N° d'ordre :

REPUBLIQUE ALGERIENNE DEMOCRATIQUE & POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR & DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE DJILLALI LIABES
FACULTE DES SCIENCES EXACTES
SIDI BEL ABBES

THESE
DE DOCTORAT

Présentée par Noureddine DOUMI

Spécialité : *Informatique*
Option : *Intelligence artificielle*

Intitulée

Extraction de connaissances à partir du texte

Soutenue le…08 Juin 2017…
Devant le jury composé de :

Président	M. Abderrahmane YOUSFATE	Pr.	Université de Sidi Bel Abbes
Examinateurs	M. Mimoun MALKI	Pr.	ESI de Sidi Bel Abbes
	M. Abdelkader GAFOUR	MCA	Université de Sidi Bel Abbes
	M. Mohammed El Amine ABDERRAHIM	MCA	Université de Tlemcen
	M. Djelloul BOUCHIHA	MCA	Université de Naâma
Directeur de thèse	M. Ahmed LEHIRECHE	Pr.	Université de Sidi Bel Abbes

Année universitaire 2016/2017

Je dédie l'effort
de ce modeste travail aux âmes de mes
défunts parents
et je dédie son contexte et l'éclairage
qu'il apporte
à ma petite famille

Remerciement

Je remercie mon Directeur de thèse M. Ahmed LEHIRECHE pour m'avoir proposé un sujet d'actualité, ouvert sur plusieurs domaines et encore vierge pour la recherche scientifique. L'engagement dans ce sujet m'a permis d'entamer différents domaines de recherche entre traitement automatique de langues, ingénierie de connaissance, théorie des automates, linguistique arabe et linguistique computationnelle, etc. Durant cette expérience j'ai confronté beaucoup de problèmes, ce qui m'a permis d'acquérir de nouvelles notions et d'évoluer en termes de connaissances. Je le remercie aussi pour le choix de la langue à traiter dans le cadre de cette thèse (la langue arabe). Cette langue qui est encore moins riche en termes d'étude théorique automatisable, d'approches scientifiques et de ressources et outils de traitement automatique, ce qui nous a donné l'opportunité à proposer des solutions dans un terrain encore vierge. Le choix de cette langue nous a poussé à revoir nos connaissances linguistiques une autre fois et d'une nouvelle vision et de redécouvrir une autre fois cette langue. Je le remercie encore pour les multiples séances de travail qui nous ont réuni cote à cote pour aboutir aux objectifs de cette thèse, que ce soit pour les différents papiers scientifiques rédigés dans le cadre de cette thèse ou pour la rédaction la thèse elle même.

Mes remerciements vont aussi à M. Denis MAUREL de l'école polytechnique de Tours pour m'avoir orienté durant toute la période de préparation de cette thèse et pour m'avoir reçu parmi les membres de son équipe de recherche BdTln et cela depuis l'année 2008. Je le remercie aussi pour les différentes séances de travail organisées pendant mes stages effectués au Laboratoire d'Informatique de l'université François Rabelais de Tours durant lesquels j'ai acquis beaucoup de notions sur l'application de la théorie des automates dans le traitement automatique des langues et particulièrement sur l'extraction d'information. Les projets entrepris par le laboratoire LI de l'université de Tours et exactement par l'équipe de BdTln m'ont permis de voir de près des projets de grande envergure et d'un impact industriel réel tels que le projet Prolexbase, Entités, Renom, Variling etc. Je le remercie aussi pour les multiples relectures de mes papiers scientifiques et les différents écrits en va-et-vient entre nous et cela malgré son calendrier très chargé.

Je tiens à remercier M. Abderhamane YOUSFATE enseignant et directeur de laboratoire à l'université Djillali Lyabes de Sidi Bel Abbes d'avoir accepté de présider le jury de ma soutenance. Il m'honore beaucoup que mon ex enseignant en graduation et en post-graduation accepte de statuer sur mon travail de thèse et de l'enrichir par ses remarques coté fondement théorique et que je les trouve toujours pertinentes et toujours les bienvenues.

Je remercie M. Mimoun MALKI enseignant et directeur de laboratoire à l'école supérieure d'informatique de Sidi Bel Abbes d'avoir accepté d'examiner ma thèse. Je profite cette occasion pour le remercier aussi pour tout ce qu'il a fait pour promouvoir la post graduation et la recherche scientifique au département d'informatique de l'université de Sidi Bel Abbes et dont j'étais l'un des bénéficiaires.

Je remercie M. Abdelkader GAFOUR enseignant chercheur à l'université de Sidi Bel Abbes d'avoir accepté d'examiner ma thèse et de statuer sur le fruit de mon effort. Je tiens à le remercier aussi pour tous les conseils et orientations qu'il nous a prodigués pendant toute notre formations au département d'informatique.

Je remercie M. Mohammed El Amine ABDERRAHIM enseignant chercheur à l'université de Tlemcen d'avoir accepté d'examiner ma thèse. Son expertise et sa longue expérience dans le domaine du TAL arabe va ajouter une valeur de crédibilité aux discussions et critiques à mon travail dans la soutenance.

Je remercie M. Djelloul BOUCHIHA enseignant chercheur à l'université de Naama d'avoir accepté d'examiner et de statuer sur ma thèse. Son expérience dans les domaines du web sémantique ainsi que l'ingénierie de connaissance vont apporter une valeur ajoutée à la discussion de ma soutenance.

Je remercie M. Sébastien PAUMIER et M. Eric LAPORTE de l'université Paris Est Marne-La-Vallée de m'avoir introduit dans leur projet Unitex/GramLab et m'avoir ouvert l'opportunité de participer au développement de cette plateforme ce qui m'a donné la chance à participer dans un projet de renommée internationale et ayant une grande communauté d'utilisateurs multilingues.

Je remercie M. Ali RAHMOUNI Directeur de laboratoire de modélisation et de méthodes de calcul à l'université de Saïda d'avoir mis à ma disposions l'endroit et matériel nécessaire pour la rédaction et à la mise au point de mes travaux de thèse.

Je profite aussi de l'occasion pour remercier Mlle Emeline LLECUIT du laboratoire ligérien de linguistique de l'université d'Orléans pour les longues discussions sur la plateforme Unitex/GramLab et sur les relectures effectuées de sa part de mes papiers scientifiques rédigés en anglais.

Je remercie M. Ahmed ABDELALI de Qatar Computing Research Institute pour les multiples écrits en messagerie électronique et les différentes séances de travail en Skype sur le TAL arabe en général et particulièrement sur la construction de ressources. Je le remercie aussi pour sa disponibilité et son intérêt qu'il donne au chercheurs en TAL arabe en Algérie.

Je remercie M. Marwane Al BAWAB de la Syrie de m'avoir aidé à comprendre les algorithmes de la morphologie arabe du système Sarf d'ALESCO.

Je remercie M. Taha ZEROUKI enseignant chercheur à l'université de Bouira de m'avoir intégré dans son projet Qutrub : conjugueur automatique de l'arabe ; et où j'ai apporté ma contribution en matière de vérification flexionnelle.

Je remercie M. Younes SAMIH du département de linguistique et des sciences d'information de l'université Heinrich Heine de Dusseldörf de m'avoir orienté sur l'utilisation des outils et de ressources en morphologie arabe tels que le système Sarf d'ALESCO que j'ai trouvés très utiles et très pertinents dans mon travail sur la construction des ressources pour l'arabe.

Je profite aussi de l'occasion pour remercier M. Gupta ANUBAV de l'équipe BdTln du laboratoire d'informatique de l'école polytechnique de Tours pour son assistance à mon égard afin de comprendre le code source de la plateforme Unitex/GramLab. Son expérience dans le développement de cette plateforme m'a procuré le courage nécessaire et m'a permis, aussi, de participer au développement de cette plateforme.

Je profite aussi de cette occasion pour remercier tous mes collègues de l'université de Saïda et du laboratoire EEDIS, notamment messieurs Moussa ALI Cherif, Maamar KHATER, Adil TOUMOUH, Aïssa FELLAH, Ahmed ZAHAF, Laouni MIMOUN et Djelloul HICHOUR pour les longues séances de discussion autour du TAL, ingénierie de connaissance, Web sémantique, méthodologies de recherche scientifique etc.

Je remercie M. Yasser YAHIAOUI et M. Abdelghani BOUZIANE de l'université de Naama pour les langues séances de travail qu'on a organisées ensemble sur la linguistique et le TAL arabe en général et sur nos futurs projets dans ce domaine.

Enfin, je ne saurais oublier mes vifs remerciements à tous les membres de ma famille qui ont toujours été à mes cotés et m'ont toujours encouragé à persévérer. Leurs encouragements

m’ont permis d’avoir l’énergie nécessaire et la patience pour mener à bien et parachever la préparation de cette thèse dans les meilleures conditions.

الملخص

موضوع هذه الأطروحة يتناول استنباط عناصر المعرفة من النص العربي. هذه المهمة أُنجزت عن طريق اكتشاف ثم استخراج العلاقات الدلالية بين أنماط الأسماء. إشكالية تحديد ثم استنباط أنماط الأسماء وكذا العلاقات الدلالية التي تربطها قد حُلّت عن طريق منهجية مستندة على القواعد. لقد قمنا بكتابة هذه القواعد والتي يحددها الخبير، على شكل مبدلات ذات حالات منتهية.

إنّ النقص الكبير في الموارد اللغوية والبرمجيات الضرورية للمعالجة الآلية للنص العربي دفعنا إلى بناء مواردنا الخاصّة وتكييف برمجيات منصة Unitex/GramLab بهدف استكمال المهمة المذكورة أعلاه. إنّ الموارد اللغوية قد بُنِيَت ثم ضُغِطت وخُزِّنت كذلك باستعمال المبدلات ذات الحالات المنتهية.

الكلمات المفتاحية

استنباط المعرفة، أنماط الأسماء، العلاقات الدلالية، المبدلات ذات الحالات المنتهية

Résumé

Dans cette thèse on aborde le sujet d'extraction des connaissances à partir du texte arabe. Cette tache a été réalisée à travers la détection et l'extraction des relations sémantiques entre les entités nommées. La problématique de repérage et d'extraction des entités nommées ainsi que les relations sémantiques les reliant a été résolue en utilisant une approche à base de règles, où les règles de l'expert sont traduites sous formes de transducteurs à états finis.
Le manque terrible des ressources linguistiques et d'outils nécessaires au TAL arabe nous a conduit à construire nos propres ressources et à l'adaptation des outils de la plateforme Unitex/GramLab afin d'accomplir les taches citées ci-dessus. Les ressources sont aussi construites et puis compressées et stockées en utilisant les transducteurs à états finis.

Mots clés
Extraction de connaissances, Entités nommées, relations sémantiques, les transducteurs à états finis.

Abstract
In this thesis we address the issue of knowledge discovery within Arabic text. This task was achieved by detecting and recognizing the semantic relations between named entities. The issue of repering and extracting the named entities as well as the semantic relations binding them is solved by using a rule-based approach where we convert the expert rules to finite state transducers.
The lack of linguistic resources and tools needed for Arabic NLP has pushed us to build our own resources and to adapt the Unitex/GramLab tools to achieve tasks mentioned above. The resources are also built, then compressed and stored using the finite state transducers.

Keywords
Knowledge discovery, Named entities, Semantic relations, Finite state transducers.

Table des matières

Chapitre II
Reconnaissance des entités nommées

Chapitre III
Relations entre les entités nommées

Partie II

Chapitre IV
Le module arabe d'Unitex/GramLab

Chapitre V
Mise en œuvre d'extraction d'entités nommées arabes

Liste des figures

Liste des tableaux

Introduction générale

Le traitement automatique de la langue arabe a suscité l'écoulement de beaucoup d'encre scientifique durant les deux dernières décennies. Le TAL arabe ou ANLP pour Arabic Natural Language Processing en anglais, a connu un engouement des scientifiques dans les grands laboratoires de recherche et les grandes universités à l'instar de l'université de Stanford et l'université de Pennsylvanie aux Etats Unis notamment après les évènements du 11 septembre 2001. Cet engouement est intensifié davantage avec l'apparition des réseaux sociaux qui peuvent être considérés comme un grand média d'échange entre le grand public sauf que ce grand public utilise pour la communication des langues vernaculaires non standards (les dialectes de l'arabe). En plus des points cités ci-dessus s'ajoute l'intérêt que donnent ALESCO (Arab League Educational, Cultural and Scientific Organization) et tous les pays arabes et particulièrement les pays de Golf à l'enrichissement et l'arabisation du contenu web et à l'informatisation de leurs gouvernements (gouvernement électronique). Un autre facteur à cet engouement, c'est la place qu'occupe la langue arabe dans le classement mondial des langues les plus utilisées sur Internet (la quatrième place devançant le français et l'allemand par cinq positions[1]). Ce classement montre d'une façon claire que cette langue est en pleine expansion en termes d'utilisateurs sur Internet. Le monde arabe en tant que marché représente une grande opportunité aux grandes entreprises internationales de recherche et de développement et les pousse à adopter cette langue dans leurs produits informatiques.

L'extraction d'information et de connaissance sont des applications les plus connues et anciennes de Traitement Automatique des Langues. Malgré leurs fins différentes, elles utilisent les mêmes approches et techniques pour réaliser ces fins. L'objet de l'extraction d'information est la création de représentation structurée d'information bien sélectionnée à partir de sources non structurées tels que des textes en langage naturel. Alors que l'objectif de l'extraction de connaissance est de construire des modèles sémantiques. Le modèle sémantique peut prendre différentes formes de la plus simple tel qu'un schéma conceptuel jusqu'à la forme la plus riche sémantiquement telle qu'une ontologie de domaine. Si on prend l'ontologie comme un modèle sémantique de référence, elle est composée entre autres de concepts et de relations sémantiques taxonomiques et non taxonomiques entre les concepts.

L'entité nommée par ses différentes formes et types constitue un pourcentage considérable dans les textes du web, notamment les articles journalistiques. La détection et repérage et extraction de cette information relève du domaine de l'extraction d'information qui trouve actuellement son intérêt dans plusieurs domaine industriels et de technologie d'information et de communication telles que la veille technologique, la veille médiatique, les systèmes de navigation, les moteurs de recherche, les système question-réponse etc. L'extraction de connaissance de sa part, a pour objectif de construire un modèle de connaissances (modèle sémantique). Ce dernier est composé de plusieurs éléments qu'on peut extraire à partir du texte brut.

Objectifs et contribution

Dans cette thèse on utilise la technique de filtrage/reconnaissance de motifs (pattern matching/recognition) pour repérer des patrons textuels bien définis. Après le repérage on classifie le contenu repéré dans des classes bien déterminées (Les types d'entités nommées). Dans le repérage ainsi que la classification on utilise une approche à base de règles. Où les règles sont de type si-alors et les patrons ou motifs à reconnaitre sont écrits sous formes de machines à états-finis (automates ou transducteurs). Précisément l'objectif de notre travail consiste à détecter et repérer les entités nommées dans un texte arabe et extraire les relations

[1] Internet World Users by Language: Top 10 Languages pour l'année 2015 consulté en janvier 2016 sur l'url http://www.internetworldstats.com/stats7.htm.

sémantiques non-taxonomiques qui les relient. Les textes choisis sont de type journalistique du fait qu'ils renferment 10% de leur contenu des entités nommées[2]. La langue traitée dans le travail courant (l'arabe) est considérée parmi les langues sous-ressourcées (under resourced) et moins denses en ce qui concerne les ressources linguistiques et les outils de traitement automatiques disponibles pour la communauté de recherche. Ce manque nous a poussés à résoudre ce problème avant de procéder à répondre aux objectifs de cette thèse. La solution consistait à choisir une plateforme TAL libre et multi langues qui est Unitex/GramLab de l'université Pars Est Marne-La-Vallée. Au lancement de cette thèse, la plateforme contenait des modules pour une vingtaine de langues mais ne contenait pas l'arabe, ce qui nous a poussé à développer un jeu de ressources linguistiques pour le TAL arabe sous cette plateforme. Le jeu consistait en un ensemble de : corpus de test, de dictionnaires électroniques, de jeu d'étiquettes morphosyntaxiques, de grammaires de flexions, de grammaires de segmentation et de fichiers de configuration.

Dans notre travail on utilise principalement les technologies à états finis ; que se soit pour la construction et le stockage des ressources lexicales ou pour le repérage et la détection des entités nommées et des relations qui les relient. Dans la construction de ressources lexicales on utilise les transducteurs à états finis pour la génération des entrées de nos dictionnaires de type LADL et puis on stocke aussi ces entrées sous formes de transducteurs. En ce qui concerne le repérage des entités nommées, on a choisi d'utiliser une approche à base de règles et les règles de type ***Si-Alors*** sont sous forme ***Si*** *une structure syntaxique locale est renfermée dans un contexte bien déterminé* ***Alors*** *on considère cette structure comme étant entité nommée et on lui attribue la classe correspondante à ce contexte*. Donc les patrons syntaxiques de repérage sont des grammaires locales exprimées sous forme de réseaux de transducteurs à états finis et ils sont appliqués au corpus textuel en cascades i.e. on applique une série de transducteurs dans un ordre bien défini de sorte que l'entrée du premier transducteur un texte brut et sa sortie un texte annoté avec certaines entités nommées et quelques balises utilitaires et cette sortie constitue une nouvelle entrée au deuxième transducteur et ainsi de suite jusqu'à l'application du dernier transducteur dans la cascades qui produit un texte annoté avec les entités nommées recherchées et leurs classes.

L'extraction de relation est le processus d'identification de mention de relations dans le texte, où une relation mentionnée est définie comme un prédicat s'appliquant sur deux arguments ou plus. Les arguments représentent des concepts et la relation prédicat décrit le type d'association ou d'interaction qui tient entre les concepts représentés par les arguments. L'extraction d'entités nommées et de leurs relations peut être vue soit comme une extraction d'instances de concepts et instances de relations (ce qu'on appelle ontology population) ou bien un apprentissage de relations entre les concepts (ce qui vient à faire de l'ontology learning). Dans les travaux de recherche sur la construction automatique ou semi automatique des ontologies la communauté de la recherche scientifique s'intéresse beaucoup plus des relations taxonomiques de type est-un. Peu de travaux et d'approches focalisent sur les relations non-taxonomiques. En conséquence, une grande partie de la sémantique spécifique au domaine est ignorée. Dans notre travail on s'intéresse aux relations sémantiques d'association non-taxonomiques. Ce processus peut être motivé par l'enrichissement du modèle sémantique telle qu'une ontologie.

L'élément du modèle sémantique concerné par cette thèse c'est les relations sémantiques non-taxonomiques qui peuvent exister entre les entités nommées repérées dans le texte. En ce qui concerne ce point l'approche utilisée est semi-automatique de sorte que le système proposé extrait la relation en utilisant des ressources lexicales et sémantiques relatives à la

[2] La page du projet français Prolex consultée en janvier 2016 à l'url http://www.cnrtl.fr/lexiques/prolex/

langue du texte (l'arabe) et des transducteurs d'annotation appliqués sur le résultat de la reconnaissance des entités nommées. Les deux processus (d'extraction d'entités nommées ou d'extraction de relation) se font en cascade de transducteurs d'annotation. Un expert linguiste valide la relation trouvée par le système et comparée à une ontologie de haut niveau (SUMO).

L'organisation des chapitres

Cette thèse est organisée en deux parties et en cinq chapitres. La première partie regroupe les trois premiers chapitres dont l'objectif est d'introduire les notions, définitions, approches et en général le bagage théorique nécessaire à la compréhension des solutions proposées dans la deuxième partie. Dans le premier chapitre, on détaille les notions en relation avec le TAL arabe et on essaye de donner une synthèse sur l'état de l'art des différents niveaux de traitement, des outils, des ressources et corpus développés jusqu'à aujourd'hui. Ce chapitre procure de son lecteur les deux principaux aspects : une synthèse de la linguistique arabe et de la particularité du TAL arabe. Dans le deuxième chapitre, on introduit la notion des entités nommées en général tout en discutant le cas de l'arabe. A chaque fois si nécessaire on donne des exemples d'éclaircissement en arabe standard moderne. A la fin de ce chapitre on fait une synthèse d'état de l'art sur les approches et sur les systèmes de reconnaissances des entités nommées arabes. Le chapitre trois entame en détail les relations sémantiques non-taxonomiques qui peuvent exister entre les entités nommées dans le texte arabe.

La deuxième partie représente notre contribution et résume notre travail dans le cadre de cette thèse. Elle est organisée en deux chapitres (quatrième et cinquième chapitre). Dans le quatrième chapitre on expose les réalisations effectuées dans le cadre de la construction de ressources linguistiques arabes sous la plateforme Unitex/GramLab et sa publication sous forme d'un module gratuit et libre pour la communauté du TAL arabe. L'approche proposée pour construire les dictionnaires électroniques pour les différentes catégories syntaxiques et le jeu d'étiquettes qui y est associé sont exposés avec des chiffres sur l'expérimentation et leur évaluation par rapport à une référence de renom. L'utilisation de ces ressources dans l'extraction de connaissance : extraction d'entités nommées et détection de relation sémantique entre elles, est donnée aussi sous forme de transducteurs à états finis. La mise en œuvre de notre travail est l'objet du dernier chapitre de cette thèse.

Chapitre I

Introduction au traitement automatique de la langue arabe

1. Introduction

La langue arabe appartient à la famille des langues sémitiques, constituées principalement de l'arabe, l'araméen, l'amharique et l'hébreu. Les langues sémitiques sont caractérisées par i) un lexique construit principalement à partir de racines trilitères et quadrilitères[3], ii) d'un système d'écriture de droite vers la gauche et iii) d'un alphabet de type Abjed[4].

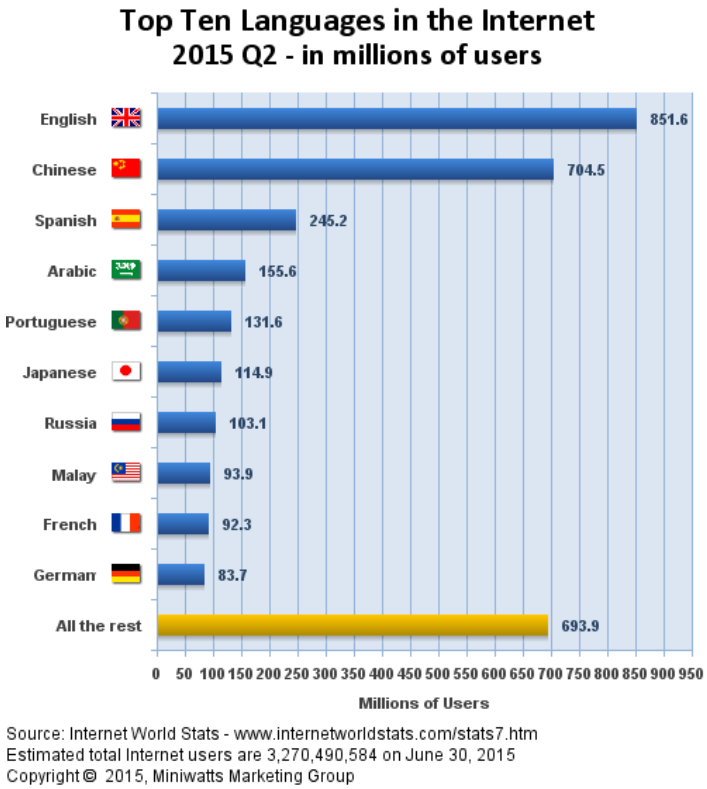


Figure 1 : Les dix langues les plus utilisées sur Internet

[3] Racines composées de trois ou quatre lettres consonnes ; par fois appelées triconsonantiques et quadriconsonantiques. Les racines biconsonantiques et pentaconsonantiques ne génèrent que la catégorie nominale (Kouloughli 1994).

[4] Alphabet composé que des consonnes.

L'arabe est la première langue sémitique en nombre de locuteurs, elle est parlée et écrite dans le monde arabe par plus de 360 millions et utilisée dans les cultes par 1.6 milliards de musulmans soit 23,4 % de la population mondiale. Elle est classée dans les dernières années comme quatrième langue mondiale en nombre d'utilisateurs dans l'Internet[5].

Utilisée par plus de 22 pays dans le monde, la langue arabe est parlée en plusieurs dialectes non écrites. La seule langue arabe utilisée par les administrations étatiques, dans les médias officiels et dans l'éducation du monde arabe c'est l'Arabe Moderne Standard (ASM).

2. Les langues arabes

La langue arabe véhiculaire est divisée en Arabe Classique (AC) et Arabe Standard Moderne (ASM), la première est la langue des textes saints de l'islam : le Coran et le Hadith et du patrimoine culturel, littéraire et scientifique de la civilisation arabo-musulmane. Cependant l'ASM est la langue officielle du monde arabe actuellement, elle est utilisée dans l'enseignement et dans les médias. La différence entre l'AC et l'ASM consiste dans le lexique : l'ASM utilise un lexique plus grand et plus moderne que celui de l'AC (Khoja 2001) et du côté de la grammaire l'ASM a abandonné l'utilisation de quelques formes compliquées de l'ancienne grammaire.

(Attia et al. 2011) ont résumé les différences entre les deux branches de la même langue dans les points suivants :

- Le lexique d'ASM est plus riche que celui d'AC à cause qu'il contient les nouveaux mots emprunté des autres langues,
- Généralement, l'ASM se conforme en règles morphologiques et syntaxiques de l'AC, sauf qu'en ASM il y a une grande tendance pour la simplification. Les écrivains modernes utilisent un sous ensemble d'une grande gamme de structures, de flexions et de dérivations existantes dans l'AC,
- L'ordre des mots classique OVS est rarement utilisé en ASM,
- En ASM, il y a une tendance à éviter la forme passive du verbe dans le cas où la forme active est possible,
- L'ordre des mots SVO relativement marginal en AC gagne davantage d'importance en ASM.

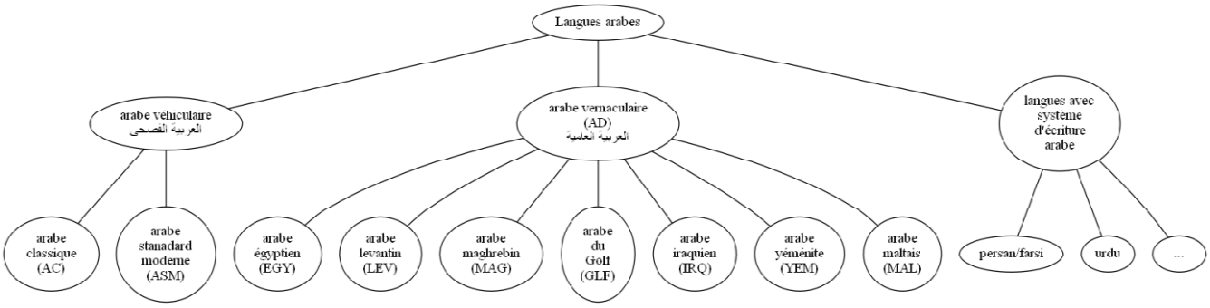


Figure 2 : Taxonomie des langues arabes

L'AC et l'ASM ont en commun d'être des langues écrites alors que les dialectes de l'arabe sont seulement parlés (langues vernaculaires); ils sont classés en sept groupes (N. Habash 2010) : arabe égyptien (EGY), arabe levantin (LEV), arabe du golfe (GLF), arabe magrébin (MAG), arabe iraquien (IRQ), arabe yéménite (YEM) et arabe maltais (MLT).

[5] Internet World Statistique à l'url http://www.internetworldstats.com (consulté en Mars 2016)

Dans la littérature on parle parfois des langues arabes pour désigner les langues non sémitiques et qui utilisent l'arabe comme système d'écriture, une partie de ces langues ont été romanisées. Les familles de ces langues sont baluche, dari, hausa, kabyle, kashmiri, kazakh, kurde, kirghize, malais, morisque, pashto, persan/farsi, punjabi, sindhi, saraiki, tatar, turc ottoman, uigur et urdu. D'autres littératures vont plus loin en considérant toute langue écrite de droite vers la gauche comme étant une langue arabe (Brun and Ehrmann 2010).

3. Le système d'écriture

L'arabe s'écrit de droite vers la gauche en utilisant deux types de caractères : les lettres et les diacritiques. Les lettres arabes renferment les consonnes et les voyelles longues alors que les diacritiques servent à vocaliser le texte i.e. déterminer la phonétique de ses mots. Quelques diacritiques à savoir فـتـحة\fat.Haħ\[6], ضمـة\Dam~aħ\ et كـسـرة\kas.raħ\ sont considérés comme des voyelles courtes. Dans ce système d'écriture les lettres sont obligatoires alors que les diacritiques sont optionnelles : l'omission des diacritiques dans le texte arabe augmente le degré d'ambiguïté pour les analyseurs morphologiques et syntaxiques.

Dans la littérature du TAL, il y a un désaccord sur le nombre de lettres arabes, ce désaccord est dû à la classification de ce qui est ou non diacritique et à l'ignorance de certaines lettres. Généralement on parle de 28 lettres mais les nombres 36 et 40 sont considérés aussi dans la littérature du domaine (Buckwalter 2004b; N. Habash 2010). Dans cette thèse on considère l'alphabet de 36 lettres car il correspond au tableau de l'Unicode arabe[7] et il est compatible avec les caractères du clavier arabe.

Les 36 lettres de l'ASM sont classées comme suit :

- 28 lettres de base,
- 6 formes possibles de la lettre Hamza,
- la lettre Ta-Marbouta et
- la lettre Alif-Maqsoura.

En plus de ce nombre de lettres, on trouve dans les textes d'ASM autres lettres non arabe telles que پ, چ, ڤ, گ pour représenter une phonétique non arabe correspondant aux lettres latines (p, j, v, g) qui n'existent pas originellement en ASM (Buckwalter 2004b). Ces lettres sont empruntées aux alphabets des langues écrites en alphabet arabe[8], tel que le *persan/farsi*.

En arabe les diacritiques sont des symboles optionnels dans l'écriture, i.e. que l'écriture arabe peut se trouver complètement, partiellement ou non diacritisée. L'ajout de diacritiques dans un texte a pour but d'aider le lecteur à lire et prononcer correctement les mots.

4. Le lexique

L'arabe est caractérisé par un lexique très riche, dérivé majoritairement à partir des racines trilitères et quadrilitères. Le nombre total de racines est estimé à environ 7000 (Al-Bawab et al. 1996; Kouloughli 1994; Al-Bawab et al. 1994; Al-Bawab 2007). Elle est caractérisé par la richesse de ses synonymes pour exprimer un concept, le Tableau 1 montre quelques exemples du nombre de mots possibles pour exprimer une notion concrète ou abstraite e.g. pour nommer l'épée, il existe 1 000 synonymes.

[6] Ici et dans ce qui suit on va utiliser la translitération/transcription des mots arabes vers leurs équivalents HSB, la translitération/transcription est une substitution d'un graphème/phonème d'un système d'écriture d'une langue source vers le système d'une langue cible, parmi ses objectifs est la lecture du mot pour un lecteur ne connaissant pas le système d'écriture de la langue source. Le HSB est choisi à cause de sa simplicité et non ambiguïté par rapport à leurs prédécesseurs.

[7] www.unicode.org/charts/PDF/U0600.pdf *(consulté en Mars 2016)*

[8] Dans la littérature on parle des langues arabes, ce sont les langues écrites en caractères arabes, il y en une vingtaine de langues telles que le persan l'urdu etc.

Dans les dictionnaires traditionnels éditoriaux le lexique arabe est organisé autour des racines ; sous chaque entrée on trouve tous les verbes, noms, adjectifs ou adverbes qui peuvent en être dérivés de l'entrée (racine). Dans le Tableau 2 nous présentons les principaux lexiques arabes pour l'ASM et l'AC[9] ; notons que les nouveaux dictionnaires tels que Elmouhit et Elghani dépassent largement le nombre de 7 000 racines en entrée, ceci est expliqué par leurs tendances à organiser le lexique arabe en lemmes ou lexèmes plutôt qu'en racines.

Lexique arabe	translitération	traduit en français	Synonymes
عَسَل	ςasal	Miel	80
أَسَد	Âasad	Lion	500
سَيْف	say.f	Epée	1 000
جَمَل	jamal	Chameau	1 000
شَرّ	šar~	Malheur	4 000

Tableau 1 : Richesse lexicale et degré d'ambiguïté sémantique du lexique arabe

Dictionnaire	auteur (s)	arabe[10]	lemmes/ racines	dérivés	mots
Elmouhit (1993)	Adib Eljemi et Sahada Elkhouri	ASM	40 000	40 000	810 000
Elghani (années 1990)	Dr Aboulazm A.	ASM	30 000	195 000	2 000 000
Tej elarouss (1965)	Mourtadha Elzebidi	ASM	11 645		3 948 160
Elouassit (1960)	Académie égyptienne de la langue arabe	ASM	7 000	30 000	450 000
Nejate elrayed (1906)	Elcheikh Ibrahim ben Nasif	ASM	142	5 629	119 176
Mouhit elmouhit (1870)	Boutrous Elboustanni	ASM	11 200	84 965	1 300 000
Elkamouss elmouhit (1414)	Elfeyrouz Abadi	AC	11 000	70 000	733 000
Lissan elarab (1311)	Ibn Mandhour	AC	9 393	158 149	4 493 934

Tableau 2 : Principaux dictionnaires éditoriaux de l'AC et de l'ASM

[9] Ce tableau est tiré du site officiel de Sakhr, une grande société de TAL arabe, http ://lexicons.sakhr.com/ consulté en mai 2016.

[10] L'historique de l'ASM a commencé au début du 19ème siècle par la modernisation de l'arabe dans le grand mouvement de renaissance arabe Nahdha (Wikipedia consulté à http ://fr.wikipedia.org/wiki/Arabe_standard_moderne en janvier 2011).

5. La morphologie arabe

La morphologie est une partie importante dans le traitement automatique des langues : la réalisation correcte et complète de l'analyse morphologique prépare le terrain et facilite les tâches qui y sont directement liées telles que la diacritisation automatique, l'étiquetage morphosyntaxique, la désambiguïsation, l'extraction de la racine, etc. ou les applications de haut niveau telles que la recherche d'information, la traduction automatique, le résumé automatique, etc.

A cause de l'utilisation croissante de l'arabe sur le web, sa morphologie a suscité un effort remarquable par les chercheurs dans les quinze dernières années citons parmi d'autres les travaux de K. R. Beesley à Xerox Research Centre Europe[11], les travaux de T. Buckwalter[12] (Buckwalter 2004b), de F. Debili (Debili et al. 2007), de A. Soudi (N. Habash 2007), de N. Habash (N. Habash 2007, 2010).

5.1. Le jeu d'étiquettes

Un jeu d'étiquette s'intéresse à l'étude des différentes catégories grammaticales (partie de discours : POS en anglais) et des traits morphosyntaxiques d'une langue. Les travaux de recherche dans ce domaine se distinguent par la granularité du jeu choisi suivant les objectifs du projet ; Citons à titre d'exemple cinq jeux d'étiquettes consacrés à la langue anglaise (parmi d'autres) :

- le jeu d'étiquettes du Brown corpus avec 87 étiquettes,
- le jeu d'étiquettes du Penn Treebank avec 45 étiquettes,
- le jeu d'étiquettes CLAWS1 avec 132 étiquettes,
- le jeu d'étiquettes CLAWS5 avec 62 étiquettes,
- le jeu d'étiquettes London-Lund avec 197 étiquettes (Al-Qrainy and Ayesh 2006; Christopher and Hinrich 2003).

Le jeu d'étiquettes arabe regroupe les trais morphologiques traditionnels sous une notation compacte, mais ce jeu, qui peut théoriquement atteindre la taille de plus de 330 000 étiquettes arabes (N. Habash 2010), n'est pas facile à cerner. Un effort de recherche limité a été investi dans le développement d'un jeu d'étiquettes arabe ; le besoin d'un tel jeu vient d'absence d'un jeu d'étiquettes exhaustif et standard (Al-Qrainy and Ayesh 2006).

Les jeux d'étiquettes arabes déjà établis sont caractérisés par la non-conformité aux recommandations EAGLES[13], ce qui est dû à la nature différente de l'arabe (langue sémitique) et des langues pour lesquelles EAGLES a été conçu, en l'occurrence les langues indo-européennes (Al-Qrainy and Ayesh 2006; Khoja et al. 2001).

Ci-après nous dressons un tableau comparatif (Tableau 3) entre quelques jeux d'étiquettes utilisés dans des projets de recherches sur le TAL arabe et qui diffèrent dans leurs applications. On remarque clairement les disparités existantes entre les différentes propositions. Par exemple le nombre de sous catégories du nom est très différent entre la 3ème ligne (103) et la 5ème ligne (11) et il en est de même pour les sous catégories du verbe entre la 1ère ligne (3) et la 3ème ligne (57).

[11] La page web de l'analyseur morphologique arabe de Xerox (consultée en mai 2016) : http ://www.cis.upenn.edu/~cis639/arabic/input/keyboard_input.html

[12] La page web de l'analyser BAMA de Buckwalter (consultée en février 2011) : http ://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp ?catalogId=LDC2004L02

[13] Groupe des experts consultatifs des standards de l'ingénierie de langage cf. l'url : http://www.ilc.cnr.it/EAGLES96/home.html

5.2. Approches de la morphologie

La morphologie est l'étude de la structure interne des mots, (N. Habash 2010) distingue deux types d'approches de la morphologie: morphologie orientée forme et morphologie fonctionnelle. La morphologie orientée forme concerne les unités composant un mot, leurs interactions les unes avec les autres et comment ils se rapportent à la forme globale de mot. En revanche, la morphologie fonctionnelle s'intéresse à la fonction des unités à l'intérieur d'un mot et comment elles affectent syntaxiquement et sémantiquement le comportement global de mot.

Appartenance	Application	Catégories	Nom	Verbe	Particules	Résiduel	Ponctuation
QAC[14]	Etiquetage morpho-syntaxique du Coran	5	34	3	21	15	12
Sawalha et al.	Général	5	29	3	13	15	11
Khoja	Etiquetage morpho-syntaxique	5	103	57	9	7	1
El-Kareh et al.	Etiquetage morpho-syntaxique	3	46	3	23	-	-
Alqrainy et al.	Vocalisation automatique	3	11	3	7	-	-

Tableau 3 : Quelques jeux d'étiquettes des projets de recherche sur le TAL arabe

5.2.1 Morphologie orientée forme

Un concept central dans la morphologie orientée forme est le morphème : la plus petite unité de sens dans une langue. Une caractéristique distinctive de la morphologie arabe et des langues sémitiques est la présence de morphèmes à schème en plus à des morphèmes par concaténation. Les morphèmes par concaténation participent à la formation du mot par un processus de concaténation séquentielle, alors que les morphèmes à schème sont entrelacés et fusionnés pour créer un mot.

Morphèmes en concaténation

Il existe trois types de morphèmes en concaténation : la base, les affixes et les clitiques. La base est indispensable pour la morphologie par concaténation, qui est nécessaire pour chaque mot. Les affixes sont attachés à la base. Il existe trois types d'affixes:

- préfixes s'attachent avant la base, par exemple أ \u0623 dans le verbe أكـتب \Âak.tubu*j'écris* (préfixe de la 1ère personne du singulier masculin ou féminin de l'inaccompli actif);
- suffixes s'attachent après la base, par exemple ت \u062A dans le verbe كـتبت\katabat.\elle *écrivait* (suffixe de la 3ème personne du singulier féminin de l'accompli actif), et

[14] QAC : projet Quranic Arabic Corpus cf l'url http://corpus.quran.com/

- circumfixes entourent la base, par exemple ي \u064A et ان \u0627 \u0646 dans le verbe يـكـتـبـان\yak.tubaAni\ *ils écrivaient* (circumfixe de la 3ème personne du duel masculin de l'inaccompli actif). Les circumfixes peuvent être considérés comme des paires liées préfixe-suffixe, l'ASM n'a pas de préfixe pur qui agit sans aucune coordination avec un suffixe.

Les clitiques s'attachent à la base après les affixes. Un clitique est un morphème qui possède les caractéristiques syntaxiques d'un mot, mais il se révèle phonologiquement lié à un autre mot. À cet égard, un clitique est nettement différent d'un affixe, qui est phonologiquement et syntaxiquement partie du mot. Les proclitiques sont des clitiques qui précèdent le mot (comme un préfixe), par exemple la conjonction و \u0648 dans le nom commun و الـكـتـاب\waAl.kitaAbu*et le livre* ou la préposition ب\u0628 dans le nom commun بـالـكـتـاب\biAl.kitaAbi*par le livre*. Les enclitiques sont des clitiques qui suivent le mot (comme un suffixe), par exemple le pronom objet ـــه\u0647 dans le verbe كـتـبـه\katabahu*il l'écrivait* (pronom de la 3ème personne du singulier masculin).

Des affixes et des clitiques multiples peuvent apparaître dans un mot. Par exemple, dans le verbe أفـيـكـتـبـونـهـا\Âafayak.tubuwnahaA*et l'écrivent-ils ?* on a :

- deux proclitiques : la particule d'interrogation أ \u0623, et la conjonction ف \u0641
- un circumfixe : ي \u064A ون \u0648\u0646 circumfixe de la 3ème personne du pluriel masculin de l'inaccompli actif et
- un enclitique: ها \u0647\u0627 pronom de la 3ème personne du singulier féminin.

Les termes préfixe et suffixe sont parfois utilisés pour désigner proclitiques et enclitiques, respectivement. Les préfixes et suffixes ont également été utilisés pour désigner toute la séquence des affixes et des clitiques attachés à une base, par exemple, dans les bases de données de l'analyseur morphologique arabe de Buckwalter (BAMA) qui traite les mots en arabe comme contenant trois éléments préfixe + base + suffixe. Par exemple, morphèmes d'affixation et de cliticisation nominaux, dans l'exemple le mot فـالـجـزائـريـان\faAl.jazaAŷiriyaAn*et les deux algériens* est analysé en ف+ ال+جـزائـري+ ان,

- la conjonction ف (proclitique),
- l'article défini ال (préfixe),
- la base NTWS جـزائـري,
- et le suffixe ان du duel masculin.

L'exemple ci-dessus serait divisé dans BAMA en أفـي + كـتـب + ونـهـا . Cela met en évidence le problème avec la définition de la base, qui peut être ad hoc et dépendant de l'implémentation. La base peut suivre ou pas un schème. Les bases qui suivent un schème sont des bases qui peuvent être formées en utilisant des morphèmes à schème par exemple كـاتـب\kaAtib*écrivain* dans le mot بـكـاتـبـه\bikaAtibih*par son écrivain;* la base كـاتـب\kaAtib*écrivain* est déduite de la racine كـتـب\ktb\ et du schème فـاعل\faAçil\. Tandis que les bases des mots ne suivant pas un schème (NTWS Non Templatic Word Stem) ne sont pas déductibles de morphèmes à schème. NTWS ont tendance à être des noms étrangers et des termes nominaux empruntés (mais jamais des verbes), par exemple la base جـزائـري\jazaAŷiriy*algérien* dans le mot فـالـجـزائـريـان.

Morphèmes à schème

Les morphèmes à schème sont de trois types qui sont tous aussi nécessaires pour créer une base d'un mot à schème : racines, patrons et vocalisme.

La racine

Le morphème racine est une séquence de trois, de quatre ou très rarement de cinq consonnes (appelées radicaux) (N. Habash 2010). La racine signifie un sens abstrait partagé par tous ses dérivés. En effet, à chaque racine correspond un champ sémantique et à l'aide de différents patrons, on peut générer une famille de mots appartenant à ce champ sémantique, par exemple la racine ك-ت-ب peut engendrer quinze mots autour de la notion de écriture.

Le patron

Le morphème patron est un modèle abstrait dans lequel les racines et vocalisme sont insérés, on représente le patron comme une chaîne de caractères y compris des symboles spéciaux pour marquer l'emplacement des radicaux et vocalismes où seront insérés. On utilise les numéros 1, 2, 3, 4 ou 5 pour indiquer les positions des radicaux et le symbole V est utilisé pour indiquer la position du vocalisme. Par exemple, le patron 1V22V3 indique que le second radical de la racine doit être doublé. Un schème peut inclure des lettres pour les consonnes et les voyelles supplémentaires, par exemple, le patron verbal V1tV2V3

Le vocalisme

Le morphème vocalisme spécifie les voyelles courtes à utiliser avec un schème. Les descriptions traditionnelles de la morphologie arabe incluaient le vocalisme dans le patron. La séparation de vocalisme a été introduite avec l'émergence de schèmes plus sophistiqués qui font abstraction de certains traits flexionnels qui varient constamment avec les patrons complexes, tels que la voix (passif ou actif)

Une base de mot est construite par l'entrelacement des trois types de morphèmes à schème. Par exemple, la base de mot كَـتَـبَ\kataba*écrire* est construite à partir du triplet (ك-ت-ب, 1v2v3, aa) de la racine ك-ت-ب, du patron 1v2v3 et du vocalisme aa

5.2.2 Morphologie fonctionnelle

Dans la morphologie fonctionnelle, on étudie les mots en fonction de leur comportement morpho-syntaxique et morpho-sémantique par opposition à la forme des morphèmes des quels ils sont construits. On distingue trois opérations fonctionnelles: flexion, dérivation et cliticisation. Dans l'Arabe, la distinction entre ces trois opérations est semblable à celle d'autres langues. Ce n'est pas surprenant puisque la morphologie fonctionnelle tend à être une manière indépendante de la langue pour caractériser les mots.

5.3. Flexion des verbes

La morphologie verbale arabe est souvent décrite comme étant très régulière. Elle est un système avec de très peu d'exceptions, presque mathématique. Les verbes se fléchissent suivant l'aspect, le mode, la voix et le sujet (personne, genre et nombre). Les Verbes arabes ont un nombre limité de patrons: 18 patrons de base trilitère et 4 patrons de base quadrilitère (N. Habash 2010; Hamlaoui 2007; Maaluf 1991). Les patrons des six premières lignes et de la dix-neuvième ligne du Tableau 4 sont des patrons de verbes primitifs مـجرد\mujar~ad\ et les autres sont des patrons des verbes augmentés مـزيـد\maziyd\.

La flexion des verbes arabes est aussi affectée par la nature du verbe i.e. s'il est défectueux ou saint, i.e. s'il contient un des caractères de défectuosité ا, و et ي ou pas.

Dans notre travail on a combiné les deux facteurs précédents de morphologie qui régissent la flexion des verbes arabes pour les classifier en un nombre de paradigmes flexionnels.

Patron1	Patron2	Exemple	En français
1a2a3	ya12a3	فَـتَـحَ / يَـفـتَـحُ \fataHa / yaftaHu\	Ouvrir
1a2a3	ya12u3	كَـتَـبَ / يَـكْـتُـبُ \kataba / yak.tubu\	Ecrire
1a2a3	ya12i3	جَـلَـسَ / يَـجْـلِـسُ \jalasa / yaj.lisu\	s'asseoir
1a2u3	ya12u3	حَسُنَ / يَـحْسُنُ	devenir beau
1a2i3	ya12a3	غَضِبَ / يَـغْـضَبُ \γaDiba / yaγ.Dabu\	se fâcher
1a2i3	ya12i3	حَسِبَ / يَـحْسِبُ \Hasiba / yaH.sibu\	Considérer
1a22a3	yu1a22i3	كَـتَّـبَ / يُـكَـتِّـبُ \kat~aba / yukat~ibu\	Dicter
1A2a3	yu1A2i3	كَـاتَـبَ / يُـكَـاتِـبُ \kaAtaba / yukaAtibu\	correspondre avec
'a12a3	yu12i3	أَجْـلَـسَ / يُـجْـلِـسُ \Âaj.lasa / yuj.lisu\	Asseoir
ta1a22a3	yata1a22a3	تَـعَـلَّـمَ / يَـتَـعَـلَّـمُ \taçal~ama / yataçal~amu\	Apprendre
ta1A2a3	yata1A2a3	تَـكَـاتَـبَ / يَـتَـكَـاتَـبُ \takaAtaba / yatakaAtabu\	Correspondre
in1a2a3	yan1a2i3	انْـطَـلَـقَ / يَـنْـطَـلِـقُ \An.Talaqa / yan.Taliqu\	Démarrer
i1ta2a3	ya1ta2i3	اكْـتَـتَـبَ / يَـكْـتَـتِـبُ \Ak.tataba / yak.tatibu\	Déclarer
i12a3a3	ya12a3i3	احْـمَـرَّ / يَـحْـمَـرُّ \AH.mar~a / yaH.mar~u\	Rougir
ista12a3	yasta12i3	اسْـتَـرْجَـعَ / يَـسْـتَـرْجِـعُ \As.tar.jaça / yas.tar.jiçu\	Récupérer
i12aw2a3	ya12aw2i3	اعْـشَـوْشَـبَ / يَـعْـشَـوْشِـبُ \Aç.šawšaba / yaç.šawšibu\	revêtir d'herbe
i12A3a3	ya12A3i3	احْـمَـارَّ / يَـحْـمَـارُّ \AH.maAr~a / yaH.maAr~u\	rougir plus
i12awaw3	ya12awiw3	اجْـلَـوَّذَ / يَـجْـلَـوِّذُ \Aj.law~ada / yaj.law~iadu\	se dépêcher
1a23a4	yu1a23i4	زَخْـرَفَ / يُـزَخْـرِفُ \zax.rafa / yuzax.rifu\	Orner
ta1a23a4	yata1a23a4	تَـزَخْـرَفَ / يَـتَـزَخْـرَفُ \tazax.rafa / yatazax.rafu\	s'orner
i12an3a4	ya12an3i4	احْـرَنْـجَـمَ / يَـحْـرَنْـجِـمُ \AH.ran.jama / yaH.ran.jimu\	se regrouper
i12a3a4a4	ya12a3i4i4	اطْـمَـأَنَّ / يَـطْـمَـئِـنُّ \AT.maÂan~a / yaT.maŷin~u\	se rassurer

Tableau 4 : Les principaux patrons des verbes arabes

Dans les deux colonnes patron1 et patron2, les chiffres représentent les position de caractère de la racine et les caractères A, ', t, n, s, w, y représentent les caractères arabes ا, أ, ت, ن, س, و, ي et les caractères a, u, i représentent les diacritiques َ, ُ, ِ respectivement.

5.4. Flexion des noms/adjectifs

Par rapport aux verbes, la morphologie nominale est beaucoup plus complexe et hétérogène. Les nominaux arabes fléchissent suivant le genre, le nombre, l'état et le cas.

5.4.1 Genre et nombre

L'arabe a deux valeurs du genre: masculin et féminin; et trois valeurs du nombre: singulier, duel et pluriel. Toutefois, en ce qui concerne leur forme, l'histoire est plus complexe. Dans 80% des nominaux arabes les genres fonctionnelles et morphémiques s'accordent, par exemple معـلـم\muç.lim*instititeur* masculin singulier, معـلـمـة\muçalimaħ*institutrice* féminin singulier, معـلـمـون\muçalimuwn*instituteurs* masculin pluriel, معـلـمـات\muçalimaAt*institutrices* féminin pluriel et dans 20% le genre et le nombre fonctionnels ne correspondent pas au genre et nombre morphémiques, Ce qui suit sont certains des patrons les plus communs de désaccord forme-fonction

- Pluriel irrégulier par exemple مـكـتـب\mak.tab*bureau* singulier masculin, مـكـاتـب\makaAtib*bureaux* pluriel masculin,

- Féminin irrégulier par exemple أزرق\Âaz.raq*bleu* masculin singulier, زرقـاء\zar.qaA'*bleue* féminin singulier,
- Incompatibilité du genre de base par exemple عيـن\ςay.n*œil* féminin avec une forme masculin et خـلـيفـة\xaliyfaħ*calife* masculin avec une forme féminin,
- Le singulier du pluriel collectif par exemple تـمـر \tam.r*datte* singulier du pluriel collectif,
- Désaccord complexe par exemple كـتـبـة\katabaħ*écrivains* masculin pluriel mais suivant la forme féminin singulier.

5.4.2 État

Les noms arabes fléchissent suivant l'état qui a trois valeurs : définie, indéfinie et construction.

5.4.3 Cas

Les noms arabes fléchissent suivant le cas qui a trois valeurs : nominatif, accusatif et génitif, le cas nominal est réalisé dans la plupart des cas par les signes diacritiques.

5.5. Dérivation

La morphologie dérivationnelle vise à créer de nouveaux mots à partir d'autres, la dérivation implique un changement de catégorie grammaticale (partie de discours). En Arabe les variantes dérivées viennent généralement d'un ensemble de relations lexicales bien définies (cf. les tableaux de patrons de dérivation du Tableau 5 au Tableau 10), La dérivation d'une forme à partir d'une autre est réalisée par un changement de patron par exemple le verbe كـتـب\katab*écrire* ayant une racine ك-ت-ب\ktb\ et un patron 1a2a3. Pour dériver le nom du participe actif on applique le patron 1A2i3 pour produire la forme كـاتـب\kaAtib*écrivain*.

Dans ce qui suit on dresse les tableaux toutes les dérivations possibles à partir d'un verbe arabe.

