UNIVERSITÉ DU QUÉBEC À MONTRÉAL

EXTRACTION DES ENTITÉS NOMMÉES PAR PROJECTION CROSS-LINGUISTIQUE ET CONSTRUCTION DE LEXIQUES BILINGUES D'ENTITÉS NOMMÉES POUR LA TRADUCTION AUTOMATIQUE STATISTIQUE

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN INFORMATIQUE

PAR

FATIMA DEFFAF

MARS 2015

REMERCIEMENTS

Je tiens tout d'abord à remercier ma directrice de recherche, Mme Fatiha Sadat, professeure à l'Université du Québec à Montréal (UQAM) d'avoir dirigé ce mémoire, pour son soutien, son aide et son encouragement durant toute la période de ce travail.

Un grand merci à Rahma Sellami, doctorante en informatique au laboratoire MIRACL à l'Université de Sfax (Tunisie) pour l'aide et les conseils qu'elle m'a apportés pour ce travail.

Je tiens également à remercier les professeurs du département d'informatique de l'UQAM pour la qualité de leur enseignement lors de ma maîtrise.

Je remercie Emad Mohamed, Mohamed Mahdi Boudabous, Wajdi Zaghouani, Samira Feddag, Habiba Chakour et les autres personnes qui m'ont aidé avec des conseils durant la période de ce travail.

Mes vifs remerciements vont également aux membres du jury pour l'intérêt qu'ils ont porté à mon travail en acceptant d'examiner ce mémoire et de l'enrichir par leurs propositions.

J'adresse aussi mes remerciements et ma profonde gratitude à mon mari Mohamed qui n'a jamais cessé de me soutenir pour que je puisse finir mes études. Merci aussi à toute ma famille.

Enfin un remerciement spécial à mes enfants Aymen, Rym et Ayoub qui sont ma source de bonheur.

TABLE DES MATIÈRES

LISTE DES FIGURES ix
LISTE DES TABLEAUX xi
LISTE DES ACRONYMES xv
RÉSUMÉ xvii
CHAPITRE I
INTRODUCTION 1
1.1 Problématique 2
1.2 Objectifs 3
1.3 Contribution 4
1.4 Organisation du mémoire 4
1.5 Règle d'écriture en langue arabe 5
CHAPITRE II
CONCEPTS DE BASE 7
2.1 La langue arabe 7
2.1.1 Aperçu général de la langue arabe 7
2.1.2 Caractéristiques et orthographe de la langue arabe 8
2.2 Les entités nommées 13
2.2.1 Définition des entités nommées 13
2.2.2 Rôle de l'entité nommée 14
2.2.3 Les formes des entités nommées 15
2.2.4 Reconnaissance des entités nommées 15
2.2.5 Classification des entités nommées 17
2.2.6 Lexème 21
2.2.7 Marqueur lexical 21
2.2.8 Lexique bilingue d'entités nommées 22

2.2.9 Translittération des entités nommées 22
2.2.10 Métriques d'évaluation des entités nommées 24
2.3 Structure et caractéristiques des entités nommées en arabe 24
2.3.1 Entités nommées de type nom de personne (EN-PERS) 25
2.3.2 Entités nommées de type nom de lieu (EN-LOC) 26
2.3.3 Entités nommées de type noms d'organisation (EN-ORG) 27
2.4 La traduction automatique 27
2.4.1 Définition de la traduction automatique 28
2.4.2 Paradigmes de la traduction automatique 28
2.4.3 Traduction automatique statistique 29
2.4.4 Le modèle de langage 30
2.4.5 Le modèle de traduction 32
2.4.6 Décodeur 36
2.4.7 Évaluation du système de TAS 36
2.5 Conclusion 37
CHAPITRE III
ÉTAT DE L'ART 39
3.1 Extraction des entités nommées 39
3.1.1 Approche symbolique (à base de règles) 40
3.1.2 Approche basée sur les corpus parallèles ou comparables 42
3.1.3 Approche par apprentissage machine 46
3.1.4 Extraction des entités nommées à partir de Wiképdia 50
3.1.5 Approche hybride 51
3.2 Traduction automatique 52
3.2.1 Les problèmes de la traduction automatique statistique 53
3.2.2 Traduction depuis ou vers la langue arabe 54
3.3 Conclusion 55
CHAPITRE IV
MÉTHODOLOGIE 57
4.1 Introduction 57
4.2 Extraction des entités nommées 58

4.2.1 Architecture générale de la solution proposée 58
4.2.2 Prétraitement des corpus 59
4.2.3 Extraction des entités nommées à partir du corpus source 61
4.2.4 Normalisation des phonèmes arabes 61
4.2.5 Algorithme d'extraction des entités nommées arabes à partir du corpus source 65
4.2.6 Méthode de projection cross-linguistique 66
4.2.7 Dictionnaire de marqueurs lexicaux 73
4.2.8 Translittération des entités nommées simples 75
4.2.9 Translittération des entités nommées composées 77
4.2.10 Annotation du corpus cible et construction de lexiques d'entités nommées 82
4.2.11 Comparaison de notre méthode de translittération avec l'état de l'art 83
4.3 Traduction automatique statistique 84
4.3.1 Translittération des entités nommées de l'arabe vers le français 84
4.3.2 Description du système de TAS 86
4.3.3 Les configurations du système de TAS 88
4.4 Conclusion 90
CHAPITRE V
ÉVALUATIONS 91
5.1 Données de l'évaluation 92
5.1.1 Corpus des Nations Unies UN 92
5.1.2 Les titres des articles de Wikipédia 93
5.1.3 Taille des lexiques bilingues d'entités nommées construits 94
5.2 Première évaluation : précision, rappel et F-mesure 94
5.2.1 Évaluation pour le corpus UN 94
5.2.2 Évaluation pour les titres Wikipédia 97
5.3 Deuxième évaluation – comparaison par rapport à Google Translate 98
5.3.1 Évaluation pour l'échantillon du corpus UN 99
5.3.2 Évaluation pour l'échantillon de titres de Wikipédia 102
5.4 Troisième évaluation – évaluation de lexique obtenu à partir de Wikipédia .. 104

5.5 Quatrième évaluation : Intégration des lexiques construits dans un système de TAS 105

5.5.1 Construction de lexiques d'ENs à partir de ressources linguistiques 105

5.5.2 Intégration des lexiques bilingues d'ENs dans le système de TAS 108

5.6 Conclusion 109

CHAPITRE VI

CONCLUSION ET PERSPECTIVES 111

PUBLICATIONS 115

APPENDICE A

TABLE DE TRANSLITTÉRATION DE L'ARABE D'APRÈS LA NORME DE BUCKWALTER 117

APPENDICE B

ÉCHANTILLON DE CORPUS PARALLÈLES ANNOTÉS 119

BIBLIOGRAPHIE 123

LISTE DES FIGURES

Figure		Page
2.1	Exemple 1 d'alignement à base de mots	33
2.2	Exemple 2 d'alignement à base de mots	33
2.3	Exemple d'alignement au niveau de phrases	35
2.4	Exemple d'une table de traduction (français - anglais)	35
4.1	Architecture de la solution d'extraction des ENs en arabe	59
4.2	Exemple d'une table de traduction (arabe-français)	88
5.1	Rappel, précision et F-mesure pour les ENs du corpus UN	95
5.2	Rappel, précision et F-mesure pour les ENs des titres de Wikipédia	98

LISTE DES TABLEAUX

Tableau		Page
2.1	Exemple d'ambiguïté d'écriture en arabe	8
2.2	Liste des signes diacritiques en arabe (Source (Zaghouani, 2009))	9
2.3	Exemples de phrases verbales en arabe	10
2.4	Exemple de phrase nominale en arabe	11
2.5	Exemple de segmentation de mots arabes	11
2.6	Les classes des proclitiques arabes	12
2.7	Les différentes écritures du phonème hamza	13
2.8	Exemples d'entités nommées	14
2.9	Exemple d'extraction des entités nommées	16
2.10	Exemple des codes d'annotation des entités nommées	17
2.11	Classes d'EN d'après la campagne MUC	18
2.12	Classes d'EN d'après la campagne ESTER	19
2.13	Classes d'EN selon Paik et al.	20
2.14	La classification d'EN retenue	21
2.15	Exemples de marqueurs lexicaux en langue arabe	22
2.16	Exemple de translittération	23
2.17	Exemples des formes d'EN-PERS d'origine arabe	25
2.18	Exemple de nombre de lexèmes d'EN en arabe et en anglais	26
2.19	Exemples de translittération des EN-ORG de l'anglais vers l'arabe	27
2.20	Exemple d'un modèle 5-grammes	31

4.1 Les phonèmes arabes 62

4.2 Normalisation des phonèmes arabes 64

4.3 Les translittérations possibles des phonèmes anglais vers l'arabe 68

4.4 Exemples des translittérations du phonème 'a' sans normalisation des phonèmes arabes 69

4.5 Exemples des translittérations du phonème 'a' après normalisation des phonèmes arabes 70

4.6 Exemple de Dic_Marqueur_Pers 74

4.7 Exemple de Dic_Marqueur_Loc 75

4.8 Exemple de Dic_Marqueur_Org 75

4.9 Exemple de translittération d'une EN simple 77

4.10 Exemple de translittération d'EN composée 81

4.11 Exemple de prétraitement avec MADA 87

4.12 Corpus d'apprentissage du modèle de traduction 89

5.1 Taille du corpus UN 93

5.2 Taille du corpus des titres de Wikipédia 93

5.3 Taille des lexiques bilingues d'ENs 94

5.4 Précision, rappel et F-mesure pour les ENs du corpus UN 95

5.5 Précision, rappel et F-mesure pour les ENs de titres de Wikipédia 97

5.6 Résultats obtenus par notre méthode de translittération 99

5.7 Résultats obtenus par Google Translate 99

5.8 Exemples des EN-PERS (du corpus UN) non traduites par Google Translate 101

5.9 Exemples d'EN-LOC mal ou non traduite par Google Translate 101

5.10 Résultats obtenus par notre méthode de translittération 102

5.11 Résultats obtenus par Google Translate 102

5.12 Exemples des EN-PERS non traduites par Google Translate 103

5.13 Comparaison de lexique des EN-PERS de titres Wikipédia avec d'autres lexiques 104

5.14 Taille du lexique des EN-PERS et EN-LOC construits à partir de ressources linguistiques 107

5.15 Score BLEU dans chaque évaluation 108

5.16 Taux des MHV dans chaque évaluation 108

A.1 Code Unicode et translittération des phonèmes arabes par le système Buckwalter 117

LISTE DES ACRONYMES

EN	Entité nommée
TALN	Traitement automatique du langage naturel
TAL	Traitement automatique des langues
EI	Extraction d'information
UN	Nations Unies (en anglais, United Nation)
EN-PERS	Entité nommée de type nom de personne
EN-LOC	Entité nommée de type nom de lieu
EN-ORG	Entité nommée de type nom d'organisation
MHV	Mot hors-vocabulaire (en anglais, Out Of Vocabulary)
TA	Traduction automatique
TAS	Traduction automatique statistique
BLEU	Score d'évaluation des systèmes de traduction automatique (en anglais, Bilingual Evaluation Understudy)

RÉSUMÉ

Ce mémoire présente une méthode d'extraction des entités nommées par projection cross-linguistique (projection inter langues ou d'une langue à une autre) en utilisant des corpus parallèles bilingues. Cette méthode consiste à automatiser la reconnaissance des entités nommées en une langue cible en exploitant des outils linguistiques d'une autre langue source.

Notre intérêt porte sur une langue à morphologie complexe, l'arabe, qui présente de grands défis en traitement automatique des langues naturelles.

La méthode de projection cross-linguistique proposée est basée sur un modèle de translittération (traduction phonétique) pour chaque entité nommée à partir de la langue source vers la langue cible. Cette méthode permet de construire des lexiques bilingues d'entités nommées.

Pour tester la performance de notre proposition, nous avons appliqué notre méthode sur un corpus extrait de Wikipédia et sur le corpus des Nations Unies (UN).

Les évaluations réalisées étaient basées dans un premier temps sur les métriques classiques, qui sont : la précision, le rappel et la F-mesure. La comparaison de nos résultats avec ceux de Google Translate montre l'utilité de la translittération des entités nommées de type nom de personne et nom de lieu.

Dans un second temps, nous avons intégré les lexiques bilingues construits dans un système de traduction automatique statistique. L'évaluation a été faite par le calcul de la valeur de score BLEU et le taux des mots MHV (mot hors-vocabulaire). Les résultats ont montré une augmentation du score BLEU et une diminution du nombre des mots MHV, ce qui illustre la performance de la procédure de translittération dans la situation où les données de test contiennent un nombre important d'entités nommées qui correspondent aux mots MHV.

Mots-clés : Traitement automatique du langage naturel, entité nommée, reconnaissance des entités nommées, annotation, projection cross-linguistique, translittération, lexique, traduction automatique statistique.

CHAPITRE I

INTRODUCTION

L'extraction d'information (EI) qui est un sous-domaine du traitement automatique du langage naturel (TALN) consiste à extraire automatiquement, à partir de données non (ou semi) structurées, des informations structurées pertinentes pour une tâche particulière[1].

Le développement rapide des technologies de l'information et de la communication et l'augmentation des volumes de données ont ouvert d'autres opportunités de manipulation de l'information. De ce fait, les recherches en EI sont devenues de plus en plus nombreuses (Elsa, 2006).

Dans ce mémoire, nous nous intéressons à l'une des sous-tâches de l'EI qui est la reconnaissance des entités nommées (ENs) (Chinchor, 1997). Cette dernière est devenue très utile pour la recherche et les applications en TALN, notamment pour la traduction automatique (TA) et la recherche d'information (Che et al., 2013). Les ENs sont des mots particuliers qui peuvent désigner les noms propres (noms de personne, noms de lieu et noms d'organisation), les expressions numériques et les expressions temporelles (Tjong Kim Sang, 2002).

Les deux approches pour l'extraction des ENs les plus utilisées sont l'approche symbolique qui se base sur l'utilisation des règles écrites manuellement et des dictionnaires d'EN, et l'autre est basée sur des méthodes statistiques comme

[1] http://en.wikipedia.org/wiki/Information_extraction

l'apprentissage machine (Maâloul, 2012). Avec l'apparition des corpus parallèles et comparables, une autre méthode d'utilisation de l'approche symbolique a été proposée. Dans ce mémoire, nous nommons cette méthode d'extraction des ENs *l'approche basée sur les corpus parallèles ou comparables.*

La procédure d'extraction des ENs par l'approche basée sur les corpus parallèles ou comparables consiste à extraire les ENs du corpus source à l'aide des outils disponibles pour la langue source et ensuite à utiliser ces ENs pour détecter les ENs du corpus cible. Cela veut dire développer une méthode de projection entre les deux langues source et cible. Cette méthode est appelée la *projection cross-linguistique* (ou inter langue ou d'une langue à une autre)[2] (Yarowsky et Ngai, 2001).

Généralement, les chercheurs utilisent les corpus parallèles avec une langue source riche en terme de ressources linguistiques ou qui contient des indices complémentaires sur les ENs, et donc peut être utilisée pour l'extraction des ENs en d'autres langues (Che et al., 2013).

1.1 Problématique

Pour certaines langues comme l'anglais, on trouve de nombreux travaux qui se concentrent sur l'extraction des ENs (Nadeau et Sekine, 2007). Toutefois, pour la langue arabe qui a une morphologie riche et complexe (Attia, 2008) le défi est grand et il y a un manque d'outils et de ressources linguistiques pour le traitement automatique de cette langue (Oudah et Shaalan, 2013). Dans ce mémoire, nous apportons une solution partielle à ce problème, en produisant des corpus en arabe annotés (étiquetés) en EN et des lexiques bilingues d'ENs pour l'anglais et l'arabe. Pour réaliser cette solution, nous allons utiliser l'approche basée sur les corpus parallèles bilingues en

[2] Dans la suite de ce mémoire, nous utilisons le terme 'projection cross-linguistique' au lieu de 'projection d'une langue à une autre'

développant une méthode de projection cross-linguistique pour extraire les ENs en arabe.

Nous abordons aussi dans ce travail le problème de l'amélioration de la traduction automatique statistique (TAS). En effet, les systèmes de TA ont réussi d'obtenir de bons résultats sur certains couples de langues comme l'anglais et le français (Do, 2011). Cependant, la traduction depuis ou vers l'arabe reste un défi jusqu'aujourd'hui (Gahbiche-Braham et al., 2014; Mallat et al., 2014). Ainsi, comme les lexiques bilingues sont des ressources linguistiques très importantes et indispensables pour la traduction humaine et la TA (Goldman et Scherrer, 2012) nous avons exploité les lexiques bilingues d'ENs construits pour améliorer la performance d'un système de TAS. Ce dernier a été construit en utilisant le décodeur Moses (Koehn et al., 2007), ainsi que d'autres outils linguistiques disponibles en ligne.

1.2 Objectifs

Les objectifs principaux de ce mémoire sont :

- L'extraction des ENs en arabe à partir de corpus parallèles bilingues en utilisant une méthode de projection cross-linguistique basée sur une technique de translittération;
- La construction de lexiques bilingues d'ENs pour la paire de langues anglais-arabe;
- La production des corpus en arabe annotés par des ENs;
- L'exploitation de la ressource Wikipédia en la considérant comme un corpus parallèle pour extraire les ENs;
- L'évaluation des lexiques bilingues d'ENs construits par des mesures telles que le taux de précision, le taux de rappel, et la F-mesure, et la comparaison des résultats obtenus avec les résultats de Google Translate;
- L'intégration des lexiques bilingues d'ENs construits à partir de corpus parallèles bilingues et des lexiques bilingues construits à partir de ressources linguistiques dans un système de TAS arabe-français avec évaluations.

1.3 Contribution

Nos contributions portent sur trois aspects distincts. Le premier aspect est le développement d'une méthode de translittération des ENs d'une langue source (arabe) vers une langue cible (anglais). Le but principal de notre choix de la langue arabe comme langue cible est la contribution aux défis de recherches sur le traitement automatique pour cette langue à morphologie riche et complexe, comparée aux autres langues comme par exemple les langues latines.

Le deuxième aspect est l'extraction des ENs à partir de Wikipédia. Cette ressource est très riche d'ENs, car son contenu est disponible en plusieurs langues, et elle peut être exploitée comme un corpus parallèle ou comparable.

Le troisième aspect est la contribution à l'amélioration des systèmes de TAS en incorporant nos résultats dans un système de TAS qui a été introduit dans la campagne d'évaluation TRAD 2014[3].

1.4 Organisation du mémoire

Après l'introduction, les objectifs de ce mémoire et la contribution, nous présentons la structure de ce mémoire.

Le chapitre II expose les notions de base de notre travail. Il comporte quatre sections. La première section est une brève étude linguistique de la langue arabe. La deuxième section présente les différentes notions liées à la compréhension du domaine de l'extraction des ENs. La troisième section décrit les caractéristiques des ENs en langue arabe. La dernière section donne les notions liées à la TA.

[3] http://www.trad-campaign.org/

Le chapitre III est un état de l'art sur les systèmes d'extraction des ENs et la TA. Pour situer l'approche que nous avons suivie, nous présentons dans ce chapitre les diverses approches pour l'extraction des ENs.

Le chapitre IV décrit notre méthodologie pour l'extraction des ENs à partir de corpus parallèles ainsi que notre démarche pour la construction du système de TAS pour évaluer les lexiques bilingues d'ENs construits.

Le chapitre V se concentre sur les évaluations de notre travail. En premier lieu, nous décrivons les données utilisées. Par la suite, nous présentons les différentes évaluations qui ont été faites avec une discussion des résultats obtenus.

Le chapitre VI est une conclusion où nous résumons le travail présenté dans ce mémoire ainsi que des perspectives pour la suite de nos travaux.

1.5 Règle d'écriture en langue arabe

Dans ce mémoire, plusieurs exemples sont écrits en caractères arabes. Pour faciliter la compréhension de ces exemples, nous avons ajouté (entre parenthèses) après chaque mot ou phrase en arabe une translittération suivant la norme de Buckwalter[4] suivie d'une traduction en français si elle n'existe pas.

La translittération de Buckwalter a été développée à Xerox par Tim Buckwalter dans les années 1990. Elle consiste à représenter les caractères arabes en latin. Elle est strictement une à une, à la différence de la majorité des autres systèmes de romanisation qui ajoutent des informations morphologiques qui ne sont pas exprimées en caractères arabes[5]. La translittération de Buckwalter a été utilisée dans plusieurs travaux en

[4] http://www.qamus.org/transliteration.htm.

[5] https://open.xerox.com/Services/arabic-morphology/Pages/buckwalter-about

TALN (Habash et al., 2007). L'appendice A illustre la translittération de chaque phonème[6] arabe suivant la norme de Buckwalter.

[6] Un phonème est la plus petite unité phonétique dans une langue qui est capable de fournir une distinction de sens. http://www.thefreedictionary.com/

CHAPITRE II

CONCEPTS DE BASE

Ce chapitre expose un ensemble de concepts de base qui sont nécessaires pour la compréhension de ce mémoire. Nous commençons par l'illustration de quelques caractéristiques de la langue arabe. Ensuite, nous présentons les notions liées aux ENs ainsi que les particularités des ENs en arabe. Enfin, la notion de la TA est présentée.

2.1 La langue arabe

Notre mémoire porte sur l'extraction des ENs en langue arabe et il est destiné à un public francophone qui ne maîtrise pas forcément l'arabe, donc il est nécessaire d'en présenter certains aspects linguistiques de cette langue.

2.1.1 Aperçu général de la langue arabe

L'arabe est une langue originaire de la péninsule arabique. Elle appartient au groupe des langues sémitiques, et elle est pratiquée sous deux formes : l'arabe dialectal et l'arabe littéraire ou littéral. L'arabe dialectal est propre à chaque pays et il correspond à la langue parlée dans le monde arabe. L'arabe littéraire est la langue officielle des pays arabes, et elle est divisée en deux types[7]. Le premier type est l'arabe classique qui est le plus ancien et c'est la langue du Coran[8] et des textes religieux. Le deuxième type est l'arabe standard moderne (ASM) qui est fondé syntaxiquement, morphologiquement et phonologiquement sur l'arabe classique. Elle est la forme utilisée dans l'écrit et l'enseignement (Gahbiche-Braham, 2013).

---

[7] http://fr.wikipedia.org/wiki/Arabe

[8] Le Coran est le livre sacré de l'islam

2.1.2 Caractéristiques et orthographe de la langue arabe

- **L'alphabet de la langue arabe**

L'arabe est une langue écrite de droite à gauche et contient 28 phonèmes et des signes diacritiques. Les voyelles en arabe correspondent aux signes diacritiques. Le dictionnaire Reverso[9] définit un signe diacritique comme « l'altération graphique d'un caractère alphabétique entrainant le plus souvent une modification du son de la lettre qui le reçoit ». Donc les signes diacritiques en arabe sont des signes ajoutés aux phonèmes, au-dessous ou au-dessus, pour avoir une prononciation différente. L'utilisation des signes diacritiques n'est pas obligatoire dans les textes en arabe moderne mais parfois, cela peut poser une ambiguïté sémantique et/ou syntaxique de sens (Gahbiche-Braham, 2013). Le tableau 2.1 illustre quelques exemples de ce type d'ambiguïté.

Tableau 2.1 Exemple d'ambiguïté d'écriture en arabe

Phrase	**Sens en français**
كتب الأستاذ (ktb Al>stA*)	Ambiguïté
كَتَبَ الأُسْتَاذُ (kataba Al>NsotaA*o)	Le professeur a écrit
كُتُبُ الأُسْتَاذُ (kutubo Al>NsotaA*o)	Les livres du professeur
ساعد (sAEd)	Ambiguïté
سَاعِد (saAEid)	Nom de personne 'Said' ou le mot 'Bras'
سَاعَدَ (saAEada)	Le verbe 'Aider'

Les signes diacritiques sont divisés en trois classes qui sont les voyelles courtes, les voyelles casuelles (Tanwinn) et les signes de syllabation.

La liste des signes diacritiques en arabe est illustrée dans le tableau 2.2.

[9] http://dictionnaire.reverso.net/francais-definition/signe%20diacritique

Tableau 2.2 Liste des signes diacritiques en arabe (Source (Zaghouani, 2009))

Illustration en arabe	**Nom**	**Prononciation ou fonction**	**Position**	**Exemple**
Voyelles courtes				
◌َ	Fatha	[a]	sur le phonème	فَ (fa)
◌ُ	Damma	[u]	sur le phonème	فُ (fu)
◌ِ	Kasra	[i]	sous le phonème	فِ (fi)
Voyelles casuelles (Tanwin)				
◌ً	Tanwin fatha	[an]	sur le phonème	فً (fan)
◌ٌ	Tanwin damma	[un]	sur le phonème	فٌ (fun)
◌ٍ	Tanwin kasra	[in]	sous le phonème	فٍ (fin)
Signes de syllabation				
◌ّ	Shadda	Doublement de consonne	sur le phonème	دَرَّسَ (dar~asa -enseigner)
◌ْ	Soukoun	Absence de voyelle	sur le phonème	دَرْسٌ (darosN - leçon)

- **Allographe des caractères arabes**

En arabe, l'allographe[10] de la plupart des phonèmes diffère selon la position d'écriture. Il y a quatre allographes : indépendant, début, milieu et fin.

Exemple

Cet exemple (Guillemin-Lanne et al., 2007) montre les différents allographes pour le phonème arabe « ع » :

- En position indépendante, lorsque le caractère est seul dans le mot. زرع (zrE - il a semé) ;

- En position initiale d'un mot. عمل (Eml - il a travaillé) ;

- En position médiane. يعمل (yEml - il travaille) ;

- En position finale. اصبع (ASbE - un doigt) ;

[10] L'allographe d'un phonème est la forme de son écriture

- **Forme des phrases arabes**

La structure de la phrase arabe est très complexe. Attia (2008) a donné dans sa thèse une présentation détaillée de la structure de la phrase arabe.

Il y a deux types de la phrase arabe : la phrase verbale et la phrase nominale[11].

- **Phrase verbale**

Pour la phrase verbale, il y a deux formes qui sont :

1- le verbe + le sujet + le complément (VSO) : cette forme est la plus utilisée dans les corpus.

2- le sujet + le verbe + le complément (SVO).

Le sujet peut être un pronom personnel (أنا، أنت، هو، هي، نحن ....) un nom propre (par exemple un nom de personne), etc.

Le complément peut être un nom, un complément circonstanciel, etc.

Le complément peut être absent dans une phrase verbale.

Exemples : Le tableau 2.3 montre quelques exemples de phrases verbales en arabe.

Tableau 2.3 Exemples de phrases verbales en arabe

Phrase en arabe	**Translittération Buckwalter**	**Traduction en français**	**Sujet**	**Verbe**	**Complément**
يكتب التلميذ الدرس	yktb Altlmy* Aldrs	l'étudiant écrit la leçon	التلميذ	يكتب	الدرس
كتب التلميذ	ktb Altlmy*	l'étudiant a écrit	التلميذ	كتب	

- **Phrase nominale**

Pour la phrase nominale, il y a une seule forme qui est : le sujet + l'attribut.

Le sujet peut être un pronom personnel, un nom propre, etc.

L'attribut peut être un adjectif qualificatif indéterminé, un complément circonstanciel, etc.

[11] http://www.arabe-gratuit.fr/cours_types_phrases.php

Exemple : Le tableau 2.4 montre un exemple de phrases verbales en arabe.

Tableau 2.4 Exemple de phrase nominale en arabe

Phrase en arabe	Translittération Buckwalter	Traduction en français	Sujet	Attribut
هو مريض	hw mryD	Il est malade	هو	مريض

- **Absence de majuscules en arabe**

Contrairement aux langues latines, la langue arabe ne possède pas la notion de majuscule. Cela rend la reconnaissance des noms propres plus difficiles (Fehri, 2012).

- **Agglutination en arabe**

La langue arabe est une langue agglutinante, c'est-à-dire les articles, les pronoms et les prépositions peuvent être collés aux noms, aux adjectifs et aux verbes. Donc les préfixes et les suffixes d'un mot en arabe peuvent être attachés à la racine. À cause de l'agglutination, un mot arabe peut correspondre à une phrase en français ou en anglais. Par exemple, le mot arabe فسأل (fs>l) correspond à la phrase en français « Il a interrogé ». Donc, la compréhension du sens d'un mot en arabe nécessite sa racinisation[12] pour extraire les différents segments qui composent le mot en arabe. Cependant parfois, même la segmentation ne résout pas le problème. C'est le cas d'une segmentation non valide comme le montre l'exemple montré dans le tableau 2.5 (Guillemin-Lanne et al., 2007).

Tableau 2.5 Exemple de segmentation de mots arabes

Segmentation valide	Segmentation non valide
فسأل (fs>l / Il a interrogé) - ف (fa / puis) : la conjonction de coordination - سأل (s>l) : le verbe interroger	Le mot فتح (ftH / Il a ouvert) ne doit pas être ssegmenter car la segmentation donne : - ف (fa / puis) : la conjonction de coordination - تح (tH) : mot n'a aucun sens.

[12] La racinisation (stemming) est le processus de transformation des mots en leur racine

En effet, la segmentation est imposée si la phrase ou le mot arabe contient des proclitiques. Un proclitique est « un clitique phonétiquement attaché au mot suivant. Les proclitiques donnent leur accent au mot suivant »[13].

Dans sa thèse de doctorat, (Mesfar, 2008) a classé les proclitiques arabes en trois classes. Le tableau 2.6 illustre ces classes avec des exemples.

Tableau 2.6 Les classes des proclitiques arabes

Classe	**Exemples**
Les proclitiques réservées aux noms et adjectifs : - l'article de définition ال (al / le ou la) - les prépositions : بِ (bi / avec), لِ (li / pour), ك (ka/ comme)	الجامعة (AljAmEp / l'Université) للأستاذ (ll>stA* / pour le professeur) كالأمير (kAl>myr / comme le prince)
Les proclitiques réservées aux verbes : - La particule du subjonctif : نصب (nasb) لِ (li / pour) - La particule du futur : سـ (sa) - La particule de l'apocopé : جزم (gazm) لِ (li / pour)	سأذهب (s>*hb / je vais aller) ليكتب (lyktb / pour écrire)
Les proclitiques générales : - Les conjonctions de coordination : ف (fa / puis), و (wa / et) - L'article d'interrogation : أ (a / est ce que) - Le marqueur de corroboration : ل التأكيد (la Alt>kyd)	فكتب (fktb / puis il a écrit) وكتب (wktb / et il a écrit) أأكتب (>>ktb / est ce que j'écris)

L'agglutination est la principale caractéristique qui pose un problème pour la langue arabe. Pour résoudre ce problème, il faut passer par l'étape de prétraitement de la phrase arabe pour extraire tous les segments existants.

