Université de Montréal

Le repérage automatique des entités nommées dans la langue arabe : vers la création d'un système à base de règles

Par

Wajdi Zaghouani

Département de linguistique et de traduction

Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures

en vue de l'obtention du grade de M.A.

en linguistique

Mars 2009

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé :

Le repérage automatique des entités nommées dans la langue arabe :
vers la création d'un système à base de règles

Présenté par :
Wajdi Zaghouani

a été évalué par un jury composé des personnes suivantes :

Marie-Claude L'homme, Président-rapporteur
Patrick Drouin, Directeur de recherche
Richard Kittredge, Co-directeur
Nie, Jian-Yun, Membre du jury

RÉSUMÉ

Notre travail relève du domaine de traitement automatique de la langue en général et de l'extraction automatique des entités nommées en particulier. Il a été réalisé dans le cadre d'un stage à Ispra en Italie et au sein du Centre commun de recherche (CCR).

On se propose d'étudier dans ce mémoire la question des entités nommées dans le contexte particulier de la langue arabe dans le but de construire un système d'extraction automatique des entités nommées.

Il existe plusieurs approches pour construire un système d'extraction automatique des entités nommées. Nous présentons le fonctionnement des systèmes à base de règles ainsi que les systèmes à apprentissage automatique. Dans ce travail, nous avons opté pour un système à base de règles linguistiques, ce qui a nécessité une recherche sur les entités nommées en arabe ainsi que sur quelques traits linguistiques de la langue. Cette étude a servi lors de l'étape de l'implémentation du système que nous avons baptisé RENAR.

Nous avons évalué les performances du système sur deux corpus représentant deux variantes régionales de l'arabe moderne. Les résultats de l'évaluation ont montré l'importance des marqueurs lexicaux et des dictionnaires pour un système à base de règles. Ce travail se veut une contribution à l'avancement des recherches dans le domaine du repérage de l'information pour la langue arabe.

Mots-clés : extraction d'information, fouille de textes, extraction des entités nommées, noms propres, langue arabe, traitement automatique de la langue, système à base de règles, constitution de corpus, évaluation.

ABSTRACT

The present work falls within the general domain of Natural Language Processing and focuses in particular on the problem of information extraction involving named entities. The research being reported here was conducted as part of an internship at the Joint Research Center (JRC) in Ispra, Italy.

Our central concern in this study is the question of named entities in the Arabic language, and our goal is to build a system for automatic extraction of named entities from Arabic text. We discuss a variety approaches that have been developed for named entity extraction, both linguistic rule-based systems and statistical machine-learning systems.

For the project being reported here, we chose to build a rule-based system, which in turn led us into detailed research regarding the properties of named entities in Arabic, as well as relevant properties of the Arabic language in general. The system that has been implemented on the basis of this research is called RENAR.

We have evaluated the performance of RENAR using text corpora in Modern Standard Arabic that have been drawn from two distinct regions. The results of the evaluation show the importance of "cue words", and of having a suitable and adequate dictionary. We hope this work will serve to advance the state of the art in automatic information extraction for Arabic.

Keywords : information extraction, text mining, named entity extraction, proper names, Arabic language, Natural Language Processing (NLP), rule-based system, corpus development, evaluation.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ACRONYMES

ENAMEX :	*Entity Name Extraction*
CCR :	*Centre commun de recherche*
EMM :	*European media monitor*
EN :	*Entités nommées*
ENAMEX :	*Entity Name Extraction*
FREQ :	*Fréquence*
HMM :	*Hidden Markov Model (Modèles cachés de Markov)*
MUC :	*Message Understanding Conference*
REN :	*Repérage des entités nommées*
RENAR :	*Repérage des entités nommées arabe*
TAL :	*Traitement automatique de la langue*

CONVENTION D'ÉCRITURE DE L'ARABE

Dans le cas où un mot ou un exemple est écrit en caractères arabes, nous avons choisir de fournir une translittération et une traduction française ou une traduction littérale française (si nécessaire) et dans l'ordre suivant :

1. Les caractères arabes.
2. Translittération selon le système de translittération QALAM[1] (voir annexe I).
3. Traduction ou translittération française.

1 Il s'agit d'un système de translittération de l'arabe créé par Heddaya *et al.*(1985).

À Anissa, Adem et mes parents,

REMERCIEMENTS

Je voudrais exprimer ma gratitude et mes remerciements à l'ensemble des personnes qui ont participé de près ou de loin à ce travail, avec leurs conseils et leurs recommandations.

Je souhaite tout d'abord remercier sincèrement mon directeur de recherche Patrick Drouin pour sa grande disponibilité, sa patience et ses précieux conseils.

Mes remerciements les plus sincères vont également à mon codirecteur Richard Kittredge pour ses encouragements et sa lecture critique de ce manuscrit.

Je ne pourrais passer sous silence les encouragements constants de mon directeur de stage, Ralf Steinberger ainsi que de tous les membres du groupe EMM, particulièrement Bruno Pouliquen pour son apport technique.

Je voudrais aussi remercier mon collègue Tim Buckwalter du Linguistic Data Consortium pour sa générosité et le partage de son expérience ainsi que Mohammed Maamouri pour son encadrement et ses encouragements.

Un grand merci pour Abderahmane Najjar pour sa lecture et ses corrections ainsi que pour sa générosité sans faille.

Enfin, j'adresse mes plus sincères remerciements, à tous mes proches et amis et particulièrement à mes parents pour le soutien, et les encouragements qu'ils m'ont apportés.

Pour finir, un merci très particulier à toi, Anissa, tu m'as toujours apporté ton soutien au jour le jour, aussi bien dans les moments de joie que dans ceux du doute et de remise en question.

1. Introduction

L'avènement de l'Internet a révolutionné le domaine de l'édition numérique de l'information. Des milliers de sites Web apparaissent chaque jour sous forme de forums de discussions, de blogues[2] ou de sites d'information continue. Avec ce flux continu de textes électroniques, il est devenu crucial pour les utilisateurs de bénéficier d'un accès rapide à l'information demandée. Ceci a motivé, en particulier, la création de systèmes d'extraction d'information qui sont devenus de plus en plus nombreux depuis quelques années.

Toutefois, plusieurs défis se dressent devant ces systèmes avant qu'ils atteignent des performances optimales. En effet, il faut tenir compte du fait que les données textuelles contiennent souvent de l'information non structurée et que les constructions langagières sont en partie imprévisibles. Enfin, avec le nouveau contexte mondial et notamment l'ouverture du monde sur les autres cultures surtout grâce à l'Internet, il est devenu essentiel d'orienter les domaines de l'extraction d'information vers le multilinguisme et l'adaptation des outils existants pour de nouvelles langues ayant parfois moins de ressources linguistiques disponibles que d'autres et nécessitant parfois des changements majeurs dans l'approche à suivre.

Ce mémoire s'insère dans ce contexte, qui est l'extraction automatique d'information et plus particulièrement le repérage automatique des noms propres, des expressions temporelles et numériques en de langue arabe. Cet ensemble est historiquement désigné par le terme « entités nommées »[3]. Depuis quelques années, la recherche dans le domaine du repérage automatique des entités nommées (EN)

---

2 Il s'agit d'un journal de bord ou un journal personnel diffusé publiquement sur Internet.

3 Le concept d'entités nommées a été introduit lors de la conférence MUC (Message understanding conference) NIST (2001), pour englober trois types de noms propres (noms de personnes, noms de lieux, nom d'organisation) ainsi que les expressions temporelles et numériques.

n'arrête pas d'évoluer et nous avons vu naître deux grandes méthodes de repérage des EN, à savoir la méthode dite à base de règles et la méthode à apprentissage automatique.

L'avancement de la recherche dans ce domaine, était souvent orienté vers les langues à plus grande diffusion comme l'anglais ou le français ou encore le japonais et ce n'est que récemment que des systèmes de repérage d'EN supportant la langue arabe ont vu le jour.

Le sujet de ce mémoire a été motivé par un stage que j'ai effectué au sein du groupe Langtech-EMM[4] qui fait partie de la Commission européenne et plus précisément du Centre commun de recherche[5] (CCR) connu aussi sous l'acronyme anglais JRC (Joint research center). Le centre est localisé géographiquement dans le nord de l'Italie, plus précisément dans la région d'Ispra.

L'objectif principal du stage était la découverte des différentes techniques employées pour le traitement automatique du langage par le groupe Langtech-EMM. Cette expérience m'a permis de découvrir et de tester les outils et les ressources linguistiques du groupe Langtech-EMM.

À la fin de mon stage, le groupe Langtech-EMM m'a offert l'opportunité de réaliser un projet dans le cadre de mon mémoire de maîtrise.

---

4 <http://langtech.jrc.it/>, consulté le 6 janvier 2007.

5 <http://www.jrc.cec.eu.int/>, consulté le 6 janvier 2007.

Lors de ce mémoire, j'étais chargé d'adapter un module de repérage des EN à la langue arabe. Le module de repérage des entités nommées est une des composantes clés du système EMM[6] (Europe Media Monitor).

Il était impératif d'étudier les différentes approches et méthodes d'extraction automatique d'EN existantes avant de commencer l'intégration du module arabe dans le système EMM. Étant donné que le système EMM a été réalisé principalement pour supporter des langues dites européennes et ayant des caractéristiques linguistiques communes, nous avons conduit une recherche complète sur les particularités linguistiques de la langue arabe ainsi que la structure des EN dans la langue arabe afin de préparer adéquatement leur adaptation dans cette langue.

En effet, l'arabe, qui est une langue sémitique ayant une morphologie complexe, requiert une attention particulière quand il s'agit du traitement automatique de la langue selon Blachère et Gaudefroy-Demombynes (1975). Dans ce mémoire, nous illustrerons l'importance de la formalisation des connaissances linguistiques avant l'élaboration d'un système de repérage d'EN.

Par ce travail, nous avons tenté de répondre à quelques questions. Peut-on se limiter à un système à base de règles quand il s'agit d'extraire automatiquement les EN de la langue arabe? Sinon, doit-on privilégier un système à d'apprentissage automatique? Nous avons essayé d'apporter quelques éléments de réponse à ces questions dans la suite de ce mémoire.

---

6 Le système EMM de veille médiatique permet de surveiller quotidiennement environ 25 000 articles de nouvelles dans 30 langues différentes dans le but d'alerter les utilisateurs sur les événements importants d'ordre politique de la journée. Il est accessible sur le site http://press.jrc.it.

Ce mémoire n'est pas présenté sous forme d'un travail unique, mais d'un ensemble de tâches qui ont pour but d'introduire la question de l'extraction des entités nommées en arabe à travers les étapes de création d'un système de repérage et d'extraction des entités nommées.

Comme ce mémoire s'adresse en premier lieu à un public francophone, qui n'est pas forcément arabophone, nous nous sommes servi du français pour expliquer les notions reliées à la langue arabe; une translittération ou une traduction suivront systématiquement les mots arabes. Pour alléger la lecture de ce manuscrit, nous avons préféré ne pas rentrer dans les détails quand il s'agit d'aborder les caractéristiques de la langue arabe, mais nous nous limiterons à quelques notions et quelques principes que nous avons jugé utile d'inclure dans ce travail.

Nous présentons d'abord les différentes sections de notre mémoire : une introduction au domaine de l'extraction des entités nommées avec un rappel historique et une définition étendue du nom propre et des entités nommées (cf. Chapitre 1), l'état de l'art des approches et des systèmes d'extraction automatique des EN existants (cf. Chapitre 2), une étude sur les caractéristiques linguistiques et morphologiques et de la langue arabe (cf. Chapitre 3) et des entités nommées en arabe (cf. Chapitre 4). Ensuite nous détaillons la mise en œuvre du système EMM et notre méthodologie pour l'intégration du module d'extraction des entités nommées en arabe (cf. Chapitre 5). Enfin, nous présentons les résultats de notre évaluation en soulignant les forces et les faiblesses de notre approche (cf. Chapitre 6).

1.1 Définitions et contexte

Dans cette section, nous présentons quelques notions de base du domaine de l'extraction d'information, ainsi que sur le nom propre, le nom commun et l'entité nommée. Ceci est nécessaire pour la compréhension de la suite de ce mémoire.

1.1.1 L'extraction de l'information

Il s'agit d'une tâche qui consiste à extraire des informations bien définies à partir d'un texte écrit, ainsi, elle permet d'identifier par exemple de l'information dans un texte et de la représenter sous forme structurée selon les besoins et les choix de l'utilisateur. Ainsi, l'extraction d'information dans le domaine de la presse, peut aider les agences de presse dans les tâches quotidiennes qui sont effectuées à la main comme la gestion des dépêches en surveillant par exemple les dépêches ayant certains mots clés.

1.1.2 Le nom propre

Dans cette section, nous présentons la notion du nom propre par rapport à celle du nom commun. Pour des raisons pratiques et par souci de clarté, nous utilisons des exemples de la langue française, quand il s'agit de recourir à des exemples.

En grammaire, le nom propre est considéré comme une sous-catégorie de 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.

Concernant les noms propres, Jonasson (1994 : 114) propose trois définitions de leurs sens :

- 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 ».
- le nom propre est vide de sens puisqu'il permet de référer sans désigner.
- 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 référant.

Enfin Boulanger et Cormier (2001 : 21), 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 ça soit pour représenter des objets, des personnes, des lieux géographiques, des marques déposées ou même des événements.

1.1.3 Les entités nommées (EN)

Après avoir caractérisé le nom propre, nous essayerons de présenter dans ce qui suit la notion d'entités nommées.

La tâche d'extraction des EN a été proposée lors de la sixième édition de la conférence MUC (Message Understanding Conference) en 1995 Grishman (1996). En 1996, une tâche similaire, mais multilingue a suivi et a été baptisée *Multilingual Entity Task*. Cette tâche concerne les systèmes d'extraction d'entités nommées en japonais, en espagnol et en portugais. Par ailleurs, Paik *et al.* (1994) et Bauer (1985) ont proposé une classification des entités nommées et des noms propres. Dans ce qui suit, nous présenterons l'approche retenue lors de la MUC, ainsi que les catégories de Paik *et al.* (1994) et de Bauer (1985).

1.1.3.1 La conférence 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 Daille *et al.* (2000).

Les systèmes d'extraction participants ont été évalués sur des domaines tels que le terrorisme en Amérique Latine lors de la MUC-3 (MUC 1991) et de la MUC-4 (MUC 1992). Lors de la MUC-5 (MUC 1993), le domaine était la fusion d'entreprises et la fabrication de circuits électroniques. Lors de la MUC-6, le domaine était les changements de dirigeants des entreprises (MUC 1995). Enfin, la MUC-7 a porté sur les accidents d'avion Chinchor (1997).

À partir de la sixième édition de la 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. Grishman et Sundheim (1996) ont présenté en détail la conférence MUC-6 tout en faisant un rappel historique des conférences précédentes en insistant sur ce qui les distingue.

La conférence MUC-7 a distingué trois types d'entités à reconnaître et à catégoriser, soit ENAMEX, NUMEX et TIMEX (Chinchor 1997).
- 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 (ex. : *Habib Bourguiba, John Glames*) ou de familles (ex. : *les Simpsons*),
- noms de lieux : ce sont des lieux définis géographiquement ou politiquement comme les villes, provinces, rivières, montagnes (ex. : *Canada, Tunis, Le Nil*),
- organisations : cette catégorie inclut les noms de gouvernements, sociétés, et autres entités organisationnelles (ex. : *UNICEF, Reporter sans frontières, IBM*).

Les entités NUMEX rassemblent les nombres et les pourcentages, les unités de mesures, les devises (ex. : *23, cinq, un dollar, 25 %, 1 hectare, 2 kg*).

Enfin, les entités TIMEX couvrent les expressions de temps et les dates (ex. : *lundi, jour de l'an, Ramadhan, Noël, Hannuka, le trois mars 1956*).

La figure 1, illustre clairement la distinction entre la notion des entités nommées, les noms propres et les classes MUC Daille *et al.* (2000).

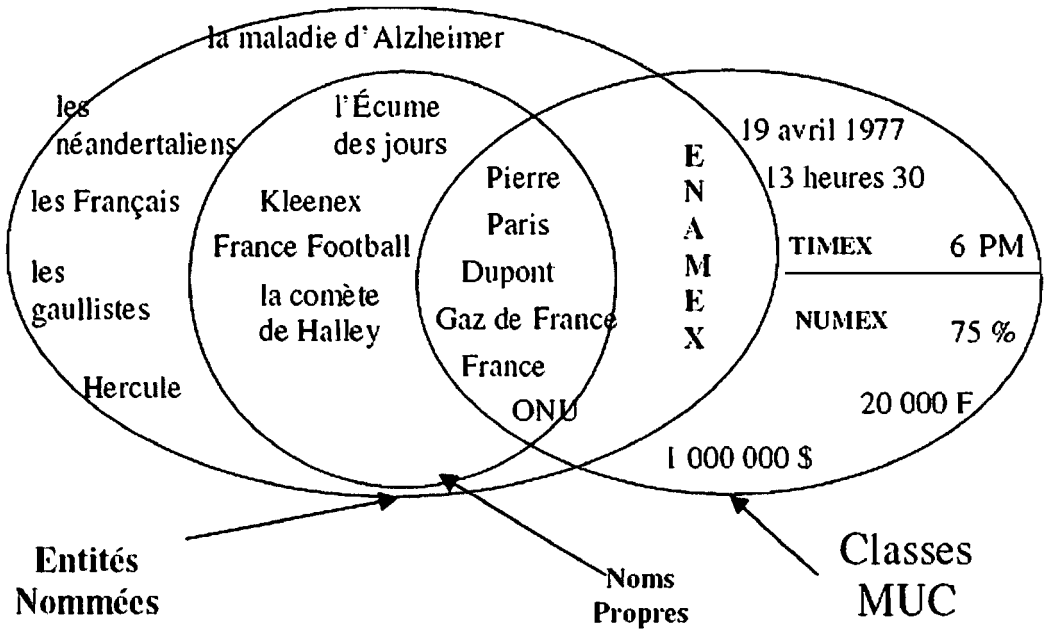


Figure 1 : Illustration de la notion d'entité nommée et les classes MUC. Daille *et al.* (2000 : 118)

1.1.3.2 La classification de Paik

La classification de Paik *et al.* (1994) regroupe ensemble les entités nommées et les entités temporelles. Cette approche, discutée par Daille *et al.* (2000), 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 :

1. **Géographique** : villes, ports, aéroports, îles, comtés ou départements, provinces, pays, continents, régions, fleuves, autres noms géographiques.
2. **Affiliation** : religions, nationalités.
3. **Organisation** : entreprises, types d'entreprises, institutions, institutions gouvernementales, organisations.
4. **Humain** : personnes, fonctions.
5. **Document** : documents.
6. **Équipement** : logiciels, matériels, machines.
7. **Scientifique** : maladies, drogues, médicaments.
8. **Temporelle** : dates et heures.
9. **Divers** : autres noms d'entités nommées.

Liste reprise intégralement de Daille *et al.* (2000 : 117).

1.1.3.3 La classification de Bauer

Bauer (1985) 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 :

1. **Anthroponymes** : les personnes individuelles ou les groupes : patronymes, prénoms, pseudonymes, gentilés, hypocoristes, ethnonymes, groupes musicaux modernes, ensembles artistiques et orchestres classiques, partis et organisations.
2. **Toponymes** : les noms de lieux : pays, villes, microtoponymes, hydronymes, oronymes, installations militaires.
3. **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'oeuvre d'art.

4. **Praxonymes** : les faits historiques, les maladies, les événements culturels.
5. **Phénonymes** : les ouragans, les zones de haute et de basse pression, les astres et les comètes.
6. **Zoonymes** : les noms d'animaux familiers.

Liste reprise intégralement de Daille *et al.* (2000 : 119).

1.1.3.4 La classification retenue

Pour la réalisation de notre système, nous avons retenu une classification qui s'inspire directement de la classification MUC-7 étant donné qu'il s'agit de la conférence qui a introduit pour la première fois, la notion d'*entité nommée*.

Nous avons classé les entités nommées en quatre catégories :

- les noms de personnes (ENAMEX dans MUC-7),
- les noms de lieux (ENAMEX dans MUC-7),
- les noms d'organisations (ENAMEX dans MUC-7),
- les entités temporelles et numériques (TIMEX et NUMEX dans MUC-7).

Dans le but de faciliter la lecture de ce mémoire, nous allons employer le terme *entité nommée* afin de désigner l'ensemble des quatre catégories mentionnées précédemment.

1.1.4 Les mesures d'évaluation

Afin de mieux comprendre la suite, nous présentons ici, les mesures d'évaluation utilisées généralement lors des tâches d'évaluation des systèmes d'extraction d'information (voir Poibeau 2001) pour plus de détails. Ainsi, il existe des

indicateurs qui permettent de mesurer les performances globales d'un système de repérage des EN : la précision et le rappel.

- La précision

La précision correspond au pourcentage des documents pertinents renvoyés par un système qui répondent effectivement à une requête. Une précision de 100 % signifie que toutes les réponses fournies par le système sont pertinentes. La précision est donnée par la formule suivante :

> Précision = nombre de réponses pertinentes du système / nombre de réponses fournies par le système

- Le rappel

Le rappel est une mesure qui calcule le pourcentage de réponses pertinentes extraites parmi les réponses pertinentes. Le rappel est donné par la formule suivante :

> Rappel = nombre de réponses pertinentes du système / nombre de réponses pertinentes réelles.

Ainsi, un rappel de 100 % signifie que toutes les occurrences pertinentes sont extraites. À ces deux notions se rajoute une mesure d'efficacité globale, la F-mesure (F-measure en anglais), initiée par Van-Rijsbergen (1979), est une mesure de synthèse qui représente la moyenne harmonique pondérée entre une précision et un rappel.

F-mesure = 2*(précision*rappel) / (précision+rappel)

2. État de l'art des systèmes d'extraction d'entités nommées

Diverses approches et systèmes d'extraction d'entités nommées ont été créés ces dernières années, dont les systèmes présentés lors des Messages Understanding Conferences. Dans ce chapitre, nous passerons en revue les diverses approches en séparant les approches à base de règles des approches à apprentissage automatique. Nous présentons aussi brièvement quelques systèmes existants pour chaque approche, dont des systèmes supportant la langue arabe.

Bien que notre intérêt ait porté en particulier sur la question de l'extraction des EN dans la langue arabe, la plupart des systèmes que nous illustrerons ont été créés et optimisés pour supporter des langues autres que l'arabe. Ceci s'explique principalement par le manque d'informations et de publications sur les systèmes d'extraction ou de repérage des entités nommées en arabe.

2.1 Les systèmes à base de règles

2.1.1 Introduction

Les systèmes à base de règles sont généralement créés à partir de règles faites à la main. Ils se basent principalement sur la description des EN grâce à des règles qui exploitent un étiquetage syntaxique, des marqueurs lexicaux et des dictionnaires de noms propres.

L'étiqueteur syntaxique reçoit en entrée un texte et produit automatiquement en sortie une version étiquetée de ce même texte. L'étiquetage consiste à produire les propriétés grammaticales d'un mot ou d'un groupe de mots dans une phrase donnée (ex. : *noms, verbes, conjonctions*) souvent accompagnée d'informations morphologiques (ex. : *genre, nombre, personne*).

Les marqueurs lexicaux sont aussi appelés *mots amorces* ou *mots déclencheurs* ou parfois *preuves contextuelles*. Il s'agit de mots ou d'indices qui entourent (à gauche ou à droite) le nom propre et qui permettent souvent de prédire sa présence.

Les règles utilisées dans ces systèmes sont décrites par des expressions de remplacement ou des expressions régulières (*Regular Expression* en anglais). Selon HIWIT (2007), « Une expression régulière permet de caractériser le format d'une chaîne de caractères [...] Le but étant de rechercher ou de remplacer un motif dans une chaîne de caractères ».

La création des règles est basée sur des marqueurs lexicaux et elle dépend généralement de l'intuition du linguiste informaticien responsable de la création de ces règles. Plusieurs opérations et tests doivent être effectués afin de s'assurer de l'efficacité des règles. Par exemple, les deux règles suivantes sont présentes dans beaucoup de systèmes à base de règles.

Article défini + titre de civilité (ex. : *M., Mme, Mlle, Mr., Dr.*) + mot majuscule -> entité nommée de type personne.

Ex. : **Le Dr. Tremblay** arrive.

Dans cet exemple, l'entité nommée détectée, ne se limite pas seulement au nom propre *Tremblay,* mais elle inclut l'article défini *Le* et le titre de civilité *Dr.* Ainsi, les EN ne se limitent pas seulement à des noms propres, mais elles peuvent inclure aussi les titres, les gentilés, etc.

Un autre exemple concerne l'usage d'un article défini (ex. *la*), d'un marqueur lexical (ex. : *ville*) et d'une information de l'étiqueteur syntaxique (ex. : *préposition*). L'exemple suivant illustre une situation où les EN comprennent des noms propres de type lieux comme des régions ou des provinces.

Article défini + Marqueur de lieu (ex. : *ville*) + préposition / *de* / + mot majuscule
-> entité nommée de type lieu.

Ex. : Ils habitent **la ville de Montréal**.

Enfin, les dictionnaires de noms propres se composent généralement d'une liste des noms et des prénoms les plus courants, des noms de lieux (villes, pays, etc.) et parfois des noms d'organisations (organismes, compagnies, etc.). Ces listes sont souvent créées à la main, toutefois, il existe des programmes informatiques qui peuvent aider à créer ces listes et ceci d'une manière semi-automatique. Les dictionnaires de noms propres sont très utilisés dans les systèmes à base de règles et même dans les systèmes automatiques.

2.1.2 Exemples de systèmes à base de règles

- Le système FUNES

Le système FUNES a été créé par Coates-Stephens (1993). Il emploie plusieurs règles syntaxiques afin de décrire la structure de l'environnement de chaque type de nom propre en anglais en se servant d'expressions régulières. Le lexique implémenté dans FUNES se compose de 2000 racines nominales et verbales ainsi que de 500 verbes avec leurs flexions.

- Le système Nominator

Nominator est créé par Wacholder *et al.* (1997). Il utilise un ensemble d'heuristiques[7] basées sur des mots en majuscules, des mots clés et la ponctuation.

7 Les heuristiques dans ce cas, peuvent se définir comme étant des indicateurs textuels qui aident le système à prendre des décisions.

De plus, ce système ne fait pas usage d'informations syntaxiques ni de listes de noms propres (à l'exception d'un petit dictionnaire). Le fait de minimiser l'emploi des ressources augmente la robustesse et la rapidité d'exécution du système selon Wacholder *et al.* (1997).

- Le système GIE

Le système GIE (Greek Information Extraction) a été développé par Karkaletsis *et al.* (1999) au sein du NCSR (National Centre of Scientific Research) avec la coopération d'un groupe de recherche en traitement automatique de la langue de l'université de Sheffield. Il s'agit d'un système d'extraction d'entités nommées pour la langue grecque. Il a été réalisé en adaptant un système existant pour la langue anglaise nommé VIE (Vanilla Information Extraction). Le système est construit autour de la plateforme d'ingénierie linguistique GATE Cunningham *et al.* (1996).

Le fait que Karkaletsis ait installé GIE sur une plateforme existante a largement facilité son travail ainsi que l'ajout de nouveaux modules spécifiques à la langue grecque. Le système GIE est composé des outils suivants :

- un analyseur morphologique,
- un outil de segmentation du texte en phrases,
- un étiqueteur morpho-syntaxique,
- des dictionnaires de noms propres,
- un analyseur syntaxique pour détecter les phrases nominales Karkaletsis *et al.* (1999).