N°	Patron arabe	Exemple	Patron
1	فِعَالَـة \fiçaAlaħ\	زِرَاعَة \ziraAçaħ\	1i2aA3aħ
2	فِعَـال \fiçaAl\	إبَـاء \ĂibaA'\	1i2aA3
3	فَـعَلاَن \façalAan\	غَلَيَـان \γalayaAn\	1a2a3Aan
4	فُـعَـال \fuçaAl\	صُدَاع \SũdaAς\	1u2aA3
5	فَـعِيل \façiyl\	رَحِيل \raHiyl\	1a2iy3
6	فُـعْلَة \fuç.laħ\	حُمْرَة \Hum.raħ\	1u23aħ
7	فُـعُولَـة \fuçuwlaħ\	سُهُولَـة \suhuwlaħ\	1u2uw3aħ
8	فَـعَـالَـة \façaAlaħ\	نَـبَـاهَة \nabaAhaħ\	1a2aA3aħ
9	فُـعُول \fuçuwl\	قُـعُود \quçuwd\	1u2uw3
10	فَـعْل \faç.l\	فَـهْم \fah.m\	1a23
11	فَـعَل \façal\	فَـرَح \faraH\	1a2a3

Tableau 5 : Patrons du gérondif (masdar) trilitère

N°	Patron arabe	Exemple	Patron
1	تَـفْـعِـيـل \taf.çiyl\	تَـطْـهِـيـر \taT.hiyr\	ta12iy3
2	تَـفْـعِـلَـة \taf.çilaħ\	تَـوْسِعَـة \taw.siçaħ\	ta12i3aħ
3	فِـعَـال \fiçaAl\	قِـتَـال \qitaAl\	1i2aA3
4	مُـفَـاعَـلَـة \mufaAçalaħ\	مُـخَـاصَـمَـة \mũxaASamaħ\	mu1a23a4aħ
5	إِفْـعَـال \ĂiH.saAn\	إِحْسَان \ĂiH.saAn\	Ăi12aA3
6	تَـفَـعُّـل \tafaç~ul\	تَـقَـدُّم \taqad~um\	ta1a22u3
7	تَـفَـاعُـل \tafaAçul\	تَـطَـايُـر \taTaAyur\	ta1aA2u3
8	إِفْـتِـعَـال \Ăif.tiçaAl\	اِشْتِـرَاك \Aiš.tiraAk\	Ăi1ti2aA3
9	إِفْـعِـلَال \Ăif.çilAal\	اِحْمِـرَار \AiH.miraAr\	Ăi12i3Aa3
10	إِنْـفِـعَـال \Ăin.fiçaAl\	اِنْـطِلَاق \Ain.TilAaq\	Ăin1i2aA3
11	اِسْتِفْـعَـال \Ais.tif.çaAl\	اِسْتِـخْـرَاج \Ais.tix.raAj\	Aisti12aA3
12	اِفْـعِـيـعَـال \Aif.çiyçaAl\	اِحْـدِيـدَاب \AiH.diydaAb\	Ai12iy2aA3
13	فَـعْـلَـلَـة \faç.lalaħ\	وَسْوَسَـة \was.wasaħ\	1a23a4aħ
14	فَـعْـلَال \faç.lAal\	دَحْـرَاج \daH.raAj\	1a23Aa4
15	تَـفَـعْـلُـل \tafaç.lul\	تَـزَحْـلُـق \daH.raAj\	ta1a23u4
16	اِفْـعِـنْـلَال \Aif.çin.lAal\	إِحْـرِنْـجَـام \ĂiH.rin.jaAm\	Ai12in3Aa4
17	اِفْـعِـلَّال \Aif.çilA~al\	اطمـئـنـان \ATmŷnAn\	Ai12i3A~a3

Tableau 6 : Patrons du masdar quadrilitère

N°	Patron arabe	Exemple	Patron
1	مَـفْـعَـل \maf.çal\	مَـنْـظَـر \man.Ďar\	ma12a3
2	مَـفْـعِـل \maf.çil\	مَـوْعِـد \maw.çid\	ma12i3

Tableau 7 : Patrons du masdar mimi

N°	Patron arabe	Exemple	Patron
1	فَـعَّـال \faç~aAl\	نَـصَّـار \naS~aAr\	1a22aA3
2	مِـفْـعَـال \mif.çaAl\	مِـقْـدَام \miq.daAm\	mi12aA3
3	فَـعُـول \façuwl\	وَدُود \waduwd\	1a2uw3
4	فَـعِـيـل \façiyl\	عَظِيم \çaĎiym\	1a2iy3
5	فَـعِـل \façil\	حَـذِر \Haðir\	1a2i3
6	فِـعِّـيـل \fiç~iyl\	سِـكِّـيـر \sik~iyr\	1i22iy3
7	مِـفْـعِـيـل \mif.çiyl\	مِـسْـكِـيـن \mis.kiyn\	mi12iy3
8	فُـعَـلَـة \fuçalaħ\	هُـمَـزَة \humazaħ\	1u2a3aħ
9	فَـاعُـول \faAçuwl\	فَـارُوق \faAruwq\	1aA2uw3
10	فُـعَّـال \fuç~aAl\	كُـبَّـار \kub~aAr\	1u22aA3
11	فَـعَّـالَـة \faç~aAlaħ\	عَلَّامَـة \çalA~amaħ\	1a22aA3aħ
12	فَـاعِـلَـة \faAçilaħ\	رَاوِيَـة \raAwiyaħ\	1aA2i3aħ
13	فُـعْـل \fuç.l\	غُـفْـل \γuf.l\	1u23
14	فَـعُـولَـة \façuwlaħ\	فَـرُوقَـة \faruwqaħ\	1a2uw3aħ
15	مِـفْـعَـل \mif.çal\	مِـحْـرَب \miH.rab\	mi12a3

Tableau 8 : Patrons de l'intensif

N°	Patron arabe	Exemple	Patron
1	أفْـعَل \Âaf.çal\	أحْمَـر \ÂaH.mar\	Âa12a3
2	فَـعْلاء \faç.lAa'\	حَمْـرَاء \Ham.raA'\	1a23Aa'
3	فَـعْلاَن \faç.lAan\	عَطْشَـان \çaT.šaAn\	1a23Aan
4	فَـعْـلَى \faç.laý\	عَطْشَى \çaT.šaý\	1a23aý
5	فَـعَل \façal\	حَسَن \Hasan\	1a2a3
6	فُـعُل \fuçul\	جُـنُب \junub\	1u2u3
7	فُـعَـال \fuçaAl\	شُجَـاع \šujaAç\	1u2aA3
8	فَـعَـال \façaAl\	جَبَـان \jabaAn\	1a2aA3
9	فَـعْل \faç.l\	ضَخْم \Dax.m\	1a23
10	فِـعْل \fiç.l\	صِفْـر \Sif.r\	1i23
11	فُـعْل \fuç.l\	صُلْب \Sul.b\	1u23
12	فَـعِل \façil\	نَـجِس \najis\	1a2i3
13	فَـا عِل \faAçil\	طَـا هِر \TaAhir\	1aA2i3
14	فَـعِـيل \façiyl\	بَـخِيل \baxiyl\	1a2iy3

Tableau 9 : Patrons de l'adjectif

N°	Patron arabe	Exemple	Patron
		Participe actif	
1	فَـا عِل \faAçil\	بَـائِـع \baAŷiç\	1aA2i3
		Participe passif	
1	مَـفْـعُـول \maf.çuwl\	مَـسْلُـول \mas.luwl\	ma12uw3
		Nom de fois\manière	
1	فَـعْـلَـة \faç.laħ\	ضَرْبَـة \Dar.baħ\	1a23aħ
2	فِـعْـلَـة \fiç.laħ\	مِـشْيَـة \miš.yaħ\	1i23aħ
		Elatif	
1	أفْـعَل \Âaf.çal\	أكْـرَم \Âak.ram\	Âa12a3
		Nom de lieu\temps	
1	مَـفْـعَل \maf.çal\	مَـطْبَـخ \maT.bax\	ma12a3
2	مَـفْـعِل \maf.çil\	مَـجْـلِس \maj.lis\	ma12i3
3	مَـفْـعَـلَـة \maf.çalaħ\	مَـسْبَـعَـة \mas.baçaħ\	ma12a3aħ
		Nom d'instrument	
1	مِـفْـعَـال \mif.çaAl\	مِـفْـتَـاح \mif.taAH\	mi12aA3
2	مِـفْـعَل \mif.çal\	مِـبْـرَد \mib.rad\	mi12a3
3	مِـفْـعَـلَـة \mif.çalaħ\	مِـكْـنَـسَة \mik.nasaħ\	mi12a3aħ

Tableau 10 : Patrons des autres catégories dérivées

5.6. Cliticisation

La cliticisation est étroitement liée à la morphologie flexionnelle. Similaire à la flexion, la cliticisation ne change pas le sens fondamental du mot. Cependant, les clitiques sont tous optionnels contrairement aux traits flexionnels qui sont tous obligatoires. En outre, la morphologie flexionnelle est exprimée en utilisant à la fois la morphologie à schème et morphologie concatenative, la cliticisation est exprimée uniquement en utilisant la morphologie concatenative.

Les clitiques arabe s'attachent à un mot de base fléchi dans un ordre précis qui peut être représenté comme suit et en utilisant les noms de catégories générales

[QST+ [CNJ+ [PRT+ [DET+ BASE +PRO]]]]

Comme tous les clitiques sont facultatifs, le mot de base fléchi est valide tel qu'il est. Au niveau le plus profond de cliticisation, on trouve DET, le déterminant (l'article défini) ال et PRO un membre de la catégorie des clitiques pronominaux. Les enclitiques pronominaux peuvent s'attacher à des noms (en tant que possessifs) ou à des verbes et des prépositions (comme des objets). Le déterminant ال ne s'attache pas aux verbes ou les prépositions.

Les enclitiques pronominaux possessifs et le déterminant ne coexistent pas dans les noms. Vient ensuite PRT, la classe des proclitiques particules selon leurs catégories grammaticales, certains de ces clitiques ne s'attachent qu'aux verbes, par exemple, la particule de futur س. Les proclitiques prépositions tels que ب et ك s'attachent généralement à des noms, à des adjectifs et à des particules telles que أنَّ mais jamais à des verbes. Un niveau d'attachement moins profond est la classe CNJ, où l'on trouve les conjonctions و et ف. elles peuvent s'attacher à n'importe quelle CS. Enfin, le niveau le peu profond d'attachement des clitiques QST, est réservé pour la particule interrogative أ qui s'attache au premier mot de toute phrase la transformant en une question.

5.6.1 Extraction des pré-bases

Commençant par le début du mot on extrait les pré-bases possibles. On relève la première lettre et on vérifie si elle forme une pré-base réalisable dans la langue arabe. La vérification se fait en s'assurant de l'existence de la pré-base extraite dans la liste des pré-bases possibles.

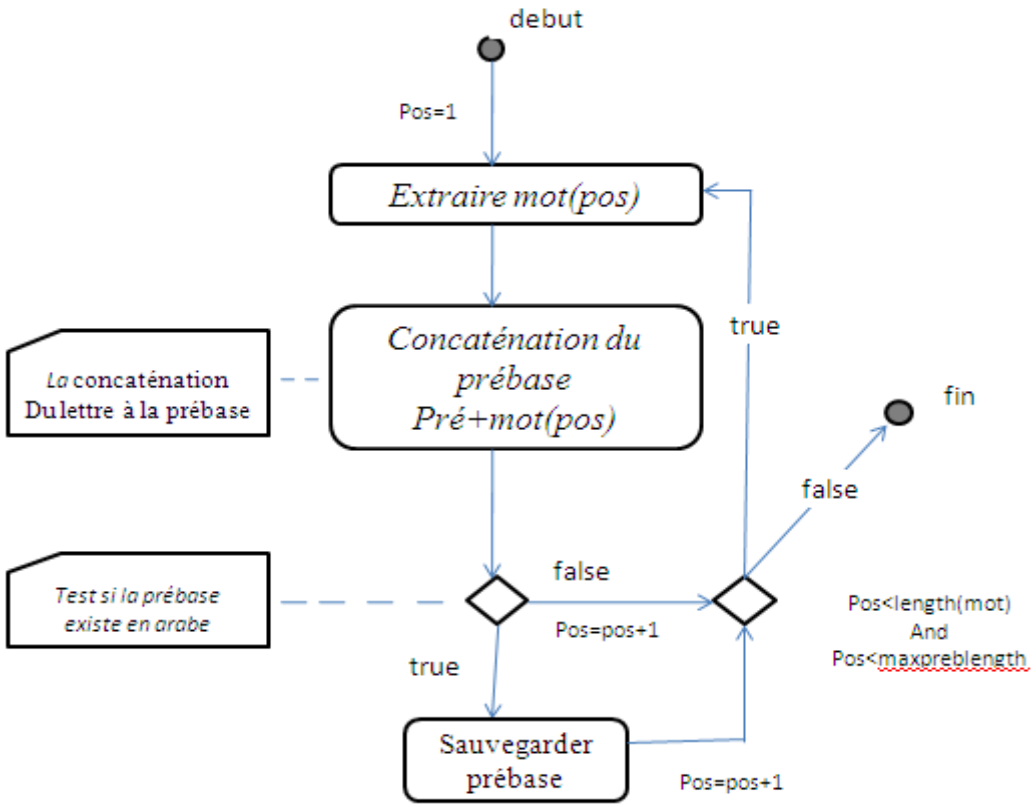


Figure 3 : Diagramme d'activité pour extraction de la pré-base

La liste est préalablement établie et chargée à partir du lexique. On concatène par la suite la lettre suivante et on reprend à nouveau la séquence de vérification. On répète cette procédure jusqu'à atteindre la longueur maximale d'une pré-base ou éventuellement la fin du mot.

5.6.2 Extraction des post-bases

Le procédé est exactement le même que pour les pré-bases, à la différence mineure que cette fois-ci on commence l'extraction par la fin du mot pour lister les post-bases éventuelles du mot.

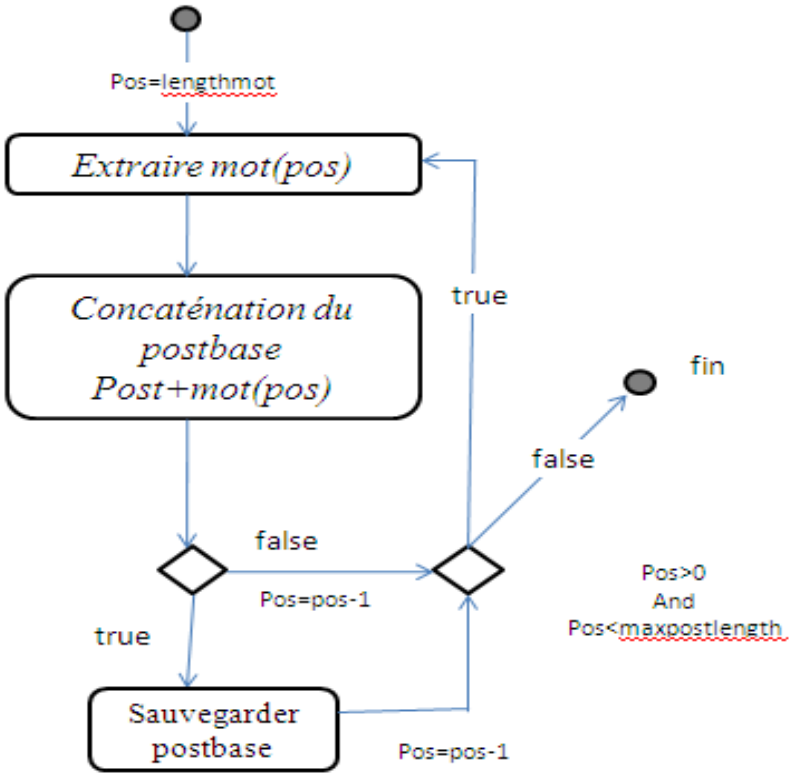


Figure 4 : Extraction des post-bases

5.6.3 Extraction des bases

À l'issu des deux manipulations précédentes on se trouve en possession de deux listings. Le premier contient les pré-bases possibles et le second contient les post-bases possibles.

L'étape suivante consiste donc à trouver les couples de pré- et post-bases compatibles et voir s'ils sont acceptés par la base résultante.

6. La syntaxe arabe

La syntaxe est la discipline linguistique qui s'intéresse à la modélisation de la façon comment les mots sont arrangés pour construire de grandes séquences dans la langue. Alors que la morphologie décrit la structure des mots à l'intérieur, la syntaxe décrit comment les mots se réunissent pour faire des expressions et des phrases. La relation entre la morphologie et la syntaxe peut être complexe en particulier pour les langues morphologiquement riches où de nombreux phénomènes syntaxiques sont exprimés non seulement en termes de l'ordre des

mots, mais aussi par la morphologie. Par exemple en arabe, le sujet d'un verbe a un *cas* nominatif et le modificateur adjectival d'un nom s'accorde avec le *cas* du nom qu'il modifie. La morphologie riche de l'arabe permet d'avoir un certain degré de liberté dans l'ordre des mots puisqu'elle peut exprimer des relations syntaxiques. Cependant, comme dans beaucoup d'autres langues, l'utilisation actuelle de la langue arabe est moins libre, en termes de l'ordre des mots, qu'elle peut en principe l'être. Dans les sections qui suivent on présente un résume de la syntaxe arabe exposé en détail dans (N. Habash 2010).

6.1. Structure de la phrase arabe

L'arabe a deux types de phrases: phrases verbales (V-Sent) et phrases nominales (N-Sent).

6.1.1 Phrase verbale

La structure prototypique d'un V-Sent est verbe-sujet-objet(s). Ceci est exprimé sous différentes formes. La forme la plus basique du V-Sent se compose seulement d'un verbe conjugué avec un sujet pronominal. Le verbe exprime la personne, le genre et le nombre du sujet.

Exemples sur V-Sent composé de verbe-sujet (pronominal)

1. كَتَبَ\kataba\-3ms/*il écrivait*
2. كَتَبْنَا\katab.naA\-1md :1fd :1mp :1fp/*nous écrivions*

Les sujets non-pronominaux apparaissent après le verbe. Le verbe s'accorde avec le sujet en personne (3) et le genre (m ou f), mais pas le nombre, qui est par défaut (s). Un verbe avec un sujet non-pronominal dans un V-Sent n'est jamais (p). Le sujet reçoit le cas du nom. Les objets pronominaux apparaissent dans les suffixes verbaux indépendamment du fait que le sujet est pronominal ou pas.

Exemples sur V-Sent composé de verbe-sujet (non-pronominal)

1. كَتَبَ الوَلَدُ/الاولادُ\kataba Alwaladu/AlAwlAdu\-3ms/*il écrivait/ils écrivaient*
2. كَتَبَت البِنْتُ/البَنَاتُ\katabat Albin.tu/AlbanaAtu\-3fs/*elle écrivait/elles écrivaient*

6.1.2 Phrase nominale

La phrase nominale arabe typique (N-Sent) a la forme Objet-Prédicat/Sujet-Complément (مبتدأ وخبر\mub.tadaÂ wxabar\). Dans N-Sent le plus simple, le sujet est généralement un nom défini, nom propre ou un pronom avec le cas *Nom* et le prédicat est un nom *Nom* indéfini, nom propre ou adjectif qui s'accorde avec le sujet en genre et en nombre.

En plus de la forme basique du prédicat nominal, le prédicat peut être un syntagme prépositionnel (SP). Le prédicat peut être un autre N-Sent, dans cette construction, l'objet du N-Sent supérieur sert comme un sujet. Le prédicat du N-Sent supérieur sera généralement référence au sujet en utilisant une référence pronominale.

Cependant, la structure du prédicat la plus intéressante utilise un V-Sent. Cette construction produit un ordre semblable à SVO (Sujet-Verbe-Objet) en arabe lorsque le sujet du V-Sent de l'intérieur renvoie au sujet du N-Sent supérieur. Ici, le sujet et le verbe sont en accord complet (genre, nombre et personne) par opposition à l'accord en genre et en personne pour le cas dans un V-Sent normal. Cette construction est parfois appelée *phrase complexe*.

Exemples de comparaison entre un N-Sent avec sa variante de V-Sent de base :

1. الأولادُ كَتَبُوا القِصَصَ\AlÂwlAdu katabuwA AlqiSaSa\-3mp/*les élèves écrivaient les histoires*
2. كَتَبَ الأولادُ القِصَصَ\kataba AlÂwlAdu AlqiSaSa\-3ms/*écrit les élèves les histoires*

Il en résulte qu'il y a en arabe, trois types de constructions verbales quand il s'agit de la façon dont le sujet est exprimé : Verbe-Sujet, Sujet-Verbe et Verbe+Sujet. Le sujet du N-Sent supérieur peut être référencé par un argument ou adjuvant à l'intérieur du prédicat V-Sent, tels que l'objet de son V-Sent ou l'objet de l'une de ses prépositions.

Exemples :

1. الكِتَابُ كَتَبَ-هُ الكَاتِبُ\AlkitaAbu kataba-hu AlkaAtibu*l'auteur écrivait le livre*
2. هَذَا البَيتُ كَتَبَ عَنْ-هُ الكَاتِبُ\haðaA Albaytu kataba ςan.-hu AlkaAtibu*l'auteur écrivait à propos de cette maison*

6.2. Structure du syntagme nominal

Le syntagme nominal arabe le plus basique (SP) est un nom ou adjectif avec/sans article défini. On distingue les modificateurs nominaux suivant :

6.2.1 Modification adjectivale

Les adjectifs arabes suivent les noms qu'ils modifient. Les adjectifs et les noms s'accordent toujours en *définitude* et en *cas*. Les adjectifs et les noms rationnels (humain) s'accordent aussi en *genre* et en *nombre*. Les adjectifs du pluriel irrégulier ont un singulier formel et un genre ad hoc basé sur la forme, mais ils sont fonctionnellement pluriels.

Exemple : le mot المَهَرَةُ\Almaharaħu/*intelligents* est *féminin singulier* par la forme mais *masculin pluriel* fonctionnellement.

1. كَاتِبٌ مَاهِرٌ-ms-ms\kaAtibũ maAhirũ*un auteur intelligent*
2. الكُتَّابُ المَهَرَةُ-mp-mp\Alkut~aAbu Almaharaħu*les auteurs intelligents*

Alors que les adjectifs des noms irrationnels (non humain) s'accordent avec les noms en *genre* et en *nombre* quand ces derniers sont singuliers ou duels ; les adjectifs des noms irrationnels pluriels sont curieusement *féminins singuliers*.

Exemples sur les adjectifs des noms irrationnels

1. مَكْتَبٌ جَدِيدٌ–ms-ms\mak.tabũ jadiydũ*un nouveau bureau*
2. مَكَاتِبُ جَدِيدَةٌ-mp-fs\makaAtibũ jadiydaħũ*des nouveaux bureaux*

6.2.2 Construction Idafa

La construction Idafa (construction génitive) est une construction possessive/génitive reliant deux noms. Le premier nom, le possesseur (مضاف\muDaAf\), grammaticalement précède et sémantiquement possède le second nom, le possédé (مضاف إليه\muDaAf Ăliyh\). Le possesseur est en état *construit* et le possédé est dans le cas *génitif*.

Les deux noms forment ensemble un syntagme nominal qui peut être la deuxième partie d'une construction Idafa différente. Cela peut être étendu récursivement créant ce qui est appelé une *chaîne Idafa*. Tous les mots d'une chaîne Idafa, à l'exception du premier mot, doivent être *génitifs* et tous les mots à l'exception du dernier, doivent être en état de *construction*.

Exemples sur la construction et chaine Idafa

1. مَكْتَبُ الـمُدِيـرِ\mak.tabu Almudiyri*bureau du directeur*
2. إبْـنُ عَمِّ جَـارِ نَـائـبِ عَمِيـدِ كُـلِّـيَـةِ الـعُـلـومِ\ib.nu çam~i jaAri naAŷbi çamiydi kul~iyaħi Alçulwmi*cousin du voisin du vice doyen de la faculté des sciences*

En plus des constructions possessives de base, la construction Idafa est utilisée dans de nombreuses constructions linguistiques en arabe comme :

- construction de quantification,
- construction adverbiale pseudo prépositionnelle,
- construction Idafa adjectivale, également connue sous le nom de fausse Idafa.

6.2.3 Construction Tamyiz

La construction Tamyiz (تـمـيـيـز\tam.yiyz*accusatif de spécification*) relie deux noms. Le premier nom, le *spécifié* (الـمُمَيَّز\Almumay~az\), dirige et gouverne le deuxième, le *spécificateur* (الـمُمَيِّز\Almumay~iz\) qualifie le premier nom. Le *spécificateur* est toujours singulier en nombre et accusatif dans le cas. En arabe, le Tamyiz est utilisé dans une variété des constructions linguistiques telles que :

- le comparatif et le superlatif,
- la spécification de mesure,
- quelques constructions numériques et
- la spécification de type

6.2.4 Construction d'apposition

Une construction d'apposition بـدل\badal\ concerne deux syntagmes nominaux qui se réfèrent à la même entité. Les têtes des deux syntagmes nominaux s'accordent au *cas*, par exemple, الـرئـيس الـجـزائـري، عبد الـعـزيـز بـوتـفـلـيـقـة\Alraŷiysu AljazaAŷiriyu, çab.duAlçaziyzi buwtaf.liyqaħ*le président algérien Abdelaziz Bouteflika*. Une construction d'apposition très commune en arabe implique le pronom démonstratif, qui habituellement précède le nom qu'il modifie bien qu'il puisse également le succède : هـذا الـكـتـاب\haðaA AlkitaAb*ce livre*.

6.2.5 Clauses relatives

Les clauses relatives modifient le nom qui les dirige. Si le nom de tête est défini, la clause relative (جـمـلـة وصل\jum.laħ waS.l\) est introduite et dirigée par un pronom relatif (اسم مـوصـول\Ais.m mawSuwl\). Lorsqu'il est présent, le pronom relatif s'accorde avec le nom qu'il modifie en *genre* et en *nombre* suivant les règles de l'accord d'adjectif (l'irrationalité obtient un accord exceptionnel).

Si le nom de tête est *indéfini*, la clause relative (appelée dans ce cas 'phrase adjectivale') n'est pas introduite avec un pronom relatif.

La clause relative *définie* et précédée par un pronom relatif peut en arabe, se présenter seule comme un syntagme nominal.

Exemples :

1. الـكِـتَـابُ الَّـذِي أُحِبُّ-هُ\AlkitaAbu Al~aðiy ÂuHibu-hu*le livre que j'aime*
2. كِـتَـابٌ أُحِبُّ-هُ\kitaAbũ ÂuHib~u-hu*un livre que j'aime*

6.2.6 Arguments nominaux

Les noms verbaux en arabe, tels que les noms déverbaux (مصدر\maS.dar*gérondif*) et les participes actifs (اسم الفاعل\Ais.m AlfaAçil\), se comportent comme des verbe dans le sens qu'ils peuvent prendre un argument d'objet *accusatif* et d'autres modificateurs verbaux.

Leur forme nominale leur permet en outre de participer à certaines des constructions nominales discutées précédemment, comme Idafa.

Exemple :

1. مَعْرِفَةُ الرَجُلِ الحَقِيقَةَ \maç.rifaħu Alrajuli AlHaqiyqaħa*le savoir de l'homme à la vérité*

6.3. Syntagme prépositionnel

Le syntagme prépositionnel arabe consiste en une préposition suivie d'un syntagme nominal. La tête du syntagme nominal est dans le cas *génitif*.

Exemple :

1. فِي البَيْتِ\fiy Albay.ti*dans la maison*

6.4. Corpus arborés arabes

Un corpus arboré (ou treebank en anglais) est une collection de phrases syntaxiquement analysées et vérifiées manuellement. Ils représentent une ressource très cruciale pour la construction et l'évaluation des parseurs statistiques. Les annotations des corpus arborés riches ont été aussi utilisées dans une variété d'applications telles que la segmentation, la diacritisation, étiquetage morpho-syntaxique, désambiguïsation morphologique, segmentation des syntagmes et étiquetage des rôles thématiques (sémantiques).

Sous la contrainte temps, la création des corpus arborés s'affronte au compromis entre la richesse linguistique et la taille du corpus. Ceci est particulièrement le cas pour les langues morpho-syntaxiquement complexes telles que l'arabe ou le tchèque (N. Habash 2010). Les représentations linguistiquement riches fournissent plusieurs caractéristiques linguistiques qui peuvent être utiles pour une variété d'applications.

Pour le cas de l'arabe, il existe deux efforts importants pour les corpus arborés avec une riche annotation :

- corpus arborée arabe de l'université de Pennsylvanie PATB (Penn Arabic TreeBank)
- corpus arboré arabe de dépendance de Prague PADT (Prague Arabic Dependency Treebank)

Les deux efforts emploient des représentations complexes et linguistiquement très riches qui nécessitent un grand entrainement humain. La quantité de détails spécifiés dans les représentations est impressionnante. Le PATB fournit non seulement la segmentation, les étiquètes morpho-syntaxiques complexes et la structure syntaxique; Il fournit également des catégories vides, la diacritisation, les choix de lemmes et quelques étiquètes sémantiques. Ces informations permettent une recherche scientifique importante dans les applications du TAL arabe; cependant, une grande partie de cette annotation riche est actuellement inutilisée dans la recherche de l'analyse syntaxique arabe, car elle est généralement considérée comme étant dérivée de la sortie de cette analyse elle-même. Par exemple, le cas nominal, qui peut être

déterminé pour l'analyse syntaxique de référence à haute précision, ne peut pas être prédit bien dans une étape pré-analyse syntaxique d'étiquetage morpho-syntaxique.

Pour résoudre ce problème, un troisième corpus arboré celui de l'université de Columbia, CATiB (Columbia Arabic Treebank), a été introduit dans le but d'accélérer l'annotation grâce à la simplification de la représentation.

7. La sémantique arabe

La sémantique est l'étude du sens des expressions linguistiques. La quantité de recherche scientifique dans les modèles computationnels de la sémantique est beaucoup plus petite que d'autres domaines du TAL. Cela est peut-être dû à sa plus grande complexité et subtilité. La recherche sur la sémantique en TAL arabe n'est pas différente.

7.1. Propbank arabe

Un Propbank (Proposition Bank) est un type de corpus annoté sémantiquement. Propbank annotent les propositions et leurs arguments sous la forme d'informations d'argument-prédicat et d'étiquettes de rôle sémantiques au dessus d'un corpus arboré syntaxique existant. Un Propbank arabe APB (Arabic PropBank) a été développé à l'Université de Colorado suivant une approche similaire à celle utilisée dans le développement des propbanks anglais et chinois. L'APB est construit au dessus de la structure syntaxique de PATB et s'y conforme.

L'APB a également accès aux étiquettes-tirets sémantiques et des annotations des lemmes présents dans le PATB.

Propbanks définit de façon cruciale un inventaire des framesets pour chaque verbe. Un frameset spécifie la signification du verbe prédictif et le nombre et les rôles de ses arguments. Les auxiliaires, qui étendent le sens de la phrase mais ne sont pas essentiels pour le verbe prédicatif, ne sont généralement pas inclus dans un frameset.

7.2. Wordnet arabe

Un wordnet est une base de données lexicale lisible par machine qui regroupe les mots en ensemble de synonymes appelés *synset*. Chaque *synset* peut être considéré comme représentant d'un sens de mot unique (sens ou concept). Un wordnet fournit généralement des définitions et des exemples généraux pour le *synset* et inclut les relations sémantiques entre eux. Les relations sémantiques, qui incluent entre autres l'hyponymie et l'hyperonymie, permettent à un wordnet d'être interprété hiérarchiquement comme une ontologie/taxonomie lexicale.

Le wordnet de Princeton PWN (Princeton WordNet) (Fellbaum 2005) est le premier wordnet créé. Il a été suivi par de nombreux projets et extensions similaires, dont le plus notable est EuroWordNet (EWN). Plusieurs projets de wordnet ont été coordonnés pour inclure des liens croisés. Cela leur permet d'être utilisés non seulement comme thésaurus monolingues informatiques sophistiqués, mais aussi comme dictionnaires.

Le projet Arabic WordNet (AWN) (El-Kateb et al. 2006) a commencé en 2006 grâce à la collaboration de plusieurs universités et entreprises. AWN est basé sur la conception et le contenu de PWN (Fellbaum 2005). Les synsets arabes sont jumelés à des synsets dans le WordNet de Princeton et sont mappables au synset dans l'ontologie de haut niveau SUMO (Suggested Upper Merged Ontology).

L'AWN a été utilisé comme la référence lexicale pour évaluer les systèmes de désambiguïsation de sens de mot arabe (WSD : Word Sens Disambiguation). Dans WSD, les

mots sont étiquetés avec leur signification particulière dans le contexte en utilisant des définitions de sens dans une ressource lexicale prédéfinie.

8. L’extraction d’information et de connaissance

L’extraction de connaissances à partir de textes est un processus non trivial qui construit un modèle de connaissances valide, nouveau, potentiellement utile et au final compréhensible à partir de textes bruts. Elle est un processus à l’intersection de plusieurs domaines dont notamment celui l’apprentissage, mais également celui du traitement automatique des langues, ou encore de la recherche d’informations et de l’extraction d’information (Toussaint 2004). Dans cette thèse, on penche sur l’extraction de relations sémantiques qui est un sujet de recherche de longue date dans le traitement du langage naturel et a été utilisé pour aider, entre autres, l'acquisition de connaissances, la recherche d'informations et de réponse aux questions. Il a également reçu beaucoup d'attention dans le domaine médical et biomédical.

9. Conclusion

Dans ce chapitre, on a essayé de donner un tour d’horizon sur le TAL arabe, car des notions et des définitions très utiles aux lecteurs ont été y exposées. L’objectif est de procurer le lecteur des notions requises à la compréhension du reste de la thèse. Le chapitre a manqué de parler sur les outils de traitement automatique de la langue arabe. Ces outils ont fleuri pendant la dernière décennie. A cet effet, on recommande le lecteur de référer à la référence (N. Habash 2010) pour une lecture assez exhaustive. Dans les chapitres qui suivent, on tente de focaliser sur l’extraction de connaissances à partir du texte arabe.

Chapitre II

Reconnaissance des entités nommées

1. Introduction

Dans ce chapitre on aborde le sujet des entités nommées, leur reconnaissance dans un texte brut et leur attribution de types. Ce chapitre sert comme un background théorique pour le chapitre V, car dans ce dernier on mettra en œuvre toutes les notions théoriques exposées dans le chapitre courant.

Comme il le note Ehrmann dans (Ehrmann 2008), le traitement des entités nommées fait actuellement figure d'incontournable en Traitement Automatique des Langues. Apparue au milieu des années 1990 à la faveur des dernières conférences MUC (Message Understanding Conferences), la tâche de reconnaissance et de catégorisation des noms de personnes, de lieux, d'organisations, etc. apparaît en effet comme fondamentale pour diverses applications participant de l'analyse de contenu et nombreux sont les travaux se consacrant à sa mise en œuvre, obtenant des résultats plus qu'honorables. Fort de ce succès, le traitement des entités nommées s'oriente désormais vers de nouvelles perspectives avec, entre autres, la désambiguïsation et une annotation enrichie de ces unités.

2. De quoi s'agit-il

Le cours de l'histoire, ou plutôt de la recherche, a voulu que l'on désigne un certain nombre des unités linguistiques de niveaux différents sous le nom d'*Entités Nommées* (EN) (*named entities* en anglais). Ces dernières correspondent traditionnellement à l'ensemble des noms propres présents dans un texte, qu'il s'agisse de noms de personnes, de lieux ou d'organisation, ensemble auquel sont souvent ajoutées d'autres expressions comme les dates, les unités monétaires, les pourcentages et autres. Contemporain des travaux en Extraction d'Information initiés au début des années 1990, le traitement des entités nommées s'articule en deux processus :

- identification ou reconnaissance de ces unités dans les textes tout d'abord,
- catégorisation ou typage selon des catégories sémantiques larges prédéfinies ensuite.

3. Les entités nommées

3.1. La quantité d'information dans les entités nommées

Du point de vue humain, la contribution des EN dans une communication facile et plus précise est évidente. Cependant, on aura besoin d'une preuve scientifique pour émettre la même affirmation. Dans le contexte de TAL stochastique on rappelle la théorie de l'information de Shannon qui stipule que la quantité d'information (self-information) contenu dans un événement x est définie par la quantité de surprise que cet événement peut apporter. Par exemple, si une personne A informe une personne B en un vendredi que '*demain c'est samedi*' donc la quantité de surprise est zéro. Alors que, si A informe B que '*le Pape Benoit*

XVI s'est converti à l'islam' alors la quantité de surprise sera grande (Benajiba 2009). Ainsi Shannon explique que la quantité d'information d'un événement *x* est inversement proportionnelle à sa probabilité d'occurrence. Et par conséquent, elle peut être exprimée par la formule de l'équation :

$$I(x) = -\log_2\big(p(x)\big) \text{ où}$$

$I(x)$: la self-information de x et

$p(x)$: la probabilité d'occurrence de l'événement x

Benajiba a mené dans (Benajiba 2009) une expérimentation sur les EN arabes afin de calculer leurs self-informations. Cette expérimentation a été effectuée en plusieurs étapes comme suit :

- Segmenter et annoter en POS (catégories syntaxiques) un corpus déjà annoté pour la tâche de reconnaissance des EN arabe,
- Calculer les probabilités d'occurrence pour les catégories : EN, Verbes, Noms communs et Mots vides.
- Et enfin, appliquer la formule citée ci-dessus pour calculer la self-information pour chaque catégorie syntaxique citée dans le deuxième point.

La Figure 5 illustre les résultats obtenus dans cette expérimentation. Elle montre que la seule catégorie syntaxique qui dépasse les EN (représentant 11% du corpus de test de l'étude) en terme de quantité d'information c'est bien les verbes.

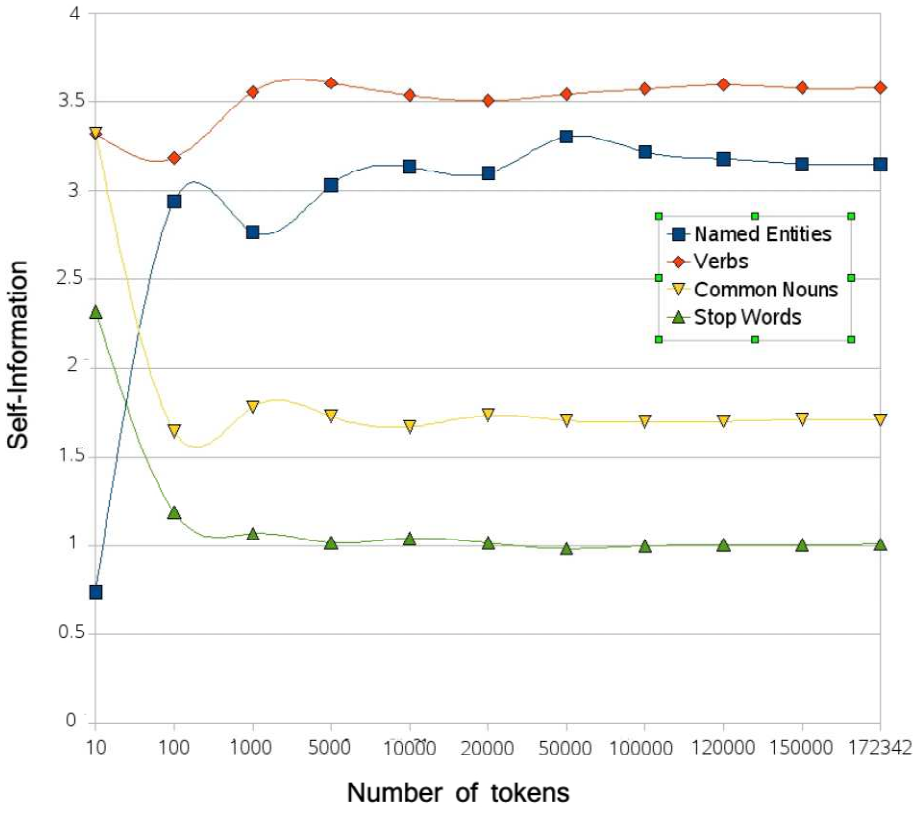


Figure 5 : Quantité d'informations des entités nommées par rapport à d'autres catégories syntaxiques (Benajiba 2009)

3.2. Discussion linguistique

En grammaire, le nom propre est en général considéré comme une sous-catégorie du nom et se distingue du nom commun. Ainsi, un nom commun est un nom employé pour designer tous les éléments d'un même ensemble. Par exemple, *animal*, *poème*, *pièce de théâtre*. Le nom commun dispose d'une définition et d'une signification et il est utilisé en fonction de cette signification. Par exemple, le nom commun *cuillère* dispose d'une définition ; et le fait d'évoquer cette définition permet à chacun d'imaginer à quoi ressemble une cuillère (W. Zaghouani 2009).

Concernant les noms propres, Jonasson dans (Jonasson 1994) propose trois définitions de leurs sens :

1. un nom propre est un prédicat de dénomination : il ne décrit pas l'objet dénoté, mais lui colle une étiquette, par exemple telle fille *est nommée Anissa*.
2. le nom propre est vide de sens puisqu'il permet de référer sans désigner.
3. le sens du nom propre est une description du référent, soit il a un sens réduit à des traits sémantiques généraux comme la distinction féminin / masculin, animé / non animé, soit il dispose d'un sens fort et il permet d'identifier clairement un référent.

Enfin Boulanger et Cormier dans (Boulanger and Cormier 2001), proposent la définition suivante : *le nom propre fait partie des éléments de nature langagière auxquels recourent les locuteurs pour produire des discours et pour construire leur image du monde ainsi que celle des réalités qui les entourent*. Ainsi, le nom propre réfère principalement à une entité unique que cela soit pour représenter des objets, des personnes, des lieux géographiques, des marques déposées ou même des événements (W. Zaghouani 2009).

Du point de vue conceptuel sémantique, les noms propres s'appuient sur les réflexions qui établissent un lien entre le langage comme ensemble de symboles signifiants et les objets ou concepts du monde réel que le langage référence. De ce point de vue, diverses théories évoquent un lien reposant, selon, sur le sens, la dénotation, la référence, la désignation, etc.

D. Nouvel dans (Nouvel 2012) reprend la thèse du mathématicien Frege qui est le premier à établir une distinction claire entre le sens et la référence. La référence pointe vers un concept, qui peut correspondre à un objet du monde réel. De manière plus abstraite, le sens est un mécanisme par lequel un signe (symbole, nom propre par exemple) peut désigner une ou plusieurs références. Il peut y avoir plusieurs sens désignant une même référence ou a contrario certains sens ne désignant aucune référence.

En outre, Frege tient également compte du fait que le sens est nécessairement lié à une représentation individuelle, chaque humain interprétant les signes selon son expérience personnelle. Il doit donc exister une convention permettant à plusieurs individus d'attribuer un sens similaire à des expressions complexes du langage naturel.

3.3. Propos définitoires

Ehrmann dans (Ehrmann 2008) expose différentes définitions de l'objet *Entité Nommée* en tant que objet linguistique et objet TAL. Cette liste de propos vient majoritairement des compagnes d'évaluation de la tâche de reconnaissance d'entités nommées. On énumère dans cette section les propos définitoires les plus distinguées à notre avis. On préfère laisser le texte d'origine (en anglais) de quelques définitions comme l'a fait l'auteur de la référence.

Définition des conférences MUC:

On the level of entity extraction, Named Entities (NE) were defined as proper names and quantities of interest. Person, organization, and location names were marked as well as dates, times, percentages, and monetary amounts

Definition des campagnes CoNLL :

Named entities are phrases that contain the names of persons, organizations and locations

Définition de T. Poibeau dans son ouvrage sur l'extraction d'information :

On appelle traditionnellement entités nommée (de l'anglais named entity) l'ensemble des noms de personnes, d'entreprises et de lieux présents dans un texte donné. On associe souvent à ces éléments d'autres syntagmes comme les dates, les unités monétaires ou les pourcentages repérables par les mêmes techniques à base de grammaires locales.

Définition du National Institute of Standards and Technology :

Named Entity: a named object of interest such as a person, organization, or location

Définition de S. Sekine:

The names of particular things or classes, and numeric expressions is regarded as an important component technology for many NLP applications.(...) the term Named Entity includes names (which is the narrow sense of Named Entity) and numeric expressions. The definition of this Named Entity is not simple, but, intuitively, this is a class that people are often willing to know in newspaper articles.

Définition de N. Friburger dans (Friburger 2002) :

En fait il semble difficile de délimiter les noms propres des autres noms; il y a une continuité entre l'ensemble des noms propres et l'ensemble des noms communs. Les informaticiens qui travaillent dans le domaine de l'extraction d'information, ont abordé le problème de manière pragmatique. Ils ont défini la notion d'entités nommées pour regrouper tous les éléments du langage définis par référence : les noms propres au sens classique, les noms propres dans un sens élargi mais aussi les expressions de temps et de quantités

Définition de la campagne ESTER

Même s'il n'existe pas de définition standard, on peut dire que les EN sont des types d'unités lexicales particuliers qui font référence à une entité du monde concret dans certains domaines spécifiques notamment humains, sociaux, politiques, économiques ou géographiques et qui ont un nom (typiquement un nom propre ou un acronyme)

Définition d'ATALApédie[15] :

Les entités nommées désignent l'ensemble des noms de personnes, de lieux, d'entreprises, etc. contenues dans un texte. On ajoute souvent à ces éléments les dates et d'autres données chiffrées. Par extension, les entités désignent parfois les éléments de base pour une tâche donnée (par exemple, les noms de gènes dans le cadre de l'étude des textes de biologie). (...) Ces séquences référentielles sont primordiales pour beaucoup d'applications linguistiques, que ce soit la recherche ou l'extraction d'information, la traduction automatique ou la compréhension de textes

[15] Association pour le Traitement Automatique des Langues (organisme français) cf. l'url http://www.atala.org/

4. Le rôle des entités nommées dans les applications TAL

4.1. Recherche d'information

Il s'agit d'identifier et de récupérer des documents pertinents à partir d'un ensemble de données en fonction d'une requête d'entrée. Une étude menée par (Guo et al. 2009) a révélé qu'environ 71% des requêtes dans les moteurs de recherche contiennent des EN. La recherche d'information peut bénéficier de la reconnaissance et classification des entités nommées (REN) en deux phases : premièrement, la reconnaissance des EN dans la requête d'une part; et reconnaître les EN dans les documents recherchés d'autre part, puis extraire les documents pertinents en tenant compte de leurs EN classées et comment elles sont reliées à la requête. Par exemple, le mot الـجزيـرة\Aljaziyrah*Aljazeera* peut être reconnu comme un nom d'organisation ou un nom correspondant au mot *île*; La classification correcte facilitera l'extraction des documents pertinents contenant ce mot (Shaalan 2014).

4.2. Système de question-réponse

C'est très similaire à la recherche d'informations mais avec des résultats plus sophistiqués. Un système de question-réponse prend en entrée des questions en langage naturel et donne en retour des réponses concises et précises. La tâche de REN peut être utilisée dans la phase d'analyse de la question afin d'y reconnaître les EN qui aideront plus tard à identifier les documents pertinents et à construire la réponse à partir des passages pertinents. Par exemple, l'EN الـشرق الأوسط\Alšar.q AlÂw.saT*Moyen-Orient* peut être classée comme un nom d'organisation (par exemple, un journal) ou comme un nom de lieu selon le contexte. Par conséquent, la classification correcte pour l'EN aidera à cibler le groupe pertinent de documents qui répondent à la requête d'entrée. En outre, les systèmes de question-réponse pourraient bénéficier substantiellement de la REN, parce que la réponse à de nombreuses questions factoïdes[16] (ou factuelles) implique des EN. Par exemple, la réponse aux questions de type مـن هـو/مـن هـي\mn. huwa/mn. hiya*Qui* concerne les personnes et les organisations, أيـن\Âay.na*Où* concerne les lieux et مـتى\mataý*Quand* implique les expressions temporelles (Shaalan 2014).

4.3. Traduction automatique

C'est la tâche de traduire automatiquement un texte d'une langue naturelle à une autre. Les EN doivent faire l'objet d'une attention particulière pour décider quelles parties d'une EN doivent être sémantiquement traduites et quelles parties doivent être phonétiquement translitérées. Souvent, ceci est dépendant du type de l'EN. Par exemple, les noms de personnes ont tendance à être translittérés[17]. Pour un nom de lieu, la partie du nom et la partie de la catégorie (par exemple montagne) est généralement translittérée et traduite respectivement. Les noms d'organisation sont complètement différents dans le sens que la plupart des constituants sont traduits (e.g. Nations Unies). La qualité du système de REN joue

[16] Dans les systèmes question-réponse on classifie les questions en plusieurs catégories : factoïde, booléenne, définition, cause/conséquence etc. Une question factoïde ou factuelle requiert une réponse sous forme d'EN, par exempe *Quelle est la capitale de l'Algérie ?*

[17] La translittération entre les langues qui utilisent des alphabets et des systèmes de son similaires est très simple. Cependant, la translittération des EN entre l'arabe et l'anglais et langues latines en général (appelée romanisation) est une tâche non triviale, principalement en raison des différences dans leurs systèmes de son et d'écriture (Al-Onaizan and Knight 2002).

un rôle important dans la détermination de la qualité globale du système de traduction automatique et, par conséquent, la traduction de l'EN est essentielle pour la plupart des systèmes d'application multilingues. En outre, la traduction des EN est très importante pour d'autres applications telles que la recherche d'informations multilingues pour extraire des EN nouvellement introduites à partir du Web et des documents d'actualité et la mise à jour régulière de la liste des paires de traduction des EN.

4.4. Catégorisation automatique de textes

Les catégories de résultats d'un moteur de recherche peuvent exploiter le REN en les classant en fonction du ratio d'entités que chaque cluster contient. Cela améliore le processus d'analyse de la nature de chaque catégorie et améliore également l'approche de catégorisation en termes de fonctionnalités sélectionnées. Par exemple, les EN d'expressions temporelles ainsi que de lieux peuvent être utilisés comme des facteurs qui donneront une indication de *quand* et *où* les événements mentionnés dans un groupe de documents se sont produits.

4.5. Système de navigation

Ces systèmes, qui facilitent la navigation à l'aide de cartes numériques, jouent maintenant un rôle important dans nos vies. Ils fournissent des indications, des informations sur les lieux voisins éventuellement liés à d'autres ressources en ligne et les conditions de trafic routier. Dans ces systèmes, les points d'intérêt (également appelés *waypoints*) sont des EN qui sont stockés dans une base de données avec leurs coordonnées géographiques. Ils font référence à des zones d'intérêt qui sont généralement d'importance pour, entre autres, les touristes, les visiteurs et les sauveteurs, en donnant la localisation des lieux tels que les parkings, les magasins, les hôpitaux, les restaurants, les universités, les écoles, les repères et ainsi de suite.

5. Les typologies des EN

La typologie des EN tente de prédéfinir des catégories et des types dans les quels on classifie les unités textuelles concernées par le processus de la REN. Le spectre des types d'EN va de la classification peu détaillée des conférences MUC dans les années 90 jusqu'aux typologies adoptées par les projets et campagnes d'évaluation REN les plus récents[18]. Ces derniers sont caractérisés par une décomposition plus fine des catégories d'EN grossièrement définies au début et de l'ajout de nouvelles catégories. Dans ce qui suit dans cette section on énumère les typologies trouvées dans la littérature pendant l'accomplissement des travaux de cette thèse.

5.1. Typologie des conférences MUC

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

Les participants étaient alors invités à développer un système qui permet l'extraction du plus grand nombre d'informations possibles sur des entités bien précises. Par la suite une évaluation est conduite en suivant la même procédure pour l'ensemble des participants. Les systèmes d'extraction participants ont été évalués sur des domaines tels que le terrorisme en Amérique Latine lors de MUC-3 (MUC 1991) et de MUC-4 (MUC 1992). Lors de MUC-5 (MUC 1993), le domaine était la fusion d'entreprises et la fabrication de circuits

[18] A titre d'exemple le projet Quaero.

électroniques. Lors de MUC-6, le domaine était les changements de dirigeants des entreprises (MUC 1995). Enfin, MUC-7 a porté sur les accidents d'avion.

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

La conférence MUC-7 a distingué trois types d'entités à reconnaître et à catégoriser, soit ENAMEX, NUMEX et TIMEX. La Figure 6 montre ces trois grandes catégories et leur limite par rapport à la grande famille des EN. On remarque clairement qu'il y a une majeure partie d'EN qui n'est pas couverte par la classification MUC.

- Les entités de type ENAMEX sont composées des noms propres, des sigles et des abréviations. Les entités ENAMEX se divisent en trois catégories :
  - personnes : les noms de personnes ou de familles,
  - noms de lieux : ce sont des lieux définis géographiquement ou politiquement comme les villes, provinces, rivières, montagnes,
  - organisations : cette catégorie inclut les noms de gouvernements, sociétés, et autres entités organisationnelles.
- Les entités NUMEX rassemblent les nombres et les pourcentages, les unités de mesures, les devises,
- Enfin, les entités TIMEX couvrent les expressions de temps et les dates.