- **Le phonème hamza**

Le phonème hamza (ء) est un phonème particulier qui s'écrit de différentes manières : seul (ء) ou combiné (sur ou sous) avec d'autres phonèmes comme montre le tableau 2.7.

[13] http://fr.wiktionary.org/wiki/proclitique

Tableau 2.7 Les différentes écritures du phonème hamza

Ecriture du phonème hamza	Exemple
Sur le phonème ا (alif) : أ	أمل (>ml / espoir)
Sous le phonème ا (alif) : إ	إحتفال (<HtfAl / célébration)
Sur le phonème و(waw) : ؤ	مسؤولية (ms&wlyp / responsabilité)
Sur le phonème ى (yaa) : ئ	رئيس (r}ys / président)

- **La kashida (tatwil)**

La kashida est un symbole (petit trait -) qui peut être ajouté à certains phonèmes pour les allonger et augmenter la distance entre eux. Ce n'est pas un phonème de l'alphabet arabe et elle n'a aucune prononciation, elle est utilisée juste pour des raisons esthétiques.

Exemple :

Pour le mot رحيم (rHym / clément), après l'ajout de la kashida au phonème ح (H), il devient رحـــيم.

2.2 Les entités nommées

Dans cette section, nous commençons à définir l'EN ensuite, nous présentons quelques notions de base liées au domaine d'extraction des ENs.

2.2.1 Définition des entités nommées

Le terme *entité nommée* (EN) est apparu en 1996 au cours de la sixième conférence MUC (Conférences sur la compréhension de messages, en anglais, Message Understanding Conference)[14]. Une EN désigne les noms de tous les personnes, organisations et lieux dans un texte (Grishman et Sundheim, 1996).

[14] http://www.nlpir.nist.gov/related_projects/muc

Même s'il n'existe pas de définition standard des ENs, plusieurs chercheurs ont proposé des définitions différentes pour cette notion. Par exemple, Le Meur et al. (2004) ont donné la définition suivante : « les ENs sont des types d'unités lexicales particuliers qui font référence à une entité du monde concret dans certains domaines spécifiques notamment humains, sociaux, politiques, économiques ou géographiques et qui ont un nom (typiquement un nom propre ou un acronyme) ».

Goldman et Scherrer (2012) ont définit l'EN comme un mot ou un groupe de mots désignant une personne, une organisation ou entreprise, un lieu, une date ou encore une expression numérique.

Daille et al. (2000) ont illustré que la notion d'EN inclut les noms propres, ainsi que les gentilés, les personnages de légendes, les maladies ou les drogues qui ne sont pas toujours considérés comme des noms propres.

À partir de ces définitions, nous pouvons conclure que les ENs sont des termes spéciaux qui désignent des sens particuliers comme les noms de personne, les noms de lieu, les noms d'organisation, les dates, et les chiffres. Ces sens ont plusieurs appellations dans le monde de recherche en extraction des ENs comme par exemples les catégories, les classes, les types, etc. Le tableau 2.8 montre quelques exemples d'EN.

Tableau 2.8 Exemples d'entités nommées

Entité nommée	**Type**
Barack Obama	Nom de personne
USA	Nom de lieu
29 Janvier 2014	Expression temporelle
UNICEF	Nom d'organisation

2.2.2 Rôle de l'entité nommée

Les ENs présentent plusieurs avantages dans le domaine de la recherche en TALN.

Elles sont par exemple utiles pour le développement des systèmes de questions/réponses, les résumés automatiques, la recherche d'information, la TA, le Web sémantique, et la bio-informatique (Mansouri et al., 2008). Les ENs sont utilisées aussi pour la réduction du taux de *mots hors-vocabulaire* (MHV) (en anglais : out of vocabulary), ceci car une partie importante des mots MHV représente des ENs et des termes techniques (Abduljaleel et Larkey, 2003; Abdel Fattah et Ren, 2008). Les MHV sont les termes rencontrés dans l'entrée et qui ne sont pas présents dans le dictionnaire ou la base de données de termes connus d'un système[15]. Les MHV sont traités dans plusieurs travaux en TALN (Bach et al., 2007; Nwesri et al., 2007; Habash, 2008).

2.2.3 Les formes des entités nommées

Il y a deux formes d'EN : Les ENs simples et les ENs composées. Chaque forme est traitée différemment.

- **Les entités nommées simples**

Une EN *simple* est une EN qui est composée d'un seul terme, comme les noms de lieu 'Canada' et 'Égypte' ou le nom de personne 'Adam'.

- **Les entités nommées composées**

Une EN *composée* est une EN qui est composée de deux ou plusieurs termes, comme par exemple le nom de personne 'Adam Smith' et le nom de lieu 'Afrique du Sud'.

2.2.4 Reconnaissance des entités nommées

La reconnaissance des ENs est la tâche de rechercher des termes qui correspondent à des ENs dans un texte, et les associer avec le type (ou la classe) approprié. Cette tâche est réalisée par l'une des approches pour l'extraction des ENs. Nous donnons une présentation de ces approches dans le chapitre III (État de l'art).

---

[15] http://en.wiktionary.org/wiki/OOV

Exemple :

Les trois phrases suivantes sont extraites d'un corpus parallèle français-anglais-Arabe. La première phrase est en français, la deuxième est en anglais et la troisième est en arabe. Les ENs dans chaque phrase sont illustrées dans le tableau 2.9.

Phrase 1 : « Mohamed Al-Shenawi, directeur général de la section d'inspection, à la banque Misr-Roumanie a déclaré.... »

Phrase 2 : «Mohamed Al-Shenawi, Director General of the Inspection Section at Misr-Romania Bank, said ...»

Phrase 3 : « .. قال محمد الشناوي مدير عام قطاع التفتيش في بنك مصر رومانيا » (qAl mHmd Al$nAwy mdyr EAm qTAE Altfty$ fy bnk mSr rwmAnyA)

Tableau 2.9 Exemple d'extraction des entités nommées

Entité nommée	**Phrase en langue**	**Type**
Mohamed Al-Shenawi	Français	Nom de personne
Mohamed Al-Shenawi	Anglais	Nom de personne
محمد الشناوي (mHmd Al$nAwy)	Arabe	Nom de personne
Banque Misr-Roumanie	Français	Nom d'organisation
Misr-Romania Bank	Anglais	Nom d'organisation
بنك مصر رومانيا (bnk mSr rwmAnyA)	Arabe	Nom d'organisation

Les types des ENs sont sélectionnés selon la classification des ENs choisie. Plusieurs classifications d'EN ont été proposées par les conférences et les campagnes d'évaluation. Nous présentons dans la prochaine sous-section quelques-unes de ces classifications.

Le résultat de la procédure de reconnaissance des ENs correspond à l'annotation des ENs (Ehrmann, 2008), ce qui produit un texte ou un corpus *annoté* par les ENs.

Les codes couramment utilisés pour l'annotation d'un texte sont l'étiquetage, le parenthésage, le balisage et la classification (Nouvel, 2012).

Exemple

Le tableau 2.10 montre l'annotation de l'exemple « Le directeur général Mohamed Al-Shanawi » selon les quatre codes. Cet exemple contient l'EN « Mohamed Al-Shanawi » qui est du type nom de personne (noté PERS).

Tableau 2.10 Exemple des codes d'annotation des entités nommées

Code d'annotation	**Phrase annotée**
Étiquetage	Le directeur général {Mohamed Al-Shenawi, +PERS }
Parenthésage	Le directeur général [PERS Mohamed Al-Shenawi]
Balisage	Le directeur général <PERS>Mohamed Al-Shenawi</PERS>
Classification	Le directeur général/PERS Mohamed Al-Shenawi PERS/

2.2.5 Classification des entités nommées

- **Classification d'après les conférences MUC**

Les conférences MUC ont été organisées et financées par DARPA (en anglais, Defense Advanced Research Projects Agency) et NOSC (en anglais, Naval Ocean System Center) dans le but d'encourager la recherche et le développement en EI (Grishman et Sundheim, 1996).

De 1987 à 1998, sept conférences MUC ont eu lieu pour traiter du problème de l'EI (Grishman et Sundheim, 1996; Chinchor, 1997). Les données des participants étaient sous forme de messages et elles étaient évaluées sur des sujets particuliers.

La tâche d'extraction des ENs a été introduite à la sixième conférence (MUC-6) (Grishman et Sundheim, 1996), puis une classification en trois classes a été proposée à la dernière conférence MUC -7 (Chinchor, 1997). Le tableau 2.11 montre ces trois classes avec leurs sous-classes.

Tableau 2.11 Classes d'EN d'après la campagne MUC

Classe	**Sous-classe**
ENAMEX	Les noms propres (noms de personne, noms de lieu et noms d'organisation)
NUMEX	Les expressions numériques
TIMEX	Expressions temporelles (date, temps)

Exemple [16]:

La phrase suivante a été annotée par un système de reconnaissance d'EN utilisé lors de la conférence d'évaluation MUC : « Henri a acheté 300 actions de la société AMD en 2006. ». Après l'extraction des ENs qui se trouvent dans cette phrase, cette dernière est étiquetée avec des balises XML comme suit :

```
<ENAMEX TYPE="PERSON">Henri</ENAMEX> a acheté <NUMEX
TYPE="QUANTITY">300</NUMEX> actions de la société <ENAMEX
TYPE="ORGANIZATION">AMD</ENAMEX> en <TIMEX
TYPE="DATE">2006</TIMEX>.
```

Cette phrase contient les ENs suivantes : Henri, 300, AMD, 2006.

- **Classification d'après la conférence CoNLL**

CoNLL (en anglais, Conference on Natural Language Learning)[17] est la conférence annuelle organisée par SIGNLL (en anglais, ACL Special Interest Group on Natural Language Learning). C'est une conférence internationale sur le langage naturel et l'apprentissage machine[18].

En 2002, le sujet principal de la conférence CoNLL était la reconnaissance des ENs, et une classification des ENs en quatre classes a été proposée. Ces classes sont les trois

[16] http://fr.wikipedia.org/wiki/Reconnaissance_d'entit%C3%A9s_nomm%C3%A9es

[17] http://www.conll.org/

[18] CoNLL 2014 : http://www.clips.uantwerpen.be/conll/cfp.html

sous-classes de la classe ENAMEX de MUC plus une quatrième classe qui regroupe toutes les entités qui n'appartiennent pas aux trois classes précédentes (Tjong Kim Sang, 2002).

- **Classification d'après la campagne ESTER[19]**

ESTER est une campagne d'évaluation des systèmes de transcription enrichie d'émissions radiophoniques en langue française. La reconnaissance des ENs est l'une des tâches évaluées dans cette campagne (Gravier et al., 2004), et une classification en sept types d'EN a été proposée. Le tableau 2.12 illustre la description de ces types[20].

Tableau 2.12 Classes d'EN d'après la campagne ESTER

Classe	**Description**
Personne	Personne, humaine, animal, fonction et civilité.
Fonction	Politique, militaire, administrative, religieuse et aristocratique.
Organisation	Politique, éducative, commerciale, non commerciale, divertissement et média et géo administrative.
Lieu	Lieu géographique naturel, région administrative, axe de circulation, adresse (adresse postale, numéro de téléphone et fax, adresse électronique) et construction humaine.
Production humaine	Moyen de transport, récompense, œuvre artistique et production documentaire.
Date et heure	Date, heure.
Montant	Âge, durée, température, longueur, aire et surface, volume, poids, vitesse et valeur monétaire.

[19] Évaluation des systèmes de transcription enrichie d'émissions radiophoniques

http://www.afcp-parole.org/camp_eval_systemes_transcription/index.html

[20] Convention d'annotation, ESTER2, 2007 *http://www.afcp-parole.org/camp_eval_systemes_transcription/docs/Conventions_EN_ESTER2_v01.pdf*

- **Classification d'après Paik et al.**

Paik et al. (1996) ont introduit une classification des ENs en neuf classes. Cette classification a été obtenue à partir d'une étude du corpus « Wall Street Journal ». Ces classes sont illustrées dans le tableau 2.13.

Tableau 2.13 Classes d'EN selon Paik et al.

Classe	**Description**
Entités géographiques	Ville, port, aéroport, île, comté, province, pays, continent, région, mer et fleuves
Affiliation	Religion, nationalité
Organisation	Entreprise, types d'entreprises, administration, administration gouvernementale
Humain	Personne, fonction
Document	Document
Équipement	Logiciel, matériel, machine
Scientifique	Maladie, drogue, médicament.
Temporelle	Date, heure
Divers	Autres types d'EN

- **La classification retenue**

Dans notre mémoire, nous avons retenu les trois sous-classes de la classification de MUC qui sont illustrées avec une description et une notation dans le tableau 2.14. Ces trois classes représentent les noms propres et elles sont communes entre toutes les classifications présentées. Les noms propres jouent un rôle important dans le langage, et dans la linguistique (Molino, 1982), et leur reconnaissance est un problème récurrent dans le TALN (Daille et al., 2000).

Tableau 2.14 La classification d'EN retenue

Classe	Description	Notation
Noms de personne	Les noms ou les prénoms de personne de n'importe quelle origine.	EN-PERS
Noms de lieu	Les noms qui désignent un lieu comme les pays, les villes, etc.	EN-LOC
Noms d'organisation	Les compagnies, les associations, les ministères, les partis, etc.	EN-ORG

Remarque

Dans la suite de ce mémoire, nous utilisons la notation mentionnée dans le tableau 2.14 pour les trois classes d'EN retenues.

2.2.6 Lexème

En linguistique, le *lexème* (en anglais, token ou lexeme) est l'unité lexicale de sens et de son qui est fixe et non dérivable dans une langue et sans distinction flexionnelle [21].

2.2.7 Marqueur lexical

Les *marqueurs lexicaux* sont des termes ou des abréviations qui précèdent ou suivent les ENs. Les marqueurs lexicaux ont un rôle important pour la reconnaissance des ENs et ils permettent de déduire l'existence et le type de l'EN. Le tableau 2.15 montre quelques exemples de marqueurs lexicaux en arabe.

[21] http://fr.wiktionary.org/wiki/lexème

Tableau 2.15 Exemples de marqueurs lexicaux en langue arabe

Marqueur lexical	**Type d'EN à détecter selon le marqueur lexical**
السيد (Alsyd / Monsieur ou Mr.) الدكتور (Aldktwr / Docteur ou Dr.)	EN-PERS.
مدينة (mdynp /Ville) شارع ($ArE / Rue)	EN-LOC
شركة ($rkp / Compagnie) جمعية (jmEyp / Association)	EN-ORG

2.2.8 Lexique bilingue d'entités nommées

Le mot *lexique* désigne l'ensemble des mots décrivant la langue d'une communauté, d'une activité humaine ou d'un locuteur, etc. [22]

Dans ce mémoire, ce terme désigne un dictionnaire de noms propres. Les ENs reconnues (en langue source et en langue cible) sont mémorisées dans des lexiques qu'on les appelle *les lexiques bilingues d'ENs.*

2.2.9 Translittération des entités nommées

La translittération est la traduction phonétique de chaque graphème (la plus petite entité) d'un système d'écriture en un graphème ou en un groupe de graphèmes d'un autre système, indépendamment de la prononciation[23].

La translittération entre les langues qui utilisent des alphabets similaires et des prononciations semblables est généralement simple, mais elle devient une tâche difficile lorsqu'il s'agit de langues totalement différentes d'un point de vue morphologique (Al-

[22] Glossaire de Linguistique Computationnelle http://ldelafosse.pagesperso-orange.fr/Glossaire/L.htm#lexique

[23] http://fr.wikipedia.org/wiki/Transcription_et_translittération

Onaizan et Knight, 2002). Par exemple, la translittération entre la paire de langues anglais-français est plus simple par rapport à la translittération entre la paire de langue anglais-arabe ou anglais-japonais.

Exemple :

Le tableau 2.16 montre un exemple de la translittération du nom de personne 'Clinton' vers le français et vers l'arabe.

Tableau 2.16 Exemple de translittération

Graphème anglais	Graphème français	Graphème arabe
C	C	ك (k)
Li	Li	لي (ly)
N	N	ن (n)
To	To	تو (tw)
N	N	ن (n)

La translittération des ENs a plusieurs avantages comme l'alignement de mots[24], la construction des lexiques bilingues ou multilingues, la TA, etc.

L'avantage le plus important de la translittération est la traduction, car même avec les traducteurs automatiques les plus utilisés, comme Google Translate[25], on n'arrive pas à traduire quelques ENs, en particulier les noms de personne. Google Translate est le service de Google pour la traduction des textes basé sur la TAS. Il possède une mémoire de traduction d'environ 200 milliards de mots provenant des corpus des Nations Unies[26].

---

[24] L'alignement de mots est l'identification des relations de traduction entre les mots. http://en.wikipedia.org/wiki/Bitext_word_alignment

[25] https://translate.google.ca

[26] http://edouard-lopez.com/fac/SciCo%20-%20S5/TAL/projet/TAL%20-%20systran%20vs.%20google%20translate.pdf

Exemple :

La traduction de l'EN-PERS 'Abdel-ellah Balqzeez' par Google Translate[27] donne l'expression 'عبد الاله Balqzeez' qui n'est pas correcte, par contre la translittération de cette EN donne le nom en arabe 'عبد الإله بلقزيز' (Ebd Al<lh blqzyz) qui est la traduction correcte.

2.2.10 Métriques d'évaluation des entités nommées

Le rappel, la précision et la F-mesure (Van Rijsbergen, 1979) sont des mesures largement utilisées dans les évaluations en TALN (Grouin et al., 2011).

La précision est le pourcentage des résultats corrects parmi les résultats obtenus.

Le rappel est le pourcentage des résultats corrects parmi les résultats qu'on doit trouver.

La F-mesure est la combinaison de la précision et du rappel et leur pondération. La formule de la F-mesure est :

$$\text{F\_mesure} = \frac{2(\text{précision} * \text{rappel})}{(\text{précision} + \text{rappel})}$$

Pour le domaine de l'extraction des ENs, les taux de la précision et du rappel sont calculés selon les formules suivantes :

$$\text{Précision} = \frac{\text{Nombre d'ENs correctement reconnues}}{\text{Nombre d'ENs reconnues}}$$

$$\text{Rappel} = \frac{\text{Nombre d'ENs correctement reconnues}}{\text{Nombre d'ENs dans le corpus}}$$

2.3 Structure et caractéristiques des entités nommées en arabe

La morphologie complexe de la langue arabe rend la tâche de la reconnaissance des ENs en arabe plus difficile par rapport aux langues européennes. Par exemple, l'absence de

[27] Traduit avec Google Translate en juin 2014

la majuscule dans cette langue pose un problème pour la reconnaissance des noms propres. Aussi, l'agglutination des mots et l'absence des voyelles imposent la segmentation des phrases arabes pour l'extraction de la bonne information.
Dans cette section, nous présentons la structure et les caractéristiques des ENs en arabe.

2.3.1 Entités nommées de type nom de personne (EN-PERS)

La structure des EN-PERS en arabe diffère selon l'origine du nom. Zaghouani (2009) a présenté dans son mémoire de maîtrise une présentation détaillée des différentes structures des noms arabes selon leur origine.

Exemple : le tableau 2.17 montre quelques exemples des formes d'EN-PERS selon leur origine.

Tableau 2.17 Exemples des formes d'EN-PERS d'origine arabe

Origine du nom	**Forme du nom**
Égypte	Le prénom + le prénom du père (peut être allongé par l'ajout de (des) prénom (s) de (s) grand (s) père (s) pour lever l'ambiguïté avec une autre personne qui a le même prénom et le même prénom de père)
Maghreb	Le prénom + le nom de famille
Mauritanie	Le prénom + 'ولد' (wld / fils de) + nom de famille

Le nombre de lexèmes de certains noms de personne d'origine arabe écrits en anglais n'est pas le même nombre de lexèmes de ces mêmes noms écrits en arabe. C'est le cas des noms qui contiennent certains termes particuliers comme par exemple les termes : ابو (>bw - abu), عبد (Ebd - Abd), etc. Le tableau 2.18 montre quelques exemples.

Tableau 2.18 Exemple de nombre de lexèmes d'EN en arabe et en anglais

Écriture anglaise	**Nombre de lexèmes**	**Écriture arabe**	**Nombre de lexèmes**
Mohamed Al Shenawi	3	محمد الشناوي (mHmd Al$nAwy)	2
Dr. Ibrahim Al Demeiri	4	الدكتور ابراهيم الدميري (Aldktwr AbrAhym Aldmyry)	3
Abdulwahab Attar	2	عبد الوهاب عطار (Ebd AlwhAb ETAr)	3
Ayatollah Mohamed Ali	3	آية الله محمد علي (|yp Allh mHmd Ely)	4

Parfois, ce problème se pose aussi pour les noms de personne d'origine non arabe, comme par exemple le nom 'Michelangelo' qui est une EN simple et qui s'écrit en arabe 'مايكل انجلو' (mAykl Anjlw) qui correspond à une EN composée de 2 mots.

Ce point complique la procédure de la translittération des noms de personne de l'anglais vers l'arabe. En effet, la translittération d'un lexème écrit en langue source vers une langue cible donne un seul lexème, ce qui est faux pour le cas présenté, et donc la procédure de translittération nécessite un certain nombre de règles pour produire la traduction adéquate.

2.3.2 Entités nommées de type nom de lieu (EN-LOC)

L'extraction des EN-LOC en arabe est moins difficile par rapport à l'extraction des EN-PERS, car les noms de lieu qui représentent les noms de pays sont standards et il y a beaucoup de ressources disponibles pour les identifier. Les autres noms de lieu sont généralement associés aux marqueurs lexicaux qui aident à reconnaître que le type de l'EN est un lieu.
L'ordre des lexèmes d'une EN-LOC en arabe peut diffère à celui de l'EN-LOC en anglais.

Exemple : L'EN-LOC en anglais 'East Slavonia' commence par le lexème 'East' suivi de 'Slovania', par contre l'EN arabe qui est 'سلافونيا الشرقية' (slAfwnyA Al$rqyp / Slavonie orientale) commence par le lexème 'سلافونيا' (Slavonia) ensuite le lexème 'الشرقية' (East).

2.3.3 Entités nommées de type noms d'organisation (EN-ORG)

Comme les EN-LOC, quelques noms d'organisation sont combinés avec les marqueurs lexicaux qui aident leur reconnaissance comme par exemple les marqueurs 'parti', 'ministère', 'compagnie' et 'banque'. Il y a d'autres organisations qui se présentent sous forme d'acronymes. En arabe, l'usage des acronymes est relativement rare si on le compare aux langues européennes comme le français et l'anglais (Zaghouani, 2009).
Les EN-ORG simples peuvent être translittérés, mais il n'y a pas beaucoup d'EN-ORG composées qui peuvent être translittérées.

Exemples : le tableau 2.19 illustre quelques exemples de translittération des EN-ORG de l'anglais vers l'arabe

Tableau 2.19 Exemples de translittération des EN-ORG de l'anglais vers l'arabe

EN en anglais	**Translittération en arabe**
Translittération des ENs simples	
Toshiba	توشيبا
Delta	الدلتا
FIFA	الفيفا
Translittération des ENs composées	
Bretton Woods	بريتون وودز
Orascom Telecom	اوراسكوم تيليكوم
Central Bank	Ne peut pas être translittérée, sa traduction donne بنك المركزي

2.4 La traduction automatique

Dans cette section, nous donnons quelques notions liées à la TA. Nous commençons par la définition de la TA suivie d'une description abrégée des trois paradigmes de TA. Enfin, l'approche de la TAS est présentée ainsi que leurs composants.

2.4.1 Définition de la traduction automatique

La notion de *traduction automatique* (TA) désigne la traduction des textes d'une langue source (originale) vers une langue cible, en utilisant des programmes informatiques, appelés systèmes de TA, sans l'intervention humaine(Loffler-Laurian, 1996).

2.4.2 Paradigmes de la traduction automatique

Il y a trois paradigmes de la TA qui sont la TA à base de règles, la TA guidée par l'exemple ou la traduction par analogie (Nagao, 1984) et la TAS (Brown et al., 1990).

La TA à base de règles nécessite une connaissance approfondie de la langue source et la langue cible. Elle est basée sur l'utilisation de règles linguistiques et de dictionnaires volumineux[28]. Les dictionnaires et les règles sont propres à chaque paire de langues utilisée. L'exemple le plus connu de systèmes de TA à base de règles est SYSTRAN[29], il y a aussi les systèmes Ariane (Boitet et Guillaume, 1982) et Eurotra (Maghi, 1981)

La TA guidée par l'exemple considère la phrase comme l'unité de traduction idéale (Carl, 2003). Elle utilise une base de données qui contient des textes ou des corpus déjà traduits. Ensuite, elle cherche les meilleurs exemples existant dans cette base de données qui sont similaires à la phrase à traduire. Puis, l'adaptation est faite pour tenir en compte la différence qui existe dans la phrase à traduire (Carl, 2003).
Mymemory (Sato et Nagao, 1990) et Babelfish[30] sont deux exemples de systèmes de TA guidée par l'exemple.
La TAS a été proposée par Brown et al. (1990) dans les années 1990 et elle est basée sur les méthodes d'apprentissage automatique. Actuellement, la TAS est la plus utilisée

[28] http://www.systran.fr/systran/entreprise/technologie/traduction-automatique/

[29] http://www.systransoft.com/

[30] http://www.babelfish.com/

dans le domaine de TA (Costa-jussa et al., 2013). Google Translate est un exemple de ce type de système de TA.

2.4.3 Traduction automatique statistique

Un système de TAS nécessite un ensemble de corpus parallèles alignés[31] dans les deux langues source et cible, et des corpus monolingues en langue cible. Il est constitué de trois composants : le décodeur, le modèle de traduction et le modèle de langues (Brown et al., 1990).

Si on suppose que S est la langue source et C est la langue cible, le modèle de traduction probabiliste est appris sur des corpus parallèles bilingues et il est noté par P(S|C), et le modèle de langue est appris sur des corpus monolingues en langue cible et il est noté par P(C).

Le décodeur est la partie centrale de tout système de traduction et il est chargé de fournir la (ou les) meilleure (s) traduction (s) possible (s) (Lavecchia, 2010). Cela veut dire la recherche d'une phrase en langue cible C ayant la probabilité P(C|S), d'être la traduction d'une phrase en langue source S.

La décomposition de la probabilité P(C|S) (en appliquant le théorème de Bayes) permet de reformuler le problème de traduction comme indique la formule (2.1).

$$P(C|S) = (P(S|C) * P(C))/(P(S)) \qquad (2.1)$$

La traduction la plus probable (optimale) C^* de la phrase source S est la traduction qui a la probabilité P(C|S) la plus élevée. Elle est obtenue par l'application de l'argument du maximum sur P(C|S) et elle est donnée par la formule (2.2).

$$C^* = \text{argmax}_C\, P(C|S) = \text{argmax}_C\, (P(S|C) * P(C)/(P(S)) \qquad (2.2)$$

[31] Un corpus parallèle aligné est un corpus parallèle ou les unités textuelles des deux parties du corpus (source et cible) sont mises en correspondance.

Comme la probabilité de la phrase source P(S) est connue en avant et n'influe pas sur le calcul de la fonction argmax, on peut l'éliminer et donc on retrouve la formule (2.3) qui est utilisée pour l'entraînement d'un système de TAS.

$C^* = \text{argmax}_C\ P(S|C) * P(C)$ (2.3)

2.4.4 Le modèle de langage

Une langue peut être modélisée statistiquement par un modèle de langage. Si on a une suite de n mots $S = (M_1, M_2, M_3, \ldots, M_n)$, la probabilité d'apparition de la suite S, notée P(S), dans un texte dépend des probabilités d'apparition des mots M_i.

Après l'apprentissage du modèle de langage sur les corpus monolingues de la langue cible, une probabilité d'apparition, notée $P(M_i)$ est attribuée à chaque mot M_i. La probabilité $P(M_i)$ dépend de l'historique des mots M_i. Pour connaître cet historique, il faut avoir les probabilités d'apparition de tous les mots $(M_1, M_2, M_3, \ldots, M_{i-1})$ qui précèdent le mot Mi.

L'historique d'un mot M_i est calculé selon le type de modèle de langage utilisé. Par exemple, si on utilise un modèle unigramme[32] on n'a pas d'historique, et donc la probabilité d'un mot ne dépend que de lui-même. La probabilité P(S) est calculée par la formule (2.4) suivante :

$P(S) = P(M_1)* P(M_2)....* P(M_n)$ (2.4)

Pour un modèle bigrammes on prend en considération qu'un seul mot précédent. Ainsi, on obtient la formule (2.5) suivante :

$P(S) = P(M_1)* P(M_2|M_1).....* P(M_n|M_{n-1})$ (2.5)

---

[32] Le terme *gramme* est utilisé pour désigner le nombre de mots dans une suite, par exemple *unigramme* est une suite d'un seul mot, *bigramme* est une suite de deux mots et pour généraliser on utilise le terme *n-gramme* pour une suit de n mots.

Et ainsi de suite, pour un modèle n-gramme, on prend on considération les n-1 mots précédents, et donc on obtient la formule (2.6) :

$$P(S) = P(M_1)* P(M_2|M_1)* P(M_3|M_1, M_2)....* P(M_n|M_1, M_2, ..., M_{n-1}) \quad (2.6)$$

$P(M_3|M_1, M_2)$: la probabilité que la suite de mots (M_1, M_2) soit suivie du mot M_3.

$P(M_n|M_1, M_2, ..., M_{n-1})$: la probabilité que la suite de mots (M_1, M_2, ...M_{n-1}) soit suivie du mot M_n.

Le nombre de grammes n influe sur les résultats d'entraînement. Plus n est grand, les meilleurs sorts de résultats sont obtenus. Généralement les chercheurs utilisent un n varie entre 1 et 5 (Gahbiche-Braham, 2013; Le, 2013).

Dans la pratique, l'implémentation du modèle de langage est faite par différents outils, comme par exemple l'outil SRILM (Stolcke, 2002).

Exemple du modèle 5-grammes :

Un modèle 5-grammes décompose la phrase 'Les entités nommées sont utiles pour la traduction automatique.' en segments de 5 mots comme montre le tableau 2.20.

