	**Sens**	**Sens Traduit**	**Mappé (%)**
I	624	551	540	281	33%
II	109	92	46	25	8%
III	36	29	26	10	5%
IV	192	138	101	44	8%
V	77	61	60	43	5%
VI	17	10	10	5	-
VII	27	17	12	8	4%
VIII	19	13	15	8	1%
IX	1	-	-	-	-
X	36	20	26	7	1%
Q_I	71	37	68	52	2%
Q_II	17	6	13	7	2%
Q_III	-	-	-	-	-
Q_IV	16	-	12	7	-

Concernant les noms et les adjectifs, le mapping est plus important que pour les verbes. Dans ce cas également, certains cas de non-correspondance sont dus à des différences conceptuelles et culturelles, comme déjà expliqué dans les exemples précédents (pour plus de détails, voir Nahli (2018a)). D'autres problèmes peuvent être identifiés tels que :

— Problèmes stylistiques : dans certains cas, le lemme n'est pas traduit ou la définition se compose d'une phrase entière qui doit être analysée et interprétée pour être traduite.

— Problèmes de culture : La vie nomade dans le désert attache une importance particulière aux chameaux et aux troupeaux. Par exemple, le sujet du verbe *aḥlab-a* est «le propriétaire de chameaux qui a donné naissance à des femelles». Il faut se souvenir que le verbe *'aḥlab-a* dérive du verbe حَلَبَ *ḥalab-a* dont la racine est *ḥlb*, dont dérive le mot حَليب *ḥalīb* «lait». Le verbe أَحْلَبَ *aḥlab-a* peut signifier «traire beaucoup de lait».

3.8. Conclusion

L'objectif principal du travail présenté dans ce chapitre est d'enrichir le lexique numérique obtenu avec les traductions anglaises et les informations sémantiques et ontologiques en liant AQAM avec des ressources anglaises préexistantes (PWN et SUMO). L'alignement a été effectué grâce au mapping automatique des lemmes arabes sur l'anglais, à travers leurs traductions obtenues à partir des deux dictionnaires Arabe- Anglais (Dictionary of Modern Written Arabic et Advanced Learner's Arabic-English Dictionary). Le but de ce travail n'est pas de défendre le relativisme linguistique. Pourtant il nous a permis de mettre en évidence certains aspects communs entre la langue arabe et la langue anglaise et il nous a permis de découvrir un certain nombre de correspondances manquantes entre les deux langues, en raison de l'organisation conceptuelle différente des lexiques des deux langues. Par conséquent, le mapping de la langue arabe sur une ontologie conçue pour une autre langue, comme l'anglais, peut soulever des problèmes fondamentaux.

L'enrichissement de la ressource digitale de AQAM avec les traductions anglaises des lemmes et des sens a été effectuée sur toutes les sections de AQAM. Comme résultats, on a obtenu un deuxièmes type de fichiers XML où les lemmes verbes, noms ou adjectifs sont accompagnés de leurs traductions anglaises (quand c'est possible). Par contre la phase pilote de mapping entre AQAM et PWN et SUMO a été effectuée sur la section

de la lettre *bā'* en produisant un fichier XML de la lettre *bā'* enrichi également avec les concepts correspondants. Tous les fichiers XML enrichis sont déposés dans l'infrastructure de *CLARIN-IT*.

Chapitre 4

Traitement et extraction des noms propres cités dans *al=qāmūs al=muḥīṭ*

4.1. Introduction

La classe des noms propres est définie par Van Langendonck comme suit :

"*A proper name is a noun that denotes a unique entity at the level of 'established linguistic convention' to make it psychosocially salient within a given basic level category [pragmatic]. The meaning of the name, if any, does not (or not any longer) determine its denotation [semantics]. [...] Proper names do not have asserted lexical meaning but do display presuppositional meanings of several kinds : categorical (basic level), associative senses (introduced either via the name bearer or via the name form), emotive senses and grammatical meanings*". (van Langendonck (2007), page 06)

"*Un nom propre est un nom qui désigne une entité unique au niveau de la «convention linguistique établie» pour la rendre psychosocialement saillante dans une catégorie donnée de niveau de base [pragmatique]. La signification du nom, le cas échéant, ne détermine pas (ou plus) sa dénotation [sémantique]. [...] Les noms propres n'ont pas de significations lexicales affirmées mais affichent des significations présupposées de plusieurs types : catégoriques (niveau de base), sens associatifs (introduits soit via le porteur du nom, soit via la forme du nom), sens émotifs et significations grammaticales*".

4.2. Défi de la reconnaissance des noms propres

L'ambiguïté des noms propres dans le texte réside à deux niveaux différents. La première ambiguïté concerne le niveau de la détection, c-à-d différencier entre les mots qui sont des noms propres et ceux qui ne le sont pas. La deuxième est séquentielle de la première, car la phase d'identification est pré-requise pour la classification, puisqu'on ne peut pas classifier un nom propre sans le détecter. Le niveau d'ambiguïté dans les deux niveaux diffère d'une langue à une autre, ainsi que d'un domaine à un autre.

Rappelons que la langue Arabe est une langue flexionnelle, avec une morphologie riche, ainsi qu'une syntaxe complexe, ce qui complique les différentes phases du traitement automatique de cette langue. La reconnaissance des noms propres est considéré comme l'un des blocs de construction d'outils et applications de traitement automatique de la langue Arabe. Malgré le progrès significatif des recherches au cours de la dernière décennie, cette tâche reste difficile en raison des caractéristiques qui concernent le traitement de la langue Arabe et qu'on essayera de résumer dans les sous-sections suivantes.

4.2.1. Manque de capitalisation

Pour les langues utilisant le script latin, la majorité des noms propres commence par une Majuscule. En fait, la majuscule aide à identifier les noms propres potentiels dans la plupart des langues écrites en script latin. L'orthographe arabe ne comporte pas de lettres majuscules pour distinguer les lettres initiales des noms propres. En fait, l'absence de cette caractéristique pour la langue Arabe augmente l'ambiguïté de distinction entre les noms propres et certains mots qui sont des noms communs ou bien des mots descriptifs qui ne sont pas des noms propres. Par exemple, le nom propre arabe (وَفَاء *wafā'*) servira de signification différente dans une phrase en fonction de son contexte. Par conséquent, une méthodologie reposant sur les dictionnaires de noms pour traiter cette problématique, ou bien la traduction anglaise de mots arabes peut être exploitée comme indicateur de fonctionnalité à cet égard.