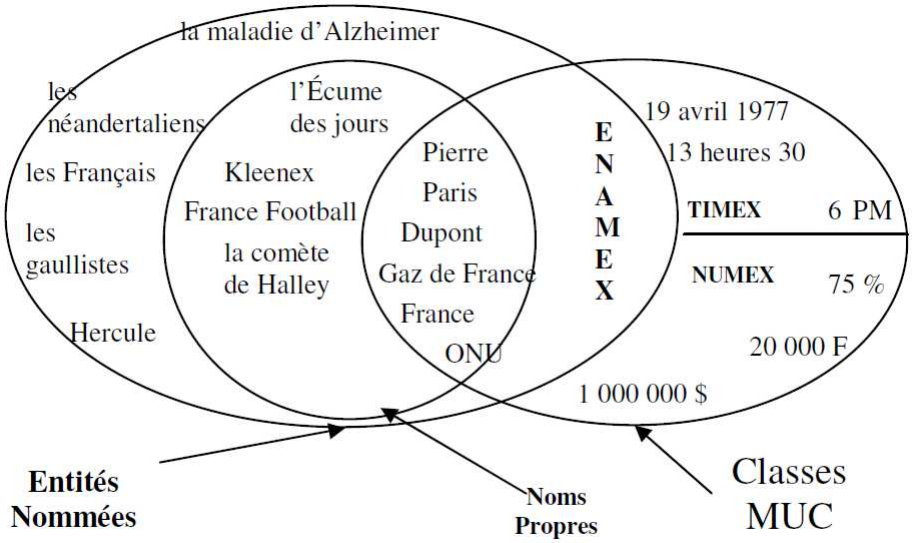


Figure 6 : Les entités nommées vs la classification MUC (Daille et al. 2000)

5.2. Typologie de Paik

La classification de Paik et al. citée par (Maurel et al. 2011) regroupe ensemble les entités nommées et les entités temporelles. Cette approche a été mise au point à la suite de l'analyse d'un corpus du Wall Street Journal. Elle comporte 30 catégories divisées en 9 classes (W. Zaghouani 2009):

- Géographique : villes, ports, aéroports, îles, comtés ou départements, provinces, pays, continents, régions, fleuves, autres noms géographiques.
- Affiliation : religions, nationalités.
- Organisation : entreprises, types d'entreprises, institutions, institutions gouvernementales, organisations.

- Humain : personnes, fonctions.
- Document : documents.
- Équipement : logiciels, matériels, machines.
- Scientifique : maladies, drogues, médicaments.
- Temporelle : dates et heures.
- Divers : autres noms d'entités nommées.

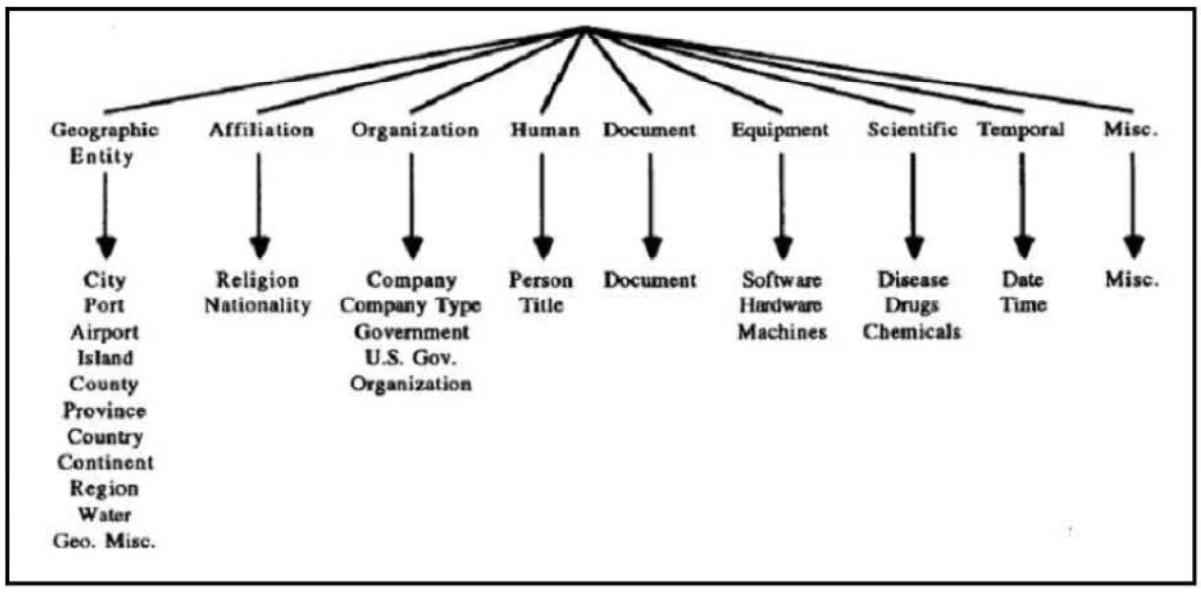


Figure 7 : Typologie de Paik et al. citée par (Maurel et al. 2011)

5.3. Typologie de Bauer

Bauer (1985) cité par (Grass, 2000) a présenté une autre catégorisation du nom propre dans le cadre de ses recherches sur la traduction. Sa classification n'inclut pas les entités temporelles et se divise en six classes et chaque classe comporte plusieurs catégories :

- Anthroponymes : les personnes individuelles ou les groupes :
  - patronymes,
  - prénoms,
  - pseudonymes,
  - gentilés,
  - hypocrites,
  - ethnonymes,
  - groupes musicaux modernes,
  - ensembles artistiques et orchestres classiques,
  - partis et
  - organisations.
- Toponymes : les noms de lieux :
  - pays, villes,
  - microtoponymes,
  - hydronymes,
  - oronymes,
  - installations militaires.

- Ergonymes : les objets et les produits manufacturés et par extension les marques, entreprises, établissements d'enseignement et de recherche, titres de livres, de films, de publications, d'œuvre d'art.

5.4. Typologie de la campagne ESTER 2

La campagne d'évaluation ESTER 2 (Evaluation des Systèmes de Transcription Enrichie d'Émissions Radiophoniques 2) s'est déroulée de janvier 2008 à avril 2009, dans la continuité de la première campagne ESTER. Ces campagnes visaient à évaluer les performances des systèmes de transcription de la parole, les performances des systèmes de segmentation en tours de paroles, et la capacité à extraire automatiquement des informations, en particulier les entités nommées. Cette troisième tâche, à laquelle se sont attelés 7 participants dans le cadre d'ESTER 2, était divisée en deux sous-tâches: la détection d'entités nommées sur transcriptions de référence et sur transcriptions automatiques (Brun and Ehrmann).

Dans le cadre d'ESTER2, il s'agissait d'extraire et de catégoriser des mentions directes d'EN, selon un guide d'annotation comprenant 7 catégories principales et 38 sous-catégories. La Figure 8 montre la hiérarchie des catégories et sous-catégories de la compagne d'évaluation ESTER 2.

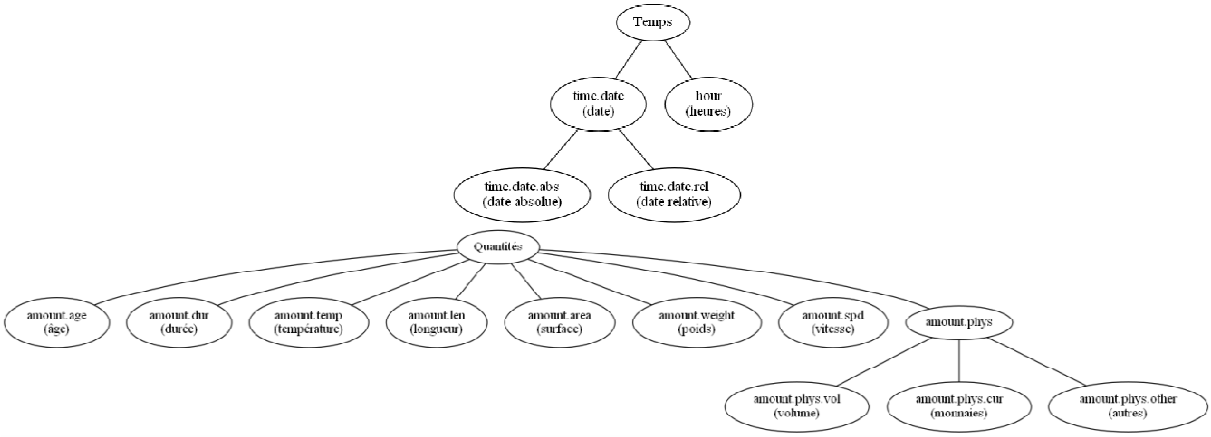


Figure 8 : Taxonomie des sept catégories d'EN de la compagne ESTER 2 (ESTER2 2007)

La principale instruction d'annotation est de considérer les entités en contexte, avec la prise en compte des phénomènes d'ambiguïtés et de métonymie: par exemple, selon les contextes, *Charles de Gaulle* doit être annoté en tant que personne (le président), véhicule (le porte-avion) ou encore lieu (l'aéroport).

L'annotation des noms de personnes inclut celle des fonctions et l'annotation des expressions temporelles couvre pour sa part un large éventail de possibilités, des classiques *Lundi matin* aux plus complexes *Il y a un peu moins de trois jours environ*. Par ailleurs, dans la mesure où l'extraction d'entités est réalisée sur des transcriptions de la parole, certains phénomènes propres à l'oral (hésitations ou répétitions) doivent être inclus dans les annotations (<pers.hum> Jacques heu Chirac </pers.hum>). Ces directives d'annotation spécifiques, combinées au nombre important de catégories à prendre en compte, complexifient la tâche d'annotation. En effet, les quantités de type âge et durée sont particulièrement difficiles à distinguer des expressions temporelles, tout comme les lieux administratifs des entités géopolitiques, puisqu'il s'agit de noms de villes ou de pays fréquemment employés en tant que l'un ou l'autre. On peut donc constater que cette tâche est plus ambitieuse que l'extraction d'EN *classique* (c'est-à-dire à la MUC).

5.5. Typologie du projet ReNom

ReNom est un projet français piloté par le laboratoire d'informatique[19] à l'université de Tours en collaboration avec le laboratoire ligérien de linguistique de l'université d'Orléans. Son objectif est d'enrichir les textes de renaissance par des informations sémantique telles que les étiquettes des catégories des EN (Maurel et al. 2013). Maurel et al. (2013) ont adoptée dans ce projet une typologie majoritairement inspiré de la norme TEI : les entités reconnues sont réparties en quatre types, les lieux géographiques (*geogName*), les lieux administratifs (*placeName*), les organisations (*orgName*) et, enfin, les personnes ou personnages (*persName*).

- Les lieux géographiques sont de deux types, d'une part, les géonymes : montagnes, plaines, plateaux, grottes... (*geo*) et, d'autre part, les hydronymes : océan, mer, rivière, lac, étang... (*hydro*).

[19] Dans le cadre des stages de perfectionnement du ministère de l'enseignement supérieur et de la recherche scientifique accordés aux enseignants chercheurs, j'ai visité ce laboratoire et j'ai assisté à plusieurs séances de travail et discussions menés par les membres de ce laboratoire dans le cadre du projet ReNom. Ces stages m'ont aidé à bien voir et comprendre la tâche de REN dans des projets réels.

- Les lieux administratifs peuvent éventuellement être typés (*ville*, *pays*, *batiment* ou *domaine*). Les deux types de lieux peuvent être imbriqués l'un dans l'autre.
- Les organisations sont partagées en trois types (*peuple*, *domaine* ou *communauté*) et n'ont pas d'identifiant associé. Elles peuvent éventuellement être imbriquées. Lorsqu'il est difficile de distinguer entre un lieu et une organisation, un double balisage est possible.
- Les personnes : Le balisage le plus simple consiste à reconnaître les personnes et à leur associer leur identifiant (*key*). Ce balisage peut être complété en interne par un balisage des prénoms (*forename*), des noms (*surname*) et des particules (*nameLink*).
- Enfin, le balisage doit être étendu par des titres ou des civilités (*roleName*) qui sont typés : titres nobiliaires, militaires ou religieux, fonctions et civilités honorifiques. Lorsqu'un titre comporte un nom de lieu ou une organisation, celui-ci est aussi balisé. Enfin, les précisions familiales (*genName*) et les surnoms (*addName*) entrent aussi dans le balisage.

5.6. Typologies utilisées pour la REN arabe

Dans la littérature, il existe trois jeux d'étiquettes (typologies) standards à usage général qui ont été utilisés pour annoter les ressources linguistiques arabes dans le domaine de la REN (Shaalan 2014). Comme il est noté dans la section 5.1 de la page 42 la conférence MUC est considérée comme l'initiateur de la tâche REN et ainsi la plupart des chercheurs (arabes et non arabes) adoptent sa typologie (Shaalan 2014).

5.6.1 Typologie de CoNLL

Comme résultat des éditions CoNLL2003 et CoNLL2005[20] (Conference of Natural Language Learning) quatre catégories d'EN ont été définies : noms de personnes, de lieux, d'organisation et divers. CoNLL suit le format BIO[21] pour étiqueter les parties de texte représentant des EN (Benajiba et al. 2007). Les annotations CoNLL sont formulées comme étant un problème à classification de mots, où chaque mot du texte reçoit une balise, indiquant si c'est le début (B) d'une EN, à l'intérieur (I) d'une EN spécifique, ou (O) en dehors de toute EN. La notation BIO est utilisée lorsque les EN ne sont pas imbriquées et ne se chevauchent pas (Shaalan 2014). Par exemple l'annotation en EN d'un système respectant le style CoNLL de la phrase فرنكفورت، أعلن اتحاد صناعة السيارات في ألمانيا\fran.kfuwr.t, Âç.lan AtiHaAd SinaAçaħ Als~ayaAraAt fiy Âl.maAnyA\ sera présentée dans le tableau ci-dessous (du fait que l'annotation CoNLL est un étiquetage de mots, un texte annoté selon ce style sera un tableau à deux colonnes : le mot et l'étiquette).

arabe	translitération	français	Etiquette EN CoNLL
فرنكفورت	fran.kfuwr.t	Frankfurt	B-LOC
،	,	,	O
أعلن	Âç.lan	a annoncé	O
اتحاد	AtiHaAd	l'association	B-ORG
صناعة	SinaAçaħ	de l'industrie	I-ORG

[20] http://www.signll.org/conll (consulté en décembre 2016)
[21] BIO : Beginning (début d'une EN), Inside (l'intérieur d'une EN) et Other ou Outside (autre qu'EN)

السيارات	Als~ayaAraAt	d'automobiles	I-ORG
في	fiy	En	O
ألمانيا	Âl.maAnyA	Allemagne	B-LOC

Tableau 11 : Exemple de la notation BIO d'EN suivant CoNLL

La notation BILOU a également été suggérée comme une alternative efficace au format BIO. Elle est utilisée pour identifier le début, l'intérieur et les derniers tokens d'un groupe multi-tokens ainsi que les groupes de longueur unitaire. Les résultats expérimentaux indiquent que la représentation BILOU des blocs de texte surpasse de manière significative en terme de performance le format BIO (Shaalan 2014).

5.6.2 Typologie d'ACE

Des ressources arabes pour l'extraction de l'information ont été développées dans le cadre du programme ACE[22] (Automatic Content Extraction). Selon les éléments d'étiquette ACE 2003, Quatre catégories sont définies: nom de personne, d'établissement, d'organisation et entités géographiques et politiques (GPE). Plus tard dans ACE 2004 et 2005, deux catégories ont été ajoutées à cet ensemble d'étiquettes: les véhicules et les armes.

6. La reconnaissance des entités nommées

6.1. Les approches pour la REN

Un certain nombre de systèmes REN arabes ont été développés en utilisant principalement deux approches: l'approche à basé de règles (basée sur la linguistique) et l'approche basée sur l'apprentissage automatique.

Les systèmes REN à base de règles s'appuient sur des règles de grammaires locales écrites à la main par des linguistes. Le principal avantage des systèmes REN à base de règles est qu'ils reposent sur un noyau de connaissances linguistiques solide. Cependant, toute maintenance ou mise à jour requise pour ces systèmes nécessite beaucoup de temps et de main-d'œuvre; Le problème sera davantage aggravé si les linguistes ayant les connaissances et le background requis ne seront pas disponibles.

D'autre part, les systèmes REN basés sur l'apprentissage automatique utilisent des algorithmes qui nécessitent de grandes masses de données annotées pour l'apprentissage et le test. Les algorithmes d'apprentissage automatique requièrent la sélection des attributs qui doivent être extraits à partir des ensembles de données annotés en EN afin de générer des modèles statistiques pour la prédiction des EN. L'avantage des systèmes REN à base d'apprentissage automatique est qu'ils sont adaptables et modifiables avec un minimum de temps et d'efforts tant qu'il est disponible des ensembles de données suffisamment grandes. De plus, si nous traitons un domaine ouvert, il est préférable de choisir l'approche apprentissage automatique, car il serait coûteux à la fois en termes de coût et de temps d'acquérir et/ou d'extraire des règles et des listes (gazetteers).

[22] Les jeux d'étiquettes ACE pour l'anglais, l'arabe et d'autres langues sont disponibles à l'adresse https://www.ldc.upenn.edu/collaborations/past-projects/ace/annotation-tasks-and-specifications (consultée en décembre 2016)

Récemment, une approche REN arabe hybride combinant l'apprentissage automatique et des approches à base de règles, a permis une amélioration significative en exploitant les décisions à base de règles d'EN en tant qu'attribut utilisé par le classifieur.

6.2. Les difficultés pour la REN arabe

La reconnaissance des entités nommées arabes rencontre des difficultés qui rendent cette tâche plus difficile par rapport aux langues indo-européennes. Et ces difficultés sont à cause des particularités de cette langue. Car c'est une langue fortement fléchie, avec une morphologie riche et une syntaxe complexe.

Shaalan, Benajiba et Zaghouani respectivement dans (Shaalan 2014; Benajiba 2009; W. Zaghouani 2009) ont résumé les particularités de la langue arabe en ce qui concerne la REN, qu'on peut citer en ce qui suit :

6.2.1 Absence de majuscule

A l'inverse des langues qui utilisent l'alphabet latin comme le français ou l'anglais, où la plupart des EN commencent par une lettre majuscule, la majuscule n'est pas une caractéristique orthographique distinctive de l'écriture arabe pour reconnaître EN tels que les noms propres, acronymes et abréviations. L'ambiguïté causée par l'absence de cette caractéristique est davantage accrue par le fait que la plupart des noms propres arabes (EN) sont indiscernables des noms communs et des adjectifs (non-EN). Ainsi, une approche basée uniquement sur la recherche d'entrées dans les dictionnaires de noms propres ne serait pas un moyen approprié de s'attaquer à ce problème, car les mots / tokens ambigus qui appartiennent à cette catégorie sont plus susceptibles d'être utilisés dans le texte comme noms non propres. Par exemple, le nom propre arabe أشرف (Ashraf) peut être utilisé dans une phrase comme un prénom, un verbe fléchi (il a supervisé) ou un superlatif (le plus honorable) (Mesfar 2008). Une EN se trouve généralement dans un contexte avec des mots de déclenchement et de repère à gauche et/ou à droite.

6.2.2 Agglutination

La nature agglutinative de l'arabe mène à des motifs différents qui créent des variations lexicales. Chaque mot peut consister en un ou plusieurs proclitiques/préfixes, une base ou une racine, et un ou plusieurs suffixes/enclitiques dans des combinaisons différentes, aboutissant à une morphologie très systématique mais compliquée. Dans d'autres langues comme le français ou l'anglais les clitiques seraient traités comme des mots séparés qui n'agglutinent pas aux mots. L'arabe a un ensemble de clitiques qui sont attachés à une EN, y compris :

- des conjonctions telles que و\wa*et* et ف\fa*et*
- des prépositions telles que ل\li*pour*, ك\ka*comme* et ب\bi*par*
- ou une combinaison des deux telles que ول\wali*et pour*.

La REN s'appuie sur les mots formant l'EN et le contexte dans lequel elle apparaît. Les mots et les contextes peuvent apparaître sous différentes formes fléchies. Afin de traiter le problème de la rareté des données sans pour autant avoir besoin de corpus d'apprentissage massif, ces morphèmes liés doivent subir un prétraitement morphologique.

Une solution consiste à omettre tous les affixes et ne conserver que le morphème racine. Par exemple, l'analyse du mot وبالجزائر\wabiAljazaAŷir*et par l'Algérie* donne الجزائر\AljazaAŷir*Algérie* comme un nom de lieu.

Une autre solution consiste à effectuer une segmentation textuelle (segmentation des clitiques) et à insérer un délimiteur entre les morphèmes constitutifs, empêchant ainsi la perte d'informations contextuelles. Comme un exemple qui montre une occurrence des morphèmes préfixes et suffixes, considérons le mot déclencheur وعاصمتها\waçaASimatuhaA*et sa capitale* qui est segmenté en trois parties : une conjonction, un nominal et un pronominal, séparées par un caractère d'espace و عاصمة ها\wa ςaASimatu haA*et capitale sa*.

6.2.3 Voyelles courtes facultatives

Le texte arabe contient des signes diacritiques, la plupart représentant des voyelles qui affectent la phonétique et donnent une signification différente à la même forme lexicale. De nos jours, la version moderne de l'arabe est écrite sans diacritiques, créant une ambiguïté un-à-plusieurs et non vocalisée-à-vocalisée, qui donne des analyses morphologiques différentes pour la même forme de surface cf. la Figure 9.

En tant que tels, la plupart des textes arabes qui apparaissent dans les médias (qu'ils soient imprimés ou numérisés) ne sont pas diacritisés. Ceci est compréhensible pour les arabophones natifs, mais pas pour un système TAL. La simplification faite en ignorant ces diacritiques a conduit à des types d'ambiguïté structurelle et lexicale parce que les diacritiques différents représentent des significations différentes. Ces ambiguïtés ne peuvent être résolues que par des informations contextuelles et une connaissance adéquate de la langue.

Par exemple, le mot قطر peut se référer

- au nom du pays *Qatar* (EN de lieu) s'il est translittéré en \qaTar\,
- au *rayon* (un mot déclencheur pour EN de mesures) s'il est translitéré en \quT.r\,
- ou au sens littéral de *distiller* s'il est translitéré en \qaT~ara\.

Malheureusement, cette solution pourrait ne pas fonctionner si l'information contextuelle est elle-même ambiguë en raison de non vocalisation. Pour considérer un autre exemple, les vocalisations probables de la forme non vocalisée مؤسسة\mŵssħ\ (Mesfar 2008) pourrait conduire à des mots déclencheurs qui dénotent deux types d'EN différents :

- مُؤَسَّسَة\muŵs~asaħ*société ou fondation*, une preuve interne d'un constituant d'un nom d'organisation ou
- مُؤَسِسَة \muŵasisaħ*fondeuse*, un mot déclencheur pour des noms de personnes.

6.2.4 Ambiguïté inhérente aux EN

A l'instar des autres langues, l'arabe fait face au problème de l'ambiguïté entre deux EN. Considérons par exemple le texte suivant : أحمد آباد رحب بالفائزين\ÂHmd ĀbAd rHb bAlfAŷzyn*Ahmed Abad a bien accueilli les gagnants*. Dans cette exemple, أحمد آباد\ÂHmd ĀbAd*Ahmed Abad* est à la fois un nom de personne et un nom de lieu, donnant ainsi lieu à une situation d'ambiguïté, où le même EN est marqué comme deux types différents d'EN. Pour résoudre les ambiguïtés, des techniques heuristiques par reconnaissance croisée des types EN, sont suggérées. Une technique heuristique, proposée par Shaalan et Raza (2009), utilise des règles heuristiques pour préférer un type EN sur l'autre. Une autre technique favorise le type d'EN pour lequel le classifieur atteint la plus grande précision.

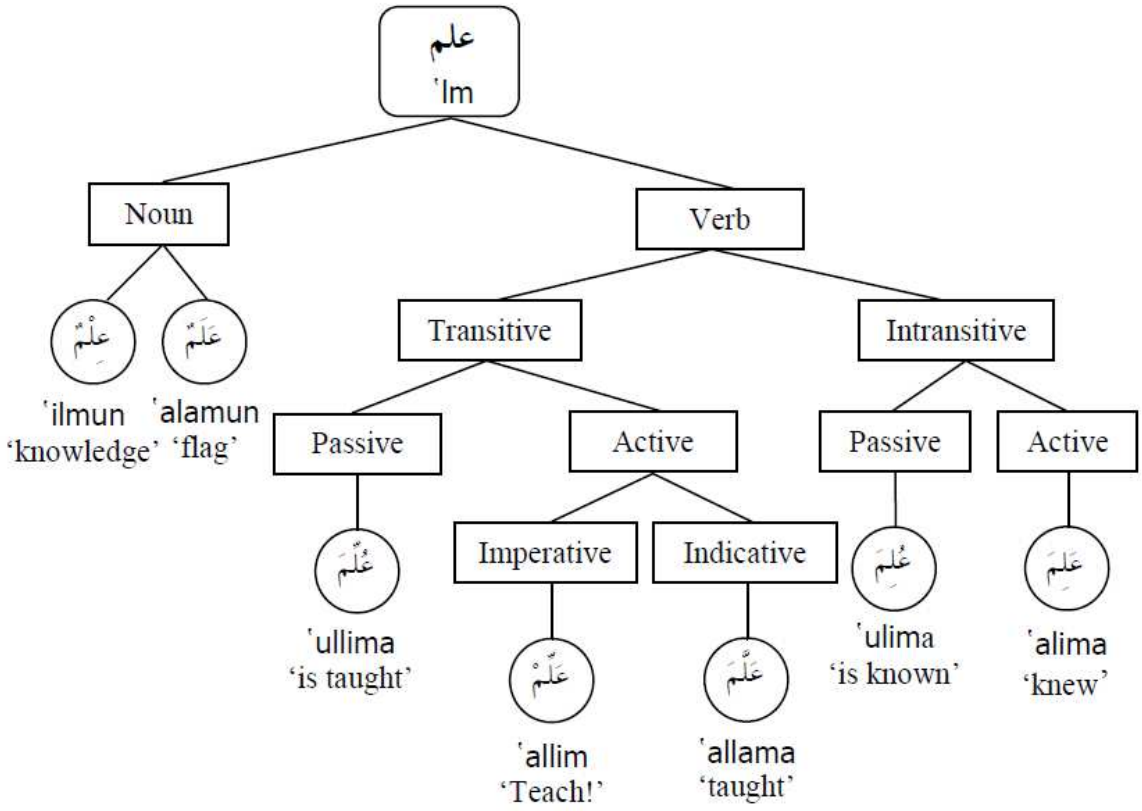


Figure 9 : Exemple d'ambiguïté causée par l'absence des voyelles courtes dans le texte arabe, extrait de (Attia 2008a).

6.2.5 Manque d'uniformité dans les styles d'écriture

L'arabe a un niveau élevé d'ambiguïté en translitération : une EN peut être translittérée de multiples façons. Cette multiplicité provient à la fois des différences entre les rédacteurs arabes et des schémas de translitération ambigus. L'absence de standardisation est critique et conduit à de nombreuses variantes du même mot qui sont orthographiées différemment mais qui correspondent toujours au même mot ayant la même signification créant ainsi une ambiguïté de plusieurs-à-un. Par exemple, translitérer (aussi connu sous le nom de *arabiser*[23]) une EN telle que la ville de Washington en une EN arabe, produit des variantes telles que واشنجطن, واشنطن, واشنغطن, وشنطن. Une des raisons en est que l'arabe a plus de sons vocaux que les langues d'Europe occidentale, ce qui peut conduire de manière ambiguë ou erronée à une EN ayant plusieurs variantes. Une solution est de conserver toutes les versions des variantes de noms avec une possibilité de les relier. Une autre solution consiste à normaliser chaque occurrence de la variante à une forme canonique. Ceci nécessite un mécanisme de matching (tel qu'un calcul de distance de chaîne) entre une variante de nom et sa représentation normalisée.

6.2.6 Erreurs d'orthographes systématiques

Les erreurs typographiques sont fréquemment faites par des rédacteurs arabes à l'égard de certains caractères. Ceci est dû soit à une similarité de caractères, soit à un désaccord inhérent

[23] arabiser c'est l'inverse de romaniser qui est la translittération ou la transcription d'une écriture non latine vers une écriture latine.

à l'égard des caractères, ce qui conduit souvent à une confusion orthographique. La première catégorie comprend le caractère *Ta-Marbuta* ة, littéralement *Ta fermée*, qui est un marqueur morphologique marquant typiquement une fin féminine; Cela est négligemment écrit de manière interchangeable avec Ha ه.

La deuxième catégorie inclut les variantes de lettres de Hamza-Alif qui sont souvent normalisées par le remplacement de la force brute par un Alif dénudé. Certains linguistes computationnels évitent d'écrire le Hamza (en particulier avec l'Alif initial), considérant cela comme un problème de restauration de Hamza qui fait partie du problème de la diacritisation automatique arabe. Comme exemple qui combine les deux types d'erreurs, considérons الجامعة الإسلامية بجدة\AljaAmiçah AlĂs.lAmiyah bijad~ah*l'université islamique à Djeddah* qui pourrait être écrite avec les deux variantes typographiques comme الجامعه الاسلاميه بجده\AljaAmiçah AlAs.lAmiyah bijad~ah\. Une technique de distance d'édition peut être utilisée pour résoudre le problème de variante d'orthographe. Il convient de noter que toutes les erreurs d'orthographe systématiques ne peuvent pas être traitées de cette manière. Par exemple, considérons la différence entre بالجامعة\biAljaAmiçah*par l'université* et بلاجامعة\bilAjaAmiçah*sans université*. Il est difficile de déterminer si cette erreur est due à la transposition des deux caractères ا (Alif) et ل (Lam), où le préfixe ال (signifie le) alors que le préfixe لا (signifie non). Cette dernière variation montre aussi un autre problème orthographique: les mots arabes *run-on* ou concaténation libre de mots, lorsque le mot immédiatement précédent se termine par une lettre non-connectée, telle que ا (Alif), د (Dal), ذ (Dhal), ر (Ra), ز (Za), و (waw) et ainsi de suite. Par exemple, la phrase suivante montre une EN de nom de personne pleinement concaténée avec son contexte environnant: الدكتورمحمدوزيرالخارجية\Alduk.tuwrmuHamadwaziyrAlxaArijiyah*Dr Mohammed le ministre des Affaires étrangères*. Ceci est compréhensible par la plupart des lecteurs mais pas par un système informatique qui doit travailler sur des mots segmentés.

6.2.7 Manque de ressources

De grandes collections de documents annotés (corpus) ainsi que des nomenclatures ou gazetteers (listes d'EN typées prédéfinies) sont d'excellentes sources sur lesquelles nous pouvons nous baser lors pour la mise en œuvre et du test des performances d'un système REN arabe. Pour que ces ressources linguistiques soient utiles, elles doivent inclure un nombre représentatif d'EN qui ne souffrent pas de la rareté.

Malheureusement, les ressources arabes disponibles pour la REN ont souvent une capacité et/ou une couverture limitées. En outre, il est coûteux de créer ou d'acquérir la licence de ces importantes ressources de REN arabe. Pour ces raisons, les chercheurs s'appuient souvent sur leurs propres corpus, qui nécessitent l'annotation et la vérification humaines. Peu de ces corpus ont été rendus libres et publiques à des fins de recherche. Alors que d'autres sont disponibles mais sous contrat de licence.

6.3. Etat de l'art des systèmes de la REN arabe

L'importance des systèmes REN arabes a été bien reconnue par la communauté comme en témoignent les publications remarquables dans ce domaine important. Dans cette section, on présente différents systèmes REN. Ils sont classés selon l'approche utilisée. Malheureusement pour le milieu de la recherche scientifique, la plupart des efforts déployés pour mettre au point des systèmes REN arabes fiables ont été entrepris à des fins commerciales. Vu que l'information sur les spécifications et les performances de ces systèmes

généralement n'est pas disponible, il est difficile d'effectuer une comparaison équitable de la performance de ces systèmes par rapport aux systèmes REN arabe proposés par la communauté de recherche.

Comme exemples de systèmes REN arabes commerciaux on a : ANEE (Coltec), IdentiFinder (BBN), NetOwlExtractor (NetOwl), Siraj (Sakhr), Clear Tags (ClearForest), Enterprise Search (FAST ESP) et InXight-Smart- Discovery-Entity- Extractor (InXight). En ce qui concerne les systèmes REN arabe disponible à la communauté de la recherche, voici une énumération du plus anciens au plus récent et qui n'est pas une liste exhaustive.

6.4. Système à base règles

6.4.1 Système PERA

Cette recherche menée par Shaalan et Raza (2007) décrit la structure des noms de personnes arabes: 'ism', 'kounya', 'nasab', 'laqab' et 'nisba'. L'ism est un nom propre donné peu de temps après la naissance. Des exemples de tels noms sont محمد\muHamad*Mohamed,* موسى\muwsaý\Moïse, إبراهيم\Ăb.raAhiym*Abraham*.

La 'kounya' est un nom honorifique ou un nom de famille qui indique le nom du père de quelqu'un أبو\Âabuw*père de* ou mère de quelqu'un أم\Âum*mère de*. Par exemple, أبو داود\Âabuw daAwud*père de David,* أم سليم\Âum saliym*mère de Salim*. Lorsque l'on utilise le nom complet d'une personne, la 'kounya' précède le nom donné, par exemple, أبو يوسف حسن\Âabuw yuwsuf Hasan*Hacene père de Joseph,* أم جعفر أمينة\Âum jaç.far Âamiynaħ*Aminah mère de Jaafar*.

Le 'nasab' indique l'héritage d'une personne par le mot ابن\Aib.n\ qui signifie *fils de* (بنت\bin.t*fille de*). Par exemple ابن عمر\Aib.nu çumar*fils d'Omar,* بنت عباس\bin.tu çabaAs*fille d'Abbas*. En usage le 'nasab' suit 'l'ism', par exemple حسن ابن فراج\Hasan Aib.nu far~aAj*Hacene fils de Ferraj,* سمية بنت خبيت\sumay~aħ bin.tu xubay.t*Soumaya fille de Khoubayt*. Plusieurs personnes historiques qui nous sont familiers par leur 'nasab' que par leur 'ism'. Exemples dignes d'être notés : l'historien ابن خلدون\Aib.nu xal.duwn*Ibn Khaldun,* le voyageur ابن بطوطة\Aib.nu baTuwTaħ*Ibn Battuta* et le philosophe ابن سينا\Aib.nu siynaA*Avicenne*.

Le 'laqab' est une combinaison de mots en une épithète, généralement religieux, ou en relation avec un trait, une description ou une qualité admirable en une personne. Exemples comme : الرشيد\Alr~ašiyd*le bien guidé* et الفاضل\AlfaADil*le proéminent*. En pratique le 'laqab' suit 'l'ism' par exemple هارون الرشيد\haAruwn Alr~ašiyd*Aaron le bien guidé.*

Enfin, la 'nisba' est un nom dérivé: du commerce, de profession, de lieu de résidence ou de naissance ou d'affiliation religieuse d'une personne. Par exemple الحلاج\AlHal~aAj*le tisseur de coton,* الجزائري\Alj~azaAŷiriy*algérien,* إسلامي\Ăs.lAmiy*islamique*. La 'nisba' suit 'l'ism' et s'il contient un 'nasab', elle suit 'nasab'.

Dans le système PERA les règles prennent la forme des expressions régulières qui contiennent les constituants de nommage cités ci-dessus, afin de reconnaitre les noms de personnes. Le système est composé de trois composants : listes, règles de grammaire et un mécanisme de filtrage. Des listes blanches de noms de personnes ont été fournies dans le composant liste afin d'extraire le matching exact des EN mis à part l'utilisation des grammaires. Après, le texte en entrée est présenté aux grammaires afin d'identifier autres EN

de personne. Enfin le mécanisme de filtrage est appliqué aux EN afin d'éloigner les noms de personnes invalides.

PERA a été évalué en utilisant les données d'ACE et de Treebank Arabic et il a obtenu une précision, rappel et f-mesure de 85.5%, 89% et 87.5% respectivement.

6.4.2 Système NERA

Comme une continuation des travaux menés par Shaalan et Raza (2007), le système NERA généralise les résultats obtenus par PERA. NERA examine les principaux chalenges posés par la REN arabe provenant de la complexité du système morphologique, des particularités du système orthographique arabe, de la non-standardisation du texte écrit, d'ambiguïté et du manque de ressources. Le système identifie les types d'EN suivants: personne, lieu, organisation, date, heure, ISBN, prix, mesures, numéros de téléphone et nom de fichiers. NERA utilise Fast ESP comme plateforme d'implémentation dont son architecture est optimisée pour les systèmes à base de règles. A l'instar de PERA, NERA a trois composantes : listes, règles sous formes d'expressions régulières et mécanisme de filtrage.

L'évaluation de ce système est faite sur un corpus construit manuellement à partir d'ACE, du web et d'organisations. Il a obtenu une f-mesure de 87.7% pour les personnes, 85.9% pour les lieux et 83.15% pour les organisations.

6.4.3 Travaux d'Alkharashi

Alkharashi (2009) a décrit la formation d'un nom de personne arabe à partir de la racine et du schème en utilisant la morphologie arabe traditionnelle et il a suggéré des ressources informatiques pertinentes. L'auteur a introduit un ensemble de tables de base de données afin d'assister la REN arabe: tables de racine-schème, d'une liste de fréquences de racines et de déclencheurs lexicaux. Un corpus a été créé à partir de noms de personnes saoudiennes avec des étiquettes spécifiques aux noms : racine de l'EN personne, des traits indiquant la possibilité d'affixation, et des caractéristiques de genre (féminin ou masculin). L'objectif principal était de reconnaître les constituants de l'EN personne, soit la forme simple, l'affixe et les connecteurs.

Par exemple, le nom du califat Omeyyade الوليد بن عبد الملك\Alwaliyd b.nu ςab.dAlmalik\ a ملك\malik\ et وليد\waliyd\ comme des noms simples, عبد\ςab.d\ et ال\Al\ comme des préfixes et بن\ban\ comme connecteur de nom.

L'étude a rapporté des observations intéressantes sur les caractéristiques des schèmes très fréquents et de leurs longueurs. Un test simple d'évaluation a été mené sur 60 000 entrées de noms de personnes générées, ce test indique comment le schème d'un nom de personne a été reconnu. Il a démontré que le schème correct apparaît à 94% du temps comme l'un des trois premiers schèmes suggérés, 86% comme l'un des deux premiers schèmes suggérés, et 69% du temps comme le premier schème suggéré.

6.4.4 Travaux d'Al Shalabi

Al-Shalabi et al. (2009) ont présenté un algorithme REN arabe pour récupérer les noms propres arabe à l'aide de déclencheurs lexicaux. Cette recherche prend en considération des modèles régionaux tels que le connecteur de nom ولد\wul.d.*fils de* utilisé en noms de personnes mauritaniennes (par exemple, الرئيس محمد ولد عبد العزيز\Alraŷiys muHamad wul.d. ςab.duAlςaziyz*le président Mohamed Ould Abdelaziz*). L'algorithme

identifie les types d'EN suivants: personnes, grandes villes, lieux, pays, organisations, partis politiques et groupes terroristes. Cependant, les recherches publiées ne portent que sur les EN de type personne. L'algorithme utilise des règles heuristiques pour prétraiter l'entrée pour nettoyer les données et supprimer les affixes. Ensuite, les éléments de preuve internes, tels que les connecteurs de nom de personne, sont utilisés pour reconnaître les EN. Le système a été évalué à l'aide de 20 documents sélectionnés au hasard du journal *Al-Raya* publié au Qatar et du journal *Alrai* publié en Jordanie. Une précision globale de 86,1% a été observée.

6.4.5 Travaux d'Attia

Attia et al. (2010) ont proposé une méthode pour acquérir un lexique EN plus riche en utilisant ArabicWordNet (El-Kateb et al. 2006) et Wikipedia arabe. Le lexique EN arabe proposé améliore les entrées lexicales dans WordNet et produit une ressource lexicale arabe bien structurée. L'objectif principal est d'extraire les substantifs instanciables de WordNet et d'identifier les catégories correspondantes dans Wikipedia arabe. Ces catégories agissent comme des déclencheurs lexicaux. Une décision est prise afin d'identifier lesquels des articles de Wikipedia de ces catégories correspondent aux EN. Elles sont ensuite extraites, connectées à ArabicWordNet, et insérées dans l'entrepôt d'EN. Dans une étape ultérieure de post-traitement, d'autres EN sont acquises en exploitant les liaisons inter-langues (ILI). Enfin, les EN acquises sont diacritisées. Cette ressource lexicale est utile pour la REN arabe car les résultats ne sont pas seulement des EN reconnus (annotées) mais aussi des synsets identifiés qui sont sémantiquement liés à eux (synonymes, sous-types, super-types, etc.).

6.4.6 Système RENAR

Zaghouani et al. (2010) ont présenté une adaptation d'un système multilingue pour inclure l'arabe, cette application c'est NewsExplorer spécialisé en recherche et extraction d'informations de l'Europe Media Monitor. Ce système comprend actuellement dix neuf langues et est capable d'analyser de grands volumes de textes d'actualités. L'architecture d'EMM-NewsExplorer est optimisée pour les systèmes à base de règles. L'adaptation a abouti à un système REN arabe basé sur des règles (RENAR), qui utilise un ensemble de règles écrites à la main et indépendantes de la langue en combinaison avec des ressources spécifiques pour l'arabe.

6.4.7 Système ARNE

Shihadeh et Neumann (2012) ont proposé un système REN arabe appelé ARNE, qui reconnaît les EN de type personne, lieu et organisation, basée uniquement sur une approche de recherche de nomenclature; Le système fournit des informations morphologiques à l'aide d'un système ElixirFM, développé par Smrz (2007)

6.4.8 Travaux d'Al-Jumaily

Al-Jumaily et al. (2012) ont proposé un système REN basé sur des règles qui peut être utilisé dans les applications Web. Le système identifie les types d'EN suivants: personne, lieu et organisation. Le système a été développé à l'aide de l'environnement GATE et fournit l'analyse morphologique arabe dans une méthode similaire à BAMA. Il intègre également différents listes de GATE, DBPedia et ANERGazet. Le système a été évalué à l'aide d'ANERcorp.

6.5. Système à base d'apprentissage automatique

6.5.1 Système ANERsys

Benajiba et al. (2007, 2008, 2009, 2010) ont exploré plusieurs techniques d'apprentissage automatique dans leurs travaux sur la REN arabe. Ils ont développé un système à base d'entropie maximale appelé ANERsys 1.0. Ils ont construit leurs propres ressources linguistiques ANERcorp et ANERgazet. Les caractéristiques lexicales, contextuelles et de listes ont été utilisées dans ce système. ANERsys identifie les catégories d'EN suivantes : personnes, lieux, organisations et divers. Toutes les expérimentations ont été menées dans le cadre de la conférence CoNLL2002. Les performances générales du système en terme de précision, rappel et f-mesure étaient 63.21%, 49.04% et 55.23% respectivement. Les difficultés d'ANERsys 1.0 consistaient dans l'identification d'EN composées de plusieurs tokens/mots. L'extension de ce travail est ANERsys 2.0 qui utilise un mécanisme à deux étapes pour la REN arabe :

- détecter le point de départ et de fin de chaque EN,
- classifier toutes les EN détectées.

Une caractéristique d'étiquetage morphosyntaxique a été exploitée pour la détection des frontières d'une EN. Les performances générales du système en terme de précision, rappel et f-mesure étaient 70.24%, 62.08% et 65.91% respectivement. Quoi que l'étape d'identification fût mauvaise avec une f-mesure de 72.03%, la performance du module de classification était très bonne avec une f-mesure de 82.22%.

Dans une tentative d'amélioration des performances du système, la technique de CRF (Conditional Random Fields) a été appliquée au lieu de l'entropie maximale. Les mêmes quatre types d'EN utilisées dans ANERsys 2.0 ont également été utilisées dans le système à base de CRF. Les travaux antécédents ne comprenaient pas des caractéristiques spécifiques à l'arabe; toutes les caractéristiques utilisées étaient indépendantes du langage. Les caractéristiques indépendantes de la langue et les caractéristiques spécifiques à l'arabe ont été utilisées dans le modèle CRF, y compris les étiquettes morphosyntaxiques, la segmentation en syntagmes (BPC : base phrase chunking), les listes et la nationalité. Le système CRF a obtenu les meilleurs résultats lorsque toutes les caractéristiques ont été combinées. Les performances générales du système en termes de précision, de rappel et de f-mesure étaient respectivement de 86,90%, 72,77% et 79,21%. L'amélioration dépend non seulement de l'utilisation du modèle CRF, mais aussi des caractéristiques spécifiques au langage supplémentaires, y compris POS et BPC.

Benajiba, Diab et Rosso (2008) ont examiné les caractéristiques lexicales, contextuelles, morphologiques, listes et syntaxiques superficielles des ensembles de données ACE à l'aide du classifieur SVM (Support Vector Machine). La performance du système a été évaluée en utilisant la validation croisée (5 fois). L'impact des différentes caractéristiques est mesuré indépendamment et en combinaison conjointe à travers différents ensembles de données standard et genres. La meilleure performance globale du système en termes de f-mesure était de 82,71% pour ACE 2003, 76,43% pour ACE 2004 et 81,47% pour ACE 2005.

Benajiba et al. (2009) ont confirmé l'importance de considérer à la fois les caractéristiques indépendantes de la langue et les caractéristiques spécifiques à l'arabe dans le système REN. Ils ont étudié l'impact des modèles SVM, ME et CRF en utilisant la même approche et les mêmes caractéristiques décrites dans le travail de 2008. La meilleure performance globale du

système en termes de f-mesure était de 83,34% pour ACE 2003, 77,61% pour ACE 2004 et 82,02% pour ACE 2005. Des conclusions et des recommandations intéressantes ont été suggérées par cette étude. Les SVM et les CRF ont obtenu des résultats très similaires, tout en dépassant le modèle ME. Une observation importante concerne le nombre de caractéristiques disponibles comme facteur principal pour le choix de l'utilisation des SVM par rapport aux CRF:

- Les SVM semblent obtenir de bons résultats avec moins de caractéristiques.
- Une autre observation importante concerne la meilleure performance obtenue en effectuant le prétraitement du texte arabe par un tokenizer de clitiques, ce qui est plus approprié compte tenu de la richesse morphologique de l'arabe.

6.5.2 Projet AQMAR

Mohit et al. (2012) ont proposé un modèle d'apprentissage pour la REN arabe à partir des textes de divers domaines comme Wikipedia et ceci dans le cadre du projet AQMAR (American and Qatari Modeling of Arabic). Ils ont utilisé un schéma d'annotation flexible qui permet l'introduction de nouvelles étiquettes d'EN. Comme Wikipedia arabe n'étant pas étiqueté en EN, ils ont adopté un apprentissage semi-supervisé pour la construction de leur propre corpus. La méthode d'apprentissage n'utilise aucune liste (gazetteer). Une fois le corpus est construit, une méthode d'apprentissage supervisée peut être utilisée pour développer et évaluer un classifieur de REN. L'espace des caractéristiques consiste en quinze caractéristiques lexicales et textuelles capturant une morphologie superficielle et un contexte local. Les caractéristiques morphologiques sont extraites des sorties de MADA.

Le système a été testé sur vingt quatre articles Wikipedia pour les combinaisons possibles de la phase d'apprentissage supervisée avec l'auto-apprentissage sur les données de Wikipedia non annotées. Les résultats expérimentaux ont montré des améliorations sur la F-mesure par le modèle orienté-rappel proposé dans les deux étapes de l'apprentissage.

6.5.3 Travaux de Koulali

Koulali et al. (2012) ont développé un système de REN arabe en utilisant une combinaison d'extracteur de patrons (un ensemble de règles d'expressions) et un classifieur SVM qui apprend les patrons à partir d'un texte étiqueté en morphosyntaxe. Le système couvre les types d'EN de la conférence CoNLL et utilise un ensemble de caractéristiques (attribut). les caractéristiques arabes incluent :

- un déterminent caractéristique ال\Al\ apparait au début des noms d'organisations. Par exemple اليونسكو\Alywnskw*UNIESCO*,
- une caractéristique basée sur le caractère qui dénote les proclitiques communs des noms,
- une caractéristique d'étiquetage morphosyntaxique,
- une caractéristique du verbe d'entourage qui dénote si l'EN est précédée ou suivie par un certains verbes.

L'apprentissage du système a été fait sur 90% des données d'ANERCorp et le test a été fait sur le reste. Le système a été testé sur différentes combinaisons de caractéristiques et le meilleur résultat pour une F-mesure moyenne globale était 83.20%.

6.5.4 Système Noor

Bidhend et al. (2012) ont présenté un système REN arabe à base de CRF appelé Noor qui extrait les noms de personnes à partir des textes religieux. Un corpus de textes religieux anciens a été développé consistant en trois types : historique, Hadith du prophète Mohammed et livres de jurisprudence. Noor-Gazet une liste (gazetteer) de noms de personnes religieuses a aussi été développée. Les noms de personnes ont été tokenisés par une étape de prétraitement. Par exemple la segmentation du nom complet حسن بن علي بن عبد الله بن المغيرة \Hasan b.nu ʕaliy b.nu ʕab.du Allh b.nu Almuγiyraħ\ produit six tokens comme suit : حسن بن علي عبد الله المغيرة\Hasan b.nu ʕaliy ʕab.du Allh Almuγiyraħ\. Un autre outil de prétraitement a été utilisé AMIRA qui a été utilisé pour l'étiquetage morphosyntaxique. L'étiquetage est enrichi par l'indication dans le cas échéant, de la présence d'une entrée Noor-Gazet d'EN de type personne. La f-mesure de la performance du système en général en utilisant un nouveau corpus de textes historiques, de Hadith et de jurisprudence, était de 99.93%, 93.86% et 75.86% respectivement.

6.6. Systèmes hybrides

L'approche hybride intègre l'approche basée sur des règles à l'approche d'apprentissage automatique dans l'objectif d'optimiser la performance globale. Abdallah et al. (2012) cité par (Shaalan 2014) ont proposé un système de REN hybride pour l'arabe. Le composant à base de règles est une réimplémentation du système NERA en utilisant GATE. Le composant de l'apprentissage automatique utilise les arbres de décision. L'espace des caractéristiques comprend les étiquettes des EN prédites par le composant basé sur des règles, d'autres caractéristiques indépendantes de la langue et des caractéristiques spécifique à l'arabe. Le système identifie les types d'EN suivants: personne, lieu et organisation. La performance de la F-mesure à l'aide de l'ANERcorp était respectivement de 92,8%, 87,39% et 86,12% pour l'EN personne, lieu et organisation.

Les travaux de Ouadah et al. (2012, 2013) sont considérés comme une continuité des travaux de Abdallah et al. (2012) par l'extension du système de la REN arabe hybride par les points suivants :

- en augmentant les EN à onze types en ajoutant temps, mesure, numéro de téléphone, nom de fichier, date, prix, pourcentage et ISBN.
- en explorant deux autres modèles d'apprentissage automatique : SVM et la régression logistique.
- en augmentant les caractéristiques à un plus grand ensemble en ajoutant des caractéristiques morphologiques et la caractéristique de la présence de majuscules dans le gloss anglais.

Les résultats expérimentaux ont montré que l'approche hybride pour la REN arabe surpasse les composants à base de règles et à base d'apprentissage quand ils sont lancés individuellement. La performance obtenue à l'aide de l'ANERcorp pour la F-mesure était de 94,4% pour l'EN personne, de 90,1% pour l'EN lieu et de 88,2% pour l'EN organisation (Shaalan 2014).

6.7. Etude comparatives entre les systèmes de la REN arabe

6.7.1 Comparaison

Afin de positionner notre travail par rapport aux travaux précédents en ce qui concerne la reconnaissance des entités nommées arabes, on expose dans la présente section une étude

comparative entre les différents travaux et systèmes de la REN arabe réalisés jusqu'ici. Le Tableau 12 ci-dessous est organisé en plusieurs colonnes représentant les critères de la comparaison qui nous sont avérés prépondérants dans cette étude. Les travaux et les systèmes de cette étude sont triés par ordre croissant de leurs années d'apparition.

Les critères de comparaison sont comme suit :

- Catégories traités : désigne le nombre de classes d'EN traitées dans le système en question, ce critères est important car il montre la diversification des cas et ainsi la complexité de la reconnaissance de l'unité qui représente l'EN. Un nombre élevé de classes (comme c'est le cas pour NERA, AQMAr et Ouadah) implique une ambiguïté élevée entre l'EN et les autres catégories grammaticales du texte et ainsi une complexité dans les règles de reconnaissances. Par contre une valeur basse (comme c'est le cas pour PERA et Noor) indique que le système a affaire avec une seule catégorie d'EN. Une valeur basse en nombre de catégories traitées doit être reflétée dans les valeurs des métriques d'évaluation, i.e. doivent être élevées par rapport à un système avec un nombre élevé de classes d'EN traitées.
- typologie : désigne la hiérarchie des classes d'EN utilisé dans le système, les typologies connues dans la littérature sont exposées dans la section 5 du chapitre courant. le recourt à des typologies propriétaires (non standards) peut exposer le système à des critiques.
- approche : désigne la solution scientifique utilisée pour aboutir aux objectifs du système, à savoir approche linguistique à base de règles ou approche statistique (apprentissage automatique, etc.). L'hybridation des deux approches est possible comme c'est le cas pour le système de Ouadah.
- formalisme/technique : désigne le formalisme utilisé pour traduire les règles dans le cas des systèmes à base de règles ou la technique utilisé pour faire l'apprentissage automatique. Dans le cas du système hybride, les règles sont utilisées comme des caractéristiques d'apprentissage.
- corpus de test : désigne le nom du corpus de test qui a été utilisé pour l'évaluation du système. Dans le cas de l'apprentissage automatique le corpus de test est utilisé à la fois à l'étape d'apprentissage et à l'étape de test.
- évaluation : ce critère désigne les métriques d'évaluation très répandues dans la littérature à savoir précision, rappel et f-mésure. N'étant pas évalué, notre système ne dispose pas de valeurs de métriques d'évaluation.