Tableau 2.20 Exemple d'un modèle 5-grammes

Les	entités	nommées	sont	utiles	pour	la	traduction	automatique	.
Les	entités	nommées	sont	utiles	pour	la	traduction	automatique	.
Les	entités	nommées	sont	utiles	pour	la	traduction	automatique	.
Les	entités	nommées	sont	utiles	pour	la	traduction	automatique	.
Les	entités	nommées	sont	utiles	pour	la	traduction	automatique	.
Les	entités	nommées	sont	utiles	pour	la	traduction	automatique	.

À partir de cette phrase composée de 10 lexèmes, les 5-grammes suivants sont construits :

1- Les entités nommées sont utiles
2- entités nommées sont utiles pour
3- nommées sont utiles pour la
4- sont utiles pour la traduction
5- utiles pour la traduction automatique
6- pour la traduction automatique.

2.4.5 Le modèle de traduction

La TAS nécessite un modèle de traduction qui permet de calculer les probabilités de traduction entre les mots, les suites de mots et les autres constituants de la phrase de la langue source vers la langue cible (Lavecchia et al., 2008).
Il y a deux types de modèles de traduction : le modèle de traduction à base de mots (Brown et al., 1990) et le modèle de traduction à base de segments[33]. Les mots et les segments sont appelés *les unités de la phrase*.

Dans la TAS, le modèle de traduction est appris sur un ensemble de corpus parallèles bilingues qui doivent être alignés au niveau des unités de phrases. L'alignement est l'identification des unités correspondantes dans les deux parties (en langue source et en langue cible) du corpus parallèle[34].

- **Modèle de traduction à base de mots**

Ce modèle est fondé sur les mots où la phrase à traduire est divisée en mots isolés et la traduction se faite mot par mot.
Dans ce modèle, l'alignement du corpus bilingue en langue source et cible se fait mot à mot. Cela veut dire que chaque mot dans la phrase en langue source a son équivalent dans la phrase en langue cible.

---

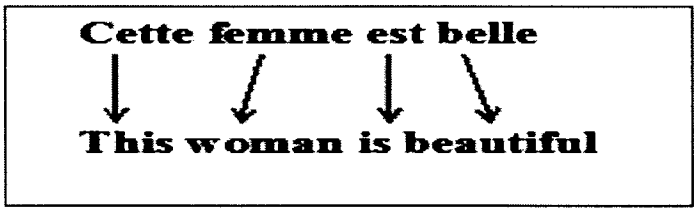
[33] Un segment peut être un mot, ou une séquence de mots contigus dans une phrase. https://interstices.info/jcms/nn_72253/la-traduction-automatique-statistique-comment-ca-marche

[34] http://en.wikipedia.org/wiki/Parallel_text

Exemples :

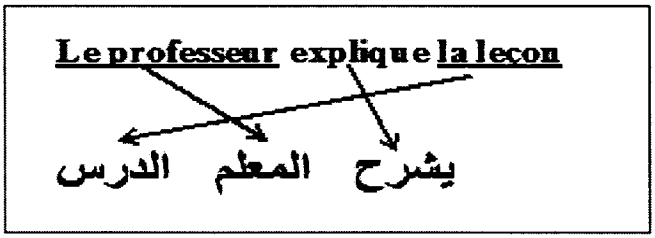
La figure 2.1 montre un exemple d'alignement à base de mots d'une phrase en français et sa traduction en anglais.

Figure 2.1 Exemple 1 d'alignement à base de mots



Dans ce premier exemple, l'ordre des mots est le même pour chacune des deux phrases. Prenons un deuxième exemple pour les langues arabe et française où l'ordre n'est pas le même pour les deux phrases (voir figure 2.2).

Figure 2.2 Exemple 2 d'alignement à base de mots



- **Les modèles IBM**

Brown et al. (1993) ont développé cinq modèles de traduction nommée IBM_1, IBM_2, IBM_3, IBM_4 et IBM_5. Chacun de ces modèles constitue à la fois un modèle de traduction et un algorithme d'alignement à base de mots. La différence entre ces modèles est les paramètres de calcul de la probabilité de traduction. Chaque modèle est une amélioration du modèle qui le précède.

Dans IBM_1, l'ordre des mots dans les deux phrases source et cible n'est pas pris en considération et donc la probabilité de traduction se définit par la traduction lexicale[35].

[35]Diapositives http://www.statmt.org/book/slides/04-word-based-models.pdf

IBM_2 améliore IBM_1 en ajoutant l'étude des positions des mots dans la phrase source et cible et donc les positions des mots affectent la probabilité de traduction. Dans IBM_3, la probabilité de traduction dépend aussi de la longueur des chaines sources et cibles, alors un mot source peut être traduit par plusieurs mots cibles. Dans IBM_4, la probabilité de traduction dépend aussi des mots sources et cibles déjà traduits. IBM_5 améliore le modèle IBM_4 avec un alignement raffiné pour corriger les déficiences.

GIZA++

Dans la pratique, il existe plusieurs outils d'alignement à base de mots. GIZA++(Och et Ney, 2003) est l'un de ces outils et il est le plus utilisé dans la TAS (Tian et al., 2011).

GIZA++ est un logiciel open-source qui est utilisé pour la construction d'alignement des cinq modèles d'IBM (Brown et al., 1993) et des modèles de Markov cachés (MMC) (Och et Ney, 2000). GIZA++ est une extension de GIZA[36] qui est l'un des modules du projet EGYPT (outil de TAS) développé par l'équipe de TAS en 1999 au Centre de traitement du langage et discours à l'Université Johns-Hopkins (Al-Onaizan et al., 1999).

- **Modèle de traduction à base de segments**

Le deuxième type du modèle de traduction est fondé sur les segments où la phrase à traduire est divisée en segments et la traduction se faite segment par segment.

Dans ce modèle, l'alignement du corpus parallèle bilingue se fait au niveau des phrases c'est-à-dire un alignement entre les segments de la phrase.
La figure 2.3 montre un exemple d'alignement au niveau des phrases pour une phrase en français et sa traduction en anglais (Brown et al., 1990).

[36] http://www-i6.informatik.rwth-aachen.de/Colleagues/och/software/GIZA++.html

Figure 2.3 Exemple d'alignement au niveau de phrases

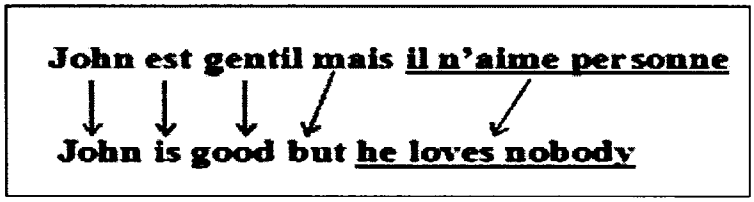


Table de traduction

Dans la TAS, suite à l'alignement des corpus parallèles bilingues, une table de phrases sources et leur traduction en langue cible est construite. Cette table est appelée *la table de traduction* (en anglais, phrase table). Elle est constituée d'un ensemble de lignes, où chaque ligne est composée de trois colonnes séparées par le symbole '|||'. La première colonne comporte l'unité (mot ou segment) en langue source, la deuxième comporte l'unité en langue cible et la dernière correspond aux probabilités de traduction P(S|C) selon le nombre de grammes utilisés dans le modèle de traduction.

La figure 2.4 montre un exemple d'une table de traduction du français à l'anglais.

Figure 2.4 Exemple d'une table de traduction (français - anglais)

la	|||	the	||| 0.6
la	|||	this	||| 0.4
femme	|||	woman	||| 0.5
femme	|||	wife	||| 0.5
la femme	|||	the woman	||| 0.6
a femme	|||	this woman	||| 0.4
de	|||	of	||| 0.6
ménage	|||	household	||| 0.8
femme de ménage	|||	maid	||| 1.0
a sauté	|||	jumped	|||0.85
a sauté	|||	skipped	|||0.15
un repas	|||	a meal	||| 1.0
sauté un repas	|||	skipped a meal	||| 1.0

2.4.6 Décodeur

La tâche de décodage consiste à chercher une phrase cible C qui est conforme avec le modèle donné et attribue à cette phrase la probabilité P(C|S) la plus élevée (Do, 2011). Cette tâche est réalisée par un décodeur qui utilise plusieurs paramètres pour calculer la fonction argmax dans la formule (2.3).

En 2004, Philipp Koehn a développé le décodeur Pharaoh (Koehn, 2004) qui était très utilisé pour la traduction jusqu'à 2007(Lavecchia, 2010) où Moses (Koehn et al., 2007), développé par Philipp Koehn et al, a pris la place de Pharaoh.

Moses

Moses (Koehn et al., 2007) est une boîte à outils de TAS à base de segments qui permet de former automatiquement des modèles de traduction pour une paire de langues, à l'aide de corpus parallèles. Moses est disponible en téléchargement gratuit sur le site Web http://www.statmt.org/moses/. La construction d'un système de TAS basé sur Moses nécessite plusieurs outils qui sont disponibles sur le site de Moses.

Il y a d'autres systèmes de décodage, comme Portage du CNRC (Sadat et al., 2005) et CDEC (Dyer et al., 2010). Mais Moses est le système le plus populaire (Kalyanee et al., 2014) et il est considéré comme un système de référence dans la communauté de TA probabiliste (Do, 2011), car il donne de bons résultats(Azab et al., 2013; Sellami et al., 2013; Gahbiche-Braham et al., 2014). Parmi les objectifs de ce mémoire l'évaluation des lexiques d'ENs dans la TAS en utilisant le décodeur Moses.

2.4.7 Évaluation du système de TAS

Pour dire qu'un système de TAS est performant ou non, il faut l'évaluer soit manuellement ou automatiquement. Cependant une évaluation manuelle prend beaucoup de temps, ce qui impose l'évaluation automatique. Parmi les métriques

d'évaluation automatique des systèmes de TAS il y a le score BLEU (Papineni et al., 2002).

Score BLEU

Le score BLEU (Papineni et al., 2002) (en anglais, BiLingual Evaluation Understudy) est une mesure très utilisée pour l'évaluation des systèmes de TA (Blain, 2013). Son principe est la comparaison des n-grammes de la traduction candidate avec les n-grammes de celle de la référence (ou les références) en calculant le nombre de correspondances. Si une TA est identique à l'une des références alors le score BLEU prend la valeur maximale qui est 100. Par contre, si aucun des n-grammes de la traduction ne correspond à aucun n-grammes de la référence, alors le score BLEU prend la valeur minimale qui est 0.

2.5 Conclusion

Dans ce chapitre, nous avons introduit les notions fondamentales pour la compréhension du domaine d'extraction des ENs et de TAS qui sont les deux domaines sur lesquels porte ce mémoire. La langue arabe et les propriétés des ENs en arabe ont été présentées aussi.

Nous présentons dans le prochain chapitre un état de l'art sur les travaux de l'extraction des ENs et de la TA. Nous concentrons notre présentation aux travaux liés à la langue arabe.

CHAPITRE III

ÉTAT DE L'ART

Ce chapitre est un état de l'art sur les travaux en extraction d'EN et en TAS. Il comporte deux sections. La première section concerne les ENs où nous décrivons chaque approche pour la reconnaissance des ENs suivie d'un ensemble de travaux. La deuxième section concerne la TAS.

3.1 Extraction des entités nommées

En TALN, il y trois approches très utilisées qui sont l'approche symbolique ou linguistique (à base de règles), l'approche statistique (ou à base d'apprentissage) et l'approche hybride (Ehrmann, 2008).

Sekine et Eriguchi (2000) ont classifié les systèmes d'extraction des ENs en trois classes selon l'utilisation de l'une de ces trois approches. Ces classes sont :

1- Système d'extraction des ENs à base de règles qui consiste à utiliser l'approche symbolique.
2- Système d'extraction des ENs à base d'apprentissage qui consiste à utiliser l'approche statistique.
3- Système d'extraction des ENs hybride qui consiste à utiliser l'approche hybride.

Nous présentons dans cette section les trois approches pour l'extraction des ENs en ajoutant la méthode basée sur les corpus parallèles ou comparables (initiée dans le chapitre I) qui est dérivée de l'approche symbolique. Nous donnons quelques exemples de travaux pour chacune de ces approches.

3.1.1 Approche symbolique (à base de règles)

L'approche symbolique, appelée aussi approche à base de règles, consiste à créer des règles manuelles pour extraire les ENs. Ces règles sont appelées les règles de reconnaissance et elles sont composées d'une suite de contraintes à appliquer sur les caractéristiques des termes du texte (Zidouni, 2010).

La construction des règles dans un texte nécessite la connaissance des propriétés grammaticales, syntaxiques et orthographiques de la langue du texte (Budi et Bressan, 2003). Ces propriétés permettent d'étiqueter le texte et de désigner les marqueurs lexicaux qui aident la reconnaissance des ENs.

Les règles peuvent être représentées par des expressions régulières (Zaghouani, 2009). Une expression régulière est une chaîne de caractères qui appelée aussi un motif et qui décrit un ensemble de chaînes de caractères selon une syntaxe précise[37].

Exemple de travaux d'extraction des ENs à base de règles

Parmi les premiers travaux d'extraction des ENs arabes, il y a le système TAGARAB (Maloney et Niv, 1998) qui a été développé en 1998 par Maloney et Niv. Celui-ci consiste à utiliser un analyseur morphologique pour isoler le nom propre en spécifiant la fin et le début de mot qui le suive.

Abuleil (2004) a développé un système à base de règles pour extraire les noms de personne à partir de systèmes de Question/Réponse. Il a déterminé un ensemble de règles en se basant sur l'étude des relations entre les mots dans la phrase. Il a utilisé un ensemble de mots-clés et de verbes spéciaux qui ont été collectés dans un autre projet présenté dans (Abuleil et Evens, 2002).
Habash et Roth (2008) ont créé un algorithme pour traiter le problème de la reconnaissance des expressions numériques en arabe. Leur travail a pour but d'identifier

[37] http://fr.wikipedia.org/wiki/Expression_rationnelle

les expressions numériques et de construire une norme de référence pour les évaluer en supportant les différentes formes (chiffre, ensemble de mots, mélange de chiffres et mots, ordre, pluriels).

Zaghouani (2009) et (2012) a adapté le module de repérage des ENs du système de veille EMM (Europe Media Monitor)[38] à la langue arabe. Le système présenté par Zaghouani a été nommé RENAR (Repérage des Entités Nommées ARabes). Il a été appliqué sur des textes écrits en arabe moderne et est fondé principalement sur un lexique et un ensemble de règles de repérage sous forme de règles manuelles. Ces règles sont regroupées dans des fichiers pour chaque classe d'EN. Le système RENAR effectue un traitement en deux étapes. La première étape est le prétraitement lexical qui consiste à segmenter le texte et à normaliser le phonème Alif hamza en Alif sans hamza, par exemple le nom أنيس (>nys / Aniss) devient le nom انيس (Anys / Aniss). La deuxième étape est le repérage des ENs où le système cherche l'existence de chaque mot dans les dictionnaires des ENs. Si le mot est trouvé dans l'un des dictionnaires, il sera retenu comme EN. Si le mot ne se trouve dans aucun dictionnaire, le système utilise des expressions régulières qui permettent de détecter les ENs.

Mesfar (2007) a développé un système de reconnaissance des ENs en arabe avec la combinaison d'un analyseur syntaxique et un analyseur morphologique. Le système de Mesfar utilise la plateforme de développement linguistique NooJ[39] et se base sur la recherche des preuves internes et externes[40] qui aident à développer des règles pour reconnaître les ENs.

---

[38] Le système EMM est un outil de regroupement des articles qui sont de différentes langues européennes, couvrant le même sujet et provenant de différents sites Web. EMM fusionne ces articles automatiquement en un seul groupe afin d'éviter la redondance des nouvelles. http://press.jrc.it

[39] http://www.nooj4nlp.net/pages/nooj.html

[40] Les preuves internes et externes sont des mots qui aident la reconnaissance des ENs. Les mots internes peuvent être contenus dans des listes de marqueurs lexicaux ou des listes de noms propres prédéfinies. Les preuves externes sont obtenues par le contexte dans lequel une EN apparaît. Ceci par l'étude des relations syntaxiques au sein d'une phrase pour attribuer le type de l'EN retenue (Mesfar, 2008).

Ben-Hamadou et al. (2010) se sont intéressés à la reconnaissance et la traduction des ENs de type noms de lieux en utilisant la plate forme Nooj. Ils ont limité leur travail au domaine du sport et ils ont utilisé une approche à base de règles en se basant sur la grammaire et la représentation lexicale de la phrase. Après l'extraction des ENs, ils ont intégré un module de translittération pour traduire les ENs extraites.
Pour une centaine de textes, Ben-Hamadou et al. ont trouvé un rappel de 95% et une précision de 97%.

Vu la complexité de la reconnaissance des noms de personne arabes, quelques travaux ont traité juste l'extraction de ce type d'EN. Par exemple, Shaalan et Raza (2007) ont présenté le système PERA (Person Name Entity Recognition for Arabic) qui consiste a extraire les ENs en arabe. Shaalan et Raza ont utilisé un ensemble de règles, un dictionnaire des noms et une grammaire régulière. Traboulsi (2009) a utilisé l'approche de la grammaire locale en utilisant un ensemble de règles et un dictionnaire. Elsebai et al. (2009) ont utilisé des règles et des mots-clés pour extraire les EN-PERS.

Poibeau et Group (2003) ont développé un système pour l'extraction des ENs multilingues (arabe, chinois, anglais, français, allemand, japonais, finlandais, malgache, persan, polonais, russe, espagnol et suédoise). Ils ont utilisé l'approche à base de règle en adaptant le même modèle pour chaque langue.

3.1.2 Approche basée sur les corpus parallèles ou comparables

Avant de présenter cette approche nous introduisons la définition du corpus parallèle et du corpus comparable.

Corpus parallèle

Un corpus parallèle est un ensemble de textes dans deux langues différentes qui sont alignés au niveau de la phrase, c'est-à-dire des textes en langue source avec leur traduction en langue cible (Afli et al., 2012).

Exemples de corpus parallèles :

- Le corpus Hansard du parlement Canadien[41] .
- Le corpus multilingue des Nations Unies UN [42].
- Le corpus du parlement européen Europarl[43].

Corpus comparable

Selon Dejean et Gaussier (2002) « Deux corpus de deux langues L1 et L2 sont dits comparables s'il existe une sous-partie non négligeable du vocabulaire du corpus en langue L1 (respectivement L2) dont la traduction se trouve dans le corpus en langue L2 (respectivement L1) ».

Ainsi, les textes du corpus comparable de la langue source partagent les mêmes idées des textes de la langue cible.

L'alignement des textes dans les corpus comparables n'est pas nécessaire.

Exemples de corpus comparables :

- Les articles de Wikipédia qui sont disponibles en plusieurs langues.
- Les articles de journaux multilingues.

Principe de la méthode basée sur les corpus parallèles ou comparables

Cette méthode consiste à exploiter les corpus parallèles ou comparables pour l'extraction des ENs en une langue cible. Elle est divisée en deux phases.

Phase 1 : l'extraction des ENs du corpus en langue source en utilisant un système d'extraction des ENs valide. La langue source est généralement celle qui possède plus

de ressources linguistiques (par exemple, l'anglais).

---

[41] http://www.isi.edu/natural-language/download/hansard/

[42] http://www.uncorpora.org/

[43] http://www.statmt.org/europarl/

Phase 2 : la projection cross-linguistique des ENs de la langue source vers la langue cible pour avoir les ENs en langue cible.

La tâche de projection est réalisée de plusieurs façons différentes comme par exemple, la translittération des ENs, la traduction des ENs, la recherche dans les dictionnaires d'EN et l'alignement au niveau de mots. Dans ce mémoire, nous nous intéressons à la méthode de translittération des ENs.

Projection cross-linguistique des ENs par la translittération

La projection cross-linguistique des ENs par la translittération consiste à utiliser une technique de translittération des ENs reconnues en langues source pour avoir les ENs en langue cible.
La traduction des ENs consiste à utiliser des dictionnaires ou des systèmes de TA pour avoir les ENs en langue cible.
Parmi les premiers travaux de la translittération des noms propres de l'arabe vers l'anglais, il y a le travail de (Arbabi et al., 1994) et celui de (Stalls et Knight, 1998).

Plus tard, Al-Onaizan et Knight (2002) ont développé un algorithme de translittération des noms de personne de l'arabe vers l'anglais. Leur algorithme consiste à associer le son et l'orthographe à l'aide de machines à états finis.
Dans la littérature, beaucoup de travaux se sont intéressés à la projection cross-linguistique en utilisant la translittération des ENs. Nous citons par exemple, le travail de Samy et al. (2005) qui ont utilisé un corpus parallèle en arabe et en espagnol, et un étiqueteur pour extraire les ENs dans le corpus espagnol. Après l'extraction des ENs en espagnol, ils ont trouvé leurs correspondances dans le corpus arabe en appliquant une technique de translittération. Leur système est composé de trois modules. Le premier module consiste à chercher les ENs de type date dans un lexique préparé auparavant puis à extraire les équivalents arabes. Le deuxième module consiste à translittérer les EN-PERS ou les EN-LOC de l'espagnol vers l'arabe. La translittération se fait en déterminant pour chaque caractère espagnol toutes les équivalents arabes possibles, et

en cherchant la meilleure combinaison dans la phrase arabe. Le dernier module consiste à chercher les EN-ORG dans un lexique préparé auparavant. Samy et al. ont obtenu un rappel de 97.5%, une précision de 84% et une F-mesure de 90%.

Aussi, Abdel Fattah et al. (2006) et Abdel Fattah et Ren (2008) ont présenté un modèle de translittération des noms propres à partir de corpus parallèles anglais-arabe. Pour cela, ils ont d'abord extrait les noms propres en anglais par l'étiqueteur CLAWS4 POS[44]. Ensuite, ils ont extrait les noms propres en arabe à l'aide d'un analyseur morphologique pour la langue arabe, puis ils ont romanisé ces noms propres. Leur méthode de translittération de l'anglais vers l'arabe est basée sur la recherche de similarités entre les noms propres en anglais et les noms propres romanisés.

Semmar et Saadane (2013) ont présenté un modèle de translittération des noms propres de l'arabe vers l'écriture latine. Leur système est basé sur la translittération de chaque consonne de nom propre en utilisant un automate à états finis. Ensuite, ils ont utilisé leur modèle de translittération pour l'alignement de mots à partir de corpus parallèles.

Les travaux de Abduljaleel et Larkey (2003), de Kashani (2007) et de Kashani et al. (2007) consistent à développer un système de translittération des noms propres avec l'utilisation d'une méthode d'entraînement de la translittération de chaque phonème. Le système de translittération présenté dans (Abduljaleel et Larkey, 2003) est de l'anglais vers l'arabe et celui dans (Kashani, 2007; Kashani et al., 2007) est de l'arabe vers l'anglais.

- **Extraction des ENs multilingues à l'aide de corpus parallèles**

Il y a quelques travaux qui se sont s'intéressés à l'extraction des ENs en plusieurs langues à l'aide de corpus parallèles multilingues. Citons par exemple, le projet d'extraction des ENs réalisé par la commission européenne du centre commun de

[44] http://ucrel.lancs.ac.uk/claws/trial.html

recherche (European Commission's Joint Research Centre JRC). Ce projet (Steinberger et al., 2011)[45] consiste à utiliser un ensemble de corpus parallèles multilingues et d'annoter automatiquement les ENs dans le corpus en anglais puis de projeter ces ENs pour les autres langues. Le même algorithme pour la projection des ENs en anglais est appliqué pour toutes les langues du corpus parallèle multilingue.

Aussi, Ehrmann et al. (2011) se sont intéressés aux corpus parallèles multilingues dans les langues : anglais, français, espagnol, allemand, tchèque et russe. Avec un système d'extraction des ENs pour l'anglais, ils ont déterminé les ENs en anglais puis ils ont projeté les ENs en anglais vers les cinq autres langues. La projection a été faite à l'aide de deux méthodes. La première méthode est la traduction de l'EN en anglais vers les autres langues en utilisant un système de TAS. La deuxième méthode est la recherche de l'EN dans une base de données d'EN multilingue.
Il y a aussi le travail de Pouliquen et al. qui ont développé un système d'extraction des noms de personne à partir de collections de presses multilingues (y inclut l'arabe) (Pouliquen et al., 2005). Ils ont intégré dans leur système un modèle de translittération en trouvant toutes les écritures possibles d'un nom de personne.

3.1.3 Approche par apprentissage machine

L'approche par apprentissage machine est très répandue dans plusieurs domaines tels que la bio-informatique, la finance, l'EI et le forage de données (Larochelle, 2009). Pour le domaine de reconnaissance des ENs, cette approche consiste à combiner des corpus d'entraînement annotés avec des algorithmes d'apprentissage machine, ce qui permet d'entraîner le modèle pour extraire les ENs.

Exemple : On a dans le corpus d'entraînement plusieurs fois le terme abrégé « Mr. » suivi d'un terme (ou plusieurs termes) qui est annoté comme étant une EN-PERS.

---

[45]Présentation en ligne http://tln.li.univ-tours.fr/Tln_Colloques/Tln_REN2011/Ehrmann-ATALA-20juin2011.pdf , Journée ATALA Entités Nommées - Lundi 20 Juin 2011.

Suite à cette observation, le système d'extraction des ENs va annoter les nouveaux termes précédés par le terme abrégé « Mr. » comme des EN-PERS.

Les algorithmes d'apprentissage peuvent se classifier selon le type d'apprentissage machine qu'ils emploient. Il y a trois types d'apprentissage machine : supervisé, semi-supervisé et non supervisé[46].

Nous présentons dans les prochaines sous-sections les types d'apprentissage machine appliqués en particulier dans le domaine de reconnaissance des ENs.

Apprentissage supervisé

L'apprentissage supervisé consiste à utiliser des exemples prédéterminés sous forme de corpus annotés d'EN pour réaliser la tâche d'extraction. Elle se déroule en deux étapes : la première étape est l'apprentissage qui consiste à fournir le corpus d'entraînement annoté. La deuxième étape consiste à concevoir des règles pour définir les types des ENs dans un texte (Sun, 2010).

Pour que les résultats soient bons dans l'apprentissage supervisé, le corpus annoté doit être de grande taille, ce qui diminue le coût de réalisation de l'apprentissage en terme de temps et d'intervention humaine.

Parmi les algorithmes d'apprentissage supervisé les plus utilisés, on trouve les chaînes de Markov cachée (Bikel et al., 1997), les arbres de décision (Sekine, 1998), l'entropie maximale (en anglais, Maximum Entropy) (Berger et al., 1996) et machine à vecteurs de support (en anglais, Support Vector Machines) (Vapnik, 1999).

Apprentissage semi-supervisé

Cette technique d'apprentissage combine des données étiquetées et des données non étiquetées. Cette combinaison permet d'améliorer la qualité de l'apprentissage, car

[46] http://fr.wikipedia.org/wiki/Apprentissage_automatique#Types_d.27apprentissage

l'intervention humaine est nécessaire pour l'annotation des données non annotées (Blum et Mitchell, 1998), mais cela influe sur le coût de cette technique qui reste élevé (Larochelle, 2009).

Apprentissage non supervisé

Contrairement à l'apprentissage supervisé, la technique d'apprentissage non supervisé ne nécessite aucune intervention humaine. Elle repose sur les ressources lexicales comme par exemple WordNet[47], sur les schémas lexicaux et sur des statistiques calculées à partir d'un corpus large non annoté, c'est-à-dire avec des données brutes qui sont considérées comme des données aléatoires (Nadeau et Sekine, 2007).

Le principe de cette méthode est la division des données en sous-groupes. Les données similaires sont associées au même groupe et les données différentes sont dans des groupes différents[48].

Exemple de travaux sur l'extraction des ENs par l'apprentissage machine

Une série de travaux de Benajiba et al. est basée sur une approche d'apprentissage machine a été réalisée. Premièrement, Benajiba et al. (2007) ont construit le corpus ANERcorp et les gazetteers[49] ANERgazet pour développer le système ANERsys. ANERsys est un système d'extraction des ENs pour la langue arabe qui est basé sur l'algorithme d'apprentissage statistique d'entropie maximale. Le corpus d'apprentissage automatique du système ANERsys est de 125 000 mots et le corpus de test est de 25 000 mots. Les résultats obtenus par ce système ont donné un rappel de 37.51%, une précision de 51.39% et une F-mesure de 43.36%.

---

[47] http://wordnet.princeton.edu/

[48] Extrait du lien http://fr.wikipedia.org/wiki/Apprentissage_non_supervis%C3%A9

[49] Un gazetteer est une liste d'EN de différentes types (Mikheev et al., 1999).

Ensuite, le système ANERsys a été amélioré à ANERsys 2.0 (Benajiba et Rosso, 2007). L'amélioration a été faite pour reconnaître les noms propres longs en combinant l'approche du maximum d'entropie avec l'étiquetage morphosyntaxique. ANERsys 2.0 a donné des résultats améliorés par rapport à l'ancienne version ANERsys. Le rappel a été amélioré à 49,04 %, la précision a été améliorée à 63,21% et la F-mesure a été améliorée à 55,23 %.

Pour améliorer encore la précision de ANERsys, Benajiba et Rosso (2008) ont utilisé un autre modèle probabiliste qui est les champs markoviens conditionnels (CMC) (en anglais, Conditional Random Fields) (Lafferty et al., 2001). Ils ont ajouté aussi la segmentation (en anglais, tokenisation) des données, ce qui amène à de meilleurs résultats.
Benajiba et al. (2008) ont utilisé les deux modèles probabilistes CMC et machine à vecteurs de support (MVS)[50] (Vapnik, 1999) pour développer un système d'extraction des ENs. Leur système intègre aussi les caractéristiques lexicales, syntaxiques et morphologiques. À l'aide des modèles CMC et MVS, ils ont attribué un classificateur à chaque type d'EN. Ensuite, ils ont combiné tous les classificateurs pour le système global d'extraction des ENs.

Dans sa thèse, Benajiba (2009) a testé la première version de ANERsys (Benajiba et al., 2007) en utilisant trois types de modèles d'apprentissage : entropie maximale, MVS et CMC. Il a obtenu les meilleurs résultats par la combinaison des trois modèles à la fois.

Un autre travail basé sur l'apprentissage machine est celui de Mohammed et Nazlia (2012) qui ont développé un système d'extraction des ENs en arabe avec l'utilisation des réseaux de neurones. Premièrement, ils ont prétraité le texte en entrée qui est en langue arabe. Ensuite, ils ont converti les phrases de ce texte en caractères romains, puis ils ont classifié les types de mots en utilisant les réseaux de neurones. Les réseaux de

[50] http://www.support-vector.net/

neurones consistent à apprendre la reconnaissance automatique des types d'EN et à prendre des décisions intelligentes basées sur les données disponibles. Le système de Mohammed et Nazlia a obtenu une F-mesure de 69.90% pour les EN-PERS, 43.30% pour les EN-LOC, et 59.20% pour les EN-ORG.