4.2.2. Manque des ressources

Les ressources en langue Arabe ne sont pas adéquates et les ressources disponibles ont une couverture limitée (Abouenour et al. (2013)). En général, le nombre de ressources linguistiques arabes disponibles gratuitement à des fins de recherche est limité. En particulier, un bon nombre des corpus disponibles ne sont pas annotés avec des entités nommées et n'ont pas un nombre suffisant d'entité nommée, ce qui les rend inaptes à l'entraînement et / ou à l'évaluation des systèmes de reconnaissance des entités nommées arabes (Shaalan (2014)). Par conséquent, les chercheurs doivent généralement déployer des efforts tangibles pour acquérir/annoter et vérifier leurs propres ressources linguistiques arabes afin de développer correctement les systèmes de reconnaissance des entités nommées arabes (Benajiba et al. (2007))

4.3. Les approches de reconnaissance des noms propres

Dans le domaine de reconnaissance des noms propres, on trouve trois approches : (i) l'approche heuristique des systèmes à base de règles, elle se base sur l'utilisation de modèles et de jeux de règles faites à la main, pour reconnaître les noms propres, elle atteint presque le même niveau de performance des annotateurs humains. (ii) l'approche de l'apprentissage automatique supervisé qui nécessitait de grandes données d'entraînement annotées pour former les classificateurs. (iii) l'approche non supervisée qui utilise de grands corpus pour dériver des éléments de réseau sans avoir besoin de règles codées à la main supplémentaires et / ou d'annotations supplémentaires (Nezda et al. (2006)).

4.3.1. L'approche à base de règle

Dans les systèmes à base de règles, les techniques linguistiques dépendent des règles et patterns de la langue Arabe, écrites pour reconnaître et extraire les noms propres. La performance de ces systèmes dépend de la qualité de ces règles dans l'identification des différents types des noms propres. Cette approche compense le manque de ressources linguistiques annotées en langue Arabe, et elle est favorisée par les résultats encourageants obtenus par divers systèmes basés sur des règles arabes. Certaines règles sont :

1. Règles grammaticales : ensemble des grammaires de langage utilisées pour reconnaître les noms propres. Normalement, ils sont les règles pour former des phrases bien

structurées ; par conséquent, l'application de règles de grammaire dans ce contexte nécessite la maîtrise de la langue en cours de traitement.

2. Les règles heuristiques complexes sont des règles plus générales qui dépendent principalement du type de l'approche globale basée sur des règles utilisée sur le système. Par exemple, elles peuvent être utilisées pour améliorer encore la capacité des mots déclencheurs appliqués dans le système, tels que les mots-clés (Elsebai et al. (2009)).

3. Les règles morphologiques sont générées sur la base du fait que la plupart des mots arabes dans une phrase sont construits normalement de la racine du mot et un autre suffixe et / ou préfixe. Tandis que les règles grammaticales agissent sur la structure de la phrase afin d'identifier les noms propres, les règles morphologiques traitent la structure des mots en examinant les stems, les affixes, et les suffixes. Quelques travaux utilisent BAMA (Buckwalter Arabic Morphological Analyser) (Buckwalter (2004)) dans l'identification. BAMA exploite principalement diverses tables telles qu'une table pour les stems, les préfixes et les suffixes arabes. Par exemple : بِفَاس *bi-fās* (à fès) ; la règle morphologique considère "ب" comme un préfixe, et le reste est la racine (فَاس / *fās*), qui est un emplacement.

L'approche basée sur des règles a été utilisée avec succès dans le développement de nombreux systèmes de traitement du langage naturel. Les systèmes qui utilisent des transformations basées sur des règles sont basés sur un noyau de connaissances linguistiques solides. L'avantage de l'approche fondée sur les règles par rapport à l'approche par corpus est évident pour : les langues moins dotées en ressources, pour lesquelles les grands corpus, éventuellement parallèles ou bilingues, avec des structures et entités représentatives ne sont ni disponibles ni facilement accessibles ; ainsi pour les langues morphologiquement riches, même avec la disponibilité des corpus, souffrent de la dispersion des données (Shaalan (2010)).

4.3.2. L'approche statistique

Les algorithmes d'apprentissage automatique (ML, Machine Learning) ont été largement utilisés pour déterminer les noms propres à partir des textes annotés, qui sont utilisés pour générer des modèles statistiques de prédiction des noms propres. L'approche la plus utilisée est basée sur l'apprentissage automatique est l'approche supervisée, qui traite le problème de reconnaissance des noms propres comme un problème de classification, et nécessite la disponibilité d'ensembles de données annotés importants et suffisants. Parmi les algorithmes d'apprentissage supervisé couramment utilisés figurent les Machines à vecteurs support (SVM, Support Vector Machine), les Conditional random fields (CRFs), l'entropie maximale (ME, Maximum Entropy), les modèles de Markov cachés (HMM, Hidden Markov Model) et les arbres de décision.

4.3.3. L'approche Hybride

L'approche hybride est l'intégration entre l'approche basée sur les règles et l'approche basée sur ML. L'intégration des composantes à base de règles et celles basées sur le ML produit un système hybride, qui vise à améliorer la performance globale de la reconnaissance des noms propres pour l'arabe. Selon (Oudah and Shaalan (2012) ; Shaalan and Oudah (2014)) les composantes basées sur des règles et celles basées sur le ML sont indépendantes et peuvent fonctionner séparément. Le traitement peut commencer du système basé sur des règles au système basé sur ML ou vice versa.