Système	année	catégories traitées	typologie	approche	formalisme / technique	corpus de test	Evaluation		
							précision	rappel	F-mesure
ANERsys 1.0	2007	4	CoNLL	apprentissage automatique	Entropie maximale	ANERcorp	63.12%	49.04%	55.23%
PERA	2007	1	noms de personnes	à base de règles	Expressions régulières	ACE et Arabic TreeBank	85.5%	89%	87.5%
NERA	2007	11	propriétaire	à base de règles	Expressions régulières	ACE et web			87.7%
ANERsys 2.0	2008	4	CoNLL	apprentissage automatique	Conditional Random Fields	ANERcorp	70.24%	62.08%	65.91%
Al-Shalabi	2009	7	propriétaire	à base de règles		Les quotidien Al-Raya et Alrai	86.1%		
ANERsys 2.0	2009	4	CoNLL-ACE	apprentissage automatique	Support Vector Machine	ACE			83.28%
RENAR	2010			à base de règles					
NERA-Gate	2012	3	personne, lieu et organisation	Hybride	Gate - arbre de décision	ANERcorp			88.77%
ARNE	2012	3	personne, lieu et organisation						
AQMAr	2012	variable		apprentissage semi supervisé	15 attributs	Wikipedia arabe			
Al-Jumaily	2012	3	personne, lieu et organisation		Gate	ANERcorp			
Koulali	2012	4	CoNLL	apprentissage automatique	Support Vector Machine	ANERcorp			83.20%
Noor	2012	1	noms de personnes		Conditional Random Fields	Texte historique, Hadith et jurisprudence			89.83%
Ouadah	2013	11	propriétaire	Hybride	Gate – SVM – régression logistique	ANERcorp			90.9%

Notre système	2015	3	Quaero	à base de règle	Transducteur à états finis	AraCorpus			

Tableau 12 : Comparaison entre les systèmes de reconnaissances des entités nommées arabes

6.7.2 Synthèse

Les valeurs des colonnes d'évaluation du tableau montrent que les systèmes hybrides sont les meilleurs en matière de performances (une f-mésure de 90.9% pour le système de Ouadah). Alors que la valeur la plus basse concerne la catégorie des systèmes à apprentissage automatique avec une f-mesure de 55.23% pour le système ANERsys 1.0.

Pour ce qui est des systèmes à base de règles, ils disposent d'une moyenne de f-mesure de plus de 87% qui est une valeur très proche de celle des systèmes hybrides.

A partir des remarques précédentes on conclut que le choix de l'approche de notre système (à base de règle) est judicieux.

7. Conclusion

Dans ce chapitre on a essayé d'exposer la problématique de la reconnaissance des entités nommées arabes et surtout les difficultés confrontées, les approches utilisées et les systèmes réalisés dans ce domaine. Au terme de ce chapitre, on conclut que la reconnaissance des entités nommées arabes est encore en stade de la recherche scientifique et n'a pas encore fait le passage tant souhaité au domaine industriel. Cette conclusion est tiré de la littérature investit pendant la réalisation des travaux de cette thèse.

Dans les perspectives, on espère que les travaux vus précédemment aboutiront à des systèmes commerciaux de haut niveau industriel.

Chapitre III

Relations entre les entités nommées

1. Introduction

Le texte brut contient de nombreuses entités réelles liées par de nombreuses relations sémantiques. L'identification des relations entre entités est d'une importance capitale pour de nombreuses tâches sur le Web telles que la recherche d'information, l'extraction d'informations et la fouille de réseaux sociaux (Bollegala et al. 2010). Une relation sémantique qui existe entre deux objets donnés (par exemple, des concepts, des mots ou des entités nommées) peut être définie de deux manières : par extension ou par intension. Une définition extensionnelle d'un concept formule son sens en spécifiant chaque objet qui relève de la définition du concept. D'autre part, une définition intensionnelle d'un concept formule son sens en spécifiant toutes les propriétés qui sont nécessaires pour atteindre cette définition.

Par exemple, considérons la relation *Acquisition* entre deux sociétés. Une définition extensionnelle de la relation *Acquisition* énumère toutes les paires d'entités entre lesquelles se trouve une relation d'*Acquisition* (par exemple (YouTube, Google), (Powerset, Microsoft), etc. Alternativement, nous pouvons exprimer la relation d'*Acquisition* intentionnellement en indiquant les différentes manières dont nous pouvons exprimer une *acquisition* entre deux entreprises X et Y telles que *X est acquise par Y, X est achetée par Y* ou *Y est vendue à X*.

Par exemple (Nakamura-Delloye 2011), la définition extensionnelle de la relation *EstPrésident* entre un individu X et une organisation Y énumère toutes les paires d'EN entretenant cette relation comme :

C(*EstPrésident*) ={(Joseph S. Blatter, FIFA), (Martin Hirsch, Emmaüs France), ...}.

Une définition intensionnelle de la relation R spécifie un ensemble de propriétés permettant d'identifier cette relation, noté P(R). Ainsi, la même relation *EstPrésident* peut être définie de manière intensionnelle par des patrons lexicaux tels que :

P(*EstPrésident*) ={"X est le président de Y", "X, président de Y", ...}

Extraire les relations entre les entités a reçu beaucoup d'attention dernièrement. Dans les moteurs de recherche niveau objet, il est particulièrement important de déduire les relations d'entités à partir du Web pour construire automatiquement, un graphe de relation d'entités pour relier toutes les informations extraites ensemble. Contrairement aux moteurs de recherche niveau document, pour lesquels un utilisateur saisit un mot clé et récupère un ensemble de documents, dans un moteur de recherche niveau objet, les utilisateurs recherchent une entité particulière ou une relation entre entités.

Dans les systèmes ouverts d'extraction d'information (Open IE), l'objectif est d'extraire un grand ensemble de tuples relationnels sans avoir besoin d'aucune intervention humaine. Les relations extraites peuvent alors être utilisées pour répondre à des questions en langage naturel. Dans le domaine biomédical, l'identification des relations entre les protéines et les maladies est utile pour découvrir les effets secondaires potentiels de divers médicaments. En

dépit de ses nombreuses applications, extraire des relations sémantiques entre entités dans un texte est difficile pour plusieurs raisons.

- Tout d'abord, une relation sémantique unique peut être exprimée à l'aide de multiples modèles lexicaux. Par exemple, mis à part le modèle *X a acquis Y*, une acquisition entre deux entreprises X et Y peut être exprimée en utilisant des modèles tels que *X a acheté Y, X a achevé son acquisition de Y*, etc.
- Deuxièmement, il peut exister plus d'une relation sémantique entre une paire d'entités. Par exemple, avant qu'une relation d'*Acquisition* soit établie entre deux sociétés, ces sociétés peuvent avoir une relation *Compétiteur*. Un système d'extraction de relation doit découvrir les différentes relations qui existent entre une paire d'entités.
- Troisièmement, les entités elles-mêmes pourraient avoir des variantes. Par exemple, *Microsoft Corp.* est souvent désigné comme *le géant du logiciel de Redmond*. Il n'est pas possible de spécifier manuellement toutes les variantes de noms différentes d'une entité.
- De plus, l'échelle et l'hétérogénéité du texte Web interdisent l'utilisation d'approches prenant beaucoup de temps et spécifiques au domaine qui nécessitent des techniques profondes de traitement de langage.

Une relation d'entité peut être établie entre deux EN ou plus, comme une personne, une organisation, un lieu ou un moment précis. Les relations entre les EN peuvent être binaire, telles que appartenance-personne ou organisation-lieu, ou peuvent impliquer davantage d'entités; Par exemple, [une personne] est dans [un lieu] à [un moment précis]. La relation entre entités est habituellement exprimée sous forme de prédicat et sert à établir des relations telles que *deux personnes travaillaient à la même organisation en même temps* (Ben Hamadou et al. 2010).

Une relation sémantique existant entre deux éléments peut être définie de deux manières : définition extensionnelle et définition intensionnelle. Une définition extensionnelle de la relation R consiste à créer la liste complète des instances de cette relation, notée C(R)

2. Définitions

L'extraction de relations entre entités nommées n'est pas un problème nouveau et a été formalisée officiellement pour la première fois en tant que tâche indépendante et réutilisable lors de la conférence Message Understandig Conference de 1998 (MUC, 1998). Le but est de détecter des relations entre entités nommées et de structurer les résultats afin d'alimenter une base de données. Plus tard, les travaux motivés par la campagne Automatic Content Extraction (ACE2004) ont fait émerger une définition : on appelle relation, un lien significatif entre entités nommées explicité dans un texte.

Krause (2011) a formalisé la définition d'une relation sémantique entre les EN sous la forme suivante :

Soit *t* un type d'entité nommée et soit EN_t l'ensemble de toutes les entités nommées de type *t*. Soit *T* un ensemble de types d'entité nommée et soit *n* = |*T*|.

Ensuite, tout ensemble R avec

$$\mathrm{R} \subseteq \times_{t \epsilon T}\ \mathrm{EN}_t$$

est appelé une relation n-aire.

Un exemple de relation sémantique est la relation de mariage :

$$R_{mariage} \subseteq EN_{personne} \times EN_{personne} \times EN_{date} \times EN_{lieu},$$

qui décrit à quelle *date* et à quel *lieu* deux *personnes* se sont mariées.

Maintenant, la tâche d'extraction de relation peut être définie comme de trouver toutes les mentions de relations sémantiques R_1, R_2, … données dans des textes en langage naturel et d'en extraire les occurences (Krause et al. 2012).

En cours de son définition des relations entre les EN, Ezzat (2010) distingue deux types de relations :

- Les relations statiques ou faits représentent essentiellement des états. Ce qu'on appelle état se caractérise par l'absence de changement. Un état qui est vrai pour un intervalle donné est vrai pour tout point de cet intervalle. C'est donc un lien stable et avéré entre deux entités nommées.

Exemple : *Arisem* est une filiale du *Groupe Thales*.

- Les évènements peuvent être assimilés à une phrase d'action et mettent en cause plusieurs entités (l'acteur, la cible et l'évènement particulier qui est défini par le prédicat et ses arguments par exemple), qui apportent une information nouvelle sur les participants et qui peuvent avoir une localisation spatio-temporelle implicite ou non.

Exemple : Le groupe *Thales* a racheté *Arisem* en *Mars 2004*.

3. Les classes de relations

La classification des relations entre les EN consiste à identifier les types de relations possibles en les affectant à des classes sémantiques bien déterminées. Dans une première approximation, on peut utiliser les types de paires possibles d'EN (PERS-ORG, PERS-PERS, PERS-LIEU, etc). On peut améliorer cette classification en tenant compte de la sémantique associée à la relation. Ainsi, plusieurs classifications ont été proposées dans la littérature et dans des compagnes d'évaluation. Dans le cadre d'ACE 2004, sept classes de relations entre les EN ont été définies. Ces classes peuvent se décliner à leur tour en sous-classes. Au total, on distingue 7 classes de relations et 23 sous-classes. Les sept classes de relations sont :

- **EMP-ORG :** ce type décrit une relation entre les EN de type *Personne* et *Organisation*, dans laquelle une personne peut par exemple être titulaire, membre ou a une fonction au sein d'une entreprise.
- **PHYSICAL** : cette catégorie inclut les sous-classes telles que
  - PHYS.Located qui capte l'emplacement physique d'une entité,
  - la catégorie PHYS.Near qui signifie qu'une EN est explicitement proche d'une autre EN,
  - et PHYS.Part-Whole qui prélève une relation physique entre des entités et leurs parties.
- **PER-SOC** (Personal/Social) : ce type décrit les relations permettant de relier deux EN de type *Personne*. Ainsi, l'ordre des arguments n'influe pas sur ce type de relations. Cette catégorie inclut les relations
  - SOC.Business désignant les relations entre des EN du domaine professionnel,
  - SOC.Family qui capte les relations familiales telles que (frère, sœur, etc.)

    - et SOC.Lasting qui concerne les relations sociales entre les entités telles que (ami, collège, etc.).
- **GPE-AFF** (GPE/Affiliation) : décrit les relations entre les entités de type *Personne*, *Organisation* et GPE (GeoPiletical Entity), lorsque plus d'un aspect de la GPE est référencé par le contexte du texte. Ce type de relation englobe la relation
    - Citoyen/Résident qui relie les EN *Personne* et GPE,
    - la relation Based-in qui relie les EN de type ORG et GPE
    - et la relation Other qui inclut d'autres relations n'appartenant pas aux deux sous classes précédentes.
- **OTHER-AFF** (Person/ORG Affiliation) : désigne la relation d'emploi entre les personnes et leurs employeurs (Employment) où on relie une EN *Personne* à une EN de type *Organisation*. Elle inclut aussi la relation
    - Ownership d'une *Organisation* par une EN de type *Personne*
    - et une relation Founder qui représente la relation entre un agent (Personne, Organisation, ou GPE) et une organisation ou GPE établis ou mis en place par cet agent.
- **ART** (Agent-Artifact) : décrit les relations entre des agents et des artéfacts. Elle renferme les relations
    - User-Owner qui désigne qu'un agent est le propriétaire d'un artéfact,
    - Inventor/Manufacturer où un agent est en relation d'inventeur/manufacturier avec un artéfact quand cet agent cause l'apparition de l'artefact
    - et Other.
- **DISC** (Discourse) : inclut une relation de partie-totalité
    - part-whole
    - et Membership, ces deux relations sont établies uniquement pour l'objectif du discours.

À ces classes, un ensemble des sous classes est affecté, comme illustré dans le Tableau 13 :

	Pers	**Org**	**GPE**	**Loc**	**Fac**
Pers	Per_Social.Bus Per_Social.Family Gen_Aff.ideology Gen_Aff.CRRE Per_Social.Lasting	Org_Aff.Employement Org_Aff.ownership Org_Aff.Student/Alum Org_Aff.Investor/shareholder Org_Aff.Membership Org_Aff.founder Org_Aff.CRRE Org_Aff.Sport_Affiliation	Physical.located Physical.near Org_Aff.Employement Org_Aff.Investor/shareholder Org_Aff.founder Gen_Aff.CRRE	Physical.located Physical.near Org_Aff.CRRE	Physical.located Physical.near Agent/Artifact.UOIM
Org		Part_whole.subsidiary Org_Aff.Investor/shareholder Org_Aff.Membership	Part_whole.subsidiary Org_Aff.Investor/shareholder Gen_Aff.Loc/Origin	Gen_Aff.Loc/Origin	Agent/Artifact.UOIM
GPE		Org_Aff.Investor/shareholder Org_Aff.Membership	Physical.near Part_whole.geographical Org_Aff.Investor/shareholder	Physical.near Part_whole.geographical	Agent/Artifact.UOIM

Loc			Physical.near Part_whole.geographical	Physical.near Part_whole.geographical	Physical.near Part_whole.geographical
Fac			Physical.near Part_whole.geographical	Physical.near Part_whole.geographical	Physical.near Part_whole.geographical

Tableau 13 : Sous classes de relation extrait de la base d'ACE2004 (Boujelben 2015).

Afin de mieux comprendre ces différentes classes, le tableau suivant présente des exemples illustratifs pour chaque classe d'une relation. Ces classes sont listées par ordre décroissant selon la fréquence de leur occurrence dans le corpus d'ACE.

Relation	Exemple
EMP-ORG	Le PDG de Microsoft
PHYS	Smith est allé à un hôtel au Brésil
GPE-AFF	Stéphanie louis, USA
PER-SOCIAL	Stéphanie est le collègue de Smith
DISC	chacun d'entre eux
ART	Arabie saoudite s'est engagée à acheter la société de télécommunications
OTHER-AFF	école du peuple français

Tableau 14 : Exemples sur les relations sémantiques extraits de (Boujelben 2015)

Kevers (2006) a développé la liste des relations à partir desquelles découlent des évènements. Chaque relation est caractérisée par une cardinalité indiquant sa fréquence d'apparition par rapport à une personne. Les auteurs supposent que toute relation implique l'existence de son inverse. A titre d'exemple, pour l'évènement *naissance*, la relation *X a pour parent Y* implique que *Y est parent de X*.

Krause et al., (2012) ont défini quelques classes de relations ciblées appartenant à trois domaines : People, business et Award/Atribution.

- **People :** inclut les relations de mariage, parent, naissance, enfant, etc.
- **Business :** désigne les relations au sein d'une organisation, création d'une entreprise, occupation d'une fonction, etc.
- **Award/Atribution :** nomination et attribution, honneur d'attribution, etc.

Les relations utilisées sont prises de la base FreeBase. Ces domaines incluent un sous-ensemble de 39 sous-classes de relations (Boujelben 2015).

D'autres travaux ont proposé des classifications spécifiques à des domaines d'étude. Par exemple, Embarek et Ferret (2007) ont défini quatre types de relations dans le domaine biomédical, comme suit :

- Traite : entre les EN *Maladie* et *Traitement*,
- Soigne : entre les EN *Maladie* et *Médicament*,

- Détecte : entre les EN *Maladie* et *Examen*,
- Signe : entre les EN *Maladie* et *Symptôme*.

Finalement, il importe de mentionner les travaux qui se sont limités à quelques couples d'EN. Parmi lesquels nous pouvons citer (Hamadou et al. 2010) qui se sont intéressés aux relations fonctionnelles (directeur, président, responsable, etc.) reliant une EN de type *Personne* à une autre EN de type *Organisation*. Santos et al. (2010) se sont limités à la relation *Famille* reliant les EN (PERS-PERS). (Nakamura-Delloye 2011) s'est intéressé à l'extraction de deux types de relations : relations entre les EN *Personne-Organisation* (PERS-ORG) et entre les EN *Personne-Personne* (PERS-PERS). Six relations ont été extraites pour le cas PERS-ORG qui sont: *président*, *directeur*, *secrétaire*, *avocat*, *porte-parole*. Et pour le cas de PERS-PERS, il a défini les relations *avocat*, porte-*parole* et *remplacer*.

Dans le même contexte, Serrano (2011) a abordé la tâche d'extraction de relations entre les couples d'EN PERS-PERS et PERS-ORG en utilisant un corpus de textes en anglais. L'extraction proposée se fonde sur quatre types de relations : pour le couple PERS-PERS, il a considéré les relations de parenté *isFamilyOf* et les relations diverses *isLinkedTo*, et pour le couple PERS-ORG, il a utilisé les relations d'appartenance *isMemberOf* et de direction *isLeaderOf*. Outre les difficultés rencontrées lors de l'identification des EN, la reconnaissance des relations entre celles-ci se heurte aussi à plusieurs obstacles. Nous énumérons les problèmes les plus importants dans la section suivante.

4. Difficultés d'extraction des relations

Du fait que l'extraction des relations entre les EN est similaire en terme de techniques de solutions proposées à l'extraction des EN, les problèmes posés sont aussi similaires. Certains d'entre eux sont liés à l'étape de détection de la relation et d'autres sont issus de l'étape d'identification de la classe correspondante. Il est à noter que l'étape de la reconnaissance d'EN peut influencer la tâche d'extraction automatique des relations sémantiques (Boujelben 2015). En effet, elle peut provoquer la non détection d'une relation existante, vu la non détection d'une EN (i.e. un des arguments de la relation). Aussi, la confusion au niveau de l'identification de la catégorie d'une EN peut aboutir également à une ambiguïté au niveau de l'identification de la classe de la relation.

Boujelben (2015) a énuméré quelques problèmes posés dans l'extraction de relation entre les EN tout en les illustrant à travers des exemples. On préfère garder les points de difficultés cités et de reporter des exemples en arabes comme suit :

- Il convient de souligner qu'une relation sémantique peut être explicite, quand elle s'exprime à travers un ou plusieurs mots de la phrase (verbe, nom, etc.) comme le montre l'exemple أحمد علي <u>منتسب</u> إلى مخبر اللسانيات الحاسوبية \ÂHmd ςly yntsb Ălý mxbr AllsAnyAt AlHAswbyħ*Ahmed Ali <u>est affilié</u> au laboratoire de linguistiques computationnelles*. Le problème ici est de distinguer entre une relation qui s'exprime à travers un mot simple, et une autre qui est identifiée par une séquence de mots comme le montre l'exemple أحمد داود أغلو <u>وزير خارجية</u> تركيا\ÂHmd dAwd Âγlw wzyr xArjyħ trkyA*Ahmed Daoud Oughlou <u>le ministre des affaires étrangères</u> de la Turquie*.
- Une relation peut aussi être implicite dans le cas où elle est déduite à travers le contexte sémantique de la phrase comme le montre l'exemple أحمد علي، مخبر اللسانيات الحاسوبية\ÂHmd ςly, mxbr AllsAnyAt AlHAswbyħ*Ahmed Ali, laboratoire de linguistiques computationnelles*. Dans ce cas, il est nécessaire d'analyser le contexte et la sémantique de la phrase pour dégager une telle relation.

- Dans la même ligne, les mots déclencheurs d'une relation reliant deux EN peuvent être localisés dans des positions différentes par rapport à l'emplacement des EN :
    - Dans le contexte droit[24] de la première EN : C'est le cas où la relation se situe parmi les premiers mots localisés avant la première EN de la phrase. Dans l'exemple في إطار زيارة عبد المالك سلال لغرداية\fy ĂTAr zyArħ ςbdAlmAlk slAl lγrdAyħ*Dans le cadre de la visite de Abdelmalek Sellal à Ghardia*, la relation est exprimée par le mot déclencheur *visite* situé avant la première EN.
    - Dans le contexte du milieu : La relation est à identifier dans l'ensemble des mots entre deux EN dans une phrase. C'est le cas du 1er exemple et le 3ème, où les mots déclencheurs *est affilié* et *le ministère des Affaires étrangères* appartiennent au contexte milieu.
    - Dans le contexte gauche : La relation est exprimée dans l'ensemble des mots se trouvant après la dernière EN de la phrase, comme le montre l'exemple فاطمة وسمية صديقتان لا تتفرقان\fATmħ wsmyħ SdyqtAn lA ttfrqAn*Fatima et Soumeya sont des amies inséparables* où le mot déclencheur *amies* est situé dans le contexte gauche.
- En outre, l'existence des EN n'implique pas systématiquement l'existence d'une relation. Cette situation concerne les cas où les deux EN ne sont pas liées syntaxiquement comme illustré par l'exemple قال مصطفى أن والده متواجد حالياً بمكة المكرمة\qAl mSTfý Ân wAldh mtwAjd HAlyAã bmkħ Almkrmħ*Mostefa a dit que son père est à la Mecque* et où les deux EN sont liées syntaxiquement comme le montre l'exemple ذهب أحمد وأخي صالح إلى الكلية\ðhb ÂHmd wÂxy SAlH Ălý Alklyħ*Ahmed et mon frère Salah sont allés à faculté*. Dans le 1er exemple, les EN *Mostefa* et *Mecque* ne sont pas reliées sémantiquement, alors qu'elles appartiennent à deux propositions différentes de la même phrase. Dans le 2ème exemple, les deux EN de type *Personne, Ahmed* et *Salah* sont non reliées sémantiquement, alors qu'elles appartiennent à la même proposition.
- De plus, on peut signaler qu'une relation peut être induite par d'autres relations explicites dans une même phrase. En effet, en considérant l'exemple سافر أحمد الى تركيا برفقة عيسى\sAfr ÂHmd Alý trkyA brfqħ ςysý*Ahmed a voyagé en Turquie avec Aissa*, nous pouvons extraire deux relations : la première exprimée par le mot déclencheur *a voyagé* entre les EN *Ahmed* et *Turquie*, et la deuxième identifiée par le mot *avec* reliant les EN *Ahmed* et *Aissa*. À partir de ces deux relations extraites, nous déduisons une troisième relation *a voyagé*) entre les EN *Aissa* et *Turquie*.
- De plus, les relations sémantiques peuvent être exprimées en forme négative comme le montre l'exemple ألغى محمد سفره الى سوريا\Âlγý mHmd sfrh Alý swryA*Mohamed a annulé son voyage en Syrie*. Ceci est exprimé par le verbe *annuler* qui a précédé le mot déclencheur de la relation *voyage*.

[24] Du fait que le système d'écriture de l'arabe est de droite vers la gauche à l'opposé des langues d'écriture de gauche vers la droite, le lecteur peut se trouver dans la confusion pour comprendre les exemples cités dans cette section. Ici on désigne le contexte droit dans le texte arabe qui correspond au contexte gauche dans le texte français et vice versa.

- Aussi, une relation peut être exprimée en forme négative en utilisant les particules de négation comme c'est illustré dans l'exemple شيماء ليست بزميلة رقية\šymA' lyst bzmylh rqyh*Shaimaa n'est pas une collègue de Rokiya*.
- En plus, l'extraction de relations se heurte au problème de relations multiples qui peuvent relier la même paire d'EN dans une même phrase. Ceci est présenté clairement dans l'exemple أحمد زميله صالح ناقش\nAqš ÂHmd zmylh SAlH*Ahmed a discuté avec son collègue Salah* où deux relations exhibées à la fois qui sont *a discuté avec* et *collègue* pour relier les EN *Ahmed* et *Salah*.
- Finalement, on mentionne le problème de discontinuité des relations multiples concernant la même EN. Ceci peut être démontré dans l'exemple منصف المرزوقي أستاذ بـ س.م.ب.هـ ببوبيني ورئيس الرابطة التونسية لحقوق الإنسان\mnSf Almrzwqy ÂstAð b_ s.m.b.h_ bbwbyny wrŷys AlrAbTħ Altwnsyħ lHqwq AlĂnsAn\ *Moncef Marzouki est professeur au SMBH de Bobigny et président de la Ligue tunisienne des droits de l'homme* où on a la même EN *Moncef Marzouki* de type *Personne* est reliée par une relation *professeur* à une EN *SMBH de Bobigny*, et *président* à l'EN *Ligue tunisienne des droits de l'homme*.

5. Problèmes spécifiques à l'extraction de relations entre les EN arabes

Outre les problèmes cités pour l'extraction des entités nommées à partir d'un texte arabe et énumérés dans la section 6.2 de la page 49 il y a d'autres difficultés spécifiques à l'arabe et liées à la tâche d'extraction de relations entre les EN.

5.1. La polysémie

La présence des formes polysémiques dans une EN en langue arabe, permet d'amplifier les difficultés de détection des relations qui la relie avec d'autres EN. Exempele أكرم هو شاب يعيش في فرنسا\Âk.ram huw šaAbũ yaçiyš fiy firan.sA\. En considérant l'exemple, la même phrase peut être analysée de deux façons différentes. En effet, dans le cas où أكرم \Âk.ram\ est considéré comme une EN de type *Personne*, la phrase sera analysée comme *Akram est un jeune qui habite en France*. Ainsi, une relation exprimée par le verbe *habite* intervient entre les deux EN *Akram* et *France*. Dans le cas, où le mot أكرم \Âk.ram\ est pris comme étant un adjectif superlatif signifiant *le plus généreux*, ce même exemple sera analysé comme suit *Le jeune le plus généreux habite en France* et par conséquent, on n'aura pas de relation.

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

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

- VSO :

سافر أحمد إلى مصر\saAfara ÂH.madu Ălý miS.ra*Ahmed s'est rendu en Egypte*.

- SVO :

أحمد **سافر** إلى مصر\ ÂH.madu saAfara Ălý miS.ra*Ahmed s'est rendu en Egypte*.

- OVS :

أحمد **سافر** إلى مصر\ Ălý miS.ra saAfara ÂH.madu\ *en Egypte, Ahmed s'est rendu.*

Dans ces exemples le mot déclencheur سافر\saAfara*se rendre* est situé avant la première EN dans le premier exemple, entre les deux EN dans les deux autres exemples.

5.3. La non voyellation des textes arabes et l'ambigüité de classification d'une relation

Ce problème concerne notamment la classification des relations entre les EN. Il provient des interprétations multiples des mots déclencheurs dépourvus de signes de voyellation. Ces interprétations affectent la classe des relations concernées.

5.4. Le manque de ponctuation dans les textes arabes

Un degré supplémentaire de difficulté est imposé à l'extraction des relations en raison d'un manque de ponctuations régulières associé à des phrases longues (i.e., renfermant plusieurs propositions reliées). Cela génère une ambigüité au niveau de l'identification des relations entre EN éloignées comme le montre l'exemple ci-dessous extrait de (Boujelben 2015). Ce problème peut être atténué par un traitement préalable de segmentation de la phrase en prépositions.

وفق ما ذكرته وكالة الأناظول للأنباء أن السيد أحمد داود أغلو وزير خارجية الجمهورية التركية التقى مع السيد فلاديمير ماكاي وزير خارجية روسيا البيضاء في اسطنبول بتاريخ 31 تشرين الأول/أكتوبر 2013 حيث استمر اللقاء قرابة 35 دقيقة كما أوضح فيه أن أنقرة أخطرت دمشق وحزب العمال الكردستاني بعملية إجلاء جنود يحرسون ضريح سليمان شاه شمال سوريا.

\wfq mA ðkrth wkAlħ AlÂnAĎwl llÂnbA' Ân Alsyd ÂHmd dAwd Âγlw wzyr xArjyħ Aljmhwryħ Altrkyħ Altqý mç Alsyd flAdymyr mAkAy wzyr xArjyħ rwsyA AlbyDA' fy AsTnbwl btAryx 31 tšryn AlÂwl/Âktwbr 2013 Hyθ Astmr AllqA' qrAbħ 35 dqyqħ kmA ÂwDH fyh Ân Ânqrħ ÂxTrt dmšq wHzb Alçmal Alkrdstany bçmlyħ ĂjlA' jnwd yHrswn DryH slymAn šAh šmAl swryA.\

Comme a déclaré l'agence Anadolu de presse, M. Ahmed Davutoglu le ministre des Affaires étrangères de la Turquie a rencontré M. Mackay Vladimir ministre des Affaires étrangères du Biélorussie à Istanbul le 31 Octobre 2013, où la réunion a demeuré environ 35 minutes et il a déclaré qu'Ankara a informé Damas et le PKK du processus d'évacuer les soldats gardant le tombeau de Sulaiman Shah au nord de la Syrie.

En examinant cet exemple, certains triplets sont faciles à extraire tels que :

- أحمد داود أغلو، وزير خارجية، الجمهورية التركية \ÂHmd dAwd Âγlw wzyr xArjyħ Aljmhwryħ Altrkyħ\ *Ahmed Davutoglu le ministre des Affaires étrangères de la Turquie* : qui montre que les deux EN de type *Personne* (أحمد داود أغلو\ÂHmd dAwd Âγlw*Ahmed Davutoglu*) et de type *Lieu* sont reliées par la relation prédite par le mot (وزير خارجية\wzyr xArjyħ*ministre des affaires étrangères*).
- ضريح سليمان شاه، شمال، سوريا\DryH slymAn šAh šmAl swryA\ *le tombeau de Sulaiman Shah au nord de la Syrie.* : qui présente une relation de localisation exprimée par le mot (شمال\šmAl*nord*) entre deux EN de type *Lieu*.

Tandis que certains autres triplets se trouvent face à des difficultés pour repérer les EN reliées. À titre d'exemple, on mentionne le triplet <أحمد داود أغلو, التقى , فلاديمير ماكاي> où les deux EN de type *Personne*, à savoir (أحمد داود أغلو\ÂHmd dAwd Âγlw*Ahmed Davutoglu*) et (فلاديمير ماكاي\flAdymyr mAkAy*Mackay Vladimir*) sont reliées par la relation exprimée par le mot (التقى\Altqý*a rencontré*). La difficulté d'extraire ce triplet est causée par les EN qui sont éloignées dans la phrase et discontinues par d'autres EN. De plus, ce même triplet peut être enrichi par d'autres EN pour inclure la date 31 تشرين الأول/أكتوبر 2013\31 tšryn AlÂwl/Âktwbr 2013 *le 31 Octobre 2013* et le lieu de la rencontre des deux ministres des affaires étrangères اسطنبول\AsTnbwl*Istanbul*. En outre, on peut révéler le problème de discontinuité des relations multiples concernant la même EN أنقرة\Ânqrħ*Ankara***,** comme le montre les deux triplets suivants :

- <أنقرة , أخطرت , دمشق> qui montre que les deux EN de type *Lieu* sont liées par la relation exprimée par le verbe أخطرت\Âax.Tarat*a informé*.
- <أنقرة , أخطرت ,حزب العمال الكردستاني> qui présente la même relation exprimée par le verbe أخطرت\Âax.Tarat*a informé* entre deux EN de type *Lieu* et *Organisation*.

6. Les approches proposées

La section suivante ainsi que ses sous-sections sont inspirées du chapitre deux de (Boujelben 2015). L'extraction des relations sémantiques entre EN se considère comme une tâche incontournable pour diverses applications comme la construction d'ontologie, les systèmes de question-réponse, le résumé automatique ainsi que l'extraction d'information. De ce fait, plusieurs travaux ont été réalisés dans le but d'extraire les relations entre EN. Pour étudier ces travaux, on présente dans les sections suivantes un aperçu des approches couramment employées tout en s'appuyant sur les exemples représentatifs des systèmes d'extraction de relations entre les EN.

Les travaux étudiés dans les sections suivantes foisonnent de diverses techniques et idées. Ils peuvent être classifiés en trois grandes approches : l'approche linguistique, l'approche statistique et l'approche hybride. La première approche repose principalement sur l'utilisation des grammaires formelles construites manuellement par un expert-linguiste. La seconde méthode à base d'apprentissage exploite des techniques statistiques pour apprendre des régularités en se basant sur un grand nombre d'exemples représentatifs des relations à étudier. Finalement, l'approche hybride a été explorée récemment. Elle consiste à combiner les deux approches précédentes afin d'améliorer la performance des systèmes d'extraction de relations entre les EN.

6.1. Approche à base de règles

L'approche linguistique est fondée sur l'intuition humaine, avec la construction manuelle des modèles d'analyse, le plus souvent sous la forme de règles contextuelles. C'est pourquoi, cette approche est appelée aussi une approche à base de règles. Elle se focalise sur l'information linguistique visant à extraire les relations entre les termes qui modélisent les phénomènes langagiers. Ce genre de méthodes prend la forme des patrons d'extraction permettant la description des enchaînements possibles des syntagmes nominaux. Ces patrons exploitent généralement des informations d'ordre morphosyntaxique comme les mots déclencheurs où des informations contenues dans des ressources (lexiques ou dictionnaires). Ce type d'approche fut largement répandu pour l'extraction de relations entre les EN, voire majoritaire durant les années 1990, au temps des premières conférences MUC avant que l'apprentissage ne fasse son apparition dans le domaine.

La plupart des travaux sur l'extraction des relations entre EN qui repose sur la méthode à base de règles, est dédiée à des domaines spécialisés. Par exemple, une part notable des travaux a porté sur l'étude du domaine biomédical à partir du constat que c'est un domaine étendu et complexe, constitué de nombreuses sous-disciplines et qui occupe une place prédominante sur le plan tant humain qu'économique. Parmi ces travaux, on cite

- le travail de (Grishman et al. 2002) qui s'intéresse à la détection d'événements épidémiques au moyen d'un transducteur à états finis.
- D'autres travaux se sont focalisés sur l'extraction des relations spatiales à partir de documents géographiques. Ces relations peuvent intervenir principalement des EN de type *Lieu*. Ces travaux ont été développés majoritairement pour les langues anglaises et chinoises.
- (Chunju et al. 2009) ont proposé un module fondé sur des règles syntaxiques formalisées selon JAPE[25] et reformulées en des grammaires par le biais de la plateforme GATE[26], pour extraire ces relations et ce en se basant sur un corpus annoté des EN géographiques et des relations spatiales existantes. Les performances atteintes par ce système sur des documents collectés de l'encyclopédie chinoise en termes de F-mesure, sont de 73,69% pour les relations de distance, de 75% pour les relations de direction et de 59,4% pour les relations topologiques. Les résultats expérimentaux ont montré que les relations spatiales sont généralement décrites à travers plusieurs chemins syntaxiques en langage naturel, en particulier celles de direction, mais que les relations typologiques sont beaucoup plus compliquées à modéliser.
- Des travaux qui ont proposé des systèmes d'extraction de relations entre les EN indépendamment du domaine. Ces systèmes sont restreints à un petit nombre de relations possibles. Nous mentionnons, tout d'abord quelques travaux qui ont cherché d'extraire en profondeur un seul type de relation sémantique tel que le travail de (Santos and Baptista 2010b) qui s'est intéressé à la relation familiale reliant la paire d'EN Personne-Personne en langue portugaise. Les auteurs se sont basés sur un ensemble de règles syntaxiques de la forme XIP[27] pour extraire et identifier la relation *famille*. L'application de ces règles sur un corpus de test collecté à partir des textes biographiques de tous les présidents de Portugal qui sont extraits de Wikipedia, a abouti à une précision élevée de 70% avec un taux de rappel très bas (33%).
- Dans le même contexte, (Serrano 2011) s'est intéressé à l'extraction des relations sémantiques entre deux paires d'EN Personne-Personne et Personne-Organisation en utilisant la plateforme de développement linguistique GATE. Ce travail a été appliqué à travers un corpus militaire. L'idée est de repérer les chemins syntaxiques entre deux entités afin de construire, par généralisation, un ensemble de patrons de relations syntaxiques spécifiques à de tel type d'entités. La première étape consiste à repérer les éléments lexicaux déclencheurs des relations à partir

[25] JAPE est le Java Annotation Patterns Engine, un composant de la plate-forme Open Source GATE. JAPE est un transducteur d'état fini qui fonctionne sur des annotations basées sur des expressions régulières. Ainsi, il est utile pour le pattern matching, l'extraction sémantique et de nombreuses autres opérations sur des arbres syntaxiques tels que ceux produits par les analyseurs de langage naturel.

[26] GATE une plateforme de développement de traitement du langage humain développée par l'Université de Sheffield et est exploitée dans une vaste variété de travaux de recherche et de projets de développement incluant l'extraction de connaissances pour l'anglais, l'espagnol, le chinois, l'arabe, le français, l'allemand, l'hindi, le cebuano, le roumain et le russe.

[27] Xerox Incremental Parser : un analyseur syntaxique de règles spécifiques permettant de lier l'analyse grammaticale à la sémantique des balises XML : http://open.xerox.com/Services/XIPParser

d’un corpus militaire. Pour chaque type de relation, une liste de lemmes, susceptible de l'annoncer est créée. Ceci permet de dégager des contextes généraux d'apparition de ces relations, servant par la suite à la construction des règles JAPE associées. Pour ce faire, une analyse de dépendances syntaxiques pour l’extraction des relations entre les entités est tout d’abord effectuée. Ensuite, l’auteur a réalisé une analyse morphologique pour obtenir le lemme de chaque mot du texte. Enfin, il a élaboré un transducteur JAPE pour représenter les règles d'extraction des relations entre les EN.

Ces méthodes basées sur l’arbre syntaxique de dépendance paraissent intéressantes. Tandis que, les relations sémantiques générées sont abstraites et non exactes. De plus, elles nécessitent une quantité de travail inutile, vu qu’on cherche à extraire seulement les relations entre les EN. De ce fait, il serait plus judicieux de localiser les phrases disposant d’au moins un couple d’EN et de ne s’intéresser qu’à leur analyse. En outre, les erreurs de l’analyse syntaxique peuvent présenter la principale source d’erreurs au niveau de l’extraction des relations entre EN. Par exemple, les syntagmes «le président du Sinn Féin, Gerry Adams», «le président palestinien Mahmoud Abbas, du Fatah» seront analysés de la même manière puisqu’ils ont la même structure syntaxique.

- Finalement, (Ezzat 2010) a proposé une approche symbolique d’extraction des relations entre les EN en se basant sur des grammaires générées d’une manière semi-automatique à partir d’un petit ensemble de phrases représentatives. La première étape de cette méthode consiste à repérer les phrases pertinentes par un linguiste à partir d’une analyse des cooccurrences d’entités qui sont repérées d’une manière automatique. Puis, l’auteur a utilisé un algorithme permettant de produire une grammaire en généralisant progressivement les éléments lexicaux exprimant les relations entre entités. Cet algorithme est de type *shift-reduce.* Il a été proposé par (Soricut and Marcu 2003) et il sert à examiner les phrases de gauche à droite en lisant les mots les uns après les autres. Par la suite, une règle de généralisation sera dégagée à partir de la quelle une grammaire sera générée. Cette méthode semi-automatique de génération d’une grammaire d’extraction de relations entre les EN représente une bonne initiative pour automatiser la tâche d’extraction des relations entre les EN. Bien que cette méthode permet un gain de temps important, elle se limite au cas des relations entre EN basées sur un prédicat verbal. De ce fait, il est nécessaire de définir plus de règles puisqu’il existe plusieurs représentations syntaxiques d’une phrase pouvant exprimer une relation. L’approche linguistique présente un niveau élevé de compréhension des textes. Cependant, elle n’est pas facile à mettre en œuvre car elle nécessite beaucoup de connaissances pour résoudre les ambiguïtés de la langue naturelle. D’autres types de travaux sur l’extraction de relations sont proposés plus récemment et ils ont montré l'efficacité des méthodes basées sur l'apprentissage statistique dans ce genre de tâches. Ceci fera l’objet de la section suivante.

6.2. Approche à base d’apprentissage

Une tendance actuelle consiste à exploiter les avancées récentes qu'a connues le domaine de l'apprentissage automatique pour résoudre de multiples tâches en TALN et en ingénierie linguistique. Le domaine de l’apprentissage automatique vise à étudier comment il est possible d'écrire des programmes qui s'améliorent par l'expérience, en se référant à des données d’apprentissage. En effet, il fait usage de techniques statistiques pour apprendre des spécificités sur de larges corpus de textes où les relations cibles ont été auparavant étiquetées appelés ainsi corpus d’apprentissage. Et par la suite, un modèle d’apprentissage sera adapté dans le but de construire automatiquement une base de connaissances à l'aide de plusieurs

modèles numériques (CRF, SVM, HMM, etc.). Nous distinguons trois types de méthodes pour l'extraction des relations sémantiques entre les EN : non supervisées, supervisées et semi-supervisées.

L'apprentissage non supervisé évoque une forme d'apprentissage effectuée à partir des données brutes (corpus non annoté). Il vise à mettre en évidence les relations reliant des EN, sans connaissance a priori de leur type et leur nombre. Ainsi, les classes de relations doivent être identifiées automatiquement à partir des textes. Cette étape est suivie probablement d'un regroupement des relations générées en fonction de leurs similarités.

Quant à l'apprentissage supervisé, il requiert des données d'apprentissage annotées, associant les données d'entrées aux résultats désirés. Un algorithme d'apprentissage supervisé (appelé aussi classifieur) apprend à classer des exemples annotés selon un modèle de classement donné. Pour atteindre ce but, il est nécessaire de disposer de deux ensembles de données : le premier, appelé corpus d'apprentissage, est destiné à l'induction du modèle de classification. Le deuxième baptisé corpus de test est destiné à la détermination de l'efficacité du modèle obtenu. L'apprentissage supervisé pour l'extraction des relations se déroule aussi en deux étapes, une phase d'apprentissage et une phase de test. La première phase consiste en l'extraction des traits d'apprentissage à partir d'un corpus annoté. L'extraction de ces informations requiert une analyse et une annotation du corpus d'entraînement. La phase de test consiste à appliquer le modèle de classification obtenu de la première phase à un corpus de test. L'objectif principal de cette approche est alors d'estimer la classe la plus appropriée à tout nouvel exemple du corpus de test. Les classifieurs les plus utilisés en extraction de relations sont : SVM, MaxEnt, CRF, arbres de décision, etc.

Finalement, vient l'apprentissage semi-supervisé qui vise à améliorer les performances en combinant les données annotées et non-annotées. En effet, de même qu'un apprentissage supervisé, ce type d'apprentissage propose un modèle permettant d'affecter des exemples à des classes prédéterminées. Cependant, cette fois-ci, on se sert également d'exemples non annotés.

6.3. Etude comparative

6.3.1 Comparaison

Dans le Tableau 15 ci-dessous on récapitule les principaux travaux dans le domaine de l'extraction des relations sémantiques entre les entités nommées. On note que dans la littérature il y a peu de travaux dans ce sujet qui traitent la langue arabe. La majorité des travaux cités dans le tableau concernent les langues latines. Les seuls travaux qui concernent l'arabe occupent les cinq dernières lignes.

En ce qui concerne les critères de comparaison, on note que la troisième colonne du tableau (domaine) est un critère décisif en relation avec les performances du système. On désigne ici par domaine l'ensemble de relations sémantiques qui peuvent traitées par le système en question. La performance du système doit être inversement proportionnelle du nombre de relations traitées.

Système	année	domaine	approche	formalisme/technique	Evaluation		
					précision	rappel	f-mesure
Grishman	2002	événements épidémiques	à base de règles	transducteurs à états finis			
Hasegawa	2004	indépendant	apprentissage non supervisé	similarité cosinus			79.5%
Zhang	2005		apprentissage non supervisé				83.5%
Chen	2005		apprentissage non supervisé	K-means, DCM et entropie			45.4%
Diem	2006		hybride	patrons lexicosyntaxiques – fréquence co-occurence	60%		
Chunju	2009	géographique	à base de règles	transducteurs à états finis - JAPE			69.36%
Santos et Batista	2010	relations familiales	à base de règles	XIP	70%	33%	
Ezzat	2010		à base de règles	grammaire formelle et algorithme shift-reduce			
Ben Abacha	2011	médical	hybride	règles pondérées - SVM	95%	94%	
Serrano	2011	militaire	à base de règles	transducteurs à états finis - JAPE			
Wang	2013		apprentissage non supervisé	similarité vectorielle, WordNet			77.3%
Ben Hamadou	2010	pers-org	à base de règles	transducteurs à états finis			70%
Alnairia	2012	spatiale	à base de règles				80.06%
Alotayq	2013	Gen-affiliation, Org-affiliation	apprentissage automatique	entropie maximale			85%
Boujelben	2015	pers, lieu, org	à base de règle	transducteurs à états finis Nooj	69.4%	58.6%	63.54%
Notre travail	2015	pers, lieu, date	à base de règles	cascade de transducteurs à états finis			

Tableau 15 : Comparaison entre les travux d'extraction des relations sématiques entre les EN.

6.3.2 Synthèse

En guise de synthèse, on note que les valeurs des métriques d'évaluation montrent la suprématie des systèmes basés sur une approche hybride à l'instar du système de Ben Abacha avec une précision de 95%. On peut aussi souligner que les travaux réalisés en langue arabe sont basés principalement sur une approche à base de règles avec une seule tentative de résoudre le problème d'extraction des relations d'une manière supervisée. Jusqu'à présent, aucun travail en langue arabe n'a utilisé une approche hybride pour l'extraction des relations entre les EN, ce qui présente un défi dans le domaine d'identification des relations pour cette langue assez riche et complexe.

7. Conclusion

Dans ce chapitre on a présenté un résumé sur les relations sémantiques entre les EN dans les différentes langues, et vu que les travaux en langue arabe sont restreints on s'est limité à exposer quelques uns. Les difficultés d'extraction spécifiques à cette langue ont été longuement discutées dans les sections de ce chapitre.

Les notions ainsi que les approches exposées du chapitre courant servent comme un bagage théorique pour la compréhension de la mise en œuvre d'extraction de relations sémantiques qui est l'objectif final de notre travail. Le lecteur doit être revenir si nécessaire pour comprendre la fin du cinquième chapitre de cette thèse.

Chapitre IV

Le module arabe d'Unitex/GramLab

1. Introduction

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

Notre tâche étant de traiter un corpus textuel, nous allons, à cet effet, utiliser la plateforme Unitex/GramLab. Cette plateforme est un logiciel de traitement automatique de corpus qui regroupe un ensemble de programmes réalisant les différentes tâches dont l'utilisateur a besoin. Nous avons été amenés, au début de nos travaux, à constater que la plateforme dont il est question bien qu'elle prend en charge une dizaine de langues[28], les ressources nécessaires pour la langue arabe n'y étant pas incluses. Nous avons été, donc, amenés à créer le module manquant, sa publication sous la licence LGPL-LR a suivi en Avril 2012, et depuis sa création il est téléchargeable par les utilisateurs intéressés, et a fait l'objet de plusieurs publications internationales comme c'est le cas pour (Doumi et al. 2016a; Doumi et al. 2013; Doumi et al. 2016b).

Le traitement automatique de la langue arabe à l'instar d'autres langues demande la construction de grandes ressources linguistiques dans le but d'effectuer des tâches différentes. Que se soit une tâche légère telle que la vérification orthographique ou lourde telle que la traduction et la compréhension automatique du texte ; ces ressources sont cruciales. Les lexiques sont reconnus comme un pré requis fondamental pour toutes les tâches de TAL (Maurel and Guenthner 2005). Construire ses propres ressources linguistiques c'est la méthode la plus économique pour le chercheur lui permettant d'acquérir ces composantes cruciales. D'un autre coté, il est difficile et au-delà du budget des chercheurs de construire des ressources à large couverture de la langue en question. C'est pour cette raison que les approches et les algorithmes proposés dans ce chapitre donnent une méthodologie de construction de ressources selon le besoin. Avant d'entamer l'explication des différentes méthodes et algorithmes proposés pour résoudre la problématique de construction de ressources nous allons introduire le bagage théorique en relation. Les sections suivantes fournissent au lecteur les notions théoriques pour comprendre le reste du chapitre.

2. Technologie à états finis

Dans la construction des ressources linguistiques, des analyseurs et des grammaires locales, on utilise la technologie à états finis, précisément les transducteurs à états finis. Pour cette raison il est nécessaire et utile d'exposer dans les sections suivantes toutes les notions théoriques et formelles en relation avec la théorie des automates. Comme il est aussi nécessaire de mettre en évidence l'usage de cette théorie dans le TAL en général et son usage

[28] Le module arabe à été ajouté le mois d'Avril 2012 en collaboration entre Alexis Neme et l'auteur de cette thèse et sous la direction d'Eric Laporte et Denis Maurel.

pour construire et stocker des dictionnaires électroniques de la langue en question (l'ASM). Les définitions qui suivent sont adaptées de (Maurel and Guenthner 2005).

Définition 1

Un automate à états finis sur l'ensemble $\mathcal{L}$ est un quintuplet $\mathcal{A}=(\mathcal{Z}, \mathcal{L}, q_0, \mathcal{F}, \delta)$ où :

- $\mathcal{Z}$ est un ensemble fini non vide d'états,
- $\mathcal{L}$ est un ensemble non vide de lettres et de diacritiques (l'alphabet)
- q_0 est un élément de $\mathcal{Z}$ (état initial)
- $\mathcal{F}$ est un sous ensemble non vide de $\mathcal{Z}$ (les états terminaux)
- δ est une relation définie de $\mathcal{Z}$ x $\mathcal{L}$ à $\mathcal{Z}$ (les transitions).

On dit qu'un élément $\ell_1\ell_2\ldots\ell_n$ est reconnu par l'automate $\mathcal{A}$ si et seulement si il correspond à une séquence de *n* transitions étiquetées respectivement par ℓ_1, ℓ_2,… et ℓ_n, commençant par l'état initial q_0 et terminant par l'état final p_n

Formellement on dit que $\ell_1\ell_2\ldots\ell_n$ est reconnu par $\mathcal{A}$ SSI:

$\exists\, p_0, p_1, \ldots, p_n \in \mathcal{Z}$, tel que :
$\forall\ i=0,1,\ldots,n\text{-}1,\ p_{i+1} \in \delta(p_i, \ell_{i+1})$
$p_0 = q_0$
$p_n \in \mathcal{F}$

On appelle $\mathcal{L}^*$ l'ensemble contenant la chaîne ε vide et toutes les séquences finies composées des éléments de $\mathcal{L}$. L'ensemble des éléments de $\mathcal{L}^*$ reconnus par $\mathcal{A}$ est appelé *le langage défini par* $\mathcal{A}$. Il est noté par L($\mathcal{A}$).

Dans le cas de l'ASM, l'ensemble de l'alphabet $\mathcal{L}$ correspond à l'intervalle d'Unicode de \u0621 à \u0652. L'intervalle arabe de base code les lettres et diacritiques standards, par contre ne code pas les formes contextuelles[29].

Par exemple, la Figure 10 représente un automate reconnaissant/stockant les sept jours de la semaine arabe, les transitions ne sont étiquetées que par les consonnes et les voyelles longues. Afin de simplifier et de ne pas avoir une figure encombrée, on a considéré que les mots reconnus par cet automate sont non voyellés[30] (sans les voyelles courtes ni le signe de gémination).