Aussi, Gahbiche-Braham et al. (2012) et Gahbiche-Braham et al. (2014) ont développé un système d'extraction des ENs basé sur l'apprentissage supervisé en utilisant les algorithmes d'apprentissage statistiques CMC avec l'outil Wapiti (Lavergne et al., 2010). Ensuite, ils ont adapté leur système à un apprentissage non supervisé (auto-apprentissage), et les résultats ont été améliorés après l'adaptation.

3.1.4 Extraction des entités nommées à partir de Wiképdia

Wikipédia est un projet d'encyclopédie collective établie sur Internet, universelle, multilingue et fonctionnant sur le principe du wiki. Wikipédia a pour objectif d'offrir un contenu librement réutilisable, objectif et vérifiable, que chacun peut modifier et améliorer[51]. C'est est une ressource riche d'EN en plusieurs langues, et beaucoup de travaux se sont intéressés à l'utiliser comme corpus de test ou d'apprentissage. Par exemple, nous citons le travail de Attia et al. (2010) qui consiste à construire un lexique de noms de personne et de noms d'organisation à partir de Wikipédia et WordNet en arabe. Attia et al. ont adapté la méthode MINELex (Multilingual, Interoperable Named Entity Lexicon) (Toral, 2009) pour la langue arabe.

Alotaibi et Lee (2012a) et Alotaibi et Lee (2012b) ont présenté une nouvelle approche d'extraction des ENs à partir de Wikipédia en arabe. Ils ont classifié les articles en un ensemble d'EN. Ils ont utilisé quatre types de classification : naïve bayésienne, naïve bayésienne multinomiale, MVS et la descente de gradient stochastique. Ils ont pris en considération dans leur classification le format des articles de Wikipédia et les caractéristiques de la langue arabe. Par la suite, ils ont complété leur projet par le

[51] Extrait de site Web http://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Accueil_principal

développement d'un système pour créer automatiquement un corpus et une liste d'EN arabes à partir de Wikipédia (Alotaibi et Lee, 2013).

Mohit et al. (Mohit et al., 2012) ont présenté le modèle ROP (Rappel-Orienté Perceptron) pour la détection des ENs à partir de Wikipédia en arabe. Dans leur travail, ils ont modifié les critères de l'apprentissage supervisé avec l'utilisation des données non annotées, ceci en intégrant une phase d'auto-entraînement. Ils ont réussi à améliorer le rappel mais avec dégradation de la précision. Ils ont développé aussi un petit corpus d'articles de Wikipédia en arabe via un schéma d'annotation des ENs. Ce corpus couvre quatre domaines thématiques : histoire, technologie, science et sport. Il est accessible en ligne sur le site Web http://www.ark.cs.cmu.edu/AQMAR.

Wikipédia est utilisée aussi pour la construction de terminologie bilingue ou multilingue (voir les travaux de (Sadat, 2010; Sadat et Terrasa, 2010; Patry et Langlais, 2011; Mohit et al., 2012; Sellami et al., 2012; Sellami et al., 2013)).

Dans ce mémoire, nous utilisons Wikipédia comme un corpus parallèle pour l'Extraction des ENs en arabe.

3.1.5 Approche hybride

L'approche hybride consiste à combiner l'approche à base de règles et l'approche d'apprentissage pour l'extraction des ENs. Cette combinaison permet de produire un système idéal qui profite des avantages de l'utilisation des deux approches : symbolique et statistique (Zribi et al., 2010). Dans cette approche, les règles sont généralement apprises automatiquement mais elles doivent être révisées par un expert (Poibeau, 2001; Mansouri et al., 2008).

Mansouri et al. (2008) ont présenté une étude comparative entre les trois approches : à base de règles, à base d'apprentissage machine et hybride. Cette étude a montré que l'approche hybride donne de bons résultats par rapport aux deux autres approches.

Parmi les systèmes d'extraction des ENs basés sur l'approche hybride, nous citons le système décrit dans (Azab et al., 2013) . Ce système consiste à automatiser le choix

entre la traduction et la translittération des ENs de l'anglais vers l'arabe. Pour cela, Azab et al. ont suivi l'approche basée sur les corpus parallèles combinée avec la technique d'apprentissage 'machine à vecteurs de support' pour entraîner leur système à choisir entre la translittération et la traduction des ENs. Le résultat de leur système est un lexique bilingue d'ENs.

Il y a aussi le système présenté dans (Zribi et al., 2010) qui est composé de deux phases : la phase d'analyse morphologique du texte en arabe et la phase d'extraction automatique de règles pour détecter les ENs selon leur type. Ce système a été combiné avec l'algorithme d'apprentissage des règles RIPPER, qui utilise un ensemble d'attributs représentant les éléments les plus influents sur le résultat d'apprentissage. Zribi et al. ont choisi d'utiliser deux types d'attributs pour l'extraction des règles : attributs morphologiques et attributs à base de lexique de noms propres. Ils ont étudié les cinq mots qui sont situés avant et après le mot à classer. Quatre caractéristiques morphologiques ont été étudiées dans leur système : la catégorie et le type du mot, le proclitique et le type du proclitique qui est rattaché à ce mot.

Aussi, le système présenté dans (Oudah et Shaalan, 2013) est basé sur l'approche hybride. Ce système est formé de deux composants. Le premier composant est à base de règles avec l'utilisation de la plateforme GATE[52]. Le deuxième composant est à base d'apprentissage machine en utilisant trois techniques d'apprentissage qui sont : l'arbre de décision, la MVS et la régression logistique. Le système développé a été testé sur le corpus ANERcorp de (Benajiba et al., 2007) et a donné un rappel de 94,9%, une précision de 94.2% et une F-mesure de 94.5%.

3.2 Traduction automatique

La recherche en TA a commencé avec l'apparition des ordinateurs. Le premier système de TA est un système présenté par IBM en 1954 qui fait la traduction des phrases du russe vers l'anglais (Elyan, 2012). Dans les années 1960, pendant la guerre froide, des

[52] GATE est disponible dans le lien http://gate.ac.uk/

besoins spécialement militaires et politiques de la traduction des articles russes vers l'anglais (ou l'inverse) apparaissent (Goudet, 2008). En 1966, le rapport ALPAC (en anglais, Automatic Language Processing Advisory Committee) concluait à l'impossibilité de fournir une TA de qualité (Grass, 2010). En 1968, Peter Toma a conçu le système de TA pour la compagnie SYSTRAN[53] qui est considérée comme l'acteur mondial et pionnier des technologies de traduction. Ce système est basé sur des règles écrites par l'humain manuellement.

Ensuite, avec l'évolution de l'utilisation d'Internet, la TA a connu une popularité dans le Web par l'intermédiaire de plusieurs systèmes de TA, comme Google Translate, Microsoft Translate[54], etc.

Comme nous avons vu dans le chapitre II, il y a trois paradigmes de la TA qui sont la TA à base de règles, la TA guidée par l'exemple et la TAS. Cependant, la TAS est la plus utilisée dans la littérature (Gahbiche-Braham, 2013).

Nous nous intéressons dans cet état de l'art aux TAS depuis ou vers la langue arabe, mais nous présentons d'abord quelques problèmes de la TAS.

3.2.1 Les problèmes de la traduction automatique statistique

Les systèmes de TAS ont connu un développement très important dans les dernières années. Cependant, il y a quelques problèmes qui se posent dans le déroulement de la procédure de réalisation d'un système de TAS. Le premier problème est lié au traitement des noms propres et les MHV. Ce problème est traité dans plusieurs travaux, car les ENs et les MHV sont très fréquents dans la majorité des textes et corpus, et une mauvaise traduction de ce type de terme peut influencer le sens de la traduction de la phrase.

Pour illustrer ce problème, prenant l'exemple suivant (Azab et al., 2013) :

---

[53] http://www.systranet.com/translate/

[54] http://www.bing.com/translator

On a la phrase suivante en anglais : 'Dudley North was an English merchant' et on veut la traduire à l'aide d'un système de TAS vers l'arabe.

Si l'EN-PERS «Dudley North » n'est pas reconnue, le système de TAS donne la traduction suivante :

كان دودلي الشمالية تاجر الانجليزية (kAn dwdly Al$mAlyp tAjr AlAnjlyzyp / Nord du Dudley est le concessionnaire de l'Angleterre) . Cette traduction est fausse et pour avoir une traduction correcte, l'EN 'Dudley North' doit être translittérée ce qui donne la traduction suivante :

كان دودلي نورث تاجر انجليزي (kAn dwdly nwrv tAjr Anjlyzy / Dudley Nord était concessionnaire anglais).

La reconnaissance des ENs constitue donc une amorce importante dans un système de TAS (Ehrmann, 2008) (Agrawal et Singla, 2010).

D'autres types de termes posent aussi un problème pour les systèmes de TAS comme par exemple les acronymes, les expressions poly lexicales et les mots composés.

Un autre problème de la TAS est lié à la faible disponibilité de corpus parallèles. Mais ce problème peut être résolu par la construction d'un corpus parallèle à partir de corpus comparables qui sont plus accessibles par rapport aux corpus parallèles.

3.2.2 Traduction depuis ou vers la langue arabe

La langue arabe est l'une des premières langues étudiées en TA (Zughoul et Abu-Alshaar, 2005). La traduction depuis ou vers la langue arabe est une tâche complexe. En effet, les termes en arabe possèdent de nombreuses variantes orthographiques, notamment sur les noms propres (ENs), ce qui multiplie les formes inconnues dans les textes (Gahbiche-Braham et al., 2012). Donc, la reconnaissance des ENs en arabe peut améliore la traduction depuis ou vers cette langue.

Il y a différentes solutions pour l'amélioration des systèmes de TA à l'aide des ENs. Une de ces solutions est l'utilisation des lexiques d'ENs. Une autre solution est

l'utilisation d'un système de translittération et d'un système de reconnaissance des ENs en langue source.

Dans la littérature, on trouve des chercheurs qui exploitent les ENs pour la TAS. Parmi les exemples, on cite les travaux (Kashani et al., 2007; Hermjakob et al., 2008; Gahbiche-Braham et al., 2014; Sellami et al., 2014) qui ont montré une bonne amélioration de la TAS en introduisant un système propre aux ENs spécifiquement pour la langue arabe.

3.3 Conclusion

Dans ce chapitre, nous avons présenté un état de l'art sur les deux domaines étudiés dans notre mémoire qui sont l'extraction des ENs et la TA.

Pour les ENs, nous avons commencé par la description des différentes approches d'extraction des ENs. L'approche symbolique est ancienne et les premiers systèmes l'utilisaient pour le développement de leurs algorithmes (Nadeau et Sekine, 2007). Mais malgré son ancienneté, elle est encore utilisée aujourd'hui en raison de sa simplicité et de sa rapidité, mais à la condition d'avoir une connaissance approfondie de la langue des ENs pour le développement des règles. Cependant, dans le cas où l'utilisation des règles manuelles ne réussit pas à reconnaître les ENs, il faut utiliser des bases de données pour rechercher les ENs. Cette hypothèse rend l'approche à base de règles limitée, car la capacité du système dépend de la taille de la base de données et le système ne peut reconnaître les nouvelles ENs qui ne s'y trouvent pas (Nguyen, 2007).

L'approche par apprentissage machine nécessite des corpus d'entraînement annotés qui ne sont pas toujours disponibles ou qui doivent être entièrement construits.

Les corpus parallèles ou comparables sont des ressources très importantes pour les applications du TALN, notamment pour l'EI, la TA et l'annotation des ENs cross-linguistique. C'est la raison pour laquelle ces ressources sont devenues très utilisées dans la littérature (Ehrmann et al., 2011).

Dans ce mémoire, pour répondre aux objectifs fixés, nous avons opté pour la méthode d'extraction des ENs basée sur les corpus parallèles pour la paire de langues anglais-arabe. Pour projeter les ENs en anglais vers l'arabe, nous avons utilisé une technique de translittération qui sera détaillée dans le chapitre IV.

Pour la TA, nous avons présenté d'abord les problèmes de la TAS ensuite, nous avons concentré sur les travaux de TAS qui utilisent la langue arabe comme langue source ou cible. Dans ce mémoire, un système de TAS a été développé pour expérimenter les lexiques bilingues d'ENs construits à l'aide de notre méthode de translittération.

CHAPITRE IV

MÉTHODOLOGIE

4.1 Introduction

L'objectif principal de notre mémoire est l'extraction des ENs par la projection cross-linguistique en utilisant des corpus parallèles pour la paire de langues anglais-arabe. Cette méthode automatique d'annotation ou d'EI consiste à exploiter des ressources et des outils disponibles pour une langue source pour l'extraction des informations pour une autre langue cible (Ben Abacha et al., 2012). Le passage par les corpus parallèles alignés en deux langues source et cible est imposé dans la méthode de projection cross-linguistique. Cette dernière est une méthode efficace aussi pour l'extraction d'informations multilingues, et elle est utilisée par exemple dans le projet JRC-Names d'extraction des ENs multilingues réalisé par la Commission européenne du centre commun de recherche (Steinberger et al., 2011).
Dans ce mémoire, notre méthode de projection consiste à développer un modèle de translittération des ENs du corpus source vers la langue cible pour extraire les ENs du corpus cible.

La contribution de notre méthode de translittération se traduit premièrement par la production de corpus annotés pour une langue complexe comme l'arabe, et deuxièmement par la construction de lexiques bilingues d'ENs pour deux langues complètement différentes du point de vue morphologique.

Pour évaluer la qualité des lexiques bilingues d'ENs construits, nous avons développé un système de TAS auquel nous introduisons ces lexiques pour améliorer la traduction.

Une partie de ce chapitre est consacrée à l'explication de la démarche suivie pour la construction de notre système de TAS.

4.2 Extraction des entités nommées

Dans cette section, nous commençons par l'illustration de l'architecture générale de la solution pour l'extraction des ENs en arabe. Ensuite, nous présentons notre méthodologie pour la construction de lexiques bilingues d'ENs.

4.2.1 Architecture générale de la solution proposée

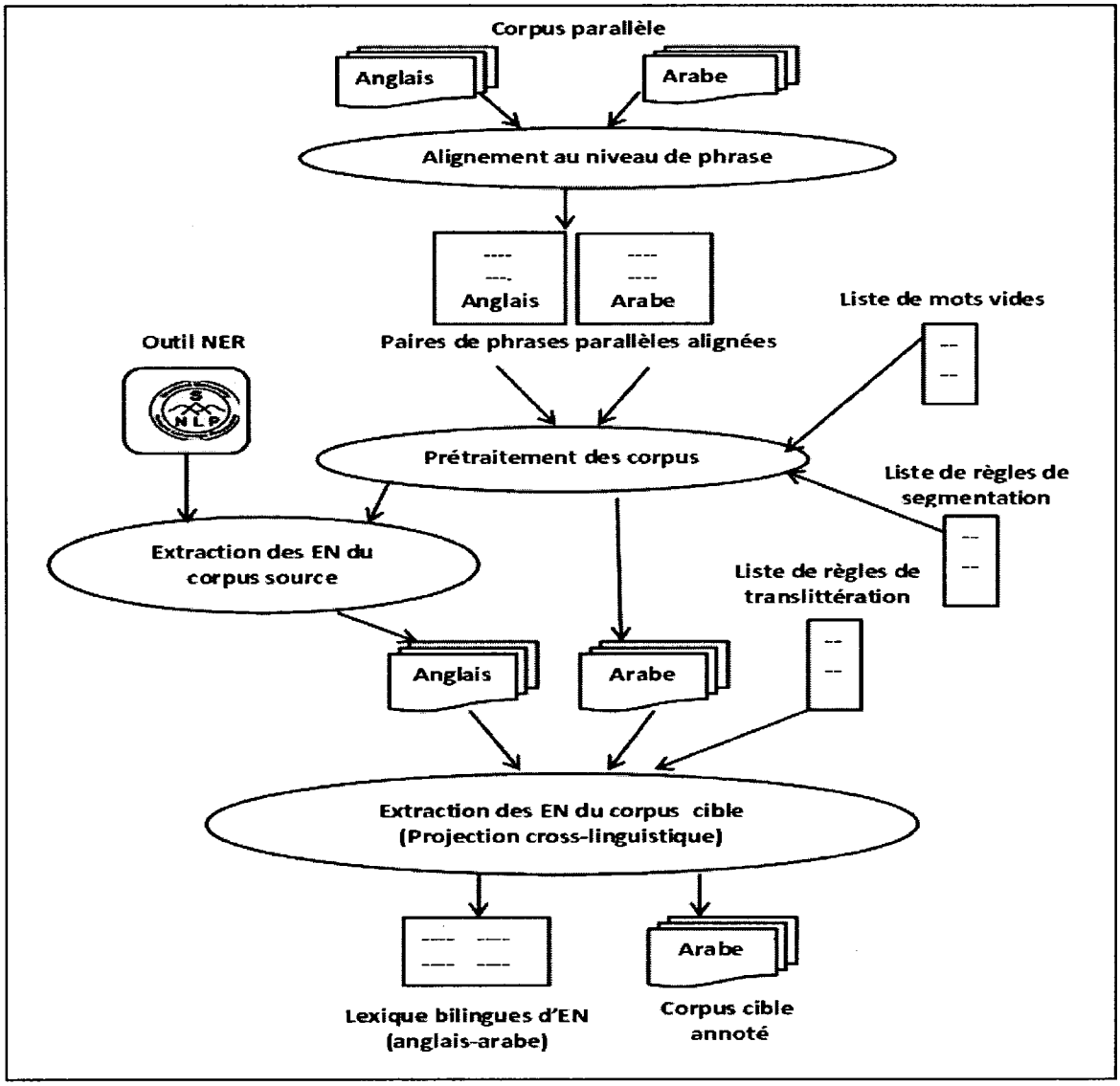
La figure 4.1 montre l'architecture générale de notre solution pour l'extraction des ENs en arabe à partir de corpus parallèles. Cette architecture s'articule autour de quatre procédures suivantes :

1- **L'alignement des corpus parallèles :** l'alignement est fait au niveau des phrases en utilisant l'outil Hunalign[55].
2- **Prétraitement du corpus source (anglais) et du corpus cible (arabe)** : le prétraitement du corpus source est fait par l'outil NER de Stanford[56], et le prétraitement du corpus cible est faite par un ensemble de règles manuelles.
3- **Extraction des ENs du corpus source** : l'extraction des ENs du corpus source est faite par l'outil NER de Stanford qui donne un corpus source annoté.
4- **Extraction des ENs du corpus cible** : cette procédure est réalisée par une méthode de projection cross-linguistique en utilisant un modèle de translittération des ENs de la langue source vers la langue cible.

[55] Hunalign est un outil d'alignement disponible en ligne dans le lien http://mokk.bme.hu/en/resources/hunalign/

[56] http://nlp.stanford.edu/

Figure 4.1 Architecture de la solution d'extraction des ENs en arabe



4.2.2 Prétraitement des corpus

Parmi les étapes les plus importantes durant le développement d'une application de TALN, il y a l'étape de prétraitement de données. L'objectif de cette étape est de traiter les données avec des processus unifiés et non une multitude de processus adaptés à tous les cas possibles (Heitz, 2006).

Prétraitement du corpus en langue source

Nous avons utilisé la langue anglaise comme une langue source à cause de sa richesse en terme de ressources linguistiques destinées pour la recherche gratuitement. L'outil NER de Stanford est l'une de ces ressources. C'est un système d'extraction des ENs basé sur l'approche statistique (Finkel et al., 2005). L'utilisation de cet outil ne nécessite qu'un texte avec des données brutes c'est-à-dire sans passer par l'étape de prétraitement qui est incluse dans l'outil NER de Stanford.

Prétraitement du corpus en langue cible

Nous avons utilisé la langue arabe comme une langue cible. Cette langue possède quelques caractéristiques particulières qui rendent nécessaire le passage par l'étape de prétraitement de données. L'agglutination des mots en arabe est l'une de ces caractéristiques. Les ENs en arabe peuvent être agglutinées avec des préfixes ou des suffixes qui nécessitent une segmentation.

Les deux étapes les plus connues dans la phase de prétraitement d'un texte sont la segmentation et la suppression des mots vides (en anglais, stop words). Nous décrivons chacune de ces étapes dans les sous-sections suivantes.

✓ **La segmentation**

Pour identifier les lexèmes de la phrase en arabe, nous avons utilisé quelques règles de segmentation. Ces règles consistent à séparer les préfixes et les suffixes de la racine. Nous avons choisi les préfixes et les suffixes qui apparient souvent avec les ENs comme par exemple l'article : ال (Al - le ou la), les phonèmes ب (b), و (w), وال (wAl).

✓ **La suppression des mots vides**

Les mots vides sont des mots qui sont très fréquents dans les textes, mais portent peu de sens et ont surtout une fonction syntaxique (Abu El-Khair, 2006). Par exemple, pour l'anglais, les prépositions, les articles et les pronoms sont des mots vides.

La suppression des mots vides est une étape importante, car elle diminue l'espace de recherche. Par exemple dans le cas d'extraction des ENs, les mots vides ne représentent jamais une EN, donc il faut les supprimer pour réduire la taille du texte. Dans notre cas, nous avons collecté une liste de mots vides à partir du Web[57], puis nous avons supprimé ces mots s'ils existent dans la phrase en arabe.

4.2.3 Extraction des entités nommées à partir du corpus source

Nous avons choisi l'outil NER de Stanford[58] pour annoter le corpus source. Cet outil prend en entrée un texte et produit le même texte annoté par les ENs avec les balises <PERSON> et </PERSON> pour les EN-PERS, <LOCATION> et </LOCATION> pour les EN-LOC et <ORGANIZATION> et </ORGANIZATION> pour les EN-ORG. L'appendice B contient un exemple d'annotation des ENs par l'outil NER de Stanford.

4.2.4 Normalisation des phonèmes arabes

La langue arabe possède 28 phonèmes pouvant être écrits en trois façons : au début du mot (position initiale), au milieu du mot (position médiane) ou à la fin du mot (position finale). Le tableau 4.1[59] illustre les différents phonèmes arabes avec leurs différentes modes d'écritures, le nom du phonème, la translittération selon la norme DIN-31635 pour la translittération de l'arabe et le son du phonème.

---

[57] Les mots vides ont été extraits à partir des liens https://code.google.com/p/stop-words/ et http://sourceforge.net/projects/arabicstopwords/

[58] Disponible en téléchargement sur le lien http://nlp.stanford.edu/software/CRF-NER.shtml#Download

[59] http://fr.wikipedia.org/wiki/Alphabet_arabe

Tableau 4.1 Les phonèmes arabes

Isolée	Initiale	Médiane	Finale	Nom	Translittération DIN-31635	Son
ء	أ, إ, ؤ, ئ			Hamza	’	ʔ
ا	—		ا	’alif	ā / â	a:
ب	بـ	ـبـ	ب	bā’	B	b
ت	تـ	ـتـ	ت	tā’	T	t
ث	ثـ	ـثـ	ث	ṯā’	ṯ	θ
ج	جـ	ـجـ	ج	Ǧīm	Ǧ	ʤ
ح	حـ	ـحـ	ح	ḥā’	ḥ	ħ
خ	خـ	ـخـ	خ	ḫā’	ḫ / h̲	x
د	—		د	Dāl	D	d
ذ	—		ذ	ḏāl	ḏ	ð
ر	—		ر	rā’	R	r
ز	—		ز	Zāy	Z	z
س	سـ	ـسـ	س	Sīn	S	s
ش	شـ	ـشـ	ش	Sīn	S	ʃ
ص	صـ	ـصـ	ص	ṣād	ṣ	sˁ
ض	ضـ	ـضـ	ض	ḍād	ḍ	dˁ, ðˁ
ط	طـ	ـطـ	ط	ṭā’	ṭ	tˁ
ظ	ظـ	ـظـ	ظ	ẓā’	ẓ	zˁ, ðˁ
ع	عـ	ـعـ	ع	‘ayn	‘ / ‘	ʔˁ
غ	غـ	ـغـ	غ	Gayn	G	ɣ
ف	فـ	ـفـ	ف	fā’	F	f
ق	قـ	ـقـ	ق	Qāf	Q	q
ك	كـ	ـكـ	ك	Kāf	K	k
ل	لـ	ـلـ	ل	Lām	L	l
م	مـ	ـمـ	م	Mīm	M	m
ن	نـ	ـنـ	ن	Nūn	N	n
ه	هـ	ـهـ	ه	hā’	H	h
و	—		و	Wāw	W	w ou u
ي		ـيـ	ي	yā’	Y	j ou i

Quelques phonèmes arabes peuvent avoir le même son en anglais par exemple, le son en anglais des phonèmes 'س' et 'ص' est 'S'.

Cette hypothèse nous a permis d'introduire une méthode de normalisation de quelques phonèmes arabes vers un seul phonème. Cette méthode est utile dans la phase de comparaison des translittérations des ENs en anglais avec les lexèmes de la phrase arabe, ce qui permet d'accélérer l'exécution de la procédure de translittération des ENs.

Dans la littérature, la normalisation du phonème hamza a été faite dans plusieurs travaux. Cette tâche simplifie le traitement des mots en arabe qui contiennent ce phonème.

Dans notre cas, nous avons ajouté à la normalisation du phonème hamza d'autres normalisations de quelques phonèmes arabes en utilisant les règles suivantes :

Règle 1 : Les phonèmes arabes qui ont une prononciation semblable en anglais sont normalisés en un seul phonème.

Exemple :

La normalisation du phonème 'ض' (D) vers le phonème 'د' (d) dans le nom de personne 'نضال' (nDAl / Nidal) donne le mot 'ندال' (ndAl / Nidal).

Règle 2 : Les phonèmes arabes qui peuvent être ensemble dans les translittérations possibles de plusieurs phonèmes en anglais sont normalisés en un seul phonème.

Exemples :

1- La normalisation des phonèmes 'ص' (S), 'ى' (Y), 'أ' (A) dans le nom de personne 'السيد أنصار ى' (Alsyd >nSArY / Mr. Ansari) donne 'انساري السيد' (Alsyd AnsAry / Mr. Ansari).

2- La normalisation du phonème 'ش' ($) dans le nom de personne 'السيد سوشاريبا' (Alsy sw$ArybA / Mr. Sucharipa) donne 'السيد سوساريبا' (Alsyd swsArybA / Mr. Suchripa)

Règle 3 : Normaliser toutes les phonèmes avec la kashida (ou tatweel -) vers des phonèmes sans kashida (Habash et Roth, 2008).

Exemple :

La normalisation des phonèmes 'آ' (|) et 'ز' (z) et la suppression du kashida dans le phonème 'يـ' (y) pour le nom de personne 'ديـفيـد ماكآدمز' (dyfyd mAk|dmz / David Makadamz) donne le nom 'ديفيد ماكادمس' (dyfyd mAkAdms / David Makadamz).

Règle 4 : Enlever tous les signes diacritiques (Habash et Roth, 2008).

Exemple :

La suppression des signes diacritiques de nom de personne 'سِبَاسْتْيَانْ كَابْوتْ' (sibaAsotoyaAno kaAbNwto / Sebastian Cabot) donne le nom de personne ' سباستيان كابوت' (sbAstyAn kAbwt / Sebastian Cabot).

Dans notre procédure de translittération, les phonèmes qui ont été normalisés sont illustrés dans le tableau 4.2.

Tableau 4.2 Normalisation des phonèmes arabes

Les phonèmes normalisés	**Phonème de normalisation**
Le phonème hamza avec toutes ses écritures (أ , إ , آ , ؤ , ئ , ء , ه)	Le phonème alif sans hamza ا
ى (y)	ي (y)
ظ (z), ض (D), ذ(*)	د (d)
ة (p), ط (T), ث (v)	ت (t)
ق (q), خ (x), ج (j)	ك (k)
ص (S), ز (z), ش ($)	س (s)

4.2.5 Algorithme d'extraction des entités nommées arabes à partir du corpus source

Dans cette sous-section, nous présentons le pseudo-code de notre méthode d'extraction des ENs en une langue source en utilisant l'approche de corpus parallèles.

Nous utilisons les notations suivantes :

L1 : langue source

L2 : langue cible

C(L1-L2) : corpus parallèle aligné pour les deux langues L1 et l2

CA(L1) : corpus en langue L1 annoté par les ENs.

CA(L2) : corpus en langue L2 annoté par les ENs.

Lex(EN-PERS) : lexique bilingue d'EN-PERS.

Lex(LOC-PERS) : lexique bilingue d'EN-LOC.

Lex(ORG-ORG) : lexique bilingue d'EN-ORG.

ENS : entité nommée en langue source

ENC-Norm : entité nommée en langue cible avec des phonèmes normalisés.

ENC : entité nommée en langue cible

CT : une combinaison de translittération d'une EN en langue source.

Remarque : Chaque phonème de l'EN en langue source peut avoir une ou plusieurs translittérations en langue cible (voir tableau 4.3). Donc on aura plusieurs possibilités de translittération de chaque lexème de cette EN. On appelle les possibilités de translittération *les combinaisons de translittération.*

Pseudo-code Extraction-EN

Début

Entrée : C(L1, L2)

Sortie : CA(L2), Lex(EN-PERS) , Lex(EN-LOC), Lex(EN-ORG).

Étape 0 : Créer des lexiques vides Lex(EN-PERS), Lex(EN-LOC), Lex(EN-ORG)

Étape 1 : Extraction des ENs en langue source

- Prétraitement du corpus source
- Annoter le corpus source par l'outil NER de Stanford pour avoir CA(L1)

Étape 2 : Extraction des ENs en langue cible

Prétraitement du corpus cible

Tant que fin du corpus en langue cible (ou langue source) non atteinte Faire

Normalisation des phonèmes de la phrase du corpus cible

Tant qu'il y a des ENS dans la phrase du CA (L1)

- Projection d'ENS vers la langue cible

Tant qu'il y a encore des combinaisons de translittération possibles d'ENS vers la langue cible Faire

1- Si une combinaison de translittération CT d'ENS existe dans la phrase du corpus cible Alors
   - ENC-Norm reçoit CT
   - ENC reçoit l'ENC-Norm après la suppression des normalisations des phonèmes
   - Ajouter la paire d'EN (ENS, ENC) à l'un des lexiques Lex(EN-PERS), Lex(EN-LOC), Lex(EN-ORG) selon le type d'ENS.
   - Annoter ENC dans la phrase du corpus cible.
   - Sortie de la boucle courante

2- Si CT n'existe pas dans la phrase du corpus cible Alors
   - Passer à la prochaine combinaison de translittération d'ENS

- Passer à la prochaine ENS dans la phrase source

Passer à la prochaine phrase en langue source et à la prochaine phrase en langue cible

Fin

4.2.6 Méthode de projection cross-linguistique

Pour réaliser la procédure de projection cross-linguistique, nous avons développé une méthode de translittération des phonèmes de l'EN en langue source vers la langue cible.