4.4. Extraction automatique des noms propres à partir de *al=qāmūs al=muḥīṭ*

Dans le contexte actuel, les dictionnaires modernes ne contiennent pas les noms propres, par contre les dictionnaires médiévales en sont riches. C'est l'une des caractéristiques spécifiques distinguant AQAM, d'où vient la nécessité de les extraire et les classer. Considérant l'entrée lexicale de l'exemple 27 composée par le mot-vedette *al=ǧanābu* et les sens qui l'accompagnent consistant en mots isolés séparés les uns des autres par des virgules.

Exemple 27.

الجَنَابُ : الفِنَاءُ، وَالرَّحْلُ، وَالنَاحِيَةُ، وَجَبَلٌ ، وَعَلَمٌ

*al=ğanāb-u : **le** patio, et **le** bagage, et **la** direction., et **une** montagne, et **un** nom propre d'une personne".*

Les trois premières définitions sont constituées de mots définis, et on peut interpréter l'entrée lexicale :

le *ğanābu* : [c'est] **le** patio, et [c'est] **le** bagage, et [c'est] **la** direction. [1]

Dans ce cas, les sens correspondent à des significations communes. Après des procédures de stemming légèr, chaque sens est considéré nom commun synonyme du lemme, comme indiqué dans l'exemple XML 4.1.

XML Example 4.1 – Conversion XML de *ğanāb*

```
<root_family c1="الجيم" c2="النُّون" c3="البَاء">
 <lexical_entry>
  <plain_text> الجَنَابُ: الفِنَاءُ، وَالرَّحلُ، وَالنَاحِيَةُ ، وَجَبَلٌ، وَعَلَمٌ</plain_text>
  <lemma>جَنَاب</lemma>
  <pos>Noun</pos>
  <sense att="فِنَاء">
   <synonym att="فِنَاء" />
  </sense>
  <sense att="رَحل">
   <synonym att="رَحل" />
  </sense>
  <sense att="نَاحِيَة">
   <synonym att="نَاحِيَة" />
  </sense>
  <root>ج ن ب</root>
 </lexical_entry>
</root_family>
```

Par contre, les deux derniers sens ne sont pas définis et se terminent par le suffixe nominatif (*un*). La définition, *wa=ğabal-un*, ne signifie pas *"et [c'est] une montagne"*, mais elle implique "et [c'est le nom propre d'] une montagne". La dernière définition,

1. Les mots entre crochets n'existent pas dans le texte arabe. Ils sont utilisés pour rendre le texte traduit plus compréhensible.

wa=ʿalam-un, est plus explicite et implique "et [c'est] le nom propre d'une personne". Dans ce cas, les procédures du stemming léger consistent à éliminer uniquement les particules possibles (conjonctions), le suffixe nominatif, par contre le lemme doit garder l'article défini : *al=ǧanāb*. En plus, et comme illustré dans l'exemple XML 4.2 ci-dessous, le lemme est accompagné de l'étiquette POS qui l'identifie comme "Nom Propre".

XML Example 4.2 – Conversion XML de *al=ǧanāb*

```
<root_family c1="الجيم" c2="النُّون" c3="البَاء">
 <lexical_entry>
  <plain_text> اَلجَنَابُ: الفِنَاءُ، وَالرَّحلُ، وَالنَاحِيَةُ، وَجَبَلٌ، وَعَلَمٌ</plain_text>
  <lemma>الجَنَاب</lemma>
  <pos>Proper_Noun</pos>
  <sense att="جَبَلٌ" />
  <sense att="عَلَمٌ" />
  <root>ج ن ب</root>
 </lexical_entry>
</root_family>
```

En résumé, un sens indique un nom propre lorsqu'il n'est pas défini et se termine par le suffixe nominatif *tanwīn*, *un*. Dans ce cas, les procédures du stemming légèr consistent à éliminer uniquement les particules possibles (conjonctions), le suffixe nominatif et le lemme doit garder l'article défini. Le tableau 4.1 présente le nombre des noms propres extrait de chaque section de AQAM.

4.5. Marquage sémantique des noms propres

La signification catégorique des noms propres qui appartient aux concepts de niveau de base est la seule signification lexicale que les noms propres semblent avoir au niveau de la convention linguistique établie (van Langendonck (2007), page 86). Ainsi, dans cette première phase de mapping, nous avons défini une liste de mots clés / déclencheurs utilisée par l'auteur d'AQAM permettant de catégoriser les noms propres. Ensuite, nous avons établi pour chaque mot déclencheur le concept de niveau de base correspondant qui existe dans SUMO. Le tableau de l'annexe 5 contient la liste des déclencheurs de noms propres trouvés dans la section *bā'*. Chaque nom propre catégorisé correspond à *une instance du*

TABLE 4.1 – statistiques des noms propres extrait à partir de AQAM

Section (bāb)	Noms Propre
hamza	93
bā'	1054
tā'	182
ṯā'	164
ǧim	336
ḥā'	346
ḫā'	116
dāl	763
ḏāl	108
rā'	1735
zāy	264
sīn	508
šīn	189
ṣād	128
ḍād	118
ṭā'	255
ẓā'	16
'ayn	494
ġayn	70
fā'	469
qāf	589
kāf	243
lām	1096
mīm	1068
nūn	874
hā'	59
wāw	373
yā'	383
Total	
28	**12095**

concept de niveau de base SUMO, par ex. homme, femme, pays, ville, rivière, etc. Ensuite, ils sont classés dans différents fichiers en fonction de la classe sémantique à laquelle ils appartiennent, par exemple : Astrologie, Bâtiment, Groupe, Localisation géographique, Nom propre humain, Eau, etc.

Dans l'exemple suivant, le mot clé *ǧabal-un* - qui signifie *une montagne* - est un déclencheur de nom propre utilisé par l'auteur d'AQAM pour désigner le nom propre de la montagne. D'où, le lemme *ǧal'ib* est mappé avec le symbole '@' comme une instance de SUMO Concept ***Mountain@***.

```
<root_family c1="الجِيم" c2="الَّام" c3="العَين" c4="البَاء">
 <lexical_entry>
  <plain_text> جَبَلٌ بِالمَدِينَةِ: جَلعِبٌ</plain_text>
  <lemma>جَلعِب</lemma>
  <pos>Proper_Noun</pos>
  <concept att="SUMO" value="Mountain@" />
 </lexical_entry>
</root_family>
```

Il convient de noter que la plupart des noms propres extraits ont été correctement mappés comme indiqué dans le tableau 4.2. En analysant les résultats obtenus, notre système est capable de reconnaître tous les noms propres des classes d'emplacement géographique, de groupe, de bâtiment et d'astrologie. Par contre, concernant les classes Cheval et Eau, nous avons obtenu 83% pour la Précision, car les motifs de surface sont ambigus et une révision manuelle est nécessaire.