[29] La table Unicode liste plus de 1260 caractères arabes pour l'écriture arabe utilisés même dans d'autres langues telles que pour persan, urdu, sindhi et les langues d'Asie central etc. En ASM, l'alphabet concerné est représenté par l'ensemble $\mathcal{L}$ = {ء, آ, أ, ؤ, إ, ئ, ا, ب, ة, ت, ث, ج, ح, خ, د, ذ, ر, ز, س, ش, ص, ض, ط, ظ, ع, غ, ـ, ف, ق, ك, ل, م, ن, ه, و, ى, ي, ً, ٌ, ٍ, َ, ُ, ِ, ّ, ْ}

[30] La longueur des mots entièrement voyelles est souvent le double de la longueur des mots non voyellés

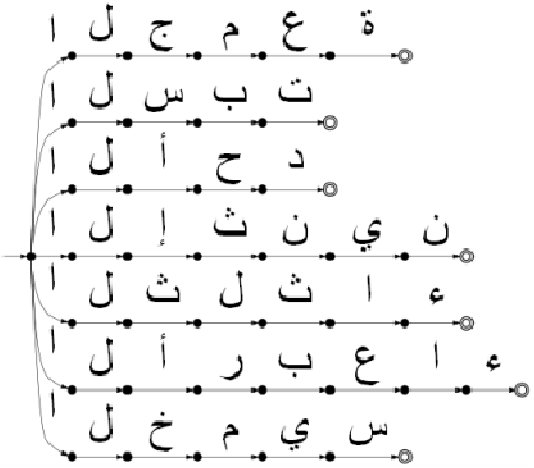


Figure 10 : Arbre lexicographique ou l'automate acyclique non déterministe qui reconnait/stocke les jours de la semaine arabe

Définition 2

Un transducteur à états finis sur l'ensemble $\mathcal{L}$ est un septuplet $\mathcal{T} = (\mathcal{Q}, \mathcal{L}, \mathcal{S}, q_0, q_f, \delta, \lambda)$

- $\mathcal{Q}$ est un ensemble fini non vide d'états,
- $\mathcal{L}$ et $\mathcal{S}$ sont deux ensembles finis non-vides de lettres et de diacritiques arabe et latins (respectivement, l'alphabet d'entrée et de sortie)
- q_0 est un élément de $\mathcal{Q}$ (état initial)
- q_f est un élément de $\mathcal{Q}$ (l'état final)
- δ est une relation définie de $\mathcal{Q}$ x $\mathcal{L}$ à $\mathcal{Q}$ (la fonction de transitions).
- λ est une fonction définie de $\mathcal{Q}$ x $\mathcal{L}$ à $\mathcal{S}^*$ (la fonction de transition de sortie)

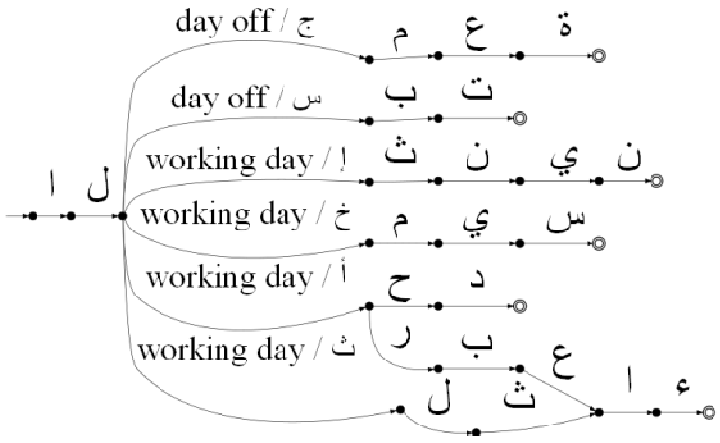


Figure 11 : Transducteur reconnaissant/stockant les sept jours de la semaine arabe et informant si le jour est ouvrable ou weekend en Algérie. Son automate sous-jacent est la déterminsation et minimisation de l'automate de la Figure 10.

L'automate $(\mathcal{Q}, \mathcal{L}, q_0, q_f, \delta)$ est appelé l'automate sous-jacent du transducteur $\mathcal{T}$, e.g., la Figure 11 montre un exemple de transducteur reconnaissant les sept jours de la semaine arabe tout en envoyant l'information qui indique si le jour est ouvrable ou weekend en Algérie. L'automate de la Figure 10 est acyclique non déterministe; ce type d'automate est utilisé pour

reconnaitre/stocker le lexique des langues naturelles, il est parfois appelé arbre lexicographique. L'automate sous-jacent du transducteur de la Figure 11 est le résultat de la déterminsation et minimisation de cet arbre lexicographique.

3. Le jeu de ressources arabe d'Unitex/GramLab

Unitex/GramLab est une plateforme open source de traitement automatique des langues, développée à l'université Paris-Est Marne-La-Vallée (Paumier, 2009). Cet outil est fondé sur la technologie des automates à nombre fini d'états (FST : Finite State Transducers) ; ainsi on peut exprimer les règles morphologiques, syntaxiques et même sémantiques d'une langue par des transducteurs, des réseaux de transitions récursifs, des grammaires algébriques et des réseaux de transitions augmentés.

Les ressources lexicales sous Unitex/GramLab doivent avoir le format et la syntaxe des dictionnaires DELA du LADL[31].

Unitex/GramLab a prévu l'ajout des modules de langues sémitiques aux dix-neuf modules de langue qui existent déjà[32]. Les options d'écriture et de lecture des textes et des graphes[33] ainsi que l'encodage (UTF16) utilisé par Unitex/GramLab permettent d'ajouter facilement la langue arabe à cette plateforme. Un package d'une langue est représenté par un répertoire qui contient toutes les ressources et les fichiers de configurations : dans les sections suivantes nous allons expliquer les différents composants du module arabe que nous avons ajoutés à Unitex/GramLab.

3.1. L'alphabet

Le package de l'arabe commence par le choix de l'alphabet et de l'ordre de cet alphabet pour effectuer un tri des unités lexicales trouvées lors d'un traitement. On a opté pour l'alphabet arabe d'Unicode[34] qui commence de \u0621 jusqu'au \u0652 dans le même ordre. Ce qui constitue les deux fichiers alphabet et alphabet_sort.

3.2. Le corpus de test

Unitex/GramLab est un logiciel sous licence LGPL[35], donc le corpus de test distribué avec le package arabe doit être libre de droits. Notre choix s'est porté sur un texte qui représente une légende arabe écrite par Ibn Toufayl (1110-1185), un philosophe et savant arabe de l'Andalousie : le roman a pour taille 96 Ko (32 pages A4) et est composé de 18 261 formes simples. Nous allons par la suite proposer un texte libre de droits et qui illustre l'ASM, i.e. un texte qui date au plus du début 19ème siècle

Notre corpus est partiellement diacritisé du fait qu'il contient des mots comprenant des signes diacritiques tels que les signes de nounation[36] et de gémination (shadda). Le reste du corpus qui représente la grande partie du texte n'est pas diacritisé.

---

[31] Laboratoire d'Automatique Documentaire et Linguistique

[32] Site officiel de l'Unitex : http ://www-igm.univ-mlv.fr/~unitex (consulté en Janvier 2011)

[33] Dans la terminologie d'Unitex/GramLab, on désigne par graphe les automates, transducteurs et réseaux de transducteurs conçus sous forme graphique.

[34] www.unicode.org/charts/PDF/U0600.pdf *(consulté en janvier 2011)*

[35] Site officiel de la fondation GNU http ://www.gnu.org/licenses/lgpl.html (consulté en janvier 2011)

[36] En arabe, addition de la consonne n à la suite de la désinence casuelle du nom, pour exprimer la modalité *indéfini*. Un nom est indéterminé grammaticalement quand il est nu, ce nom pouvant alors, selon son type, être affecté ou non de la nûnation (...) (CNRTL-CNRS 2012).

3.3. Le jeu d'étiquettes

Pour le choix du jeu d'étiquettes, nous avons opté pour un ensemble de 17 catégories grammaticales (cf. le Tableau 16). Ces catégories sont le résultat d'une étude comparative approfondie des jeux d'étiquettes citées dans le Tableau 3. Notons que parmi ces catégories il y a des catégories fermées et des catégories ouvertes. Les traits morphosyntaxiques sont une partie importante du jeu d'étiquettes, dans notre travail nous avons opté pour un ensemble de 21 traits.

Dans Unitex/GramLab, le jeu d'étiquettes est introduit dans des fichiers de configuration pour être utilisé ultérieurement dans des tâches de validation telle que la vérification de conformité des dictionnaires. La Figure 12 représente le fichier *morphology* de notre package arabe d'Unitex/GramLab.

```
Arabic
<CATEGORIES>
Nb : <E>, s, p, d
Gen : <E>, m, f
Temps : A, I, F, P
Pers : <E>, 1, 2, 3
Cas : <E>, a, u, i, A, U, I, o
Def : d, i
<CLASSES>
Nc : (Nb,<var>), (Gen,<var>), (Def,<var>), (Cas,<var>)
Npr : (Nb,<var>), (Gen,<var>), (Cas,<var>)
Adj :(Nb,<var>), (Gen,<var>), (Def,<var>), (Cas,<var>)
Adv :
v : (Temps,<var>), (Pers,<var>), (Nb,<var>), (Gen,<var>),
(Cas,<var>)
ve : (Temps,<var>), (Pers,<var>), (Nb,<var>), (Gen,<var>),
(Cas,<var>)
Prsl : (Pers,<var>), (Nb,<var>), (Gen,<var>)
Dmst : (Nb,<var>), (Gen,<var>), (Cas,<var>)
Rltf : (Nb,<var>), (Gen,<var>), (Cas,<var>)
Prps :
CnjCrd :
Apcp :
Sbjc :
Crbr :
Intrg :
Rstr :
Evdt :
```

Figure 12 : Le jeu d'étiquettes

4. La construction des dictionnaires DELA

Comme il a été indiqué auparavant, les ressources lexicales sous Unitex/GramLab doivent respecter le formalisme DELA de LADL, qui spécifie la structure des dictionnaires électroniques et de leurs contenus. On trouve dans DELA trois types de dictionnaires (Blondine Courtois 1994-1995) :

- dictionnaires de mots simples, DELAS et DELAF,

- dictionnaires de mots composés, DELAC et DELACF,

- dictionnaires phonémiques, DELAP et DELAPF.

Dans cette thèse, nous nous intéresserons seulement au premier type : les dictionnaires DELAS et DELAF pour construire le lexique arabe nécessaire au traitement automatique de la langue arabe sous Unitex/GramLab. Nous rappelons que les mots composés dans la langue arabe représentent un pourcentage important du lexique mais, le traitement de cette partie du lexique est régi par des règles morphologiques différentes. Une autre étude aura pour objet spécifique la morphologie polylexicale.

Dans notre dictionnaire chaque forme fléchie a deux entrées, une entrée entièrement voyellée et l'autre non voyellée et avec signe de gémination. Mais cela à notre avis ne résout pas le problème de la voyellation car les unités lexicales dans un texte de l'ASM se présentent sous trois formes ou plus (a) formes non voyellée ni signe de gémination (b) formes non voyellée et avec signe de gémination s'il se présente dans l'unité (c) formes partiellement voyellée. Le problème de la voyellation partielle peut être résolu en intégrant un algorithme qui calcule toutes les formes de voyellation possibles d'un mot, ce qui est possible grâce à l'entrée entièrement vocalisée de notre dictionnaire. Il faudra voir si nous devons stocker ces formes ou les calculer pendant le traitement.

Dans notre travail, le lexique compte deux grands groupes distincts : les catégories fermées et les catégories ouvertes.

4.1. Les catégories fermées

En arabe, à l'instar des autres langues il y a des catégories grammaticales ouvertes, capables d'être enrichies dans le temps et en fonction de l'évolution de la langue, et il y a les catégories grammaticales fermées qui sont limitées en nombre dont les fonctions syntaxiques sont bien connues, constituées majoritairement des mots outils (cf. Tableau 17).

Code	Catégorie	Catégorie d'origine
V	Verbe	Verbe
Ve	Verbe d'état	Verbe
Nc	Nom commun	Nom
Npr	Nom propre	Nom
Adj	Adjectif	Nom
Adv	Adverbe	Nom
Prsl	Pronom personnel	Nom
Dmst	Pronom démonstratif	Nom
Rltf	Pronom relatif	Nom
Intrg	Article d'interrogation	Particule
CnjCrd	Conjonction de coordination	Particule
Prps	Préposition	Particule
Sbjc	Particule de subjonctif	Particule
Evdt	Marqueur d'évidentialité	Particule
Apcp	Particule de l'apocopé	Particule
Rstr	Particule de restriction	Particule
Crbr	Marqueur de corroboration	Particule

Tableau 16 : Les catégories grammaticales du module arabe d'Unitex/GramLab

Les catégories fermées par leur nature limitée et non extensible sont communes entre l'AC et l'ASM.

Dans cette thèse, nous considérons comme catégories fermées tous les pronoms personnels, relatifs et démonstratifs, toutes les particules ainsi que les verbes d'état.

Le Tableau 17 dresse l'ensemble des catégories considérées dans notre travail comme catégories fermées. Ces catégories sont divisées en trois groupes.

Article	**Catégorie**	**Code**	**Nombre**
Verbe d'état	Verbe d'état	Ve	13
Pronoms personnels	Nom	Prsl	32
Pronoms démonstratifs	Nom	Dmst	34
Pronoms relatifs	Nom	Rltf	20
Particule de l'apocopé	Particule	Apcp	21
Conjonction de coordination	Particule	CnjCrd	8
Marqueur de corroboration	Particule	Crbr	7
Marqueur d'évidentialités	Particule	Evdt	7
Article d'interrogation	Particule	Intrg	15
Préposition	Particule	Prps	22
Particule de restriction	Particule	Rstr	9
Particule de subjonctif	Particule	Sbjc	6

Tableau 17 : La liste des catégories fermées

4.1.1 Les verbes d'état

Les verbes d'états sont des verbes particuliers, cependant leur flexion est analogue aux verbes normaux. Donc un verbe d'état possède une forme canonique et un nombre bien déterminé de formes fléchies (voir Figure 13). Les lemmes des verbes d'état sont cités dans tous les livres de la grammaire arabe sous le titre de أخوات كان\ÂaxawaAt kaAna\ et ils sont au nombre de 13 (Al-Afghani 1971).

صِرْتُ,صار.Ve:A1sm:A1sf
صِرْنَا,صار.Ve:A1dm:A1df:A1pm:A1pf
صِرْتَ,صار.Ve:A2sm
صِرْتِ,صار.Ve:A2sf
صِرْتُمَا,صار.Ve:A2dm:A2df
صِرْتُمْ,صار.Ve:A2pm
صِرْتُنَّ,صار.Ve:A2pf
صَارَ,صار.Ve:A3sm
صَارَتْ,صار.Ve:A3sf
صَارَا,صار.Ve:A3dm
صَارَتَا,صار.Ve:A3df
صَارُوا,صار.Ve:A3pm
…

Figure 13 : Extrait du DELAF d'un verbe d'état

Dans notre cas, les verbes d'état sont intégrés dans le dictionnaire des formes simples (DELAS) ; e.g. une entrée de ce dictionnaire a la forme suivante :

صا ر,$Ve2

Où SaAra représente le lemme non vocalisé, le signe $ indique le mode morphologique d'Unitex/GramLab et Ve2 représente le nom du paradigme flexionnel dans lequel le verbe d'état SaAra (devenir) se fléchit, c'est un graphe de flexion qui représente un transducteur.

Par exemple, le graphe Ve2 de la Figure 14 génère 96 formes fléchies, on note ici que les verbes d'états se fléchissent en fonction du : sujet (personne, genre et nombre), de l'aspect et du mode ; on note aussi qu'il y a des verbes d'état qui ne se fléchissent pas avec tous les aspects (Al-Afghani 1971); le Tableau 18 montre les types de verbes d'états et le nombre de leurs formes fléchies.

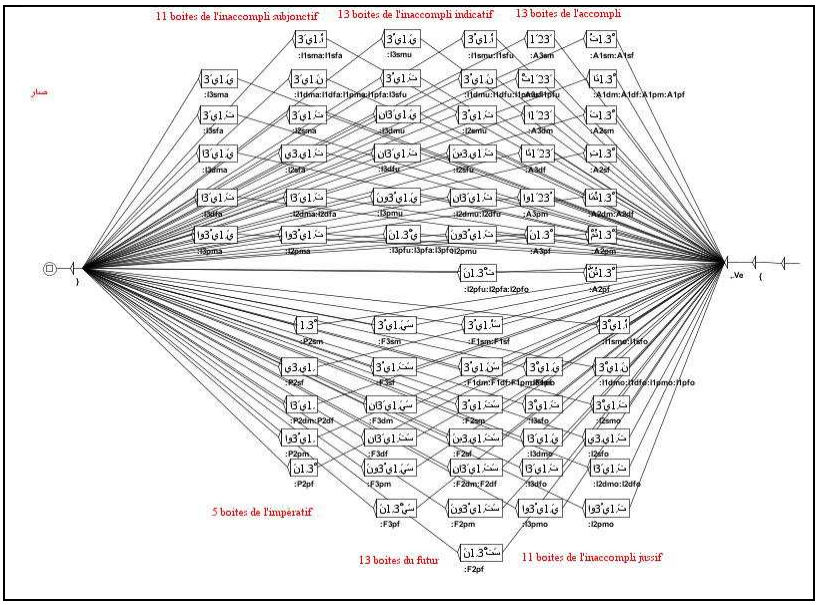


Figure 14 : Le graphe de flexion du verbe d'état صار \SaAra

Verbe d'état	**Aspects de flexion**	**Nombre de formes fléchies**
كـان \kaAna\	Tous	96
صا ر \SaAra\	Tous	96
أصبح \ÂaS.baHa\	Tous	96
بـات \baAta\	Tous	96
ظل \Ďal~a\	Tous	96
أمـسى \Âam.saý\	Tous	96
أضحى \ÂaD.Haý\	Tous	96
لـيس \lay.sa\	Accompli	18
مـابـرح \maAbariHa\	Accompli, inaccompli	36
مـا ز ال \maAzaAla\	Accompli, inaccompli	36
مـافـتـئ \maAfatiŷa\	Accompli, inaccompli	36
مـا د ام \maAdaAma\	Accompli	18
مـا انـفـك \maAAn.fak~a\	Accompli, inaccompli	36

Tableau 18 : Les verbes d'état arabes

4.1.2 Les pronoms et particules

Les formes fléchies des pronoms et particules du Tableau 17 sont calculés manuellement et ajoutées directement dans le dictionnaire DELAF. Ceci est dû au nombre limité de formes fléchies pour les pronoms et l'absence des formes fléchies pour les particules.

A titre d'exemple les pronoms démonstratifs sont fléchis suivant le genre, le nombre et le cas, i.e. 18 formes fléchies (2 genres x 3 nombres x 3 cas), les pronoms personnels sont fléchies en genre, en nombre et en personne (2 genres x 3 nombres x 3 personnes). Le nombre peut augmenter si on prend en considération les formes tolérantes aux erreurs d'écriture, e.g. le pronom démonstratif هؤلاء\haŵulA'i*ceux-ci* peut être écrit au moins en trois formes : هؤلاء \haŵulA'i\ (29 700 000 occurrences Google), هاؤلاء \haAŵulA'i\ (140 000 occurrences Google) et هأولاء \haÂuwlA'i\ (78 300 occurrences Google)[37].

4.2. Les Catégories ouvertes

On désigne par catégories ouvertes les catégories qui évoluent suivant l'évolution de la langue, toutes les catégories non citées dans la section des catégories fermées sont considérées comme ouvertes, en l'occurrence les noms et les verbes. Avant de parler de chacune de ces catégories, présentons le format du dictionnaire DELAS.

4.2.1 Le dictionnaire DELAS

Les entrées de ce dictionnaire représentent les formes canoniques des verbes et des nominaux et leurs paradigmes flexionnels correspondants. Ce dictionnaire servira plus tard à la génération du dictionnaire final. Une entrée doit contenir le lemme du verbe ou du nom/adjectif et le nom du graphe qui représente le paradigme flexionnel, la Figure 15 ci-dessous représente un échantillon de notre DELAS.

```
درس,$V1
كتب,$V1
أصر,$V10
أقر,$V10
تسامح,$V11
```

Figure 15 : Exemple de DELAS

4.2.2 Les verbes

Comme il a été mentionné à la section 5.3 de la page 24, la flexion des verbes arabes est régulière, mais le nombre de cas dépend des patrons et de la structure saine ou défectueuse du verbe. Pour notre dictionnaire les verbes viennent d'un corpus de test puis un expert linguiste à travers une interface graphique lemmatise et introduit le lemme du verbe à notre programme, un schème du lemme est calculé puis recherché dans la liste des verbes déjà établis; si le schème du verbe existe on l'ajoute le verbe directement au DELAS, sinon on calcule son graphe d'une façon automatique. La Figure 16 résume la méthode suivie pour produire les entrées dictionnaire DELAF ou DELAS. Les algorithmes 1, 2, 3 et 4 donnent plus de détails sur la partie automatique de cette méthode (Doumi et al. 2016a). Pour ajouter

[37] Résultat d'une recherche sur www.google.com effectuée en février 2011

un nouveau verbe au dictionnaire, l'utilisateur/l'annotateur[38] est guidé par un curseur dans le texte. Il parcourt le texte mot à mot quand il rencontre un nouveau verbe il introduit dans le système deux formes fléchies : la première correspond au lemme arabe i.e. le verbe à l'accompli et à la troisième personne du masculin singulier et la deuxième forme correspond à l'inaccompli de la même personne, genre et nombre. Le reste de la méthode peut être résumée comme suit :

1. A partir des deux formes fléchies mentionnées précédemment, le schème flexionnel est calculé (cf. l'Algorithme 1 et le paragraphe correspondant).
2. Le lemme du verbe et le schème calculé sont comparés à ceux des verbes déjà traités et stockés
3. S'ils s'accordent, i.e. le verbe est déjà traité ou le paradigme flexionnel (le transducteur flexionnel ou le graphe flexionnel dans le langage d'Unitex/GramLab) est déjà connu donc il ne reste qu'ajouter le verbe au DELAS et DELAF.
4. Sinon l'algorithme calcule le nouveau transducteur (cf. l'Algorithme 3 et la section correspondante)

4.2.3 Le schème flexionnel

Comme il est montré dans l'algorithme de la Figure 16, nous avons proposé un schème flexionnel qui définit la façon selon laquelle un verbe se fléchit : ainsi deux verbes ayant le même schème flexionnel sont obligatoirement fléchis de la même façon, même si une différence reste au niveau des consonnes qui composent la base. Ce schème est calculé en partant de l'idée que les caractères d'un verbe susceptibles d'être affectés par la flexion sont :

- Au début de verbe : ا, أ, إ, آ, و
- Au milieu de verbe : ا
- A la fin de verbe : ا, ى, أ, و, ي, ن, ت.
- Tous les signes diacritiques : ً, ٌ, ٍ , ْ , ّ , ً , ٌ , ٍ ّ ْ

Le schème flexionnel représente le paradigme flexionnel dans une notation compacte et transparente. Le calcul de cette forme compacte se base sur quatre principes :

1. A la différence de la morphologie arabe, la flexion dans notre algorithme repose sur le *lemme* plutôt que sur la *racine*
2. Comme il est noté, la flexion arabe affecte un sous ensemble de consonnes et voyelles (courtes et longues) du lemme dans des positions bien déterminées
3. On peut classifier les verbes arabes sur la base de l'idée que si leurs consonnes et voyelles sont affectées par la flexion, on peut déterminer les classes représentant les paradigmes flexionnels
4. On peut déterminer le paradigme flexionnel d'un verbe si on connait son schème flexionnel.

Algorithm 1 : Calculating inflectional pattern
Input : verb lemma and 3rd person masculine imperfect active word form
Output : inflection pattern

---

scheme ← ''
for (i=0 to length(lemma)-1) scheme ← scheme + coding(lemma[i], Table1)
scheme ← scheme + ',,'

---

[38] L'utilisateur du système propose ou l'annotateur doit être au moins un arabe natif s'il n'est pas expert en arabe. Manuellement, il lemmatise les verbes corpus et donne les cinq formes fléchies si nécessaire. L'annotation du corpus peut être effectuée hors ligne.

for (i=0 to length(word_form-1) scheme ← scheme + coding(word_form[i], Table1)
return scheme

Pour calculer le schème flexionnel on utilise le tableau de correspondances ci-dessous.

Caractère	ا	ى	و	ي	أ	ؤ	ئ	ء	آ	ت	ن	َ	ُ	ِ	ّ	ْ	autre
Schème	A	Y	U	I	H	V	W	h	M	t	n	a	u	i	s	o	c

Tableau 19 : Correspondance entre caractères dans le calcul du schème flexionnel

Ci-dessous, des exemples sur le résultat de calcul des schèmes flexionnels des verbes ; pour chaque exemple on donne les deux formes fléchies demandées à l'utilisateur dans l'algorithme 1, leurs translitérations HSB et la traduction en anglais.

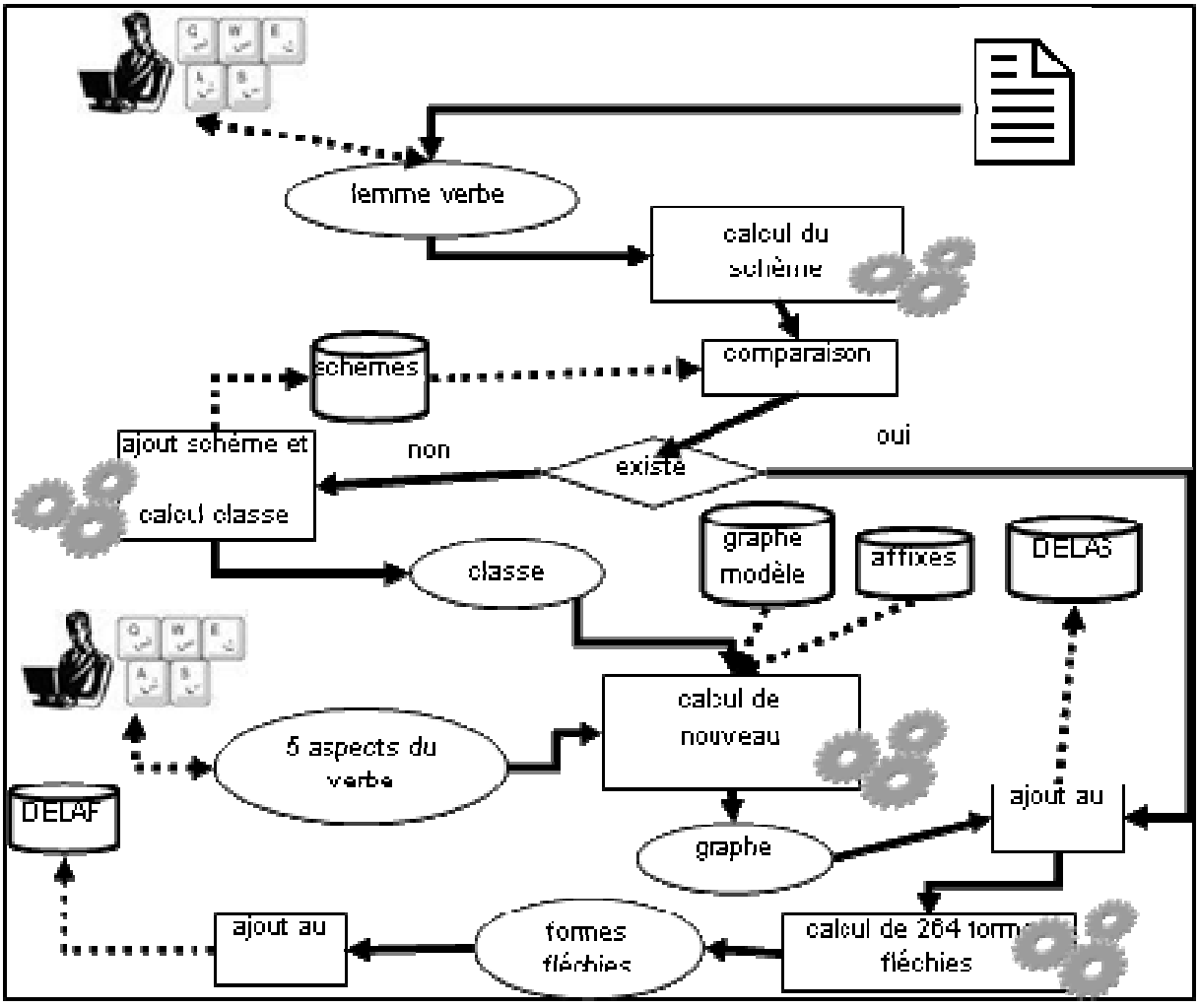


Figure 16 : Algorithme général de génération automatique des transducteurs et de flexion des verbes

Exemples de schème flexionnel

a) كَـتَبَ يَـكْـتُبُ \kataba yak.tubu\ *to write*

inflection pattern(كَـتَبَ يَـكْـتُـبُ)=cacaca cacocucu
b) كَـتَبَ يَـكْـتِـبُ \kataba yak.tibu\ *to prescribe*
inflection pattern(كَـتَبَ يَـكْـتِـبُ)=cacaca cacocicu
c) وَلَّـى يُـوَلِّـي \wal~aý yuwal~iy\ *to crown*
inflection pattern(وَلَّـى يُـوَلِّـي)=cacsaY cucacsiI
d) أدَّى يُـؤدِّي \Ad~aý yuŵd~iy\ *to lead*
inflection pattern(أدَّى يُـؤدِّي)=HcsaY cuOcsiI

4.2.4 La classe flexionnelle

Dans le but de générer le paradigme flexionnel sous Unitex/GramLab (le graphe) d'une façon automatique, on trie les verbes arabes en de multiples niveaux ; en premier lieu, on doit déterminer la classe du verbe selon l'algorithme 2, dans cet algorithme on détermine la liste des affixes à appliquer dans l'algorithme 3. Les affixes à concaténer avec le radical de la forme fléchie dépendent des consonnes et des voyelles qui composent le lemme et la deuxième forme fléchie utilisée dans l'algorithme 1. La Figure 17 montre un exemple des affixes de la première classe i.e. la classe des verbes de type كـتـب \kataba\.

Algorithm 2 : Calculating inflectional class
Input : verb lemma and 3rd person masculine imperfect active word form
Output : affixes

switch (lastCharacter(lemma), lastCharacter(word_form))
  case 'ى', 'ي': { affixes ← affixes2}
  case 'ّ', 'ّ': { affixes ← affixes3}
  case 'ا', 'و': { affixes ← affixes4}
  case 'ى', 'ى': { affixes ← affixes5}
  default : { affixes ← affixes1}
end switch
return affixes

```
**********************************
*Fichier des affixes du verbe de type :كَـتَبَ
**********************************
Nombre d'entrées : 184
Accompli actif : 0-12
Inaccompli actif : 13-86
Impératif : 87-96
Accompli passif : 97-109
Inaccompli passif : 110-183

0,-,-,1,ْتُ
1,-,-,1,ْنَـا
2,-,-,1,ْتَ
3,-,-,1,ْتِ
4,-,-,1,ْتُـمَـا
```

Figure 17 : Exemple de liste d'affixes

4.2.5 Génération automatique des graphes

Sous la plateforme Unitex/GramLab on peut élaborer manuellement un graphe à travers un éditeur graphique, e.g. l'édition d'un graphe comme celui de la Figure 19 peut dépasser une heure. Dans notre travail, le nombre de graphes de flexion est en fonction des types et sous types de verbes arabes. Ainsi ce nombre est important et par conséquent la réalisation manuelle de ce nombre de graphes devient un fardeau lourd pour l'utilisateur ; d'où l'idée de générer automatiquement ces graphes. Un algorithme a été proposé pour réaliser cette tâche, il a besoin de trois entrées : la classe de flexion, la liste des affixes et le graphe modèle. La Figure 17 montre un exemple de liste des affixes et la Figure 18 un exemple de graphe modèle.

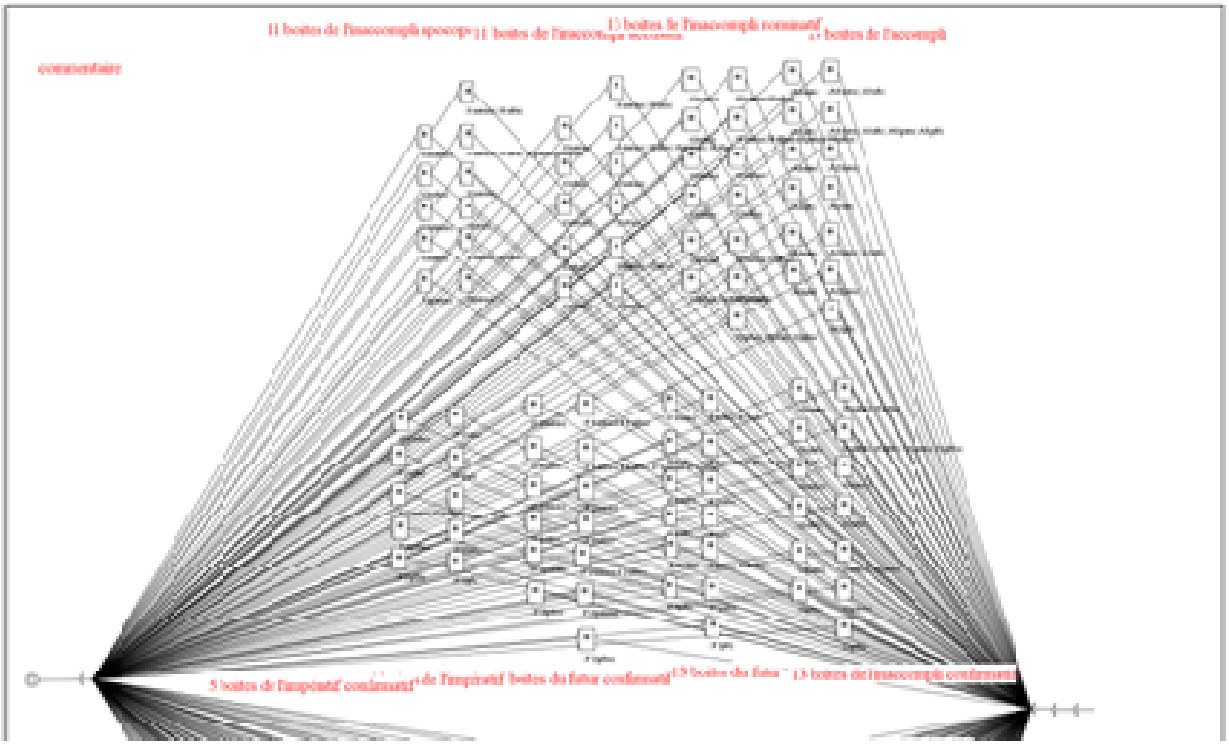


Figure 18 : Graphe modèle

Le graphe modèle est une structure vide qui peut être transformée en un transducteur de flexion. L'opération de transformation consiste à remplacer les *entrées* de ses boites[39] par des chaines adéquates et laisser les *sorties* telles qu'elles sont. La structure vide contient 184 boites contenant en *entrée* (ce qui se trouve à l'intérieur de la boite) un astérisque * et en *sortie* (ce qui se trouve au dessous de la boite) des codes flexionnels selon le cas. En utilisant une classe java, on génère le graphe de flexion d'une catégorie de verbes. La Figure 19 montre un exemple du résultat de la génération automatique d'un transducteur de flexion à partir du graphe modèle de la Figure 18. Le transducteur obtenu est transducteur de flexion d'un seul paradigme flexionnel des verbes arabes représenté par le verbe كـتـب\kataba\ *écrire*

[39] Dans la terminologie d'Unitex/GramLab, la boite d'une graphe est équivalente à une séquence de transitions du transducteur. Elle contient la séquence de caractères d'entrée qui doit être reconnue. Sous la boite il y a une autre séquence de caractères préfixée par ':' et qui doit être produite en cas où la reconnaisse est réussite. Pour de plus ample explications, on réfère le lecteur à (Paumier 2014).

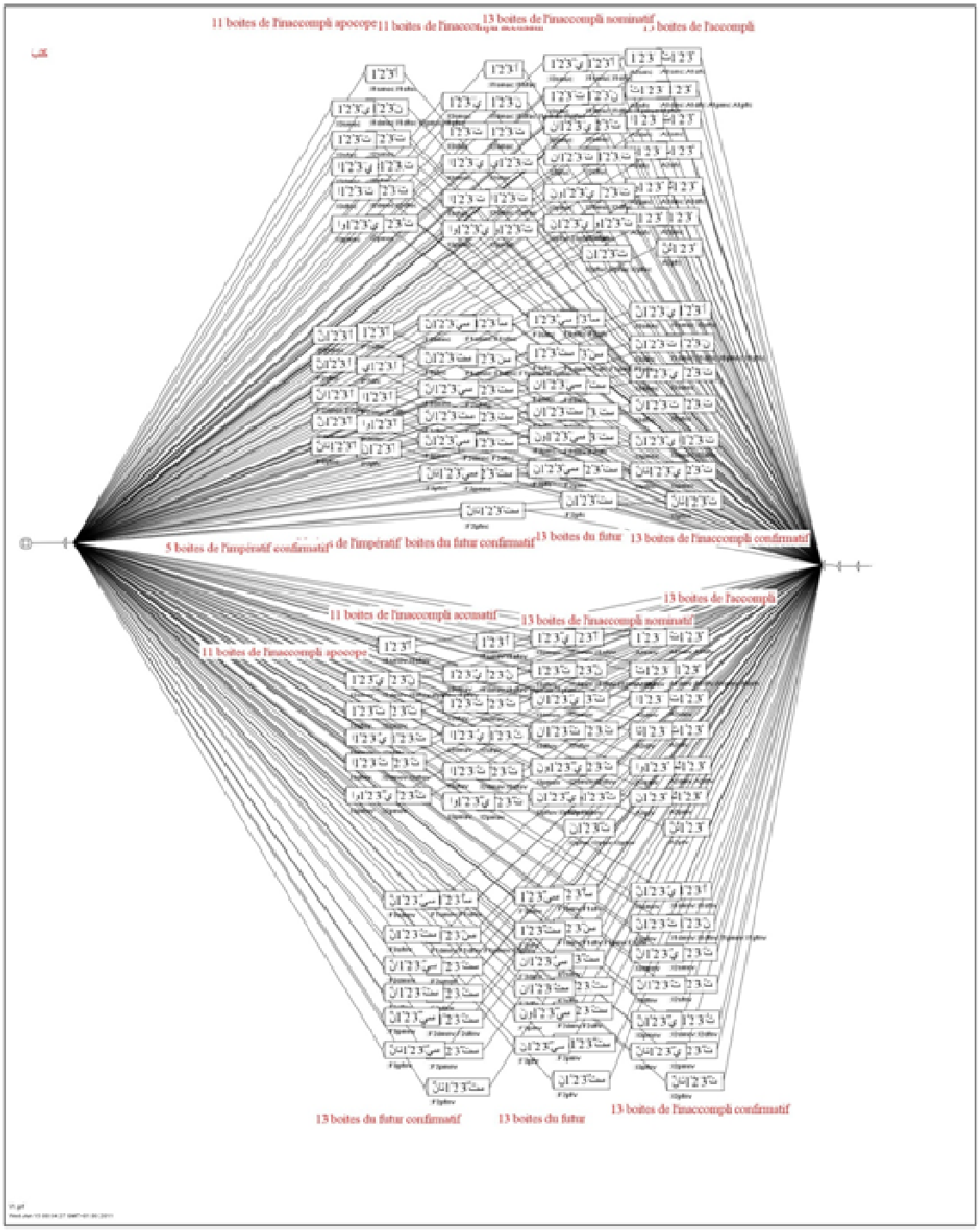


Figure 19 : Un des graphes engendrés à partir du graphe de la Figure 18

Pour expliquer le contenu du graphe de la Figure 19 on préfère extraire une seule boite qui représente un seul chemin parmi 184 chemins possibles de l'automate de flexion. La boite est représentée par la Figure 20. Par exemple la première boite (la plus en haut à droite) contient 1ثُ23ّ qui est calculée par la classe et qui a remplacé * dans le graphe modèle. La sortie de la boite reste telle qu'elle /:A1smc:A1sfc pour dire (l'accompli de l'actif de la 1ère personne du singulier masculin et féminin)

Le graphe calculé ne représente pas le paradigme d'un seul verbe mais un ensemble de verbe, pour détecter tous les verbes ayant le même paradigme flexionnel nous avons conçu un petit algorithme qui se base sur la manière dont le verbe est formé.

Figure 20 : La boite de la partie haute la plus à gauche du graphe de la Figure 19, le graphe contient 184 boites.

Notre système de génération n'est pas actuellement disponible sous Unitex/GramLab, mais le sera dès que possible.

4.2.6 Les noms/adjectifs

En arabe les noms communs et les adjectifs sont traités comme une seule classe : les nominaux. Dans notre travail on a distingué entre les deux (cf. le jeu d'étiquettes de la Figure 12). La flexion des nominaux est plus complexe que les verbes (N. Habash 2010) ; cette complexité à notre avis revient au grand nombre de paradigmes flexionnels des nominaux. Les noms et les adjectifs se fléchissent selon leurs classes flexionnelles. e.g. un déverbal est assez régulier par rapport à un nom primitif.

La flexion des nominaux est en fonction

- du genre (2 valeurs : masculin et féminin),
- du nombre (3 valeurs : masculin, duel et pluriel),
- du cas (3 valeurs : nominatif, accusatif et génitif),
- de la nounation (3 valeurs : nounation de *fetha*, nounation de *dhamma* et nounation de *kasra*),
- de la définitude (2 valeurs : défini et indéfini)
- et de la construction (2 valeurs : en construction et pas de construction).

Ces traits morphologiques ne se combinent pas toujours entre eux e.g. la nounation ne se combine pas avec le duel, le pluriel régulier masculin ou le défini. Ainsi le nombre de formes fléchies d'un nominal ne dépasse pas 63 formes, comme cela est montré sur la Figure 12 les traits morphologiques sont codés dans le fichier *morphology*, par exemple la ligne 8 de la Figure 12 indique que la forme déclinée قـرارين\qaraArayn*deux décisions* est morphologiquement ambiguë, elle peut être :

1. smia : un nom singulier, masculin, indéfini, accusatif
2. ou smii : un nom singulier, masculin, indéfini, génitif.

La Figure 22 représente l'algorithme général de flexion des noms/adjectifs, on remarque qu'il est similaire à celui des verbes avec quelques différences

4.2.7 L'algorithme de flexion des nominaux

A travers une interface, un linguiste parcours un corpus et pour chaque mot s'il est nom/adjectif, il introduit son lemme, on calcule son schème en utilisant le même algorithme que les verbes. Ce schème est recherché dans la liste des schèmes des noms qui existent, s'il

est trouvé on l'ajoute au DELAS. Sinon le linguiste détermine sa classe et introduit son pluriel et son féminin en cas d'irrégularité comme il est expliqué à la section 5.4 de la page 25. En utilisant ces entrées et en consultant des graphes modèles on calcul automatiquement le graphe de flexion du nom/adjectif courant. Le graphe sera appliqué sur le lemme pour générer les 63 formes fléchies nominales.

قـرار,قَـرَارُ.Nc :smiu
قـرار,قَـرَارَ. Nc :smia
قـرار,قَـرَارِ.Nc :smii
قـرار,قَـرَارٌ.Nc :smiU
قـرار,قَـرَاراً.Nc :smiA
قـرار,قَـرَارٍ.Nc :smiI
قـرار,قَـرَارَانِ.Nc :dmiu
قـرار,قَـرَارَيـنِ.Nc :dmia :dmii

Figure 21 : Exemple de flexion des noms communs

5. Le traitement de la cliticisation

Les entrées dans notre dictionnaire de formes fléchies sont stockées sans clitiques alors que les unités lexicales d'un texte traité se présentent sous forme de cliticisation complexe. Comme indiqué au paragraphe 2.2.4 un mot renferme plusieurs niveaux de clitiques qui sont en fonction de la catégorie grammaticale de la base.

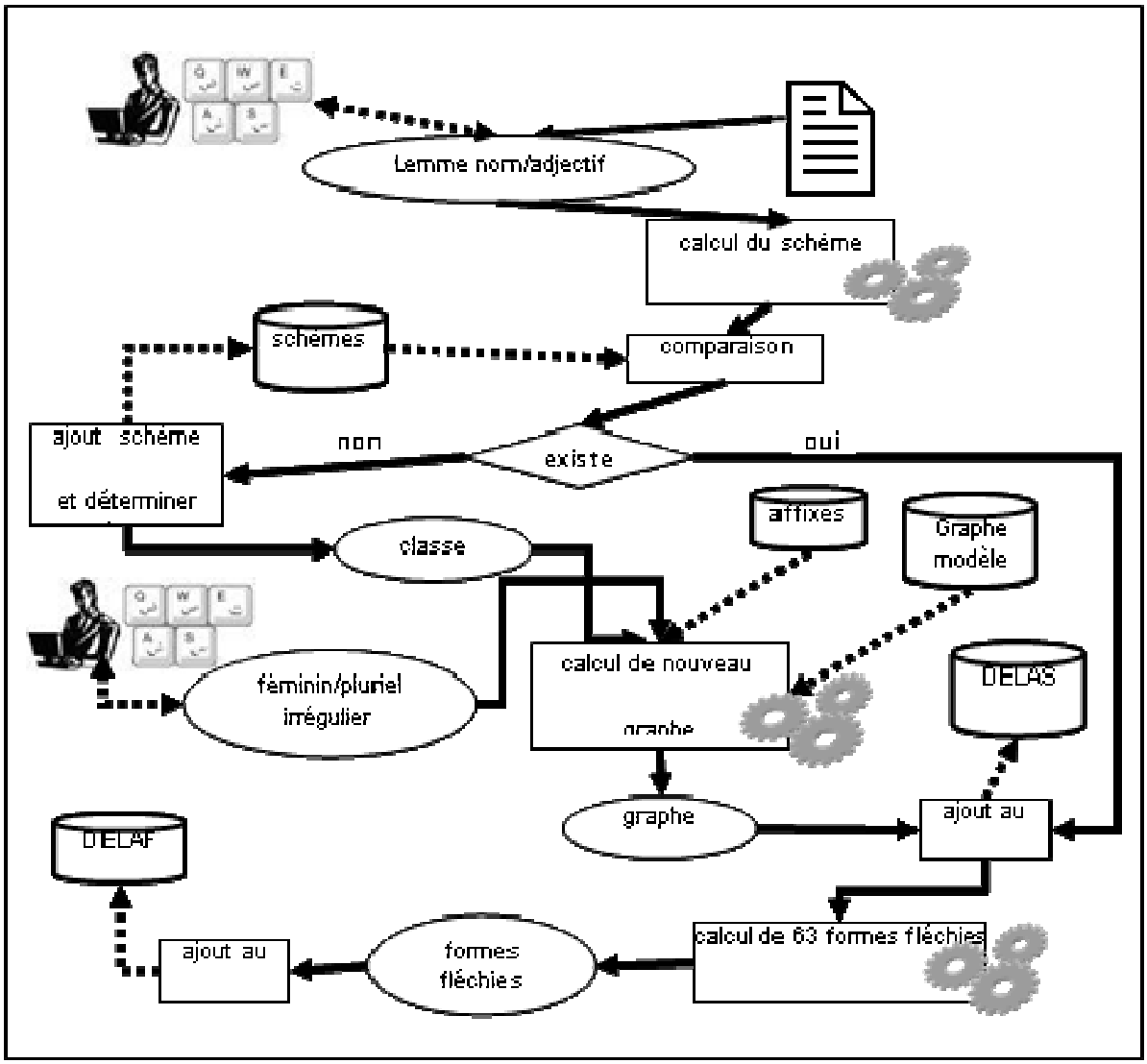


Figure 22 : Algorithme de flexion des nominaux

Dans notre cas la cliticisation est traité sous forme d'automate de reconnaissance. La Figure 23 présente un graphe de reconnaissance des unités lexicales contenant des clitiques qui peuvent s'attacher à un verbe.

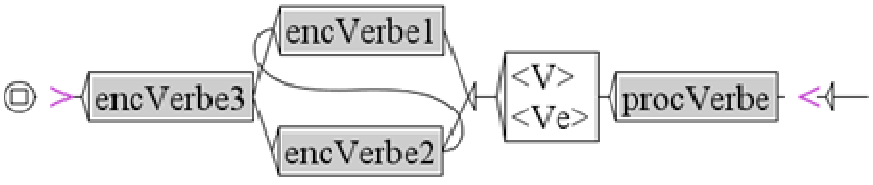


Figure 23 : Graphe de reconnaissance des unités lexicales contenant des clitiques attachées à un verbe.

6. Résultats d'expérimentation

L'expérimentation de notre système de génération automatique des graphes a donné les résultats cités dans le Tableau 20. Dans ce tableau les 400 premiers verbes du corpus de test sont regroupés par groupes de 50 verbes, on remarque dans la colonne 4 (nouveaux graphes) que le nombre de nouveaux graphes décroit avec l'accroissement des verbes traités (à l'exception de la ligne 8) et on remarque dans la colonne 5 (taux de graphes/verbe) que le taux décroit presque de la moitié. En augmentant davantage le nombre de verbes traités on arrivera à un point où le nombre de nouveaux graphes tend vers zéro, on couvrira donc tous les graphes possibles. Si on atteint ce stade, le traitement d'un nouveau verbe deviendra plus facile et consistera seulement à lui attribuer son graphe déjà établi.

Groupes de 50 verbes	Total verbes	Total graphes	Nouveaux graphes	Taux graphes/verbe	Taux d'accroissement des graphes	Total d'échec	Taux d'échec
1	50	21	21	0.42	100%	0	0%
2	100	33	12	0.33	36.36%	0	0%
3	150	46	13	0.31	28.26%	0	0%
4	200	53	7	0.27	13.21%	1	0.50%
5	250	63	10	0.32	15.87%	1	0.40%
6	300	70	7	0.23	10%	1	0.33%
7	350	72	2	0.21	2.78%	1	0.29%
8	400	91	9	0.23	9.89%	8	2%

Tableau 20 : Résultats de la génération automatique des graphes

Ce qui a été dit pour les verbes est applicable sur les noms et adjectifs et avec un nombre de formes fléchies nettement moindre (au plus 63 au lieu de 264 dans le cas des verbes).

7. Conclusion

D'après l'expérimentation qui a touché 16853 verbes/verbes d'état, 1780 noms/adjectifs, 181 pronoms et 164 particules en forme vocalisée et non vocalisée; on conclut que l'arabe se prête facilement à exprimer sa morphologie complexe en technologie à nombre fini d'état à travers la plateforme Unitex/GramLab. Un premier package arabe est disponible avec la version 3 d'Unitex/GramLab. Le Tableau 21 donne le contenu actuel[40] des dictionnaires.

Catégorie	Nombre des entrées DELAS	Nombre des entrées DELAF vocalisées	Nombre des entrées DELAF non vocalisées	Couverture
Verbes d'état	13	768	902	100%
Verbes	16842	4446288	6632397	+70%
Noms et adjectifs	1780	86553	101757	inconnue
Adverbes		86	95	+90%
Pronoms		181	198	+90%
Particules		164	205	+90%
Noms de personnes			8353	inconnue
Noms de pays			802	100%
Noms de villes			7977	inconnue

[40] Le contenu du mois fevrier 2017

Total	18635	4534040	6752686

Tableau 21 : Contenu des dictionnaires disponibles sous Unitex/GramLab

Les dictionnaires et les graphes conçus seront étendus par des dictionnaires de mots polylexicaux et des dictionnaires de noms propres afin d'utiliser Unitex/GramLab pour des applications de haut niveau telles que la reconnaissance des entités nommées dans les textes arabes.

Chapitre V

Mise en œuvre d'extraction de relation entre entités nommées arabes

1. Introduction

Dans ce chapitre on entame l'approche et les méthodes utilisées pour mettre en œuvre la détection et l'extraction des relations entre entités nommées arabes. Cela veut dire parcourir le texte brut qui représente le corpus à traiter et appliquer toutes les règles pour :

1. prétraiter le texte,
2. détecter et catégoriser les entités nommées,
3. repérer et catégoriser les relations sémantiques qui les relient et
4. donner du sens à la relation détectée.

Ce travail a requis la conception et la mise en œuvre de plusieurs programmes, ressources linguistiques, grammaires locales et transducteurs. La Figure 24 de la page suivante montre les différents niveaux nécessaires à la réalisation de l'approche.

Comme il est montré dans l'approche, le traitement se fait en pipeline comme c'est le cas pour toutes les applications TAL. A chaque niveau ou étape dans un niveau on utilise des programmes et des ressources linguistiques qui sont majoritairement développés dans le cadre de cette thèse. Les étapes indiquées dans la figure de l'approche en flèche pointillée (comme l'étape 2 et 5) actuellement sont partiellement réalisées et nécessitent dans le futur une amélioration dans les idées et dans la réalisation. L'étape numéro 10 correspond actuellement à un traitement manuel validé par un expert linguiste. Les différentes étapes seront expliquées au long des sections qui vont suivre.

2. Le prétraitement de corpus

Le corpus utilisé dans notre travail de détection des entités nommées et des relations est connu sous le nom *d'araCorpus* et qui peut être trouvé dans le web[41]. Ce corpus est un ensemble d'articles journalistiques du monde arabe, compilé au CRL (Computer Research Laboratory) à l'université de New Mexico (Abdelali et al. 2005). Ce corpus dans sa totalité contient plus de 113 millions de mots mais dans notre travail on s'est confiné seulement à une partie contenant 5 millions de mots.

[41] Le corpus est disponible gratuitement au lien http://aracorpus.e3rab.com/argistestsrv.nmsu.edu/AraCorpus.tar.gz (dernier accès en septembre 2016).