Notre méthode de translittération nécessite un corpus parallèle bilingue. Elle consiste (1) à trouver, pour chaque phonème de l'EN en langue source, toutes les translittérations possibles en langue cible et (2) à trouver la bonne combinaison des phonèmes translittérés dans la phrase cible.

Pour la translittération de l'anglais vers l'arabe, les voyelles peuvent n'avoir aucune translittération vers l'arabe. Ceci, car les voyelles peuvent être les équivalents des signes diacritiques en arabe qui ont été enlevés dans la phase de normalisation des phonèmes arabes. Par exemple, les voyelles 'o', 'a' et 'e' dans le nom de personne 'Mohamed' n'ont aucune translittération vers l'arabe, et donc la translittération de nom de personne 'Mohamed' vers l'arabe est le nom de personne محمد (mHmd).

D'autres phonèmes comme le 'c', 'h' et 's' peuvent n'avoir aucune translittération vers l'arabe. Les deux phonèmes 'c' et 'h' combinés (ch) peuvent être prononcées comme 'h' tout seul ou 'c' tout seul ou les deux 'ch'. Le phonème 's' se trouve dans plusieurs cas à la fin de l'EN et donc il ne se prononce pas comme, par exemple, le nom de lieu 'Athens' qui est translittéré vers l'arabe à 'أثينا' (>vynA / Athènes).

Les translittérations possibles des phonèmes de la langue anglais vers la langue arabe sont illustrées dans le tableau 4.3.

Tableau 4.3 Les translittérations possibles des phonèmes anglais vers l'arabe

Phonème	**Translittérations possibles vers l'arabe** [60]
A	Null, \u0627 (ا), \u064A (ي), \u062A (ت), \u0639 (ع), \u0639\u0627 (عا), \u0627\u064A (اي)
B	\u0628 (ب)
C	\u0633 (س), \u0643 (ك), \u062A (ت), \u062A\u0633(تس)
D	\u062F (د), \u062A (ت), Null
E	Null, \u0639 (ع), \u0627 (ا), \u064A (ي), \u0648 (و), \u0639\u064A (عي), \u0627\u064A (اي)
F	\u0641 (ف)
G	\u063A (غ), \u0643 (ك), \u064A (ي), Null
H	Null, \u062D (ح), \u064A (ي), \u062A (ت), \u0627(ا)
I	Null, \u064A (ي), \u0627 (ا), \u0639 (ع), \u0627\u064A (اي)
J	\u0643 (ك), \u0627 (ا), \u064A(ي)
K	\u0643 (ك)
L	\u0644 (ل), \u064A (ي), Null
M	\u0645 (م)
N	\u0646 (ن)
O	Null, \u0648 (و), \u0627(ا), \u0627\u0648 (او), \u0639 (ع)
P	\u0628 (ب), \u0641 (ف)
Q	\u0643 (ك)
R	\u0631 (ر), \u063A (غ), \u0627 (ا)
S	\u0633 (س), \u062A (ت), Null
T	\u062A (ت), \u062F (د)
U	Null, \u0627 (ا), \u0648 (و), \u0648\u064A (وي), \u064A\u0648 (يو)
V	\u062A (ت), \u0641 (ف)
W	\u0648 (و), \u0641 (ف)
X	\u0643 (ك), \u0627\u0643\u0633 (اكس), \u0643\u0633 (كس)
Y	Null, \u0627 (ا), \u064A (ي)
Z	\u0633 (س), \u062F (د)

[60] Représentation par le code Unicode

Exemple :

Le tableau 4.4 illustre des exemples de translittérations du phonème 'a' avant de passer par l'étape de normalisation des phonèmes arabes.

Tableau 4.4 Exemples des translittérations du phonème 'a' sans normalisation des phonèmes arabes

Phonème	Translittération Buckwalter	Anglais	Arabe
ا	A	**A**hmed	احمد (**A**Hmd)
أ	>	**A**hmed	أحمد (**>**Hmd)
آ	|	**A**dem	آدم (**|**dm)
ؤ	&	Mu**a**yad	مؤيد (m**&**yd)
ة	P	Amin**a**	امينة (Amyn**p**)
ع	E	**A**dnan	عدنان (**E**dnAn)
عا	EA	**A**del	عادل (**EA**dl)
ه	H	Dalil**a**	دليله (Dlyl**h**)
يا	yA	Gi**a**comelli	غياكوميلي (g**yA**kwmyly)
ى	Y	Salm**a**	سلمى (slmY)
ء	'	Haif**a**	هيفاء (hyfA')

La normalisation des phonèmes arabes permet de diminuer le nombre de translittérations possibles de quelques phonèmes anglais. Cela réduire donc le nombre de combinaisons des translittérations possibles d'une EN.

Par exemple, dans le tableau 4.4, après la normalisation des phonèmes arabes, nous avons éliminé les translittérations vers les phonèmes ء, ى, ه , أ, آ, ؤ, ة.

Le tableau 4.5 illustre quelques exemples des différentes translittérations possibles du phonème 'a' après la normalisation des phonèmes arabes.

Tableau 4.5 Exemples des translittérations du phonème 'a' après normalisation des phonèmes arabes

phonème	**Translittération Buckwalter**	**Exemples**		
		Anglais	**Arabe non normalisé**	**Arabe Normalisé**
ا	A	**A**hmed	أحمد (>Hmd)	احمد (**A**Hmd)
		Adem	آدم (|dm)	ادم (**A**dm)
		Mu**a**yad	مؤيد (m&yd)	مايد (m**A**yd)
		Amin**a**	امينة (Amyn**p**)	امينا (Amyn**A**)
		Dalil**a**	دليله (dlyl**h**)	دليلا (dlyl**A**)
		Salm**a**	سلمى (slmY)	سلما (slm**A**)
		Haif**a**	هيفاء (hyfA')	هيفاا (hyfAA)
ع	E	**A**dnan		عدنان (**E**dnAn)
عا	EA	**A**del		عادل (**EA**dl)
يا	yA	Gi**a**comelli		غياكوميلي (g**yA**kwmyly)

Comparaison d'une combinaison de translittération avec un lexème en arabe

Après la détection des ENs du corpus source (anglais), on procède par une projection sur le corpus cible (arabe). Pour cela, chaque translittération de l'EN du corpus source est comparée aux lexèmes de la phrase du corpus cible.

Les combinaisons de translittération de l'EN en langue source peuvent n'avoir aucun sens en langue cible et il faut choisir la meilleure d'entre elles qui existe dans la phrase alignée du corpus cible. Un lexème en langue cible correspond à la meilleure combinaison de translittération, s'il satisfait l'une des conditions suivantes :

Condition 1 : Égalité d'une combinaison avec un lexème arabe

Le cas où nous trouvons un lexème dans la phrase arabe qui est exactement égale à la combinaison testée est le cas le plus simple dans notre méthode de translittération.

Exemples :

Supposons qu'on a les deux phrases suivantes extraites d'un corpus parallèle bilingue (anglais-arabe).

Phrase 1 : « Mohamed Al-Shenawi, Director General of the Inspection Section at Misr-Romania Bank, said … »

Phrase 2 : « .. قال محمد الشناوي مدير عام قطاع التفتيش في بنك مصر رومانيا » (qAl mHmd Al$nAwy mdyr EAm qTAE Altfty$ fy bnk mSr rwmAnyA / Mohamed Al-Shenawi, directeur général de la Section d'inspection au Misr-Roumanie Bank, a déclaré)

L'EN-PERS 'Mohamed Al-Shenawi' est translittérée en arabe à 'محمد الشناوي' (mHmd Al$nAwy), et ce terme existe dans la phrase alignée en arabe donc l'EN arabe est ' محمد الشناوي'.

Condition 2 : Similarité d'une combinaison avec un lexème arabe

Si la combinaison testée n'existe pas dans la phrase en arabe, nous vérifions s'il y a une similarité entre cette combinaison et un des lexèmes de la phrase arabe. Ce cas est plus utilisé pour les EN-PERS. Notre méthode de la similarité est inspirée de la technique de distance d'édition ou distance de Levenshtein (Levenshtein, 1966)[61].

Pour notre cas, la similarité consiste à ajouter ou à enlever un ou deux phonèmes à la fin ou au début de la combinaison. Le problème de l'agglutination dans les ENs en arabe est traité dans ce cas. Par exemple on peut trouver des ENs attachées aux conjonctions au début comme 'و' (waw, et) et 'ب' (b, en) donc on teste, pour chaque combinaison,

[61] La distance de Levenshtein est une distance mathématique permettant de mesurer la similarité entre deux chaines de caractères.

l'existence d'un équivalent dans la phrase arabe avec l'ajout au début de l'une de ces conjonctions.

Exemples :

Ex 1 : L'EN-PERS 'Salman' est translittérée à 'السلمان' (**Al**slmAn) qui est égale à la combinaison 'سلمان' (slmAn) avec l'ajout de l'article de définition 'ال' (Al - le) au début.

Ex 2 : Pour les noms de lieu, on peut trouver dans la phrase arabe le terme 'بامريكا' (bAmrykA – en Amérique) qui est équivalent à l'EN-LOC 'امريكا' (AmrykA), mais avec la suppression de la conjonction 'ب' (b , en) au début.

Ex 3 : On peut trouver dans la phrase arabe le terme 'وبيتر' (wbytr) qui est équivalent à l'EN-PERS 'Peter', mais avec l'ajout de la conjonction 'و' (w, et) au début.

Ex 4 : L'EN-PERS 'Hasna' est translittérée à 'حسنا' (HsnA), mais dans la phrase arabe son équivalent est 'حسناء' (HsnA'), donc on doit ajouter le phonème 'ء' (') à la fin.

Condition 3 : Égalité d'une combinaison avec plusieurs lexèmes arabes

Les langues arabe et anglaise sont complètement différentes du point de vue morphologique. Parfois un terme en anglais correspond à deux termes en arabe. Nous avons intégré ce cas dans notre procédure de translittération. Donc si les deux conditions précédentes ne donnent aucun terme en arabe, nous testons s'il y a deux termes adjacents qui sont égaux à la combinaison testée.

Exemple :

L'EN-LOC 'Higashiyamato' correspond à l'EN composée 'هيغاشي ياماتو' (hygA$y yAmAtw).

Condition 4 : Similarité d'une combinaison avec plusieurs lexèmes arabes

Comme dans la condition 2, une similarité peut exister dans le cas où une EN correspond à deux lexèmes en arabe.

Exemple :

L'EN-LOC 'Higashiyamato' est l'équivalent de l'EN en arabe 'هيغاشي ياماتو' (hygA$y yAmAtw), mais dans la phrase arabe on peut trouver cette EN attachée à la conjonction 'و' (w, et) au début, alors on doit enlever cette conjonction.

4.2.7 Dictionnaire de marqueurs lexicaux

Généralement les ENs sont associées avec des marqueurs lexicaux qui aident leur reconnaissance. Par exemple, les marqueurs lexicaux 'Mr', 'Mrs', 'Dr' et 'Dre' sont associés aux EN-PERS.

À l'aide des marqueurs lexicaux, nous pouvons éviter l'ambiguïté dans la reconnaissance du type de l'EN. Par exemple, l'expression 'Airport Mohamed V' (مطار محمد الخامس - mTAr mHmd AlxAms) contient un nom de personne qui est 'Mohamed', mais elle contient aussi le terme 'Airport' qui caractérise une EN-LOC et non une EN-PERS.

Les marqueurs lexicaux ne peuvent pas être translittérés, mais plutôt ils peuvent être traduits. Pour cela nous avons créé un dictionnaire (anglais-arabe) des marqueurs lexicaux pour chaque type d'EN. Ces dictionnaires contiennent toutes les traductions possibles en arabe[62] de chaque marqueur lexical en anglais. Ensuite, nous cherchons l'une de ces traductions qui existe dans la phrase en arabe. Nous avons nommé ces dictionnaires *Dic_Marqueur_Pers*, *Dic_Marqueur_Loc, Dic_Marqueur_Org* pour les EN-PERS, les EN-LOC, les EN-ORG respectivement.

Le dictionnaire Dic_Marqueur_Loc contient aussi les noms de pays qui ne peuvent être translittérés, par exemple la translittération de l'EN-LOC 'Egypt' donne le terme 'اجيبت' (Ajybt), par contre l'EN en arabe qui correspond à cette EN est 'مصر' (mSr).

---

[62] Nous avons utilisé Google Translate pour trouver les différentes traductions possibles en arabe d'un marqueur lexical en anglais.

Les tableaux 4.6, 4.7 et 4.8 illustrent quelques exemples de contenu du dictionnaire Dic_Marqueur_Pers, Dic_Marqueur_Loc et Dic_Marqueur_Org respectivement.

Tableau 4.6 Exemple de Dic_Marqueur_Pers

Terme anglais	**Terme arabe**	**Translittération en Buckwalter**
Mr.	السيد	Alsyd
Mrs.	السيدة	Alsydp
Ms.	السيدة	Alsydp
Dame	السيدة	Alsydp
Dr.	الدكتور	Aldktwr
Dre.	الدكتورة	Aldktwrp
King	الملك	Almlk
Queen	الملكة	Almlkp
Prince	الامير	AlAmyr
Brigadier	العميد	AlEmyd
President	الرئيس	Alr}ys
.....		

Remarque

Si une EN-PERS débute par un marqueur lexical, nous enlevons ce marqueur de l'EN, mais nous cherchons dans Dic_Marqueur_Pers sa traduction qui existe dans la phrase arabe pour l'ajouter à l'EN en arabe.

Tableau 4.7 Exemple de Dic_Marqueur_Loc

Terme anglais	Terme (s) arabe	Translittération en Buckwalter
Algeria	الجزائر	AljzA}r
Egypt	مصر	mSr
Camp	معسكر, مخيم	Mxym, mEskr
Park	متنزه, حديقة	Hdyqp, Mtnzh
District	دائرة, حي, منطقة, مقاطعة	mqATEp, mnTqp, Hy, dA}rp
Region	إقليم, منطقة	mnTqp, <qlym
Airport	مطار	mTAr
Republic	جمهورية	Jmhwryp
Province	محافظة,اقليم, مقاطعة	mqATEp, Aqlym, mHAfZp
River	وادي, نهر	Nhr, wAdy
...		

Tableau 4.8 Exemple de Dic_Marqueur_Org

Terme anglais	Terme (s) arabe	Translittération en Buckwalter
Committee	لجنة	Ljnp
Unit	وحدة	wHdp
Company	مؤسسة, شركة	$rkp, m&ssp
Institution	مؤسسة	m&ssp
Party	طرف, حزب	Hzb, Trf
Movement	حركة	Hrkp
F^çc^çv^^^f^^ Organization	مؤسسة, منظمة	mnZmp, m&ssp
.....		

4.2.8 Translittération des entités nommées simples

La procédure de translittération des ENs simple est la même pour les trois classes d'EN étudiées dans ce mémoire. Dans cette section, nous donnons le pseudo-code de cette procédure qui prend en entrée l'EN en langue source et produit l'EN en langue cible. Nous utilisons les notations suivantes :

ENS : Une EN simple en langue source;

CT(ENS) : Une combinaison de translittération de l'ENS;

ENC-Norm : Une EN simple en langue cible avec des phonèmes normalisés

ENC : Une EN simple en langue cible avec des phonèmes non normalisés;

Procédure Translittération-EN-Simple (ENS) : ENC

Début

Si ENS existe dans le dictionnaire Dic_Marqueur_Loc Alors

ENC reçoit l'une des traductions d'ENS qui existe dans la phrase en arabe.

Sinon

Tant qu'il y a encore des CT(ENS) Faire

S'il existe un lexème dans la phrase arabe qui est équivalent à CT(ENS) Alors

- ENC-Norm reçoit CT(ENS);
- Enlever la normalisation des phonèmes de l'ENC-Norm, et ENC reçoit cette nouvelle EN;
- Sortie de la boucle Tant que;

Sinon

- Passer à la prochaine combinaison de translittération de l'ENS;

Fin tant que

Fin Si

Retourner ENC;

Fin

Remarque : La valeur 'null' désigne que le phonème en anglais n'a aucune translittération en arabe. Cette valeur est supprimée lorsqu'on cherche la comparaison dans la phrase arabe.

- **Exemple de translittération d'une EN simple**

Phrase en anglais : ".....<PERSON>Mr. Nidal</PERSON>....."

Phrase en arabe : ".....السيد نضال....." (....Alsyd nDAl .....)

Après l'étape de la normalisation, la phrase en arabe devient : "... السيد ندال....." (...Alsyd ndAl.....)

En translittérant chaque phonème de L'EN-PERS 'Nidal' suivant les translittérations présentées dans le tableau 4.3, on trouve les combinaisons de translittération illustrées dans le tableau 4.9.

Tableau 4.9 Exemple de translittération d'une EN simple

Combinaison	**Explication**
ن null د null ل (ندل)	Cette combinaison n'existe pas dans la phrase arabe normalisée.
ن null د null (ندي)	Cette combinaison n'existe pas dans la phrase arabe normalisée.
ن nullد null null (ند)	Cette combinaison n'existe pas dans la phrase arabe normalisée.
ن null د ا ل (ندال)	Cette combinaison existe dans la phrase arabe normalisée, donc c'est la meilleure combinaison. Le terme équivalent sans phonèmes normalisés est l'EN-PERS 'السيد نضال' (Alsyd nDAl).

Dans cet exemple, s'il n'y a pas l'étape de la normalisation, le nombre de combinaisons de translittérations possibles augmente, car les phonèmes 'ض' (\u0636) et 'ظ' (\u0638) doivent être ajoutés dans la liste de translittérations possibles du phonème 'D' (tableau 4.3), ce qui donne la liste suivante (\u0636, \u0638, \u062F, \u062A, Null) au lieu de (\u062F, \u062A, Null). Donc l'étape de la normalisation diminue le nombre de combinaisons de translittérations possibles de l'EN en langue source.

4.2.9 Translittération des entités nommées composées

Pour la translittération des ENs composées, il y a deux procédures. La première procédure est pour l'extraction des EN-PERS et EN- LOC, et la deuxième procédure est pour l'extraction des EN-ORG. Généralement les EN-ORG composées ne peuvent être translittérées, mais en utilisant le lexique Dic_Marqueur_Org, nous pouvons translittérer quelques EN-ORG.

Nous présentons dans cette section, les pseudo-codes de l'extraction des ENs composées. Nous utilisons les notations suivantes :

N : nombre entier ;

ENS : une EN composée en langue source ;

LS : lexème en langue source ;

LC : lexème en langue cible ;

ENC : une EN composée en langue cible avec des phonèmes non normalisés;

MS : marqueur lexical en langue source

MC : marqueur lexical en langue cible

Procédure Translittération-EN-PERS-LOC-Composée (ENS) : ENC

Début

N reçoit le nombre de lexèmes de l'ENS ;

Si ENS contient un marqueur lexical MS parmi les marqueurs lexicaux de Dic_Marqueur_Loc Alors

- Décrémenter la valeur de N ;
- Chercher dans Dic_Marqueur_Loc la traduction de MS qui existe dans la phrase arabe soit MC ;

Si MS n'est pas le dernier lexème dans ENS Alors

- LS reçoit le lexème qui suive MS dans ENS ;
- LC reçoit Translittération-EN-Simple (LS) ;
- ENC reçoit MC concaténée avec LC et les (N – 1) lexèmes qui suivent LC dans la phrase arabe non normalisée; (les lexèmes sont séparés par un espace).

Sinon (si MS est le dernier lexème dans ENS)

S'il existe un seul lexème dans la phrase arabe non normalisée qui correspond à MC Alors

- ENC reçoit MC concaténé avec les (N – 1) lexèmes qui suivent MC dans la phrase arabe non normalisée ;

Sinon

- LS reçoit le lexème qui précède MS dans ENS ;
- LC reçoit Translittération-EN-Simple (LS) ;
- ENC reçoit MC concaténée avec LC et les (N – 1) lexèmes qui suivent LC dans la phrase arabe non normalisée ;

Sinon (l'ENS ne contient aucun lexème de Dic_Marqueur_Loc, l'EN peut être EN-LOC ou EN-PERS)

- LS reçoit le premier lexème d'ENS ;
- LC reçoit Translittération-EN-Simple (LS) ;
- ENC reçoit le lexème LC concaténé avec les (N – 1) lexèmes qui le suivent dans la phrase arabe non normalisée;

Fin si

Retourner ENC ;

Fin

Remarque 1 : Pour les EN-LOC, si le lexème qui suit MS appartient aussi au Dic_Marqueur_Loc, nous décrémentons encore la valeur de N et nous ajoutons aussi à l'ENC la traduction de ce lexème. Nous avons limité la recherche à trois lexèmes de Dic_Marqueur_Loc.

Remarque 2 : Pour la valeur de N, nous ne comptons pas les marqueurs lexicaux des EN-PERS, les articles et les prépositions.

Remarque 3 : Comme nous avons vu, l'ordre des lexèmes d'une EN en arabe peut être différent de celui de l'EN en anglais. Ce cas est traité en cherchant la combinaison des lexèmes qui existe dans la phrase en arabe.

Exemple :

L'équivalent en arabe de l'EN 'Province of Misiones' est 'محافظة ميسيونس' (mHAfZp mysywns). Cet équivalent est trouvé en translittérant le terme 'Misiones' vers l'arabe, et en combinant cette translittération avec la traduction du terme 'Province', qui est 'محافظة' (mHAfZp) (celle qui existe dans la phrase arabe).cependant, pour l'EN 'South America', nous trouvons l'EN en arabe أمريكا الجنوبية (>mrykA Aljnwbyp). Dans ce cas l'ordre des lexèmes de l'EN anglaise n'est pas le même que celui de l'EN arabe donc on doit chercher l'ordre qui existe dans la phrase arabe.

Remarque 4 : Il y a des cas particuliers pour les noms d'origine arabe. En effet, quelque un de ces noms commencent ou contiennent un terme qui désigne que c'est un nom arabe, comme les termes suivants : abd, abdel, abd el, abdal, abd al, ebdul, ebd ul, ebd, ebdal, ibn, ben, abu, abuel, abu el, etc. L'écriture en anglais de ce genre de noms contient, souvent, une liaison '-'. Nous avons traité ce cas pour ne pas avoir des ENs en arabe incomplètes ou avec des termes en plus. Par exemple, l'EN en anglais 'Khalid Abu-Ismail' contient 2 lexèmes qui sont 'Khalid' et 'Abu-Ismail'. L'équivalent en arabe de cette EN est 'خالد أبو إسماعيل' (xAld >bw <smAEyl) et elle contient 3 lexèmes qui sont : 'خالد' (xAld), 'أبو' (>bw) et 'إسماعيل' (<smAEyl).

Pour avoir le même nombre de lexèmes pour les deux ENs en langue source et cible, la liaison '-' a été remplacée par un espace.

- **Exemple de translittération d'une EN composée**

La phrase extraite du corpus anglais annoté est : ".... <PERSON> Aziz bin Saud bin Nayef Al-Saud </PERSON>....."

La phrase extraite du corpus arabe est : " ....العزيز بن سعود بن نايف آل سعود....." (....AlEzyz bn sEwd bn nAyf |l sEwd .....).

La phrase en arabe après la normalisation des phonèmes est " ... العسيس بن سعود بن نايف ال سعود.... " (AlEsys bn sEwd bn nAyf Al sEwd).

Après la suppression de la liaison '-', on trouve le nombre de lexèmes de l'EN en anglais N = 7. On teste si l'une des combinaisons de translittérations possibles de premier terme 'Aziz' existe dans la phrase arabe. On trouve les combinaisons illustrées dans le tableau 4.10.

Tableau 4.10 Exemple de translittération d'EN composée

Combinaison	**Explication**
nullس nullس (سس)	Cette combinaison n'existe pas dans la phrase arabe normalisée.
nullس null د (سد)	Cette combinaison n'existe pas dans la phrase arabe normalisée.
Null سيس (سيس)	Cette combinaison n'existe pas dans la phrase arabe normalisée.
...	Autres combinaisons qui n'existent pas dans la phrase arabe normalisée
عسيس	Cette combinaison existe dans la phrase arabe normalisée, avec l'ajout de l'article de définition 'ال' (Al), ce qui donne 'العسيس' (AlEsys). En ajoutant les N -1 = 6 termes qui existent après le terme 'العسيس', nous trouvons 'العسيس بن سعود بن نايف ال سعود'. Après l'enlèvement de la normalisation, l'EN devient 'العزيز بن سعود بن نايف آل سعود'

Dans cet exemple, s'il n'y a pas l'étape de la normalisation, le nombre de combinaisons de translittérations possibles augmente, car le phonème 'ز ' (\u0632) doit être ajoutée dans la liste des translittérations possibles du phonème 'Z' (tableau 4.3), ce qui donne la liste suivante (\u0632, \u0633, \u062F) au lieu de (\u0633, \u062F).

Procédure Translittération-EN-ORG-Composée (ENS) : ENC

Début

N reçoit le nombre de lexèmes de l'ENS ;

Si ENS contient un marqueur lexical MS parmi les marqueurs lexicaux de Dic_Marqueur_Org Alors

- Chercher dans Dic_Marqueur_Org la traduction de MS qui existe dans la phrase arabe soit MC ;
- ENC reçoit MC concaténé avec les (N – 1) lexèmes qui suivent MC dans la phrase arabe non normalisée ; (les lexèmes sont séparés par un espace).

Retourner ENC ;

Fin

Remarque :

Pour les EN-ORG, nous devons assurer qu'il n'y a pas une autre EN en anglais dans la même phrase qui contient le même marqueur lexical MS, car il arrive que l'ordre des ENs dans la phrase en anglais ne soit pas le même que dans la phrase en arabe, et dans ce cas on n'extrait pas la bonne EN en arabe.

Exemple :

L'EN 'Electoral Assistance Unit' contient un terme parmi les termes du dictionnaire Dic_Termes_Org qui est 'Unit'. L'équivalent en arabe de ce terme, qui est 'وحدة' (wHdp), existe dans la phrase en arabe, donc nous ajoutons les deux termes arabes qui le suivent et nous obtenons 'وحدة المساعدة الانتخابية' (wHdp AlmsAEdp AlAntxAbyp).

4.2.10 Annotation du corpus cible et construction de lexiques d'entités nommées

Parmi les objectifs de ce mémoire, l'annotation du corpus cible et la construction de lexiques bilingues d'ENs. Pour cela, nous avons utilisé l'annotation de balisage pour étiqueter les ENs extraites dans le corpus cible. Les balises utilisées sont <PERS> et

</PERS> pour les EN-PERS, <LOC> et </LOC> pour les EN-LOC et <ORG> et </ORG> pour les EN-ORG.

L'appendice B illustre un exemple de deux échantillons d'un corpus en anglais et d'un corpus en arabe annotés par les ENs après l'application de notre méthode d'extraction des ENs.

Pour la construction des lexiques bilingues d'ENs, elle se fait au fur et à mesure de l'extraction des ENs en langue cible. À chaque fois qu'une EN en langue cible est extraite, la paire (EN en langue source, EN en langue cible) est ajoutée dans le lexique d'ENs selon le type de l'EN en langue source.

4.2.11 Comparaison de notre méthode de translittération avec l'état de l'art

Dans la littérature, la procédure de projection cross-linguistique des ENs d'une langue source vers une langue cible est réalisée dans quelques travaux avec d'autres paires de langues et dans de grande envergure tels que le projet JRC-Names de la Commission européenne du centre commun de recherche (Steinberger et al., 2011)[63].

Cette méthode de translittération a été utilisée dans le travail de Samy et al. pour la paire de langues espagnol-arabe (Samy et al., 2005). Cependant, dans ce mémoire, nous avons ajouté la méthode de normalisation d'un ensemble de phonèmes en langue cible vers un seul phonème en même langue. Cette normalisation a permis de diminuer le nombre de combinaisons de translittérations possibles d'une EN, ce qui réduit le nombre de phonèmes en langue cible. La méthode de normalisation des phonèmes en une langue source vers une langue cible est efficace dans le cas des langues différentes du point de vue morphologique comme le cas de l'anglais et l'arabe. En effet, pour cette paire de langues, si nous ne normalisons pas les phonèmes arabes, quelques phonèmes

[63] Présentation en ligne http://tln.li.univ-tours.fr/Tln_Colloques/Tln_REN2011/Ehrmann-ATALA-20juin2011.pdf , Journée ATALA Entités Nommées - Lundi 20 Juin 2011

anglais peuvent être translittérés par plusieurs phonèmes arabes ce qui augmente le nombre de combinaisons de translittérations possibles.

4.3 Traduction automatique statistique

Les lexiques d'ENs construit dans ce mémoire ont été évalués dans un système de TAS à base de segments en utilisant le décodeur Moses.

Dans notre méthode d'extraction des ENs à partir de corpus parallèles bilingues, nous nous sommes intéressés à la paire de langues anglais-arabe. Cependant, dans le but de participer à la campagne d'évaluation TRAD 2014[64] qui exige des travaux pour la paire de langues arabe-français, nous avons ajouté la langue française à notre étude. Pour cela, nous avons utilisé une technique simple pour translittérer les ENs extraites de l'arabe vers le français, en utilisant l'anglais comme une langue pivot. Une langue pivot est une langue intermédiaire utilisée pour faciliter les traductions d'un même texte dans plusieurs langues[65].

Dans cette section, nous présentons d'abord cette technique de translittération, ensuite nous décrivons la méthodologie suivie pour la construction de système de TAS.

4.3.1 Translittération des entités nommées de l'arabe vers le français

Notre technique de translittération d'une EN de l'arabe vers le français en utilisant l'anglais comme une langue pivot est basée sur :

- **Une liste des ENs en anglais** : pour avoir cette liste, nous avons utilisé les lexiques d'ENs construits par la méthode de projection cross-linguistique.

[64] La campagne d'évaluation TRAD est un cadre commun pour rendre compte des performances actuelles des systèmes de traduction automatique pour le couple de langues arabe-français. http://www.trad-campaign.org/

[65] http://fr.wikipedia.org/wiki/Langue_pivot

- **Un corpus parallèle anglais-français** : nous avons exploité les mêmes corpus parallèles qui sont utilisés pour l'extraction des ENs arabes, mais pour la paire de langues anglais-français.