TABLE 4.2 – Statistiques des noms propres extraits

Classe	**Précision**	**Rappel**	**F-Mesure**
Emplacement géographique	100%	100%	100%
Groupe	100%	100%	100%
Bâtiment	100%	100%	100%
Astrologie	100%	100%	100%
Nom Propre	99%	99%	99%
Eau	83%	100%	90%
Cheval	83%	100%	90%

Dans l'exemple XML 4.3, le mot-vedette *ḥallāb* est "*[un nom propre de]* ***un cheval*** *[de la tribu] banī taglab.*" Ainsi, le mot déclencheur *faras-un* "un cheval" indique que le

mot-vedette est un Nom propre, et il est automatiquement lié comme instance du concept de niveau de base SUMO ***Horse***.

XML Example 4.3 – Exemple de **faras-un** comme Nom Propre.

```
<root_family c1="الحَاء" c2="اللَّام" c3="البَاء">
 <lexical_entry>
  <plain_text> .فَرَسٌ لِبَنِي ثَعْلَبَ :حَلَّابٌ</plain_text>
  <lemma>حَلَّابٌ</lemma>
  <pos>Proper_Noun</pos>
  <sense att=".فَرَسٌ لِبَنِي ثَعْلَبَ">
   <concept att="SUMO" value="Horse@" />
  </sense>
  <root>ح ل ب</root>
 </lexical_entry>
</root_family>
```

Dans d'autres cas, comme dans l'exemple XML 4.4, le système classe le lemme *Al=ğānib* comme instance du concept de cheval mais le motif de surface *faras-un* "cheval" fait partie d'une phrase explicative de bon sens, "*[est] un cheval large entre les jambes*".

XML Example 4.4 – Exemple de **faras-un** comme partie d'une phrase explicative.

```
<root_family c1="الجِيْم" c2="النُّون" c3="البَاء">
 <lexical_entry>
  <plain_text> الجَانِبُ :المُجْتَنَبُ، المَحْقُورُ، وَفَرَسٌ بَعِيدُ مَا بَيْنَ الرِّجْلَيْنِ</plain_text>
  <lemma>جَانِب</lemma>
  <pos>ADJ</pos>
  <sense att="مُجْتَنَب" />
  <sense att="مَحْقُور" />
  <sense att="فَرَسٌ بَعِيدُ مَا بَيْنَ الرِّجْلَيْنِ" />
  <root>ج ن ب</root>
 </lexical_entry>
</root_family>
```

Un contrôle manuel est également nécessaire lors de la confrontation avec les exemples XML 4.5 où le motif de surface *mā'-un*" une eau "indique un nom propre de" *[une] eau*

(c'est-à-dire) de banī al=‘anbari" et 4.6 où le même motif de surface fait partie de la phrase explicative "*ce que vous voyez [à] midi comme [de] l'eau*".

XML Example 4.5 – Exemple de l'**eau** comme Nom Propre

```
<root_family c1="الهَمْزَه" c2="الزَّاي" c3="البَاء">
 <lexical_entry>
  <plain_text> إِزَابٌ، بِالكسر :مَاءٌ لِبَنِي العَنبَرِ</plain_text>
  <lemma>إِزَاب</lemma>
  <pos>Proper_Noun</pos>
  <sense att="مَاءٌ لِبَنِي العَنبَرِ">
   <concept att="SUMO" value="Water@" />
  </sense>
  <root>ء ز ب</root>
 </lexical_entry>
</root_family>
```

XML Example 4.6 – Exemple de l'**eau** comme partie d'une phrase explicative

```
<root_family c1="السِّين" c2="الرَّاء" c3="البَاء">
 <lexical_entry>
  <plain_text> اَلسَّرَابُ :مَا تَرَاهُ نِصفَ النَّهَارِ كَأَنَّهُ مَاءٌ</plain_text>
  <lemma>سَرَاب</lemma>
  <pos>Noun</pos>
  <sense att="مَا تَرَاهُ نِصفَ النَّهَارِ كَأَنَّهُ مَاءٌ" />
  <root>س ر ب</root>
 </lexical_entry>
</root_family>
```

4.6. Conclusion

Chaque approche de reconnaissance des noms propres a ses avantages et ses faiblesses. Les systèmes à base de règles reposent sur des règles grammaticales acquises auprès des linguistes. Par conséquent, toute maintenance appliquée aux systèmes basés sur des règles demande beaucoup de travail et de temps, en particulier si les linguistes possédant les

connaissances et les antécédents requis ne sont pas disponibles. Par contre, les systèmes basés sur ML utilisent des algorithmes d'apprentissage qui utilisent l'ensemble sélectionné de caractéristiques extraites à partir des ensembles de données annotées avec des entités nommées pour la construction de classifieurs prédictifs. Le principal avantage des systèmes basés sur ML, qu'ils sont adaptables et modifiables avec un minimum de temps et d'efforts tant que les datasets suffisamment importants sont disponibles, et ce qui n'est pas le cas pour l'Arabe classique. Donc nous avons choisi d'utiliser l'approche à base de règles linguistiques car elle donne des résultats similaires à ceux des annotateurs humains, aussi pour contribuer à remédier la rareté des corpus annotés pour l'arabe classique.

Comme résultats de cette dernière phase du travail, le dossier sans traduction de chaque section de AQAM contient également les fichiers des noms propres divisés en fonctions des concepts auxquels ils ont été liés (noms propres de personnes, emplacement géographique, groupe, bâtiment, astrologie, cheval et eau). Ces fichiers sont également disponibiles dans *CLARIN-IT* infrastructure [2].