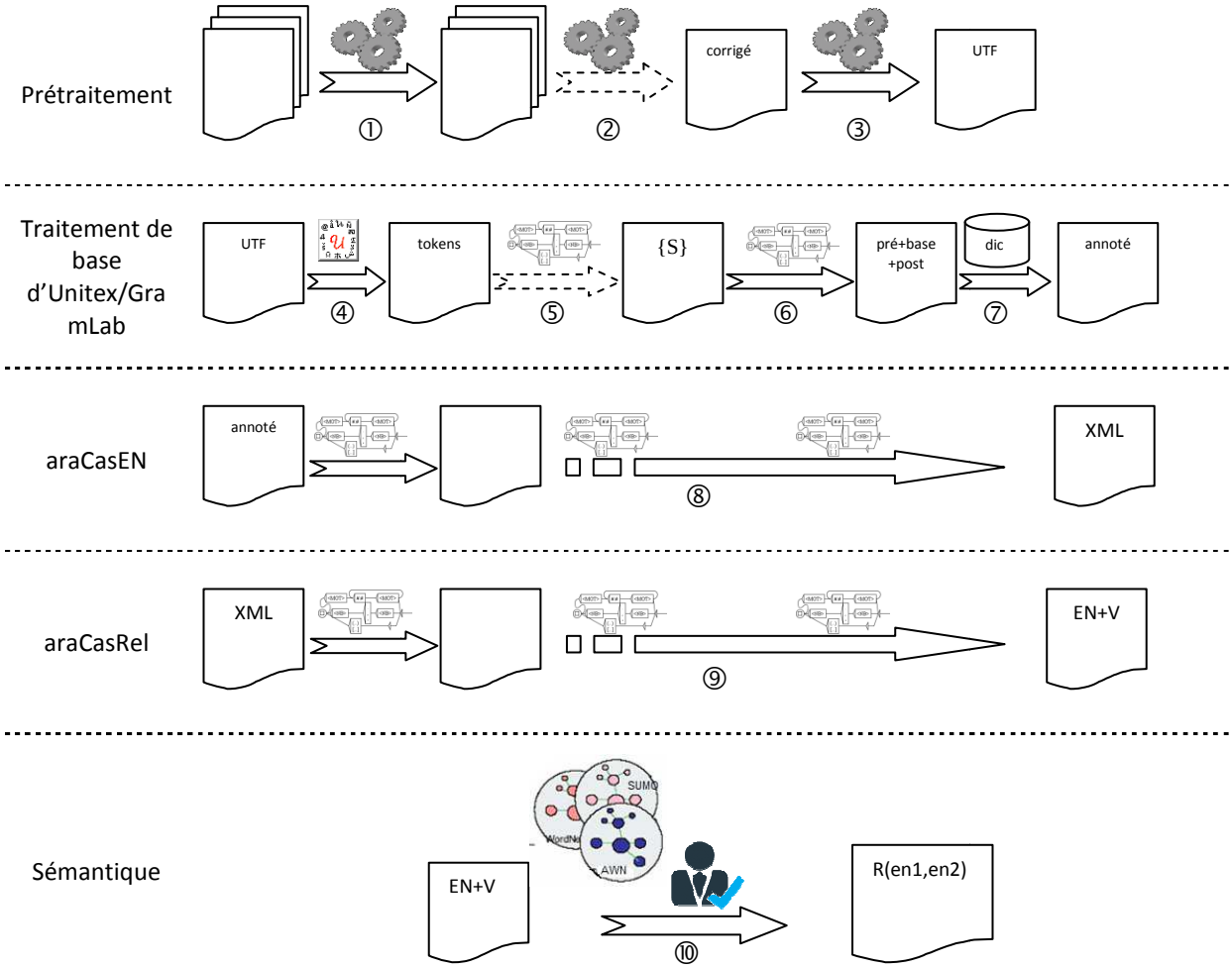


Figure 24 : Schéma général des différents niveaux de l'approche proposée

Le traitement de corpus est précédé par un prétraitement dont l'objectif est de normaliser quelques formes qui peuvent y exister et corriger d'autres. Ce prétraitement est réalisé en utilisant des programmes hors la plateforme Unitex/GramLab, ce sont des programmes des classes de la Figure 54 (notamment les méthodes de la classe *ArabicWord*). Ce prétraitement correspond aux étapes 1 et 2 du premier niveau de la Figure 24 et consiste à :

- Eliminer les parties de textes en double car on a remarqué qu'il existe dans le corpus des paragraphes qui ont été répétés,
- Supprimer le caractère d'allongement (kashida ou tatouil) i.e. le caractère (\u0640), dans la rédaction ce caractère est utilisé pour distendre un mot pour qu'il convienne à une mise-en-forme bien donnée e.g. le mot كـتب \kataba*écrire* peut être écrit sous différentes formes, comme suit : كـتـــــــــــــــــــــــب, كـــــــــــــــــــتب de longueur 20 et 15 respectivement au lieu de 3. Les trois formes sont lexiquement équivalentes donc la suppression ou le maintien du caractère kashida sont équivalents. L'objectif de l'élimination du caractère kashida est de permettre un matching entre les mots du texte et ceux du dictionnaire,
- Supprimer les signes diacritiques à l'exception du signe de gémination (shadda). Du fait que les diacritiques sont ajoutés à un mot pour résoudre une éventuelle ambiguïté ; la suppression de ces caractères apporte un degré d'ambiguïté

suplémentaire au texte. Mais pour une simplicité de traitement l'utilisateur peut choisir ou non un prétraitement qui supprime les diacritiques (prétraitement facultatif). On note que le maintient du signe de gémination (la shadda) est dû à sa fonction qui est le redoublement du caractère sur lequel il apparait. Par exemple le verbe مـدّ\mad~a*tendre* est de longueur 2 (sans diacritiques) mais il est considéré comme un verbe trilitère à cause de redoublement du caractère د : le dernier caractère avec gémination,

- La correction des lettres non arabes (par exemple perse) en leurs équivalents arabes,
- La conversion des signes de ponctuation latins en leurs équivalents arabes. Par abus, le rédacteur arabe a souvent tendance d'utiliser des signes de ponctuation latins au lieu des signes de ponctuation arabes. e.g. souvent le virgule ou le point virgule latins (, ;) sont écrits au lieu du virgule et point virgule arabes (، ؛).

Le traitement de caractère de kashida a été réalisé en premier lieu sous forme d'un programme java hors la plateforme Unitex/GramLab puis on a proposé une deuxième alternative de le faire sous forme de transducteur principal appelant des sous-graphes. La Figure 25 montre un des cinq sous-graphes qui traitent le caractère de kashida. Dans ce traitement on a la possibilité de revenir à tout moment à la forme d'origine (avec kashida) i.e. faire l'extraction des EN et leurs relations et puis revenir au texte d'origine.

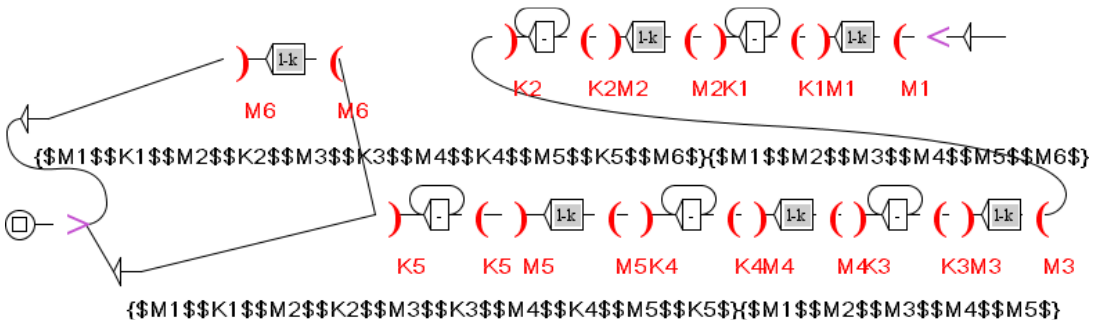


Figure 25 : Traitement de kashida par un transducteur sauvegardant la forme d'origine pour en revenir

Le transducteur de la Figure 25 suppose qu'il existe un caractère de kashida (ou une séquence) dans cinq positions différentes au maximum dans le token à traiter. Dans le graphe principal on appelle quatre d'autres transducteurs, chacun traite le cas où le caractère de kashida se trouve dans une, deux, trois ou quatre positions dans le token.

3. Le traitement

3.1. La segmentation des clitiques

Le phénomène de l'agglutination dans la morphologie arabe augmente le degré d'ambiguïté lexicale et accroit la complexité de l'analyse d'un mot en ses composants. Comme il a été discuté dans la section 5 en page 21, un mot arabe est grossièrement composé de pré-base, base et post-base. Ces derniers composants eux même composés des affixes et des clitiques arabes qui peuvent s'agglutiner en plusieurs niveaux. La Figure 26 montre les composants d'un mot arabe et la Figure 27 donne un exemple de l'analyse d'un mot arabe en affixes et en clitiques.

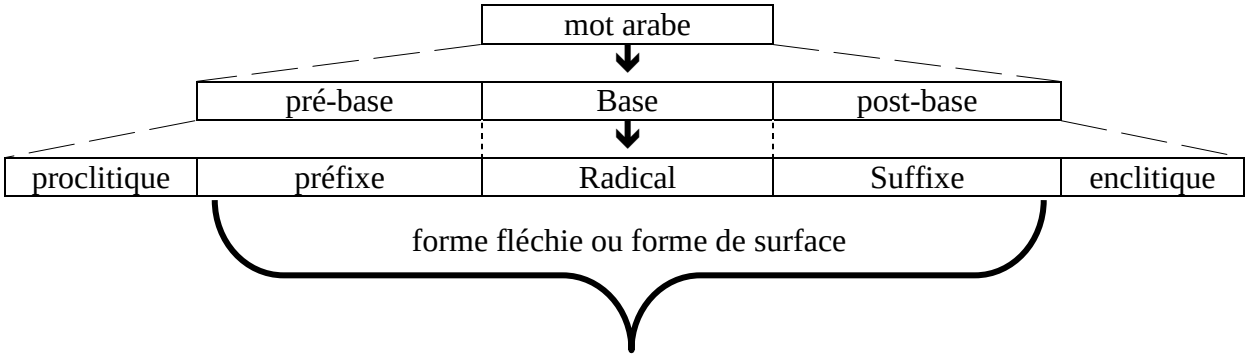


Figure 26 : Morphologie d'un mot arabe

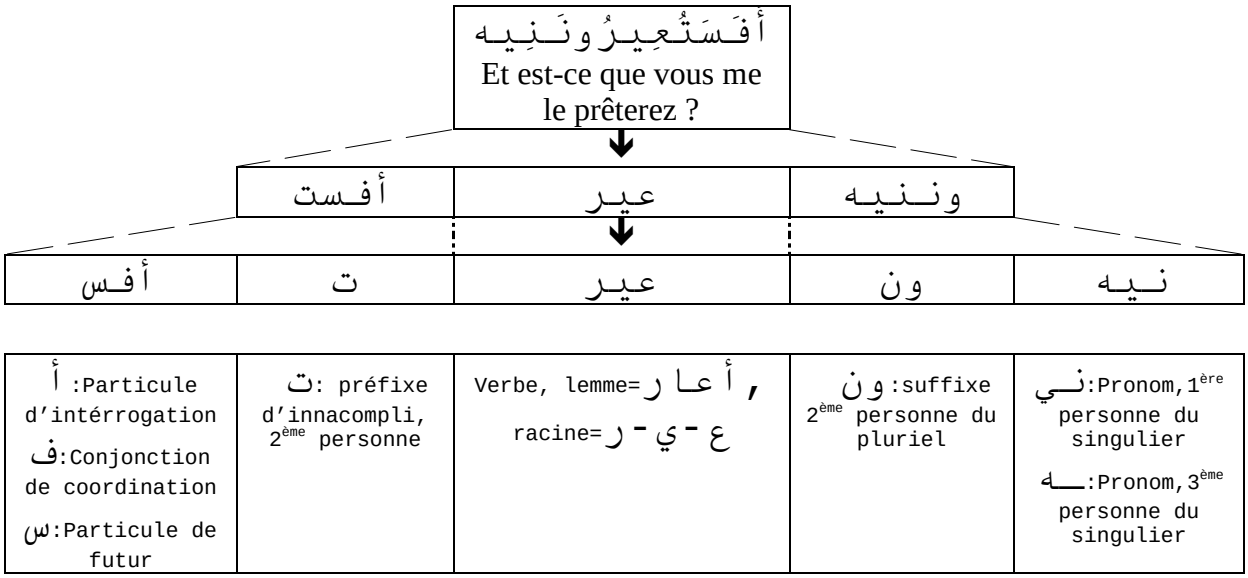


Figure 27 : Exemple d'agglutination/analyse des clitiques et des affixes dans un mot arabe

Pour réaliser la segmentation des clitiques dans notre travail, on s'est servi des dictionnaires DELA qu'on a construit (cf. section 4 de la page 82). Une entrée de ces dictionnaires contient les affixes et le radical comme il est montré dans la Figure 26, donc tout ce qui reste à segmenter c'est bien les clitiques.

Le traitement de la segmentation a été réalisé en trois procédures, chacune correspond à un transducteur dans une cascade, comme suit :

- Le cas de la présence seulement des proclitiques (variable Pref dans la figure) dans le token (cf. la Figure 28)
- Le cas de la présence seulement des enclitiques (variable suff dans la figure) dans le token (cf. la Figure 29)
- Le cas de la présence à la fois, des proclitiques et des enclitiques dans le token (cf. la Figure 30)

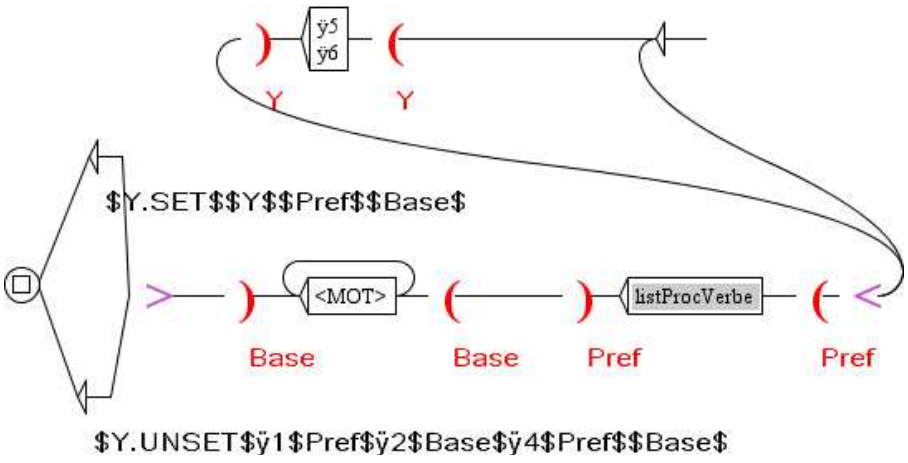


Figure 28 : Segmentation des proclitiques des verbes

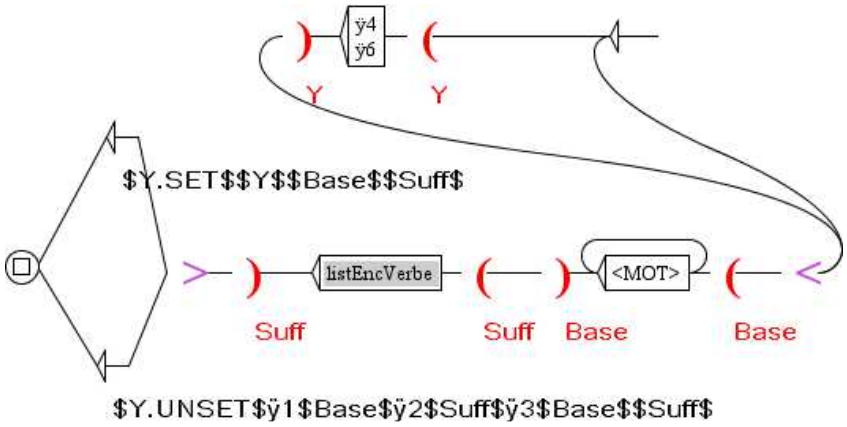


Figure 29 : Segmentation des enclitiques des verbes

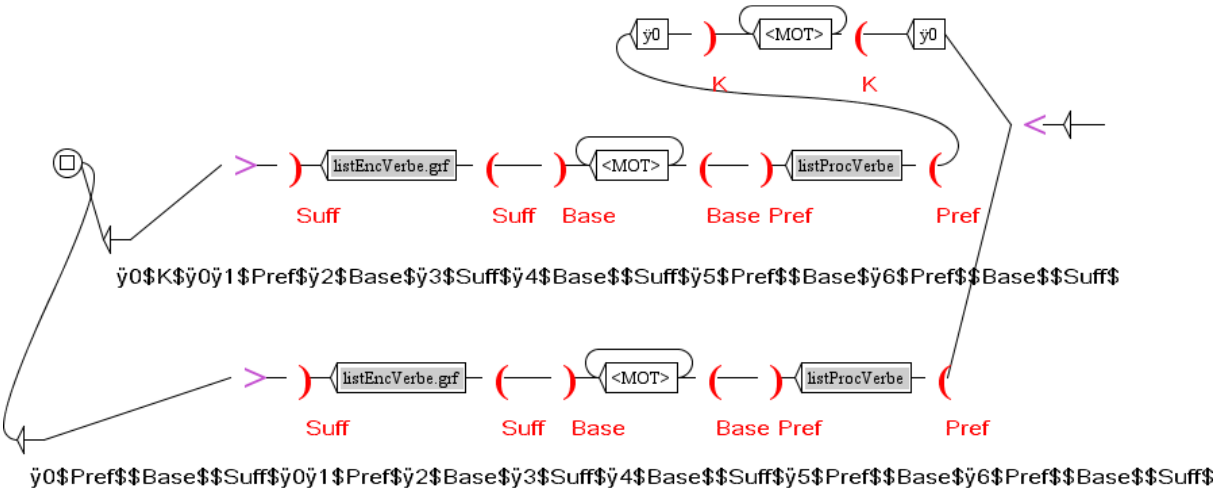


Figure 30 : Segmentation des clitiques des verbes

3.2. La segmentation en phrases

La segmentation d'un texte arabe en phrases révèle une grande importance dans le TAL arabe, car tout traitement qui repose sur la structure syntaxique nécessite une segmentation du texte en petite structure syntaxique que nous l'appelons ici phrase. Le symbole qui dénote la fin d'une phrase dans la plateforme Unitex/GramLab est {S} pour dire *sentence* (phrase en anglais). Dans notre tokenizer de phrase de la Figure 31 on se base sur les signes de ponctuation pour effectuer la segmentation.

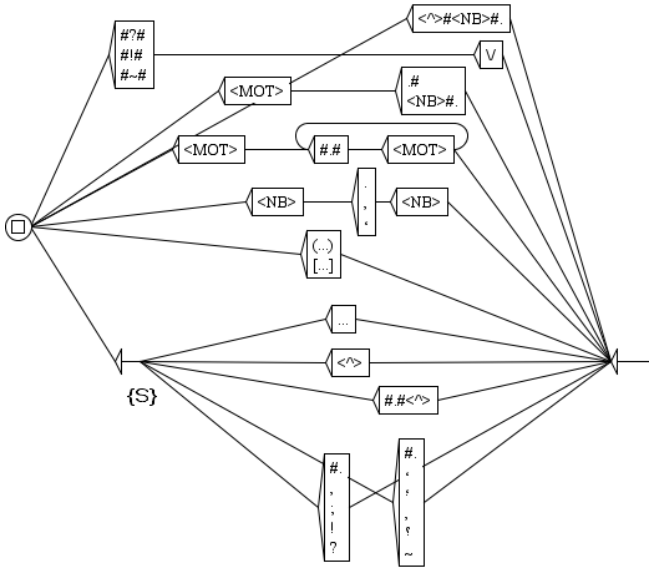


Figure 31 : Transducteur de segmentation de corpus en phrases

4. La détection des entités nommées arabes

Pour détecter les entités nommées arabes on applique une cascade de transducteurs. Chaque élément dans cette cascade (transducteur) apporte un changement sur le texte brut. Ce changement servira par la suite aux autres transducteurs qui viennent après celui-ci. Pendant l'annotation on utilise la typologie Quaero qui respecte la norme TEI.

4.1. TEI

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

4.2. L'annotation des entités nommées arabes

4.2.1 Les noms de personnes

Les noms de personnes peuvent se trouver dans un texte arabe sous différentes formes. Les noms de personnes dans un texte de l'arabe classique se caractérisent par l'absence de la structure moderne (*first name*, *middle name* et *last name*) alors que les noms de personnes de l'arabe standard moderne varient entre les régions du monde arabe : les noms de la région du

Maghreb diffèrent des noms de l'Egypte et ceux-ci sont différents des pays de Golf. Les noms de la famille royale des pays de Golf diffèrent des noms de personnes qui n'appartiennent pas à la famille royale. La Figure 32 montre les différentes catégories de noms de personnes qui peuvent exister dans un texte arabe. On s'intéresse dans notre travail à la détection des entités nommées de type noms de personnes dans les textes de l'ASM.

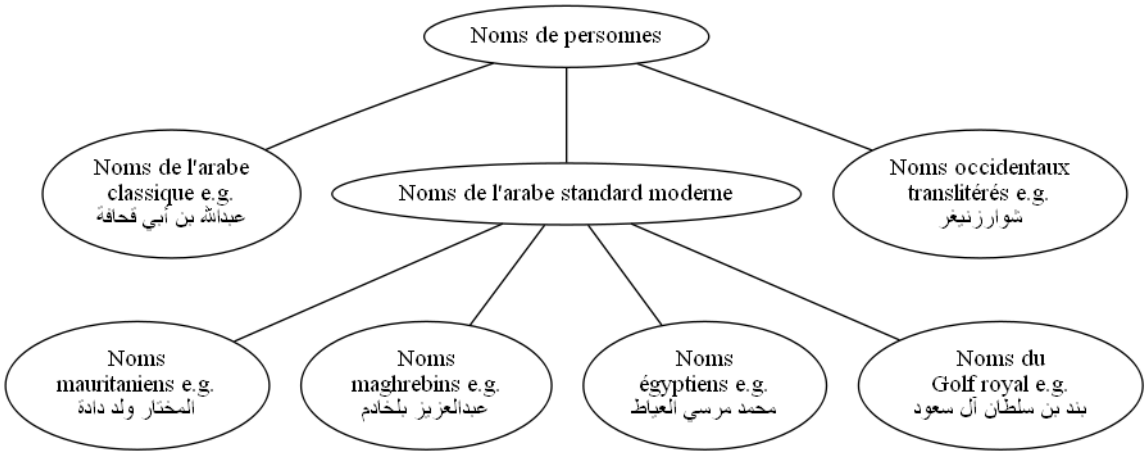


Figure 32 : Taxonomie des noms de personnes dans un texte arabe

Les noms de personnes occidentaux dans un texte journalistique arabe se caractérisent par le problème de translitération (N. Habash 2010). La translitération du même nom peut être se produite sous différentes variantes et cela dépend de la personne (le journaliste ou le rédacteur en général) qu'elle soit anglophone ou francophone. Dans la littérature, le problème de *Schwartzenegger* réfère au cas où une seule écriture en anglais correspond à plusieurs translitérations en arabe. Ici une seule écriture du gouverneur de la Californie peut apparaitre en arabe en شوارزنيغر\šuwAr.ziniγar\, شوارزنغر\šuwAr.ziniγar\, شوارزنيجر\šuwAr.ziniyjar\ et شوارتزنيجر\šuwAr.tziniyjar\ parmi plusieurs d'autres. Une variante du même problème est le cas de *Mozart* où différentes écritures préservant des particularités de prononciations apparaissent : موزارت\muwzaAr.t\ (anglophonique) et موزار\muwzaAr\ (francophonique) (N. Habash 2010).

A l'instar des autres langues naturelles un nom de personne dans un texte arabe est caractérisé par une preuve externe et une preuve interne. L'utilisation de ces preuves facilitera l'opération de la détection.

e.g. pour annoter un texte contenant la phrase suivante

تحت الرعاية السامية لفخامة الرئيس عبدالعزيز بوتفليقة...\tHt AlrçAyħ AlsAmyħ lfxAmħ Alrŷys ςbdAlçzyz bwtflyqħ…*sous le haut patronage de son excellence le président Abdelaziz Bouteflika*

Le résultat de l'annotation doit être comme suit :

```
<rawText>تحت الرعاية السامية ل</rawText>
        <pers.ind>
                <title>فخامة
                        <func.ind>الرئيس</func.ind>
                </title>
                <name.first>عبد العزيز</name.first>
                <name.last>بوتفليقة</name.last>
```

```
        </pers.ind>
<rawText>...</rawText>
```

La granularité de l'annotation est en fonction de son objectif final. On peut dans l'exemple précédent réduire l'annotation comme suit :

```
<rawText>تحت الرعاية السامية ل</rawText>
        <pers.ind>
                <title>فخامة الرئيس</title>
                <name>عبد العزيز بوتفليقة</name>
        </pers.ind>
<rawText>...</rawText>
```

Pour la partie *titre* de la preuve interne des noms de personnes on a conçu un transducteur qui balise cette partie par l'étiquette *title*. Le transducteur est montré dans la Figure 33, ce petit transducteur peut générer tous les chemins possibles des titres de noblesse utilisé dans les articles journalistiques. La liste du Tableau 22 montre une partie des expressions qui peuvent être reconnues par le transducteur de la Figure 33. Ce transducteur appelle des sous transducteurs (en gris dans la figure) et qui peuvent eux même appeler d'autres transducteurs et ainsi de suite. Le transducteur de la Figure 33 peut reconnaitre en tout, 21840 expressions différentes (ce graphe orienté contient 21840 chemins possibles y compris ceux des sous transducteurs).

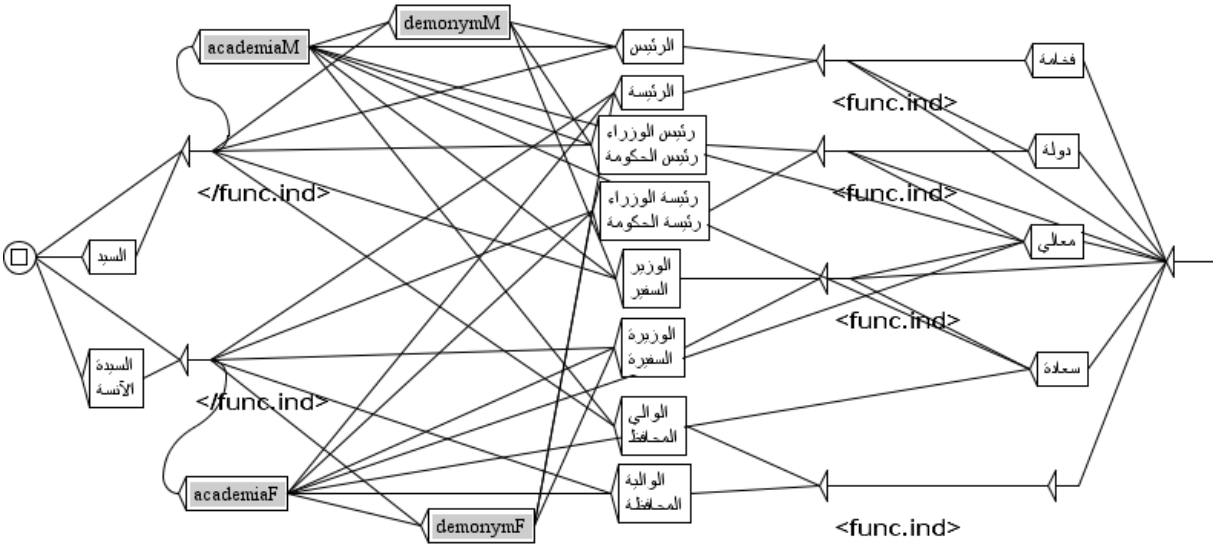


Figure 33 : Transducteur de reconnaissance de la partie <title> de nom de personne de la politique

فخامة الرئيس
فخامة الرئيس السيد
فخامة الرئيسة
فخامة الرئيسة السيدة
فخامة الرئيسة الآنسة
دولة الرئيس
دولة الرئيس السيد
دولة رئيس الوزراء
دولة رئيس الوزراء السيد
دولة رئيس الوزراء الدكتور
دولة رئيس الوزراء الدكتور السيد
دولة رئيس الوزراء المهندس
دولة رئيس الوزراء المهندس السيد
دولة رئيس الحكومة
دولة رئيس الحكومة السيد
دولة رئيس الحكومة الدكتور
دولة رئيس الحكومة الدكتور السيد
دولة رئيس الحكومة المهندس
دولة رئيسة الوزراء الدكتورة
دولة رئيسة الوزراء الدكتورة الآنسة
سعادة الوزير
سعادة الوزير السيد
سعادة الوزير الدكتور
سعادة الوزير الدكتور السيد
سعادة الوزير المهندس السيد
سعادة الوزير البروفيسور
سعادة الوزير البروفيسور السيد
سعادة السفير
سعادة السفير السيد
سعادة السفير الدكتور
سعادة السفير الدكتور السيد
سعادة السفير البروفيسور السيد
سعادة الوزيرة
سعادة الوزيرة السيدة
سعادة السفيرة المهندسة
سعادة السفيرة المهندسة السيدة
سعادة السفيرة المهندسة الآنسة
معالي رئيس الوزراء
معالي رئيس الوزراء السيد
معالي رئيس الوزراء الدكتور
معالي الدكتورة الآنسة
معالي الأستاذة الدكتورة
معالي الأستاذة الدكتورة السيدة
المحافظ المهندس
المحافظ المهندس السيد
المحافظة المهندسة الآنسة

Tableau 22 : Un exemple de la liste des expressions <title> qui peuvent être reconnues par le transducteur de la Figure 33

Les expressions citées dans le Tableau 22 montrent la puissance des transducteurs dans la reconnaissance ; un simple graphe (transducteur) peut couvrir une large variante de structures syntaxiques au nombre de 21840 structures. De cette façon, on peut couvrir toutes les preuves externes d'une entité nommée de type *nom de personne*. La preuve externe sert comme un déclencheur de reconnaissance pour le transducteur i.e. si le celui-ci rencontre une telle

structure linguistique il peut conclure que la suite du texte est une entité nommée de type *nom de personne*. Le processus de reconnaissance nécessite aussi les preuves internes. Une preuve interne dans le cas des noms de personnes peut consister en une entrée des dictionnaires des prénoms de la plateforme Unitex/GramLab.

4.2.2 Le dictionnaire des prénoms

Les prénoms arabes correspondent au *first name* des noms occidentaux ; sauf qu'ils ne sont pas caractérisés par une majuscule comme il est le cas des langues indo-européennes. Cette absence de caractère distinctif rend la reconnaissance de ce sous-type d'entité nommée davantage difficile. Une solution consiste à collecter une liste de prénoms arabes sous forme d'un dictionnaire accessible par les programmes de la plate forme Unitex/GramLab.

Pour notre travail dans cette thèse on a opté à la collection des prénoms arabes à partir des sites web proposant des prénoms pour les nouveau-nés dans le monde arabe. Le Tableau 23 donne une idée sur le nombre des entrées de ce dictionnaire. Cette opération nous a permis de compiler une liste de prénoms des différentes régions du monde arabe (Maghreb, Egypte, Golf, Levantin et Yémen). Le Tableau 24 montre un extrait de la liste des prénoms arabes compilée sous forme d'un dictionnaire LADL sous la plateforme Unitex/GramLab.

Puis la liste est traitée par des programmes pour calculer les différentes variantes d'un prénom i.e. calculer les différentes formes possible rencontrées dans le texte journalistique brut. A titre d'exemple la huitième ligne du Tableau 24 représente une entrée qui compte quatre variantes (les lignes 8, 9, 10 et 11 du même tableau).

Genre	Nombre
Variantes des prénoms pour nouveau-nés de genre féminin	4392
Variantes des prénoms pour nouveau-nés de genre masculin	3963
Total	8355

Tableau 23 : Contenu du dictionnaire <prenom.dic>

Les variantes sont calculées en utilisant une classe java (cf. le diagramme de classes de l'annexe A, exactement le package dz.utms.cs.anlp.LADL et quelques méthodes de la classe ArabicWord du package dz.utms.cs.anlp.morphology) ; l'idée des variantes repose sur les erreurs orthographiques répandues dans les textes journalistiques.

Par exemple dans la tradition typographique égyptienne la voyelle longue ي\u064A est remplacée par ى\u0649. Le Tableau 25 montre quelques exemples de ces erreurs. Le tableau n'est pas exhaustif et ne couvre pas toutes les erreurs ; car les erreurs typographiques des corpus journalistiques arabes nécessitent tout un module pour la correction orthographique et qui doit passer le premier avant tout traitement automatique de la langue arabe.

```
عبد الوارث,عبد الوارث.Np+Hum:ms
عبد الواسع,عبد الواسع.Np+Hum:ms
عبد الواسع,عبد الواسع.Np+Hum:ms
عبد الوتر,عبد الوتر.Np+Hum:ms
عبد الوتر,عبد الوتر.Np+Hum:ms
عبد الودود,عبد الودود.Np+Hum:ms
عبد الوكيل,عبد الوكيل.Np+Hum:ms
عبد الولي,عبد الولي.Np+Hum:ms
```

```
عبد الولى,عبد الولى.Np+Hum:ms
عبد الولي,عبد الولي.Np+Hum:ms
عبد الولى,عبد الولى.Np+Hum:ms
عبد الوهاب,عبد الوهاب.Np+Hum:ms
اثار,اثار.Np+Hum:fs
آداب,آداب.Np+Hum:fs
اداب,اداب.Np+Hum:fs
آذار,آذار.Np+Hum:fs
اذار,اذار.Np+Hum:fs
آذان,آذان.Np+Hum:fs
اذان,اذان.Np+Hum:fs
آسة,آسة.Np+Hum:fs
آسية,آسية.Np+Hum:fs
اسية,اسية.Np+Hum:fs
آسيه,آسيه.Np+Hum:fs
اسيه,اسيه.Np+Hum:fs
آصال,آصال.Np+Hum:fs
اصال,اصال.Np+Hum:fs
آفاق,آفاق.Np+Hum:fs
افاق,افاق.Np+Hum:fs
آكام,آكام.Np+Hum:fs
اكام,اكام.Np+Hum:fs
آلاء,آلاء.Np+Hum:fs
الاء,الاء.Np+Hum:fs
آمال,آمال.Np+Hum:fs
امال,امال.Np+Hum:fs
آمنة,آمنة.Np+Hum:fs
امنة,امنة.Np+Hum:fs
آمنه,آمنه.Np+Hum:fs
امنه,امنه.Np+Hum:fs
آية,آية.Np+Hum:fs
اية,اية.Np+Hum:fs
آيه,آيه.Np+Hum:fs
ايه,ايه.Np+Hum:fs
أبحار,أبحار.Np+Hum:fs
ابحار,ابحار.Np+Hum:fs
أبعاد,أبعاد.Np+Hum:fs
ابعاد,ابعاد.Np+Hum:fs
أبية,أبية.Np+Hum:fs
ابية,ابية.Np+Hum:fs
```

Tableau 24 : Extrait du dictionnaire des prénoms <prenoms.dic> utilisé dans les travaux de reconnaissance des EN de type nom de personne sous la plateforme Unitex/GramLab.

En plus des corrections calculées du Tableau 25, d'autres variantes sont ajoutées, par exemple, les prénoms composés de deux unités lexicales ou plus. L'exemple ci-dessous montre le traitement effectué pour les prénoms polylexicaux.

Exemple

L'entrée عبد العزيز\çbdAlçzyz\ donne deux variantes : la première sans espace entre les deux unités lexicales عبد\çbd\ et العزيز\Alçzyz\ et la deuxième avec espace entre les deux unités عبد العزيز\çbd Alçzyz\.

Origine	أ	إ	آ	ة	ي	ئ	ئ	يء	ؤ	ؤ
Unicode	\u0623	\u0625	\u0622	\u0629	\u064A	\u0626	\u0626		\u0624	\u0624
Erreur	ا	ا	ا	ه	ى	ى	ىء	ئ	وء	و
Unicode	\u0621	\u0621	\u0621	\u0647	\u0649	\u0649		\u0626		\u0648

Tableau 25 : Tableau de correspondance entre les erreurs typographiques des corpus et leurs corrections possibles

4.2.3 Les dates

Dans la catégorisation Quaero, l'entité nommée *temps* est divisée en deux classes : date et heure. Ces deux classes peuvent être absolues ou relatives. La Figure 34 montre les classes et sous classes de la catégorie *temps* dans le projet Quaero dont nous suivons son annotation dans notre travail. On note que ce travail s'intéresse beaucoup plus sur les dates et leur imbrication dans les entités nommées de types *lieu*.

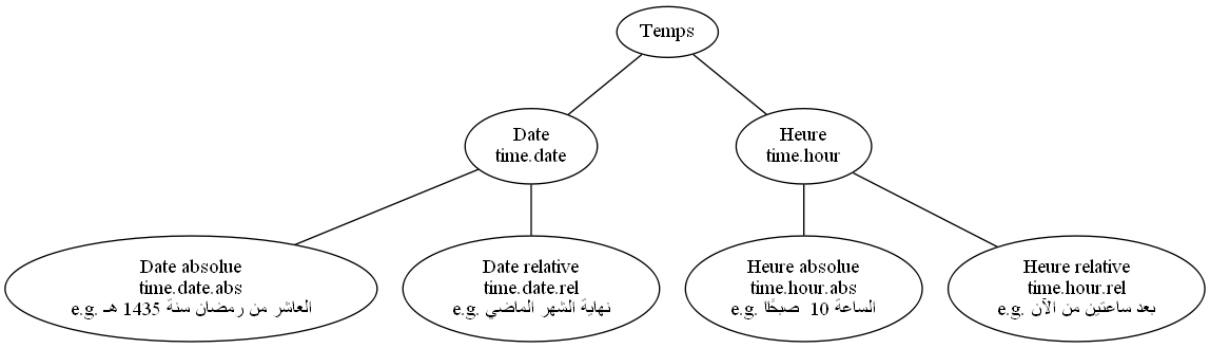


Figure 34 : Les classes et sous classes de la catégorie temps dans le projet Quaero

Les dates dans un texte arabe peuvent apparaitre sous différentes formes en chiffres et en lettres. Les chiffres peuvent se trouver en hindi ou en arabe et les noms des mois peuvent se trouver en hijri, en grégorien ou en syriaque. La Figure 38 montre un automate qui reconnait des séquences de texte représentant des nombres cardinaux et ordinaux arabes en lettres. Tous les mots ou groupes de mots faisant référence aux jours de la semaine, aux mois, aux années ou à des événements calendaires (عيد الاستقلال\çyd AlAstqlAl\, \AlÂDHý AlmbArk\ الأضحى المبارك…) sont annotés si et seulement s'ils réfèrent à une date ou une période unique. La sous-classe *date* regroupe deux types de date : date absolue et date relative.

Date absolue

Une date est considérée absolue lorsque la date du document annoté n'est pas nécessaire pour la déterminer précisément (ESTER2 2007).

Dans notre travail, les dates sont annotées en *date absolue* ou *date relative* selon le cas. Ci-dessous des exemples de date absolue que nos transducteurs peuvent les reconnaitre et annoter.

1. 24-02-1970

2. ب تاريخ التاسع و العشرين من شهر كانون الثاني / يناير من عام ألف و تسعمائة و خمس و تسعين ميلادي\b tAryx AltAsç w Alçšryn mn šhr kAnwn AlθAny / ynAyr mn çAm Âlf w tsçmAŷħ w xms w tsçyn mylAdy\
3. منتصف شهر غشت عام ١٩٩٥م\mntSf šhr γšt çAm ١٩٩٥m\

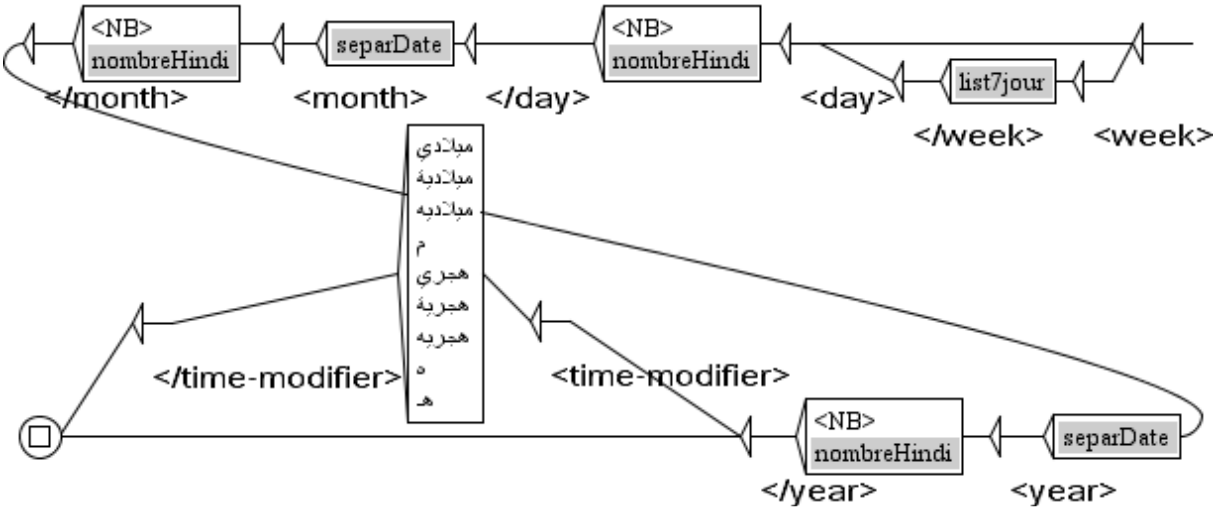


Figure 35 : Transducteur d'annotation des dates absolues en chiffres

A titre d'éclaircissent, les exemples sont annotés par les transducteurs de la Figure 35 et la Figure 36 comme suit :

<time.date.abs> <day>24</day> - <month>02</month> - <year>1970</year> </time.date.abs>
ب تاريخ <time.date.abs> <day> التاسع و العشرين</day> من شهر <month> يناير / كانون الثاني</month> من عام <year> ألف و تسعمائة و خمس و تسعين</year> <time-modifier>ميلادي </time-modifier> </time.date.abs>
<time.date.abs> <day>منتصف</day> شهر <month> غشت</month> عام <year> ١٩٩٥ </year> <time-modifier>م</time-modifier> </time.date.abs>

On note que les dates peuvent prendre des formes non habituelles et que le transducteur de la Figure 36 ne peut pas les reconnaitre. A titre d'exemple on donne les expressions suivantes :

1. في اواخر ستينيات القرن الماضى\fy AwAxr stynyAt Alqrn AlmADy\
2. بداية القرن الحادي عشرة الهجري\bdAyħ Alqrn AlHAdy ςšr Alhjry\

Dans ce cas le transducteur de la Figure 37 traite les exemples ci-dessus en les annotant comme suit :

<time.date.abs> <year>نهاية تسعينيات</year> <century> القرن الماضي</century> </time.date.abs>
<time.date.abs> <year>اوائل</year> <century> القرن الثالث عشرة</century> <time-modifier> الميلادي</time-modifier> </time.date.abs>

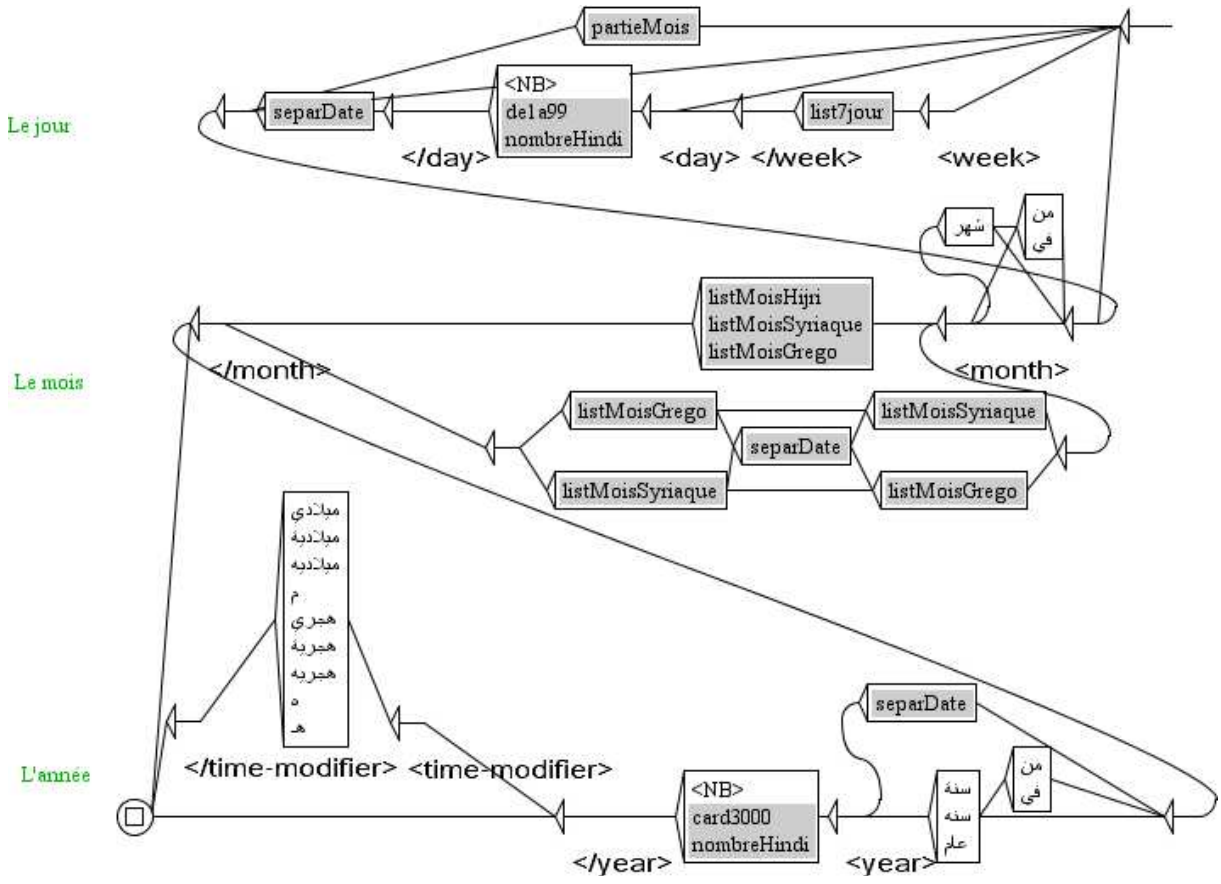


Figure 36 : Transducteur d'annotation des dates en lettres et chiffres absolues

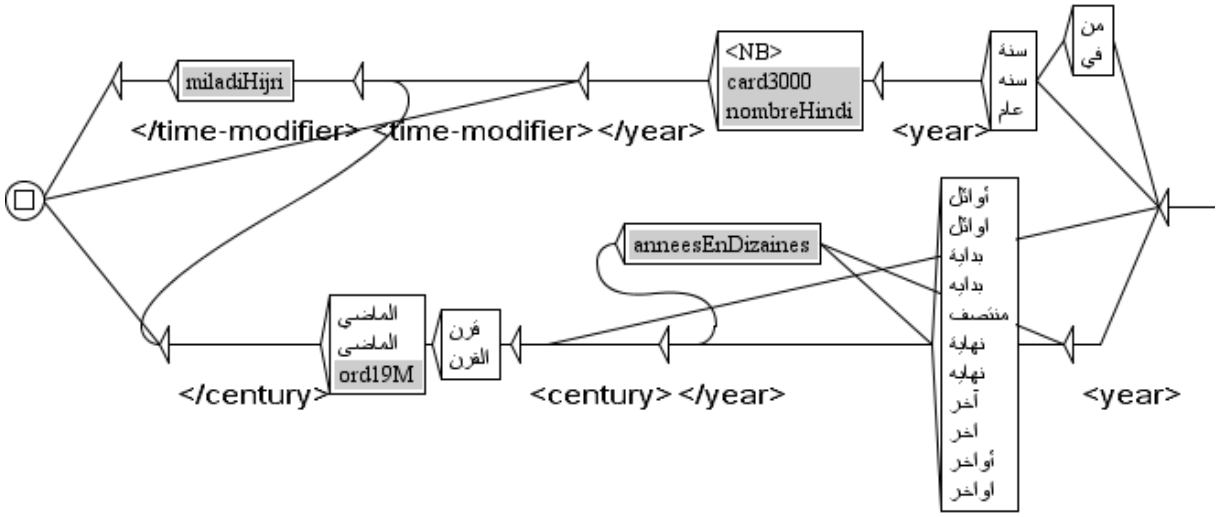


Figure 37 : Transducteur d'annotation des dates absolues sous des formes irrégulières

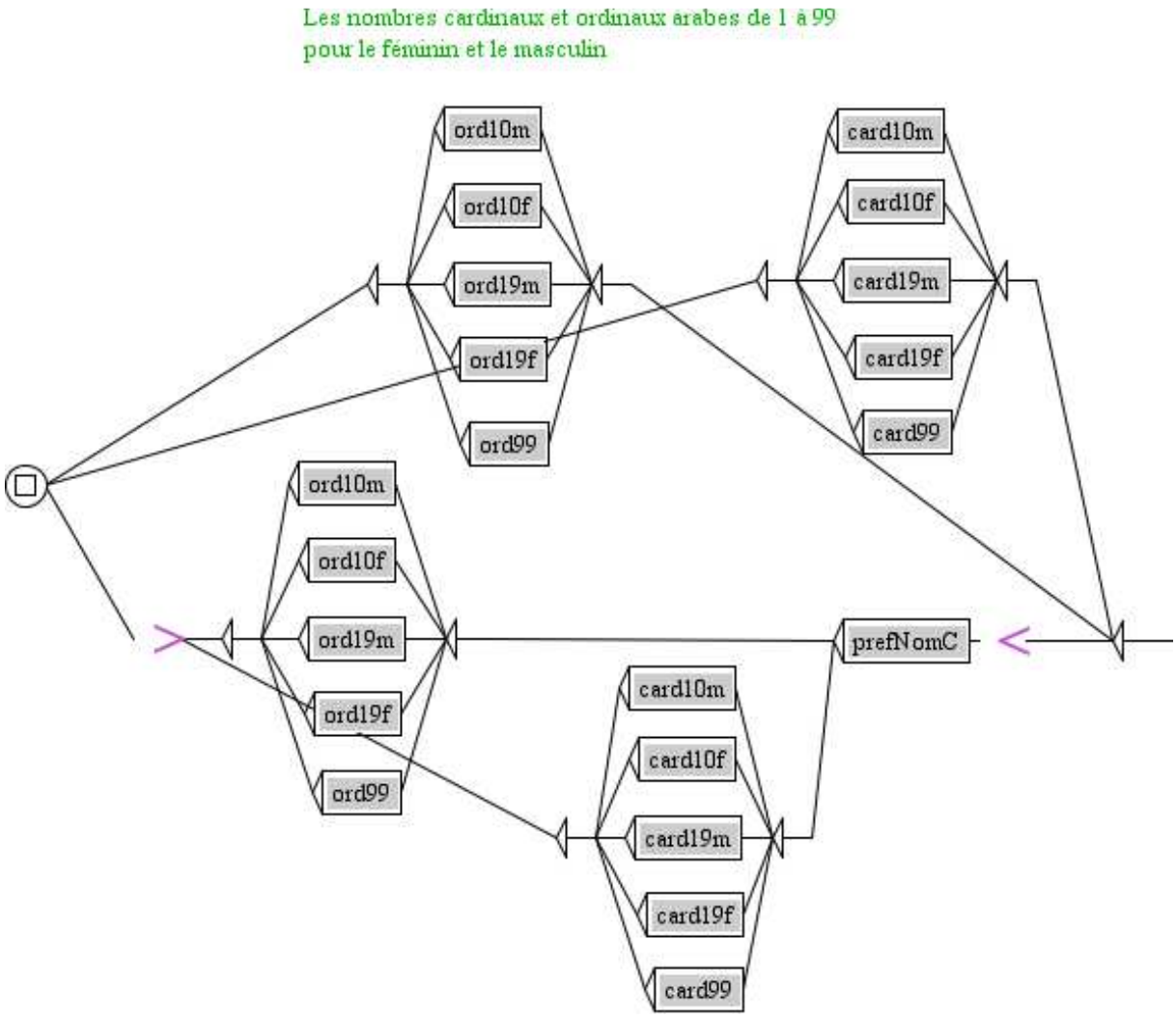


Figure 38 : Automate de reconnaissance des nombres arabes ordinaux et cardinaux de 1 à 99

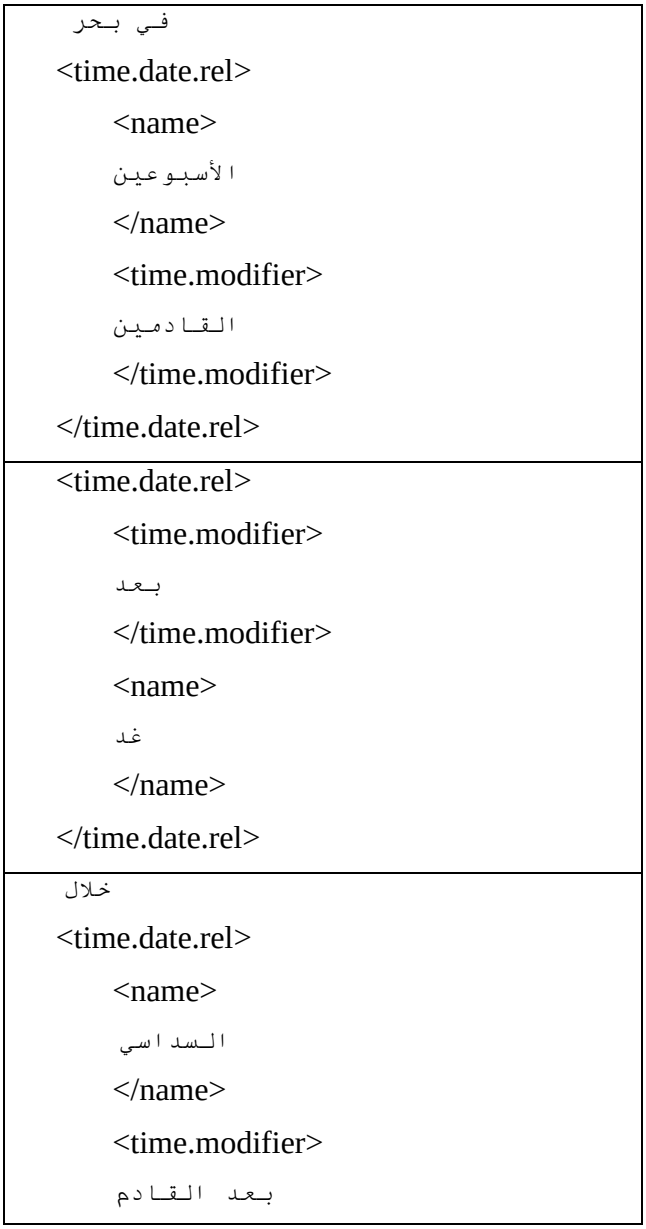
Date relative

Une date est considérée comme relative lorsque la date du document annoté est indispensable pour la déterminer précisément (ESTER2 2007). Dans notre travail on s'intéresse beaucoup plus aux dates absolues car elles s'incluent dans les noms de lieux en imbrication. Par contre les dates relatives se trouvent dans les textes journalistiques sous formes d'expressions temporelles pures i.e. elles ne se trouvent jamais incluses dans d'autres entités nommées.

