UNIVERSITÉ DU QUÉBEC À MONTRÉAL

TRADUCTION AUTOMATIQUE STATISTIQUE ET ADAPTATION AU DOMAINE DES MÉDIAS SOCIAUX

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN INFORMATIQUE

PAR

FATMA MALLEK

AVRIL 2017

À mon petit Mohamed !

REMERCIEMENTS

Merci Dieu, le tout puissant et le miséricordieux, qui m'a donné la patience et la force pour accomplir ce travail de maîtrise.

Je tiens tout d'abord à remercier ma directrice de recherche, Mme Fatiha Sadat, pour ses conseils, son soutien précieux, sa patience, ses encouragements et aussi surtout pour avoir cru en moi.

Je tiens également à remercier les professeurs du département d'informatique de l'Université du Québec à Montréal, pour la qualité de leurs enseignements lors de ma maîtrise.

Mes plus vifs remerciements aux membres du jury, pour l'intérêt qu'ils ont eu vis à vis de mon travail de maîtrise et pour leurs conseils qui m'ont permis d'aller vers l'avant.

Je remercie également tous mes amis et collègues au laboratoire de recherche en Gestion, Diffusion et Acquisition des Connaissances (GDAC), qui m'ont apporté leur soutien moral et intellectuel. Aussi un grand remerciement à Émilie Dazé pour l'aide à la révision et à la mise en page de ce mémoire.

Une grande reconnaissance à mes parents, sans qui je n'aurais jamais pu devenir ce que je suis. Un grand merci à ma famille, sans oublier mes deux frères, pour tous les efforts et les sacrifices qu'ils ont fait pour nous et pour toujours !

Le mot merci ne suffit pas pour remercier mon mari, pour son soutien, sa compréhension et ses encouragements. Aussi, un grand merci à mon petit Mohamed, pour l'amour, le plaisir et la joie qu'il fait entrer dans notre vie !

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

ASM	Arabe Standard Moderne
BLEU	*Bilingual Evaluation Understudy* (métrique d'évaluation des systèmes de traduction automatique)
LM	*Language model* (langage de modèle)
OOV	*Out of Vocabulary* (mots hors-vocabulaire)
PBSMT	*Phrase Based System Machine Translation*
POS	*Part of Speech*
SMT	*Statistical Machine Translation*
SRILM	*SRI Language Modeling toolkit*
TA	Traduction Automatique
TAL	Traitement Automatique de Langue
TALN	Traitement Automatique de Langage Naturel
TAS	Traduction Automatique Statistique
TM	*Translation Model* (modèle de traduction)
UN	*United Nation* (Nations Unies)
WBSMT	*Word Based System Machine Translation*

RÉSUMÉ

Le besoin de communiquer en plusieurs langues est devenu une nécessité dans un monde envahi par les nouvelles technologies de communication et les médias sociaux, comme les blogues, les wikis, les microblogues, etc. Ainsi, Twitter constitue une source continue et illimitée de données en langage naturel, qui est particulièrement non structurée et hautement bruitée, ce qui la rend difficile à traiter avec les approches classiques de Traitement Automatique du Langage Naturel (TALN).

Ce travail de recherche consiste donc en l'élaboration d'un système de traduction automatique statistique à base de segments pour la traduction des *tweets* d'une langue à morphologie riche et complexe, l'arabe vers l'anglais.

Notre premier intérêt est le prétraitement des *tweets* hautement bruités pour la langue source (arabe) et la langue cible (anglais). Cette phase comprend la normalisation, la segmentation des mots ainsi que l'adaptation des outils linguistiques existants pour le traitement de ces deux langues naturelles au domaine des médias sociaux.

Notre deuxième intérêt est l'incorporation de données hors domaine lors de l'entraînement des deux modèles de traduction et de langue, afin de concevoir un système de traduction automatique statistique performant pour les *tweets*.

Nos évaluations confirment notre thèse selon laquelle le prétraitement des langues source et cible améliore la performance du système de traduction automatique statistique. De plus, l'utilisation d'un système hybride du domaine et hors domaine pour l'entraînement des modèles de langues ainsi qu'une optimisation des poids du décodeur Moses avec un corpus de développement du domaine a donné un système de traduction automatique statistique plus efficace, pour les *tweets* de la langue arabe vers l'anglais.