2. http ://hdl.handle.net/20.500.11752/ILC-97 (18/12/2019)

Conclusion et Perspectives

Dans cette thèse, on a décrit les étapes et la méthodologie suivies pour créer une nouvelle ressource lexicale pour l'arabe classique basée sur la conversion et l'enrichissement du dictionnaire arabe classique AQAM dans un format lisible par machine. Le processus de conversion est basé sur le codage, la segmentation, l'extraction et représente autant que possible des informations lexicales explicites ensuite, la classification des entrées AQAM suivant leurs catégories grammaticales.

On a également décrit notre approche pour lier le lexique arabe mentionné à des références externes, telles que PWN et SUMO, à travers un lexique arabe-anglais. Cette partie de notre travail a été expérimentée sur la section de la consonne *bā'* (البَاء). L'alignement d'ontologie inter-linguistique obtenue entre AQAM, PWN et SUMO prouve que chaque langue a son environnement linguistique et son contexte culturel spécifique. Cela rend donc nécessaire d'ajouter des concepts qui prennent en considération les aspects historiques et culturels de la langue arabe.

Ensuite, les ressources résultant de ce travail sont déposées dans *CLARIN-IT* infrastructure [3], où il y a un fichier descriptif du lexique AQAM et auquel sont liés les dossiers correspondant aux sections. Chaque dossier correspond à une section et contient :

— un fichier nom_section.TXT contenant le texte brut de la section enrichi d'indicateurs qui détermine le début des chapitres ; des familles de racines et des entrées lexicales.

— un dossier contenant des fichiers XML divisés selon des catégories grammaticales d'entrées lexicales (Verbe ; Noms ; Adjectifs).

3. http ://hdl.handle.net/20.500.11752/ILC-97 (18/12/2019)

— Un dossier contenant les noms propres qui sont présentés dans des fichiers séparés en fonction de leurs classes sémantiques correspondantes (plante, nourriture, animal, nom propre, géographique, eau, groupe et autres).

— un dossier contenant des fichiers XML de verbes, noms et adjectifs enrichis de traductions anglaises.

En perspective, concernant la conversion AQAM, nous prévoyons de contrôler et d'optimiser le système à base de règles et les expressions régulières utilisées dans l'extraction d'informations, puis de vérifier et de valider davantage les résultats générés automatiquement.

Rappelons que la ressource conçue est codée dans un format XML générique, et dans Nahli et al. (2016) et Khalfi et al. (2016) nous avons présenté deux études de conversion d'AQAM au format standard comme LMF (Lexical Markup Framework) et LEMON (LExicon Model for ONtology), révélant que ces normes ne couvrent pas toutes les spécificités de la langue arabe, comme pour leur module morphologique qui doit être étendu afin de représenter certaines caractéristiques morphologiques spécifiques à l'arabe (*maṣdar*, imparfait, paradigme, racine, etc.). Ainsi, cela nous pousse à réfléchir à l'utilisation du codage TEI (Text Encoding Initiative) qui est plus convivial pour les données historiques.

Par ailleurs, concernant le niveau sémantique, nous avons l'intention de continuer à enrichir la ressource lexicale, et d'utiliser d'autres ressources bilingues basées sur l'arabe classique afin d'obtenir plus de couverture et d'augmenter le taux de mapping. En fait, la désambiguïsation est très compliquée et prend du temps, d'où la nécessité de trouver des moyens d'aide dans le processus de contrôle, en utilisant des techniques de mesure de similarité sémantique basées sur les connaissances.

Enfin, nous envisageons de continuer à enrichir notre ressource lexicale, dans le but de l'utiliser dans d'autres projets de recherche.

Annexes

.1. Listes des suffixes nominatives et transitives

TABLE 3 – Suffixes Nominatives et Transitives

Nominative Suffixes		Transitive Suffixes	
Arabic Form	Buckwalter Transliteration	Arabic Form	Buckwalter Transliteration
◌ً	*a*	هُ	*hu*
◌َتْ	*ato*	هَا	*haA*
◌َتِ	*ati*	هُمْ	*humo*
◌ْتُ	*otu*	هُمَا	*humaA*
◌ْتَ	*ota*	هُنَّ	*hun~a*
◌ْتِ	*oti*	كَ	*ka*
◌ْتُنَّ	*otun~a*	كِ	*ki*
◌ْتُمْ	*otumo*	كُمَا	*kumaA*
◌ْنَا	*onaA*	كُنَّ	*kun~a*
◌ُوا	*uwA*	كُمْ	*kumo*
◌َا	*aA*	نَا	*naA*
-	-	نِي	*niy*

.2. Liste des lettres solaires utilisées

Script	Transliteration	Letter	Transliteration
ت	*t*	التَّاء	*tā'*
ث	*ṯ*	الثَّاء	*ṯā'*
د	*d*	الدَّال	*dāl*
ذ	*ḏ*	الذَّال	*ḏāl*
ر	*r*	الرَّاء	*rā'*
ز	*z*	الزَّاي	*zāy*
س	*s*	السّين	*sīn*
ش	*š*	الشّين	*šīn*
ص	*ṣ*	الصَّاد	*ṣād*
ض	*ḍ*	الضَّاد	*ḍād*
ط	*ṭ*	الطَّاء	*ṭā'*
ظ	*ẓ*	الظَّاء	*ẓā'*
ل	*l*	اللَّام	*lām*
ن	*n*	النّون	*nūn*