L'outil de repérage des entités nommées de GIE se compose de son côté des éléments suivants :

- un module de recherche simple basé sur la liste des entités nommées des dictionnaires,

- un outil basé sur des règles de grammaire locales faites à la main pour reconnaître les entités nommées introuvables dans le dictionnaire,

- un module qui permet de personnaliser l'outil pour pouvoir couvrir d'autres domaines, ce module emploie une technique d'apprentissage automatique et permet la création et l'entraînement des données ainsi que la création automatique de nouvelles règles.

- Le système SPRACH-R

Renals *et al.* (1999) ont créé le système SPRACH-R afin de participer à la campagne d'évaluation Hub4 et de représenter l'université de Sheffield. Il s'agit d'un système à base de règles employant des automates à états finis Audibert (2007). Le système procède tout d'abord à la segmentation du texte en phrases. Ensuite, un module de reconnaissance d'entités nommées se charge de comparer, grâce à des règles sous forme d'automates, les mots du texte avec la liste des mots du dictionnaire des noms propres (personnes, lieux, organisations). Par ailleurs, une version modifiée de l'étiqueteur morpho-syntaxique de Brill (1995) a été développée afin d'assigner une partie de discours à chaque entrée. Enfin, un module de décision basé sur une grammaire locale se charge d'analyser l'information de sortie des modules précédents afin d'attribuer l'étiquette finale correspondante à chaque EN.

2.1.3 Conclusion

Un des avantages de l'approche à base de règles est que le contenu de la base des connaissances linguistiques est facilement accessible. Il est donc possible de la modifier ultérieurement grâce à sa transparence et à son accessibilité. Un des défauts de cette approche est que la plupart du temps, le lexique et les règles faites à

la main sont souvent optimisés pour un certain type de textes; des textes journalistiques par exemple. Ceci implique donc un enrichissement du lexique et la réécriture de certaines règles à chaque changement du type de textes, notamment les textes scientifiques et ceux traitant de la médecine. Cette approche peut s'avérer coûteuse en terme de temps de développement.

Enfin, il est possible de bâtir des systèmes très performants basés seulement sur les méthodes à base de règles Bogers (2004), mais leur réussite dépend de plusieurs éléments :

- ils doivent être adaptés manuellement pour chaque nouveau type de textes,
- toutes les règles et le lexique doivent être réécrits pour chaque langue,
- la performance du système dépend directement du niveau et de l'expérience du linguiste chargé de la création des règles,
- certaines règles sont difficiles à produire à cause de leur complexité et le linguiste doit faire usage de quelques règles *ad hoc* et de plusieurs expressions régulières assez complexes pour tenter d'extraire correctement le maximum d'EN.

2.2 Les systèmes à apprentissage automatique

2.2.1 Introduction

Les systèmes à apprentissage automatique sont conçus pour avoir une certaine « intelligence » lors de la prise des décisions. Ils se distinguent ainsi des systèmes à base de règles qui ne font qu'appliquer les règles injectées préalablement. Pour entraîner un système, il faut d'abord, lui fournir des données, appelées couramment corpus d'entraînement, et par la suite appliquer un algorithme d'apprentissage qui va permettre de construire automatiquement une base de connaissances. Le système sera alors prêt à fonctionner et à prendre des décisions d'une manière « autonome ».

Avant de pouvoir créer ce type de système, il existe plusieurs éléments à considérer avant de lancer l'opération d'apprentissage :

- la sélection du type de données qui va être employé lors du processus d'entraînement,

- choix du corpus d'entraînement, qui doit être représentatif de l'ensemble des données qui vont être traitées ultérieurement,

- choix de la meilleure méthode (algorithme) d'apprentissage, qui dépend généralement du type de la tâche demandée,

- exécution de l'algorithme sur un sous-ensemble du corpus d'apprentissage afin de vérifier les performances de l'algorithme et ajuster les paramètres au besoin, avant de valider l'opération et de commencer le processus d'apprentissage proprement dit.

Plusieurs approches d'apprentissage automatique ont été développées depuis quelques années. Nous nous inspirons dans ce travail de la classification de Bogers (2004) qui distingue les méthodes d'apprentissage selon qu'elles soient supervisées ou non. Dans ce qui suit, nous énumérons les principales approches automatiques d'extraction des EN.

Dans l'approche automatique par apprentissage supervisé, il s'agit de faire un apprentissage à partir d'un corpus d'entraînement, préalablement préparé. Cette méthode, qui requiert tout de même une intervention humaine considérable, se distingue de l'apprentissage non supervisé dans lequel l'algorithme procède à un apprentissage avec une intervention humaine minimale.

La méthode non supervisée fait usage, entre autres, de la technique du *clustering* qui consiste à regrouper automatiquement les entités similaires Jackson et Moulinier (2002). Enfin, il y a les approches dites mixtes, appelées aussi hybrides, combinant l'apprentissage automatique et l'écriture des règles à la main.

Après la description de chaque approche, nous décrirons brièvement quelques systèmes existants dans la mesure où nous disposons d'assez d'information sur leur fonctionnement.

2.2.2 Apprentissage supervisé (*supervised learning*)

Il s'agit d'une méthode d'apprentissage qui nécessite une plus grande intervention humaine à chacune des étapes de l'opération d'apprentissage.

Ce type d'apprentissage requiert en général un volume important de données pour les besoins de l'entraînement. La performance du système augmentera proportionnellement avec la quantité et la qualité du corpus d'apprentissage. Nous illustrons dans ce qui suit, quelques approches populaires d'apprentissage supervisé.

- Les modèles de Markov cachés (*Hidden Markov Model*)

Les modèles de Markov cachés (HMM) sont basés sur une approche purement probabiliste très populaire. Elle procure généralement d'assez bons résultats comme le souligne Zhou *et al.* (2000) qui font usage de cette approche afin d'implémenter leur système de repérage d'EN. Selon Merialdo (1995 : 10) : « Les modèles de Markov constituent un des types de modèles probabilistes les plus utilisés, en raison de leur simplicité et de leur efficacité ». Sans entrer dans les détails des formules mathématiques derrière cette méthode, nous expliquerons brièvement le principe de cette approche.

Si on suppose qu'au départ nous avons choisi un modèle pour chaque catégorie d'entités nommées, bien évidemment l'apprentissage se fait à partir d'un corpus d'entraînement, où tout le texte, y compris les entités nommées, a été préalablement étiqueté et catégorisé.

La première étape de l'apprentissage consiste à garder en mémoire un modèle pour chaque mot dans le corpus d'entraînement. Imaginons maintenant que le modèle HMM rencontre la phrase suivante dans un corpus d'évaluation :

Dhahaba ali ilaa **bariz** « Ali est allé à **Paris »**

Le mot *bariz* qui suit immédiatement la préposition *illaa* n'a pas été reconnu par le modèle HMM, vu que ce dernier n'était pas observé dans le corpus d'apprentissage. Dans ce cas, le modèle HMM va employer le contexte précédent afin d'analyser le mot inconnu. Le contexte précédent dans notre exemple consiste en la préposition *illa* « à ». Si, on suppose que dans notre corpus d'apprentissage, la préposition *illa* est fréquemment suivie d'un nom de lieu. Dans ce cas, le système va attribuer la catégorie nom de lieu au mot *bariz* en se basant sur le modèle probabiliste du contexte d'apparition de la préposition *illa* dans le corpus d'apprentissage.

Afin d'illustrer cette approche, nous avons choisi de présenter le système IdentiFinder qui est réalisé par BBN Bikel *et al.* (1997) et Boisen *et al.* (2000). BBN est une compagnie qui a investi beaucoup d'efforts dans le développement d'outils linguistiques multilingues avec un intérêt croissant pour la langue arabe. Le système IdentiFinder est un système basé sur l'apprentissage supervisé avec des modèles de Markov cachés. Pour reconnaître les noms propres, l'apprentissage est réalisé sur des corpus préalablement étiquetés. C'est donc un modèle probabiliste assez simple. Initialement, IdentiFinder ne disposait pas d'un dictionnaire ou d'une liste de noms propres. Par la suite, dans le but d'améliorer ses performances, IdentiFinder a été doté d'un très grand dictionnaire de noms propres.

La performance de ce système est en général très proche des systèmes à base de règles pour les langues ayant la distinction majuscule / minuscule et elle est même meilleure dans les langues n'ayant pas cette distinction comme l'arabe Bikel *et al.* (1999 : 1).

- Les arbres de décision

Il s'agit d'une des méthodes les plus populaires de l'apprentissage supervisé.
Elle se présente sous forme d'un arbre ayant un ensemble de branches. Chaque branche correspond à une décision et dispose d'un certain poids et d'une certaine probabilité pour prendre une décision donnée (voir Figure 2)

Cette méthode requiert la création d'un corpus d'entraînement. Chaque nœud dans l'arbre spécifie un test et toutes les autres feuilles de l'arbre venant de ce nœud correspondent à une valeur possible de cette instance.

La classification des nouvelles instances se fait par un parcours à travers les branches de l'arbre et chaque trait correspondant jusqu'à l'obtention d'un résultat satisfaisant. Paliouras *et al.* (2000) ont montré que l'emploi de l'arbre de décision pour la génération des règles de décision peut dans certain cas surpasser les systèmes à base de règles.

Un des avantages de cette technique est la possibilité de générer automatiquement l'ensemble des décisions sous forme de règles facilement compréhensibles par le linguiste. De plus, cette méthode est assez fluide du point de vue du traitement et du calcul informatique. Par contre, faute d'un bon et d'un assez large corpus d'entraînement, cette méthode risque d'être peu performante.

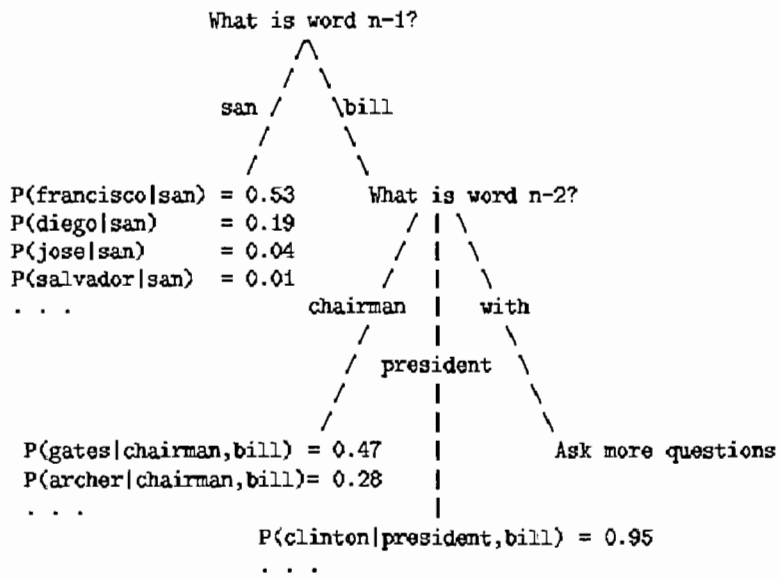


Figure 2 : Exemple d'un arbre de décision Borthwick (1999 : 11)

Gallippi (1996) a créé une stratégie d'acquisition des noms de personnes à base d'arbres de décision. Il utilise des heuristiques (aidées par la morphologie, le lexique ou la syntaxe) et tente d'en acquérir automatiquement de nouvelles (voir Tableau I). Gallippi applique les arbres de décisions de l'anglais directement sur les textes espagnols et japonais (les arbres spécifiques à l'espagnol et au japonais sont créés ensuite grâce aux résultats obtenus avec les arbres de décisions de l'anglais).

Le système de Gallippi a obtenu une F-mesure de 94,00 % sur l'anglais, 89,20 % sur l'espagnol et enfin 83,10 % sur le japonais. Pour plus de détails sur les résultats voir le tableau II qui inclut les résultats de Gallippi comparés à d'autres systèmes.

Type	Feature	Example	How many
Part of Speech	Proper Noun Common Noun	"Aristotle" "philosophy"	NA NA
Designator	Company Person Location Date	"Corp.", "Ltd." "Mr.", "President" Country, State, City Month, Day of week	100 E, 110 S, 60 J 70 E, 70 S, 43 J 520 E, 900 S, 570 J 56 E, 19 S, 19 J
Morphology	Capitalization Company Suffix Word Length	"A-", "B-" "-corp", "-tec" WL>8, WL<3	1 E, 1 S, 0 J 5 E, 0 S, 30 J 4 E, 4 S, 2 J
List	Companies Persons Locations Nationalities Keyword(s)	"IBM", "AT&T" "Smith", "Michael" "Gulf of Mexico" "Japanese" "based in", "said he"	0 E, 100 S, 7K J 21K E, 21K S, 185K J 20 E, 20 S, 2K J 220 E, 0 S, 0 J 44 E, 49 S, 54 J
Template	Company Person Location Date Proper Name	< NNP CN_desig > < P_Desig NNP > < NNP L_desig > < MM Num , Num > < NNP NNP >	210 E, 210 S, 210 J 90 E, 95 S, 90 J 190 E, 190 S, 190 J 17 E, 18 S, 70 J 140 E, 140 S, 140 J
Special Purpose	Lngst Cm Substr Duplicated PNs	"VW" <- Volkswagen DUP_2+, DUP_5+	1 E, 1 S, 1 J 5 E, 5 S, 2 J

Tableau I : Extrait des heuristiques employées par Gallippi Gallippi (1996 : 425)

System	Lang.	Class	R	P	P&R
Rau	English	Com	NA	95	NA
PNP (McDonald)	English	Com Pers Loc Date	NA	NA	"Near 100%"
Panglyzer	Spanish	NA	NA	80	NA
MAJESTY	Japanese	Com Pers Loc	84.3 93.1 92.6	81.4 98.6 96.8	82.8 95.8 94.7
MNR (Gallippi)	English	Com Pers Loc Date (Avg)	97.6 98.2 85.7 100	91.6 100 91.7 100	94.5 99.1 88.6 100 94.0
MNR	Spanish	Com Pers Loc Date (Avg)	74.1 97.4 93.1 100	90.9 79.2 87.5 100	81.6 87.4 89.4 100 89.2
MNR	Japanese	Com Pers Loc Date (Avg)	60.0 86.5 80.4 90.0	60.0 84.9 82.1 94.7	60.0 85.7 81.3 92.3 83.1

Tableau II : Résultats du système de Gallippi et d'autres systèmes Gallippi (1996 : 428)

2.2.3 Apprentissage paresseux (*lazy learning*)

Plusieurs méthodes d'apprentissage paresseux existent, nous illustrons une méthode qui était populaire lors de la MUC-6.

Ce modèle commence par l'enregistrement des données du corpus d'entraînement dans la mémoire et procède par la suite à la comparaison des nouvelles requêtes avec celles qu'il vient de garder en mémoire, ensuite, à l'aide d'une formule mathématique, il calcule le degré de similarité des lettres composant les deux requêtes.

Enfin, le système va attribuer à la nouvelle requête la catégorie de l'exemple d'entraînement le plus proche. Cette approche mobilise d'importantes ressources mémoire pour gérer toutes les opérations d'enregistrement et de comparaison.

Supposons maintenant que nous disposons d'un corpus d'entraînement fictif limité à deux éléments (voir Tableau III).

Liste des éléments dans le corpus d'entraînement	**Catégorie attribuée manuellement**
Canada	Pays
Mohamed	Personne

Tableau III : extrait d'un corpus d'entraînement

Observons alors les nouvelles requêtes suivantes :

Requête 1 : Kannada

Requête 2 : Canada

Requête 3 : Muhammed

Requête 4 : Kandahar

Suite au calcul de la distance d'édition[8] entre ces 4 requêtes et les 2 éléments dans le corpus d'entraînement, le système va attribuer aux nouvelles requêtes, les catégories les plus probables en respectant un seuil de similarité bien déterminé.

Résultat de la requête 1 : Kannada, mot inconnu, atteint le seuil de similarité orthographique avec Canada. Il reçoit donc la catégorie pays.

Résultat de la requête 2 : Canada, mot trouvé, il reçoit la catégorie pays.

Résultat de la requête 3 : Muhammed, mot inconnu, mais proche orthographiquement de Mohamed, il reçoit donc la catégorie personne.

Résultat de la requête 4 : Kandahar, mot inconnu, le seuil de similarité n'est pas atteint à la différence de Kannada. Il n'est proche d'aucun élément dans le corpus, donc la catégorie n'est pas trouvée.

Après l'observation des résultats fictifs de sortie, on constate que la méthode de calcul de similarité entre les exemples précédents est assez stricte puisqu'elle n'accepte de catégoriser que les mots inconnus qui sont très proches des mots figurants dans la liste. Ce qui n'est pas le cas du mot Kandahar qui est jugé loin du mot Canada.

Les méthodes de calcul de similarité diffèrent d'un système à un autre et il incombe au concepteur de choisir la méthode appropriée pour son système après la réalisation des tests nécessaires.

---

8 La distance d'édition ou distance de Levenshtein mesure la similarité entre deux chaînes de caractères. Elle calcule le nombre minimal de caractères qu'il faut remplacer, insérer ou supprimer afin de passer d'une chaîne à l'autre Levenshtein (1966 : 707–710).

- Le système de De Meulder et Daelemans

De Meulder et Daelemans (2003) ont employé le système d'apprentissage TIMBL afin de repérer les noms de personnes dans des journaux allemands et anglais. Ils ont combiné un corpus d'apprentissage avec un lexique. Les premiers résultats ont montré que le lexique n'a apporté aucune amélioration des performances sur l'anglais; par contre, il a pu améliorer les résultats sur l'allemand.

2.2.4 Apprentissage légèrement supervisé (*weakly supervised*)

Cette méthode est parfois appelée *apprentissage limité*. Parmi les principales techniques utilisées, nous citons le *bootstrapping* Borthwick (1999). Cette méthode n'a besoin que d'un nombre limité de données injectées préalablement afin de fonctionner correctement.

L'injection des données appelées *semence* (*seed rules*), peut être sous forme d'une liste de noms de personnes par exemple. Le système procède alors à l'analyse des phrases contenant un certain type d'entités nommées, ensuite, le système retient les marqueurs lexicaux dans le contexte immédiat de ces entités. En répétant ce processus, un grand nombre de noms de personne peut être repéré. Ce principe a été employé par Riloff et Jones (1999), Yanarber *et al.* (2000) et Collins et Singer (1999).

- Le système de Collins et Singer

Collins et Singer (1999) ont créé un système basé sur l'apprentissage légèrement supervisé travaillant sur des textes en anglais. Les règles sont au nombre de 7 et sont extrêmement simples. Elles ressemblent aux règles décrites ci-dessous :

Montréal est un lieu; *Québec* est un lieu; *Canada* est un lieu; tout nom qui contient *Mr.* est un nom de personne; tout nom qui contient *Incorporated* ou *Inc.* est un nom d'organisation; *I.B.M.* est une organisation; *Microsoft* est une organisation;

Ces 7 règles de base permettent à l'algorithme d'apprentissage de se déclencher et de déduire de nouvelles règles. Par exemple, avec la phrase suivante :

Mr. Alfredo.

Le système infère que *Alfredo* est un nom de personne, car il est précédé par le titre *Mr.* qui prédit un nom de personne dans les règles et ainsi et suite.

2.2.5 Apprentissage non supervisé (*unsupervised learning*)

Selon Candillier (2006 : 250), l'apprentissage non supervisé « consiste à former différents groupes à partir d'un ensemble de données, de telle manière que les données considérées comme les plus similaires soient associées au même groupe et qu'au contraire les données considérées comme différentes se retrouvent dans des groupes distincts, permettant ainsi d'extraire de la connaissance à partir de ces données ». L'approche typique dans cette méthode d'apprentissage est le clustering (le regroupement automatique d'un ensemble de données similaires). Les EN sont groupées automatiquement en sous-ensembles selon la similarité de leur contexte immédiat.

- Le système de Cuchiarelli et Velardi

L'apprentissage non supervisé a été employé par Cuchiarelli et Velardi (1999), dans le but d'acquérir les marqueurs lexicaux indiquant la présence des noms propres et de les appliquer par la suite sur un corpus non étiqueté afin de détecter les noms

propres. Le but de Cuchiarelli et Velardi était d'employer cette approche en guise de complément à leur système basé sur un apprentissage supervisé.

2.2.6 Systèmes d'apprentissage statistique

Dans cette section, nous présenterons deux systèmes à apprentissage statistique.

- Le système de Cucerzan et Yarowski

Cucerzan et Yarowski (1999) ont créé un système de repérage des noms propres qui utilise des probabilités et un ensemble de données d'entraînement très réduit. L'utilisateur du système doit fournir des exemples de noms de personnes, de prénoms, ainsi que de noms de lieux (les noms d'organisations ne sont pas gérés). L'algorithme d'apprentissage se base sur ces exemples pour pouvoir générer automatiquement sa base de connaissance. La base de connaissance peut aussi contenir des informations morphologiques comme les suffixes et les préfixes.

Dans des langues comme le roumain et le turc, les suffixes sont de bons indicateurs pour repérer les entités nommées (par exemple les suffixes des noms en *escu* en roumain indiquent un nom de personne). Le système de Cucerzan et Yarowski (1999) va essayer dans un premier temps d'analyser les informations morphologiques dans les langues ayant ces traits, sinon le système va prendre seulement en compte le contexte immédiat des noms propres afin de détecter les EN.

- Le système de Borthwick

Le système MENE proposé par Borthwick (1999) réalise une reconnaissance statistique des entités nommées avec un algorithme appelé *entropie maximale*. Le système MENE peut facilement être adapté à d'autres langues. Ainsi, Borthwick

(1999) a réussi à adapter son système de l'anglais au japonais, sans avoir une connaissance de la langue japonaise, en l'espace de trois jours seulement.

Les résultats obtenus sur le corpus japonais étaient supérieurs de six points à la moyenne des autres systèmes Borthwick (1999), ce qui est très prometteur. Borthwick a utilisé un corpus annoté de 294 000 mots pour le japonais; il recommande un minimum de 300 000 mots pour adapter son système à toute nouvelle langue Borthwick (1999). De plus, Borthwick propose aux auteurs des systèmes à base de règles de combiner leurs systèmes avec MENE afin d'améliorer sensiblement leurs résultats. La combinaison pourra alors se faire comme une deuxième passe du processus de repérage des EN : la première sera dans ce cas celle du système à base de règles et la deuxième sera l'application de MENE sur les résultats du système à base de règles Borthwick (1999).

2.2.7 Systèmes hybrides

Pour construire un système de repérage d'entités nommées, les systèmes hybrides se distinguent par leur méthode qui consiste à combiner les techniques d'apprentissage automatique (à partir d'un corpus) à des méthodes à base de règles (création de règles à la main). Les modèles hybrides sont de plus en plus populaires et sont employés dans plusieurs systèmes comme celui de Cucchiarelli et Velardi (1999).

- Le système de Fourour

Le système Nemesis a été proposé par Fourour (2002). Ce système est basé sur des règles de grammaire ainsi que sur un module d'apprentissage ; de plus, il exploite des lexiques spécialisés. Nemesis réalise une première reconnaissance de noms

propres à l'aide du lexique et d'un ensemble de marqueurs lexicaux. Ensuite, des expressions régulières[9] sont appliquées au texte pour extraire les noms propres.

- Le système de Dalianas et Aström

SweNam extrait des EN dans la langue suédoise Dalianas et Aström (1998). Ce système combine des techniques d'apprentissage (sur environ 10 000 articles) et des règles pour construire un système de repérage d'entités nommées. Étant donné que le nom propre en suédois est facilement identifiable à partir des suffixes, SweNam cherche les suffixes connus pour extraire et typer les noms propres (ex. : le suffixe – *son* dans le nom Richardson).

- Le système de Senellart

Senellart (1998) automatise partiellement la construction des transducteurs[10] pour la reconnaissance des noms propres en français. Les transducteurs sont construits à l'aide d'une concordance dont les parties pertinentes sont choisies à la main et sont intégrées automatiquement dans des transducteurs qui vont reconnaître les noms propres. Les groupes nominaux décrits dans les transducteurs sont composés de noms de personnes et de leurs contextes (titres, fonctions, métiers, etc.). Ils sont ensuite appliqués à de nouveaux documents pour repérer les noms propres.

- Le système de Poibeau

Poibeau (1999) utilise aussi des transducteurs dans le module SemTex du projet

---

9 Les expressions régulières sont employées dans ce cas pour déterminer si une chaîne de caractères répond ou pas à un modèle donné. Elles permettent aussi d'opérer certaines manipulations afin de transformer une chaîne de caractères.

10 Un transducteur à états finis est un dispositif algorithmique qui représente un ensemble de séquences en entrée et qui leur associe des séquences produites en sortie. Un transducteur permet de modifier les séquences en entrée (effacement, remplacement) Balvet (2001).

ECRAN[11] : ce système repère principalement les noms d'entreprises et les noms de leurs dirigeants afin de faire de la veille économique. SemTex fait l'extraction des noms propres avec des patrons prédéfinis en anglais et en français dans des journaux comme *Le Monde* et *Herald Tribune*. Un seul transducteur contient des appels à d'autres transducteurs qui décrivent les grammaires des entités nommées suivantes : email, URL, date, lieu, personne, compagnie.

Il est à noter que SemTex utilise les noms propres qu'il a appris avec ces grammaires pour extraire les mots inconnus qui leur sont identiques ou proches et de les étiqueter avec le même type.

Notons enfin qu'un problème se pose au système SemTex ainsi qu'aux systèmes employant des transducteurs contenant toutes les grammaires à l'image de SemTex. Le problème est le suivant : si deux règles de même longueur d'un transducteur peuvent s'appliquer sur une séquence du texte, le résultat va dépendre de l'ordre dans lequel les règles du transducteur ont été compilées.

- Le système de Mikheev

Le système LTG Mikheev *et al.* (1998), a été le plus performant pour la langue anglaise lors de la MUC 7. LTG a obtenu un rappel de 93,6 % pour une précision de 95,00 % sur les entités ENAMEX. LTG dispose d'un analyseur morphosyntaxique. De plus, il intègre un module nommé *Lttok* qui permet de repérer des noms propres candidats. Ensuite, le module *FsgMatch* utilise les résultats de *Lttok* pour extraire les entités nommées et les catégoriser. Dans ce qui

11 ECRAN (Extraction of Content Research At Near-market) est un projet européen d'extraction d'information portant sur le français, l'anglais et l'italien. Ce projet propose d'offrir un accès filtré à la masse d'information textuelle délivrée par la télévision et les ordinateurs personnels Poibeau (1997).

suit, nous illustrons l'opération d'extraction des entités nommées du système LTG à travers les différentes étapes du processus.

Étape 1 : Passage des règles les plus sûres (*sure-fire rules*). Après un étiquetage en parties du discours du texte, cette étape applique des règles qui contiennent des marqueurs lexicaux. Les noms de lieux sont reconnus par leur contexte, par exemple « à » suivi de « Philadelphie », et grâce à un dictionnaire.