Cette technique de translittération exploite le fait que l'anglais et le français utilisent des alphabets similaires. Cela rend la translittération de l'anglais vers le français plus facile qu'entre deux langues qui sont différentes dans leur écriture et morphologie comme l'arabe et le français.

La translittération de l'anglais vers le français se base sur le calcul de la similarité entre l'EN en anglais et chaque mot dans la phrase française alignée. Donc pour chaque EN en anglais du lexique (anglais-arabe) construit, nous cherchons dans quelle phrase est située cette EN dans le corpus en anglais. Ensuite, nous cherchons la phrase en français alignée dans le corpus parallèle anglais-français. Puis nous calculons la similarité de chaque mot dans cette phrase, et nous choisissons le mot qui a le meilleur score.

La similarité entre une EN en anglais et un mot en français a été réalisée par la technique Editex (Zobel et Dart, 1996) et la distance de Levenshtein (Levenshtein, 1966). Editex(Zobel et Dart, 1996) définit neuf groupes phonétiques correspondent souvent à une prononciation similaire qui sont : g0(a, e, i, o, u, y, h, w), g1(b, p), g2(c g j k q), g3(d, t), g4(l), g5(m, n), g6(r), g7(f, v), g8(s, x, z). Le calcul de la similarité entre les mots se base sur ces 9 groupes de phonèmes et la distance de Levenshtein.

Le choix du meilleur score de similarité nous donne le mot en français équivalent à l'EN en anglais. Ensuite, en utilisant le lexique anglais-arabe, nous choisissons l'EN en arabe équivalent pour avoir la paire d'EN arabe-français.

4.3.2 Description du système de TAS

Le système construit est basé sur la boite à outils libre Moses (Koehn et al., 2007) en utilisant ses paramètres par défaut. Moses permet de construire un système de TAS par segments.

Prétraitement des corpus source et cible

Parmi les conditions pour la réalisation d'un système de TAS, l'utilisation des corpus parallèles alignés en deux langues source et cible et des corpus monolingues pour la langue cible.

Pour notre cas, la langue source est l'arabe et la langue cible est le français. Les corpus d'entraînement et de test ont été prétraités pour chaque langue.

Pour les corpus en français, un simple prétraitement a été réalisé par la transformation des majuscules en minuscules et la segmentation des phrases.

Pour les corpus en arabe, soit pour les données d'entraînement ou pour les données de développement et de test, le prétraitement a été fait avec l'analyseur morphologique MADA (Habash et al., 2009).

- **MADA**

MADA (en anglais, Morphological Analysis and Disambiguation for Arabic) est un analyseur morphologique disponible librement. C'est un outil de désambiguïsation pour la langue arabe.

La première étape de prétraitement d'un texte en arabe avec MADA est la translittération de ce texte par l'encodage Buckwalter. La deuxième étape est l'ajout des informations lexicales et morphologiques pour lever l'ambiguïté des mots. Le résultat de cette étape est un fichier segmenté et translittéré suivant la norme de Buckwalter.

Les termes arabes ont été segmentés par le système D3 (Habash et Sadat, 2006). Le système D3 sépare tous les clitiques en trois classes.

D1- La classe des conjonctions : و (w) et ف (f)

D2- La classe des particules : ل (l), ك (k), ب (b), et س (s)

D3- La classe des articles de définition ال (Al) et tous les pronoms enclitiques.

Exemple (Habash et al., 2009):

On a la phrase suivante en arabe : 'وسينهي الرئيس جولته بزيارة الى تركيا' (wsynhY Alr}ys jwlth bzyArp <lY trkyA / Le président va terminer sa tournée par une visite en Turquie.)
Le résultat du prétraitement avec MADA de cette phrase est illustré dans le tableau 4.11.

Tableau 4.11 Exemple de prétraitement avec MADA

Segments de la phrase	wsynhY	Alr}ys	Jwlth	bzyArp	<lY	trkyA
Classe D1	w+ synhy	Alr}ys	Jwlth	bzyArp	<lY	trkyA
Classe D2	w+ s+ ynhy	Alr}ys	jwlth	b+zyArp	<lY	trkyA
Classe D3	w+ s+ ynhy	Al+ r}ys	jwlp + 3ieme Pronom	b+ zyArp	<lY	trkyA

Le modèle de langue

Pour implémenter le modèle de langue pour la langue cible (français), l'outil SRILM (Stolcke, 2002) a été utilisé. Afin d'avoir de bons résultats d'entraînement, nous avons utilisé le modèle de langue en 5-grammes.

Le modèle de traduction

Un alignement par mot a été fait en utilisant le système GIZA++ (Och et Ney, 2003) dans les deux sens des langues des corpus. Nous avons choisi 50 mots comme une longueur maximale des phrases.

La figure 4.2 montre une partie de la table de phrases utilisée dans notre système de TAS. La phrase arabe est translittérée suivant la norme de Buckwalter.

Figure 4.2 Exemple d'une table de traduction (arabe-français)

```
Al+ AEtmAdAt Al+ mxSSp l+ ||| à celui de ||| 0.000677461 2.9934e-13 0.00388649 2.63774e-08
Al+ AEtmAdAt Al+ mxSSp mn Al+ Sndwq ||| les crédits alloués au moyen du Fonds ||| 0.147687 8.36738e-05 0.147687 6.4742e-11 2.718
Al+ AEtmAdAt Al+ mxSSp ||| des crédits alloués ||| 0.00369217 0.00876646 0.00263726 0.000428061 2.718
Al+ AEtmAdAt Al+ mxSSp ||| des crédits approuvés au titre ||| 0.147687 1.8346e-05 0.00263726 1.51814e-09 2.718
Al+ AEtmAdAt AlmnqHp l+ ||| les crédits révisés pour ||| 0.435516 0.0311052 0.0378709 0.000281919 2.718
Al+ AEtmAdAt AlmnqHp l+ ||| montant révisé des crédits ouverts pour ||| 0.275143 0.0137832 0.167478 7.1305e-06 2.718
Al+ AEtmAdAt AlmnqHp llgyldr ||| le crédit révisé dans ||| 0.0738433 7.37109e-10 0.0738433 6.57125e-06 2.718
Al+ AEtmAdAt Hsb ||| des crédits ouverts par ||| 0.0070327 0.00162003 0.0738433 0.000674679 2.718
Al+ AEtmAdAt Hsb Al+ ||| milliers ||| 4.74571e-05 1.26095e-06 0.0738433 0.001834 2
Al+ AEtqAd Al+ ||| la conviction ||| 0.00351223 0.0196402 0.0348413 0.00803394 2.718
```

4.3.3 Les configurations du système de TAS

Pour la réalisation de nos expériences pour la campagne d'évaluation TRAD 2014, nous avons utilisé deux systèmes différents. Le système de base et le système à base de lexiques d'ENs. Ces deux systèmes sont décrits comme suit :

- **Système de base (système 1)**

Ce système est considéré comme la base à qui le système 2 sera comparé. Les données utilisées pour la réalisation des expériences dans le système de base ont été limitées aux données parallèles fournies par la campagne TRAD. La campagne d'évaluation TRAD

fournit des données d'apprentissage, de développement et d'évaluation pour les participants. Toutes les données fournies sont sous encodage UTF-8, et elles ont trois formats qui sont : les données d'apprentissage, les données de développement et les données de test.

Les corpus d'apprentissage utilisés pour le modèle de traduction sont illustrés dans le tableau 4.12.

Tableau 4.12 Corpus d'apprentissage du modèle de traduction

Corpus	**Description**
News Commentary	Obtenu par la campagne d'évaluation TRAD 2012 pour l'apprentissage des modèles de langue et de traduction du système de base. Ce corpus comporte 90 753 paires de phrases.
Nist08	Contient 813 paires phrases.
MultiUN	Corpus des Nations Unies

Les corpus d'apprentissage du modèle de langue sont ceux de modèle de traduction (voir tableau 4.12) plus le corpus Europarl français[66].

Les données d'apprentissage sont composées de plus de 3.5 millions de mots en français et en arabe.

Pour les données de développement, nous avons utilisé les données de test obtenues lors de la première édition de TRAD 2012 (Sellami et al., 2013). Ces données comportent 423 (10 000 mots) phrases en langue source avec 4 références

Les données de test de notre système sont les données de test fournies par la deuxième édition de TRAD 2014. Elles sont composées de 352 phrases arabes avec deux références.

[66] http://www.europarl.europa.eu/plenary/en/home.html?language=fr

- **Système à base de lexiques bilingues d'ENs (système 2)**

Dans ce système, nous ajoutons aux données du système de base les lexiques d'EN-PERS et d'EN-LOC pour les deux langues française et arabe. Nous ajoutons aussi deux lexiques construits à partir des ressources linguistiques :Geonames, DBPedia et JRC-name (voir section 5.5.1).

4.4 Conclusion

Dans ce chapitre, nous avons présenté notre méthodologie qui est composée de deux parties. La première partie est l'extraction des ENs en arabe à partir de corpus parallèles bilingues. La deuxième partie est la construction d'un système de TAS basé sur le décodeur Moses.

Notre méthode de translittération des ENs de l'anglais vers l'arabe est basée en premier sur la recherche de toutes les translittérations possibles de chaque phonème de l'EN anglais vers l'arabe, ensuite sur la recherche de la meilleure translittération produite et existante dans la phrase alignée du corpus cible (arabe).

Notre solution pour l'extraction des ENs a permis d'annoter le corpus cible par les ENs et de construire des lexiques bilingues d'ENs.

Pour le système de TAS construit, nous avons présenté une méthode de translittération des ENs de l'arabe vers le français en utilisant l'anglais comme une langue pivot pour avoir des lexiques bilingues d'ENs pour la paire de langues arabe-français. Nous avons décrit aussi les deux configurations du système de TAS qui sont utilisés pour l'évaluation des lexiques d'ENs construits. Le chapitre suivant présente cette évaluation ainsi que trois autres évaluations pour tester la performance de notre méthode de translittération des ENs.

CHAPITRE V

ÉVALUATIONS

Pour tester la performance de notre méthode de projection cross-linguistiques d'une langue source vers une langue cible en utilisant les corpus parallèles alignés, nous avons évalué les lexiques bilingues d'ENs résultants et les deux corpus annotés à travers plusieurs évaluations.

Les lexiques et les corpus annotés peuvent être évalués en se basant sur l'une des façons suivantes :

1- La plateforme collaborative d'Amazon (en anglais, Amazon Mechanical Turk)[67] qui est une application web de crowdsourcing lancé par Amazon.com en 2005. L'objectif de cette plateforme est la réalisation des tâches rémunérées contre des évaluations manuelles effectuées par des humains. Les tâches en question ne doivent pas dépendre d'un support physique.

2- Un expert en linguistique qui maîtrise bien la langue source et la langue cible.

3- Par une application de TALN.

Dans ce mémoire, nous avons choisi les deuxième et troisième solutions. Pour réaliser la deuxième solution, nous avons demandé à un expert linguistique qui maîtrise les langues arabe et anglaise d'évaluer les corpus annotés et les lexiques bilingues d'ENs construits. Cette évaluation a été exprimée par les métriques classiques d'évaluation qui sont le rappel, la précision et la F-mesure.

---

[67] https://www.mturk.com/mturk/

Pour la troisième solution, l'application TALN consiste l'incorporation des lexiques d'ENs dans le système de TAS construit. Nous avons limité notre évaluation aux EN-PERS et EN-LOC. Nous avons ajouté deux autres lexiques construits à partir des ressources linguistiques JRC-Names, Geonames et DBPedia. Cette évaluation est exprimée par la valeur du score BLEU (Papineni et al., 2002) et le nombre des mots MHV.

Nous avons utilisé deux évaluations supplémentaires. La première consiste en une comparaison avec le système de traduction Google Translate[68]. La deuxième consiste en une comparaison de lexiques bilingues d'ENs obtenus par le corpus des titres de Wikipédia avec les deux lexiques d'ENs présentés dans (Mohit et al., 2012), (Azab et al., 2013) et (Alotaibi et Lee, 2013).

Dans ce chapitre, nous commençons par la présentation des données de l'évaluation (les corpus). Ensuite, nous détaillons toutes les évaluations dans l'ordre suivant : la première évaluation qui est faite par l'expert linguistique. La deuxième évaluation qui est la comparaison avec Google Translate. La troisième évaluation qui est la comparaison de lexique obtenu à partir de Wikipédia avec d'autres lexiques d'ENs. La dernière évaluation est l'incorporation des lexiques d'ENs dans un système de TAS.

5.1 Données de l'évaluation

Dans cette section, nous présentons la description des corpus parallèles que nous avons utilisés. Nous indiquons aussi la taille des lexiques bilingues d'ENs obtenus par l'application de notre méthode d'extraction des ENs sur les deux corpus présentés.

5.1.1 Corpus des Nations Unies UN

Les corpus parallèles de l'UN (Organisation des Nations Unies) sont disponibles dans six langues : arabe, chinois, anglais, français, russe et espagnol. Ils sont très utilisés par

[68] Ces évaluations ont été faites en ligne en juin 2014

les chercheurs, surtout dans le domaine de la reconnaissance des ENs, car ces corpus contiennent un grand nombre d'EN. Dans les corpus UN, on peut trouver des EN-PERS de plusieurs origines. Ceci représente un bon exemple pour tester notre méthode de translittération des ENs. Les corpus UN sont disponibles en ligne via le système de documents officiels de l'UN[69].

Le corpus UN utilisé est dans les deux langues arabe et anglaise. Nous avons utilisé l'outil Hunalign[70] pour aligner ce corpus au niveau des phrases.

La taille du corpus UN est illustré dans le tableau 5.1.

Tableau 5.1 Taille du corpus UN

Langue	**Nombre de phrases**	**Nombre de mots**
Anglais (corpus source)	3,882,645	118,875,041
Arabe (corpus cible)	3,882,645	104,215,163

5.1.2 Les titres des articles de Wikipédia

L'utilisation de la ressource Wikipédia a été limitée aux titres, arabes et anglais, des articles de Wikipédia. Pour cela nous avons utilisé les deux archives[71] en anglais et en arabe de titres Wikipédia. Le document obtenu est aligné au niveau de phrases et sa taille est illustrée dans le tableau le tableau 5.2.

Tableau 5.2 Taille du corpus des titres de Wikipédia

Langue	**Nombre de phrases**	**Nombre de mots**
Anglais (corpus source)	137,968	348,265
Arabe (corpus cible)	137,968	347,903

[69] http://documents.un.org.

[70] http://mokk.bme.hu/en/resources/hunalign/

[71] http://dumps.wikimedia.org/arwiki/latest/

5.1.3 Taille des lexiques bilingues d'entités nommées construits

Le tableau 5.3 montre la taille en terme du nombre d'EN de chacun des lexiques construits par notre méthode de projection cross-linguistique pour le corpus UN et les titres de Wikipédia.

Tableau 5.3 Taille des lexiques bilingues d'ENs

	Corpus UN	**Titres Wikipédia**
Lexique d'EN-PERS	227,299	131,576
Lexique d'EN-LOC	14,075	1966
Lexique d'EN-ORG	56,636	5362

5.2 Première évaluation : précision, rappel et F-mesure

La première évaluation consiste à évaluer les résultats obtenus par notre méthode de translittération pour le corpus UN et le corpus des titres de Wikipédia. Pour cette évaluation, nous avons demandé à un expert linguistique qui maîtrise les deux langues anglaise et arabe d'évaluer les corpus annotés et les lexiques bilingues d'ENs. La tâche de l'expert consistait à vérifier si les ENs ont été correctement translittérées de l'anglais vers l'arabe.

Les résultats de cette évaluation ont été traduits par le calcul des valeurs des métriques classiques qui sont la précision, le rappel et la F-mesure.

5.2.1 Évaluation pour le corpus UN

Résultats obtenus

Nous commençons nos évaluations par la présentation des résultats obtenus par l'évaluation du corpus UN en entier (3,882,645 phrases). Le tableau 5.4 illustre les valeurs de la précision, du rappel et de la F-mesure pour les trois types d'EN.

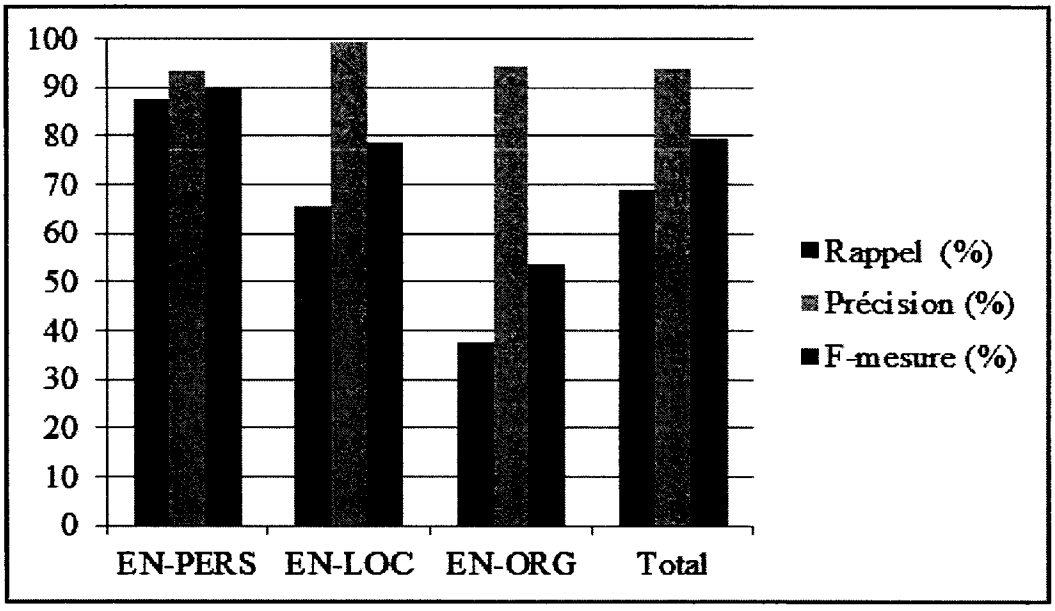
Tableau 5.4 Précision, rappel et F-mesure pour les ENs du corpus UN

	EN-PERS	EN-LOC	EN-ORG	Total
Nombre d'ENs dans le corpus	242,799	21,398	141,559	405,756
Nombre d'ENs bien translitérées	212,073	13,987	53,336	279,396
Nombre d'ENs mal translitérées	15,226	88	3300	18614
Rappel (%)	87.34	65.36	37.67	68.85
Précision (%)	93.30	99.37	94.17	93.75
F-mesure (%)	90.22	78.85	53.81	79.39

Discussion des résultats

En nous basant sur les résultats illustrés dans le tableau 5.4 et la figure 5.1, nous discutons les résultats obtenus à partir du corpus UN pour chaque type d'EN.

Figure 5.1 Rappel, précision et F-mesure pour les ENs du corpus UN



Pour les EN-PERS, la valeur du rappel trouvée est 87.34%. Cette valeur est la meilleure par rapport aux autres valeurs du rappel pour les deux autres types d'EN. Cela montre que notre méthode de translittération réussit à trouver plus d'EN-PERS. En effet, la prononciation des noms de personne est généralement la même pour n'importe quelle langue, et pour trouver l'équivalent d'un nom de personne écrit en une langue source, une translittération de ce nom de la langue source vers la langue cible donne généralement la bonne traduction. Dans notre cas, le corpus UN contient beaucoup d'EN-PERS de différentes origines et la majorité de ces ENs ont été correctement translittérées (précision de 93.3%). Donc on peut dire que notre méthode de translittération est efficace, peu importe l'origine du nom de la personne. Par exemple, le nom 'Carl August Fleischhauer' est translittéré en arabe à 'كارل أوغست فلايشاور' (kArl >wgst flAy$Awr) et le nom d'origine arabe 'Mohamed Salah Dembri' est translittéré en arabe à 'محمد صلاح دمبري' (mHmd SlAH dmbry).

Cependant, il y a un nombre d'EN-PERS qui sont mal translittérées. Cela revient à l'écriture différente de ces ENs en arabe telle qu'on peut avoir des phonèmes en plus ou en moins. Par exemple, le nom de personne 'Mrs. Kalajdzisalihovi' doit être traduit en arabe par le nom 'كالاجساليهوفيتش' (kAلاjsAlyhwfyt$) qui a deux phonèmes de plus à la fin 'تش' (t$). Une solution pour ce problème est l'élargissement du nombre de translittérations des phonèmes anglais vers l'arabe pour avoir plus de chance de trouver la meilleure translittération.

En comparant les valeurs de précision et du rappel pour les EN-PERS avec la littérature, nous pouvons dire que notre méthode de translittération est idéale pour les EN-PERS. Par exemple, dans le projet JRC-Names (Pouliquen et al., 2005) (pour la langue arabe) la précision obtenue est 89.3% et le rappel obtenu est rappel = 83.3%.

Pour les EN-LOC, la valeur du rappel trouvée est 65,36%, car il y a beaucoup des EN-LOC qui ne peuvent pas être translittérées et donc il faut les traduire. Cependant, la précision obtenue pour ce type d'EN est très élevée (99.37%) ce qui montre la fiabilité de notre méthode de translittération pour les EN-LOC.

Pour les EN-ORG, la valeur du rappel trouvée est la plus faible par rapport aux autres types d'EN, et elle a influé sur les résultats globaux. En effet, les EN-ORG sont généralement composées de plusieurs termes ou elles sont sous forme d'acronymes, ce qui rend leur translittération plus difficile. Une solution pour ce type d'EN est la traduction au lieu de la translittération. La plupart des EN-ORG reconnues ont été correctement translitérées (précision de 94.17%). Donc le dictionnaire des marqueurs lexicaux Dic_Marqueur_Org a joué un rôle important dans la procédure de translittération des EN-ORG.

5.2.2 Évaluation pour les titres Wikipédia

Résultats obtenus

L'évaluation des 13000 titres de Wikipédia nous a donné les résultats illustrés dans le tableau 5.5.

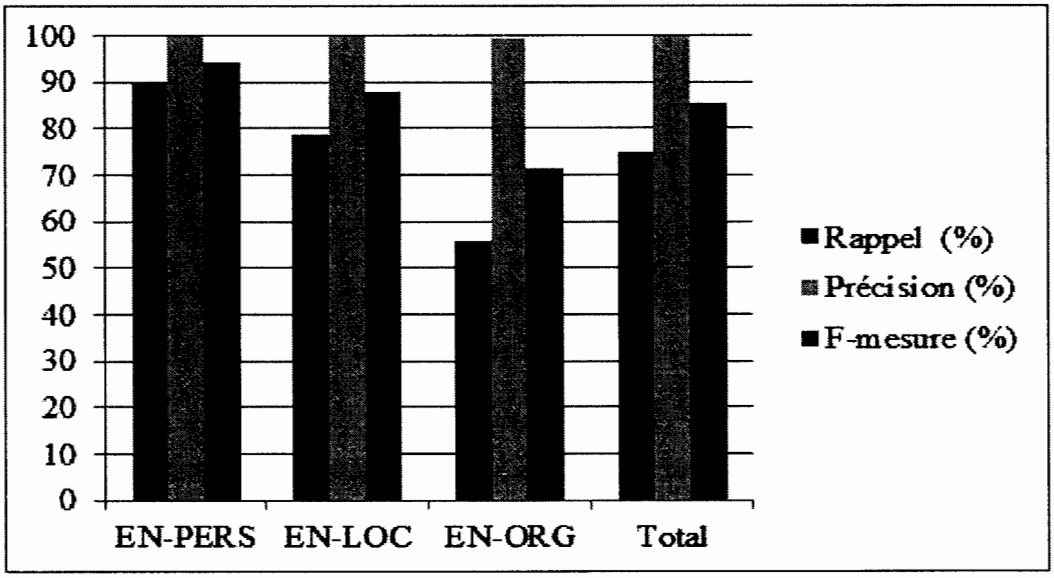
Tableau 5.5 Précision, rappel et F-mesure pour les ENs de titres de Wikipédia

	EN-PERS	**EN-LOC**	**EN-ORG**	**Total**
Nombre d'EN dans l'échantillon	3033	6894	3795	13,722
Nombre d'ENs bien translitérées	2719	5424	2118	10,261
Nombre d'ENs mal translitérées	5	4	15	24
Rappel (%)	89.64	78.67	55.81	74.77
Précision (%)	99.81	99.92	99.29	99.76
F-mesure (%)	94.45	88.03	71.45	85.47

Discussion des résultats

En nous basant sur les résultats illustrés dans le tableau 5.5 et la figure 5.2, nous discutons les résultats obtenus à partir des titres Wikipédia pour chaque type d'EN.

Figure 5.2 Rappel, précision et F-mesure pour les ENs des titres de Wikipédia



La valeur de la précision est de plus de 99% pour les trois types d'EN, ce qui montre que les ENs ont été correctement translittérées. La valeur pour les EN-ORG est la plus basse par rapport aux autres types d'EN, mais elle est mieux que la valeur du rappel obtenue pour les organisations dans le corpus UN. Cela est dû aux phrases courtes des titres de Wikipédia alors que les phrases du corpus UN sont plus longues. La probabilité de trouver une EN dans un titre Wikipédia est égal à 1 la plupart des temps.

5.3 Deuxième évaluation – comparaison par rapport à Google Translate

Dans cette section, nous présentons une comparaison de notre méthode de translittération par rapport au système de traduction Google Translate[72].

Dans cette évaluation, nous considérons un échantillon du corpus UN de 278410 phrases et un échantillon de titres Wikipédia de 13000 phrases. Ensuite, nous comparons les lexiques bilingues d'ENs construits selon notre méthode de

[72] Google translate daté de juin 2014

translittération avec les lexiques construits par la traduction des ENs en utilisant le traducteur Google Translate.

5.3.1 Évaluation pour l'échantillon du corpus UN

Les résultats obtenus

Les tableaux 5.6 et 5.7 illustrent les résultats obtenus par notre méthode de translittération et par Google Translate respectivement.

Tableau 5.6 Résultats obtenus par notre méthode de translittération

	EN-PERS	EN-LOC	EN-ORG	Total
Nombre d'ENs dans l'échantillon	17,553	3359	20,569	41,481
Nombre d'ENs bien translitérées	15,069	2187	5613	22,869
Nombre d'ENs mal translitérées	524	33	445	1002
Rappel (%)	85.84	65.1	27.28	55.13
Précision (%)	96.63	98.51	92.65	95.8
F-mesure (%)	90.91	78.39	42.14	69.98

Tableau 5.7 Résultats obtenus par Google Translate

	EN-PERS	EN-LOC	EN-ORG	Total
Nombre d'ENs dans l'échantillon	17,553	3359	20,569	41,481
Nombre d'ENs bien traduites	16,273	2981	19,966	39,220
Nombre d'ENs mal traduites	764	80	501	1345
Rappel (%)	92.70	88.74	97.06	94.54
Précision (%)	95.51	97.38	97.55	96.68
F-mesure (%)	94.08	92.85	97.3	95.6

Discussion des résultats obtenus

Pour les EN-PERS, Google Translate a obtenu un rappel et une F-mesure meilleurs que ceux de notre méthode de translittération. Cependant, nous avons obtenu une valeur de

précision (96.63) légèrement mieux que celle de Google Translate (95.51%). De ces résultats, on peut déduire que l'ajout d'une technique de translittération à un système de TA peut améliorer la traduction, ceci est expliqué par plusieurs raisons comme par exemples :

1- Les EN-PERS peuvent avoir plusieurs écritures différentes, et il arrive que la mémoire de traduction de Google Translate ne contienne pas toutes ces écritures, et donc soit il ne peut pas traduire certaines EN-PERS ou il donne une traduction fausse. Par exemple, le nom de personne 'Hussin' a plusieurs façons d'écriture comme 'Hussayn', 'Husayn', 'Hussein', 'Hussiayn', 'husyn', Husiayn', .etc. qui représentent une traduction pour le nom arabe 'حسين'. Mais Google Translate ne trouve pas la traduction pour les noms 'husyn', 'Hussiayn' et 'Husiayn' par contre la translittération de ces noms donne la traduction adéquate.

2- Les noms de personnes sont parfois mal écrits dans un texte (ou corpus). Dans ces cas, Google Translate peut donner une traduction erronée. Par exemple, nous avons trouvé que le nom 'Naguib Sawiris' est écrit dans le corpus UN comme 'Naguib Saweiras'. Google Translate donne une traduction partielle pour le nom 'Naguib Saweiras», car il ne trouve pas la traduction pour le mot 'Saweiras', mais il donne la bonne traduction pour le mot 'Sawiris' qui est 'ساويرس'. Cependant, une translittération donne la traduction adéquate pour les deux écritures. Dans notre évaluation, il y a beaucoup d'ENs qui sont partiellement traduites par Google Translate.

3- Il y a des EN-PERS qui contiennent des termes qui ont un sens particulier comme le nom 'Patrick Sale' qui est traduit par Google Translate à 'باتريك بيع' (bAtryk byE / Patrick vente) qui est fausse, car dans ce cas le mot 'Sale' ne désigne pas la 'vente' et il doit être translittéré à 'سيل' (syl).

Le tableau 5.8 montre quelques exemples d'EN-PERS qui n'ont pas bien traduites par Google Translate, mais bien translittérées par notre méthode.

Tableau 5.8 Exemples des EN-PERS (du corpus UN) non traduites par Google Translate

EN-PERS	Google Translate	Notre méthode	Remarques
Abdel-Ellah Balqzeez	عبد اللهBalqzeez	عبد الإله بلقزيز (Ebd Al<lh blqzyz)	Le terme 'Balqzeez' est introuvable par Google Translate.
Miryam Husayn Fahdi	مريم حسين Fahdi	مريم حسين فهدي (mrym Hsyn fhdy)	Le terme 'Fahdi' est introuvable par Google Translate.
Mrs. Hawa Ahmed Youssof	السيدة حواء أحمد Youssof	السيدة حوا أحمد يوسف (Alsydp HwA >Hmd ywsf)	Le terme 'Youssof' est introuvable par Google Translate.

Comme pour les EN-PERS, la précision pour les EN-LOC (98.51%) est un peu mieux que celui de Google Translate (97.38%) ce qui montre l'efficacité de la translittération pour ce type d'EN aussi. Nous avons trouvé qu'il y a quelques EN-LOC qui sont correctement translittérées par notre méthode par contre elles ne sont pas traduites par Google Translate ou elles sont partiellement traduites. Le tableau 5.9 montre quelques exemples de ces noms. On a remarqué que Google Translate ne donne pas la bonne traduction s'il s'agit des EN-LOC qui contiennent des noms d'origine arabe qui nécessitent généralement une translittération au lieu de traduction.