Les exemples :

- في بحر الأسبوعين القادمين\fy bHr AlÂsbwçyn AlqAdmyn\au cours des deux semaines prochaines
- بعد غد\bçd γd\après demain
- خلال السداسي بعد القادم\xlAl AlsdAsy bçd AlqAdm\au cours du semestre l'après prochain
- منذ السنة الماضية\mnð Alsnħ AlmADyħ\depuis l'année passée

Le graphe de la Figure 40 va annoter les exemples ci-dessous comme suit :



</time.modifier </time.date.rel>

On note que le graphe de la Figure 40 est appelé par un graphe plus général, c'est celui de la Figure 41 et il appelle le sous-graphe de la Figure 39. Ce dernier contient des masques lexicaux comme des entrées dans sa boite. Prenons la première entrée <Adj:f.ماضي> qui désigne toutes les formes fléchies non-vocalisées du lemme ماضي\mADy\passé en tant que *adjectif* (Adj) au *féminin* (f).

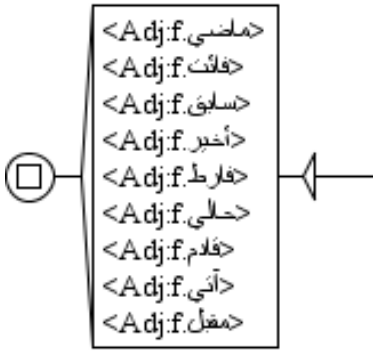


Figure 39 : Sous-graphe des modifier date relative en féminin <modifierDateRelF>

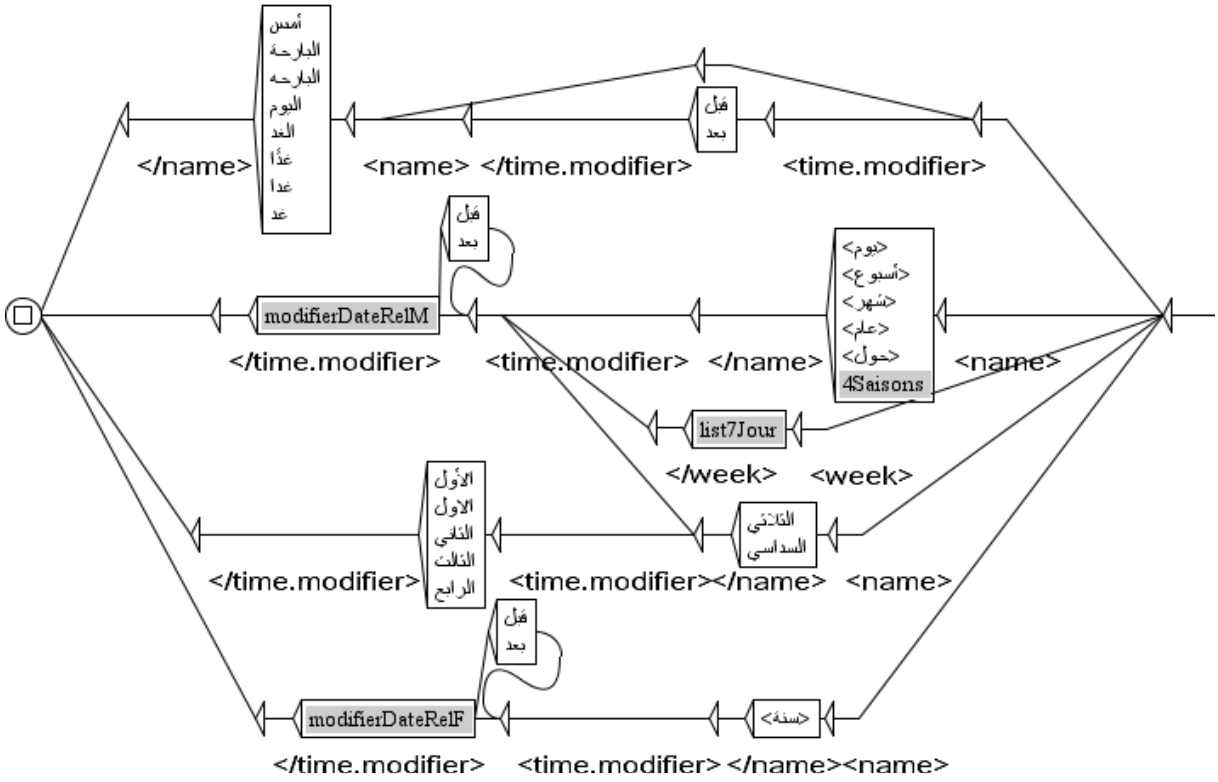


Figure 40 : Sous-graphe de reconnaissance des dates relatives

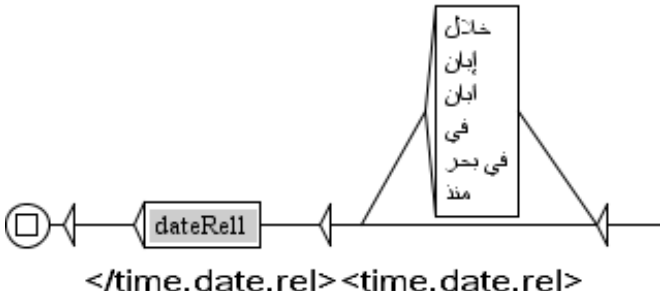


Figure 41 : Transducteur de reconnaissance des dates relatives

Ce masque lexical reconnaît toutes les formes fléchies de cet adjectif qui sont au nombre de 36 formes différentes. Les formes changent en fonction de leurs traits flexionnels, pour un adjectif arabe la flexion est en fonction du genre, nombre, cas, et définitude. Ce nombre est calculé en tenant compte qu'il y a 2 genres (masculin et féminin), 3 nombres (singulier, duel et pluriel), 3 cas (accusatif, nominatif et génitif sans nounation dans le cas du défini et avec nounation dans le cas de l'indéfini) et 2 définitudes (indéfini et défini).

Figure 42 : Dictionnaire DELAF de l'adjectif <قادم>

En calculant les différentes combinaisons on aura 2x3x3x2=36. La Figure 42 montre un exemple du dictionnaire des formes fléchies (DELAF) de l'entrée de l'adjectif قادم\qAdm\prochain. Chaque entrée du dictionnaire est constituée de quatre composantes :

1. la forme fléchie en bleue,
2. la forme canonique ou lemme en rouge,
3. la catégorie syntaxique de la forme fléchie en vert,
4. et les traits flexionnels de la forme en marron.

Si on prend par exemple l'entrée de l'avant la dernière ligne de la Figure 42 qui est :

Dans cet exemple la forme fléchie est entièrement vocalisée (toutes les consonnes sont diacritisées) la production de son équivalent sans voyelles courtes (devocalisé) est possible et facile. Dans notre travail on utilise des méthodes java de la classe *ArabicWord* de la Figure 54 pour supprimer toutes les diacritiques à part le signe de gémination. On désigne les méthodes *removeDiacritics*, *removeDiacriticsButShadda* et *removeDiacrticsShaddaTatweel*.

La forme canonique ou lemme est la forme la plus réduite du mot en question, en arabe cela correspond au masculin singulier indéfini nominatif dans le cas des nominaux (nom commun, adjectif, démonstratif, relatif,…). Dans notre dictionnaire, on préfère l'inclure sans voyelle.

La catégorie syntaxique (POS en anglais) dans notre exemple correspond à l'adjectif (Adj).

Les traits flexionnels (ou code flexionnel) est une suite de lettres, chacune représente un trait. L'ordre suivi est : genre-nombre-cas-définitude. Dans l'exemple :

- genre = f (féminin)
- nombre = d (duel)
- cas = u (nominatif)
- définitude = r (défini)

Le code flexionnel respecte le jeu d'étiquettes conçu pour les dictionnaires (cf. la section 3.3 de la page 82) et pour plus de détails sur ce jeu d'étiquettes le lecteur est orienté à l'annexe B vers la fin de la thèse.

Le dictionnaire des formes fléchies nous sert dans la reconnaissance des entités nommées à reconnaitre, par exemple toutes les formes possibles que peut prendre une expression temporelle d'une date relative. Si on revient à la Figure 40, dans son deuxième et troisième chemin, le transducteur peut reconnaitre toutes les formes possibles de la date relative خلال الأسبوع الماضي\xlAl AlÂsbwς AlmADy\au cours de la dernière semaine.

Par exemple, grâce aux formes fléchies du *nom commun* أسبوع\Âsbwς\ et de l'*adjectif* ماضي\mADy\ qui sont stockées dans les dictionnaires DELAF_N.dic (dictionnaires des formes nominales entièrement vocalisées) et DELAF_N_d.dic (dictionnaire des formes nominales devocalisées) il peut reconnaitre les expressions suivantes :

- الأسبوعان الماضيان
- أسبوعان ماضيان
- الأسبوعين الماضيين
- أسبوعين ماضيين
- الأسابيع الماضية
- أسابيع ماضية
- plus les formes entièrement vocalisées des formes ci-dessus.

Comme il a été indiqué dans la sous-section 4.2.6 de la page 92 la flexion des noms communs et adjectifs est régie par les mêmes règles. Dans notre travail, la flexion de 1317 noms communs et 463 adjectifs a nécessité de concevoir 364 graphes de flexions (transducteurs de flexion) pour les noms communs et 113 graphes de flexion pour les adjectifs. Ce grand nombre de graphes est dû à la non régularité de flexion des nominaux (par rapport aux verbes). Comme il a été indique à la sous-section 4.2.5 de la page 90, on a conçu un algorithme de génération semi automatique des transducteurs de flexion afin de réduire le temps de conceptions d'un graphe. Car le temps de la conception manuelle d'un graphe comme celui de la Figure 19 nécessite au moins deux heures ou plus. Les algorithmes de génération de graphes de verbes et des nominaux sont expliqués sous la sous-section 4.2.1 de la page 86 et codés en java comme il est montré dans la Figure 55 et exactement les méthodes des classes : *FlexionNcAdj*, *FlexionVerbe* et *LADL*.

Le contenu du dictionnaire des noms communs et adjectifs du module arabe de la plateforme Unitex/GramLab est montré dans le Tableau 26 ci-dessous.

Catégorie syntaxique	Forme canonique	Forme fléchie vocalisée	Forme fléchie non vocalisée
Adjectif	463	29169	37053
Nom commun	1317	81726	99021
Total	1780	110895	136074

Tableau 26 : Contenu en noms communs et adjectifs du module arabe d'Unitex/GramLab

4.2.4 Les lieux

Les entités nommées de type *lieux* se divisent en cinq sous-classes : administratif, géographique, voies, bâtiments et adresses. Une sixième sous-classe (autre) est ajoutée pour regrouper le reste des lieux qui ne peuvent pas être annotés par les cinq classes (Quaero 2013). La Figure 43 montre les classes et sous-classes des entités nommées de type lieux selon la classification Quaero.

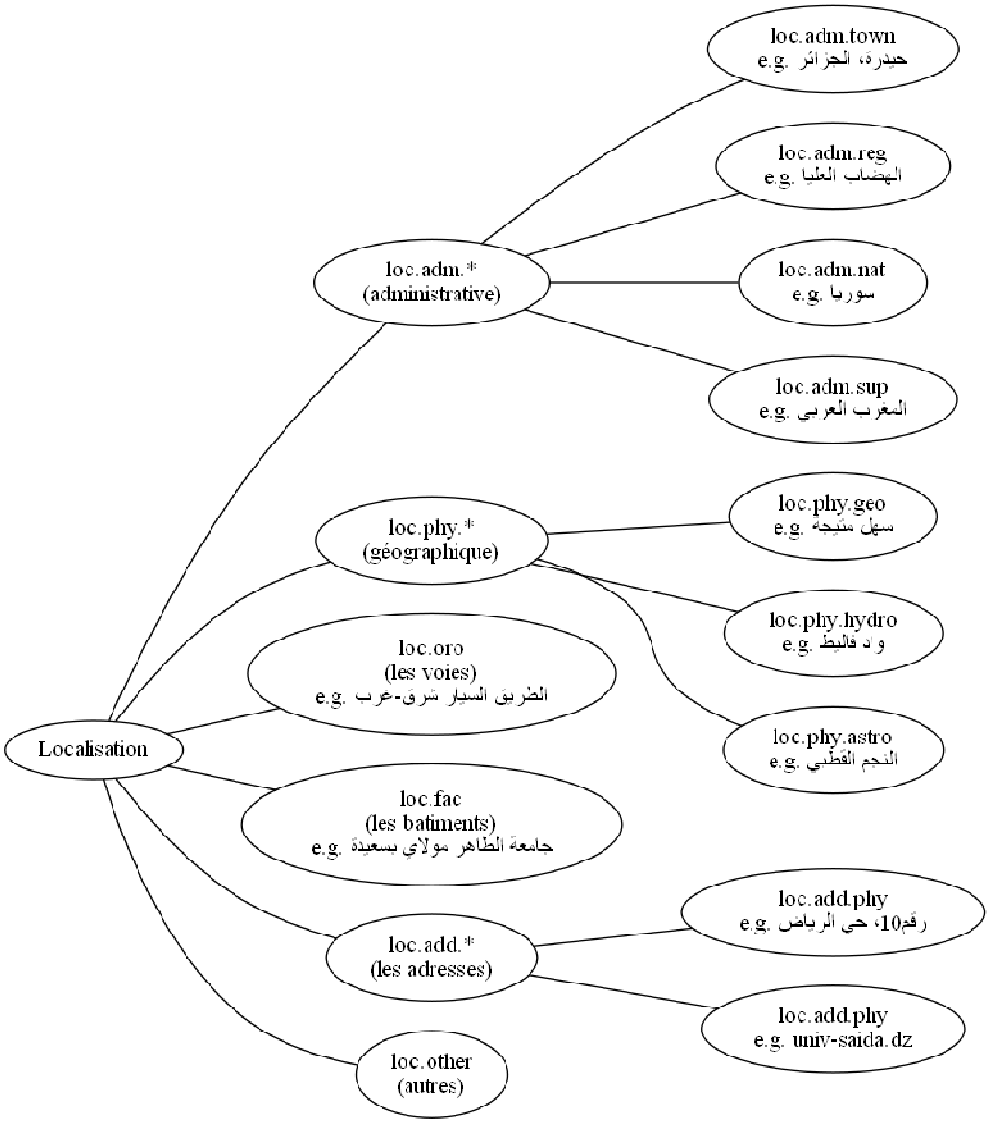


Figure 43 : Les sous-classes de la catégorie <Lieux> suivant Quaero

Elles se trouvent généralement dans les textes en imbrication avec des *noms de personnes* ou des *dates* i.e. elles renferment dans leurs expressions des instances de ces entités nommées. Dans ce cas, pour les détecter on doit passer les transducteurs de détection dans un ordre bien précis en ce qu'on appelle une cascade de transducteurs (Maurel et al. 2011).

Si on prend le quatrième exemple ci-dessous, مضيق جبل طارق\mDyq jbl TArq*étroit de Gibraltar* ; c'est une expression d'entité nommée de type lieux *hydro* composée d'une EN de type lieux *geo* elle-même composée d'une EN de type *pers*. Trois niveaux d'imbrication rendent la détection plus vigilante.

Exemples :

- جامعة الأمير عبد القادر الإسلامية
- ملعب 5 جويليا
- مطار جدة الدولي
- مضيق جبل طارق
- متحف زبانة ب وهران

L’annotation de ces exemples doit être de la manière suivante :

<loc.fac> <kind> جـامـعـة </kind> <pers.ind> <titile> الأمـيـر </title> <name> عبد القادر </name> </pers.ind> <modifier> الإسلامـيـة </modifier> </loc.fac>	*pers imbriqué dans loc (2 niveaux d’imbrication)*
<loc.fac> <kind> مـلـعـب </kind> <time.date.abs> <day> 5</day> <month> جـويـلـيـا </month> </time.date.abs> </loc.fac>	*date imbriquée dans loc (2 niveaux d’imbrication)*
<loc.fac> <kind> مـطـا ر </kind> <loc.adm.town> جـد ة </loc.adm.town> <modifier>الـد ولـي </modifier> </loc.fac>	*loc imbriqué dans loc (2 niveaux d’imbrication)*
<loc.phy.hydro> <kind> مـضـيـق </kind> <loc.phy.geo> <kind> جـبـل</kind> <pers.ind> <name>طـا رق</name> </pers.ind> </loc.phy.geo> </loc.phy.hydro>	*pers imbriqué dans loc imbriqué dans loc (3 niveaux d’imbrication)*
<loc.fac> <kind> مـتـحـف </kind> <pers.ind> <last.name> زبـا نـة </last.name> </pers.ind> ب <loc.adm.town>و هـر ا ن </loc.adm.town> </loc.fac>	*pers et loc imbriqués dans loc (2 niveaux d’imbrication)*

Le repérage des entités nommées imbriquées nécessitent l’exécution des transducteurs l’un après l’autre en une cascade, donc nécessitent la conception des transducteurs ainsi que le choix de l’ordre de leur exécution. Par exemple pour détecter une EN de type *lieux* qui renferme une EN de type *personne*, on doit passer le transducteur de la reconnaissance de l’EN la plus emboitée (*personne*) puis appeler le transducteur de l’EN qui emboite (*Lieux*). En général, dans une cascade, on passe les transducteurs les moins ambigus puis les plus ambigus.

En arabe par exemple, si on compare le *degré d’ambiguïté* entre les noms communs (gérondif masdar) et noms propres de lieux (ville, département, pays, continent, etc…) on le

trouve inferieur à celui entre les noms communs (adjectifs et gérondifs) et noms propres de personnes. Car un grand pourcentage des nouveaux nés dans le monde prennent comme prénom des adjectifs et des gérondifs e.g. جميل\jmyl\beau, أمينة\Âmynħ\honnête, وسيم \wsym\handsome, نبيلة\nbylħ\pairesse, منتصر\mntSr\victorieux, إيمان\ĂymAn\foi, إسلام\ĂslAm\soumission, خلود\xlwd\immortalité, وئام\wŷAm\concorde et فرح \frH\contentement.

Par contre les dates absolues sont les moins ambiguës donc on les passe les premieres dans la cascade de repérage des EN de type lieux. La Figure 46 montre une capture d'écran du système CasSys (Maurel and Friburger 2013) (cf. la sous-section 4.2.6 de la page 120) qu'on a utilisé dans notre travail et l'ordre d'exécution des transducteurs qu'on a choisi pour notre cascade de reconnaissance des *lieux* (appelée araCasEN).

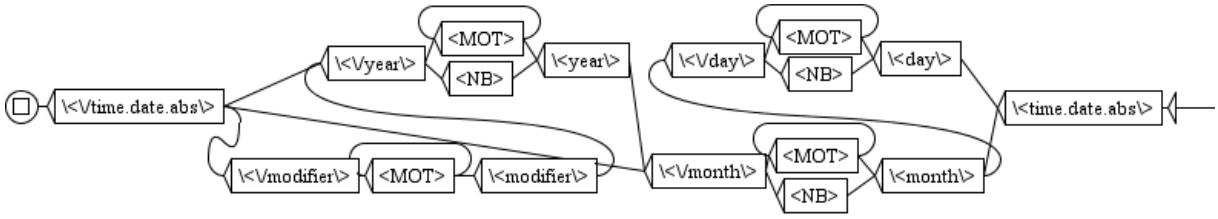


Figure 44 : Transducteur de reconnaissance des dates absolues balisées <dateTag.grf>

4.2.5 Dictionnaire des lieux

Afin de renforcer la reconnaissance des lieux et afin de les utiliser comme preuve interne dans les transducteurs on a compilé un dictionnaire des noms propres des lieux. La principale source des ces informations de lieux c'est le wikipedia. Comme il est montré dans la Figure 45 à l'intérieur des boites du transducteur, on a utilisé le masque lexical <Np+xxx> où :

- Np indique que la catégorie syntaxique de l'entrée qui doit être un *nom propre* (cf. le jeu d'étiquettes dans l'annexe B),
- xxx représente la catégorie sémantique du lieu.

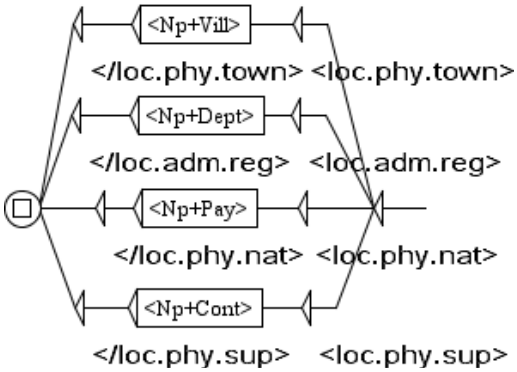


Figure 45 : Transducteur de reconnaissance des lieux (villes, départements, pays et régions)

Les entrées du dictionnaire des lieux sont tolérantes aux erreurs i.e. dans ce dictionnaire on a calculé toutes les formes possibles que peut prendre un nom de lieux en prenant en considération les erreurs typographiques les plus répandus que les journalistes arabes commettent dans leurs articles journalistiques (cf. le Tableau 25). A ce problème

typographique s'ajoute le fait qu'un lieu peut avoir différentes façons d'appellation. Comme le montre le Tableau 27, le pays Algérie peut être appelé de deux façons :

- الجزائر\Al.jazaAŷir*Algérie* ou encore
- الجمهورية الجزائرية الديمقراطية الشعبية\Al.jam.huwriyaħu Al.jazaAŷiriyaħu Ald~iymuq.raATiyaħu Alš~aç.bĨyaħ*République Algérienne démocratique et populaire*

Le Tableau 27 montre le contenu actuel des dictionnaires des lieux et qui sont disponibles librement sous la plateforme Unitex/GramLab. En ce qui concerne la couverture, le contenu actuel couvre tous les noms des pays du monde et la majorité des noms de villes du monde arabe, de l'Europe, de la Russie et de l'Amérique du nord avec leurs départements.

Dictionnaire des lieux	Nombres d'entrées	Exemple
Les pays	795	e.g. pour l'Algérie on a deux entrées : • الجمهورية الجزائرية الديمقراطية الشعبية,.Np+Pay • الجزائر,.Np+Pay
Les régions (Continents, régions de continents e.g. Maghreb)	46	e.g. pour l'Afrique on a plusieurs entrées : • أفريقيا,.Np+Cont • إفريقيا,.Np+Cont • إفريقية,.Np+Cont • etc…
Les départements (états, willayas, fédérations, etc.)	525	e.g. ماساتشوستس,.Np+Dept
Les villes (village, ville, capitales)	7102	e.g. بوسطن,.Np+Vill
Total	8468	

Tableau 27 : Contenu des dictionnaires des noms de lieux

4.2.6 Cascade de transducteurs

Le système CasSys (Maurel and Friburger 2013) permet de concevoir une cascade de transducteurs et de les exécuter suivant leur ordre. Comme on a indiqué dans les sections précédentes et comme il est montré dans la Figure 46 ci-dessous qui représente l'ensemble de transducteurs de la cascade araCasEN1 qui reconnait les lieux qui peuvent se trouver en imbrication avec des dates et des noms de personnes.

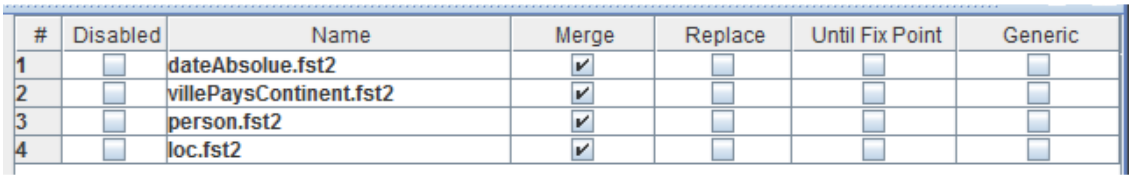


Figure 46 : L'ordre d'exécution de transducteurs dans la cascade <araCasEN1> sous le système CasSys

Dans ce qui suit, on va montrer les différentes exécutions de la cascade araCasEN1 sur un texte contenant des expressions de noms de lieux renfermant des noms de personnes et des dates (en rouge dans le texte ci-dessous).

انطلق المتظاهرون من ساحة الأمير عبد القادر وتواصلت مسيرتهم عبر شارع أول نوفمبر حتّى وصلت إلى نقطة النهاية

Le premier transducteur qui va être exécuté sur le texte brut ci-dessus est le transducteur des dates absolues montré dans la Figure 47 et qui lui-même appelle les transducteurs de la Figure 35 et la Figure 36 ci-dessus.

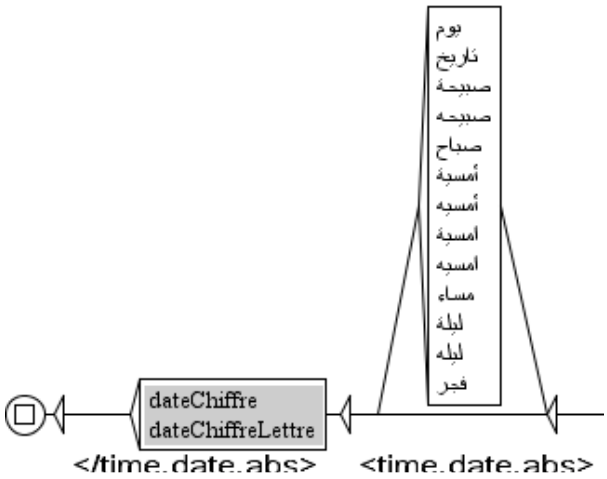


Figure 47 : Transducteur principal des dates absolues

Ce transducteur transforme le texte brut en un premier texte partiellement traité qui doit avoir la forme suivante :

انطلق المتظاهرون من ساحة الأمير عبد القادر وتواصلت مسيرتهم
<time.date.abs><day>أول</day> عبر شارع
<month>نوفمبر</month></time.date.abs> حتّى وصلت إلى نقطة النهاية

Ce dernier texte va être passé au transducteur de la Figure 45 qui traite les noms de lieux sous forme de noms de ville, de pays etc. Ce dernier ne change rien dans le texte car il ne rencontre aucune instance de des EN recherchées.

A l'étape suivante le texte est passé au transducteur de Figure 48 qui lui-même appelle ses sous-graphes de reconnaissance des différentes parties qui composent un nom de personne telles que la partie *titre* qui est reconnue par le transducteur de la Figure 33 ci-dessus.

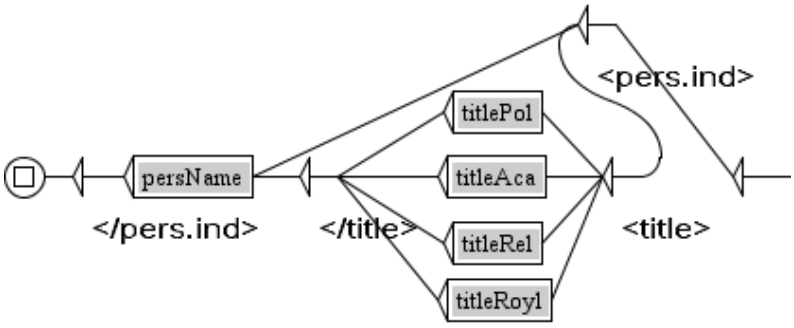


Figure 48 : Transducteur d'annotation des noms de personnes

Le résultat de l'exécution du transducteur de la Figure 48 sur le texte précédent est le texte ci-dessous. On remarque bien que les EN de types date et nom de personne sont bien annotées alors que les expressions exprimant les lieux (en rouge dans le texte) attendent toujours le passage du transducteur en question pour finaliser le repérage.

انطلق المتظاهرون من ساحة
<pers.ind><title><func.ind>الأمير</func.ind></title><name.first> عبد
القادر</name.first></pers.ind> وتواصلت مسيرتهم عبر شارع
<time.date.abs><day>أول</day> <month>نوفمبر</month></time.date.abs>
حتّى وصلت إلى نقطة النهاية

Le dernier transducteur de la cascade araCasEN1 c'est celui de la détection des lieux comme le montre la Figure 49. Dans cette figure on utilise des listes de preuve interne comme par exemple le cas de la première boite du premier chemin du transducteur ou du sous-graphe *listFac* du deuxième chemin. La liste en vert (commentaire en Unitex/GramLab) représente le contenu de ce sous-graphe.

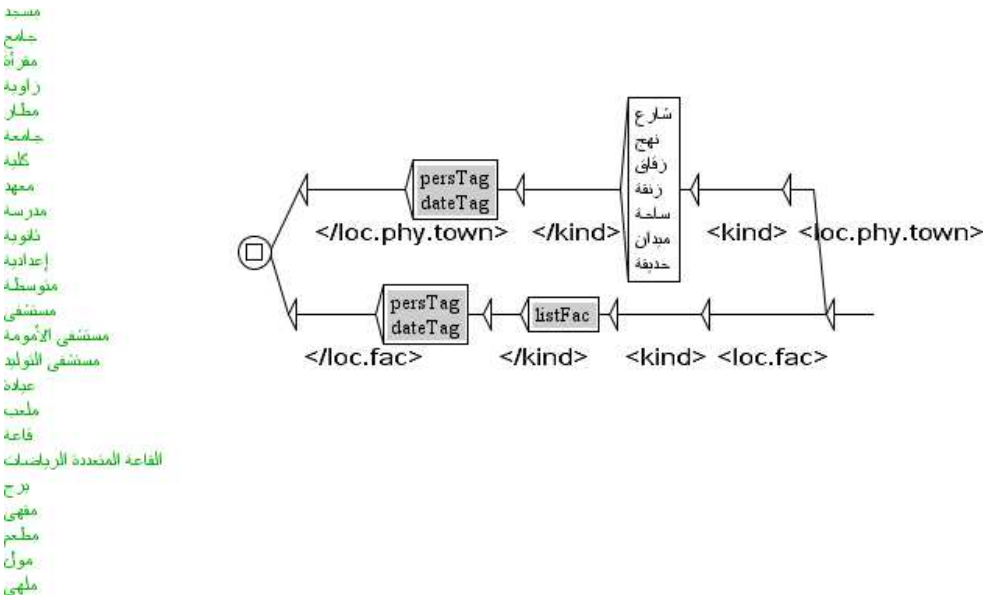


Figure 49 : Transducteur principal de la reconnaissance des lieux

Le résultat final de la cascade sur le texte est montré dans le texte balisé ci-dessous. Tout est balisé à part les expressions qui n'expriment aucune entité nommée (en bleu).

انطلق المتظاهرون من<loc.phy.town><kind> ساحة </kind>
<pers.ind><title><func.ind>الأمير</func.ind></title><name.first> عبد
القادر</name.first></pers.ind></loc.phy.town> وتواصلت مسيرتهم
عبر<loc.phy.town><kind> شارع <time.date.abs><day>أول</day> </kind>
<month>نوفمبر</month></time.date.abs></loc.phy.town> حتّى وصلت
إلى نقطة النهاية

```
انطلق المتظاهرون من
<loc.phy.town>
    <kind>ساحة</kind>
    <pers.ind>
        <title>
            <func.ind>الأمير</func.ind>
        </title>
        <name.first> عبد القادر</name.first>
    </pers.ind>
</loc.phy.town>
و تواصلت مسيرتهم عبر
<loc.phy.town>
    <kind>شارع</kind>
    <time.date.abs>
        <day>أول</day>
        <month>نوفمبر</month>
    </time.date.abs>
</loc.phy.town>
حتى وصلت إلى نقطة النهاية
```

Figure 50 : Le résultat final l'application de la cascade araCasEN1sur le texte brute de la page 120

5. La détection de relation entre les EN arabes

La détection de relations entre les EN nécessite que celles-ci soient déjà complètement repérées dans le texte et puis on passe le système de détection de relations qu'on a nommé araCasRel. Ce dernier est aussi réalisé en cascade.

La cascade d'araCasRel commence par cacher dans le texte, le balisage XML que araCasEN a réalisé en remplaçant l'élément XML par la classe principale d'EN de Quaero correspondante. Cette procédure facilitera davantage le traitement de corpus et augmentera la lisibilité du texte.

Par exemple l'élément XML :

```
<loc.phy.town>
        <kind> ساحة</kind>
        <pers.ind>
                <title>
                        <func.ind> الأمير</func.ind>
                </title>
                <name.first> عبد القادر</name.first>
        </pers.ind>
</loc.phy.town>
```

devient une étiquette simple <loc>.

Ainsi les textes bruts suivants deviennent comme il est indiqué dans la deuxième colonne du Tableau 28 de la page suivante.

Texte brut	Résultat d'araCasRel
القمة العربية : استقبل رئيس الجمهورية السيد عبد العزيز بوتفليقة ب مقر رئاسة الجمهورية وزير الخارجية السوري السيد فاروق الشرع	: <event> استقبل <pers> ب<loc> <pers>
و افاد بيان ل وزارة التجارة ان جعبوب استقبل الوزير الصيني على هامش أشغال الدورة الخامسة ل الجنة المشتركة الجزائرية الصينية ل التعاون التي انعقدت في الجزائر	و افاد بيان ل<org> ان <pers> استقبل <pers> على هامش أشغال <event> التي انعقدت في <loc>
كما استقبل محافظ بنك الجزائر مر فوقا ب نظيره التونسي من قبل وزير المالية التونسي السيد محمد رشيد كشيش قبل أن يقوم الوفد الجزائري...	كما استقبل <func> مر فوقا ب<func> من قبل <pers> قبل أن يقوم الوفد الجزائري...
كما التقى السيد بلخادم ب السيد باك نام سون وزير خارجية جمهورية كوريا الديمقراطية	كما التقى <pers> ب<pers>
ست مؤسسات جزائرية تشارك في صالون دولي ب اسبانيا حول صناعة زيت الزيتون	ست <org> تشارك في <event> ب <loc> حول صناعة زيت الزيتون
مفرزة من القوات البحرية تشارك في استعراض بحري ب فرنسا	<func> تشارك في <event> ب <loc>
الصحراء الغربية : كاتبة الدولة الإسبانية للتعاون تزور مخيمات اللاجئين الصحراويين	: <loc> <func> تزور <loc>

Tableau 28 : Traitement partiel d'araCasRel

Les verbes seront ensuite lemmatiser en passant un transducteur utilisant la technique des variables d'entrée (Paumier 2014) en mode *repalce*. Ce transducteur utilise les entrées du dictionnaire DELAF passé en premier lieu (application des dictionnaires par le programme Dico) pour les annoter avec toutes les informations de l'entrée. Parmi les informations de l'entée on trouve le lemme du verbe. Le mode *replace* permet de remplacer la forme fléchie du verbe par son lemme. Par exemple dans la dernière ligne du Tableau 28 dans le balisage <loc> : <func> تزور <loc>; la forme fléchie تزور\tazuwru*elle visite* est lemmatisée en زار\zaAra*visiter* ce qui simplifie le travail d'ultérieur. Donc le balisage devient <loc> : <func> زار <loc>.

En utilisant Arabic WordNet (El-Kateb et al. 2006), on cherche par lemme, le verbe qui représente la relation. La Figure 51 représente Arabic WordNet browser dans lequel on a introduit le lemme زار\zaAra*visiter* dans le champ *Arabic word* pour rechercher ses sens dans les trois ontologies : Arabic WordNet, SUMO et Princeton WordNet.

La recherche du sens de ce lemme a donné quatre synsets d'AWN affichés dans l'espace *Arabic word senses*. Chacun des synsets représente un sens bien déterminé et composé d'un

ensemble de mots synonymes. Dans l'affichage chaque ligne représente un synset. Par exemple, le troisième synset est composé de six mots synonymes alors que le dernier synset est composé d'un seul mot.

Dans notre approche, l'expert linguiste doit trancher à quel synset doit appartenir notre verbe représentant la relation sémantique recherchée entre EN. Pour ce faire il se base sur sa propre connaissance linguistique pour désambiguïser le sens.

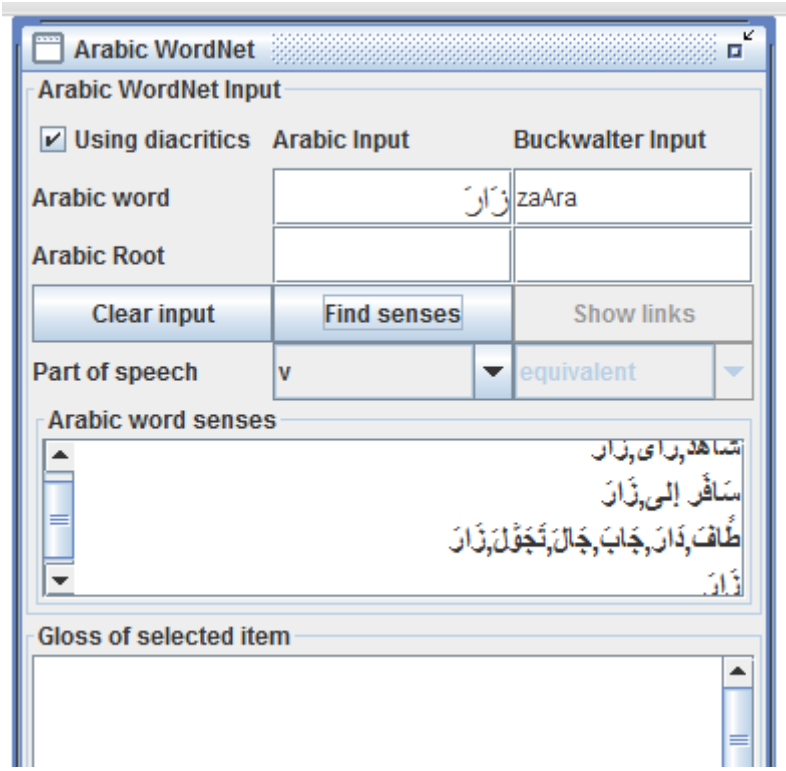


Figure 51 : L'entrée zaAra dans Arabic WordNet

Dans l'exemple de la Figure 51, si l'expert choisit par exemple de valider le premier synset comme equivalent en sens à notre verbe alors le concept d'ontology SUMO correspondant va être SocialInteraction comme il est montré dans le Tableau 29 ci-dessous.

```
::: SocialInteraction  ::: (SUBSUMING)

"The subclass of IntentionalProcess that involves interactions
between CognitiveAgents."

SocialInteraction is a subclass of IntentionalProcess.
ChangeOfPossession is a subclass of SocialInteraction.
Communication is a subclass of SocialInteraction.
Contest is a subclass of SocialInteraction.
Cooperation is a subclass of SocialInteraction.
Meeting is a subclass of SocialInteraction.
Pretending is a subclass of SocialInteraction.
```

Tableau 29 : Gloss d'un concept de l'ontologie SUMO

La Figure 52 réprésente l'interface d'Arabic WordNet à partir de laquelle l'expert linguite peut désambiguiser le sens. La figure montre le mapping entre l'AWN et PWN en utilisant l'ontologie de haut niveau SUMO.

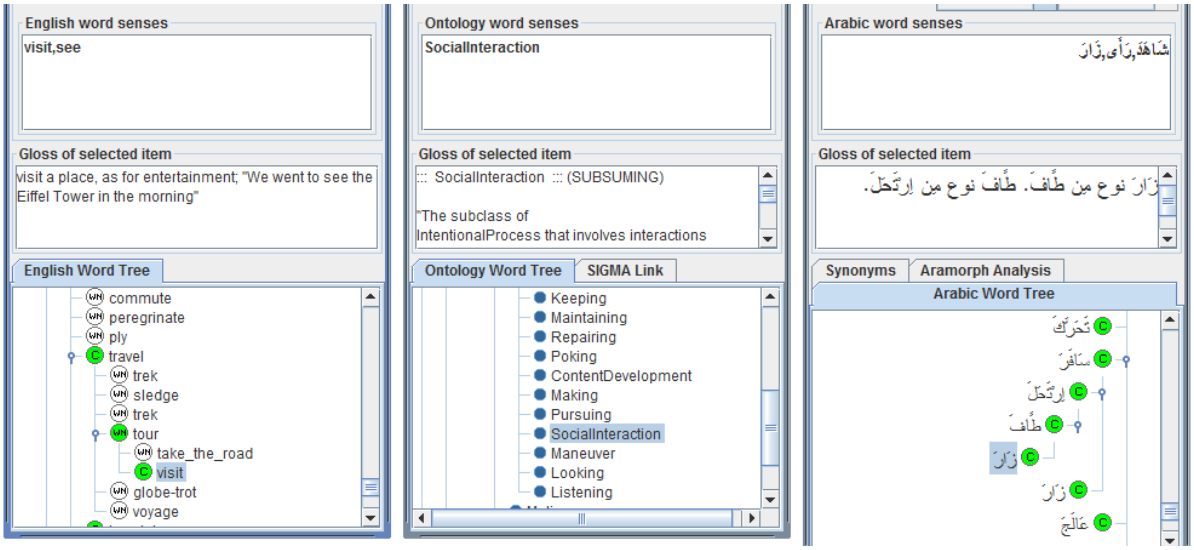


Figure 52 : Désambiguisation du sens de zaAra en utilisant les trois ontologies : AWN, SUMO et PWN

6. Evaluation

L'évaluation de notre système nous a poussés à développer un éditeur d'annotation en entités nommées arabes. La Figure 53 représente une capture d'écran de cet outil contenant un texte en cours d'annotation. On remarque clairement que l'hiérarchie de la typologie Quaero est disponible en menu contextuel. Cette manière de faire facilite l'opération à l'annotateur qui doit être un expert en étiquetage des entités nommées arabes.

Le résultat de l'annotation est un texte balisé en XML. Le projet Quaero propose l'annotation des EN en types/sous-types et composants. Par exemple dans la deuxième ligne du texte de la Figure 53 l'occurrence آسيـا\Ās.yaA*Asie* est annotée comme sous-type *loc.adm.reg* et comme composant *name*.

Une veritable évaluation d'un système de reconnaissance des entités nommées doit être effectuée en comparaison avec un corpus manuellement annoté par des experts dans des campagnes d'évaluation. Le manque de telles ressources pour la langue arabe nous a privé d'effectuer une évaluation objective de notre système de reconnaissance d'entités nommées et d'extraction de relations les reliant.

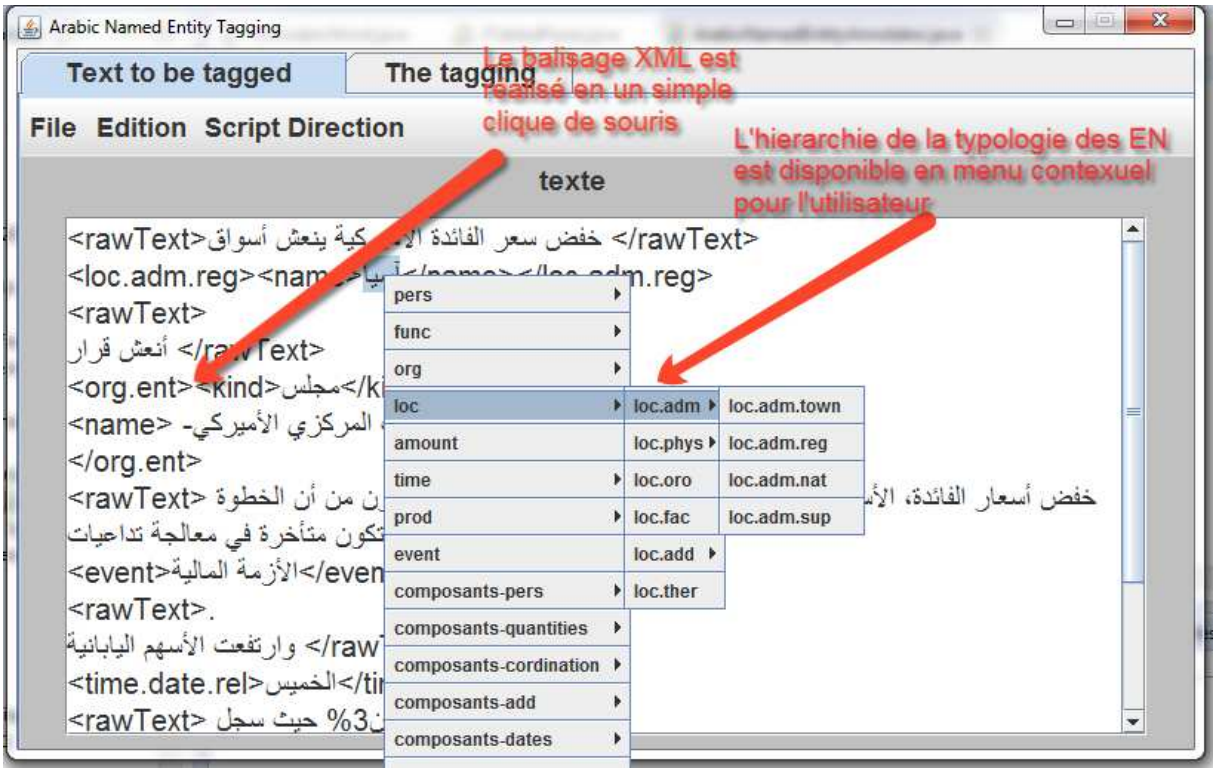


Figure 53 : Editeur d'annotation en EN arabes respectant la typologie de Quaero

7. Conclusion

L'expérimentation de nos ressources et de notre système nous a permis de dégager un ensemble de remarques et de conclusions. Ces conclusions peuvent être résumées dans les limites de notre système ainsi que des ressources construites. Les limites sont :

- couverture des ressources des nominaux très réduite,
- manque d'un correcteur automatique des erreurs typographiques et orthographiques robuste,
- manque d'un système de bonne désambiguïsation lexicale,
- manque d'une bonne segmentation des clitiques,
- manque d'une bonne segmentation de phrases.

Comme il est mentionné dans la section précédente, le manque de ressources d'évaluation nous a empêché d'évaluer les performances de notre système d'où la création d'un outil de construction de corpus annoté en EN. Ce travail ainsi que les limites de notre système et nos ressources linguistiques restent comme des projets en perspectives.

Conclusion générale

Les objectifs de cette thèse étaient multiples, la réalisation de ces objectifs nous a conduits à entamer plusieurs disciplines en même temps et à résoudre beaucoup de problèmes. La langue traitée dans le cadre de cette thèse est lexiquement riche, morphologiquement complexe et d'un degré d'ambiguïté élevé. A ces défis s'ajoute le manque terrible en ressources et outils de traitement automatique ouverts à la communauté de TAL arabe. Cette situation a rendu la tache très difficile aux chercheurs du domaine.

Pour notre cas, nous étions contraints à ouvrir tout un chantier à part, juste pour préparer le terrain à la réalisation du sujet. Au début, ce chantier de construction de ressources n'était pas un objectif, mais le cours des choses le rendait inévitable.

Ce chantier avait pour objet de combler le manque observé en termes de ressources et d'outils de traitement de corpus, ainsi le choix c'était sur la plateforme Unitex/GramLab qui à l'époque ne contenait pas de module pour l'arabe. Cette plateforme par sa nature libre et open source nous a présenté à la fois une opportunité et un alternatif : c'est une opportunité pour un chercheur en TAL arabe pour participer au développement d'un logiciel à un impact important et d'une grande envergure. Cette opportunité lui permet aussi de développer des ressources et des outils et de les partager avec ses confrères du domaine afin de les valider. Et c'est un alternatif aux ressources et outils propriétaires payant qui sont au-delàs du budget d'un chercheur ou d'un doctorant.

Le développement du module arabe d'Unitex/GramLab nous a permis de relire la langue arabe et de la voir d'un nouvel angle ce qui nous a procuré de nouvelles connaissances sur cette langue très riche et ses phénomènes langagiers. La construction de ressources de cette langue nous a fourni une vision claire sur le système morphologique arabe.

L'utilisation des technologies à états finis comme formalisme pour modéliser les ressources linguistiques et les règles de leurs construction d'une part et les règles de notre système d'extraction d'entités nommées et des relations les reliant d'autre part nous a démontré la puissance de ce formalisme et nous a ouvert des perspectives sur son utilisation pour modéliser d'autres phénomènes langagiers (cf. section perspectives ci-dessous).

L'étude des entités nommées arabe particulièrement les noms propres de personnes nous a ouvert des perspectives sur plusieurs applications qui peuvent bénéficier de cette tâche (cf. perspectives), ainsi qu'à l'application des extracteurs sur des textes de domaines bien précis. Ici on désigne les textes qui contiennent une terminologie bien limitée tels que les textes écrits en langage juridique dans les institutions juridiques.

D'après les travaux qu'on a entrepris pour la langue arabe dans le cadre de cette thèse ou les travaux qu'on a côtoyé ; que ce soit en langue française dans les projets du laboratoire LI de l'université de Tours ou les travaux sur la langue arabe du laboratoire MIRACL de l'université de Sfax, on conclut que l'utilisation de cascade de transducteurs donne une grande souplesse dans le traitement et rend la tâche beaucoup plus facile et maitrisable. Cette modularité en traitement de corpus offre à l'utilisateur tous les avantages connus de cette stratégie tels que la réutilisabilité de transducteur, l'aisance de débogage, la facilité de maintenance, l'évolutivité selon le besoin etc. Le système CasSys développé au laboratoire LI a montré aussi une grande performance et aisance en manipulation de cascades de transducteurs.

Au cours de la réalisation de l'extraction de connaissances à partir du texte, on a remarqué que l'intérêt de cette tache par rapport à l'ingénierie d'ontologie est grandiose. Que se soit pour la détection, repérage et classification des EN ou pour l'extraction des relations sémantiques non-taxonomiques les reliant, l'intérêt de cette extraction consiste à la

conversion de la donnée et de la connaissance d'une forme non structurée et non exploitée vers une forme lisible et exploitable par l'ordinateur. Malheureusement la non disponibilité de corpus annotés en EN et en relations sémantiques nous a privé de l'évaluation réelle de notre système.

Quant à l'approche scientifique utilisée pour atteindre nos objectifs (l'approche linguistique déterministe à base de règles au lieu de l'approche statistique non déterministe) on a confirmé ce que la littérature affirme. L'approche déterministe est couteuse en effort et en temps. Elle demande une grande et profonde connaissance du problème à résoudre et demande une couverture totale des règles régissant les phénomènes étudiés. D'où le recourt inévitable aux experts du domaine pour se procurer la connaissance requise.