Étape 2 : Reconnaissance partielle. Cette étape est réalisée par l'interaction de deux modules. Le premier module se charge de collecter les entités déjà identifiées dans le document. Ensuite le système génère des variantes d'entités nommées en changeant l'ordre des mots. Par exemple, dans le cas suivant, l'entité « *Jules Lafontaine Inc.* » suggère un nom d'une organisation, les variantes de cette entité comme « *Jules Lafontaine* », « *Lafontaine Inc.* », « *Lafontaine* » ou « *Jules* » vont recevoir le type organisation comme catégorie, mais cet étiquetage n'est pas définitif. Le second module utilise un algorithme probabiliste pour finaliser l'étiquetage des noms propres.

Étape 3 : Règles relâchées (Rules relaxation). Ce sont des règles plus souples en terme de contraintes contextuelles. Par exemple, un prénom (déjà présent dans le dictionnaire des prénoms) suivi d'un ou plusieurs mots en majuscules et inconnus sera suffisant pour catégoriser la séquence comme un nom de personne.

Étape 4 : Deuxième reconnaissance partielle. Lorsque toutes les ressources du système ont été utilisées, la reconnaissance partielle annote les noms propres restants de manière probabiliste.

Étape 5 : Traitement des grands titres des journaux : En anglais, la plupart des grands titres dans les journaux sont rédigés entièrement en lettres majuscules. Dans ce cas, les lettres majuscules ne peuvent pas servir comme des indices pour repérer

les noms propres, c'est pourquoi Mikheev *et al.* (1998) ont implémenté des règles et un algorithme probabiliste pour traiter ces cas particuliers.

2.3 Systèmes de repérage des entités nommées pour la langue arabe

Après avoir illustré les principales tendances, outils et techniques employés pour le repérage des EN, nous présentons brièvement dans cette section quelques systèmes de repérages des EN pour la langue arabe.

Nous tenons d'abord à souligner que les informations et les publications sur les systèmes de repérage des EN arabe étaient assez rares comme l'affirment Benajiba *et al.* (2007). Quand elles existent, elles sont sous forme de brochures destinées à la commercialisation du produit. La plupart des intervenants dans ce domaine sont en effet des compagnies privées et il est normal dans ce cas de ne pas divulguer d'informations confidentielles sur l'architecture du produit pour des raisons stratégiques et commerciales. Toutefois, nous avons pu récolter des informations sur quelques produits, soit à travers les sites Web, soit en contactant directement le responsable de la recherche et du développement de la compagnie concernée.

Certains systèmes parmi ceux énumérés dans la section précédente ont été adaptés pour supporter la langue arabe et d'autres ont été créés exclusivement pour la langue arabe.

2.3.1 La compagnie Inxight

Depuis quelques années la compagnie Inxight (2009) a montré un intérêt particulier pour la langue arabe et un effort particulier à été effectué pour l'adaptation des outils existants pour l'arabe. C'est le cas pour Inxight ThingFinder qui est un système de repérage des EN.

ThingFinder est un système à apprentissage automatique qui associe des informations syntaxiques et morphologiques grâce à un analyseur morphosyntaxique. Il réalise une analyse grammaticale locale avec deux grammaires concurrentes : une grammaire spécifique écrite à la main et une générale qui a été automatiquement extraite d'un corpus de référence où les entités nommées étaient connues.

ThingFinder se distingue aussi par son module Pattern Builder, qui est un système interactif de création de règles personnalisées par l'utilisateur. Ainsi, selon son domaine et ses besoins particuliers, un utilisateur peut ajouter à sa guise des règles et des expressions régulières.

Enfin, ThingFinder permet de détecter une grande variété d'EN, environ 25 catégories, puisqu'il ne se limite pas seulement aux grandes catégories, mais prend aussi en charge les sous-types. Par exemple au lieu de se limiter seulement aux noms de lieux, il inclut également des sous catégories de lieux comme *pays, ville, village, rivière, montagne*.

2.3.2 La compagnie Apptek

La compagnie Apptek (2009) se spécialise dans le secteur du traitement automatique de la langue arabe. Depuis quelques années, elle a lancé le système NameFinder qui est un système d'extraction des entités nommées pour la langue arabe. Leur système combine les méthodes statistiques aux règles linguistiques. Il comprend un dictionnaire de noms propres et fait usage des marqueurs lexicaux et repose sur une méthode de recherche exhaustive (*Brut force lookup*). Il s'agit d'un algorithme de recherche simple, qui consiste à énumérer tous les candidats possibles et à sélectionner celui qui satisfait le plus aux critères de recherche. Par ailleurs, Apptek offre la possibilité de mise à jour du lexique grâce à une base de données dynamique. Par ailleurs, une base de données ethnolinguistique est

présente afin de fournir des informations supplémentaires sur les noms de personnes.

2.3.3 La compagnie SRA

La compagnie SRA (2009) a produit le système Net Owl Extractor disponible pour la langue arabe. Net Owl intègre Robotag qui est un système multilingue d'apprentissage basé sur les arbres de décision, ce qui lui permet de générer une phase de révision des règles : cette phase permet de choisir l'interprétation la plus probable pour un texte. Le poids d'une règle dépend de sa longueur, de son numéro d'ordre et du type de nom propre qu'elle repère. Par ailleurs, Net Owl fait usage des règles linguistiques de base et du contexte immédiat et dispose d'un dictionnaire de noms propres assez volumineux. Enfin, Net owl permet de reconnaître 7 catégories principales d'EN ainsi que 70 sous-catégories.

2.3.4 La compagnie Basistech

À l'inverse des autres compagnies, le site Web de Basistech (2009) dispose de quelques informations techniques sur leurs produits, notamment sur Rosette Named Entity Extractor (REX), illustré dans la Figure 3. REX est basé sur des modèles statistiques ainsi que sur des règles linguistiques complexes et se sert d'outils de prétraitement linguistique comme un analyseur morphologique et un étiqueteur syntaxique.

Les modèles statistiques de REX ont été construits à partir de larges corpus de langue arabe. Toutefois, il est toujours possible à l'utilisateur d'intervenir afin entraîner spécifiquement le système pour couvrir une variété de textes. Enfin, grâce à des règles syntaxiques, REX est capable de délimiter les syntagmes nominaux au sein d'un texte afin de pouvoir mieux isoler les entités nommées.

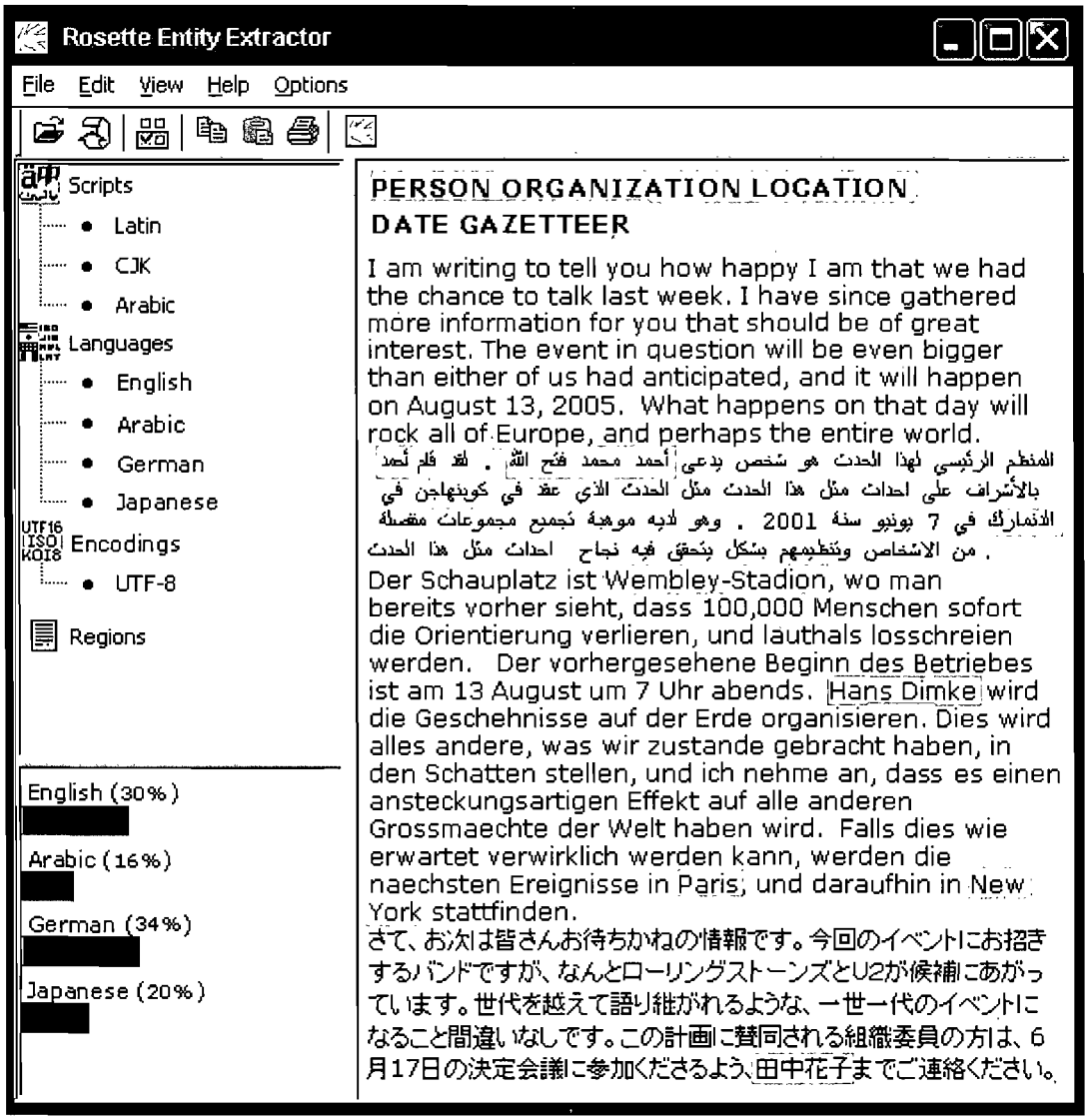


Figure 3 : Capture d'écran du système Rosette Named Entity Extractor Basistech (2009)

2.3.5 La compagnie Sakhr

La compagnie Sakhr (2009) est considérée comme la première compagnie privée dans le monde à se spécialiser dans le traitement automatique de la langue arabe et cela depuis une vingtaine d'années. Pour l'extraction des entités nommées, le système Siraj est basé sur le moteur de traitement linguistique Arabic Linguistic

Engine (ALE). Peu d'information est disponible sur l'architecture de Siraj, mais nous postulons qu'il est basé en partie sur une plateforme riche d'outils de traitement automatique comme un analyseur morphologique et un étiqueteur des parties du discours.

Nous avons pu tester la version démo de Siraj, disponible en ligne sur le site de la compagnie (voir Figure 4). Le test de Siraj nous a permis d'avoir une idée sur les performances de l'outil ainsi que sur son fonctionnement. Sans prétendre évaluer l'outil exhaustivement, ce test nous a permis de constater que Siraj permet de repérer correctement un très grand nombre d'EN.

Figure 4 : Illustration de système Siraj (Sakhr 2009)

2.3.6 La compagnie LANGUAGE ANALYSIS SYSTEMS (LAS)

Le site Web de la compagnie LAS (2009) ne dispose pas d'assez d'informations sur leur système de repérage d'entités nommées. Le système de LAS permet d'effectuer une analyse complète des noms propres détectés ; notamment en fournissant des informations ethnolinguistiques et sémantiques. LAS dispose d'une base de données multilingue d'environ un million de noms propres.

2.3.7 La compagnie ClearForest

La compagnie ClearForest (2009) dispose d'un outil d'extraction de l'information qui inclut un module de repérage d'entités nommées pour l'arabe : ClearForest Extraction Modules. Selon la fiche du produit, cet outil se distingue des autres systèmes, que ce soit des systèmes basés sur les méthodes d'apprentissage automatique ou à base de règles, par un système d'étiquetage flexible et extensible qui peut intégrer des modules linguistiques propres selon le domaine et les besoins du client. De plus, selon le site Web de la compagnie, le système de Clearforest permet de fournir des informations sur les relations entre les entités nommées trouvées, afin d'établir le réseau social de chaque entité.

2.3.8 Le système ANERsys

Benajiba *et al.* (2007) ont crée ANERsys, un système de repérage d'EN pour la langue arabe. Ce système est basé sur une méthode d'apprentissage statistique qui emploie l'algorithme d'entropie maximale. L'apprentissage automatique de ANERsys a été effectué sur un corpus de 125 000 mots. Dans le but d'améliorer les performances de leur système, Benajiba *et al.* (2007) ont combiné leur approche à un lexique qui a été construit manuellement à partir de plusieurs sites de nouvelles en ligne. Le lexique comprend 1 950 noms de lieux, 1 920 noms de personnes et 262 noms d'organisations.

Une évaluation du système a été réalisée sur un corpus de nouvelles en lignes totalisant 25 000 mots. Le système a obtenu une précision globale de 63,21 % et un rappel global de 49,04 % soit une F-mesure de 55,23 %.

2.4 Autres travaux

Quelques projets ont été spécifiquement développés pour la reconnaissance du nom propre en arabe en excluant les autres catégories d'entités comme les lieux ou les organisations. C'est pour cette raison que nous avons inclus ces travaux dans une section à part.

- Samy *et al.* (2005) ont combiné un outil d'extraction d'entités nommées en espagnol avec un corpus parallèle aligné (espagnol et arabe). L'opération consiste à extraire dans un premier temps les noms propres du texte en espagnol. Ensuite, les noms propres espagnols trouvés vont être alignés à ceux en arabe grâce à une simple méthode de translitération [12] des caractères en arabe vers des caractères non latins. Enfin, il faut noter qu'un désavantage de cette méthode est l'obligation d'avoir un corpus parallèle afin qu'elle soit fonctionnelle.

- Abuleil (2004) de son côté a présenté une méthode d'extraction des noms propres en arabe dans le but d'alimenter une base de données qui peut servir au sein de systèmes de questions-réponses. Le système commence par sélectionner les phrases qui peuvent contenir des noms propres. Ensuite, une analyse des mots de chaque phrase retenue permet de créer des graphes représentant la relation entre les mots au sein de la phrase. Enfin, des règles vont repérer et classer les noms propres avant de les enregistrer dans une base de données.

---

12 La translittération est l'opération consistant à faire correspondre aux symboles graphiques d'un système d'écriture, des symboles d'un autre système, et ce, indépendamment de la prononciation. Source : United Nations Statistical Division (2002 : 63).

- Maloney et Niv (1998) ont développé l'outil TAGARAB qui est un outil de repérage des noms propres en arabe basé sur une analyse morphologique du mot. L'analyseur morphologique permet à TAGARAB d'isoler le nom propre en spécifiant sa limite gauche grâce à des informations comme la déclinaison du mot ou sa partie du discours.

2.5 Conclusion

Nous avons vu, dans ce chapitre, trois grands types de méthode de repérage d'entités nommées : la méthode à base de règles, les méthodes par apprentissage automatique et les méthodes hybrides.

Les systèmes les plus communs sont les systèmes à base de règles. Ce type de systèmes exige un grand investissement dans l'effort de développement, mais ils ont pour avantage de fournir de très bons résultats Friburger (2002 : 40). Les systèmes à base de règles se basent généralement sur l'emploi d'une liste de marqueurs lexicaux et d'un dictionnaire combiné avec des règles faites à la main. Les systèmes à apprentissage minimisent la description linguistique et obtiennent en général des résultats plus faibles. Les systèmes hybrides sont très intéressants ; on peut citer en exemple les résultats du système LTG, qui a obtenu la meilleure performance lors de la MUC-7 pour la langue anglaise avec une F-mesure de 93,39 % Mikheev *et al.* (1998).

Nous présentons notre propre système d'extraction d'entités nommées dans le chapitre 5 et l'évaluation des résultats dans le chapitre 6. Dans les chapitres 3 et 4, nous reviendrons sur quelques caractéristiques linguistiques de la langue arabe et sur la composition des entités nommées en arabe. Le fait d'avoir conduit cette recherche sur la langue arabe et sur les structures des entités nommées en particulier va nous permettre d'implémenter notre outil et de formaliser les règles de repérage.

3. Particularités de la langue arabe pour le TAL[13]

La langue arabe est une langue sémitique qui s'écrit et se lit de droite à gauche. Ayant un alphabet de 28 lettres et un système morphologique très complexe Blachère et Gaudefroy-Demombynes (1975)

Notre travail porte sur l'arabe moderne standard qui est la langue écrite dans la presse de nos jours, dans les articles et les publications scientifiques ainsi que dans le discours officiel dans l'ensemble des pays arabes.

Avant de pouvoir commencer les démarches pour la création d'un système d'extraction de l'information pour l'arabe, plusieurs particularités de la langue l'arabe doivent être prises en compte vu leurs impacts sur la création du système. En effet, certaines particularités linguistiques de la langue arabe posent des difficultés, en particulier quand il s'agit du domaine du repérage d'entités nommées. De plus, la langue arabe possède certains atouts, qui, une fois connus et exploités, peuvent contribuer énormément au succès de tout système de traitement automatique de la langue.

Une analyse linguistique rigoureuse doit donc être faite au préalable; c'est dans ce cadre que nous avons orienté ce chapitre. Une importance particulière sera accordée à tous les éléments de la langue ayant une incidence directe sur l'extraction d'information qui sont susceptibles de poser des difficultés du point de vue informatique, mais aussi linguistique. Nous avons examiné de près quelques traits de la langue arabe. Nous résumons dans ce qui suit, les points clés de la langue arabe à considérer en vue de l'implémentation du système de repérage des entités

---

13 Traitement automatique de la langue.

nommées. Notre recherche a porté sur plusieurs aspects de la langue, comme la morphologie, l'orthographe et la ponctuation.

3.1 Le système morphologique de l'arabe

Le mot arabe se divise traditionnellement en trois catégories principales, le nom, le verbe et les particules Blachère et Gaudefroy-Demombynes (1975). La langue arabe se caractérise par une morphologie dérivationnelle très complexe, ainsi l'ensemble des mots composant la langue arabe sont pratiquement tous dérivés de racines en employant des patrons ou des gabarits Vergyri *et al.* (2004 : 1).

Il existe en arabe, environ 10 000 racines de trois, quatre ou cinq lettres. Les racines de trois lettres composent approximativement 85 % des mots de la langue De Roeck et Al-Fares (2000 : 1). L'exemple suivant illustre le cas d'une racine typique de trois lettres. Par ailleurs, nous avons inclus dans le Tableau IV, quelques formes dérivées de la racine arabe KTB.

Racine :

ك ت ب KTB = la notion d'écriture

Formes	Prononciation	Sens en français
كتاب	/kitab/	« Livre »
مكتبة	/maktaba/	« Librairie »
مكتب	/maktab/	« Bureau »
كاتب	/katib/	« Ecrivain »
كتب	/kataba/	« A écris »

Tableau IV : Illustration de quelques formes dérivées de la racine arabe ك ت ب (KTB)

De plus, l'arabe dispose d'une morphologie très riche ayant plusieurs traits qui peuvent aider dans la détection de la catégorie grammaticale. Ainsi, certains traits peuvent nous renseigner sur la partie du discours, par exemple pour distinguer le

verbe du nom. Dans l'exemple suivant, nous illustrons la segmentation d'un nom commun :

وللمكتبات /walilmaktabat/ « Et pour les librairies »
و+ل+ال+مكتبة+ات
wa+li+al+maktaba+at
Et+pour+les+librairies+pluriel

Par ailleurs, il existe des traits qui sont spécifiques au nom et d'autres au verbe comme le genre, le nombre, le temps, l'aspect à l'instar du verbe *faire* dans l'exemple suivant :

وسنفعلها /wasanaf'aluhaa/ « et on va la faire »
و+ س+ ن+ فعل+ ها
/ wa+sa+na+f'alu+ha / « et+on+nous+faire+elle »

L'observation des exemples précédents illustre l'importance de l'analyse morphologique de la langue arabe dans le cadre de la réalisation d'un système d'extraction d'information comme affirmé par Larkey *et al.* (2002).

3.2 L'absence de majuscules

Un des traits particuliers du système d'écriture arabe et des langues sémitiques en général par rapport aux langues latines est l'inexistence de la distinction entre lettres minuscules et majuscules. Ceci a une incidence particulière sur les systèmes de REN pour la langue arabe Samy (2005) puisque l'usage des lettres majuscules dans des langues comme le français ou l'anglais permet de facilement détecter les noms propres sans recourir à beaucoup d'efforts.

3.3 Les signes diacritiques

De nos jours, les signes diacritiques ou signes de vocalisation ont pratiquement disparu du texte arabe écrit Debili et Achour (1998 : 1), surtout dans les textes électroniques. Ainsi, avec l'absence des voyelles courtes, les mots peuvent être ambigus et poser d'énormes difficultés pour les systèmes d'extraction d'information (voir section 3.6). Outre le point souscrit ou suscrit obligatoire qui sert à distinguer les lettres ambiguës, l'usage a fait que les signes de vocalisation et les signes de syllabation ne sont plus employés, exception faite des textes religieux ou didactiques où on les emploie encore. Ces signes, au nombre de trois, sont appelés aussi voyelles courtes : la *damma* [u], la *fatha* [a] et la *kasra* [i]. Cette pratique a provoqué une ambiguïté très importante qui touche environ 74 % des mots de la langue.

Les voyelles courtes servent à indiquer le cas dans la flexion nominale. Les noms peuvent recevoir jusqu'à trois désinences différentes : [a], [i], et [u].

Par ailleurs, si un mot est indéfini, il prend généralement des désinences en [-an],[-un],[-in] appelé *taniwin* ou voyelles casuelles. Elles sont notées par des diacritiques spéciaux sous forme d'une double voyelle courte, par exemple la *fatha* [a], devient dans le cas du *taniwin* [aa], mais elle est prononcée [au]. Le Tableau V illustre les différents signes diacritiques dans la langue arabe.

Illustration en arabe	Type	Prononciation / fonction
Voyelles courtes		
ـُـ	damma	[u]
ـَـ	fatha	[a]
ـِـ	kasra	[i]
Voyelles casuelles (Tanwin)		
ـًـ	tanwiin fatha	[an]
ـٌـ	tanwiin damma	[un]
ـٍـ	tanwiin kasra	[in]
Signes de syllabation		
ـّـ	shadda	Doublement de consonne
ـْـ	soukoun	Absence de voyelle

Tableau V : Liste des signes diacritiques en arabe

Illustration de l'usage des voyelles courtes :

la *damma*, l'équivalent de [u], est écrit sur la lettre, par exemple :

تُ [tu]

جُ [ju]

نُ [nu]

La *fatha*, l'équivalent de [a], est écrit lui aussi sur la lettre :

تَ [ta]

جَ [ja]

نَ [na]

La *kasra*, l'équivalent de [i], est écrit sous la lettre, comme dans les exemples suivants :

تِ [ti]

ج [ji]

نِ [ni]

Comme c'est le cas pour les signes de vocalisation, les signes de syllabation ne sont pas obligatoirement écrits. Ils permettent cependant une meilleure compréhension du texte et s'utilisent parfois quand le mot n'est pas vocalisé. On distingue les deux signes suivants : le *soukoun* et la *shadda.*

Le *soukoun* est un signe diacritique représenté par un petit cercle placé au dessus de la lettre. Il peut indiquer qu'une consonne donnée ne peut être suivie d'une voyelle. Elle sert aussi à représenter les diphtongues. Ainsi, une *fatha* suivie de la lettre /و / (waw), indique la prononciation de cette lettre en /aw/.

الأوْسط /aalaaw-saT/ « le moyen ».

Enfin, la *shadda* est un signe diacritique ressemblant à la lettre minuscule [w]. Elle sert principalement à indiquer qu'une consonne est géminée, ce qui est l'équivalent d'un doublement de consonne. Elle est placée au dessus de la consonne en question. Elle est aussi employée dans les textes où les diacritiques sont absents pour limiter l'ambiguïté.

Par exemple, on note le doublement de la lettre رّ [ra] dans l'exemple suivant :

مدرسة /madrasat/ « école »

مدرّسة /mudarrisat/ « enseignante »

Il faut noter que si on enlève les voyelles courtes (a,o,u et i) des exemples précédents, il sera impossible de faire la différence entre les deux mots et ils seront affichés comme suit : /mdrst/.

Bien que traditionnellement la langue arabe se base en grande partie sur les signes diacritiques pour distinguer un mot d'un autre, un sens d'un autre, ou une fonction grammaticale d'une autre, l'usage a changé. Il est de plus en plus rare de trouver les signes diacritiques dans le texte arabe écrit, mis à part les manuels scolaires destinés à l'enseignement de l'arabe et les textes sacrés comme le Coran et la Bible. Puisque notre recherche porte sur le texte arabe écrit, il est primordial de se pencher sur cet aspect de la langue écrite pour évaluer les répercussions d'une telle pratique. Ceci va nous permettre de prendre les mesures nécessaires lors de l'implémentation du système.

Voici un cas d'ambiguïté : le mot arabe مارس /maars/ écrit sans signes diacritiques peut avoir les significations suivantes :

- un nom propre : *le mois de mars* ou *la planète Mars*, prononcé /maaris/,
- un verbe : le verbe *pratiquer* conjugué à la 2e personne du passé composé, prononcé /marasa/.

Comme on a pu le constater, le sens de ce mot dépend de la prononciation du locuteur, qui ajoute les signes diacritiques en fonction du contexte qui entoure le mot. Dans l'exemple qui précède, on s'aperçoit que lors de la lecture, le locuteur ajoute la voyelle courte /i/ et met à la fin un *soukoun* (signe marquant l'absence d'une voyelle), ce qui donne un nom propre مارسْ /maaris/ « la planète Mars ».

Par contre si le locuteur décide de faire la 2e lecture du mot en ajoutant deux voyelles courtes /a/ au milieu et à la fin, le sens du mot va changer et correspond à un verbe conjugué. Ainsi, l'absence de voyelles courtes a des incidences importantes sur la compréhension du texte écrit Debili et Achour (1998 : 1). Seul un locuteur expérimenté peut faire la distinction entre les mots en ayant recours à sa propre connaissance de la langue et au contexte du mot en question.

Cette question est particulièrement importante quant il s'agit de repérer des EN. En effet, le fait d'omettre les signes de vocalisation constitue un obstacle pour le développement d'un bon outil d'extraction d'entités nommées. Seul un développement de règles linguistiques rigoureuses pourra atténuer l'effet d'une telle pratique.

3.4 Le système numérique

En observant les textes écrits en arabe, nous avons constaté une double norme dans l'usage des chiffres. Ceci dépend principalement du pays d'origine (voir Tableau VI). En effet, dans les pays d'Afrique du Nord, les chiffres arabes standards sont employés systématiquement, alors que cet usage est différent dans la plupart des pays arabes du Moyen-Orient, de l'Égypte et de l'Arabie Saoudite où l'usage des chiffres dits « indiens » est en vigueur.

À l'inverse de l'alphabet, les chiffres s'écrivent de gauche à droite, mais se lisent tout de même de droite à gauche. Au niveau de la lecture, le nombre est lu, en commençant par la plus petite valeur; par exemple, 21 se lit un et vingt. Du point de vue du traitement informatique, nous devons tenir en compte de cette variation dans les normes des systèmes numériques et détecter la présence de toutes les variantes des chiffres employées dans le texte arabe. À défaut de préparer le système pour ces particularités, on risque d'avoir des difficultés lors de la construction des règles.