Tableau 5.9 Exemples d'EN-LOC mal ou non traduite par Google Translate

EN	Google Translate	Notre méthode	Remarques
Al-genayen quarter	حي genayen	حي الجنابن (Hy AljnAyn)	La traduction du terme 'genayen' est introuvable par Google Translate
Qota	Non traduit	قوته (qwth)	La traduction du terme 'Qota' est introuvable par Google Translate
Estadio Omnilife	استاد Omnilife	ملعب أومنيلايف (mlEb >wmnylAyf)	La traduction du terme 'Omnilife' est introuvable par Google Translate
Zahmin	Zahmin	زحمين (zHmyn)	La traduction du terme 'Zahmin' est introuvable par Google Translate

Pour les EN-ORG, malgré la valeur élevée de notre précision, nos résultats ne sont pas compétitifs par rapport à ceux de Google Translate. Ceci, car, notre rappel est très faible par rapport à celui de Google Translate. Cela montre qu'une traduction suffit pour ce type d'EN.

5.3.2 Évaluation pour l'échantillon de titres de Wikipédia

Résultats obtenus

Les tableaux 5.10 et 5.11 illustrent les résultats obtenus par notre méthode de translittération et par Google Translate respectivement pour l'échantillon du corpus des titres de Wikipédia.

Tableau 5.10 Résultats obtenus par notre méthode de translittération

	EN-PERS	**EN-LOC**	**EN-ORG**	**Total**
Nombre d'ENs dans l'échantillon	3033	6894	3795	13,722
Nombre d'ENs bien translitérées	2719	5424	2118	10,261
Nombre d'ENs mal translitérées	5	4	15	24
Rappel (%)	89,64	78.67	55.81	74.77
Précision (%)	99.81	99.92	99.29	99.76
F-mesure (%)	94.45	88.03	71.45	85.47

Tableau 5.11 Résultats obtenus par Google Translate

	EN-PERS	**EN-LOC**	**EN-ORG**	**Total**
Nombre d'ENs dans l'échantillon	3033	6894	3795	13,722
Nombre d'ENs bien traduites	2489	6793	3601	12,883
Nombre d'ENs mal traduites	510	13	160	683
Rappel (%)	82.06	98.53	94.88	93.88
Précision (%)	82.99	99.80	95.74	94.96
F-mesure (%)	82.52	99.16	95.30	94.41

Discussion des résultats

En nous basant sur les résultats illustrés dans les tableaux 5.10 et 5.11, nous discutons les résultats pour chaque type d'EN.

Pour les EN-PERS, le rappel, la précision et la F-mesure de notre méthode sont meilleurs que ceux de Google Translate. Donc notre méthode est performante pour les EN-PERS notamment si les phrases du corpus parallèle utilisé sont courtes comme le cas des titres de Wikipédia. Concernant Google Translate, il y a quelques EN-PERS qui ne sont pas traduites ou elles sont mal traduites. Le tableau 5.12 montre quelques exemples.

Tableau 5.12 Exemples des EN-PERS non traduites par Google Translate

EN-PERS	**Google Translate**	**Notre méthode**	**Remarques**
John Maffey	جون Maffey	جون مفي	Le terme 'Maffey' n'est pas traduit par Google Translate.
Jose Chinantequilla	جوس Chinantequilla	جوس شينانتيكيلا	La traduction de terme 'Chinantequilla' est introuvable par Google Translate.

Pour les EN-LOC, le rappel et la F-mesure de Google Translate sont meilleurs que nos résultats. Notre précision (99.92%) est un peu plus élevée à celle de Google Translate (99.80%). Il y quelques EN-LOC qui contiennent des termes non traduits par Google Translate comme par exemple, l'EN-LOC 'San Servan' qui est traduite en 'سان Servan' par contre la translittération donne 'سان سيرفان' (sAn syrfAn).

Pour les EN-ORG, notre méthode de translittération a obtenu ses plus mauvaises performances par rapport aux autres types d'EN. Cependant, les résultats obtenus par les titres de Wikipédia sont mieux par rapport aux résultats obtenus par le corpus UN.

Pour cette évaluation, nous pouvons constater que le type, en terme de longueur de phrases, du corpus utilisé influe sur les résultats de notre méthode de translittération. Les bons résultats étaient pour le corpus de titres Wikipédia qui a de courtes phrases.

5.4 Troisième évaluation – évaluation de lexique obtenu à partir de Wikipédia

La troisième évaluation est simple, elle consiste à comparer le lexique des noms de personnes obtenu par les titres de Wikipédia avec les deux lexiques décrits dans (Mohit et al., 2012), (Azab et al., 2013)[73] et (Alotaibi et Lee, 2013)[74]. Ces lexiques sont en anglais et en arabe et ils ont été construits à partir de Wikipédia. Le lexique de Mohit et al. est construit par un système d'extraction des ENs à partir de Wikipédia en utilisant l'approche basée sur l'apprentissage machine. Le lexique de Alotaibi et Lee est construit par une méthode de classification des articles de Wikipédia selon des classes prédéfinies d'EN.

Tableau 5.13 Comparaison de lexique des EN-PERS de titres Wikipédia avec d'autres lexiques

Nombre d'ENs de notre lexique sans doublons	**Nombre d'ENs qui existent dans notre lexique et dans les lexiques de Mohit et Alotaibi**	**Nombre d'ENs qui existent dans notre lexique et n'existent pas dans les lexiques de Mohit et Alotaibi**
9973	1723 (17.28%)	8250 (82.72%)

Les résultats illustrés dans le tableau 5.13 montrent que notre lexique d'EN-PERS extrait de Wikipédia est un complément aux deux lexiques de Mohit et Alotaibi avec 8250 nouvelles ENs. Cela parce que Wikipédia est une ressource très dynamique et en croissance rapide et des articles sont souvent ajoutés dès leur survenance (Bunescu et Pasca, 2006). Donc le nombre des ENs qui existent dans Wikipédia augmente au fur et à mesure que de nouvelles pages sont ajoutées.

[73] http://nlp.qatar.cmu.edu/resources/NETLexicon/ et http://www.ark.cs.cmu.edu/ArabicNER/

[74] http://www.cs.bham.ac.uk/~fsa081

5.5 Quatrième évaluation : Intégration des lexiques construits dans un système de TAS

Les ENs et les mots MHV composent un problème pour la TA et leur traduction peut donner un mauvais sens de la phrase. Par exemple, le nom de personne d'origine arabe 'أمل' (Amal) a le sens 'espoir' en français; Un système de TA peut donner la traduction 'espoir' à ce nom de personne qui n'est pas la bonne traduction. Une translittération de ce type de termes résout ce problème.

Pour évaluer la qualité de notre méthode de translittération, nous avons incorporé les deux lexiques d'EN-PERS et d'EN-LOC dans un système de TAS en utilisant le décodeur Moses.

Pour avoir des lexiques d'ENs plus riches, nous avons ajouté à ces deux lexiques d'ENs deux autres lexiques bilingues d'EN-PERS et d'EN-LOC qui sont construits à partir de JRC-Names, (Steinberger et al., 2011), DBPedia (Lehmann et al., 2012) et Geonames[75] .

Nous avons participé dans la campagne d'évaluation TRAD 2014 avec le système de TAS construit en utilisant les lexiques bilingues d'ENs.

Avant de présenter les résultats de la quatrième évaluation et la discussion, nous décrivons la méthode de construction des lexiques bilingues d'ENs à partir des ressources linguistiques.

5.5.1 Construction de lexiques d'ENs à partir de ressources linguistiques

Cette sous-section décrit la méthode suivie pour la construction de deux lexiques d'EN-PERS et d'EN-LOC à partir de trois ressources linguistiques qui sont JRC-Names

[75] http://www.geonames.org/

(Steinberger et al., 2011), DBPedia (Lehmann et al., 2012) et Geonames[76]. Une brève description de ces ressources est présentée aussi.

Geonames

Geonames est une ressource qui contient des EN-LOC. C'est une base de données géographique de la toponymie qui couvre tous les pays et compte plus de huit millions de noms de lieux. Geonames est disponible gratuitement sur Internet, et présente les données dans plusieurs langues, mais il y a plusieurs ENs qui ne sont disponibles qu'en anglais.
Nous avons utilisé les API de Geonames pour extraire les ENs.

JRC-Names

JRC-Names (Steinberger et al., 2011) est une ressource multilingue des ENs de type nom de personne et noms des organisations. C'est un outil disponible gratuitement suivant le lien http://langtech.jrc.ec.europa.eu/JRC-Names.html

Pour extraire les ENs à partir de JRC-Names, nous avons utilisé le code source de JRC-names (JRC-Names Java source code) disponible en ligne.

DBPedia

Le projet DBPedia[77] a été lancé par l'Université Libre de Berlin et l'Université de Leipzig, en collaboration avec l'entreprise OpenLink Software. Le premier ensemble de données accessibles au public a été publié en 2007[78]. Ensuite, DBPedia est devenu un outil pour la construction des dictionnaires d'ENs ou de termes généraux, par exemple (Al-Jumaily et al., 2012) ont utilisé DBPedia comme ressource dans leur système d'extraction des ENs en langue arabe.

---

[76] http://www.geonames.org/

[77] http://dbpedia.org/About

[78] http://fr.wikipedia.org/wiki/DBpedia

Le principe de ce DBPedia est l'extraction des informations structurées à partir de Wikipédia. Le contenu des articles de Wikipédia est sous un format général et non structuré. DBPedia extrait les informations à partir de ces articles, mais sous une forme structurée et normalisée dans un format du web sémantique (Lehmann et al., 2012; Mendes et al., 2012).

L'utilisation de DBPedia se fait par les graphes RDF (Manola et al., 2004) et les requêtes SPARQL[79]. Un graphe RDF (Manola et al., 2004) est un modèle de graphe servant à formuler les ressources du Web. Il est particulièrement adapté à représenter des métadonnées sur les ressources du Web, comme le titre, l'auteur, la date de modification d'une page Web, etc.
Nous avons interrogé les graphes RDF de DBPedia par le langage SPARQL (Protocol and RDF Query Language) (Prud'hommeaux et Seaborne, 2008). SPARQL est un langage de requêtes qui permet la mise à jour (recherche, ajout, modification et suppression) des données RDF disponibles à travers Internet.

Volume des lexiques construits à partir des ressources linguistiques

Le tableau ci-dessous illustre le nombre d'EN extrait à partir de chaque ressource linguistique utilisée.

Tableau 5.14 Taille du lexique des EN-PERS et EN-LOC construits à partir de ressources linguistiques

	Geonames	**JRC-Names**	**DBPedia**	**Total**
EN-PERS	-	9159	41,956	51,115
EN-LOC	2142	-	-	2142

[79] http://www.w3.org/TR/rdf-sparql-query/

5.5.2 Intégration des lexiques bilingues d'ENs dans le système de TAS

- **Résultats obtenus**

Les évaluations ont été faites par le calcul de la valeur de score BLEU (Papineni et al., 2002) et le taux des mots MHV.

Le score BLEU obtenu dans les deux configurations[80] de notre système de TAS est illustré dans le tableau 5.15. Le taux des MHV et le nombre des mots non reconnus sont montrés dans le tableau 5.16.

Tableau 5.15 Score BLEU dans chaque évaluation

Expérience	**Score BLEU**
Système de base (système 1)	24.37%,
Système à base de lexiques d'ENs (système 2)	24.92%

Tableau 5.16 Taux des MHV dans chaque évaluation

Expérience	**Pourcentage de MHV**	**Nombre de mots non reconnus**
Système de base (système 1)	2.99%	314
Système à base de lexiques d'ENs (système 2)	2.63%	276

- **Discussion des résultats**

Nous constatons que l'introduction des lexiques d'ENs a amélioré la valeur de score du système de base. Avec ces lexiques d'ENs, le système de TAS n'a pas besoin de traduire les ENs contenues dans les lexiques d'ENs, ce qui évite une mauvaise traduction de ces ENs. Cela confirme que la translittération des ENs joue un rôle important dans la TAS.

[80] Ces configurations ont été présentées dans la section 4.3.3

Dans le tableau 5.16, l'introduction des lexiques d'EN au système 2 diminue le taux des mots MHV (de 2.99% à 2.63%) et donc le nombre des mots non reconnus (de 314 mots à 276 mots). Les mots MHV reconnus par le système 1 existent dans l'un des lexiques introduits, car beaucoup d'ENs correspondent à des MHV et les lexiques introduits ont permis de les reconnaître et ainsi d'utiliser leurs traductions dans le système de TAS.

Exemples :

L'EN-PERS 'المختار' (AlmxtAr) a été traduite par le système de base (système 1) en 'choisis' qui est une fausse traduction, mais cette EN a été bien reconnue dans le système à base de lexiques d'ENs (système 2) et elle est traduite en 'Al-Mokhtar'. Une des phrases en arabe qui contiennent cette EN est :

'كان 'المختار' قد دشّن رحلة فرار دراماتيكيّة' (AlmxtAr" qd d$~n rHlp frAr drAmAtyky~p)

La traduction obtenue pour cette phrase est : 'Al-Mokhtar' a lancé le voyage dramatique'.

Parmi les mots MHV reconnus après l'introduction de lexiques d'ENs le mot 'أميمة' (>mymp) qui est une EN-PERS et a été reconnue à 'Omima'.

5.6 Conclusion

Dans ce chapitre, nous avons présenté quatre méthodes différentes pour évaluer les résultats obtenus dans ce mémoire. Les trois premières évaluations étaient destinées à évaluer les résultats obtenus par notre méthode de translittération des ENs de l'anglais vers l'arabe.

La première méthode consiste à évaluer les lexiques d'ENs construits par la projection cross-linguistique. Les corpus de test sont le corpus UN et le corpus de titres de Wikipédia. Les résultats de cette évaluation ont montré la performance de la méthode de translittération proposée notamment pour les EN-PERS et les EN-LOC.

Une comparaison de notre méthode de translittération des ENs a été faite avec la traduction du système Google Translate. Notre méthode a obtenu une précision meilleure que celle de Google Translate pour les noms de personnes, mais un rappel inférieur de celui de Google Translate. Ce qui montre que l'intégration d'un module de translittération dans un traducteur automatique est très efficace pour la TA.

La troisième évaluation montre que le lexique obtenu à partir de Wikipédia à compléter d'autres lexiques d'ENs extraites de Wikipédia.

La dernière évaluation montre que le problème des mots MHV et des ENs dans la TA peut être résolu par l'incorporation de lexiques d'ENs dans le système de TAS. Une amélioration en terme de score BLEU a été réalisée par notre système de TAS en introduisant les lexiques bilingues d'ENs. Cette dernière évaluation a été présentée lors des évènements internationaux TRAD 2014[81].

[81] http://www.trad-campaign.org/

CHAPITRE VI

CONCLUSION ET PERSPECTIVES

Ce mémoire consiste en premier à développer une méthode de projection cross-linguistique pour l'extraction des ENs à partir de corpus parallèles alignés, et la construction de lexiques bilingues d'ENs pour deux langues différentes source et cible. En deuxième, ces lexiques ont été incorporés dans un système de TAS.

Dans la première partie de ce mémoire, nous avons choisi la langue anglaise comme langue source et la langue arabe comme langue cible. La langue arabe appartient aux langues à morphologie complexe et la reconnaissance des ENs en cette langue reste un défi dans le domaine du TALN. L'objectif principal visé par ce projet est la contribution pour soulever ce défi. Cet objectif est réalisé par le développement d'une méthode de projection cross-linguistique pour l'extraction des ENs en arabe en se basant sur les corpus parallèles. La méthode de projection présentée se base sur un modèle de translittération des ENs d'une langue source vers une langue cible. Ce modèle consiste à trouver toutes les translittérations possibles d'un phonème en langue source vers la langue cible. Une méthode de normalisation basée sur des règles a été introduite pour normaliser quelques phonèmes arabes vers un seul phonème.

Notre méthode de translittération peut être appliquée à n'importe quelle paire de langues en changeant juste les phonèmes de translittération et les règles de la méthode de normalisation des phonèmes.

Le résultat de notre méthode de translittération était l'annotation du corpus cible et la construction de trois lexiques d'ENs pour les types : noms de personne, noms de lieu et noms d'organisation.

Pour tester la performance de notre méthode de translittération, nous avons utilisé deux corpus parallèles complètement différents. Le premier corpus est le corpus UN caractérisé par ses phrases longues. Le deuxième corpus, extrait de Wikipédia, constitue en un ensemble de titres parallèles des articles de Wikipédia. Contrairement au corpus UN, le corpus de titres de Wikipédia est caractérisé par ses phrases courtes. Ces deux corpus sont riches en terme d'EN ce qui donne un bon exemple pour le corpus test. Plusieurs évaluations ont été faites sur les résultats obtenus par l'application de notre méthode de translittération sur ces deux corpus. Parmi ces évaluations, une évaluation a été faite pour comparer notre modèle de translittération des ENs avec la traduction des ENs par le système Google Translate et une autre évaluation a été faite pour tester l'efficacité de l'incorporation des lexiques d'ENs dans les systèmes de TAS.

Ces deux évaluations ont montré que la translittération des ENs est très utile pour la TA, on arrive à trouver qu'il y a des noms de personne ou des noms de lieu qui ne peuvent pas être traduits par un traducteur automatique, comme Google Translate. Par contre, une translittération de ces noms donne la traduction adéquate. Aussi pour le problème des MHV dans les systèmes de TA, nos résultats d'évaluation ont montré une diminution de taux des MHV après l'utilisation des lexiques d'ENs.

Notre première évaluation qui a été faite par un expert linguistique a obtenu une bonne précision pour les EN-PERS et EN-LOC. Cela explique l'efficacité de notre méthode de translittération des ENs pour ces deux types d'EN.

Les principales contributions de ce travail peuvent se résumer à travers les points suivants :

- Réalisation d'une méthode de translittération des ENs de l'anglais vers l'arabe qui peut être généralisée à d'autres paires de langues.
- Construction de lexiques bilingues d'ENs pour deux langues différentes.
- Annotation de corpus en arabe avec les ENs reconnues.
- Amélioration d'un système de TAS, en utilisant les lexiques bilingues d'ENs.

Concernant les perspectives, nous souhaitons dans le futur l'amélioration de notre solution pour obtenir de meilleurs résultats pour les EN-ORG. En effet, ce type d'EN se présente généralement sous forme d'acronymes et ne peuvent être translittérés, mais doivent plutôt être traduits en utilisant une méthode spécifique à ce genre de problèmes.

Nous souhaitons aussi généraliser notre solution pour être appliquée sur des corpus comparables qui sont plus disponibles par rapport aux corpus parallèles.

PUBLICATIONS

1. Fatima DEFFAF, Fatiha SADAT. Construction automatique de lexiques bilingues d'entités nommées à partir du corpus parallèle (anglais-arabe). In proceedings of ACFAS 2014, Montréal, QC, Canada. May 12-16, 2014.

2. Sellami Rahma, **Deffaf Fatima**, Sadat Fatiha and Belguith Hadrich Lamia. Improved Statistical Machine Translation by Cross-Linguistic Projection of Named Entities Recognition and Translation. In Proceedings of the 16th International Conference on Intelligent Text Processing and Computational Linguistic **Cicling 2015**, Cairo, Egypt, Apr. 14-20, 2015.

APPENDICE A

TABLE DE TRANSLITTÉRATION DE L'ARABE D'APRES LA NORME DE BUCKWALTER

Tableau A.1 Code Unicode et translittération des phonèmes arabes par le système Buckwalter[82]

Code Unicode	Buckwalter	Glyph		Code Unicode	Buckwalter	Glyph
\u0621	'	ء		\u062E	X	خ
\u0622	|	آ		\u062F	D	د
\u0623	>	أ		\u0630	*	ذ
\u0624	&	ؤ		\u0631	R	ر
\u0625	<	إ		\u0632	Z	ز
\u0626	}	ئ		\u0633	s	س
\u0627	A	ا		\u0634	$	ش
\u0628	B	ب		\u0635	S	ص
\u0629	P	ة		\u0636	D	ض
\u062A	T	ت		\u0637	T	ط
\u062B	V	ث		\u0638	Z	ظ
\u062C	J	ج		\u0639	E	ع
\u062D	H	ح		\u063A	G	غ
\u0640	_	ـ		\u0641	F	ف

[82] http://languagelog.ldc.upenn.edu/myl/ldc/morph/buckwalter.html

Code Unicode	Buckwalter	Glyph
\u0642	Q	ق
\u0643	K	ك
\u0644	L	ل
\u0645	M	م
\u0646	N	ن
\u0647	H	ه
\u0648	W	و
\u0649	Y	ى
\u064A	Y	ي
\u064B	F	ً
\u064C	N	ٌ
\u064D	K	ٍ
\u064E	A	َ
\u064F	U	ُ
\u0650	I	ِ
\u0651	~	ّ
\u0652	O	ْ

APPENDICE B

ÉCHANTILLON DE CORPUS PARALLÈLES ANNOTÉS

B.1 Échantillon du corpus UN

B.1.1 Échantillon en langue arabe

وأود أن انتهز هذه الفرصة أيضا لأشكر سلفي، <PERS>السيد ريكاردو لاغوريو</PERS> من <LOC>الأرجنتين</LOC>، الذي تمكن من تهيئة جو من التعاون ساعدنا على الاحتفاظ بتوافق الآراء حول القضايا الرئيسية المطروحة على هذه اللجنة وعلى توسيع نطاقه، الأمر الذي اعتبره واحدا من أكبر منجزاتنا على مدى السنوات القليلة الماضية.

ولقد عقدت النية على أن أواصل السير على هذا الطريق الذي شقه لنا.

واسمحوا لي أيضا أن أعبر عن مدى غبطتي إذ أجد معي أعضاء مكتب موقرين مثل <PERS>السيد كونيك</PERS> من <LOC>بولندا</LOC>، و<PERS>السيد نييتو</PERS> من <LOC>الأرجنتين</LOC>، والسيد تشوكوي من <LOC>كينيا</LOC> بوصفهم نواب الرئيس، وكذلك <PERS>الدكتور هولوهان</ PERS> من <LOC>ايرلندا</LOC> بوصفه المقرر.

وإنني لعلى ثقة من أنه سيكون بوسعنا، بالاستفادة من معرفتهم المتعمقة بالقضايا المطروحة على اللجنة، أن نقطع شوطا بعيدا صوب بلوغ الأهداف التي علينا بلوغها في هذه الدورة.

وأود كذلك أن أرحب بالسيد <PERS>ماركو فيانيللو كيودو</PERS>، المساعد الجديد للأمين العام لشؤون الاعلام، الذي أثار اعجابنا بالفعل بتفانيه وحكمته وطاقته، وكذلك بمهاراته الشخصية والمهنية.

وإننا نعرب للسيد <PERS>فيانيللو كيودو</PERS> ولموظفي إدارته عن رغبتنا في مواصلة التعاون الوثيق والمثمر.

وأخيرا لا آخرا، أود أن أرحب بوفدي جمهورية <LOC>كوريا</LOC> و<LOC>السنغال</LOC>، وهما أجد أعضاء لجنتنا.

ولا شك في أن مشاركتهما ستعزز موقف هذه اللجنة وستزيد من تنوع عضويتها.إننا نعيش فترة من التغيرات المثيرة والبالغة الأهمية؛ فترة مليئة أيضا بآمال كبار معلقة على <ORG>الأمم المتحدة</ORG>.

وفي مسألة تتصل بذلك، تجدر أيضا ملاحظة أن<ORG> مكتبة داغ همرشولد</ORG> قد أعدت مؤخرا نسخة مستوفاة من قائمة مراجعها السابقة المتعلقة بقضية <LOC>فلسطين</LOC>، والتي نشرت أصلا في عام 1976.

B.1.2 Échantillon en anglais

I would also like to take this opportunity to thank my predecessor, <PERSON>Mr. Ricardo Lagorio</PERSON> of <LOCATION>Argentina</LOCATION>, who was able to develop an atmosphere of cooperation that helped us to maintain and broaden the consensus on major issues before this Committee that I consider as one of our major achievements over the past few years.

It is my intention to continue on the path that he blazed.

Permit me also to say how delighted I am to have such distinguished fellow officers as <PERSON>Mr. Konik</PERSON> of <LOCATION>Poland</LOCATION>, <PERSON>Mr. Nieto</PERSON> of <LOCATION>Argentina</LOCATION> and <PERSON>Mr. Chokwe</PERSON> of <LOCATION>Kenya</LOCATION> as Vice-Chairmen, as well as <PERSON>Dr. Holohan</PERSON> of <LOCATION>Ireland</LOCATION> as the Rapporteur.

Benefiting from their profound knowledge of the issues before this <ORGANIZATION>Committee</ORGANIZATION>, I am sure that together we shall be able to go a long way towards meeting the objectives before us at this session.

I would also like to welcome <PERSON>Mr. Marco Vianello Chiodo</PERSON>, the new Assistant Secretary-General for Public Information, who has already impressed us with his dedication, wisdom and energy, as well as with his personal and professional skills.

To <PERSON>Mr. Vianello Chiodo</PERSON> and to the staff of the Department we extend our wishes for our continued close and fruitful cooperation.

Last but not least, I would also like to welcome the delegations of the Republic of <LOCATION>Korea</LOCATION> and <LOCATION>Senegal</LOCATION>, the newest members of our Committee.

Their participation certainly further enhances the standing of this Committee and broadens the diversity of its membership. We live in a period of exciting and momentous changes; a period also filled with great expectations placed on the <ORGANIZATION>United Nations</ORGANIZATION>.

On a related matter, it should also be noted that the <ORGANIZATION>Dag Hammarskjold Library</ORGANIZATION>, in response to a request made by the <ORGANIZATION>General Assembly</ORGANIZATION> at its forty-sixth session, has recently prepared an update to its earlier bibliography on the question of <LOCATION>Palestine</LOCATION>, originally published in 1976.

B.2 Échantillon de titres de Wikipédia

B.2.3 Échanillon en langue arabe

<PERS>عبد الحميد بن باديس</PERS>
الدولة العثمانية
مرينيون
قالب:نقطة النطق
ويكيبيديا:استكشاف
ويكيبيديا:انتخابات مجلس إدارة مؤسسة <ORG>ويكيميديا</ORG>
ملحق:قائمة أعلام الموسيقى الكلاسيكية
القدس
شكيم
تصنيف:بحيرات
إيطاليا
تونس
تاريخ <LOC>فلسطين</LOC>
تصنيف:رؤساء <LOC>مصر</LOC>
<LOC>السودان</LOC>
<LOC>بريطانيا</LOC> (توضيح)
أدولف هتلر

<LOC/>غينيا<LOC>
<LOC/>الصومال<LOC>
قالب:أوبك
ملحق:قائمة الأحزاب السياسية في <LOC/>إسرائيل<LOC>
<PERS/>عبد العزيز بن باز<PERS>

B.2.4 Échanillon en langue anglais

<PERSON>Abdelhamid Ben Badis</PERSON>
Ottoman Empire
Marinid dynasty
Template:Place of articulation
Wikipedia:Explore
Wikipedia:Elections for the <ORGANIZATION>Board of Trustees</ORGANIZATION> of the <ORGANIZATION>Wikimedia Foundation</ORGANIZATION>
List of classical music composers by era
Jerusalem
<LOCATION>Shechem</LOCATION>
Category:Lakes
<LOCATION>Italy</LOCATION>
<LOCATION>Tunisia</LOCATION>
History of <LOCATION>Palestine</LOCATION>
Category:Presidents of <LOCATION>Egypt</LOCATION>
<LOCATION>Sudan</LOCATION>
<LOCATION>Britain</LOCATION>
Adolf Hitler
<LOCATION>Guinea</LOCATION>
<LOCATION>Somalia</LOCATION>
Template:<ORGANIZATION>OPEC</ORGANIZATION>
List of political parties in <LOCATION>Israel</LOCATION>
<PERSON>Abd al-Aziz ibn Baz</PERSON>

BIBLIOGRAPHIE

Abdel Fattah Mohamed et Ren Fuji. 2008. «English-Arabic Proper-Noun Transliteration-Pairs Creation». *Journal of the American Society for Information Science and Technology,* Vol.59(10), P.1675-1687.

Abdel Fattah Mohamed, Ren Fuji et Kuroiwa Shingo. 2006. «Machine transliteration». *In Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation, China, Novembre 2006,* P.370-373.

Abduljaleel Nasreen et Larkey Leah S. 2003. «Statistical Transliteration for English-Arabic Cross Language Information Retrieval». *In Proceedings of the Twelfth ACM International Conference on Information and Knowledge Management, New Orleans, USA, 2003,* P.139-146.

Abu El-Khair Ibrahim. 2006. «Effects of stop words elimination for Arabic information retrieval: a comparative study». *International Journal of Computing & Information Sciences,* Vol.4(3), P.119-133.

Abuleil Saleem. 2004. «Extracting Names From Arabic Text For Question-Answering Systems». *In Proceedings of the 7th International Conference on Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval. Avignon (Vaucluse), France, 2004,* P.638-647.

Abuleil Saleem et Evens Martha. 2002. «Extracting an Arabic Lexicon from Arabic Newspaper Text». *Computers and the Humanities,* Vol.36(2), P.191-221.

Afli Haithem , Barrault Loïc et Schwenk Holger. 2012. «Traduction automatique à partir de corpus comparables: extraction de phrases parallèles à partir de données

comparables multimodales». *In Proceedings of the Joint Conference JEP-TALN-RECITAL, Grenoble, France, juin 2012,* Vol.2 P.447-454.

Agrawal Neeraj et Singla Ankush. 2010. «Using Named Entity Recognition to improve Machine Translation». *Technical report, Standford University, Natural Language Processing,* P.1-10.

Al-Jumaily Harith, Martínez Paloma, Martínez-Fernández José L. et Goot Erik van der. 2012. «A real time Named Entity Recognition system for Arabic text mining». *Journal of Language Resources and Evaluation,* Vol.46(4), P.543-563.

Al-Onaizan Yaser, Curin Jan, Jahr Michael , Knight Kevin, Lafferty John, Melamed Dan, Och Franz Josef, Purdy David , Smith Noah A. et Yarowsky David. 1999. «Statistical Machine Translation». *Technical Report, Summer Workshop on Language Engineering, Center for Speech and Language Processing, Baltimore, MD, Université Johns-Hopkins, USA, 1999,* P.1-42.

Al-Onaizan Yaser et Knight Kevin. 2002. «Machine transliteration of names in Arabic text». *In Proceedings of the ACL workshop on Computational approaches to semitic languages, Philadelphia, PA, USA, 6-12 juillet 2002,* P.1-13.