Perspectives

Au fur et à mesure de la préparation de cette thèse plusieurs points ont surgis et devenus des projets de recherche à petite, moyenne et grande taille. Par manque de temps et afin de ne pas se perdre dans le travail et de ne pas perdre les objectifs principaux de cette thèse, on les a planifié comme des perspectives de court, moyen et long terme. On les résume dans les points suivants :

- Le module arabe d'Unitex/GramLab dans son état actuel n'est pas complet pour faire toutes les tâches du TAL arabe attendues par un chercheur du domaine, ce besoin nous oblige à concevoir des solutions aux problèmes posés, on cite quelques unes :
    - un module de correction des erreurs orthographiques fréquentes,
    - un transducteur de segmentation des clitiques pour résoudre l'obstacle d'agglutination,
    - un algorithme de matching entre le token du corpus, partiellement vocalisé ou contenant une séquence de caractère d'allongement (la kashida) et l'entrée normalisée du dictionnaire,
    - un algorithme de compression spécial pour les dictionnaires arabes,
    -
- L'apparition dans les dernières années de plusieurs nouveaux outils de TAL arabe sous forme de logiciels, API, plug-ins ou module par de grand laboratoires et centres de recherche dans les universités de renommée mondiale nous invite à utiliser ces produits dans nos travaux futurs
- Penser à appliquer les techniques de compressions des automates utilisées originellement dans le stockage des informations biologiques, dans le stockage des informations lexicales de la langue arabe
- Réflichir à la conception des transducteurs de reconnaissance des noms de personnes de l'arabe classique tels que les noms des narrateurs du Hadith et ainsi on peut intégrer ce module dans un logiciel d'authentification du Hadith qui peut avoir de grandes retombés sur la science du Hadith
- Nous nous fixons, à l'avenir, de limiter le domaine des textes à traiter afin de bien maitriser le lexique et la terminologie qui y sont utilisés. Cette restriction nous garantira une bonne qualité de résultat et un taux de couverture très élevé. Notamment si on vise les institutions algériennes qui utilisent la langue arabe standard moderne dans leurs correspondances administratives et quotidiennes telles que la justice, la sureté nationale, la défense et la gendarmerie nationale.
- Vu l'engouement de chercheurs du domaine concernant les ressources linguistiques de l'arabe dialectal, nous espérons contribuer à la construction des dites ressources. On vise exactement à la compilation de petits lexiques de l'arabe algérien et de

tenter de formaliser les règles qui régissent la flexion des verbes, la déclinaison des noms et des adjectifs. Parmi nos plans futurs, nous essayons d'utiliser les machines à états finis dans cette formalisation en raison de la souplesse qu'on a trouvée lors de la compilation des dictionnaires de l'ASM.

- Avec l'évolution fulgurante des outils de TAL arabe ces toutes dernières années, nous espérons en faire usage et les intégrer dans nos travaux. En tous cas nous prédisons un bel avenir pour le TAL arabe et dans un futur très proche.

Bibliographie

Abbes, R. (2004). *La conception et la réalisation d'un concordancier électronique pour l'arabe. Thèse de Doctorat,* L'institut national des sciences appliquées de Lyon,

Abbes, R., Dichy, J., & Hassoun, M. The architecture of a standard Arabic lexical database: some figures, ratios and categories from the DIINAR.1 source program. In *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, Geneva, Switzerland, 2004* (pp. 15-22). 1621810: Association for Computational Linguistics

Abdelali, A., Cowie, J., & Soliman, H. S. (2005). *Building A Modern Standard Arabic Corpus.* Paper presented at the Workshop on Computational Modeling of Lexical Acquisation, Croatia, 25-28 July

Abdelrahman, S., El-Arnaoty, M., Magdy, M., & Fahmy, A. (2010). Integrated Machine Learning Techniques for Arabic Named Entity Recognition. *International Journal of Computer Science Issues, 7*(4).

Abdulhay, A. (2012). *Constitution d'une ressource sémantique arabe à partir de corpus multilingues alignés. Thèse de Doctorat,* Université de Grenoble,

Al-Afghani, S. (Ed.). (1971). *الموجز في قواعد اللغة العربية\Almuwjaz fiy qawaAçid All~uγaħ Alçarabiyaħ\ Le résumé des règles de la langues arabe*. Damas, Syrie: Dar Al-fikr.

Al-Bawab, M. (2007). Arabic derivation and inflection algorithms. *Sarf system documentation*. Tunisia: ALESCO : Arab League Educational, Scientific and Cultural Organization.

Al-Bawab, M., Merayati, M., Mir Alam, Y., & Al-Tayene, M. H. (1994). A computerized morpho-syntactic system of Arabic. *The Arabian Journal of Science and Engineering*, 19.

Al-Bawab, M., Merayati, M., Mir Alam, Y., & Al-Tayene, M. H. (1996). *Statistics on Arabic verbs in the computational lexicon*. Lebanon: Librairie Du Liban Publishers.

Al-Jumaily, H., Martinez, P., José, M.-F., & Goot, E. (2012). A real time named entity recognition system for Arabic text mining. *Language Resources and Evaluation, 46*(4), 543–563.

Al-Kalak, A., Al-Bawab, M., Merayati, M., Mir Alam, Y., Al-Attar, S., Mohtasib Bellah, H., et al. (2007). Sarf Arabic Morphology System. (1.0 ed.). Tunisia: ALESCO.

Al-Kharashi, I. A. Person Named Entity Generation and Recognition for Arabic Language. In *Proceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo, 2009* (pp. 205-208)

Al-Najem, S. R. (2007). Inheritance-based Approach to Arabic Verbal Root-and-Pattern Morphology. In A. Soudi, A. v. d. Bosch, & G. Neumann (Eds.), *Arabic computational morphology: Knowledge-based and Empirical Methods* (Vol. 38, pp. 67-88, Text, Speech and Language Technology): Springer.

Al-Onaizan, Y., & Knight, K. Machine transliteration of names in Arabic text. In *Proceedings of the ACL-02 Workshop on Computational Approaches to Semitic Languages (SEMITIC 2002), Stroudsburg, PA, 2002* (pp. 1-13)

Al-Qrainy, S., & Ayesh, A. (2006). Developing a tagset for automated POS tagging in Arabic. *WSEAS transactions on computers, 5*(11), 2787-2792

Al-Shalabi, R., Kanaan, G., Al-Sarayreh, B., Khanfar, K., AIGhonmein, A., Talhouni, H., et al. Proper noun extracting algorithm for the Arabic language. In *Proceedings of International Conference on IT to Celebrate S. Charmonman's 72nd Birthday, Bangkok, 2009* (pp. 28.21–28.29)

Alkharashi, I. Person named entity generation and recognition for Arabic language. In *Proceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo, 2009* (pp. 205–208)

Attia, M. (2008a). *Handling Arabic Morphological and Syntactic Ambiguity within the LFG Framework with a View to Machine Translation*. University of Manchester,

Attia, M. (2008b). *Handling Arabic Morphological and Syntactic Ambiguity within the LFG Framework with a View to Machine Translation. PhD thesis,* University of Manchester,

Attia, M., Pecina, P., Toral, A., Tounsi, L., & Van Genabith, J. A lexical database for modern standard Arabic interoperable with a finite state morphological transducer. In C. Mahlow, & M. Piotrowski (Eds.), *Procedding of Second International Workshop, SFCM Systems and Frameworks for Computational Morphology, Zurich, Switzerland,, August 26, 2011 2011* (pp. 98-118): Springer

Attia, M., Toral, A., Tounsi, L., Monachini, M., & Genabith, J. v. An automatically built named entity lexicon for Arabic. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010), Valletta, 2010* (pp. 3,614–613,621)

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

Beesley, K. R. Arabic morphology using only finite-state operations. In *Proceedings of the Workshop on Computational Approaches to Semitic languages, 1998* (pp. 50-57): Association for Computational Linguistics

Beesley, K. R., & Karttunen, L. Finite-state non-concatenative morphotactics. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, 2000* (pp. 191-198): Association for Computational Linguistics

Beesley, K. R., & Karttunen, L. (2003). Finite-state morphology: Xerox tools and techniques. *CSLI, Stanford*.

Ben Hamadou, A., Odile, P., & Héla, F. (2010). Multilingual extraction of functional relations between Arabic named entities using NooJ platform. *In HAL Archives at http://hal.archives-ouvertes.fr/*, 1-10, doi:hal-00547940.

Benajiba, Y. (2009). *Arabic Named Entity Recognition. PhD thesis,* Universidad Politécnica de Valencia, Valencia, Spain.

Benajiba, Y., Diab, M., & Paolo Rosso, M. Arabic named entity recognition using optimized feature sets. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008), Stroudsburg, PA., 2008a* (pp. 284–293)

Benajiba, Y., Diab, M., & Paolo Rosso, M. Arabic named entity recognition: An SVM-based approach. In *Proceedings of Arab International Conference on Information Technology (ACIT 2008), Hammamet, 2008b* (pp. 16–18)

Benajiba, Y., Diab, M., & Rosso, P. (2009a). Arabic named entity recognition: A feature-driven study. *IEEE Transactions on Audio, Speech, and Language Processing, 17*(5), 926–934.

Benajiba, Y., Diab, M., & Rosso, P. (2009b). Using Language Independent and Language Specific Features to Enhance Arabic Named Entity Recognition. *The International Arab Journal of Information Technology, 6*(5).

Benajiba, Y., & Rosso, P. ANERsys 2.0: Conquering the NER task for the Arabic language by combining the maximum entropy with POS-tag information. In *Proceedings of Workshop on Natural Language-Independent Engineering, 3rd Indian International Conference on Artificial Intelligence (IICAI-2007), Mumbay, 2007* (pp. 1,814–811,823)

Benajiba, Y., & Rosso, P. Arabic named entity recognition using conditional random fields. In *Proceedings of the Workshop on HLT & NLP within the Sixth International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, 2008* (pp. 143–153)

Benajiba, Y., Rosso, P., & Ruiz, J. M. B. ANERsys : An Arabic named entity recognition system based on maximum entropy. In A. Gelbukh (Ed.), *Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2007), Berlin, 2007* (pp. 143–153): Springer-Verlag Berlin Heidelberg

Benajiba, Y., Zitouni, I., Diab, M., & Rosso, P. Arabic named entity recognition: Using features extracted from noisy data. In *Proceedings of the ACL 2010 Conference Short Papers, ACLShort 2010, Stroudsburg, PA., 2010* (pp. 281–285)

Bidhend, M., BehrouzMinaei-Bidgoli, & Jouzi, H. Extracting person names from ancient Islamic Arabic texts. In *Proceedings of Language Resources and Evaluation for Religious Texts (LRE-Rel) Workshop Programme, Eight International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, 2012* (pp. 1–6)

Bollegala, D., Matsuo, Y., & Ishizuka, M. (2010). *Relational Duality: Unsupervised Extraction of Semantic Relations between Entities on the Web.* Paper presented at the 19th international conference on World wide web, WWW2010, Raleigh, North Carolina, USA, April 26-30

Bosch, A. v. d., Marsi, E., & Soudi, A. (2007). Memory-based Morphological Analysis and Part-of-speech Tagging of Arabic. In A. Soudi, A. v. d. Bosch, & G. Neumann (Eds.), *Arabic Computational Morphology: Knowledge-based and Empirical Methods* (Vol. 38, Text, Speech and Language Technology): Springer.

Boujelben, I. (2015). *Une méthode hybride d'extraction des relations entre les entités nommées : Application à la langue arabe*. Université de Sfax, Stax - Tunisie.

Boulaknadel, S. (2008). *Traitement Automatique des Langues et Recherche d'Information en langue arabe dans un domaine de spécialité: Apport des connaissances morphologiques et syntaxiques pour l'indexation*. *Thèse de Doctorat*, Université de Nantes,

Boulanger, J. C., & Cormier, M. C. (2001). *Le nom propre dans l'espace dictionnairique général* (Études de métalexicographie). Tübingen, Niemeyer.

Brun, C., & Ehrmann, M. (2010). *Un système de détection d'entités nommées adapté pour la campagne d'évaluation ESTER 2.* Paper presented at the La 17e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2010), Montréal (Canada),

Buckwalter, T. (2004a). Buckwalter Arabic Morphological Analyzer Version 2.0. *catalog number LDC2004L02*: LDC.

Buckwalter, T. Issues in Arabic orthography and morphology analysis. In *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, 2004b* (pp. 31-34): Association for Computational Linguistics

Buckwalter, T. (2007). Issues in Arabic Morphological Analysis. In A. Soudi, A. v. d. Bosch, & G. Neumann (Eds.), *Arabic computational morphology: Knowledge-based and Empirical Methods* (Vol. 38, pp. 23-41, Text, Speech and Language Technology): Springer.

Cahill, L. (2007). A Syllable-based Account of Arabic Morphology. In A. Soudi, A. v. d. Bosch, & G. Neumann (Eds.), *Arabic computational morphology: Knowledge-based and Empirical Methods* (Vol. 38, pp. 45-66, Text, Speech and Language Technology): Springer.

Carrasco, R. C., Daciuk, J., & Forcada, M. L. (2007). An implementation of deterministic tree automata minimization. In *Implementation and Application of Automata* (pp. 122-129): Springer.

Carrasco, R. C., Daciuk, J., & Forcada, M. L. (2009). Incremental construction of minimal tree automata. *Algorithmica, 55*(1), 95-110.

Cavali-Sforza, V., & Soudi, A. (2007). Arabic Computational Morphology : A trade-off Between Multiple Operations and Multiple Stems. In A. Soudi, A. v. d. Bosch, & G. Neumann (Eds.), *Arabic computational morphology: Knowledge-based and Empirical Methods* (Vol. 38, pp. 89-114, Text, Speech and Language Technology): Springer.

Christopher, D. M., & Hinrich, S. (2003). *Foundations of statistical natural language processing* (Vol. 6ème édition): Massachusetts Institute of Technology.

Chunju, Z., Xueying, Z., Wenming, J., Qijun, S., & Shanqi, Z. Rule-Based Extraction of Spatial Relations in Natural Language Text. In *proceedings de la conference internationale CiSE 2009, China, 2009* (pp. 1–4)

Clark, A. (2007). Supervised and Unsupervised Learning of Arabic Morphology. In A. Soudi, A. v. d. Bosch, & G. Neumann (Eds.), *Arabic computational morphology: Knowledge-based and Empirical Methods* (Vol. 38, pp. 181-200, Text, Speech and Language Technology): Springer.

Clément, L., Lang, B., & Sagot, B. Morphology based automatic acquisition of large-coverage lexica. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC04), Lisbon, Portugal, 2004* (pp. 1841-1844)

CNRTL-CNRS (2012). Outils et ressources pour un traitement optimisé de la langue. http://www.cnrtl.fr/definition/nounation. Accessed Fevrier 2017.

Cohen, W. W. Learning trees and rules with set-valued features. In *Proceedings of the 13th National Conference on Artificial Intelligence (AAAI 96) / 8th Conference on Innovative Applications of Artificial Intelligence (IAAI 96), Portland, Aug 04-08 1996* (Vol. 1 and 2, pp. 709-716)

Courtois, B. (1994-1995). *Buts et méthodes de l'élaboration des dictionnaires électroniques du LADL* (Vol. Théories et pratiques du lexique, Cahier du CIEL). Université Paris 7 Denis Diderot: Centre Interlangue d'Études en Lexicologie.

Courtois, B., & Silberztein, M. (1990). Dictionnaires électroniques du français. *Langue française, 87*(1), 3-4.

Daciuk, J. (1998). *Incremental construction of finite-state automata and transducers, and their use in the natural language processing*. PhD thesis, Technical University of Gdańsk,

Daciuk, J. (2003). Comparison of construction algorithms for minimal, acyclic, deterministic, finite-state automata from sets of strings. In *Implementation and Application of Automata* (pp. 255-261): Springer.

Daciuk, J., Mihov, S., Watson, B. W., & Watson, R. E. (2000). Incremental construction of minimal acyclic finite-state automata. *Computational Linguistics, 26*(1), 3-16, doi:Doi 10.1162/089120100561601.

Daciuk, J., Watson, B. W., & Watson, R. E. Incremental construction of minimal acyclic finite state automata and transducers. In *Proceedings of the International Workshop on Finite State Methods in Natural Language Processing, 1998* (pp. 48-56): Association for Computational Linguistics

Dada, A., & Ranta, A. (2006). Implementing an open source Arabic resource grammar in GF*. In M. A. Mughazy (Ed.), *Perspectives on Arabic linguistics XX, Papers from the twentieth annual symposium on Arabic linguistics* (Vol. 290, pp. 209-232, Amsterdam studies in the theory and history of linguistic science. Series IV, Current issues in

linguistic theory). Amsterdam, The Netherlands & Philadelphia, USA: John Benjamins Publishing Company.
Daille, B., Fourour, N., & Morin, E. (2000). Catégorisation des noms propres : une étude en corpus. *Cahiers de Grammaire, 25*(Sémantique et Corpus), 115-129.
Darwish, K., & Oard, D. W. (2007). Adapting Morphology for Arabic Information Retrieval. In A. Soudi, A. v. d. Bosch, & G. Neumann (Eds.), *Arabic computational morphology: Knowledge-based and Empirical Methods* (Vol. 38, pp. 245-262, Text, Speech and Language Technology): Springer.
Debili, F., Ben Tahar, Z., & Souissi, E. Analyse automatique vs analyse interactive: un cercle vertueux pour la voyellation, l'étiquetage et la lemmatisation de l'arabe. In *Proceedings of the 14ème Conférence TALN & 11ème Rencontre RECITAL, Toulouse, 2007* (pp. 347-356): IRIT Press
Diab, M., Hacioglu, K., & Jurafsky, D. (2007). Automatic Processing of Modern Standard Arabic Text. In A. Soudi, A. v. d. Bosch, & G. Neumann (Eds.), *Arabic computational morphology: Knowledge-based and Empirical Methods* (Vol. 38, pp. 159-179, Text, Speech and Language Technology): Springer.
Dichy, J., & Farghaly, A. (2007). Grammar-Lexis Relations in the Computational Morphology of Arabic. In A. Soudi, A. v. d. Bosch, & G. Neumann (Eds.), *Arabic computational morphology: Knowledge-based and Empirical Methods* (Vol. 38, pp. 115-140, Text, Speech and Language Technology): Springer.
Dichy, J., & Farghaly, A. A. S. (2003). *Roots & Patterns vs. Stems plus Grammar-Lexis Specifications : on what basis should a multilingual lexical database centered on Arabic be built ?* Paper presented at the MT-Summit IX workshop on machine translation for Semitic languages, New Orleans, USA,
Doumi, N., Lehireche, A., Maurel, D., & Abdelali, A. (2016a). A Semi-Automatic and Low Cost Approach to Build Scalable Lemma-based Lexical Resources for Arabic Verbs. *International Journal of Information Technology and Computer Science(IJITCS), 8*(2), 1-13, doi:DOI: 10.5815/ijitcs.2016.02.01.
Doumi, N., Lehireche, A., Maurel, D., & Ali Cherif, M. (2013). *La conception d'un jeu de ressources libres pour le TAL arabe sous Unitex.* Paper presented at the TRADETAL2013, Colloque international en Traductologie et TAL, Oran - Algeria, 5-6 may
Doumi, N., Lehireche, A., Maurel, D., & Bouziane, A. (2016b). *FSM-based Resources and Tools for Modern Standard Arabic Processing.* Paper presented at the The 6th. world Congress on Electrical Engineering, Computer Science and Information Technology (WCECIT2016), Barcelona, Spain, September 8-10
Dreyer, M., & Eisner, J. Discovering morphological paradigms from plain text using a Dirichlet process mixture model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011* (pp. 616-627): Association for Computational Linguistics
Dufournaud, N., Demonet, M.-L., Uetani, T., Vincent, T., Le Rolle, V., Bontemps, L., et al. (2012). Manuel d'encodage TEI - Renaissance et temps modernes. http://www.bvh.univ-tours.fr/XML-TEI/ManuelWeb/co/0Manuel%20Renaissance_Book.html. Accessed Aout 2016.
Durrett, G., & DeNero, J. Supervised Learning of Complete Morphological Paradigms. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACLHLT), Atlanta, GA, 2013* (pp. 1185-1195)
Ehrmann, M. (2008). *Les entités nommées, de la linguistique au TAL: Statut théorique et méthodes de désambiguïsation. Thèse de Doctorat,* Université Paris 7 - Denis Diderot,

El-Kateb, S., Black, W., Vossen, P., Farwell, D., Pease, A., & Fellbaum, C. Arabic WordNet and the challenges of Arabic. In *Proceedings of Arabic NLP/MT Conference, London, 2006* (pp. 15–24)

El-Maarouf, I. (2011). *Formalisation de connaissances à partir de corpus: Modélisation linguistique du contexte pour l'extraction automatique de relations sémantiques*. *Thèse de Doctorat*, Université de Bretagne Sud,

Embarek, M., & Ferret, O. Une expérience d'extraction de relations sémantiques à partir de textes dans le domaine médical. In *Proceedings de la 14ème conférence sur le Traitement Automatique des Langues Naturelles TALN2007, Toulouse, 2007* (pp. 37–46)

ESTER2 (2007). Entités Nommées, Dates, heures et montants. *Convention d'annotation, version 0.1*.

Ezzat, M. Acquisition de grammaire locale pour l'extraction de relations entre entités nommées. In *Les actes de la 14ème conférence sur le Traitement Automatique des Langues Naturelles TALN2010, Montréal, 2010* (pp. 19–23)

Farber, B., Freitag, D., Habash, N., & Rambow, O. (2008). *Improving NER in Arabic Using a Morphological Tagger*. Paper presented at the International Conference on Language Resources and Evaluation (LREC08), Marrakech, Morocco,

Fellbaum, C. Wordnet and wordnets. In K. Brown (Ed.), *Encyclopedia of Language and Linguistics, Oxford, 2005* (pp. 665–670): Elsevier

Fradin, B. (2010). Recherches actuelles en morphologie. In J. Stichauer (Ed.), *La forme et le sens. Actes de l'école doctorale de Podèbrady, Février 2006* (pp. 151-166). Prague, République tchèque: Univerzita Karlova v Praze Filozofická fakulta / Asociace Gallica.

Friburger, N. (2002). *Reconnaissance automatique des noms propres : application à la classification automatique de textes journalistiques*. Thèse de doctorat, Université François Rabelais, Tours.

Gadalla, H. A. H. (2000). *Comparative morphology of standard and Egyptian Arabic* (LINCOM studies in Afro-Asiatic linguistics, Vol. 05). Muenchen: Lincom Europa.

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

Grishman, R., Huttunen, S., & Yangarber, R. (2002). Information Extraction for Enhanced Access to Disease Outbreak Reports. *Biomedical Informatics, 35*(4), 23–246.

Grouin, C., Galibert, O., Rosset, S., Quintard, L., & Zweigenbaum, P. (2011). *Mesures d'évaluation pour entités nommées structurées*. Paper presented at the QDC 2011/EvalECD 2011, Brest, France, 25 janvier

Guessoum, A., & Zantout, R. (2007). Arabic Morphological Generation and its Impact on the Quality of Machine Translation to Arabic. In A. Soudi, A. v. d. Bosch, & G. Neumann (Eds.), *Arabic computational morphology: Knowledge-based and Empirical Methods* (Vol. 38, pp. 287-302, Text, Speech and Language Technology): Springer.

Guo, J., Xu, G., Cheng, X., & Li, H. Named entity recognition in query. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009), New York City, 2009* (pp. 267–274)

Habash, N. (2007). Arabic morphological representations for machine translation. In A. Soudi, A. v. d. Bosch, & G. Neumann (Eds.), *Arabic computational morphology: Knowledge-based and Empirical Methods* (pp. 263-285, Text, Speech and Language Technology): Springer.

Habash, N. (2010). *Introduction to Arabic natural language processing* (Synthesis Lectures on Human Language Technologies): Morgan & Claypool.

Habash, N., & Rambow, O. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Proceedings of the 43rd Annual Meeting of ACL, Ann Arbor, Michigan, 2005* (pp. 573-580)

Habash, N. Y., Soudi, A., & Buckwalter, T. (2007). On Arabic Transliteration. In A. Soudi, A. v. d. Bosch, & G. Neumann (Eds.), *Arabic computational morphology: Knowledge-based and Empirical Methods* (Vol. 38, pp. 15-22, Text, Speech and Language Technology): Springer.

Hamadou, A. B., Piton, O., & Fehri, H. (2010). Multilingual Extraction of functional relations between Arabic Named Entities using NooJ platform. *HAL*, doi:hal-00547940.

Hamlaoui, A. (2007). شَذَا العُرْف في فَنِ الصَرْف */šaðaA Alçur.f fy fani AlSar.f/* (1st ed.). Beirut, Lebanon: Resalah Publishers.

Internet World Users By Language : Top 10 Languages (2015). http://www.internetworldstats.com/stats7.htm. Accessed Feb. 2016.

Jonasson, K. (1994). *Le nom propre. Constructions et interprétations.* (Champs Linguistiques). Louvain-la-Neuve: Editions Duculot.

Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics* (2nd Edition ed.): Prentice-Hall.

Kenneth, R. B. (2001). *Finite-State Morphological Analysis and Generation of Arabic at Xerox Research: Status and Plans.* Paper presented at the EACL2001 workshop on Arabic Language Processing: Status and Prospects, Toulouse, France,

Kevers, L. L'information biographique : modélisation, extraction et organisation en base de connaissances. In *Proceedings de la conférence internationale RECITAL 2006, Leuven, 2006* (pp. 680–689)

Khalid, M. A., Jijkoun, V., & De Rijke, M. The Impact of Named Entity Normalization on Information Retrieval for Question Answering. In *Advances in Information Retrieval, Berlin Heidelberg, 2008*: Springer

Khoja, S. APT: Arabic Part-of-Speech Tagger. In *Proceedings of the Student Workshop at the 2nd Meeting of the NAACL, (NAACL'01), Carnegie Mellon University, Pennsylvania, June 2001* (pp. 20-25)

Khoja, S., Garside, R., & Knowles, G. A tagset for the morphosyntactic tagging of Arabic. In *Proceedings of Corpus Linguistics, Lancaster, UK, 2001*

Koehn, P. (2004). Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In R. E. Frederking, & K. B. Taylor (Eds.), *Machine Translation: From Real Users to Research, Proceedings* (Vol. 3265, pp. 115-124, Lecture Notes in Computer Science).

Koulali, R., Meziane, & Abdelouafi. A contribution to Arabic named entity recognition. In *Proceedings of the 10th International Conference on ICT and Knowledge Engineering, Morocco, 2012* (pp. 46–52)

Kouloughli, D. E. (1994). *Grammaire de l'arabe d'aujourd'hui*: Pocket.

Krause, S., Xu, F., Uszkoreit, H., & Leser, U. (2012). Relation Extraction with Massive Seed and Large Corpora.

Larkey, L. S., Ballesteros, L., & Connell, M. E. (2007). Light Stemming for Arabic Information Retrieval. In A. Soudi, A. v. d. Bosch, & G. Neumann (Eds.), *Arabic computational morphology: Knowledge-based and Empirical Methods* (Vol. 38, pp. 221-243, Text, Speech and Language Technology): Springer.

Maâloul, M. H. (2012). *Approche hybride pour le résumé automatique de textes: Application à la langue arabe. Thèse de Doctorat*, Université Aix-Marseille,

Maaluf, L. (1991). المُنْجِد في اللُّغَة و الاعلام */Almun.jid fy All~uγaħ w AlAçlAm/ (31st edition)*. Beirut, Lebanon: Dar El-Machreq SARL Publishers.

Manning, C., & Schütze, H. (2003). *Foundations of Statistical Natural Language Processing* (6ed.). Cambridge, MA: MIT Press.

Maurel, D., & Friburger, N. (2013). *CasSys : Un système libre de cascade de transducteurs.* Paper presented at the TALN-RECITAL, Les Sables d'Olonne, 17-21 Juin

Maurel, D., Friburger, N., Antoine, J.-Y., Eshkol-Taravella, I., & Nouvel, D. (2011). Cascades de transducteurs autour de la reconnaissance des entités nommées. *Traitement automatique des langues, 52*(1), 27.

Maurel, D., Friburger, N., & Eshkol-Taravella, I. (2013). Recherche d'entités nommées dans des textes de la Renaissance. *Traitement automatique des langues, 54*(2), 24.

Maurel, D., & Guenthner, F. (2005). *Automata and dictionaries* (Vol. 6, Texts in computing). London: King's college.

Mesfar, S. (2008). *Analyse morpho-syntaxique automatique et reconnaissance des entités nommées en arabe standard*. PhD thesis, Université de Franche-Comté,

Mihov, S. (1999). Direct construction of minimal acyclic finite states automata. *Annuaire de l'Universite de Sofia St. Kl. Ohridski, Faculté de mathématiques et Informatique, 92*(2).

Mohit, B., Schneider, N., Bhowmick, R., Oflazer, K., & Smith, N. Recall-oriented learning of named entities in Arabic wikipedia. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL) 2012, Stroudsburg, PA., 2012* (pp. 162–173)

Mohri, M. Compact representations by finite-state transducers. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics, 1994* (pp. 204-209): Association for Computational Linguistics

Mohri, M. (1996). On some applications of finite-state automata theory to natural language processing. *Natural Language Engineering, 2*(1), 61-80.

Mohri, M. (1997). Finite-state transducers in language and speech processing. *Computational Linguistics, 23*(2), 269-311.

Mohri, M., Pereira, F., & Riley, M. (2002a). Weighted finite-state transducers in speech recognition. *Computer Speech & Language, 16*(1), 69-88.

Mohri, M., Pereira, F., & Riley, M. (2005). Weighted automata in text and speech processing. *arXiv preprint cs/0503077*.

Mohri, M., Pereira, F., & Riley, M. (2008). Speech recognition with weighted finite-state transducers. In *Springer Handbook of Speech Processing* (pp. 559-584): Springer.

Mohri, M., Pereira, F. C. N., & Riley, M. D. (2002b). Systems and methods for determinization and minimization a finite state transducer for speech recognition. *Acoustical Society of America Journal, 111*(1), 21-21.

Moore, E. F. (1956). Gedanken-experiments on sequential machines. *Automata studies, 34*, 129-153.

Müller, S. (2015). *Grammatical theory: From transformational grammar to constraint-based approaches* (Textbooks in Language Sciences). Berlin, Germany: Language Science Press.

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Journal of Linguisticae Investigationes, 30*(1), 3-26.

Nakamura-Delloye, Y. (2011). *Extraction non-supervisée de relations basée sur la dualité de la représentation.* Paper presented at the TALN 2011 : Traitement Automatique des Langues Naturelles, Montpellier, France, Juin 2011

Neme, A. A. A lexicon of Arabic verbs constructed on the basis of Semitic taxonomy and using finite-state transducers. In *Proceedings of the International Workshop on Lexical Resources, Slovenia, 2011* (pp. 78-85)

Nouvel, D. (2012). *Reconnaissance des entités nommées par exploration de règles d'annotation: Interpréter les marqueurs d'annotation comme instructions de structuration locale*. *Thèse de Doctorat,* Université François Rabelais de Tours,

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics, 29*(1), c-51.

Oudah, M. M., & Shaalan, K. A Pipeline Arabic Named Entity Recognition Using a Hybrid Approach In *Proceedings of COLING 2012, Mumbai, December 2012* (pp. 2159–2176)

Paumier, S. (2014). *Unitex manual for version 3.1*. http://www-igm.univ-mlv.fr/~unitex/UnitexManual3.1.pdf: Institut Gaspard-Monge (IGM), University of Paris-Est Marne-la-Vallée, France.

Poibeau, T. (2003). *Extraction Automatique d'Information* (Du texte brut au web sémantique): Hermès.

Quaero (2013). Quaero en bref. http://www.quaero.org/quaero-en-bref/. Accessed Aout 2016.

Refaat, K. S., & Madkour, A. An Optimized Method for Arabic Cross-Document Named Entity Normalization. In *Proceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo, 2009* (pp. 209-212)

Rosset, S., Grouin, C., & Zweigenbaum, P. (2011). Entitées nommées structurées : guide d'annotation Quaero. (pp. 82): LIMSI.

Ryding, K. C. (2005). *A Reference Grammar of Modern Standard Arabic*. New York, United States of America: Cambridge University Press.

Saleh, I. M. H., & Habash, N. (2009). *Automatic extraction of lemma-based bilingual dictionaries for morphologically rich languages.* Paper presented at the Third Workshop on Computational Approaches to Arabic Script-based Languages at the MT Summit XII, Ottawa, Canada,

Santos, D., & Baptista, N. M. J. Extraction of Family Relations between Entities. In *INForum 2010, 2010a* (pp. 54–560)

Santos, D., & Baptista, N. M. J. Extraction of Family Relations between Entities. In L. S. Barbosa, & M. P. Correia (Eds.), *INForum 2010, 2010b* (pp. 549–560)

Sawalha, M. S. S. (2011). *Open-source resources and standards for Arabic word structure analysis: Fine grained morphological analysis of Arabic text corpora*. PhD thesis, School of computing, University of Leeds, UK.

Sawalha, M. S. S., & Atwell, E. S. تـوظيف قـواعد الـنحو والـصرف فـي بـنـاء مـحلل صرفـي لـلـغة الـعربـيـة /twĎyf qwAçd AlnHw wAlSrf fy bnA' mHll Srfy llγħ Alçrbyħ/ (Adapting Language Grammar Rules for Building Morphological Analyzer for Arabic Language). In *Proceedings of the ALECSO Arab League Educational Cultural and Scientific Organization workshop on Arabic morphological analysis, Damascus, Syria, 26 April 2009*: King Abdul-Aziz City of Technology (KACT) and Arabic Language Academy

Serrano, L. (2011). Modélisation d'une ontologie de domaine et des outils d'extraction de l'information associés pour l'anglais et le français. *dumas-00569002*.

Shaalan, K. (2014). A Survey of Arabic Named Entity Recognition and Classification. *Computational Linguistics, 40*(2), 469-510, doi:10.1162/COLI_a_00178.

Shaalan, K., & Raza, H. Person name entity recognition for Arabic. In *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, Stroudsburg, PA., 2007* (pp. 17–24)

Shaalan, K., & Raza, H. (2009). NERA: Named entity recognition for Arabic. *Journal of the American Society for Information Science and Technology, 60*(8), 652–663.

Shihadeh, C., & Neumann, G. ARNE: A tool for named entity recognition from Arabic text. In *Fourth Workshop on Computational Approaches to Arabic Script-based*

Languages (CAASL4), located at the Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA), San Diego, CA., 2012 (pp. 24–31)

Silberztein, M. NooJ: an oriented object approach. In J. S. Royauté, M. (Ed.), *Proceedings of the INTEX pour la Linguistique et le Traitement Automatique des Langues, Actes des 4èmes et 5èmes journées INTEX, Bordeaux, mai 2001 et Marseille, mai 2002, Besançon, 2004*: Presses universitaires de Franche-Comté

Smrz, O. ElixirFM - Implementation of Functional Arabic Morphology. In *Proceedings of the Computational Approaches to Semitic Languages: Common Issues and Resources (ACL2007), Prague, Czech Republic, 2007* (pp. 1-8)

Smrz, O. (2007). *Functional Arabic Morphology : Formal System and Implementation*. *PhD thesis*, Charles university, Prague.

Soricut, R., & Marcu, D. Sentence Level Discourse Parsing using Syntactic and Lexical Information. In *Les actes de la conference internationale North American Chapter of the Association for Computational Linguistics on Human Language Technology, 2003*

Soudi, A., Neumann, G., & Bosch, A. v. d. (2007). Arabic computational morphology: Knowledge-based and Empirical Methods. In A. Soudi, A. v. d. Bosch, & G. Neumann (Eds.), *Arabic computational morphology: Knowledge-based and Empirical Methods* (Vol. 38, pp. 3-14, Text, Speech and Language Technology): Springer.

Sowa, J. F., & Majumdar Kyndi, A. K. (2015). *Natural Language Understanding.* Paper presented at the Data Analytics Summit II, Harrisburg University, USA, 14 December

Tolone, E. (2011). *Analyse syntaxique à l'aide des tables du Lexique-Grammaire du français*. *Thèse de Doctorat*, Université Paris-Est,

Tounsi, L., Bouchou, B., & Maurel, D. A compression method for natural language automata. In *Proceeding of the 2009 conference on Finite-State Methods and Natural Language Processing: Post-proceedings of the 7th International Workshop FSMNLP, 2009* (pp. 146-157)

Toussaint, Y. (2004). Extraction de connaissances à partir de textes structurés. *Document numérique, Vol. 8*(2004/3), 11-34, doi:10.3166/dn.8.3.11-34.

Watson, B. W. (2001). A taxonomy of algorithms for constructing minimal acyclic deterministic finite automata. *South African Computer Journal*(27), 12-17.

Watson, B. W., & Daciuk, J. (2003). An efficient incremental DFA minimization algorithm. *Natural Language Engineering, 9*(1), 49-64.

Zaghouani, W. (2009). *Le repérage automatique des entités nommées dans la langue arabe : vers la création d'un système à base de règles*. Université de Montréal, Montréal, Canada.

Zaghouani, W. (2014). *Critical Survey of the Freely Available Arabic Corpora*. Paper presented at the Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools - LREC2014, Reykjavik, Iceland, 27 May

Zaghouani, W., Bruno Pouliquen,, & Mohamed Ebrahim, a. R. S. Adapting a resource-light highly multilingual named entity recognition system to Arabic. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010), Valletta, 2010* (pp. 563–567)

Zanned, L. (2004). Root formation and polysemic organization in Arabic lexicon : a probabilistic model. In M. T. Alhawary, & E. Benmamoun (Eds.), *Perspectives on Arabic Linguistics XVII–XVIII, Papers from the Seventeenth and Eighteenth Annual Symposia on Arabic Linguistics* (Vol. 264, pp. 85-116, Amsterdam studies in the theory and history of linguistic science. Series IV, Current issues in linguistic theory). Amsterdam, The Netherlands & Philadelphia, USA: John Benjamins Publishing Company.

Annexe A

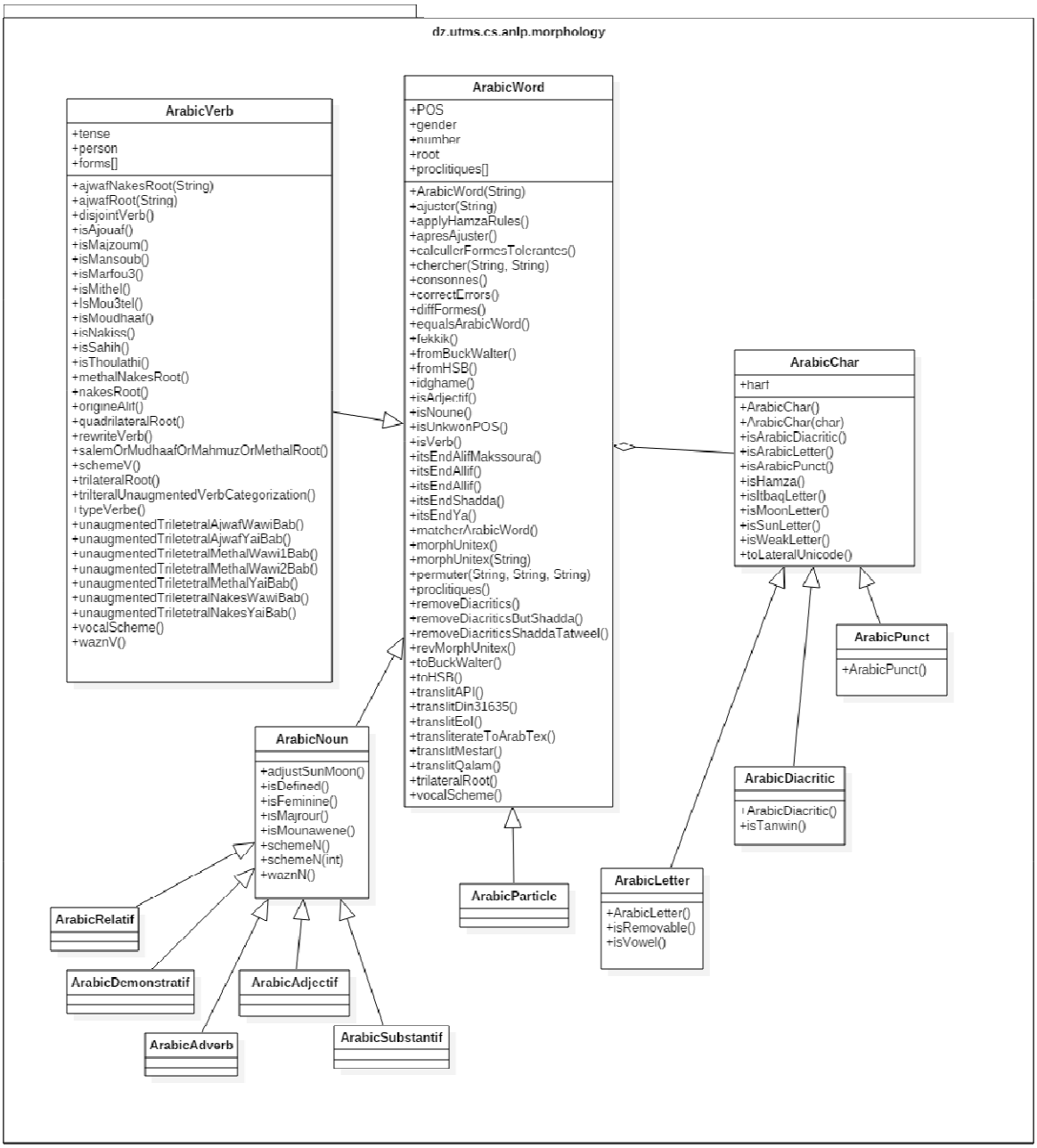


Figure 54 : Package des classes java de la morphologie arabe utilisées pour réaliser les différents traitements nécessaires à la construction des ressources linguistiques du module arabe de la plateforme Unitex/GramLab

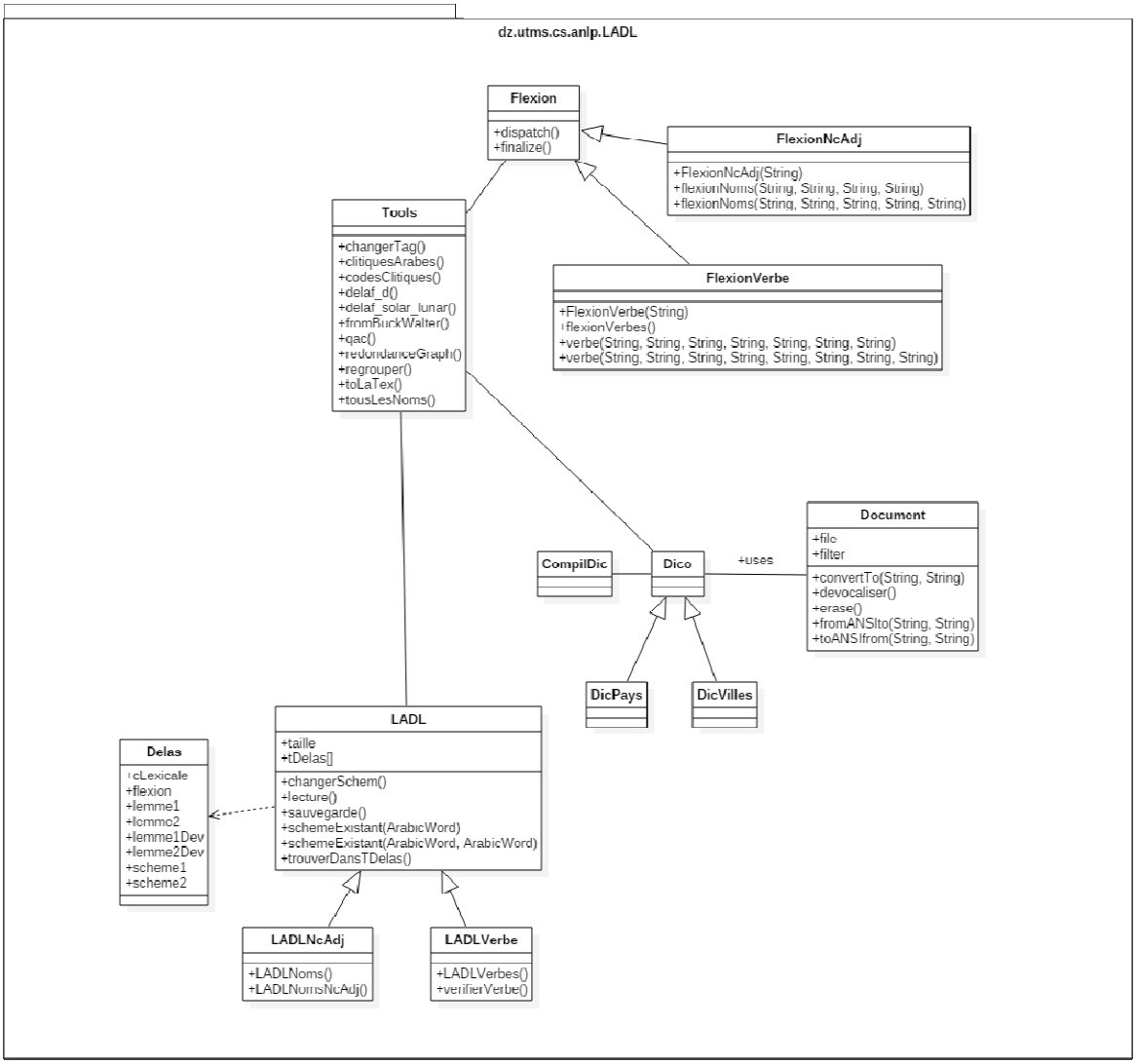


Figure 55 : Package java pour la construction des différents dictionnaires électroniques pour l'accomplissent des tâches de reconnaissance des entités nommées arabes et les relations les reliant

Annexe B

Le jeu d'étiquètes utilisé dans le module arabe d'Unitex/GramLab

Code	English	French	Arabic
		Part-of-speech	
V	Verb	Verbe	فعل
Ve	State verb	Verbe d'état	أخوات كان
Nc	Noun	Nom commun	اسم موصوف
Np	Proper noun	Nom propre	اسم علم
Adj	Adjective	Adjectif	صفة
Adv	Adverb	Adverbe	حال، ظرف حال، ظرف مكان، ظرف زمان
Prsl	Personal pronoun	Pronom personnel	ضمير منفصل
Dmst	Demonstrative pronoun	Pronom démonstratif	اسم إشارة
Rltf	Relative pronoun	Pronom relatif	اسم موصول
Intrg	Interrogative particle	Article d'interrogation	حرف استفهام
CnjCrd	Coordinating conjunction	Conjonction de coordination	حرف عطف
Prps	Preposition	Préposition	حرف جر
Sbjc	Accusative particle	Particule de subjonctif	حرف نصب
Evdt	Verb-like particles	Marqueurs d'évidentialité	أخوات إن
Apcp	Apocopative particle	Particule de l'apocopé	حرف جزم
Rstr	Exceptive particle	Particule de restriction	أداة استثناء
Crbr	Emphasis particle	Marqueur de corroboration	حرف تأكيد
Anl	Annuler	Particule d'annulation	حرف نسخ
Apl	Vocative particle	Particule d'appel	حرف نداء
Ftr	Particle of futurity	Particule de futur	حرف استقبال
Cslt	Causative particle	Particule de causalité	حرف تعليل
Ngt	Negative particle	Particule de négation	حرف نفي
Srmn	Jurative particle	Particule de serment	حرف قسم
Attn	Attention particle	Particule d'attention	حرف تنبيه
Rpns	Answer particle	Particule de réponse	حرف جواب
Cndt	Apocopative answer particle	Particule de condition	حرف شرط
Inct	Incitement particle	Particule	حرف تحضيض

		d'incitation	
Explc	Explanation particle	Particule d'explication	حرف تفسير
		Syntax	
		Transitivity	
Tr	Transitive	Transitif	متعدي
Intr	Intransitive	Intransitif	لازم
		Morphological features	
		Mood/Case	
A	Subjunctive /Accusative	Subjonctif/Accusatif	منصوب/مفتوح
A		Nonnation d'accusatif	تنوين الفتحة
U	Indicative /Nominative	Indicatif/Nominatif	مرفوع/مضموم
U		Nonnation de nominatif	تنوين الضمة
I	Genitive	Génitif	مجرور
I		Nonnation de génitif	تنوين الكسرة
O	Jussive	Jussif (apocopé)	مجزوم
<E>			
		Definiteness	
R	Definite	Défini	معرف
N	Indefinite	Indéfini	نكرة
		Gender	
M	Masculine	Masculin	مذكر
F	Feminine	Féminin	مؤنث
		Number	
S	Singular	Singulier	مفرد
D	Dual	Duel	مثنى
P	Plural	Pluriel	جمع
		Person	
1	First person	1ère personne	متكلم
2	Second person	2ème personne	مخاطب
3	Third person	3ème personne	غائب
		Tense	
A	Past	Accompli	ماضي
I	Present	Inaccompli	مضارع
F	Future	Futur	مستقبل
P	Imperative	Impératif	أمر
		Emphasize	
N	Emphatic verb	Emphatique	مؤكد
<E>	Non-emphatic verb	Non emphatique	غير مؤكد
		Construction	
C		Construction	مضاف
<E>		Pas de construction	غير مضاف
		Voice	
C	Active voice	Active	مبني للمعلوم
V	Passive voice	Passive	مبني للمجهول

Semantic categories			
Hum	Human	Humain	إنسان
Vill	Town	Ville ou village	مدينة أو قرية
Cont	Continent	Continent ou partie de continent	قارة، شبه قارة أو جزء من قارة
Dept	State	Département, état, fédération ou willaya	مقاطعة، فدرالية أو ولاية
Pay	Country	Pays	دولة

La structure de l'étiquette contenant les traits morphologiques, syntaxiques et sémantiques et respectant le formalisme DELA de LADL

Code	Catégorie	Ordre des traits flexionnels	Exemple
V	Verbe	Temps (aspect), Personne, Genre, Nombre, Mode, Emphase, Voix (forme)	**V:A1fsc** signifie verbe conjugué en accompli (passé) avec la 1ère personne du féminin singulier en voix (forme) active. **V:I1fsuc** signifie verbe conjugué en inaccompli (présent) avec la 1ère personne du féminin singulier en mode indicatif et la voix (forme) active. **V:F1fsv** signifie verbe conjugué en futur avec la 1ère personne du féminin singulier en voix (forme) passive. **V:F1msnv** signifie verbe conjugué en futur avec la 1ère personne du masculin singulier emphatique (énergétique) en voix (forme) passive.
Ve	Verbe d'état	Temps (aspect), Personne, Genre, Nombre	**Ve:A1ms** signifie verbe d'état conjugué en accompli (passé) avec la 1ère personne du masculin singulier.
Nc	Nom commun	Genre, Nombre, Cas, Définitude, Construction	**Nc:msunc** signifie nom masculin singulier nominatif non défini et en construction **Nc: msir** signifie nom masculin singulier génitif et défini
Np	Nom propre	Genre, Nombre	**Np+Hum:ms** signifie nom propre d'humain masculin singulier
Adj	Adjectif	Genre, Nombre, Cas, Définitude	**Adj: msun** signifie adjectif masculin singulier nominatif et non défini

Adv	Adverbe		
Prsl	Pronom personnel	Genre, Nombre, Personne	**Prsl: ms1** signifie pronom personnel masculin singulier pour la 1ère personne
Dmst	Pronom démonstratif	Genre, Nombre	**Dsmt: ms** signifie pronom démonstratif masculin singulier
Rltf	Pronom relatif	Genre, Nombre	**Rltf:ms** signifie pronom relatif masculin singulier
Intrg	Article d'interrogation		
CnjCrd	Conjonction de coordination		
Prps	Préposition		
Sbjc	Particule de subjonctif		
Evdt	Marqueur d'évidentialités		
Apcp	Particule de l'apocopé		
Rstr	Particule de restriction		
Crbr	Marqueur de corroboration		
Anl	Particule d'annulation		
Apl	Particule d'appel		
Ftr	Particule de futur		
Cslt	Particule de causalité		
Ngt	Particule de négation		
Srmn	Particule de serment		
Attn	Particule d'attention		

Rpns	Particule de réponse
Cndt	Particule de condition
Inct	Particule d'incitation
Explc	Particule d'explication

Annexe C

La translitération HSB (Habash-Soudi-Buckwalter)

N°	Unicode	arabe	HSB	API
1	0x0621	ء	ʼ	ʔ
2	0x0622	آ	Ā	ʔaː
3	0x0623	أ	Â	
4	0x0624	ؤ	ŵ	
5	0x0625	إ	Ă	
6	0x0626	ئ	ŷ	
7	0x0627	ا	A	aː
8	0x0628	ب	b	b
9	0x0629	ة	ħ	a, at
10	0x062A	ت	t	t
11	0x062B	ث	θ	θ
12	0x062C	ج	j	d͡ʒ, g, ʒ
13	0x062D	ح	H	ħ
14	0x062E	خ	x	x
15	0x062F	د	d	d
16	0x0630	ذ	ð	ð
17	0x0631	ر	r	r
18	0x0632	ز	z	z
19	0x0633	س	s	s
20	0x0634	ش	š	ʃ
21	0x0635	ص	S	sˤ
22	0x0636	ض	D	dˤ
23	0x0637	ط	T	tˤ

24	0x0638	ظ	ď	ðˤ, zˤ
25	0x0639	ع	ʕ	ʕ
26	0x063A	غ	ɣ	ɣ
27	0x0640	ـ	_	
28	0x0641	ف	f	f
29	0x0642	ق	q	q
30	0x0643	ك	k	k
31	0x0644	ل	l	l
32	0x0645	م	m	m
33	0x0646	ن	n	n
34	0x0647	ه	h	h
35	0x0648	و	w	w, u:
36	0x0649	ى	ý	a:
37	0x064A	ي	y	j, i:
38	0x064B	ً	ã	
39	0x064C	ٌ	ũ	
40	0x064D	ٍ	Ĩ	
41	0x064E	َ	a	
42	0x064F	ُ	u	
43	0x0650	ِ	I	
44	0x0651	ْ	.	
45	0x0652	ّ	~	