Type	Exemple
Chiffres arabes standards (Tunisie, Algérie, Maroc)	0 1 2 3 4 5 6 7 8 9
Chiffres arabes *variantes occidentales* (Égypte, Syrie, Palestine.)	٠ ١ ٢ ٣ ٤ ٥ ٦ ٧ ٨ ٩
Chiffres arabes *variantes orientales* (Pakistan, Afghanistan.)	٠ ١ ٢ ٣ ۴ ۵ ۶ ٧ ٨ ٩

Tableau VI : Les différents systèmes numériques

Les particularités que nous avons décrites dans cette section vont servir directement lors de la création du fichier de règles de repérage des entités de type numérique et temporel.

3.5 Le cas de la lettre hamza

La hamza est une lettre qui s'écrit comme un diacritique. Sa présence est motivée par des raisons historiques. D'un point de vue phonologique, elle correspond au coup de glotte / ʔ /. La lettre hamza peut s'écrire de différentes manières, seule ou avec un support, le choix de son support est dicté par des règles orthographiques :

- seule : ء
- combinée avec d'autres lettres :
    - sur et sous la lettre alif أ et إ,
    - sur la lettre waw ؤ,
    - sur la lettre yaa ئ.

De nos jours, on observe une diminution importante de l'usage de cette lettre dans certaines publications. Ceci est en partie causé par la méconnaissance des règles d'orthographe de la hamza. Par ailleurs, beaucoup de locuteurs préfèrent l'éviter afin d'éviter de produire des textes comportant des erreurs typographiques tout en sachant qu'un texte produit sans la hamza sera toujours aussi lisible par des lecteurs ayant une connaissance de base de la langue arabe de la même manière qu'ils arrivent à lire des textes sans voyelles courtes. Dans d'autres cas par contre, nous avons remarqué sa présence dans des contextes où elle est devrait être absente selon les règles d'orthographe des grammairiens de l'arabe classique par exemple يرأ /yaraa/ « voire » pour يرى /yary/ « voire ».

À cause de la complexité des règles d'écriture de la hamza (voir Tableau VII), des fautes d'orthographe sont généralement présentes dans le texte arabe écrit. Ceci est en partie causé par le manque de connaissance de l'auteur des règles de la hamza et

par souci de rapidité. En effet, le fait d'omettre les diacritiques permet un gain de temps considérable lors de l'édition d'un texte. Ceci aura cependant une incidence particulière sur le sens et la prononciation du mot, ce qui risque de gêner beaucoup les systèmes d'extraction d'information, plus particulièrement ceux de REN à cause de l'ambiguïté éventuelle que l'absence de la hamza peut créer.

Type	Graphie
La lettre alif	ا
La lettre alif avec hamza dessus	أ
La lettre alif avec hamza dessous	إ
La lettre waw avec hamza dessus	ؤ
La lettre yaa avec hamza dessus	ئ
La hamza seule	ء

Tableau VII : Illustration de l'écriture de la hamza et de l'Alif seule et en combinaisons

Nous avons prévu la présence d'erreurs de la hamza dans les textes que nous avons testés avec notre outil et un algorithme de normalisation résoudra l'ambiguïté générée par les fautes de la hamza (voir section 5.2)

3.6 L'ambiguïté dans la langue arabe

Puisque la langue arabe se distingue par sa morphologie dérivationnelle et flexionnelle très riche, l'ambiguïté est très présente. Dans ce qui suit, on présente quelques cas d'ambiguïté.

L'ambiguïté dérivationnelle

Le mot قاعدة /qaa'idat/ qui est une forme dérivée de la racine قعد /q'd/ devient une forme ambiguë avec une seule dérivation qui inclut l'ajout de la voyelle longue alif

après la deuxième lettre et l'ajout du suffixe nominal ة /t/. La forme dérivée peut donc se lire en arabe de diverses façons :

- un principe /qaa'idat/,
- une règle /qaa'idat/,
- une base militaire /qaa'idat/,
- un nom d'une organisation /aalqaa'idat/.

L'ambiguïté due à l'absence des signes diacritiques du texte écrit

L'absence des voyelles courtes dans le texte arabe est un grand facteur d'ambiguïté, comme l'illustre l'exemple du Tableau VIII :

Catégorie	Translittération	Graphie	Signification
Verbe	/bayyana/	بين	« a déclaré/démontré »
Verbe	/bayyanna/	بين	« elles [féminin] ont déclaré/démontré »
Adjectif	/bayyin/	بين	« clair/évident »
Préposition	/bayna/	بين	« entre/parmi »
Préposition	/biyin/	بين	« avec un Yen »

Tableau VIII : Exemple d'ambiguïté cause par l'absence des voyelles courtes

Les exemples en arabe cités dans cette section sont tirés de Habash (2005 : 56). Le sens du mot وجد /wjd/ peut changer selon les règles de segmentation adoptées. Même sans segmentation, il peut avoir deux sens distincts avec l'absence des signes diacritiques.

Sans segmentation

وجد :

première lecture : « il a trouvé »,

deuxième lecture : « amour ».

Avec segmentation

و+جد :

« et+grand père ».

Première lecture : conjonction de coordination /w/ attachée au nom commun /jadd/ ayant pour sens « grand père de » avec l'ajout de la voyelle courte /a/.

« et+du sérieux ».

Deuxième lecture : conjonction de coordination / w / attachée au nom commun /jidd/ ayant pour sens « du sérieux » avec l'ajout de la voyelle courte /i/.

Ambiguïtés lexicales

Nous avons remarqué que la plupart des prénoms arabes correspondent à une unité d'une autre catégorie morphosyntaxique. Ainsi, un nom propre peut correspondre à un adjectif, à un nom ou même à un verbe, ce qui peut poser des difficultés aux systèmes de repérage d'EN. Nous avons déjà vu un exemple avec le nom *Qaida* qui peut être un nom propre désignant une organisation ou un nom commun qui signifie « une règle ou un principe ».

3.7 L'ordre des mots en arabe

La phrase arabe se divise en deux types : la phrase verbale et la phrase nominale. La phrase nominale se compose en général d'un sujet et d'un attribut ; le sujet précède toujours l'attribut comme dans l'exemple qui suit :

الأولاد جائعون

/aal-awladwu/ (sujet) /jai'wuna/ (attribut) « les enfants ont faim ».

Concernant la phrase verbale, l'ordre de la phrase arabe standard obéit généralement à l'ordre VSO Al-Chartouni (1986) comme dans l'exemple suivant :

أكل الولد السلطة /akala/ (verbe) /al-waladwu/ (sujet) /assalaTa/ (objet)

Cette phrase se traduit littéralement comme suit :

a mangé l'enfant la salade,

« l'enfant a mangé la salade ».

Dans la langue arabe, si on met un mot au début de la phrase, c'est qu'il y a une intention de focaliser sur ce mot. D'un autre côté, on a généralement tendance à mettre vers la fin de la phrase, le mot qui rime le mieux ou qui soit le plus long Blachère et Gaudefroy-Demombynes (1975). Ceci peut expliquer l'existence d'autres structures mentionnées par Mahfoudhi (2002), notamment l'ordre SVO, très employé dans la forme emphatique dans l'arabe dialectal à l'instar de l'arabe tunisien :

الولد أكل السلطة /al-waladwu/ (sujet) /akala/ (verbe) /assalaTa/ (objet),

l'enfant a mangé la salade,

« l'enfant a mangé la salade ».

La structure VOS existe aussi, mais avec une fréquence moindre. Elle sert exclusivement à exprimer l'emphase sur le sujet :

أكل السلطة الولد /akala/ (verbe) /assalaTa/ (objet) /al-wladwu/ (sujet),

a mangé la salade l'enfant,

« l'enfant a mangé la salade ».

Enfin, la structure OVS qui est rare est employée parfois pour exprimer la focalisation sur le sujet :

السلطة أكل الولد /assalaTa/ (objet) /akala/ (verbe) /al-wladwu/ (sujet),

la salade a mangé l'enfant,

« l'enfant a mangé la salade ».

Le fait d'avoir plusieurs combinaisons possibles dans l'ordre des mots en arabe aura une incidence sur notre outil de repérage des entités nommées même si certaines combinaisons sont assez rares. Étant donné que l'ordre d'apparition des indices textuels est très important pour un tel système, le fait d'aborder un ordre assez libre risque de provoquer la multiplication des combinaisons de nos règles. Ce problème va contribuer à l'augmentation de l'effort consacré à la formalisation de notre grammaire locale.

3.8 Conclusion

Le fait d'avoir étudié quelques traits linguistiques de la langue arabe, ainsi que les difficultés posées par l'orthographe et la morphologie, va nous permettre d'implémenter correctement notre système de repérage des entités nommées.

Dans le chapitre 4, nous poursuivrons les descriptions entreprises dans le présent chapitre pour faire ressortir les caractéristiques linguistiques des entités nommées dans la langue arabe afin d'en faire usage dans l'étape de formalisation des règles.

4. Les entités nommées dans la langue arabe

Afin de pouvoir créer les règles de repérage des EN de manière efficace, nous avons mené une étude sur la structure des noms propres arabes. Étant donné la complexité de la structure des noms de personnes en arabe et de l'orientation actuelle du système EMM[14], l'essentiel de cette section y est consacré. Le reste des catégories sera présenté d'une manière plus brève.

Nous reviendrons par la suite sur la méthodologie de création des règles dans le chapitre suivant. Pour réaliser cette étude, nous avons constitué un corpus de 30 articles de nouvelles en ligne appartenant à l'édition arabe du site www.aljazeera.net et totalisant 22 960 mots et 480 phrases.

Nous avons essayé de choisir des articles couvrant des événements dans les différents pays arabes afin d'avoir une meilleure représentation des variantes régionales des EN. Ainsi, nous avons parcouru manuellement chaque document et nous avons systématiquement classé les entités nommées trouvées dans un classeur Excel. Pour chaque entrée nous avons inscrit le type (personne, organisations ou lieux). Nous avons pu dénombrer 1100 EN reparties entre 500 noms de personnes, 270 noms de lieux et 330 noms d'organisations.

Nous avons par la suite organisé notre classeur par type d'EN afin de pouvoir étudier la structure de chaque type. Dans ce qui suit, nous présenterons le résultat de cette étude.

---

14 L'objectif du système EMM est orienté en grande partie vers l'extraction des noms de personnes. Nous avons tout de même réalisé une recherche sur toutes les catégories d'entités nommées.

4.1 La structure des noms de personnes

Au sein de la société arabe traditionnelle, chaque individu est distingué par un ensemble de qualificatifs qui déterminent précisément son identité. Le prénom, reçu à la naissance, n'est que le premier des éléments constitutifs de son nom. Ainsi, le nom arabe est composé de plusieurs parties, dont l'ordre n'est pas systématiquement observé, et certains éléments peuvent être élidés. De plus, il n'existe pas de règles strictes stipulant la composition des noms de personnes en arabe. Mais généralement, un nom complet arabe doit être composé d'au moins un prénom et un nom de famille, donc deux mots au minimum, et peut comporter jusqu'aux six mots et plus. La composition du nom varie aussi d'un pays à un autre. Par exemple, la composition du nom dans la région du Maghreb est en général limitée à deux, trois ou quatre mots, qui sont le prénom simple ou composé et le nom de famille simple ou composé. Dans les pays arabes de la région orientale et ceux du golfe Persique, il faut généralement rajouter le patronyme.

Dans les pays du petit Maghreb[15], le nom de famille peut être un patronyme, un gentilé[16] ou un surnom. Par contre dans les pays de la région orientale, il peut être composé d'un patronyme, d'un gentilé ou simplement du nom de la famille ou la tribu, comme le cas de la famille آل سعود */Al Saoud/*, qui est la famille royale saoudienne. Dans le cas unique de la Mauritanie, les noms doivent avoir la particule ولد */Ouild/* « fils de » entre le prénom et le patronyme ou le nom famille. Les exemples suivants illustrent la situation :

(1) **Nom au Maghreb :** محمد هنداوي Mohamed Hindawi

15 Le petit Maghreb regroupe la Tunisie, le Maroc et l'Algérie. Le grand Maghreb ou tout simplement «Maghreb», comprend en plus de ces 3 pays, la Mauritanie et la Libye.

16 Un gentilé est le mot désignant les habitants d'un lieu ou d'une identité nationale ou ethnique.

(2) **Nom de la Mauritanie** : محمد ولد هنداوي Mohamed Ouild Hindawi

(3) **Nom de la région orientale :** محمد بن أحمد الهنداوي Mohamed Bin Ahmed AlHindawi

Le Tableau IX illustre l'ordre d'apparition théorique des diverses composantes d'un nom de personne complet en arabe.

Titre	Konia « Surnom »	Prénom	Patronyme	Nisba « Gentilé »	Nom de famille
/Al-imam/	/Sayf ad-Dawla/	/Mohammed/	/Abou Ahmed/	/Al Tounisi/	/Al-ahmar/
الإمام	سيف الدولة	محمد	أبو أحمد	التونسي	الأحمر

Tableau IX : Composition d'un nom de personne en arabe

Dans ce qui suit, nous illustrons les différentes composantes du nom de personne en arabe. Afin de réaliser ce travail, nous avons tout d'abord trié la liste des noms de personnes dans notre classeur. Ensuite, nous avons isolé les différentes composantes de ces noms de personnes comme l'illustre le tableau IX. Enfin, nous avons trié les entrées avant de supprimer les entrées multiples.

Cette étude nous a servi par la suite comme référence dans l'élaboration des règles de repérage des EN que nous exposerons dans le chapitre 5.

4.1.1 Titre (صفة Sifa)

Il peut s'agir d'un titre honorifique, par exemple (إمام /Imam/, شيخ /Sheikh/,سيدي /Sidi/, مولاي /Moulai/, etc.). Ces titres sont en quelque sorte l'équivalent de certains titres français comme (*Monseigneur, Mr., son altesse, Pape*, etc.).

4.1.2 Le surnom (كنية Konia)

Il s'agit d'un surnom qui peut avoir une valeur métaphorique, par exemple إبن الأرض /ibn aal aardh/ qui veut dire littéralement « l'enfant de la terre », qui est généralement employé pour qualifier quelqu'un de « grand voyageur ». La Konia peut être honorifique et se rapporte par exemple à la religion ou au pouvoir, par exemple عماد الدين /'imaad ad-din/ « le pilier de la religion », سيف الدولة /sayf ad-dawla/ « le sabre de l'état ». À ces éléments peut s'ajouter éventuellement l'indication du métier exercé, par exemple فريد الدين العطار /farid ad-din aal'attar/ « Farid Addin le parfumeur ». De plus, la Konia peut s'inspirer des pratiques dialectales comme en Égypte où on donne le surnom de أبو علي *Abou Ali* à toute personne portant officiellement le prénom de حسن Hassan.

Il existe une autre pratique dans le pays arabe de la région orientale qui consiste à nommer le père avec le prénom de son premier enfant en y ajoutant la particule أبو Abou, ainsi l'expression أبو فلان /aabwu fulaan/ « père de qqn. » où qqn. est le nom du fils ou de la fille aînée comme dans l'exemple en 1 ci-dessous. Pour une femme, cela prend la forme de أم فلان /aum fulaan/ « *mère de qqn.* » où qqn. est le nom du fils ou de la fille aînée comme dans l'exemple en 2. Ainsi ياسر عرفات *Yasser Arafat* est connu pour être أبو عمار /aabwu *eammar/* qui est sa Konia, ce qui veut dire littéralement « le père de Ammar » :

1- أبو أحمد /aabwu aaHmad/ « père d'Ahmed »,

2- أم كلثوم /aaum kulthwum/ « mère de Koulthoum ».

4.1.3 Le prénom (إسم Ism)

Le prénom ou du moins ce que nous appelons ainsi aujourd'hui, est le nom proprement dit. C'est ce qui est devenu le prénom dans les états civils de type napoléonien. Il est la seule dénomination de l'identité intime de l'individu et il peut être soit simple soit composé :

- simple : منير Mounir, وائل Wael, إلياس Ilyas, يوسف Yousef, منصف Moncef.
- Composé : عبد الله Abd Allah ; عبد الحكيم Abd al-hakim.

La plupart des prénoms arabes sont des mots arabes doués de sens. Ils signalent habituellement le bon caractère de l'individu. Par exemple, *Karim* veut dire le généreux, ce même mot est aussi un adjectif et il peut être une source de confusion dans un système de repérage d'entités nommées, lorsqu'il est assez difficile de deviner l'usage d'un mot surtout avec l'absence des voyelles courtes (signes diacritiques). Ce problème a poussé certains journaux à mettre les noms propres entre guillemets afin de lever toute équivoque. Généralement, le contexte et la syntaxe permettent de lever l'ambiguïté.

Exemple d'un nom : أسامة Oussama, « type de lion »,

exemple d'un adjectif : كريم Karim, « le généreux »,

exemple d'un verbe : يزيد Yazid, « augmenter », au présent, conjugué à la 3ème personne du singulier.

Il serait intéressant d'observer la structure du prénom à travers les variantes sociales et régionales. Ainsi, beaucoup d'Arabes de confession musulmane ont une tendance à choisir des prénoms ayant la combinaison de la particule عبد abd suivi d'un autre mot : /abd / qui signifie « esclave de X » où X est un mot décrivant الله Allah « le nom employé pour désigner Dieu », souvent un des 99 attributs de Dieu. Par exemple le prénom عبد الله /'abd-aallah/ « l'esclave de Dieu ».

De plus, on observe une divergence dans l'attribution des prénoms selon le groupe religieux auquel on s'identifie. Par exemple, les musulmans sunnites emploient fréquemment des prénoms comme محمد Mohamed, أحمد Ahmed, عيشة Aicha ou عمرOmar. Par contre, les musulmans Shiite ont une préférence particulière pour des prénoms comme علي Ali, حسن Hassan, حسين Houssein ou فاطمة Fatma.

D'un autre côté, les Arabes de confession chrétienne ont tendance à utiliser des prénoms s'inspirant de la Bible. Par exemple les prénoms des apôtres comme Jean ou Paul ou des prénoms d'origine européenne et particulièrement française (dans la région du Levant[17]). Citons par exemple des prénoms comme celui de جورج حبش Georges Habash ou كامي شامون Camille Chamoun.

4.1.3.1 Les cas des prénoms composés

Suite à l'observation des prénoms composés en arabe dans notre corpus, nous avons constaté qu'il existe une certaine régularité dans la composition de ces derniers comme dans les exemples suivants :

بن لادن **Bin** Laden / إبن بطوطة **Ibn** Batouta / آية أحمد **Ait** Ahmed

Il existe une liste finie de particules qui se trouvent à l'initiale de ces prénoms comme (بن Bin, إبن Ibn, آية Ait, أم Um, عبد Abd, ولد Ouild.). Nous avons compilé une telle liste afin de nous en servir lors de la création des règles. Il existe par ailleurs des cas de prénoms où la particule fait partie intégrante du mot comme dans les prénoms suivants :

أية علوان **Ait**alouane / بن جراد **Bin**jrad /إبن بشير **Ibn**bachir

Ainsi, nous avons créé des règles spécialement conçues pour détecter ces particules.

4.1.3.2 La formation des prénoms en arabe

Certains noms de personnes en arabe se forment en suivant des patrons bien précis. Le fait de connaître ces patrons peut contribuer à leur détection. Dans le Tableau X, les trois prénoms sont formés suivant des patrons réservés généralement pour les noms propres.

---

17 On parle d'États du Levant pour désigner tout particulièrement le Liban, la Syrie, la Palestine et la Jordanie.

Racines	Patrons	Exemples
K.T.B	C1a-C2i-C3	KaTiB
HMD	C1a-C2i-C3	HaMiD
SDM	C1a-C2-C3a-C4	SaDdaM

Tableau X : Exemples de patrons pour les prénoms arabes

Les lettres en majuscules (Ci), désignent les consonnes de base de la racine. Les lettres (a, i) désignent les voyelles et les consonnes en minuscules sont en fait des consonnes de dérivation comme le d dans صدام Saddam.

4.1.4 Le patronyme ou nom de filiation (نسب Nasab)

Il s'agit du prénom d'un grand parent d'une personne donnée qui se compose du mot إبن Ibn ou بن Ben « fils de » ou بنت Bint « fille de », et du prénom du père comme dans إبن عبد العزيز Ibn Abd Alaziz « le fils de Abd Alaziz » ou بنت محمد Bint Mohammed « la fille de Mohammed ».

Le prénom de la mère est plus rarement mentionné comme dans إبن أم مكتوم Ibn om maktoum « fils de la mère de Maktoum ». Notons le cas du prophète عيسى Issa « Jésus » mentionné de nombreuses fois dans le Coran sous le nom de عيسى إبن مريم Issa Ibn Mariam « Jésus fils de Marie ».

L'emploi de deux نسب Nasab successifs, par la mention du nom du père puis celui du grand-père, permet de déterminer plus sûrement une identité ; exemple إبن إدريس إبن العباس Ibn Idris Ibn Al-Abbas « Fils d'Idris lui-même fils de Abbas ».

4.1.5 Le nom d'origine (نسبة Nisba)

La Nisba, est en fait une extension de la définition des gentilés. Elle inclut entre autres : le lignage ou l'origine sociale de la personne, elle est généralement sous forme adjectivale se référant à un pays ou à une ville, par exemple البغدادي *Al-Baghdadi*, est celui originaire de Baghdâd et الزغواني *Alzaghouani* qui est originaire de la ville de Zaghouan. Nous nous attarderons un peu sur cette classe des noms

propres qui est très productive en arabe. Elle suit la plupart du temps une logique bien précise.

Tout d'abord, les Nisbas doivent avoir la lettre ي /Yaa/ accompagnée de la voyelle courte /i/ vers la fin du mot, ce qui indique sémantiquement l'appartenance du prénom X à l'endroit Y. Parfois l'article défini /al/ précède la Nisba comme dans l'exemple suivant :

(1) صدام حسين التكريتي Saddam Hossein (X) al-Tikriti (Y).

Il existe donc, une relation d'appartenance de X vers Y التكريتي /aal-tikryiti/ veut dire celui venant de la ville de Tikrit. Les Nisbas sont créées à partir de noms communs qui peuvent en général exprimer une certaine affiliation ou une appartenance à un pays, à une religion ou à un groupe ethnique bien déterminé. Elle peut aussi se retrouver sous forme d'adjectif exprimant le métier de la personne. Par ailleurs, il existe des règles de formation de la Nisba qui sont plus complexes comme dans les cas suivants :

le cas des noms communs composés de deux ou trois lettres comme dans les deux exemples suivants :

(1) le nom commun حي /Hay/ « vivant » se transforme en la Nisba حيوي /*Hayawi*/ « le vivant ».

(2) Le nom propre علي /'alyi/ se transforme en la Nisba العلوي /*Alawi*/ « celui qui appartient à la secte des Alaouites », avec l'ajout de la lettre /w/ et la voyelle courte /i/.

Le cas où la Nisba se dérive des noms composés comme dans l'exemple suivant :

(1) le prénom أبو بكر /aabu bakr/ se transforme en la Nisba أبو بكري /aabu bakr**yi**/ « celui qui appartient à la tribu du calife Abou Bakr », avec l'ajout de

la lettre ي /Yaa/ et de la voyelle courte /i/. Le cas où la Nisba est dérivée d'un nom de lieu composé, on assiste à un phénomène d'ellipse avec la disparition du deuxième mot. Par exemple, les noms des lieux composés comme حضر موت /Hadhra maw-t/ qui devient simplement حضري /Hadhryi/ « celui qui habite Hadhramaout ».

4.1.6 Le nom de famille (لقب Laqab)

Il s'agit d'un mot attribué à une famille pour la distinguer des autres familles. Dans la langue arabe, اللقب le Laqab réfère généralement à une classe sociale ou simplement à une description physique ou morale d'une famille donnée. Par exemple le nom de famille الأحمر /aal-aHmar/ qui veut dire « le rouge » ou جاوحدو /jaawaH-du/ qui signifie « celui qui est venu tout seul ».

4.1.7 Les emprunts dans les prénoms arabes

Vu l'importance de la distribution géographique des pays arabes et leurs liens avec les pays musulmans non arabophones, plusieurs noms de personnes trouvent leurs origines dans la langue perse (voir Tableau XI) ou la langue turque (voir Tableau XII).

Prénom arabe après emprunt	Signification en perse
Jihan جيهان	« le monde »
Nevin نفين	« moderne »
Shahinez شاهيناز	« la fierté du roi »
Nermine نرمين	« douce »

Tableau XI : Prénom d'origine perse

La langue turque, à son tour, a influencé la langue arabe à travers la présence de l'Empire ottoman dans les pays arabes. Nous retenons l'exemple du suffixe turque /–ji/ employé dans la langue arabe pour exprimer un adjectif de métier. Le mot

arabe قهوة /qahwat/, qui veut dire café, est devenu قهواجي /qahwaji/, qui est le nom généralement attribué aux gérants des cafés.

Prénom arabe avant emprunt	Emprunt turc	Signification
Hikma حكمة	Hikmatt حكمت	« la sagesse »
Marwa مروى	Mirvatt مرفت	« le nom d'une montagne »
Niimah نعمة	Niimatt نعمت	« bénédiction »

Tableau XII : Prénom d'origine turc

4.1.8 Les noms propres d'origine étrangère

La plupart des noms propres d'origine étrangère sont transcrits en arabe simplement sur une base phonologique, c'est ce qu'on appelle la translittération. Par exemple, *Georges W. Bush* sera translittéré par جورج دبليو بوش /jwurj dabilyu bush/ et *Monica Seles* par مونيكا سيليش /monyika syilyish/. Toutefois, il faut noter qu'il n'existe pas toujours de correspondance entre les lettres qui composent le mot étranger et les lettres de l'alphabet arabe, puisque dans la langue arabe, il n'y a pas d'équivalent pour les lettres P, V ou G, ainsi *Paul* devient بول /bwul/, *Valérie* devient فاليري /falyiryi/ et *Gordon* devient جوردن /Jwurdwun/.

Il y a aussi une différence notable dans la translittération vers l'arabe des noms propres étrangers qui dépend en partie des connaissances linguistiques du traducteur. Suite à une requête avec le moteur de recherche Google, nous avons pu constater que la translittération en arabe du nom de l'ex-premier ministre français, *Édouard Balladur* s'écrit en arabe de deux manières différentes puisque le phonème français /u/ n'a pas un équivalent direct en arabe. La variante إدوار بلادير /'id-war baladyir/ a été trouvée 64 fois. Par contre, إدوار بلادور /'id-war baladwur/ a été trouvée 1200 fois. Par ailleurs, nous avons remarqué que la prononciation de lettre finale /d/ du mot *Édouard*, a été omise dans plusieurs cas d'une manière arbitraire. On ose ainsi avancer qu'un locuteur natif de l'arabe et ayant le français comme

langue seconde ou ayant une connaissance suffisante de la phonologie du français va omettre la lettre d, ce qui est en accord avec la prononciation française du prénom *Édouard* إدوار /`id-war/. Par contre, un arabe ayant vécu dans un milieu anglophone aura tendance à prononcer la /d/ finale à l'instar de la prononciation anglaise du prénom *Edward* إدوارد / `id-ward /.

Enfin, il existe des noms propres étrangers qui ne sont pas translittérés étant donné qu'ils ont une valeur historique particulière et qu'il existe déjà un équivalent dans la langue arabe. Par exemple, les noms des prophètes bibliques comme *Jésus* qui se traduit par son équivalent arabe عيسى /'yisa/. Nous avons ajouté ce type de noms propres à notre lexique des noms propres d'une manière systématique.