TABLE 4 – lettres solaires

.3. Exemple de certains tags/mots clés utilisés

Tags	SUMO basic level concept	Semantic Class
اسْمٌ - صحابيٌّ - صحابِيَّاتٌ - صحابِيَّة - شاعرٌ - محدثٌ - (مَلِكٌ ل) رجلٌ - أبو قَبيلة - أسماءٌ - مِن أسْمائِهِنَّ - لَقَبُ - لُقِّبُ أبو قَبيلَةٍ - امْرَأَةٌ - [اسمُ رَجُلٍ] - [والِدُ + مضاف إليه] - شَيخٌ - قارِئٌ - (بنتُ + مضاف إليه) - (أبو بَطْنٍ) - مُتَكَلِّمٌ - كُنْيَةُ - جَدُّ - زاهِدٌ - عَلَمٌ - مَلِكٌ - فَرْدٌ - مَوْلىً - مُؤَرِّخٌ - تابِعِيٌّ - أخْبارِيٌّ	Human	Proper Name
قَبِيلَةٌ - جَمَاعَةٌ - بَطْنٌ - قَبائِلُ - بُطُونٌ - قَوْمَةٌ	Group	Group
د – بَلَدٌ ة - قَرْيَةٌ - قَرْيَتانِ - [كُورَةٌ] - مَكَّةُ - اسْمُ كُورَةٍ + مَدِينَةُ ع - مَوْضِعٌ - مَوَاضِعُ - مَوْضِعَانِ - المَكَانُ - نَاحِيةٌ (بَيْنَ – قُرْبَ - وَرَاءَ - مِنْ) - أَرْضٌ - مَحَلَّةٌ - جَبَلٌ - جَبَلانِ - أجْبُلٌ - وجِبالٌ حَجَرٌ - صَخْرَةٌ هَضَبَة - هِضَابٌ	Nation City Region Neighborhood Mountain Rock Hill	Geographic
مَاءٌ - [مَاءٌ لِ] - مَاءةٌ - مُوَيْهَةٌ - مِيَاهٌ - مَسِيلٌ بِئْرٌ - الآبَارُ نَهْرٌ - وادٍ - وادِيانِ سَبَخَةٌ	Creek MineOrWell River Swamp	Water
نباتٌ - نَبْتٌ - عُشْبَةٌ - بَقْلَةٌ - ما يَنْبُتُ شَجَرٌ - [شُجَيْرَةٌ] - كُلُّ شَجَرٍ - نَخْلٌ	Plant Botanical Tree	Plant
طَعَامٌ - [مِن طَعَامٍ] شَرَابٌ - [شَرَابٌ مِن] تَمْرٌ - مِن تَمْرٍ	Food Beverage Date Fruit	Food
النَّحْلُ سَمَكٌ الذِّئْبُ - [الذِّئْبُ + نعتٌ] - اسْمُ الذِّئْبِ فَرَسٌ م - فَرَسٌ لِ - أفْراسٌ لِ - [فَرَسُ + مضاف إليه] - اسْمُ فَرَسٍ... قُنْفُذٌ - دُوَيْبَّةٌ ...	Bee Fish Canine Horse Animal	Animal
داءٌ - بَثْرٌ – مَرَضٌ دَواءٌ	Disease Or Syndrome Biologically Active Substance	Disease
لُعْبَةٌ شَهْرٌ اسمُ جِنِّيٍّ - قَبِيلَةٌ مِنَ الجِنِّ - اسْمُ الشَّيْطانِ كَوْكَبٌ ثَوْبٌ	Game Time Interval Cognitive Agent Planet Clothing	Others

TABLE 5 – Exemple de tags/mots clés pour l'extraction des noms propres

Publications

1. M. Khalfi, A. Zarghili and O. Nahli, "A new rich Lexical Resource for Classical Arabic", *COMPUSOFT : An International Journal of Advanced Computer Technology*, vol. 9, no. 10, pp. 3863-3885, Oct. 2020.
   Retrieved from https ://www.ijact.in/index.php/ijact/article/view/1196
2. M. Khalfi, O. Nahli and A. Zarghili, "Classical dictionary Al-Qamus in lemon", in : *2016 4th IEEE International Colloquium on Information Science and Technology (CiSt)*, Tangier, 2016, pp. 325-330, doi : 10.1109/CIST.2016.7805065.
3. O. Nahli, F. Frontini, M. Monachini, F. Khan, A. Zarghili and M. Khalfi, 2016. "Al qamus al muhit, a medieval arabic lexicon in lmf", in : *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, European Language Resources Association (ELRA), Paris, France. pp. 943–950.

Bibliographie

Abou-sikkīn, a.m., 1981. al=ma‘ājim al=‘arabiyyat - madārisuhā wa=manāhijuhā. 2nd edition ed., al=fārūq al=ḥadīṯat li aṭ=ṭibā‘at wa an=naŝr.

Abouenour, L., Bouzoubaa, K., Rosso, P., 2013. On the evaluation and improvement of arabic wordnet coverage and usability. Language Resources and Evaluation 47, 891–917. URL : https://doi.org/10.1007/s10579-013-9237-0, doi :10.1007/s10579-013-9237-0.

Al-fīrūz’ābādī, 1998. al-qāmūs al-muḥīṭ. mu’assasat ar-risālah, Beirut.

Al-Zoghby, A.M., Elshiwi, A., Atwan, A., 2018. Semantic Relations Extraction and Ontology Learning from Arabic Texts—A Survey. Springer International Publishing, Cham. pp. 199–225. URL : https://doi.org/10.1007/978-3-319-67056-0_11, doi :10.1007/978-3-319-67056-0_11.

Alginahi, Y., Binali, A.M.A., Dekkak, M., Kushk, A., 2017. A computerized reversible arabic transliteration system 43, 759–776. doi :https://doi.org/10.1007/s13369-017-2737-2.

AlKhalifa, M., Rodríguez, H., 2009. Automatically extending ne coverage of arabic wordnet using wikipedia, in : the 3rd international conference on Arabic language processing CITALA’09.

Alkhatib, M., Monem, A.A., Shaalan, K., 2017. A rich arabic wordnet resource for alhadith al-shareef. Procedia Computer Science 117, 101 – 110. URL : http://www.sciencedirect.com/science/article/pii/S1877050917321555, doi :https://doi.org/10.1016/j.procs.2017.10.098. arabic Computational Linguistics.

Alqrainy, S., Muaidi, H., Alkoffash, M.S., 2012. Article : Context-free grammar analysis for arabic sentences. International Journal of Computer Applications 53, 7–11. Full text available.

Ameur, M.S.H., Meziane, F., Guessoum, A., 2017. Arabic machine transliteration using an attention-based encoder-decoder model. Procedia Computer Science 117, 287 – 297. URL : http://www.sciencedirect.com/science/article/pii/S1877050917321774, doi :https://doi.org/10.1016/j.procs.2017.10.120. arabic Computational Linguistics.