Alotaibi Fahd et Lee Mark. 2012a. «Mapping Arabic Wikipedia into the Named Entities Taxonomy». *In Proceedings of the 24th International Conference on Computational Linguistics (COLING), Mumbai, India, décembre 2012,* P.43-52.

Alotaibi Fahd et Lee Mark. 2013. «Automatically Developing a Fine-grained Arabic Named Entity Corpus and Gazetteer by utilizing Wikipedia». *In proceedings of the sixth International Joint Conference on Natural Language Processing, Nagoya, Japan, octobre 2013,* P.392-400.

Alotaibi Fahd et Lee Mark. 2012b. «Using Wikipedia as a Resource for Arabic Named Entity Recognition». *In Proceedings of the 4th International Conference on Arabic Language Processing (CITALA), Rabat, Morocco, mai 2012,* P.1-8.

Arbabi Mansur, Fischthal Scott M., Cheng Vincent C. et Bart Elizabeth. 1994. «Algorithms for Arabic name transliteration». *IBM Journal of Research and Development,* Vol.38 (2), P.183-194.

Attia Mohamed. 2008. «Handling Arabic Morphological and Syntactic Ambiguity within the LFG Framework with a View to Machine Translation». Thèse de doctorat en philosophie, *Université de Manchester.*

Attia Mohamed, Toral Antonio, Tounsi Lamia, Monachini Monica et Genabith Josef van. 2010. «An automatically built Named Entity lexicon for Arabic». *In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC), Valletta, Malta, mai 2010,* P.3614-3621.

Azab Mahmoud, Bouamor Houda , Mohit Behrang et Oflazer Kemal. 2013. «Dudley North visits North London: Learning When to Transliterate to Arabic». *In Proceedings of the conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Atlanta, Georgia, juin 2013,* P.439-444.

Bach Nguyen, Noamany Mohamed, Lane Ian et Schultz Tanja. 2007. «Handling OOVWords In Arabic ASR Via Flexible Morphological Constraints». *In Proceedings of the eighth conference of Interspeech, Antwerp Belgium, août 2007,* P.2373-2376.

Ben-Hamadou Abdelmajid, Piton Odile et Fehri Héla. 2010. «Recognition and translation Arabic-French of Named Entities: case of the Sport places». *Computing Research Repository (CoRR), Electronic Edition, 1002.0481,* P.1-8.

Ben Abacha Asma, Zweigenbaum Pierre et Max Aurélien. 2012. «Extraction d'information automatique en domaine médical par projection inter-langue : vers un passage à l'échelle». *In Proceedings of the Joint Conference JEP-TALN-RECITAL, TALN, Grenoble, France, juin 2012,* Vol.2 P.15-28.

Benajiba Yassine. 2009. «Arabic Named Entity Recognition», Thèse de doctorat, Université polytechnique de Valence.

Benajiba Yassine, Diab Mona et Rosso Paolo. 2008. «Arabic Named Entity Recognition using Optimized Feature Sets». *In Proceedings of the Empirical Methods in Natural Language Processing (EMNLP), Waikiki, Honolulu, Hawaii, octobre 2008,* P.284-293

Benajiba Yassine et Rosso Paolo. 2007. «ANERsys 2.0: Conquering the NER task for the Arabic language by combining the maximum entropy with POS-tag information». *In Proceedings of the 3rd Indian International Conference on Artificial Intelligence, IICAI-2007, Pune, India, décembre 2007,* P.1814-1823.

Benajiba Yassine et Rosso Paolo. 2008. «Arabic named entity recognition using conditional random fields». *In Proceedings of the Conference on Language Resources and Evaluation (LREC), Marrakech Morocco, mai 2008.*

Benajiba Yassine, Rosso Paolo et Benedi Ruiz José Miguel. 2007. «ANERsys: An Arabic Named Entity Recognition System Based on Maximum Entropy». *In Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing, Mexico City, Mexico, février 2007,* P.143-153.

Berger Adam , Della Pietra Stephen et Della Pietra Vincent. 1996. «A maximum entropy approach to natural language processing». *Computational Linguistics,* Vol.22(1), P.39-71.

Bikel Daniel, Miller Scott, Schwartz Richard et Weischedel Ralph .1997. «Nymble: a High-Performance Learning Name-finder». *In Proceedings of the Fifth Conference on Applied Natural Language Processing, Washington, USA, 31 mars - 3 avril, 1997,* P.194-201.

Blain Frédéric. 2013. «Modèles de traduction évolutifs». Thèse de doctorat en informatique, *Université du Maine*

Blum Avrim et Mitchell Tom. 1998. «Combining labeled and unlabeled data with co-training». *In Proceedings of the Workshop on Computational Learning Theory (COLT), Morgan Kaufmann, juillet 1998,* P.92-100.

Boitet Christian et Guillaume Pierre. 1982. «Ariane-78: an integrated environment for automated translation and human revision». *In Proceedings of the Ninth International Conference on Computational Linguistics (COLING), Prague, juillet 1982,* P.19-27.

Brown Peter, Della Pietra Stephen, Della Pietra Vincent et Mercer Robert L. 1993. «The Mathematics of Statistical Machine Translation: Parameter Estimation». *Computational Linguistics - Special issue on using large corpora,*Vol.19(2), P.263-311.

Brown Peter F., Cocke John , Della Pietra Stephen A., Della Pietra Vincent J. , Jelinek Fredrick , Lafferty John D., Mercer Robert L. et Roossi Paul S. 1990. «A statistical approach to machine translation». *Computational Linguistics,*Vol.16(2), P.79-85.

Budi Indra et Bressan Stéphane. 2003. «Association Rules Mining for Name Entity Recognition». *In Proceedings of the Fourth International Conference on Web Information Systems Engineering (WISE), Rome, Italy, 10-12 décembre 2003,* P.325-328.

Bunescu Razvan et Pasca Marius. 2006. «Using Encyclopedic Knowledge for Named Entity Disambiguation». *In Proceesings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Trento, Italy, avril 2006,* P.9-16

Carl Michael. 2003. «Introduction à la traduction guidée par l'exemple (Traduction par analogie)». *In proceedings of 10th Conference of TALN, Batz-sur-Mer, France, juin 2003,* P.11-26.

Che Wanxiang , Wang Mengqiu, Manning Christopher D. et Liu Ting. 2013. «Named Entity Recognition with Bilingual Constraints». *In Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), Atlanta, juin 2013,* P.1-11.

Chinchor Nancy. 1997. «MUC-7 Named Entity Task Definition». *In Proceedings of the Seventh Message Understanding Conference (MUC-7), Fairfax, Virginia: USA, 19 Avril- 1 mai 1998.*

Costa-jussa Marta R. , Henriquez Carlos A. et Banchs Rafael E. 2013. «Evaluating Indirect Strategies for Chinese–Spanish Statistical Machine Translation: Extended Abstract». *In Proceedings of the Twenty-Third international joint conference on Artificial Intelligence, Beijing, août 2013,* P.3142-3145.

Daille Béatrice, Fourour Nordine et Morin Emmanuel. 2000. «Catégorisation des noms propres : une étude en corpus». *Cahiers de Grammaire,*Vol.25 P.115-129.

Dejean Hervé et Gaussier Éric. 2002. «Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables». *Lexicometrica, Alignement lexical dans les corpus multilingues,* P.1-22.

Do Thi-Ngoc-Diep. 2011. «Extraction de corpus parallèle pour la traduction automatique depuis et vers une langue peu dotée». Thèse de Doctorat en sciences en informatique, *Université de Grenoble.*

Dyer Chris, Lopez Adam et Ganitkevitch Juri. 2010. «cdec: A Decoder, Alignment, and Learning Framework for Finite-State and Context-Free Translation Models». *In Proceedings of the ACL System Demonstrations, Uppsala, Sweden, juillet 2010,* P.7-12.

Ehrmann Maud. 2008. «Les entités nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation». Thèse de doctorat en Linguistique théorique, descriptive et automatique, *Université Paris 7.*

Ehrmann Maud, Turchi Marco et Steinberger Ralf. 2011. «Building a Multilingual Named Entity-Annotated Corpus Using Annotation Projection». *In Proceedings of Recent Advances in Natural Language Processing, Hissar, Bulgaria, septembre 2011,* P.118-124.

Elsa Tolone. 2006. «Extraction d'Entités Nommées par les Graphes d'Unitex». *Technical report, I.G.M., Université de Marne-la-Vallée, septembre 2006,* P.1-43.

Elsebai Ali, Meziane Farid et Belkredim Fatma_Zohra. 2009. «A Rule Based Persons Names Arabic Extraction System». *In the 11th Communications of the IBIMA, Cairo, Egypt, janvier 2009,* Vol.11 P.53-59.

Elyan Jean. 2012. «La traduction automatique a 58 ans, une initiative IBM/Georgetown University». *Article en ligne* *http://www.lemondeinformatique.fr/actualites/lire-la-traduction-automatique-a-58-ans-une-initiative-ibm-georgetown-university-47429.html*.

Fehri Héla. 2012. «Reconnaissance automatique des entités nommées arabes et leur traduction vers le français». Thèse de doctorat en informatique, *Université de Sfax, Faculté des Sciences Économiques et de Gestion.*

Finkel Jenny-Rose , Grenager Trond et Manning Christopher. 2005. «Incorporating non-local information into information extraction systems by Gibbs sampling». *In Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL), University of Michigan-Ann Arbor, juin 2005,* P.363-370

Gahbiche-Braham Souhir. 2013. «Amélioration des systèmes de traduction par analyse linguistique et thématique Application à la traduction depuis l'arabe». Thèse de doctorat en informatique, *Université Paris Sud.*

Gahbiche-Braham Souhir, Bonneau-Maynard Hélène, Lavergne Thomas et Yvon François. 2012. «Repérage des entités nommées pour l'arabe : adaptation non-supervisée et combinaison de systèmes». *In Proceedings of the conference of JEP-TALN-RECITAL, Grenoble, France, juin 2012,* Vol.2 P.487-494.

Gahbiche-Braham Souhir, Bonneau-Maynard Hélène et Yvon François. 2014. «Traitement automatique des entités nommées en arabe : détection et traduction». *TAL (Traitement Automatique des Langues),* Vol.54(2), P.101-132.

Goldman Jean-Philippe et Scherrer Yves. 2012. «Création automatique de dictionnaires bilingues d'entités nommées grâce à Wikipédia». *Cahiers de linguistique française,* Vol.30(11), P.213-227.

Goudet Jean-Luc. 2008. «Traduction automatique : les années où tout a changé». *Article en ligne http://www.futura-sciences.com/.*

Grass Thierry. 2010. «À quoi sert encore la traduction automatique ?». *Article en ligne http://www.cahiersdugepe.fr/index1367.php.*

Gravier G, Bonastre J.-F, Geoffrois E, Galliano S, Mctait K et Choukri K . 2004. «The ESTER evaluation campaign of rich transcription of French broadcast news». *In Proceedings of the 4th international Conference on Language Resources and Evaluation (LREC), Lisboa, Portugal, mai 2004,* P.885-888.

Grishman Ralph et Sundheim Beth. 1996. «Message Understanding Conference - 6: A Brief History». *In Proceedings of the 16th conference on Computational linguistics (COLING), Copenhagen, Denmark, août 1996,* P.466-471

Grouin Cyril, Galibert Olivier, Rosset Sophie, Quintard Ludovic et Zweigenbaum Pierre. 2011. «Mesures d'évaluation pour entités nommées structurées». *In Ateliers joints QDC'2011 - EvalECD'2011. Évaluation des méthodes d'Extraction de Connaissances dans les Données, Brest, France, janvier 2011,* P.49-62.

Guillemin-Lanne Sylvie , Debili Fathi, Ben Tahar Zied et Gaci Chafik. 2007. «Reconnaissance des entités nommées en arabe». *In Colloque Veille Stratégique Scientifique et Technologique (VSST)- Systèmes d'information élaborée, Bibliométrie, Marrakech, Morocco, octobre 2007.*

Habash Nizar. 2008. «Four Techniques for Online Handling of Out-of-VocabularyWords in Arabic-English Statistical Machine Translation». *In Proceedings of Association for Computational Linguistics-Human Language Technologies (ACL-HLT), Columbus, Ohio, USA, juin 2008,* P.57-60.

Habash Nizar, Rambow Owen et Roth Ryan. 2009. «MADA+TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization». *In Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt, Avril 2009,* P.102-109.

Habash Nizar et Roth Ryan. 2008. «Identification of Naturally Occurring Numerical Expressions in Arabic». *In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC), Marrakech, Morocco, mai 2008,* P.3330-3336.

Habash Nizar et Sadat Fatiha. 2006. «Arabic Preprocessing Schemes for Statistical Machine Translation». *In Proceedings of the Human Language Technology Conference of the NAACL, Companion, New York City, USA, juin 2006,* P.49-52.

Habash Nizar, Soudi Abdelhadi et Buckwalter Tim. 2007. «On Arabic Transliteration». *Arabic Computational Morphology Text, Speech and Language Technology,* Vol.38 P.15-22.

Heitz Thomas. 2006. «Modélisation du prétraitement des textes». *In Proceedings of the JADT -8es Journées internationales d'Analyse statistique des Données Textuelles, Besançon, France, avril 2006,* P.1-8.

Hermjakob Ulf, Knight Kevin et Daume III Hal. 2008. «Name Translation in Statistical Machine Translation Learning When to Transliterate». *In Proceedings of Association for Computational Linguistics - Human Language Technologies (ACL-HLT), Columbus, Ohio, USA, juin 2008,* P.389-397.

Kalyanee Kanchan Baruah, Pranjal Das, Abdul Hannan et Shikhar Kr Sarma. 2014. «Assames-English bilingual machine translation». *International Journal on Natural Language Computing (IJNLC) juin 2014,* Vol.3(3), P.73-82.

Kashani Mehdi M. 2007. «Automatic transliteration from arabic to english and its impact on machine translation». Mémoire de Master en science, *Université Simon Fraser, Colombie-Britannique.*

Kashani Mehdi M., Joanis Eric, Kuhn Roland, Foster George et Popowich Fred. 2007. «Integration of an Arabic Transliteration Module into a Statistical Machine Translation System». *In Proceedings of the Second Workshop on Statistical Machine Translation, Prague, Czech Republic, June 2007,* P.17-24.

Koehn Philipp. 2004. «Pharaoh: A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models». *In Proceedings of the 6th conference of the Association for Machine Translation in Americas, Washington, USA, 28 septembre - 2 octobre 2004,* Vol.3265 P.115-124.

Koehn Philipp, Hoang Hieu, Birch Alexandra, Callison-Burch Chris, Federico Marcello, Bertoldi Nicola, Cowan Brooke, Shen Wade, Moran Christine, Zens Richard, Dyer Chris, Bojar Ondrej, Constantin Alexandra et Herbst Evan. 2007. «Moses: Open Source Toolkit for Statistical Machine Translation». *In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, Prague, Czech Republic, 23-30 juin 2007,* P.177-180.

Lafferty John, McCallum Andrew et Pereira Fernando. 2001. «Conditional random fields: Probabilistic models for segmenting and labeling sequence data». *In Proceedings of 18th International Conference on Machine Learning, San Francisco, CA, USA, 28 juin - 1 juillet 2001,* P.282-289.

Larochelle Hugo. 2009. «Étude de techniques d'apprentissage non-supervisé pour l'amélioration de l'entraînement supervisé de modèles connexionnistes». Thèse de doctorat en informatique, *Université de Montréal.*

Lavecchia Caroline. 2010. «Les triggers inter-langues pour la Traduction Automatique Statistique». Thèse de doctorat en informatique, *Université Nancy.*

Lavecchia Caroline, Smaili Kamel et Langlois David. 2008. «Une alternative aux modèles de traduction statistique d'IBM: Les triggers inter-langues». *In Proceedings*

of 15eme conférence sur le Traitement Automatique des Langues Naturelles (TALN), Avignon. inria-00285275, version 1, Avignon, France, juin 2008.

Lavergne Tomas, Cappé Olivier et Yvon François. 2010. «Practical Very Large Scale CRFs». *In Proceedings of 48th Annual Meeting Association for Computational Linguistics (ACL), Uppsala, Sweden, juillet 2010,* P.504-513.

Le Hai Son. 2013. «Continuous space models with neural networks in natural language processing». Thèse de doctorat en informatique, *Université Paris Sud.*

Le Meur Céline, Galliano Sylvain et Geoffrois Edouard. 2004. «Conventions d'annotations en Entités Nommées - ESTER». *Article en ligne* http://www.afcp-parole.org/ester/docs/convention_en_old.pdf.

Lehmann Jens, Isele Robert, Jakob Max , Jentzsch Anja , Kontokostas Dimitris, Mendes Pablo N., Hellmann Sebastian , Morsey Mohamed, Kleef Patrick van, Auer Soren et Bizer Christian. 2012. «DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia». *Semantic Web Journal, 2012,* P.1-29.

Levenshtein Vladimir. 1966. «Binary codes capable of correcting deletions, insertions and reversals». *Soviet Physics-Doklady, 10,*Vol.6 P.707-710.

Loffler-Laurian A.M. 1996. «La traduction automatique». *Presses universitaires du Septentrion, ISBN: 9782859395025. URL:* http://books.google.ca/books?id=Fm8ecQQQKSkC

Maâloul Mohamed Hédi. 2012. «Approche hybride pour le résumé automatique de textes. Application à la langue arabe». Thèse de doctorant en informatique, *Université de Provence - Aix-Marseille I.*

Maghi King. 1981. «Eurotra–a european system for machine translation». *Journal Lebende Sprachen,*Vol.26(1), P.12-14.

Mallat Souhey, Ben Mohamed Mohamed Achraf , Hkiri Emna, Zouaghi Anis et Zrigui1 Mounir. 2014. «Semantic and Contextual Knowledge Representation for Lexical Disambiguation: Case of Arabic-French Query Translation». *Journal of Computing and Information Technology,* Vol.22(3), P.191-215.

Maloney John et Niv Michael. 1998. «TAGARAB: A Fast, Accurate Arabic Name Recognizer Using High-Precision Morphological Analysis». *In Proceedings of the Workshop on Computational Approaches to Semitic Languages, Montreal, Canada, Août 1998,* P.8-15.

Manola Frank , Miller Eric et McBride Brian. 2004. «RDF Primer, W3C Recommendation». *article en ligne* http://www.w3.org/TR/2004/REC-rdf-primer-20040210/.

Mansouri Alireza, Affendey Lilly-Suriani et Mamat Ali. 2008. «Named Entity Recognition Approaches». *IJCSNS International Journal of Computer Science and Network Security,* Vol.8 P.339-344.

Mendes Pablo N., Jakob Max et Bizer Christian. 2012. «DBpedia: A Multilingual Cross-Domain Knowledge Base». *In Proceedings of the International Conference on Language Resources and Evaluation, LREC, Istanbul, Turkey, 21-27 mai 2012.*

Mesfar Slim. 2007. «Named Entity Recognition for Arabic Using Syntactic Grammars». *Natural Language Processing and Information Systems,* Vol.4592 P.305-316.

Mesfar Slim. 2008. «Analyse morphosyntaxique automatique et reconnaissance des entités nommées en arabe standard ». Thèse de doctorat en informatique, *Université de Franche-Comté, école doctorale 'langages, espaces, temps, sociétés'.*

Mikheev Andrei , Moens Marc et Grover Claire. 1999. «Named Entity Recognition without Gazetteers». *In Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics (EACL), University of Bergen, Bergen, Norway, 8-12 juin 1999,* P.1-8.

Mohammed Naji F. et Nazlia Omar. 2012. «Arabic Named Entity Recognition Using Artificial Neural Network ». *Journal of Computer Science,* Vol.8(8), P.1285-1293.

Mohit Behrang, Schneider Nathan, Bhowmick Rishav, Oflazer Kemal et Smith Noah A. 2012. «Recall-Oriented Learning of Named Entities in ArabicWikipedia». *In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, 23-17 avril 2012,* P.162-173.

Molino Jean. 1982. «Le nom propre dans la langue». *In: Langages, le nom propre, 16e année,* Vol.16(66), P.5-20.

Nadeau David et Sekine Satoshi. 2007. «A survey of named entity recognition and classification». *Linguisticae Investigaciones,* Vol.30(1), P.3–26.

Nagao Makoto. 1984. «A framework of a mechanical translation between Japanesse and Anglish by analogy principale». *In Proceedings of Artificial and Human Intelligence, North-Holland, Amsterdam, 1984,* P.173-180.

Nguyen Trong Khanh. 2007. «Extraction des entités nommées vietnamiennes». *Rapport technique. Institution de la francophonie pour l'informatique, Juillet 2007.*

Nouvel Damien. 2012. «Reconnaissance des entités nommées par exploration de règles d'annotation, Interpréter les marqueurs d'annotation comme instructions de structuration locale». Thèse de doctorat en informatique, *Université François Rabelais De Tours.*

Nwesri Abdusalam F.A., Tahaghoghi S.M.M. et Scholer Falk. 2007. «Finding Variants of Out-of-Vocabulary Words in Arabic». *In Proceedings of the 5th Workshop on Important Unresolved Matters, Association for Computational Linguistics, Prague, Czech Republic, 23-30 juin 2007,* P.49-56.

Och Franz Josef et Ney Hermann . 2000. «Improved Statistical Alignment Models». *In Poceedings of the 38th Annual Meeting of the Association for Computational Linguistics, Hongkong, China, octobre 2000.,* P.440-447.

Och Franz Josef et Ney Hermann. 2003. «A Systematic Comparison of Various Statistical Alignment Models». *Computational Linguistics,*Vol.29(1), P.19-51.

Oudah Mai et Shaalan Khaled. 2013. «Person Name Recognition Using the Hybrid Approach». *Natural Language Processing and Information Systems,*Vol.7934 P.237-248.

Paik Woojin , Liddy Elizabeth D., Yu Edmund et McKenna Mary. 1996. «Categorizing and standardizing proper nouns for efficient information retrieval». *Corpus Processing for Lexical Acquisition, The MIT Press journals, Cambridge, 1996,* P.61-73

Papineni Kishore , Roukos Salim , Ward Todd et Zhu Wei-Jing. 2002. «BLEU: a method for automatic evaluation of machine translation». *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, 6-12 juillet 2002,* P.311-318.

Patry Alexandre et Langlais Philippe. 2011. «Identifying Parallel Documents from a Large Bilingual Collection of Texts: Application to Parallel Article Extraction inWikipedia». *In Proceedings of the 4th Workshop on Building and Using Comparable Corpora, Portland, Oregon, juin 2011.*

Poibeau Thierry. 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.*

Poibeau Thierry et Group the INaLCO Named Entity. 2003. «The Multilingual Named Entity Recognition Framework». *In Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics (EACL), Budapest, Hungary, 12-17 avril 2003,*Vol.2 P.155-158.

Pouliquen Bruno, Steinberger Ralf, Ignat Camelia, Temnikova Irina, Widiger Anna, Zaghouani Wajdi et Zizka Jan. 2005. «Multilingual person name recognition and transliteration». *Journal CORELA - Cognition, Representation, Langage. Numeros speciaux, Le traitement lexicographique des noms propres,*Vol.3 P.1-24.

Prud'hommeaux Eric et Seaborne Andy. 2008. «SPARQL Query Language for RDF, . W3C Recommendation 2008». *article en ligne http://www.w3.org/TR/rdf-sparql-query/.*

Sadat Fatiha. 2010. «Exploiting a Multilingual Web-based Encyclopedia for Bilingual Terminology Extraction». *In Proceedings of the 24th Pacific Asia Conference on Language, Information and computation (PACLIC), Sendai, Japan, 4-7 novembre 2010,*Vol.7614 P.519-526.

Sadat Fatiha, Johnson Howard, Agbago Akakpo, Foster George, Martin Joel et Tikuisis Aaron. 2005. «Portage : A phrase-based machine translation system». *In Proceedings of the ACL Worskhop on Building and Using Parallel Texts, Ann Arbor, juin 2005,* P.129-132.

Sadat Fatiha et Terrasa Alexandre. 2010. «Exploitation de Wikipédia pour l'Enrichissement et la Construction des Ressources Linguistiques». *In proceedings of TALN 2010, Montréal, Canada, 19-23 juillet 2010.*

Samy Doaa , Moreno Antonio et Guirao José M. 2005. «A Proposal For An Arabic Named Entity Tagger Leveraging a Parallel Corpus». *In Proceedings of Recent Advances in Natural Language Processing (RANLP), Borovets, Bulgaria, 21-23 September 2005,* P.459-465.

Sato Satoshi et Nagao Makoto. 1990. «Toward memory-based translation». *In Proceedings of the 13th conference on Computational linguistics (COLING), Université d'Helsinki, Finland, 20-25 août 1990,*Vol.3 P.247-252

Sekine Satoshi. 1998. «Nyu: Description of the Japanese NE System Used For Met-2». *In Proceedings of the Seventh Message Understanding Conference (MUC-7), Fairfax, Virginia, USA, 19 avril- 1mai 1998,* P.1-6.

Sekine Satoshi et Eriguchi Yoshio. 2000. «Japanese Named Entity Extraction Evaluation- Analysis of Results». *In Proceedings of Computational Linguistics (Coling), Saarbrücken, Germany, 31 juillet- 4 août 2000,* P.25-30.

Sellami Rahma, Sadat Fatiha et Belguith Hadrich Lamia. 2014. «Mining Named Entity Translation from Non Parallel Corpora». *In Proceedings of the Twenty-Seventh International Florida Artificial Intelligence Research Society Conference, Florida, USA, 21- 23 mai 2014,* P.219-224.

Sellami Rahma, Sadat Fatiha et Hadrich-Belguith Lamia. 2012. «Exploiting Wikipedia as a Knowledge Base for the Extraction of Linguistic resources: Application on Arabic-French Comparable Corpora and Bilingual Lexicons». *In Proceedings of the Fourth Workshop on Computational Approaches to Arabic Script-based Languages (CAASL4) at AMTA 2012, San Diego, novembre 2012,* P.72-79.

Sellami Rahma, Sadat Fatiha et Hadrich-Belguith Lamia. 2013. «Traduction Automatique Statistique à partir de corpus comparables : Application au couple de

langues arabe-français». *In Proceedings of CORIA, Neuchâtel, Switzerland, 3-5 avril 2013,* P.431-440.

Semmar Nasredine et Saadane Houda. 2013. «Using Transliteration of Proper Names from Arabic to Latin Script to Improve English-Arabic Word Alignment». *In proceedings of the sixth International Joint Conference on Natural Language Processing, Nagoya, Japan, octobre 2013,* P.1022-1026.

Shaalan Khaled et Raza Hafsa. 2007. «Person Name Entity Recognition for Arabic». *In Proceedings of the 5th Workshop on Important Unresolved Matters, Prague, Czech Republic, juin 2007.,* P.17-24.

Stalls Bonnie Glover et Knight Kevin. 1998. «Translating Names and Technical Terms in Arabic Text». *In Proceedings of the Workshop on Computational Approaches to Semitic Languages, Stroudsburg, USA, août 1998,* P.34-41.

Steinberger Ralf, Pouliquen Bruno, Kabadjov Mijail, Belyaeva Jenya et Goot Erik van der. 2011. «JRC-NAMES: A Freely Available, Highly Multilingual Named Entity Resource». *In Proceedings of the 8th International Conference Recent Advances in Natural Language Processing (RANLP), Hissar, Bulgaria, 12-14 septembre 2011,* P.104-110.

Stolcke Andreas. 2002. «SRILM - an Extensible Language Modeling Toolkit». *In proceedings of the 7th International Conference on Spoken Language Processing (ICSLP), USA, 16-20 septembre 2002,* P.901-904.

Sun Bowen. 2010. «Named entity recognition Evaluation of Existing Systems», *Mémoire de maîtrise en système d'information. Université norvégienne de sciences et de technologie (NTNU).*

Tian Liang , Wong Fai et Chao Sam. 2011. «Word Alignment Using GIZA++ on Windows ». *In proceedings of Machine Translation Summit XIII, Xiamen, China, septembre 2011,* P.369-372.

Tjong Kim Sang Erik F. 2002. «Introduction to the CoNLL-2002 Shared Task : Language-Independent Named Entity Recognition». *In proceedings of the 6th conference on Natural language learning- COLING-02,*Vol.20 P.1-4.

Toral Antonio. 2009. «Enrichment of Language Resources by Exploiting New Text and the Resources Themselves. A case study on the acquisition of a NE lexicon». Thèse de doctorat en linguistique, *Université d'Alacant.*

Traboulsi Hayssam. 2009. «Arabic Named Entity Extraction: A Local Grammar-Based Approach». *In Proceedings of the International Multiconference on Computer Science and Information Technology (IMCSIT), Mrągowo, Poland, 12-14 octobre,* P.139-143.

Van Rijsbergen C. J. . 1979. «Information Retrieval. Butterworth». *Article en ligne* http://www.dcs.gla.ac.uk/Keith/Preface.html.

Vapnik Vladimir. 1999. «An Overview of Statistical Learning Theory». *IEEE Transactions on neural networks, septembre 1999,*Vol.10(5), P.988-999.

Yarowsky David et Ngai Grace. 2001. «Inducing multilingual POS taggers and NP brackets via robust projection across aligned corpora.». *In Proceedings of the North American Chapter of the Association for Computational Linguistics NAACL, Pittsburgh, PA, USA, 2-7 juin 2001,* P.1-8.

Zaghouani Wajdi. 2009. «Le repérage automatique des entités nommées dans la langue arabe: vers la création d'un système à base de règles». Mémoire de M.A. en linguistique, *Université de Montréal.*

Zaghouani Wajdi. 2012. «RENAR: A Rule-Based Arabic Named Entity Recognition System». *ACM Transactions on Asian Language Information Processing,* Vol.11(1), P.1-13.

Zidouni Azeddine. 2010. «Modèles graphiques discriminants poour l'étiquetage de séquences : application à la reconnaissance d'entités nommées radiophonique». Thèse de doctorat en informatique, *Université de la Méditerranée, France.*

Zobel Justin et Dart Philip. 1996. «Phonetic String Matching: Lessons from Information Retrieval». *In Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, 18-22 août 1996,* P.166-172.

Zribi Ines, Mezghani-Hammami Souha et Hadrich Belguith Lamia. 2010. «L'apport d'une approche hybride pour la reconnaissance des entités nommées en langue arabe». *In Proceedings of TALN 2010 : 17e conférence sur le Traitement Automatique des Langues Naturelles, Montréal, Canada, 19-23 juillet 2010,* P.1-6.

Zughoul Muhammad Raji et Abu-Alshaar Awatef Miz'il 2005. «English/Arabic/English Machine Translation: A Historical Perspective». *Meta : journal des traducteurs,* Vol.50(3), P.1022-1041.