4.1.9 L'ambiguïté des noms propres en arabe

Puisque les noms propres arabes sont la plupart du temps porteurs d'un sens particulier, il arrive souvent qu'ils présentent une ambiguïté que seul le contexte permet de résoudre. Par exemple, le nom de l'ex-président syrien حافظ الأسد *Hafez Al-Assad* veut littéralement dire « le protecteur du lion ». Ainsi même un locuteur natif méconnaissant cette personnalité politique peut tomber dans l'erreur et juger qu'il ne s'agit pas d'un nom propre, mais un syntagme nominal. Un autre exemple est celui du président égyptien حسني مبارك *Hosni Moubarak*, où le mot مبارك /mwubaarak/ signifie « le béni ».

Afin de résoudre partiellement cette ambiguïté, il faut étudier le contexte immédiat du nom propre et vérifier l'existence des marqueurs lexicaux qui peuvent précéder ou suivre ce dernier. Ainsi le prénom حسني *Hosni*, qui n'est pas ambigu en lui-même, permet d'écarter le doute et de confirmer que مبارك *Moubarak* est aussi un nom propre. Par contre dans le cas ou les éléments constituants les noms propres sont ambigus, il faut vérifier l'existence des marqueurs lexicaux de la catégorie titre, par exemple le président حافظ الأسد *Hafez Al-Assad*, le titre *président* permet de lever l'ambiguïté dans ce cas. Toutefois, il est très fréquent qu'un prénom soit

dépourvu des marqueurs lexicaux. Dans ce cas, il sera difficile de le détecter par des règles simples. Nous reviendrons en détail sur cette question quand nous parlerons de la création des règles et des difficultés rencontrées dans le chapitre 5.

4.2 La structure des noms de lieux

Les noms de lieux comprennent plusieurs catégories telles que les villes, les pays, les villages, les montagnes et les fleuves. Dans la langue arabe, les noms de lieux proviennent de diverses langues comme le français, l'anglais, le turc, etc.

La liste des noms de lieux connus et existants dans le monde est généralement une liste relativement stable dans la mesure où les noms de lieux ne changent pas souvent. Toutefois, à l'instar des noms de personnes, certains noms de lieux sont ambigus. Voici deux exemples qui illustrent ces propos :

- le mot *Tunisie* en arabe تونس /twunis/ ou *Algérie* الجزائر /aaljazayir / qui peuvent désigner le pays ou la capitale (la distinction existe en français entre la Tunisie et Tunis et l'Algérie et Alger),

- le mot *Maroc* en arabe المغرب /aalmaghrib/ désigne soit le pays qui est le *Maroc* soit la grande région du Maghreb en Afrique du Nord. Parfois afin de lever l'ambiguïté, on appelle le *Maroc* المغرب الأقصى /almagh-rib alaq-sa/ qui signifie l'occident lointain et المغرب الكبير /almagh-rib al-'arabi/ pour désigner la région entière du Maghreb.

Les exemples ci-dessus n'ont pas de répercussions majeures sur les performances d'un système de repérage des entités nommées puisque l'ambiguïté reste tout de même entre deux catégories de noms de lieux. Par contre, la tâche se complique quand un nom de lieu dispose de plusieurs significations n'ayant aucun rapport les unes avec les autres.

Par exemple, le mot *Yémen* en arabe اليمن /aalyaman/ peut être soit le pays, soit le nom commun signifiant la chance.

La traduction du sens des noms des pays étrangers en équivalents arabes est une pratique courante et il est important de prévoir l'effet de cette pratique sur notre outil de repérage des entités nommées. Nous pouvons citer les trois exemples suivants :

- les États Unis الولايات المتحدة /aalwilayat almwuttahidat/, qui signifie littéralement « les états qui sont unis »,

- les Pays-Bas الأراضي المنخفضة /aalaaraaDyi aalmunkhafiDat/, qui signifie « les basses terres », est utilisé en alternance avec le mot هولندا /hwulandaa/, qui reste le plus fréquemment employé,

- le Cap-Vert الرأس الأخضر /aalraa's aalaa'kh-dhar/ qui signifie littéralement « la tête verte ».

4.3 La structure des noms d'organisations

Cette catégorie d'entités nommées inclut les noms des gouvernements, des compagnies et de toutes les entités organisationnelles physiques ou morales. Les noms d'organisations sont assez nombreux et sont difficilement quantifiables puisque leur apparition et leur disparition dépendent de la situation dans le monde. Il arrive aussi qu'une organisation alterne entre l'usage d'une forme longue et d'une forme courte de son nom. Par exemple منظمة الأمم المتحدة /mwunaZamat aalaaumam al muttahida/ « Organisation des Nations Unies » qui est une forme longue, peut exister dans un autre texte avec une forme plus courte comme الأمم المتحدة /alaaumam aalmuttahida/ « les Nations Unies ».

Dans la langue arabe, comme dans la plupart des langues, les noms d'organisations peuvent être soit simples (un seul mot), soit complexes (deux mots ou plus). En outre, l'usage des acronymes en arabe est relativement faible si on le compare aux langues européennes comme le français et l'anglais.

En fait, les acronymes constituent environ 1 % de l'ensemble des entités nommées trouvées dans notre corpus. Ceci est dû au fait que la langue arabe n'a jamais adopté cette forme de réduction des noms propres. Dans les rares exemples trouvés, il s'agit plutôt d'acronymes étrangers. Le rédacteur arabe, se contente souvent de transcrire phonétiquement l'acronyme comme dans le cas de l'acronyme *CIA* qui s'écrit en arabe سي آي آيه /syi aay aah/.

Les formes à reconnaître sont plus variées et plus complexes que celles des noms de personnes puisqu'il existe des conventions régissant la composition des noms de personnes en arabe. Par contre, les noms d'organisations en arabe sont souvent choisis d'une manière arbitraire et n'obéissent pas à des règles strictes. Ainsi, n'importe quel mot peut théoriquement faire partie de cette catégorie, cependant, dans la plupart des cas observés, les noms d'organisation se composent de noms communs, d'adjectifs, de noms propres et de mots étrangers.

Les noms d'organisations de type simple (composé d'un seul mot), peuvent être sous forme d'un nom de personne ou de n'importe quel nom propre, par exemple تبريد *Tabrid,* اعمار *Iimar,* بارك *Barak*, etc.

Par ailleurs, les noms d'organisations étrangères se composent souvent d'un mot introducteur en arabe comme *compagnie* الشركة /aalsharika/, *organisation* المنظمات /aal munaZamat/, *association* الجمعيات /aaljam-'iiat/ et le reste des éléments de l'entité sont simplement sous forme d'une translittération simple du nom étranger comme dans la *Société Business United* qui sera traduite en arabe par شركة بيزنيس يونايتد /sharikat biz-nis yunay-tid/.

En observant la liste des noms d'organisations originaires des pays arabes, nous avons parfois constaté l'insertion de mots d'origine anglaise ou française dans la composition de l'entité. Par exemple *Ras Al Khaimah Ceramics* رأس الخيمة سيراميكس /raasu aal khay-ma siramiks/ ou Mohamed Ben Abdallah Pièces Auto محمد بن عبد الله بياس أوتو /muhammad bin abdaalaah biyaas aautwu/. Ce type de combinaison risque de compliquer la tâche des systèmes de repérage des EN vu l'irrégularité et l'inconsistance dans l'usage de l'alternance codique et de l'emprunt dans l'arabe écrit.

Enfin, il existe des noms de compagnies qui peuvent être confondus avec des noms de personnes, puisque le nom de la compagnie est formé simplement du nom et du prénom d'une personne. Par exemple, *Mona Ibrahim* est un nom d'un magasin de vêtement. Le Tableau XIII présente d'autres exemples de noms d'organisations.

Modèle du nom propre d'une organisation	**Exemple avec translittération de l'arabe**	**Traduction littérale**
Nom de personne	/muna ibrahyim/ منى إبراهيم	Mona Ibrahim
Nom de personne + type de profession en anglais	/aal madanyi tay-lwurz/ المدني تايلورز	Les tailleurs Al Madani
Nom commun simple	/tab-ryid/ تبريد	Refroidissement
Nom de personne + type de produit	/Raas aal khay-mat siramyiks/ رأس الخيمة سيراميكس	Les céramiques Ras Al Khaima
Nom composé complètement en arabe	/mataar dubay aal dualyi/ مطار دبي الدولي	L'aéroport international de Dubaï
Nom de personne + type de produit	/Muhamm-id dawud biyaas aauTu/ محمد داوود بياس أوتو	Mohamed Daoud pièces auto
Usage de l'arabe et l'anglais en même temps	/sharikat habbat aal barakat aand kwu/ شركة حبة البركة و كو	La société Habbat al baraka & compagnie

Tableau XIII : Illustration de quelques noms d'organisation en arabe

4.4 Les entités temporelles et numériques

4.4.1 Les entités temporelles

Les entités temporelles incluent les dates (ex. : *le 14 mars*) y compris les jours de fête religieuse ou officielle (Jour de l'an, Achoura, Noël, etc.) et tout autre expression exprimant le temps (ex. : *hier, la semaine prochaine, dans quelques instants, midi, une heure et demi*). La plupart des entités temporelles dans la langue arabe sont identifiables grâce à une liste de marqueurs lexicaux comme par exemple *jour, mois, année*, etc. Concernant les dates, il existe une différence dans l'usage des calendriers qui varient d'un pays arabe à un autre. Dans notre corpus de travail, la plupart des dates étaient affichées en chiffres arabes et parfois en toutes lettres.

Ex. : *12 يونيو yuniyu / حيزران huzayran 1989 « 12 Juin 1989 »*

Dans l'exemple ci-dessus du journal en ligne الجزيرة *Al-Jazeera,* le jour est écrit en chiffres arabes suivi du mois en format du calendrier solaire syriaque et grégorien, suivi de l'année en chiffres arabes. Ce système, qui est relativement simple, est assez courant. Cependant, chaque journal peut choisir sa propre formule de datation parmi les formules possibles. Il faudra donc savoir repérer toutes les variations possibles pour les dates selon les trois principaux calendriers que nous présenterons dans la suite. Nous tiendrons aussi compte de deux systèmes de numération, sans oublier que les chiffres peuvent être écrits en toutes lettres.

4.4.1.1 Le calendrier grégorien

Dans ce calendrier, les noms des douze mois sont transcrits en lettres arabes. Les chiffres correspondant aux jours et aux années peuvent figurer dans les deux systèmes d'écriture, arabe ou indien. Ce système est particulièrement employé dans les pays du Maghreb.

Ex. : *2007 نوفمبر 01 – 01 من نوفمبر /min nwufam-bar 2007/ « 01 novembre 2007 »*

4.4.1.2 Le calendrier solaire syriaque

Dans ce système, les douze mois correspondent aux mois du calendrier grégorien, mais le premier mois de l'année est novembre et le dernier est octobre. Même si la tendance est d'utiliser le système grégorien, ce calendrier est beaucoup utilisé, surtout dans les pays arabes du Moyen-Orient.

Ex. : *2007 تشرين الثاني 02 - 02 /tishryin aalthaanyi 2007/ « 02 novembre 2007 »*

4.4.1.3 Le calendrier lunaire musulman

Ce calendrier est basé sur la datation musulmane de l'Hégire (date correspondant à la migration du prophète محمد *Mohammed* de Médine vers La Mecque). Il débute en 622 après J.-C. qui est l'an 1. Ce calendrier comprend 12 mois qui sont utilisés surtout dans des contextes religieux ou historiques comme, par exemple, le mois de Ramadan. Son usage n'est pas rare et il arrive souvent de trouver deux types de calendrier dans un même texte. Le calendrier musulman est le calendrier officiel du Royaume d'Arabie saoudite.

Ex. : *21 من شوال 1428 هـ 21 -21 /min shawaal 1428 H/ « 21 de Shawal 1428 du Hégire »*

4.4.1.4 Les jours de la semaine

Les jours de la semaine dans la langue arabe sont habituellement précédés par le mot يوم /yawm/ qui signifie « jour ». De plus, l'article défini /Al/ doit être obligatoirement collé au nom du jour (voir Tableau XIV).

	Graphie	Translittération	Signification
Le mot يوم */yawm/* « jour » précède le nom du jour .	الاثنين	Al-Ithnain	Lundi
	الثلاثاء	Al-Thoulathaa	Mardi
	الاربعاء	Al-Irbouaa	Mercredi
	الخميس	Al-Khamis	Jeudi
	الجمعة	Al-Joumouaa	Vendredi
	السبت	Al-Sabt	Samedi
	الأحد	Al-Ahad	Dimanche

Tableau XIV : Illustration des jours de la semaine en arabe

4.4.2 Les entités numériques

Les entités numériques incluent principalement les systèmes de mesures (poids, distance, volume, vitesse), les pourcentages, ainsi que les devises. La liste des entités numériques peut être plus longue selon les définitions ; pour les besoins de ce mémoire, nous nous contenterons des trois principales entités numériques, qui sont les systèmes de mesures, les devises et les pourcentages.

4.4.2.1 Les unités de mesure

Quand les unités de mesure sont employées dans un texte en arabe, l'usage des abréviations est systématique. Dans ce cas, on peut trouver des expressions comme : *dix kilos -10* كيلو */kilwu /, 3 tonnes -3* طن */Tonn /, 25 cm -25* صم */Sm/*. Par exemple, la course de 100 mètres est transcrite en arabe par سباق مئة متر / *sibaaq mi'at mitr/.*

4.4.2.2 Les pourcentages et les devises

L'usage des pourcentages et des expressions monétaires se caractérise par l'emploi d'un signe particulier, le signe du pourcentage % dans le premier cas et les symboles monétaires pour les devises : $ pour le *dollar*, € pour l'*euro,* ¥ pour le *Yen*. Il arrive aussi que les symboles monétaires soient omis pour être remplacés par

l'équivalent en arabe, qui est simplement une translittération du mot latin. Ainsi, l'usage dans la langue arabe des systèmes de mesures, des pourcentages ou des devises suit une convention et des règles d'écriture bien établies. Cette approche facilite la formulation des règles de repérage pour cette catégorie d'entités nommées.

4.5 Conclusion

Puisque la description linguistique des règles d'extraction des EN nécessite un formalisme bien particulier, nous avons étudié dans ce chapitre les divers aspects des entités nommées dans la langue arabe.

Une grande partie de ce chapitre a été réservée aux noms de personnes pour leur importance dans la langue arabe d'une part et afin de respecter les directives de notre projet d'une autre part, mais sans oublier le reste des catégories non moins complexes comme les noms d'organisations ou les noms de lieux ainsi que les entités temporelles et numériques.

Dans le chapitre suivant, nous présenterons le système EMM ainsi que l'architecture de l'outil de repérage des entités nommées que nous avons développé, tout en montrant les différentes étapes de notre méthodologie.

5. Méthodologie

Lors du stage qui s'est déroulé au Centre commun de recherche (CCR) en Italie et plus précisément au groupe Langtech-EMM, nous avons été chargés d'adapter le module de repérage des entités nommées du système de veille EMM à la langue arabe.

L'originalité de notre approche réside dans le fait qu'il s'agit, au moment de sa parution[18], du premier système de repérage d'EN pour la langue arabe entièrement à base règles, se basant sur un lexique et n'intégrant aucune approche statistique (Apptek 2009 ; Basistech 2009 ; Benajiba *et al.* 2007) ou d'apprentissage automatique (Inxight 2009).

À l'image des autres systèmes de repérage des EN à base de règles mentionnés dans la section 2.1.2, le repérage des EN avec notre système est fondé principalement sur un lexique sous forme de dictionnaires et sur un ensemble de règles de repérage sous forme d'expressions régulières faites à la main, d'où la nécessité de construire un lexique ciblé spécifiquement pour l'extraction des divers types d'EN ainsi que pour la création de règles qui permettent l'usage du lexique et les indicateurs textuels pour le repérage des EN.

Étant donnée la complexité de la langue arabe que nous avons présentée dans la section 3, un prétraitement lexical spécialement dédié à la segmentation et à la normalisation du texte arabe était rendu nécessaire.

Dans ce qui suit, nous présentons le système EMM ainsi que l'architecture du module de repérage des EN pour la langue arabe, désormais baptisé RENAR. RENAR est un acronyme de (**R**epérage des **E**ntités **N**ommées **AR**abes); cet acronyme a été choisi afin de faciliter la lecture dans ce mémoire. Par la suite, nous

18 La première mise en ligne de RENAR a été effectuée le 10 septembre 2005.

présenterons la méthodologie suivie pour l'implémentation des différentes composantes de RENAR.

5.1 Présentation du système EMM

Le système de veille médiatique EMM parcourt quotidiennement une moyenne de 25 000 articles en 30 langues différentes provenant de plus de 900 sites Web. Il permet de regrouper les articles couvrant le même sujet, fusionnant ces articles automatiquement en un seul groupe afin d'éviter la redondance des nouvelles en n'affichant qu'un seul événement pour plusieurs articles traitant le même sujet.

Le système fonctionne comme un système d'information et d'alerte médiatique en temps réel permettant d'avoir un aperçu quotidien des nouvelles à travers le monde, tout en affichant les informations sous un format lisible et permettant l'accès rapide au contenu pertinent d'un article donné. Nous avons inclus une capture d'écran de l'environnement EMM dans l'Annexe II.

Derrière le système EMM il existe une panoplie d'outils et de modules, comme les modules de représentation graphique, les outils d'analyse statistique ou le module de repérage des EN. Dans la suite de ce mémoire, nous nous intéresserons exclusivement au module de repérage des EN qui a été adapté avec succès à huit langues.

5.2 Description de l'outil RENAR

L'outil RENAR fait partie du module de repérage des entités nommées du système EMM. Il s'agit du module de repérage des EN à base de règles que nous avons créé spécifiquement pour la langue arabe. Une capture d'écran montrant l'environnement EMM après l'intégration de RENAR est incluse dans l'annexe III.

Dans ce qui suit, nous présenterons l'architecture et le fonctionnement de cet outil. Un diagramme représentant l'architecture de RENAR est présenté dans la figure 5.

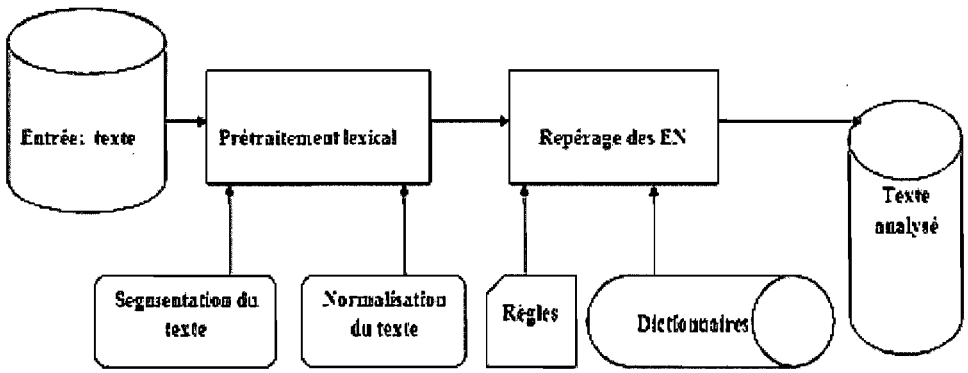


Figure 5 : Architecture de RENAR

Entrée

Lors de cette étape, le texte brut est saisi en entrée.

Le prétraitement lexical

Il s'agit de l'étape préalable au processus de repérage des EN par le système RENAR. Le prétraitement lexical permet de préparer le texte brut pour son analyse linguistique. La phase de prétraitement sous RENAR se fait en deux étapes : d'abord, la segmentation du texte en phrases, ensuite, la normalisation de son orthographe.

La segmentation est donc la première phase où le système effectue le découpage du texte en phrases et en mots en se servant des règles de segmentation basées sur la ponctuation et la morphologie comme l'illustre l'exemple suivant à travers les différentes transformations :

وللمكتبـــــات /walilmaktabat/ « et pour les librairies ».

Transformation 1 : suppression de la conjonction de coordination و wa « et ».

للمكتبـــــات /lilmaktabat/ « pour les librairies ».

Transformation 2 : suppression de la préposition ل /li/ « pour » et de l'article défini ال / aal / « le ».

مكتبـــــات /maktabat/ « librairies ».

Transformation 3 : suppression de la Kashida[19] « __ ».

مكتبات /maktabat/ « librairies ».

Ensuite vient l'étape de la normalisation du texte qui sert à uniformiser certains mots ayant une orthographe proche, soit parce qu'il s'agit d'erreurs de transcription assez courantes, soit que la convention d'écriture choisie soit propre à une région.

C'est donc pour des raisons pratiques que nous avons instauré ce module qui comprend plusieurs règles inspirées de Shaalan et Hafsa (2008). Par exemple, la normalisation de la lettre Alif-hamza en une seule forme d'alif sans la hamza comme dans l'exemple suivant :

أبي /'aabi/, « mon père » devient ابي /aabi/ « mon père »

Une fois l'étape de prétraitement lexical terminée, le texte prétraité doit passer par l'étape de repérage des EN.

Le repérage des EN

19 La Kashida n'est pas une lettre de l'alphabet arabe. Il s'agit d'un allongement graphique de certaines lettres de l'alphabet qui sert notamment dans les textes poétiques.

L'opération d'extraction des EN se fait entièrement lors de cette étape. L'opération de repérage des EN avec l'outil RENAR est divisée en deux étapes.

La première étape est basée sur la consultation directe du lexique qui se compose de plusieurs dictionnaires. Lors de cette étape, l'outil commence par la comparaison de chaque entrée dans le texte brut avec chacune des entrées des différents dictionnaires que nous avons construits (voir section 5.3.2).

Une fois une EN reconnue grâce à un dictionnaire, elle sera automatiquement retenue sans passer par la deuxième étape, qui est réservée exclusivement à la détection des EN ne figurant pas dans le lexique.

La deuxième étape repose sur des fichiers de règles écrites à la main sous forme d'expressions régulières qui permettent de détecter les EN grâce aux dictionnaires et à la liste des marqueurs lexicaux qui sont des indices textuels permettant de localiser les EN dans un texte (voir section 5.3.3).

Sortie

Les EN, trouvées à la fin de la première et la deuxième étape, sont sauvegardées dans un fichier de sortie avant leurs affichages par l'interface du système EMM.

5.3 La création du lexique et des règles

Dans cette section, nous présentons l'ensemble des outils que nous avons employés afin de pouvoir automatiser certaines procédures lors des différentes étapes de la création du lexique et des règles. Ensuite, nous présenterons la démarche suivie pour la création du lexique et des règles.

5.3.1 Les outils

5.3.1.1 L'éditeur Textpad

Nous avons choisi cet éditeur de texte pour réaliser les différentes tâches liées à la manipulation des textes, les recherches, la formalisation des règles et l'écriture du code du langage de programmation PERL. Nous avons installé Textpad sur une station Microsoft Windows XP version arabe afin d'avoir une meilleure prise en charge de la langue arabe. Textpad permet d'éditer des documents de très grande taille, ce qui constitue un avantage lors de la création et de l'édition de fichiers volumineux comme ceux des dictionnaires. Textpad inclut aussi des fonctionnalités pratiques comme la possibilité d'insérer des balises pour accéder plus rapidement à une zone déterminée dans le texte, ce qui permet aux fonctions d'édition d'être applicables uniquement aux lignes sélectionnées.

Textpad a été d'un apport considérable dans l'automatisation de plusieurs tâches spécialement avec la fonctionnalité Macro qui permet d'enregistrer une série d'actions, ce qui nous a permis d'éviter de faire manuellement plusieurs tâches répétitives d'édition.

5.3.1.2 Le langage PERL

PERL, qui signifie en anglais *Practical Extraction and Report Language*, est un langage créé par Larry Wall en 1987. PERL est bien connu pour ses expressions régulières qui font partie intégrante du langage, ce qui offre une grande facilité de manipulation. Il s'agit d'un langage très pratique surtout pour des tâches de manipulation de texte. C'est pour cette raison que PERL a été choisi pour faire partie du système EMM. Par ailleurs, l'opération de repérage des EN est entièrement gérée par un script écrit en PERL.

5.3.1.3 Le logiciel HTTrack

Le logiciel HTTrack est un aspirateur de pages Web. Il s'agit d'un outil permettant de copier tout le contenu d'un site Web sur le disque dur, en récupérant la structure originale du site et tous les fichiers HTML, images, sons, etc. Nous avons choisi le logiciel HTTrack puisqu'il est gratuit et qu'il est entièrement configurable. Nous avons surtout employé cet aspirateur afin de télécharger des pages Web de différents sites arabes afin d'observer la structure des noms propres dans les différentes régions et les différents dialectes du monde arabe.

5.3.1.4 Le concordancier AConCorde

AConCorde est un concordancier, il s'agit d'un logiciel indispensable aux travaux sur corpus. Il a pour fonction principale la recherche de mots dans un corpus et l'affichage du contexte d'apparition de ces mots avec leur fréquence. Cette opération est plutôt connue par le terme anglais KWIC (key word in context). L'utilisation d'un concordancier permet d'extraire les concordances des mots selon les spécifications et les paramètres de la recherche.

Pour les besoins de notre travail, nous avons choisi le logiciel AConCorde version 0.4 pour sa prise en charge du standard Unicode et de la langue arabe, (voir Figure 6). Le fait de travailler avec un concordancier nous a permis de faire des recherches sur les contextes d'apparition des noms propres dans notre corpus. Ceci nous a permis de mieux formaliser nos règles de repérage.

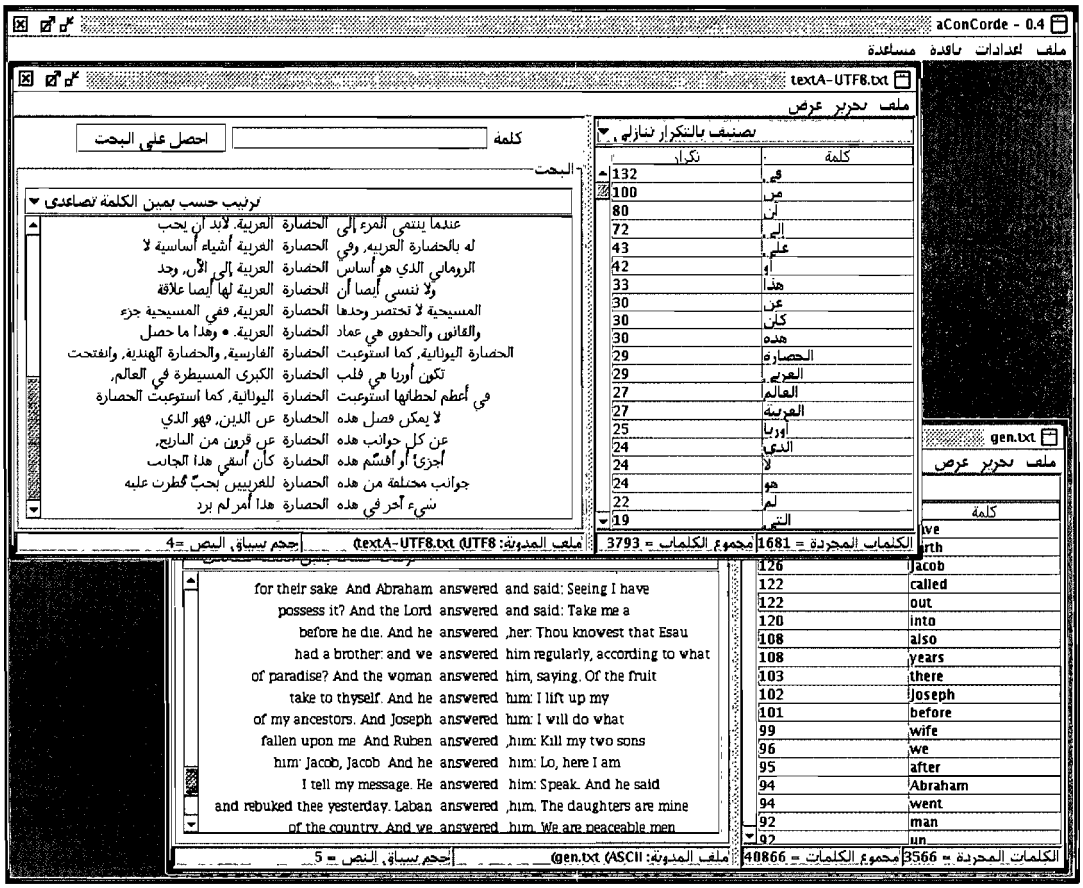


Figure 6 : Illustration du logiciel AConCorde 0.4

5.3.2 La création et l'implémentation du lexique

Dans cette section, nous présentons notre lexique ainsi que la méthodologie suivie lors de sa création. Le lexique que nous avons créé est reparti entre plusieurs dictionnaires selon la catégorie de l'entité nommée à extraire. Ainsi, nous avons des dictionnaires pour les noms de personnes, les noms de lieux, ainsi que pour les noms d'organisations.