Attia, M., Pecina, P., Toral, A., Tounsi, L., van Genabith, J., 2011. A lexical database for modern standard arabic interoperable with a finite state morphological transducer, in : Mahlow, C., Piotrowski, M. (Eds.), Systems and Frameworks for Computational Morphology, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 98–118.

az=zabīdī, 2001. tāĝ al-ʻarūs min ĝawāhir al-qāmūs. Maṭbaʻat al-Kuwayt.

Baʻlabakkī, R., 2014. The Arabic Lexicographical Tradition : From the 2nd/8th to the 12th/18th Century. Handbook of Oriental Studies, Brill. URL : https://books.google.co.ma/books?id=WtGcoAEACAAJ.

Benajiba, Y., Rosso, P., BenedíRuiz, J.M., 2007. Anersys : An arabic named entity recognition system based on maximum entropy, in : Gelbukh, A. (Ed.), Computational Linguistics and Intelligent Text Processing, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 143–153.

Buckwalter, T., 2004. Issues in Arabic orthography and morphology analysis, in : Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, COLING, Geneva, Switzerland. pp. 31–34. URL : https://www.aclweb.org/anthology/W04-1606.

Darwish, K., Mubarak, H., Abdelali, A., 2017. Arabic diacritization : Stats, rules, and hacks, in : Proceedings of the Third Arabic Natural Language Processing Workshop, Association for Computational Linguistics, Valencia, Spain. pp. 9–17. URL : https://www.aclweb.org/anthology/W17-1302, doi :10.18653/v1/W17-1302.

Del Gratta, R., Nahli, O., 2014. Enhancing arabic wordnet with the use of princeton wordnet and a bilingual dictionary, in : 2014 Third IEEE International Colloquium in Information Science and Technology (CIST), pp. 278–284. doi :10.1109/CIST.2014.7016632.

Duwairi, R., Al-Refai, M., Khasawneh, N., 2007. Stemming versus light stemming as feature selection techniques for arabic text categorization, in : 2007 Innovations in Information Technologies (IIT), pp. 446–450.

Elkateb, S., Black, W., Rodríguez, H., Alkhalifa, M., Vossen, P., Pease, A., Fellbaum, C., 2006. Building a WordNet for Arabic, in : Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), European Language Resources Association (ELRA), Genoa, Italy. URL : http://www.lrec-conf.org/proceedings/lrec2006/pdf/805_pdf.pdf.

Elsebai, A., Meziane, F., Belkredim, F.Z., Hassiba, U., Bouali, B., 2009. Rule based persons names arabic extraction system, in : Communications of the IBIMA, pp. 53–59.

Fellbaum, C., 1998a. A Semantic Network of English : The Mother of All WordNets. Springer Netherlands, Dordrecht. pp. 137–148. URL : https://doi.org/10.1007/978-94-017-1491-4_6, doi :10.1007/978-94-017-1491-4_6.

Fellbaum, C., 1998b. Wordnet : An electronic lexical database. MIT Press .

Gibbon, D., 2000. Computational Lexicography. Springer Netherlands, Dordrecht. pp. 1–42. URL : https://doi.org/10.1007/978-94-010-9458-0_1, doi :10.1007/978-94-010-9458-0_1.

Ide, N., Véronis, J., 1995. Encoding Dictionaries. Springer Netherlands, Dordrecht. pp. 167–179. URL : https://doi.org/10.1007/978-94-011-0325-1_14, doi :10.1007/978-94-011-0325-1_14.

Khalfi, M., Nahli, O., Zarghili, A., 2016. Classical dictionary al-qamus in lemon, in : 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt), pp. 325–330. doi :10.1109/CIST.2016.7805065.

Ladislav, B., Petr, D., Vojtech, O., 2012. Knowledge Base Modeling and Design Procedure. IOS Press. pp. 331–343.

van Langendonck, W., 2007. Theory and Typology of Proper Names. Studies and monographs], Bod Third Party Titles. URL : https://books.google.co.ma/books?id=B7jOGz7zMnwC.

Miller, G.A., 1995. Wordnet : A lexical database for english. Commun. ACM 38, 39–41. URL : http://doi.acm.org/10.1145/219717.219748, doi :10.1145/219717.219748.

Nahli, O., 2018a. Arabic language alignment with english ontologies : Some ontological reflections, in : 5th IEEE International Congress on Information Science and Technology, CiSt 2018, Marrakech, Morocco, October 21-27, 2018, pp. 254–260. URL : https://doi.org/10.1109/CIST.2018.8596580, doi :10.1109/CIST.2018.8596580.

Nahli, O., 2018b. Vers une ontologie de la culture arabo-musulmane PhD in Civiltà del 'Asia e del 'Africa, within the framework "Civiltà Islamica". Ph.D. thesis. The Doctor of Philosophy degree was obtained on 13/02/2018.

Nahli, O., Frontini, F., Monachini, M., Khan, F., Zarghili, A., Khalfi, M., 2016. Al qamus al muhit, a medieval arabic lexicon in lmf, in : Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), European Language Resources Association (ELRA), Paris, France. pp. 943–950.

Namly, D., Bouzoubaa, K., 2015. Lmf conversion of an editorial dictionary : the case of the contemporary arabic dictionary, in : Journée d'étude Ressources langagières de l'arabe pour le TAL : construction, standardisation, gestion et exploitation, Institut d'Etudes et de Recherches pour l'Arabisation, Rabat.

Nezda, L., Hickl, A., Lehmann, J., Fayyaz, S., 2006. What in the world is a shahab ? : Wide coverage named entity recognition for Arabic, in : Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), European Language Resources Association (ELRA), Genoa, Italy. URL : http://www.lrec-conf.org/proceedings/lrec2006/pdf/368_pdf.pdf.

Niles, I., Pease, A., 2003. Linking lexicons and ontologies : Mapping wordnet to the suggested upper merged ontology, in : PROCEEDINGS OF THE 2003 INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE ENGINEERING (IKE 03), LAS VEGAS, pp. 412–416.