5.3.2.1 Format des fichiers du lexique

Les dictionnaires qui forment notre lexique sont intégrés chacun dans un fichier texte respectant les conditions de format suivantes :

- une seule entrée par ligne,

- les entrées sont séparées par des balises qui permettent de catégoriser l'entrée (voir l'exemple de l'annexe VIII),

- chaque fichier est encodé selon le standard Unicode[20] avec le format UTF-8.

Une opération de formatage avec l'éditeur TextPad est appliquée sur chaque fichier, cette opération, qui vise le nettoyage des fichiers, permet la suppression des sauts de ligne, des espaces supplémentaires, des tabulations supplémentaires, ainsi que les lignes éventuellement vides.

5.3.2.2 La création des dictionnaires des noms et des prénoms

Afin de créer les dictionnaires des noms et des prénoms, nous avons compilé manuellement une liste qui comprend 14 600 noms et prénoms arabes provenant de plusieurs sites Web couvrant diverses régions du monde arabe.

Nous avons utilisé le logiciel HTTrack afin de télécharger automatiquement les pages Web. Ensuite, afin de faciliter l'édition du texte, nous avons converti les pages HTML au format texte. Cette conversion a été réalisée en employant le script PERL html2text.pl[21] qui permet d'éliminer les balises HTML dans un fichier donné, afin de retenir uniquement le corps du texte.

Après la conversion des textes, nous avons commencé l'étape de fouille manuelle des textes. Cette étape nous a permis de collecter et de sélectionner dans une

---

20 Il s'agit d'une norme informatique, qui vise à donner à tout caractère de n'importe quel système d'écriture de langue un nom et un identifiant numérique, et ce, de manière unifiée. <http://unicode.org/standard/standard.html>, consulté le 11 juin 2008.

21 Le script est disponible à l'adresse suivante : <http://search.cpan.org/~awrigley/html2text-0.003/html2text.pl>, consulté le 09 juillet 2008.

première étape une liste de noms et de prénoms potentiels qui vont être triés automatiquement par la suite afin de ne garder que les entrées uniques.

Recherche des candidats potentiels avec la commande GREP

Nous avons employé la commande GREP[22] de l'environnement du système d'exploitation LINUX, afin de rechercher les noms de personnes dans le fichier des noms et des prénoms potentiels.

Pour compléter la recherche avec la commande GREP, nous avons eu recours à des indices textuels afin de rechercher uniquement les lignes contenant éventuellement un nom ou un prénom comme dans l'exemple suivant, où l'indice textuel est le mot *président*.

grep président fichier.txt

La commande ci-dessus, permet d'obtenir les lignes qui font partie du fichier fichier.text et qui contiennent le mot *président*.

Cette opération de recherche doit se faire sur l'ensemble des textes et le résultat est exporté dans un fichier de résultats qui n'inclut que les lignes contenant l'un des mots clés de nos expressions de recherche GREP.

Nettoyage du fichier des résultats

Une validation manuelle est effectuée sur chaque entrée du fichier des résultats, chaque fois qu'un nom ou qu'un prénom est localisé, il est systématiquement

22 La commande UNIX GREP permet de rechercher une chaîne de caractères dans un fichier en affichant la ligne qui contient la chaîne recherchée.

intégré dans une feuille Excel qui comporte deux colonnes : la première est réservée pour les prénoms et la deuxième est réservée pour les noms de famille.

Par la suite, nous avons trié chaque colonne afin de ne pas retenir les entrées multiples. Enfin, nous avons exporté le contenu de chaque colonne unique dans un fichier dictionnaire correspondant.

Le dictionnaire des prénoms

Ce dictionnaire se compose d'une liste des prénoms arabes les plus courants ainsi que d'un bon nombre de prénoms non arabes. Nous avons tenu compte des prénoms simples (les prénoms se composant d'un seul mot) et des prénoms composés qui peuvent comprendre jusqu'à quatre mots. Après la compilation et le tri de ce dictionnaire, nous avons pu dénombrer une liste d'environ 12 000 prénoms uniques.

Ensuite, dans le but d'étendre la couverture de la liste des prénoms, nous avons acquis une liste de prénoms arabes (9 000 entrées) ; cette liste qui comprend les prénoms arabes les plus fréquents nous a été fournie par la CJK Dictionary Institute (CJKI)[23], une compagnie spécialisée dans le développement de ressources linguistiques.

La liste de CJK a été fusionnée avec la liste que nous avons déjà compilée. Cette fusion nous a permis d'avoir une version finale du dictionnaire des prénoms qui comprend désormais 17 000 entrées uniques (4000 entrées de la liste CJK figuraient dans notre liste initiale).

Le dictionnaire des noms

Ce dictionnaire a été compilé de la même manière que celui des prénoms. Par contre, ce dictionnaire n'inclut qu'un nombre limité de noms de famille d'origine

23 <http://www.kanji.org/cjk/index.htm>, consulté le 11 janvier 2008.

non arabe. Les noms de famille composant ce dictionnaire sont majoritairement d'origine arabe, ils sont souvent composés et ils peuvent comprendre jusqu'à six mots pour les cas les plus longs. Ce dictionnaire se compose de 2 600 noms de famille.

5.3.2.3 La création du dictionnaire des organisations

Ce dictionnaire comprend une liste de plusieurs compagnies et organismes connus à travers le monde. Nous avons suivi la même procédure de collecte que celle employée avec les noms de personnes, et qui est basée sur la collecte des données à partir des sites Web.

Ainsi, nous nous sommes servis du même ensemble de fichiers HTML que nous avons rassemblé lors de la recherche des noms de personnes.

Nous avons recueilli les noms d'organisations en faisant des recherches simples avec la commande GREP sur chaque fichier, en nous servant toujours des indices textuels comme (la compagnie, l'agence, l'organisation, etc.). Enfin, après avoir réuni les noms d'organisation dans un classeur Excel, nous avons trié cette liste avant d'exporter les entrées uniques dans le fichier des dictionnaires des noms d'organisation. Cette opération nous a permis de créer une liste d'environ 4 000 noms d'organisations.

5.3.2.4 La création du dictionnaire des noms de lieux géographiques

La liste des lieux géographiques dans le monde est relativement stable et il convient de se munir d'une liste des noms de lieux les plus connus, à l'instar des noms de pays et des principales villes dans le monde.

Ainsi, afin de créer le dictionnaire des lieux géographiques, nous avons utilisé une liste existante de noms de lieux en arabe. La liste, baptisée KNAB, provient d'une base de données développée par l'Institut de la langue estonienne[24] KNAB et qui dispose d'une base de données multilingue comprenant une liste de 2 200 entrées en langue arabe. Le contenu de cette liste est reparti entre les noms de pays, les noms de villes, les noms de montagnes ainsi que les noms de rivières. Toutes les entrées de la liste ont été vérifiées manuellement avant leur importation dans un fichier dictionnaire formaté pour qu'il soit conforme au formatage requis par nos dictionnaires (cf. section 5.3.2.1).

5.3.3 La création et l'implémentation des règles

La formalisation des règles de repérage exige une connaissance approfondie de la distribution des EN dans la phrase arabe. De ce fait, nous avons étudié dans le chapitre 4 la distribution des marqueurs lexicaux dans le texte arabe afin de mettre à nu la structure des EN en arabe. Il s'agit d'une étape préalable à la création et à l'implémentation des règles de repérage. Nous allons présenter dans ce qui suit le résultat de cette étude en corpus ainsi que les différents fichiers de règles que nous avons créés dans le cadre de ce projet.

5.3.3.1 La distribution des marqueurs lexicaux dans le corpus

Dans cette section, nous reviendrons sur la question des marqueurs lexicaux (présentée dans la section 2.1.1). D'abord, nous définissons deux sous-types de marqueurs lexicaux. Ensuite, nous présentons le résultat de l'étude que nous avons faite en corpus qui porte sur la fréquence d'apparition des marqueurs lexicaux selon

---

24 On peut consulter la base de données sur le site suivant :
<http://www.eki.ee/knab/p_mm_en.htm>, consulté le 3 mars 2007.

le type. Enfin, nous verrons les limites de l'utilisation de ces marqueurs comme preuve à travers des exemples.

5.3.3.1.1 Les marqueurs lexicaux internes et externes

Nous adapterons dans cette section la notion employée par Mcdonald (1993) qui propose un système de repérage des noms propres fondé sur la notion de marqueurs lexicaux internes et de marqueurs lexicaux externes.

Les marqueurs lexicaux internes font partie intégrante du nom propre et se trouvent donc à l'intérieur de ce dernier. Les marqueurs lexicaux internes sont généralement issus d'une liste finie et permettent de repérer les noms propres comme dans les exemples suivants :

جريدة الصباح /jaryidatu aSabaaH/ « **journal** le matin »,

منظمة الشباب و الطفولة /MunaZamatu ashabab wa atufula/ « **organisation** de la jeunesse et de l'enfance ».

Le marqueur lexical interne est généralement localisé au début, ou à la fin des noms propres comme dans le cas du nom de compagnie *limité* ou *incorporé* (Inc. ou Ltée).

Pour leur part, les marqueurs lexicaux externes aident le lecteur à avoir plus d'informations sur les personnes ou les organisations citées. Elles peuvent se situer à gauche ou à droite du mot comme dans les exemples suivants :

مدينة واشنطن /madyinatu waashinTun/ « la ville de Washington »,

الدكتور عمران /aalduktwur `umraan/ « le docteur Omrane »,

وائل الطالب التونسي /waa'il aaltaalib attwunisyi/ « Wael, l'étudiant tunisien ».

Les marqueurs lexicaux externes sont des éléments très importants à prendre en compte, surtout pour désambiguïser la catégorie exacte du nom propre. Par exemple, un nom de personne peut être en même temps un nom de compagnie, comme dans le cas de la *compagnie Tremblay auto* où le mot *Tremblay* n'est plus considéré comme un nom de personne, mais une partie intégrante du nom de l'organisation.

5.3.3.1.2 Présentation et analyse des résultats

Afin de pouvoir établir les spécifications linguistiques pour créer nos règles de repérage d'une manière adéquate, nous avons conduit une analyse en corpus de la structure des différents types d'entités nommées ainsi que des marqueurs lexicaux.

L'analyse a été réalisée sur un ensemble d'articles de nouvelles issus des sites Web de journaux en ligne comme le Libanais النهار Annahar et du site d'information continue www.aljazeera.net. L'ensemble du corpus contient 30 articles dans lesquels nous avons dénombré 1 100 noms propres. L'extraction des noms propres a été faite tout d'abord d'une manière semi-automatique avec des expressions régulières qui permettent la localisation des lignes contenant les noms propres potentiels. Dans un deuxième temps, une extraction manuelle complémentaire a permis d'extraire les noms propres qui ne disposent pas de marqueurs lexicaux et qui n'ont pas été repérés par les expressions régulières.

Enfin, une validation de la liste des noms propres a été faite à la main avec l'éditeur de texte Textpad. La validation est rendue possible grâce à la taille sensiblement réduite de la liste. Parmi les 1 100 noms propres que nous avons détectés, nous avons pu constater une légère domination, en terme de fréquence, des noms de personnes qui sont suivis des noms d'organisations et enfin des noms de lieux qui arrivent en dernière position (voir Tableau XV pour plus de détails).

Type des EN	Fréquence dans le corpus		Fréquence des EN disposant de marqueurs lexicaux	
	Fréq.	%	Fréq.	%
Personnes	500	45,45 %	440	88 %
Lieux	270	24,54 %	81	30 %
Organisations	330	30 %	257	78 %
Total	1 100	100 %	778	65 % (moyenne)

Tableau XV : Distribution des EN dans le corpus

Dans ce qui suit, nous illustrerons quelques exemples de marqueurs lexicaux internes et externes qui couvrent les trois catégories d'entités nommées retenues dans cette étude : les noms de personnes, les noms de lieux et les noms d'organisations (Tableaux XVI, XVII, XVIII, XIX).

Noms de personnes :

الأستاذ خالد عبد الله الثاني الذي قال /aal aa'ustaadh khaalid `bdallah aalthaani aalathi qaala/ « le professeur Khaled Ben abdallah AlThani qui a dit ».

AlOstad « Le professeur »	Altahni « AlThani »	Alathi Qala « qui a dit »
Marqueur lexical externe	Marqueur lexical interne	Marqueur lexical externe

Tableau XVI : Exemple de marqueurs lexicaux pour les noms de personnes

Noms de lieux : مدينة واشنطن /madyinatu washinTun/ « la ville de Washington ».

مدينة madinatu « ville de »
Marqueur lexical externe à droite

Tableau XVII : Exemple d'un marqueur lexical pour les noms de lieux

Noms d'organisations :

شركة وحيد و اخوان لتدي /sharikat waHyid wa'ikhwaan iltidyi/ « la compagnie Wahid et frères Ltée ».

sharikat « la compagnie »	iltidyi « Ltée »	wa'ikhwaan « et frères »
Marqueur lexical externe	Marqueur lexical interne	Marqueur lexical interne

Tableau XVIII : Exemple de marqueurs lexicaux pour les noms d'organisations

صندوق النقد الدولي /Sunduq aalnaqd aalduwali/ « le fond monétaire international ».

Sunduq « fond »	aalduwali « international »
Marqueur lexical interne	Marqueur lexical interne

Tableau XIX : Exemple de marqueurs lexicaux pour les noms d'organisations

Nous avons aussi calculé la fréquence des marqueurs lexicaux accompagnant les noms propres selon la catégorie des EN. Le Tableau XX ci-dessous, détaille les différents types de preuves selon la catégorie de l'entité nommée.

Catégories des EN	**Contexte de droite présent uniquement**		**Contexte de gauche présent uniquement**		**Contexte de droite et de gauche présent**		**Total de tous les contextes**	
	Fréq.	%	Fréq.	%	Fréq.	%	Fréq.	%
Personnes	288	65,45 %	42	9,54 %	110	25 %	440	100 %
Lieux	81	100 %	0	0 %	0	0 %	81	100 %
Organisations	88	34,24 %	27	10,50 %	142	55.25 %	257	100 %
Grand total	457		69		252		778	

Tableau XX : Distribution des marqueurs lexicaux selon le contexte

En parcourant le Tableau XX, nous pouvons déduire qu'une grande partie des EN peut être repérée grâce à la présence de marqueurs lexicaux de droite à l'instar de la catégorie des noms de personnes où 65,45 % des marqueurs lexicaux sont localisés à droite. L'observation de ces contextes révèle que la plupart sont soit des titres, soit des fonctions (*Mr, le docteur, le policier*).

Enfin, il arrive souvent qu'une entité soit accompagnée d'une preuve conjointe dans le contexte de gauche et le contexte de droite. Ceci arrive presque exclusivement avec les noms d'organisations 55,25 % et, dans un degré moindre, avec les noms de personnes 25 %.

Les noms de lieux ne présentent généralement pas de variation et ils ne disposent que rarement de marqueurs lexicaux. Dans les rares cas où des preuves sont présentes, il s'agit de marqueurs lexicaux internes qui se positionnent à la droite du mot dans 100 % des cas.

5.3.3.1.3 Quelques particularités des marqueurs lexicaux

Lors de cette étude, nous avons observé quelques cas de marqueurs lexicaux qui méritent une attention particulière. Citons par exemple le cas des prépositions في fi « dans » et ب bi « dedans ». Ces deux prépositions peuvent paraître de prime abord comme des marqueurs lexicaux fiables dans le repérage de certains noms de lieux comme le cas suivant :

في باريس /fi bariz/ « à Paris »

في سان فانسان /fyi saan faansaan/ « à Saint-Vincent »

La règle qui permet de détecter le nom de lieu « Saint-Vincent » dans l'exemple ci-dessus, est formalisée de la manière suivante :

في /fi/ « dans » + nom inconnu -> noms inconnus= nom de lieu

La règle peut se lire comme suit : chaque fois que le système rencontre la séquence /fi/ suivi immédiatement d'un nom inconnu (non présent dans nos dictionnaires), le nom inconnu peut être automatiquement reconnu comme étant un nom de lieu.

Sur la base de nos observations, nous avons constaté que ce type de règles peut provoquer des erreurs de repérage. En effet, la préposition /fi/ précède fréquemment des noms communs ainsi que d'autres catégories comme dans l'exemple suivant :

في خطابه /fyi khiTaabihi/ « dans son discours ».

Le fait d'éliminer ce genre de règles aura certainement une incidence sur le rappel, puisque nous risquons de perdre un indice de repérage important. D'un autre côté, leur absence fera certainement augmenter la précision globale du système.

5.3.3.2 La présentation des règles de repérage

Les règles de repérage du système RENAR sont au nombre de 142 et elles sont reparties sur 4 fichiers de règles comme l'illustre le Tableau XXI.

Catégorie du fichier des règles	Nombre de règles
Personnes	52
Lieux	12
Organisations	32
Les entités temporelles et numériques	46
Total	**142**

Tableau XXI : Distribution des règles selon la catégorie d'EN

Dans cette section, nous présentons tout d'abord le format requis par les différents fichiers de règles. Par la suite, nous présentons individuellement les fichiers de règles pour chacune des 4 catégories d'EN (noms de personnes, noms d'organisation, noms de lieux, dates et expressions numériques). Enfin, nous illustrons certaines difficultés rencontrées lors de la création de ces règles.

5.3.3.2.1 Format des fichiers des règles

Le fichier des règles doit obligatoirement commencer par la balise ouvrante <nameshape>. La balise fermante </nameshape>, sert à son tour pour signifier la fin du fichier. Les règles sont séparées entre elles par des sauts de lignes. Il n'existe

pas de limite quant au nombre de règles dans un fichier. L'édition du fichier de règles peut se faire avec un éditeur de texte standard.

Les règles sont sous forme de patrons qui reflètent la structure sous-jacente des EN à extraire. Le repérage des règles se fait dans l'ordre de leur apparition dans le fichier, nous reviendrons sur l'importance de l'ordre de passage des règles dans la section 5.3.3.2.6.

5.3.3.2.2 Le fichier des règles pour les noms de personnes

Le fichier des règles pour le repérage des noms de personnes se compose de 52 règles. Ce fichier se divise en deux sections (un extrait des règles est inclus dans l'annexe IV). Dans la première section, nous avons placé les règles de repérage. Dans une deuxième section, nous avons placé une liste de titres spécialement dédiés pour le repérage des noms de personnes. La liste des titres a été créée manuellement en se basant sur nos connaissances linguistiques, ainsi que sur des observations faites sur le corpus. La liste est divisée en sous-groupes à l'instar des titres militaires, titres de civilité, professions (ex. : *colonel, général, professeur, docteur, madame, frère*).

Nous illustrons dans ce qui suit un exemple d'une règle simple extraite du fichier des règles pour les noms de personnes.

${firstnames}\b{lastnames}

Cette règle permet de détecter les EN qui commencent par un prénom figurant dans le dictionnaire des prénoms (appel avec l'expression {firstnames}) et qui sont immédiatement suivis par un nom de famille figurant dans le dictionnaire des noms de famille (appel avec l'expression {lastnames}). Le caractère $ au début de la règle signifie que la chaîne à extraire commence par la catégorie à droite du signe dollar. Ensuite, l'expression \b, qui représente l'abréviation anglaise de *blank* équivaut au caractère espace dans la règle.

La règle citée doit pouvoir détecter des EN comme celle en gras dans l'exemple suivant :

السيد **أحمد عيسى** / aalsyd **'ahmid 'isa**/ « monsieur **Ahmed Issa** »

Dans ce cas, la règle commence par la recherche du prénom أحمد *Ahmed* dans le dictionnaire des prénoms. S'il est trouvé, le système vérifie si le mot suivant figure à son tour dans la liste des noms de famille, le cas échéant, l'entité nommée sera détectée correctement.

Dans le cas où le nom de famille عيسى *Issa* ne figure pas dans le dictionnaire des noms de famille, cette règle va échouer. Néanmoins, il est toujours possible de repérer cet EN grâce à une règle qui emploie les titres à l'instar de la règle suivante :

${title}\b{firstnames}\b {unknown}

Cette règle permet de détecter les EN commençant par un marqueur lexical de type *titre* représenté ici par l'expression {title} et qui est dans ce cas le titre de civilité *Monsieur*. Ensuite, l'expression {firstnames} permet de repérer dans ce cas un prénom connu comme le prénom *Ahmed* dans notre cas. Enfin, l'emploi de l'expression de repérage des mots inconnus {unknown}, permet de repérer le nom de famille Issa et de terminer l'opération de repérage de cette EN.

5.3.3.2.3 Le fichier des règles pour les noms d'organisations

De même qu'avec les fichiers de règles pour les noms de personnes, nous avons créé un fichier de règles pour le repérage des noms de compagnies et des organisations. La structure de ce fichier de règles est similaire à celui des noms de personnes. Un extrait est inclus dans l'Annexe V. Le fichier se divise en deux sections. La première section comporte les règles de repérage et la deuxième section inclut les marqueurs lexicaux reliés exclusivement au lexique des noms

d'organisations et des compagnies comme les abréviations (*Inc, ltée*) ou des expressions comme و إخوان /waikhwan/ « et frères » et و شركائه /washourakaouh/ « et associés ».

Par ailleurs, afin de créer les règles de repérage pour les noms d'organisations, nous avons utilisé le logiciel aConCorde afin d'étudier le contextes d'apparition de 1 200 noms d'organisations, ce qui nous a permis de générer les patrons les plus représentatifs des organisations d'origine arabe et non arabe.

Dans ce qui suit, nous présentons un exemple d'une règle très productive :

${cues_org}\b{firstnames}\b{patronym}\b{lastnames}\b{cues_org2}

Cette règle permet de détecter un nom d'organisation commençant par un marqueur lexical externe représenté ici par l'expression {cues_org} et qui est suivi par un prénom connu {firstnames} et d'un patronyme connu {patronym} suivi à son tour d'un nom de famille connu {lastnames} et se terminant par un marqueur lexical interne {cues_org2}. Cette règle permet le repérage des noms de compagnies comme celui de l'exemple suivant :

شركة محمد أبو الحمداني و إخوان /sharikatu muwHamad aabuw aalHamadani waikhwaan/ « la compagnie Mohamed Abou Alhamadani et frères. ».

5.3.3.2.4 Le fichier des règles pour les lieux géographiques

Le fichier des règles pour le repérage des lieux géographiques est différent des deux fichiers décrits précédemment (l'Annexe VI donne un extrait). Cette distinction s'explique par le fait que la tâche dans ce cas consiste principalement en la détection des noms de lieux connus dans le dictionnaire.

Nous avons choisi cette approche étant donné que la liste était très complète et qu'elle couvrait approximativement tous les pays et les capitales du monde. Nous

avons estimé qu'il serait très risqué de créer des patrons pour détecter des lieux qui ne font pas partie de la liste. Prenons l'exemple simple suivant :

${geo_name_simple}|{geo_name_complex}

Cette règle permet de détecter les noms de lieux simples ou complexes. L'expression {geo_name_simple} fait appel à la liste des noms de lieux simples dans le dictionnaire, tandis que l'expression {geo_name_complex} fait appel à la liste des noms de lieux complexes (composés de 2 mots ou plus). L'opérateur | entre les deux expressions est un opérateur de choix entre plusieurs alternatives, il fait correspondre l'une des expressions placées avant ou après l'opérateur.

5.3.3.2.5 Le fichier des règles pour les entités temporelles et numériques

Afin de repérer les entités temporelles et les entités numériques, nous avons créé un fichier de règles qui inclut à la fois les règles et le lexique nécessaire pour le repérage des entités temporelles et des entités numériques (voir Annexe VII).

Concernant les entités numériques, le fichier des règles permet le repérage des expressions monétaires (monnaies et devises), des mesures, des distances, des masses et des volumes.

Par exemple, concernant le lexique servant au repérage des unités de mesure, nous avons inclus des abréviations en arabe comme كغ *kg* pour *kilogramme* ou des unités comme هكتار *hectare* et أكر *acre*.

Pour ce faire, nous avons compilé manuellement un lexique de base pour chaque catégorie en traduisant du français vers l'arabe les unités de mesure disponibles sur le site du Bureau international des Poids et mesures[25].

25 <http://www.bipm.org/fr/si/si_brochure/>, consulté le 3 avril 2007.

Une fois que nous avons incorporé les entrées du lexique dans le fichier des règles, des expressions régulières simples ont permis l'identification et l'extraction des entités numériques. Dans ce qui suit, nous illustrerons une ligne extraite du fichier des règles qui permet de repérer dans cet exemple une entité numérique.

${number}/b{measure}

Dans la règle ci-dessus, l'expression *number* permet de repérer un nombre précédant l'unité de mesure, tandis que l'expression *measure* entre accolades renvoi à la liste d'expressions de mesure que nous avons préalablement compilée. Cette règle simple permet de repérer systématiquement des expressions comme : 85 كغ « 85 kg » et 130 صم «130 cm ».

Une autre section dans le fichier des règles a été réservée au repérage des entités temporelles qui apparaissent principalement sous forme de dates et d'expressions temporelles.

Nous avons tout d'abord construit un lexique composé de termes simples ainsi que de termes complexes suite à une étude des différentes expressions temporelles dans un échantillon de 30 textes issus du journal en ligne Aljazeera.net. Nous avons collecté manuellement 190 entrées (100 termes simples et 90 termes complexes). Nous avons inclus un échantillon de ces termes dans le Tableau XXII.

Termes simples	Termes complexes
ساعة /saa'a/ « heure »	بعد الزوال /ba'da azawaal/ « après-midi »
دقيقة /dakika/ « minute »	السنة القادمة /asana aalqaadima/ « l'année prochaine »
يوم /yawm/ « jour »	عيد رأس السنة /ra'is asana/ « jour de l'an »
رمضان /ramaDaan/ « ramadan »	عيد الميلاد /'id aalmyilaad/ « Noël »
أسبوع /'asbuw'/ « semaine »	عيد القيامة /'id aalqiyaama/ « la fête de Pâques »

Tableau XXII : Exemples de termes reliés à la notion du temps

Par la suite, nous avons défini les règles de repérage en nous basant sur le lexique des expressions temporelles afin de créer des expressions régulières qui permettent la détection des entités temporelles des plus simples aux plus complexes.

Par exemple, l'expression temporelle 2007 عيد القيامة qui se traduit par « la fête de Pâques 2007 », peut être détectée par la règle suivante :

${dates-complex}/b{number}

l'expression *dates-complex* entre accolades renvoie à la liste des expressions temporelles complexes et permet de détecter dans ce cas *la fête de Pâques*. Enfin, l'expression *number* permet de détecter le nombre 2007, ce qui permet de détecter l'entité numérique au complet.

5.3.3.2.6 L'ordre de passage des règles

Dans les quatre fichiers de règles que nous avons décrits, l'ordre de passage des règles revêt un aspect très important pour la stratégie de repérage.

D'abord, nous avons décidé de placer les règles les plus longues et les plus complexes en premier, et ceci, dans tous les fichiers de règles sans exception dans le but de laisser les règles les plus simples vers la fin.

Cette approche permet de s'assurer que nous détecterons les EN les plus longues en premier et d'éviter ainsi de détecter partiellement des EN comme dans l'exemple suivant :

شركة وحيد و إخوان إلتدي /sharikatu waHyid wa 'aikhwaan iltidyi/ « la compagnie Wahid et frères Ltée ».

Dans cet exemple, nous avons un cas où le nom de l'organisation est composé de 4 mots séparés par des espaces. Supposons que nous avons mis les règles les plus

courtes en premier et les plus longues en dernier comme l'illustre le Tableau XXIII.

Ordre	Règles	Explication	Entité reconnue
1	${cues_org}/b{person_ name}	Marqueur lexical externe d'organisation + nom de personne	**EN partiellement reconnue** شركة وحيد/sharikatu waHyid/ « La compagnie Wahid »
2	${cues_org}/b{person_ name}/b{cues_org2}/b {cues_org2}	Marqueur lexical externe d'organisation + nom de personne + marqueur lexical interne d'organisation + marqueur lexical interne d'organisation	**EN partiellement reconnue** شركة وحيد و إخوان /sharikatu waHyid wa 'aikhwaan/ « La compagnie Wahid et frères »
3	${cues_org}/b{person_ name}/b{cues_org2}/b {cues_org2}/b{cues_or g2}	Marqueur lexical externe d'organisation + nom de personne + marqueur lexical interne d'organisation + marqueur lexical interne d'organisation + marqueur lexical interne d'organisation	**EN entièrement reconnue** شركة وحيد و إخوان إلتدي /sharikatu waHyid wa 'aikhwaan iltidyi/ « La compagnie Wahid et frères Ltée »

Tableau XXIII : Illustration de l'importance de l'ordre des règles

L'exemple du Tableau XXIII, nous montre l'impact et l'importance de l'ordre de passage des règles. Ainsi, il était nécessaire de mettre les règles les plus longues en premier et les règles les plus courtes en dernier. Cette approche permet d'éviter la détection partielle des EN comme c'est le cas avec la règle 1 et la règle 2 dans l'exemple du tableau XXIII.

Par ailleurs, nous avons fait aussi quelques choix sur le type d'entités nommées qui doit être détecté en premier. Dans ce cadre, nous avons jugé utile de détecter en

premier les EN de type personnes qui figurent déjà dans le dictionnaire de noms et de prénoms connus. Le fait de trouver en même temps le nom et le prénom d'une personne dans le dictionnaire, nous procure la certitude que l'EN en question est bien détectée. Aussi, ce choix nous évite un possible chevauchement de règles entre le fichier des règles des noms de personnes et celui des organisations.

En effet, les fichiers de règles des noms de personnes et celui des noms d'organisations comportent beaucoup de règles assez proches. Ceci est dû au fait que les noms d'organisations comportent souvent des noms de personnes et de familles. Il était donc nécessaire d'inclure des règles similaires dans les deux fichiers tout en veillant à ce que la priorité soit accordée au fichier des noms de personnes. Ainsi, les noms de personnes sont détectés en premier, par la suite viennent les noms d'organisations et enfin les noms de lieux. Les entités numériques et entités temporelles sont détectées dans une étape ultérieure et elles sont séparées pour des raisons techniques[26] du processus de repérages des entités nommées.

5.3.3.2.7 Quelques difficultés rencontrées lors de la création des règles

Il était parfois difficile de créer des règles à cause de la complexité des noms de personnes en arabe. Une difficulté vient du fait qu'il était impossible de prévoir la limite gauche du nom propre. En d'autres termes, il est impossible de prévoir à l'avance la composition du nom propre en terme de nombre de mots afin de savoir où s'arrêter lors du repérage afin de ne pas inclure des mots en dehors du nom propre en question. L'étude sur la structure des noms de personne en arabe (cf. section 4) nous a révélé que ce dernier peut s'étendre jusqu'à 8 voire 10 mots dans certains cas. Par conséquent, nous avons plafonné les règles pour le repérage des noms de personnes à une combinaison de 10 mots au maximum.

---

26 Le module de repérage des entités numériques et des dates est séparé du module principal dans l'architecture du système EMM.

De plus, il était très difficile de couvrir avec notre grammaire locale l'ensemble des variantes régionales des noms d'organisations étant donné que les appellations et les standards diffèrent d'un pays à un autre et d'une culture à une autre. Nous avons donc limité notre couverture aux principales organisations de renommée internationale ainsi que celles provenant du monde arabe.

5.4 Conclusion

Nous avons illustré, dans ce chapitre, l'architecture du système EMM ainsi que l'outil RENAR. Ensuite, nous avons présenté l'approche suivie pour la création du lexique et des règles de repérage. La section suivante s'intéresse à l'évaluation des performances de l'outil RENAR sur un corpus de textes journalistiques.

6. Évaluation

Le présent chapitre vise la présentation et l'interprétation des résultats obtenus par le système RENAR lors de son évaluation. Dans la section 6.1, nous commençons par la présentation des corpus d'évaluation. Enfin, nous présentons et nous interprétons les résultats de l'évaluation dans la section 6.2.

6.1 Constitution des corpus

Les corpus d'évaluation ont été constitués à partir de textes journalistiques tirés de journaux arabes en ligne couvrant deux régions distinctes du monde arabe.
Pour cela, le corpus global[27] est découpé en deux sous-corpus selon la région couverte.

Nous avons choisi le journal tunisien الصباح Assabah[28] pour le sous-corpus Maghreb et le journal الأنوار Alanwar[29] pour le sous-corpus Levant. Le fait d'avoir choisi des corpus assez variés est motivé par notre souci d'évaluer la performance du système avec deux variétés de langue arabe.

Afin de réaliser l'évaluation de notre outil, nous avons balisé manuellement toutes les entités nommées présentes dans les deux sous-corpus. Par la suite, nous les avons comparées à celles trouvées par le système.

---

27 Le terme corpus global désigne ici l'ensemble composé du sous-corpus Maghreb et du sous-corpus Levant.

28 <http://www.assabah.com.tn/>, consulté le 29 avril 2008.

29 <http://www.alanwar.com/ar/>, consulté le 3 avril 2008.

Le nombre des EN balisées manuellement s'élève à 1751. Les tableaux XXIV et XXV ci-dessous résument le nombre d'occurrences des trois types d'entités nommées respectivement dans le corpus global et dans chaque sous-corpus.

Type des EN	**Distribution**	
	Fréq.	%
Personnes	804	45,92 %
Lieux	433	24,73 %
Organisations	514	29,35 %
Total	1 751	100 %

Tableau XXIV : Distribution des EN balisées dans le corpus global

Les textes choisis contiennent un nombre assez important d'entités nommées et permettent de constituer des corpus d'évaluation assez hétérogènes. Nous avons réuni un ensemble de 35 textes d'environ 34 000 mots au total. Le sous-corpus Maghreb qui comprend 17 316 mots est légèrement plus volumineux que le sous-corpus Levant qui comprend 16 684 mots.

Type des EN	**Distribution**						
	Personnes		Lieux		Organisations		Grand total
	Fréq.	%	Fréq.	%	Fréq.	%	
Maghreb	431	53,60 %	204	47,12 %	293	57,00 %	928
Levant	373	46,40 %	229	52,88 %	221	43,00 %	823
Total	804		433		514		1 751

Tableau XXV : Distribution des EN balisées dans chaque sous-corpus

6.2 Résultats

Dans cette section, nous présentons les résultats obtenus par RENAR sur nos deux sous-corpus. Les tableaux XXVI et XXVII illustrent le nombre d'occurrences des EN qui ont été identifiées correctement, les occurrences mal identifiées ainsi que celles qui n'ont pas été identifiées. Les valeurs de ce tableau vont nous servir pour l'évaluation des performances du système.

Maghreb	**Personnes**	**Lieux**	**Organisations**
EN correctement trouvées et catégorisées	270	150	93
EN mal identifiées / catégorisées	55	16	47
EN non trouvées	106	38	153
Total des entités nommées dans le texte	431	204	293

Tableau XXVI : Résultat de l'extraction par nombre d'occurrences de chaque catégorie d'entités nommées sur le sous-corpus Maghreb

Dans un premier temps, notre évaluation ne va pas inclure la catégorie des entités temporelles et des entités numériques. Ainsi, nous avons délibérément voulu isoler ces deux catégories de nos premiers calculs afin d'avoir une idée claire sur les performances du système (avec ou sans les entités temporelles et entités numériques).

Levant	**Personnes**	**Lieux**	**Organisations**
EN correctement trouvées et catégorisées	265	174	91
EN mal identifiées / catégorisées	25	14	32
EN non trouvées	83	41	98
Total des entités nommées dans le texte	373	229	221

Tableau XXVII : Résultat de l'extraction par nombre d'occurrences de chaque catégorie d'entités nommées sur le sous-corpus Levant

En joignant les résultats globaux obtenus sur le corpus global, on obtient une précision globale de 84,66 % et un rappel de 59,56 % soit une F-mesure de 69,92 % (voir tableau XXVIII).

	Le corpus global
Précision globale	84,66 %
Rappel global	59,56 %
F-mesure	69,92 %

Tableau XXVIII : Précision, rappel et F-mesure globaux obtenus sur le corpus global

Ainsi, l'outil RENAR a obtenu un rappel de 64,39 % et de 55,28 % respectivement sur le sous-corpus Levant et sur le sous-corpus Maghreb et une précision de 88,18 % et de 81,29 % respectivement sur le sous-corpus Levant et sur le sous-corpus Maghreb (voir tableau XXIX). La F-mesure est de l'ordre de 65,80 % sur le sous-corpus Maghreb et de 74,43 % sur le sous-corpus Levant.

	Maghreb	**Levant**
Précision globale	81,29 %	88,18 %
Rappel global	55,28 %	64,39 %
F-mesure	65,80 %	74,43 %

Tableau XXIX : Précision, rappel et F-mesure globaux pour chaque sous-corpus

Nous avons inclus, dans un deuxième temps, le résultat obtenu exclusivement sur les entités numériques et les entités temporelles dans les tableaux XXX et XXXI.

Maghreb	**Entités temporelles**	**Entités numériques**
EN correctement trouvées et catégorisées	116	83
EN mal identifiées / catégorisées	5	6
EN non trouvées	3	4
Total des entités nommées dans le texte	124	93

Tableau XXX : Résultat de l'extraction par nombre d'occurrences des entités numériques et temporelles sur le sous-corpus Maghreb

Levant	Entités temporelles	Entités numériques
EN correctement trouvées et catégorisées	108	70
EN mal identifiées / catégorisées	4	5
EN non trouvées	2	3
Total des entités nommées dans le texte	114	78

Tableau XXXI : Résultat de l'extraction par nombre d'occurrences des entités numériques et temporelles sur le sous-corpus Levant

Les tableaux XXXII et XXXIII présentent cette fois, les résultats de l'évaluation qui comprend les entités temporelles et les entités numériques. L'observation de ces résultats montre une augmentation significative des performances sur les deux sous-corpus par rapport aux résultats obtenus dans les tableaux XXVIII et XXIX qui n'ont pas pris en considération l'évaluation des entités temporelles et les entités numériques.

Ainsi, la F-mesure obtenue sur le sous-corpus Maghreb est passée de 68,00 % à 71,70 % et celle obtenue sur le sous-corpus Levant est passée de 74,36 % à 78.53 % ce qui constitue une augmentation d'environ 4 points pour la F-mesure globale qui est passée de 71,18 % à 74,95 %.

	Le corpus global
Précision globale	87,17 %
Rappel global	65,74 %
F-mesure	74,95 %

Tableau XXXII : Précision, rappel et F-mesure globaux obtenus sur le corpus global (incluant les entités numériques et temporelles)

De plus, nous constatons que les résultats du tableau XXXIII sont à la faveur du sous-corpus Levant avec une F-mesure de 78,53 %, soit 7 points de plus que la F-mesure de 71,70 % obtenue sur le sous-corpus Maghreb.

	Maghreb	Levant
Précision globale	84,66 %	89,84 %
Rappel global	62,18 %	69,75 %
F-mesure	71,70 %	78.53 %

Tableau XXXIII : Précision, rappel et F-mesure globaux pour chaque sous-corpus (incluant les entités numériques et temporelles)

L'observation des erreurs obtenues a révélé des cas de mauvaise catégorisation causée par la forte ambiguïté de certains mots en arabe, des erreurs de reconnaissance partielles de l'entité et des cas d'inclusion de mots ne faisant pas partie de l'entité en question. Enfin, l'absence de marqueurs lexicaux a provoqué plusieurs erreurs de catégorisation.

Une analyse minutieuse des cas difficiles est alors envisagée afin de trouver la meilleure stratégie de repérage. Le tableau XXXIV ci-dessous, détaille le rappel et la précision obtenus pour chaque type d'entités dans les deux sous-corpus y compris les entités temporelles et les Entités numériques.

				Les entités temporelles et numériques	
	Personnes	**Lieux**	**Organisations**	**Entités temporelles**	**Entités numériques**
Maghreb					
Précision	83,07 %	90,36 %	66,42 %	95,86 %	93,25 %
Rappel	62,64 %	73,52 %	31,74 %	93,54 %	89,24 %
Levant					
Précision	91,37 %	92,55 %	73,98 %	96,42 %	93,33 %
Rappel	71,04 %	75,98 %	41,17 %	94,73 %	89,74 %

Tableau XXXIV : Précision et rappel pour les deux sous-corpus (incluant les entités numériques et temporelles)

En parcourant la liste des erreurs de repérage des EN obtenues lors de l'évaluation, nous avons pu constater plusieurs points qui, une fois implémentés, vont améliorer

la performance du système. Il semble que le module de repérage des noms d'organisations mérite une attention particulière étant donné que cette catégorie a obtenu le résultat le plus faible par rapport aux autres catégories comme l'illustre le Tableau XXXIV avec une précision de 66,42 % et un rappel de seulement 31,74 % sur le sous-corpus Maghreb.

D'un autre côté, la prise en charge partielle de nos dictionnaires pour le sous-corpus Maghreb a réduit encore le score étant donné que la variété de l'arabe standard du Maghreb diffère de celle du Levant.

6.2.1 Résultats du repérage des noms de personne

En observant les résultats obtenus sur les noms de personnes dans le tableau XXXIV, nous avons noté que la précision est assez faible sur le sous-corpus Maghreb, 83,07 %. Ceci représente une performance inférieure d'environ 8 points par rapport à celle obtenue sur le sous-corpus Levant qui lui obtient une précision de 91,37 % ce qui constitue une bonne performance.

De même, le rappel obtenu sur le sous-corpus Maghreb est relativement bas avec 62,64 %, soit 8 points de moins que celui obtenu sur le sous-corpus Levant avec 71,04 %.

L'analyse de ces résultats nous fait observer une certaine constance dans l'écart des performances du système sur les deux sous-corpus. En effet, les résultats obtenus sur le sous-corpus Levant devancent systématiquement ceux obtenus sur le sous-corpus Maghreb. Ce qui pourrait en partie s'expliquer par le style particulier de l'arabe journalistique employé dans la presse du Maghreb ; notamment dans le journal tunisien الصباح Assabah qui se caractérise par une forte présence des expressions dialectales.

Compte tenu des moyens limités alloués à ce projet ainsi que du court laps de temps consacré à son développement, on peut considérer que le résultat obtenu avec les noms de personnes constitue une performance honorable malgré les lacunes comme l'absence de repérage d'environ 269 noms de personnes dans les deux sous-corpus.

Parmi les noms de personnes non détectés, on trouve surtout des EN ayant des patrons que nous avons omis d'inclure dans nos règles ou qui n'ont été reconnues que partiellement faute d'avoir, soit un dictionnaire de noms propres plus complet, soit une plus grande liste des marqueurs lexicaux.

Après l'observation de la liste des entités non reconnues, spécifiquement sur le sous-corpus Maghreb, nous avons constaté qu'un nombre assez élevé de noms propres n'était pas encadré de marqueurs lexicaux. Nous avons aussi constaté un usage irrégulier des marqueurs lexicaux. Enfin, nous avons noté l'emploi de noms de personnes typiquement maghrébins qui ne figurent pas dans nos dictionnaires.

6.2.2 Résultats du repérage des noms de lieux

La performance obtenue sur les noms de lieux dans le corpus global est la meilleure parmi les trois catégories d'EN. L'outil RENAR a pu obtenir une précision de 90,36 % sur le sous-corpus Maghreb et de 92,55 % sur le sous-corpus Levant. On note une légère différence dans les performances obtenues sur le sous-corpus Levant qui s'explique par le faible impact du style de la langue arabe du Maghreb dans une catégorie comme les noms de lieux. Par ailleurs, le rappel obtenu sur les deux sous-corpus affiche une légère avance de 3 points sur le sous-corpus Levant qui atteint 75,98 % pour seulement 73,52 % sur le sous-corpus Maghreb.

L'observation des noms de lieux dans le corpus global nous a montré l'existence de contextes ambigus suggérant parfois des noms d'organisations à la place des noms de lieux, mais la plupart des noms de lieux n'ont pu être détectés faute d'une liste plus complète.

6.2.3 Résultats du repérage des noms d'organisations

C'est sur les noms d'organisations que RENAR a obtenu ses plus mauvaises performances. À l'instar de ce qui s'observe pour les noms de personnes, les résultats sont à la faveur du sous-corpus Levant. Ainsi, le système obtient une faible performance de l'ordre de 66,42 % et de 73,98 % respectivement pour la précision sur le sous-corpus Maghreb et le sous-corpus Levant.

Le rappel est aussi faible sur les deux sous-corpus, soit 31,74 % sur le Maghreb et 41,17 % sur le Levant. En d'autres termes, 330 noms d'organisations n'ont pu être repérés dans le corpus global sur un total de 514, ce qui montre que l'outil RENAR éprouve une difficulté particulière avec cette catégorie précise d'entités nommées (indépendamment du corpus).

Cette chute importante des performances peut s'expliquer en partie par le peu de temps que nous avons pu consacrer à cette partie. Par ailleurs, la catégorie des noms d'organisations est particulièrement difficile à extraire.

Ceci s'explique par le fait qu'il n'était pas évident de pouvoir identifier avec précision la limite gauche du nom de l'organisation en terme de mots. De plus, il s'agit d'une catégorie ouverte et très productive et qui ne cesse de s'étendre quotidiennement. Ceci rend le dictionnaire que nous avons compilé, qui est de taille modeste, incapable de couvrir un grand nombre d'organisations. Enfin, la grammaire permettant de repérer les noms d'organisations est de loin la plus complexe et la plus imprévisible par rapport aux deux autres catégories d'EN. Afin

d'atteindre de meilleures performances sur cette catégorie, une augmentation de la couverture du lexique est nécessaire.

6.2.4 Résultats du repérage des entités temporelles et numériques

Les résultats de l'évaluation faite sur les entités temporelles et numériques ont révélé que le système obtient ses meilleures performances avec cette catégorie pour les deux sous-corpus (voir tableau XXXV). Ainsi, sur le sous-corpus Maghreb la précision était de 95,86 % pour les entités temporelles et de 93,25 % pour les entités numériques tandis qu'on obtient sur le sous-corpus Levant des performances très proches avec une précision de 96,42 % sur les entités temporelles et 93,33 % sur les entités numériques.

	Entités temporelles	**Entités numériques**
Maghreb		
Précision	95,86 %	93,25 %
Rappel	93,54 %	89,24 %
F-mesure	94,68 %	91,20 %
Levant		
Précision	96,42 %	93,33 %
Rappel	94,73 %	89,74 %
F-mesure	95,56 %	91,49 %

Tableau XXXV : Précision, rappel et F-mesure globaux sur les entités temporelles et les entités numériques

Ces résultats montrent que les règles que nous avons implémentées ont bien fonctionné. De plus, un rappel élevé a été obtenu avec les entités temporelles sur le corpus global avec respectivement 93,54 % et 94,73 % sur le sous-corpus Maghreb et le sous-corpus Levant. Enfin, une légère baisse du rappel est observée sur les entités numériques, avec respectivement 89,24 % et 89,74 % sur le sous-corpus Maghreb et sur le sous-corpus Levant. Cette baisse est attribuée encore une fois à la taille assez réduite du dictionnaire des entités numériques.

6.3 Conclusion

Cette évaluation nous a permis de tester les performances de l'outil RENAR sur deux corpus couvrant deux variétés régionales de l'arabe journalistique. L'analyse des résultats de cette expérience nous a permis de comprendre mieux les raisons de la baisse des performances de l'outil par rapport à certaines catégories d'EN ainsi que la disparité des résultats entre les deux corpus. L'ébauche de ce travail doit nous permettre d'envisager les développements futurs permettant à l'outil RENAR d'améliorer davantage ses performances.

7. Conclusion

Rappel sur les objectifs de ce mémoire :

- présenter la méthodologie des systèmes de repérage des entités nommées à travers l'état de l'art des systèmes existants d'une manière générale et dans le cadre particulier de la langue arabe,

- créer un module de repérage des EN pour la langue arabe sous forme d'une composante d'un système multilingue d'extraction des EN,

- évaluer les performances du système créé.

Tout d'abord, nous avons montré l'importance de la formalisation des connaissances linguistiques de la langue qu'on s'apprête à adapter surtout dans un environnement optimisé pour des langues très différentes de l'arabe.

En outre, nous avons souligné l'importance de mener une étude sur la composition des EN dans la langue arabe comme étape préalable à la création des règles de repérage. Ainsi, nous nous sommes attardés sur l'étude de la structure et de la composition des noms de personnes tout en tenant compte des variantes régionales du nom propre arabe ainsi que d'autres éléments clés comme les titres, les surnoms, le nom d'origine, etc.

Ensuite, nous avons présenté la composition des noms de lieux dans la langue en évoquant la question de l'ambiguïté de certains noms de lieux en arabe. S'agissant des noms d'organisations, nous avons présenté leurs structures tout en insistant sur les difficultés rencontrées lors de la création des règles pour cette catégorie.

Enfin, nous avons montré que la catégorie des entités temporelles et numériques pose moins de difficultés quand il s'agit de créer les règles de repérage étant donné que les entités temporelles et numériques appartiennent à une liste fermée ayant un nombre limité d'entrées.

Afin de créer l'outil RENAR, nous avons employé une méthode à base de règles pour plusieurs raisons. D'abord, pour les résultats encourageants obtenus par les systèmes à base de règles lors des compagnes d'évaluation et notamment des conférences MUC Bogers (2004). Ensuite, parce que l'environnement EMM est optimisé pour les systèmes à base de règles. Enfin, le fait de réaliser un système à base de règles assure la portabilité de ce dernier, ainsi que la lisibilité des règles de la grammaire locale. Cette lisibilité permet plus facilement l'amélioration, la personnalisation et la mise à jour éventuelle du système.

Nous avons présenté l'architecture d'un outil de repérage automatique des entités nommées pour la langue arabe ainsi que la méthodologie suivie à partir de la préparation des données jusqu'à la création et l'implémentation des règles et des dictionnaires. Nous avons mis l'accent sur l'importance d'avoir en main un ensemble d'outils permettant la préparation adéquate du texte, notamment par le prétraitement lexical à travers les modules de segmentation et de normalisation du texte.

Par ailleurs, une étude sur la distribution en corpus des marqueurs lexicaux a été réalisée. À la lumière de celle-ci nous avons créé et implémenté nos règles. Les résultats obtenus ont confirmé l'importance des marqueurs lexicaux pour un système à base de règles. Nous avons présenté par la suite la structure des règles de repérage ainsi que les difficultés rencontrées lors de leurs implémentations.

Une évaluation a été conduite afin de mesurer les performances du système. Nous avons choisi pour cette fin, deux corpus couvrant deux variétés régionales de

l'arabe journalistique moderne, soit le corpus Maghreb et le corpus Levant. Les résultats obtenus ont montré une meilleure performance sur le Levant, sur lequel nous avons obtenu une F-mesure de 78,53 % par rapport à la F-mesure de 71,70 % obtenue sur le corpus Maghreb. Toutefois, les résultats restent encourageants pour un premier système qui peut être amélioré ultérieurement.

En résumé, dans la perspective d'une prochaine amélioration de l'outil RENAR, nous avons retenu quelques recommandations. D'abord, il serait primordial d'inclure un module de résolution des ambiguïtés surtout avec l'importance de l'ambiguïté dans la langue arabe que nous avons montrée dans les sections 3.6 et 4.1.9

De plus, il serait important d'enrichir nos règles en tenant compte des contextes particuliers que le système n'a pas pu détecter. Ceci concerne particulièrement les règles de repérage des noms d'organisation. En outre, il serait important de retravailler la grammaire locale en tenant compte de certaines particularités de l'arabe journalistique du Maghreb ainsi que d'autres variantes régionales de la langue arabe.

Par ailleurs, il est recommandé d'enrichir les différents dictionnaires, avec une attention particulière à la liste des noms de lieux qui dépend quasi totalement du dictionnaire ainsi que la liste des noms d'organisations qui contient un nombre insuffisant d'entrées. D'autre part, une révision systématique des mots déclencheurs, plus particulièrement des mots ambigus qui génèrent fréquemment des erreurs, permettra au système d'atteindre une meilleure performance.

L'intégration d'un antidictionnaire sous forme d'une liste de mots servant de filtres doit contribuer certainement à l'amélioration des performances du système en filtrant directement certaines catégories syntaxiques « indésirables » dans la composition des entités nommées (les *verbes* ou les *pronoms* par exemple).

Enfin, dans la perspective d'une évaluation plus complète du système, il serait intéressant de faire des tests supplémentaires sur des corpus de taille plus larges couvrant d'autres variantes régionales de l'arabe journalistique moderne.

Bibliographie

Abuleil, S. (2004). « Extracting Names from Arabic Text for Question-Answering Systems », dans *Proceedings of Coupling approaches, coupling media and coupling languages for information retrieval (RIAO 2004)*. Avignon, France, p. 638- 647.

Al-Chartouni, R. (1986). *mabaadiu al arabijati (Les bases de l'arabe)*. Beyrouth, Dar El-Machreq, s.p.

Apptek. (2009). Le site Web de la compagnie Apptek. En ligne. <http://www.apptek.com/index.php/name-finder-free-text-proper-noun-identification>. Consulté le 12 janvier 2009.