Oudah, M., Shaalan, K., 2012. A pipeline Arabic named entity recognition using a hybrid approach, in : Proceedings of COLING 2012, The COLING 2012 Organizing Committee, Mumbai, India. pp. 2159–2176. URL : https://www.aclweb.org/anthology/C12-1132.

Pease, A., Niles, I., Li, J., 2002. The suggested upper merged ontology : A large ontology for the semantic web and its applications. AAAI Technical Report WS-02-11 .

Piasecki, M., Szpakowicz, S., Broda, B., 2009. A Wordnet from the Ground Up. Wroclaw University of Technology Press. URL : http://www.plwordnet.pwr.wroc.pl/main/content/files/publications/A_Wordnet_from_the_Ground_Up.pdf. (ISBN 978-83-7493-476-3).

Regragui, Y., Abouenour, L., Krieche, F., Bouzoubaa, K., Rosso, P., 2016. Arabic wordnet : New content and new applications, in : Proceeding of the 8th Global Wordnet Conference (GWN 2016), Bucharest, Romania.

Rodríguez, H., Farwell, D., Farreres, J., Bertran, M., Alkhalifa, M., Martí, M.A., Black, W., Elkateb, S., Kirk, J., Pease, A., Vossen, P., Fellbaum, C., 2008a. Arabic wordnet : Current state and future extensions, in : The fourth global WordNet conference, pp. 387–405.

Rodríguez, H., Farwell, D., Ferreres, J., Bertrán, M., Alkhalifa, M., Martí, M.A., 2008b. Arabic wordnet : Semi-automatic extensions using bayesian inference, in : the the 6th Conference on Language Resources and Evaluation LREC2008.

Salmoné, H.A., 1978. An advanced learner's Arabic-English dictionary : incl. an Engl. index. Libr. du Liban. URL : https://books.google.co.ma/books?id=DQwEvwEACAAJ.

Sami, B., D, M.W.W., 2010. Aralex : A lexical database for modern standard arabic. Behavior Research Methods 42, 481 – 487. URL : https://doi.org/10.3758/BRM.42.2.481, doi :10.3758/BRM.42.2.481.

Shaalan, K., 2010. Rule-based approach in arabic natural language processing. International Journal on Information and Communication Technologies 3, 11—-19. URL : http://www.ieee.ma/IJICT/IJICT-SI-Bouzoubaa-3.3/3%20-%20Khaledl.pdf.

Shaalan, K., 2014. A survey of Arabic named entity recognition and classification. Computational Linguistics 40, 469–510. URL : https://www.aclweb.org/anthology/J14-2008, doi :10.1162/COLI_a_00178.

Shaalan, K., Habash, N.Y., 2010. Introduction to arabic natural language processing (synthesis lectures on human language technologies). Machine Translation 24,

285–289. URL : https://doi.org/10.1007/s10590-011-9087-8, doi :10.1007/s10590-011-9087-8.

Shaalan, K., Oudah, M., 2014. A hybrid approach to arabic named entity recognition. Journal of Information Science 40, 67–87. URL : https://doi.org/10.1177/0165551513502417, doi :10.1177/0165551513502417, arXiv:https://doi.org/10.1177/0165551513502417.

Wehr, H., 1976. A Dictionary of Modern Written Arabic. 3rd edition ed.. Spoken Language Service, Inc. ISBN 0-87950-001-8, p. XI.

Zerrouki, T., Balla, A., 2017. Tashkeela : Novel corpus of arabic vocalized texts, data for auto-diacritization systems. Data in Brief 11, 147 – 151. URL : http://www.sciencedirect.com/science/article/pii/S2352340917300112, doi :https://doi.org/10.1016/j.dib.2017.01.011.

Centre d'Etudes Doctorales : Sciences et Techniques de l'Ingénieur

Etablissement : Faculté des Sciences et Techniques de Fès

Résumé de la thèse

Actuellement, les grandes ressources lexicales acquièrent une pertinence potentielle élevée pour les systèmes d'information et le besoin de ressources lexicales dans les domaines du Traitement Automatique des Langues Naturelles (TALN) est primordial.

Pour contribuer à répondre à ces besoins pour la langue Arabe qui souffre d'une grave pénurie de ce genre de ressources numériques, nous construisons une ressource lexicale à base du célèbre dictionnaire *al=qāmūs al=muḥīṭ* (AQAM). En utilisant une approche basée sur des règles, nous avons conçu un système qui permet d'extraire des informations morpho-syntaxiques, sémantiques et lexicales du célèbre dictionnaire. Nous avons donc obtenu une version numérisée et structurée d'AQAM, enrichie d'informations explicites morpho-syntaxiques et lexicales.

Dans cette thèse, nous décrivons les étapes suivies pour la conversion d'AQAM en un format lisible par machine, en commençant par la phase de segmentation du texte, puis l'extraction des lemmes et des sens, suivie de la phase d'identification de la catégorie grammaticale (Part Of Speech), l'extraction d'informations morpho-syntaxiques et lexicales.

Le dictionnaire est très riche de noms propres, nous avons donc dû faire face à une méthodologie automatique permettant de les identifier et les extraire. La tâche d'extraction des noms propres est effectuée sur la base d'une liste de balises définie constituée de mots-clés utilisés par l'auteur d'AQAM.

De plus, la ressource obtenue est enrichie par des traductions anglaises du lemme et des sens simples qui l'accompagnent à l'aide de deux dictionnaires bilingues Arabe-Anglais.

Ensuite, nous présentons un aperçu d'un alignement d'expérience de la section de la lettre ***bā'*** sur le WordNet de Princeton (PWN) et Suggested Upper Merged Ontology (SUMO).

Cette expérience s'est avérée intéressante car elle a révélé que le mappage d'une ressource lexicale arabe sur une ressource anglaise montre des points communs entre les deux langues, mais elle permet surtout de souligner les non-équivalences entre elles.

Toutes les ressources obtenues sont représentées au format XML et distribuées sous licence gratuite sous la plate-forme CLARIN-IT.

Mots clés : Extraction d'informations, lexique arabe, Al Qamus Al Muhit, dictionnaire lisible par machine, ressource lexicale arabe.