Audibert, L. (2007). *UML 2.0 Notes de cours*. Institut universitaire de technologie (IUT). En ligne. <http://laurent-audibert.developpez.com/Cours-UML/html/index.html>. Consulté le 20 mars 2007.

Balvet, A. (2001). « Le système INTEX : une plate-forme pour les grammaires locales ». En ligne. <http://atala.biomath.jussieu.fr/je/011215/Balvet.ppt>. Consulté le 16 Janvier 2008.

Basistech. (2009). Le site Web de la compagnie Basistech. En ligne. <http://www.basistech.com/entity-extraction/>. Consulté le 16 janvier 2009.

Bauer, G. (1985). *Namenkunde des Deutschen*. Berlin,Germanistische Lehrbuchsammlung, 356 p.

Benajiba, Y., Rosso, P., et J. Benedíruiz (2007). ANERsys: An Arabic Named Entity Recognition System Based on Maximum Entropy . En ligne.

<http://www.springerlink.com/content/5g6n298843878701/>. Consulté le 20 mars 2009.

Bikel, D.-M., Miller, S., Schwartz, R., et R. Weischedel (1997). « Nymble : a high-performance learning name finder », dans *Proceedings of the 5th Conference on Applied Natural Language Processing*. Washington, USA, p. 159-168.

Bikel, D.-M., Schwartz, R. et R. Weischedel (1999). « An Algorithm that Learns What's in a Name », *dans Machine Learning,* vol. 34, nº 1-3, p. 211-231.

Blachère, R. et M. Gaudefroy-Demombynes (1975). *Grammaire de l'arabe classique*, Paris, Maisonneuve & Larose, 508 p.

Bogers, T. (2004). *Dutch Named Entity Recognition : Optimizing Features, Algorithms, and Output*, mémoire de maîtrise, Tilburg University.

Boisen, S., Crystal, M.-R., Schwartz, R., Stone, R. et R. Weischedel (2000). « Annotating resources for Information Extraction », dans *Proceedings of the 2nd International Conference on Linguistic Resources and Evaluation (LREC'2000)*. Athens, Greece, p. 1211-1214.

Borthwick, A. (1999). *A Maximum Entropy Approach to Named Entity Recognition*, thèse de doctorat, NewYork University.

Boulanger, J.-C. et M.-C. Cormier (2001). *Le nom propre dans l'espace dictionnairique général. Études de métalexicographie*. Tübingen, Niemeyer, 214 p.

Brill, E. (1995). « Transformation-Based Error-Driven Learning and Natural Language Processing : A Case Study in Part of Speech Tagging », dans *Computational Linguistics*, december 1995, vol. 21, nº 4, p. 543-566.

Candillier, L. (2006). *Contextualisation, Visualisation et Evaluation en Apprentissage Non Supervise*, Thèse de doctorat, Université Charles de Gaulle, Lille 3.

Chinchor, N.-A. (1997). Overview of MUC-7 / MET-2. En ligne. <http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html#appendices>. Consulté le 3 mai 2007.

Clearforest. (2009). Le site Web de la compagnie Clearforest. En ligne. <http://www.clearforest.com/Technology/ExtractionModules.asp>. Consulté le 30 janvier 2009.

Coates-Stephens, S. (1993). « The Analysis and Acquisition of Proper Names for the Understanding of Free Text », dans *Computers and the Humanities*, vol. 26, n° 5-6, p. 441-456.

Collins, M. et Y. Singer (1999). « Unsupervised models for named entity classification », dans *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. Singapore, p. 189-196.

Cucchiarelli, L. et P. Velardi (1999). « Adaptability of linguistic resources to new domains : an experiment with proper noun dictionaries », dans *Proceedings of the Vextal Conference*. Venice, Italy, p. 25-30.

Cucerzan, S. et D. Yarowsky (1999). « Language independent named entity recognition combining morphological and contextual evidence », dans *Proceedings of 1999 Joint SIGDAT Conference on EMNLP and VLC*. Singapore, p. 90-99.

Cunningham, H., Wilks Y. et Gaizauskas R. (1996). « GATE - a General Architecture for Text Engineering », dans *Proceedings of the 16th Conference on Computational Linguistics (COLING-96)*. Copenhagen, Denmark, p. 1057-1060.

Daille, B., Fourour, N. et E. Morin (2000). « Catégorisation des noms propres : une étude en corpus ». *Cahiers de Grammaire*, n° 25, p. 115-129.

Dalianas, H. et E. Aström (1998). « SweNam - A swedish Named Entity Recognizer. Its construction, training and evaluation », dans *Proceedings of MUC-7*. En ligne. <http://www.itl.nist.gov /iad/894.02/related_projects/muc/>. Consulté le 22 juin 2006.

Debili, F. et H. Achour (1998). « Voyellation automatique de l'arabe », dans *Proceedings of the Workshop on Computational Approaches to Semitic Languages*. Montreal, Canada, p. 42-49.

De Meulder, F. et W. Daelemans (2003). « Memory-based named entity recognition using unannotated data », dans *Proceedings of the 7th conference on Natural language learning at HLT-NAACL*. Edmonton, Canada, vol. 4, p. 208-211.

De Roeck, A.-N. et W. Al-Fares (2000). « A Morphologically Sensitive Clustering Algorithm for Identifying Arabic Roots », dans *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*. Hong Kong, p. 199-206.

Fourour, N. (2002). « Nemesis, un système de reconnaissance incrémentielle des entités nommées pour le français », dans *Actes de TALN'2002*. Nancy, France, p. 265-274.

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

Gallippi, A. (1996). « Learning to Recognize Names Across Languages », dans *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*. Copenhagen, Denmark, p. 424-429.

Grishman, R. (1996). MUC-6. En ligne. <http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html>. Consulté le 10 juin 2006.

Grishman, R. et B. Sundheim (1996). « Message Understanding Conference - 6 : A Brief History », dans *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*. Copenhagen, Danemark, p. 466-471.

Habash, N. (2005). « Introduction to Arabic Natural Language Processing », dans *Summer School on Human Language Technology*. En ligne. http://www.clsp.jhu.edu/ws2005/calendar/documents/HabashJuly6.ppt. Consulté le 10 juin 2007.

Heddaya, A., Hamdy, W. et M.-H. Sherif (1985). « Qalam : A Convention forMorphological Arabic-Latin-Arabic Transliteration ». En ligne. <http://langs.eserver.org/qalam.txt>. Consulté le 15 Avril 2006.

HIWIT. (2007). Tutoriaux – Cours. En ligne. <http://www.hiwit.org/tutoriaux/php/regex.html>. Aznet SARL. Consulté le 5 septembre 2007.

Inxight. (2009). Le site Web de la compagnie Inxight. En ligne. <http://www.inxight.com/products/smartdiscovery/ee/index.php>. Consulté le 10 janvier 2009.

Jackson, P. et I. Moulinier (2002). *Natural Language Processing for Online Applications Text Retrieval, Extraction and Categorization*. Amsterdam, John Benjamins Publishing Co, 225 p.

Jonasson, K. (1994). *Le nom propre. Constructions et interprétations*. Louvain-la-Neuve, Champs Linguistiques, Editions Duculot, 256 p.

Karkaletsis, V., Spyropoulos, C. et G. Petasis (1999). « Named Entity Recognition form Greek Texts : the GIE Project », dans *Advances in Intelligent Systems : Concepts, Tools and Applications, Kluwer Academic Publishers*, p. 131-142.

Larkey, L.-S, Ballesteros L. et M.-E. Connell (2002). « Improving Stemming for Arabic Information Retrieval : Light Stemming and Co-occurrence Analysis », dans *Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval (SIGIR 2002)*. Tampere, Finland, p. 275-282.

LAS. (2009). Le site Web de la compagnie LAS. En ligne. <http://www.las-inc.com/>. Consulté le 28 janvier 2009.

Levenshtein, V. I. (1966). « Binary codes capable of correcting deletions, insertions, and reversals », *dans Soviet Physics Doklady*, vol. 10, n° 8, p. 707-710.

Mahfoudhi, A. (2002). « Agreement lost, agreement regained! A minimalist account of word order and agreement variation in Arabic », dans *California Linguistic Notes*, vol. 27 n° 2 (2002).

Maloney, J. et M. Niv (1998). « TAGARAB : A Fast, Accurate Arabic Name Recogniser Using High Precision Morphological Analysis » dans *Proceedings of the Workshop on Computational Approaches to Semitic Languages*. Montréal, Canada, p. 8-15.

Mcdonald, D.D. (1993). « Internal and external evidence in the identification and semantic categorization of proper names », dans *Corpus Processing for Lexical Acquisition*, sous la dir. de B. Boguraev et J. Pustejovsky, MIT Press. Cambridge(Mass.), p. 61–76.

Merialdo, B. (1995). « Modèles probabilistes et étiquetage automatique », *Traitement automatique des langues, traitements probabilistes et corpus*, vol. 36, n° 1-2, 1995, p. 7-22.

Mikheev, A., Grover, C. et M. Moens (1998). « Description of the LTG system used for MUC -7 », dans *Proceedings of 7th Message Understanding Conference* (MUC-7).En ligne. <http://www.itl.nist.gov/iad/894.02/related_projects/muc/>. Consulté le 7 novembre 2007.

MUC. (1991). *Proceedings of the 3rd Message Understanding Conference*, Morgan Kauffmann.

MUC. (1992). *Proceedings of the 4th Message Understanding Conference*, Morgan Kauffmann.

MUC. (1993). *Proceedings of the 5th Message Understanding Conference*, Morgan Kauffmann.

MUC. (1995). *Proceedings of the 6th Message Understanding Conference*, Morgan Kauffmann.

NIST. (2001). *Introduction to Information Extraction.* En ligne. <http://www.itl.nist.gov/iaui/894.02/related_projects/muc/index.html>. Consulté le 7 novembre 2007.

Paik, W., Liddy, E.D., Yu, E. et M. McKenna (1994). « Categorizing and Standardizing Proper Nouns for Efficient Information Retrieval », dans B. Boguraev, & J. Pustejovsly (eds), Corpus Processing for Lexical Acquisition, MIT Press, chapitre 4.

Paliouras, G., Karkaletsis, V., Petasis G. et C. D. Spyropoulos (2000). « Learning Decision Trees for Named-Entity Recognition and Classification », dans *Proceedings of the 14th European Conference on Artificial Intelligence*. Berlin, s.p.

Poibeau, T. (1997). *ECRAN : projet européen d'extraction d'information.* En ligne. <http://www.dcs.shef.ac.uk/research/ilash/Ecran/>. Consulté le 13 Avril 2007.

Poibeau, T. (1999). « Repérage des entités nommées : un enjeu pour les systèmes de Veille », dans *Actes du Colloque TIA '99 : Terminologie et Intelligence Artificielle.* Nantes, p. 43-51.

Poibeau, T. (2001). « Deconstructing Harry, une évaluation des systèmes de repérage d'entités nommées », *Revue de la société d'électronique, d'électricité et de traitement de l'information.* En ligne. <http://www-lipn.univ-paris13.fr/~poibeau/articles/ree.ps>. Consulté le 10 juillet 2007.

Renals, S., Gotoh, Y., Gaizauskas, R. et M. Stevenson (1999). « Baseline IE-NE experiments using the SPRACH/LaSIE system », dans Proceedings of DARPA Broadcast News Workshop. Herndon, Virginia, P. 47-50.

Riloff, E. et R. Jones (1999). « Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping », dans *Proceedings of the 16th national Conference on Artificial Intelligence (AAAI-99)*. Orlando, Florida, p. 474-479.

Sakhr. (2009). Le site Web de la compagnie Sakhr. En ligne. <http://sakhr.com/>. Consulté le 19 janvier 2009.

Samy, D. (2005). « Named Entities : Structure and Translation. A study based on a Parallel Corpus (Arabic-English-Spanish) », dans Recent Advances in Natural Language Processing (*RANLP-2005)*. En ligne. http://www.corpus.bham.ac.uk/PCLC/NamedEntitiesParallelCorpus.doc. Consulté le 30 juillet 2007.

Samy, D., Moreno, *A.* et J.-M. Guirao (2005). « A Proposal for an Arabic Named Entity Tagger Leveraging a Parallel Corpus » dans *International Conference RANLP.* Borovets, Bulgaria, p. 459-465.

Senellart, J. (1998). « Locating Noun Phrases with Finite State Transducers », dans *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and Proceedings of the 17th International Conference on Computational Linguistics*. Montréal, Québec, p. 1212-1219.

Shaalan F. K. et R. Hafsa (2008). « Arabic Named Entity Recognition from Diverse Text Types », *dans Proceedings of the 6th International Conference GoTAL*. Gothenburg, Sweden, p. 440-451.

SRA. (2009). Le site Web de la compagnie SRA. En ligne. <http://www.sra.com/netowl/entity-extraction/>. Consulté le 15 janvier 2009.

United Nations Statistical Division, (2002). *Glossary of terms for the standardization of geographical names*, United Nations. New York, United States, p.63-65.

Van-Rijsbergen, C. (1979). *Information Retrieval*. 2nd edition, London, Butterworths, 208 p.

Vergyri, D., Kirchhoff, K., Duh, K. et A. Stolcke (2004). « Morphology-Based Language Modeling for Arabic Speech Recognition », dans *International Conference on Spoken Language Processing (ICSLP)*. Jeju Island, Korea, p. 2245-2248

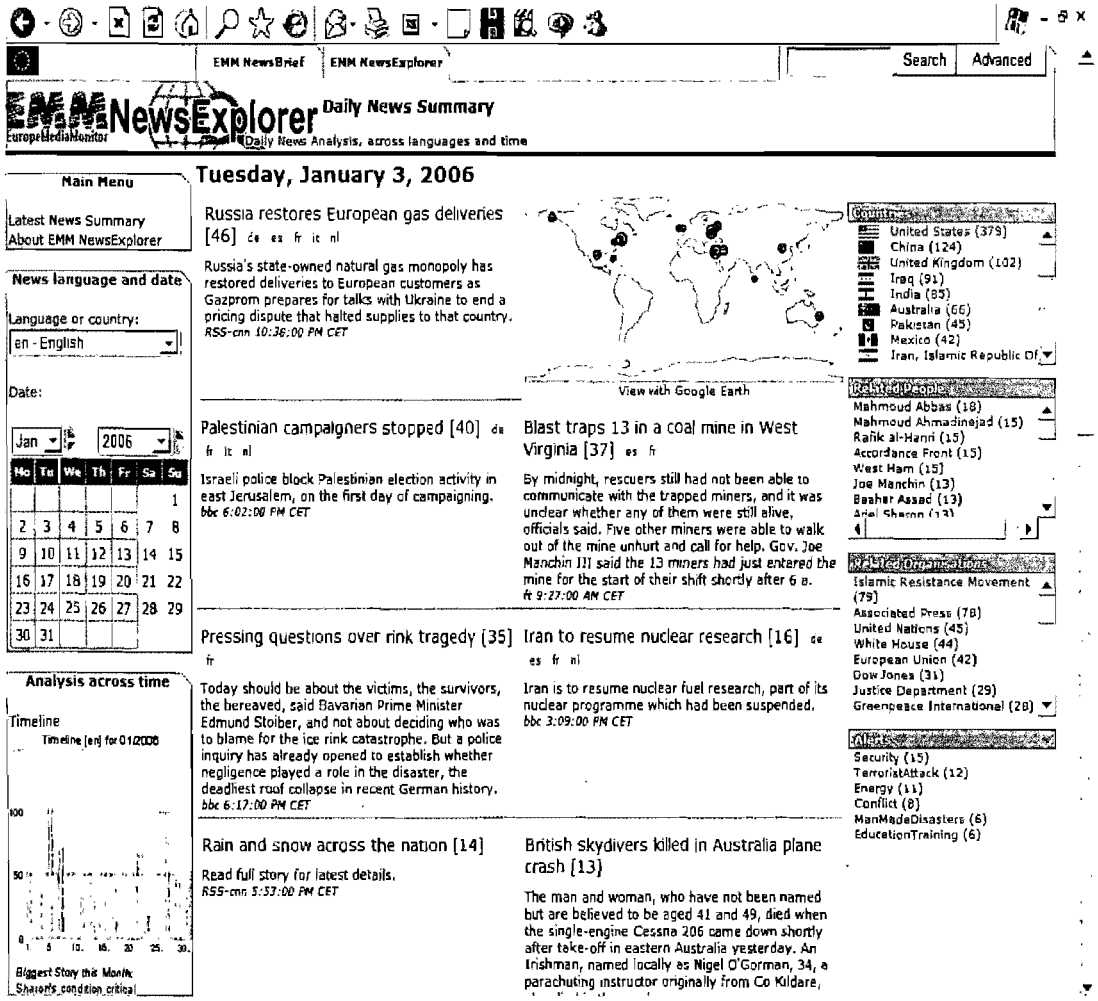
Wacholder, N., Ravin, Y. et M. Choi (1997). « Disambiguation of Proper Names in Text », dans *Proceedings of the 5th Conference on Applied Natural Language Processing*. Washington, D.C., p. 202-208.

Zhou, G.-D, Su, J. et T.-T. Guan (2000). « Hybrid Text Chunking », dans *Proceedings of CoNLL'2000*. Lisbon, Portugal, p. 163-165.

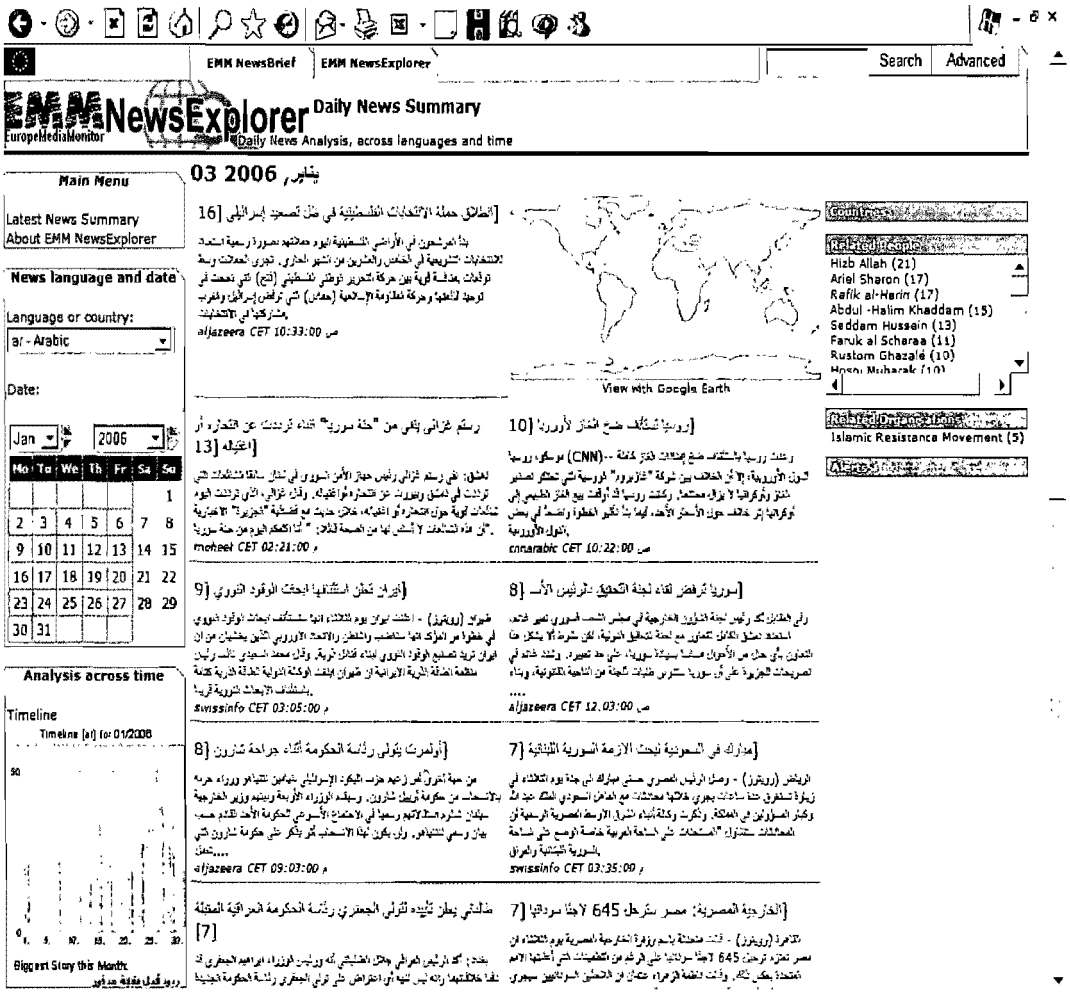
Annexes I -Illustration du système QALAM de translittération de l'arabe

Lettre	Qalam
ء	'
ا	Aa
ب	B
ت	T
ث	Th
ج	J
ح	H
خ	Kh
د	D
ذ	Dh
ر	R
ز	Z
س	S
ش	Sh
ص	S
ض	D
ط	T
ظ	Z
ع	`
غ	Gh
ف	F
ق	Q
ك	K
ل	L
م	M
ن	N
ه	H
و	W
ي	Y
ة	h,t
ى	Ae
لا	La

Annexes II -Illustration de l'environnement EMM



Annexes III -Illustration de RENAR



Annexes IV -Extrait du fichier de règles pour les noms de personnes

```
# first name + first name
\b${firstnames_1}\b${firstnames_3}\b
\b${firstnames_3}\b${firstnames_4}\b
${firstnames_1}\b${firstnames_3}\b${firstnames_3}\b
\b${firstnames_6}\b${firstnames_7}\b${firstnames_8}\b
\b${firstnames_5}\b${firstnames_6}\b${firstnames_7}\b${firstnames_8}\b

# first name + last name
${firstnames}\b+${lastnames}
#first + max4words + last name
${firstnames}\b+(\w+\b){0,4}${lastnames}
${firstnames}\b+${middlenames}\b+${firstnames}\b+(\w){5,16}
${firstnames}\b+(^ببال|^وال|^فال|^بال|^لال)(\w){4,16}
${firstnames}\b+(${middlenames}\b+)(\w){5,16}

#### first + max4w + middle + max4w + last
${firstnames}\b+(\w+\b){0,4}(${middlenames}\b+)+(\w+\b){0,4}${lastnames}
${firstnames}\b+(\w+\b){0,4}(${middlenames}\b+)+(\w+\b)
# title + something + last names
${^titles}\b+\w+\b+${lastnames}+
${^titles}\b+\w+\b+(${middlenames}\b+)*${lastnames}+
```

Annexes V -Extrait du fichier de règles pour les noms de lieux

```
# Arabic preffixes rules
# all the variations of AL in arabic
^(ال|أل|آل|و ال|و الأ|ب|ول|بال|ل|لل|ا|الإ|الأ|الآ|الا|و)^          #alirak -> irak
(ي|يون|ون|ويتان|يين|يان|ة|ية|يتان|يات)$          # irakian ->irak


#Tunisian --> Tunisia  or Tunisians --> Tunisia
##Other rules  Alcanadioun -> Canada

^($(ال|لل|فال|كال|و ال)?(.+)(ية|ي|يون|ين|ة|ويتان|يات|ون|يين|يان|يتان)$)^          2$|2$
^لل          ال
^إ          ا|إ
^ا          إ|أ
^أ          ا|أ
```

Annexes VI -Extrait du fichier de règles pour les noms d'organisations

```
<nameshape>
#######Rule 5
(ال|الأ)?${firstnames}\b${firstnames}
(ال|الأ)?${firstnames}\b
(ال|الأ)?${firstnames}

#######Rule 6
${cues_org}\b{name_org}\b{cues_org }
${cues_org}\b{firstnames}\b{patronym}\b{lastnames}\b{cues_org}
######Rule 7

(ال|الأ)?${middlenames}\b((و|ل|ب)?${middlenames_2}|${gentils}(و|ل|ب)?\w{3,16})
\b((و|ل|ب)?${middlenames_3}|${gentils}|${middlenames_Geo})\b((و|ل|ب)?${midd
lenames_4}|${gentils}(و|ل|ب)?\w{3,16}|${middlenames_Geo})\b((و|ل)?${middlena
mes_5}|${gentils})\b
######Rule 8
(ال|الأ)?${middlenames}\b((و|ل|ب)?${middlenames_2}|${gentils})\b((و|ل|ب)?${mid
dlenames_3}|${gentils}|${middlenames_Geo})\b((و|ل|ب)?${middlenames_4}(
و|ل|ب)?\w{3,16}|${gentils}|${middlenames_Geo})\b((و|ل)?${middlenames_5}|${ge
ntils})\b((ب|ل)?${middlenames_6}|(ب|ل)?${gentils})\b
</nameshape>
```

Annexes VII -Extrait du fichier de règles pour les entités temporelles et les entités numériques

```
<prep>
        of => من
        on => في
</prep>
<units>
        day => يوم اليوم
        month => شهر الشهر
        year => سنة عام السنة العام
        date => التقويم تاريخ
<units>

<ynb>
        1900 => ألف و تسعمائة  أل_ف و_تسعمئة
        2000 =>  ألفين
        1400 =>ألف_و_أربعة_مئة  ألف_و_أربع_مئة
</ynb>
<holidays>
        04.07  =>      الاستقلال_عن_بريطانيا      #indepedance day USA
        00.00  =>      الجمعة_المقدس              #Holy Friday
        00.00  =>      العاشر_من_ذي_الحجة      العاشر_من_شهر_ذي_الحجة #10th of zull al
hajja
        00.00  =>      العيد_لوطني                                #national day
        00.00  =>      المولد_النبوي الشريف عيد_المولد_النبوى      المولد_النبوي
        المولد_النبوي عيد_ميلاد_الرسول #Mawlid nabawi
        00.00  =>      اليوم_التذكاري                  #rememberance day
        00.00  =>      إعلان_قيام_الجمهورية  يوم_الجمهورية  ذكرى_إنشاء_الجمهورية
        اعلان_الجمهورية          #republic day
```

```
        00.00  =>  ذكرى_لثورة                    #revolution day
        01.01  =>  رأس_السنة    عيد_رأس_السنة_الميلادية    #new year
        00.00  =>  عاشوراء              #Ashoura
        00.00  =>  عيد_الأضحى    عيد_الأضحى_المبارك    #Aid Al Idha
        00.00  =>  عيد_الفطر    عيد_الفطر_المبارك    #Aid Al Idha
        00.00  =>  عيد_الاستقلال    احتفاء_بالاستقلال    #indepedance day
        00.00  =>  عيد_الفصح    اثنين_الفصح    #Thanksgiving
        25.12  =>  يوم_الميلاد    كريستماس    #Christmas
        00.00  =>  الأول_من_محرم              #First of muharram
        00.00  =>  الخميس_المقدس              #holy Thursday
        00.00  =>  يوم_مارتن_لوثر_كينغ              #martin lutherking day
        00.00  =>  عيد_الأم          #mother day
        25.12  =>  هانوكا  الهانوكا          #hanouka

</holidays>
```

Annexes VIII -Extrait du dictionnaire des entités temporelles et les entités numériques

<temperature>
درجة مئوية
فهرنهايت
سلسوس
</temperature>
<thermal_conductivity>
وات لكل متر
لكل متر وات
وات في المتر
</thermal_conductivity>
<time>
يوما
أيام
ساعات
دقيقة
دقائق
ثانية
</time>
<volume>
برميل
الباريل
باريل
لتر
لترات
أوقية
</volume>