Mots clés : médias sociaux, *tweets*, traduction automatique statistique à base de segments, modèle de langue, modèle de traduction, normalisation.

INTRODUCTION

La traduction automatique (TA) et ses différentes applications intéressent les chercheurs œuvrant dans le domaine du traitement automatique des langues naturelles (TALN). À l'ère de l'information numérique et avec l'émergence des médias sociaux, la traduction automatique des textes publiés en ligne est devenue un besoin concret des entreprises et un objet d'intérêt important pour la recherche en informatique. Un exemple fameux est le microblogue Twitter, qui est devenu rapidement très populaire [1]. Depuis son lancement en 2006, ce média social n'est pas seulement utilisé pour des raisons personnelles par les particuliers, mais aussi comme outil médiatique par des organisations en tous genres, y compris plusieurs agences et organisations gouvernementales, dont des canadiennes (Gotti *et al.*, 2014). Face à la fluidité des données publiées sur Twitter et sur les médias sociaux en général, plusieurs travaux de recherche se sont penchés sur l'analyse et le traitement de ce type de données dans le domaine du TALN (Jehl, 2010; Jehl *et al.*, 2012; Ling, 2015).

En effet, les fondateurs de Twitter ont récemment pris conscience de l'importance de la traduction automatique des *tweets* au regard de la diversité des langues parlées par ses utilisateurs. C'est en janvier 2015 que Twitter a ajouté à ses fonctions

---

1. Ce microblogue a pour particularité de permettre à ses utilisateurs de publier des messages qui ne dépassent pas 140 caractères. Suivant les statistiques sur le site de Twitter, ce site de média sociaux compte jusqu'à 310 millions utilisateurs actifs par mois (mars 2016) publiant 500 millions de *tweets* par jour, et ce, dans plus de 40 langues (https ://about.twitter.com/fr/company) (dernière consultation le 8 juillet 2016)

la possibilité de traduire les *tweets* publiés, en intégrant le traducteur automatique Bing à son site [2]. Cependant, la qualité des traductions est plutôt faible en comparaison de celles produites par un traducteur humain professionnel [3]. Cette situation a incité les chercheurs à étudier les méthodes de traduction des données issues des médias sociaux afin de pallier les problèmes engendrés par la nature de ces données, qui sont peu structurées et fortement bruitées, c'est-à-dire qu'elles contiennent des caractères spéciaux, des fautes d'orthographe et des graphies variables que la machine ne peut pas reconnaître et traduire.

Problématique

Les données diffusées par les médias sociaux, notamment les *tweets*, sont très riches et dynamiques. Cependant, le traitement de ces données par l'application des méthodes habituelles de TALN donne des résultats peu concluants en raison de la nature bruitée et peu structurée de ces données (Farzindar et Roche, 2013). Pour offrir des résultats intéressants, le traitement de ce type de données nécessite une bonne analyse et des étapes de prétraitement qui dépendent de la langue et de la taille souvent très grande des corpus traités.

Par ailleurs, le traitement d'une langue morphologiquement riche et complexe comme l'arabe est une tâche difficile en TALN (Habash et Sadat, 2012). Les problèmes rencontrés par les applications habituelles en TALN pour l'arabe sont plus importants, car les textes issus de cette langue comportent beaucoup de caractères informels et leur contenu est fortement bruité. Plusieurs phénomènes ont été observés dans le cas de la traduction de la langue arabe dans les textes tirés des

---

2. http ://www.journaldugeek.com/2015/01/23/twitter-lance-la-traduction-automatique-des-tweets-avec-bing/ (dernière consultation le 8 juillet 2016)

3. https ://support.twitter.com/articles/20172133 (dernière consultation le 8 juillet 2016)

médias sociaux, par exemple le problème d'Arabizi, qui consiste en l'utilisation des lettres latines pour exprimer un message écrit en arabe (Darwish, 2014). Aussi, l'arabe utilisé dans les médias sociaux est souvent un mélange entre l'arabe standard moderne (ASM) et l'arabe dialectal, c'est-à-dire l'arabe tel que parlé dans une région spécifique du monde arabe. En effet, les arabophones ont davantage tendance à s'exprimer en utilisant leur langue maternelle qu'en utilisant l'arabe standard, en plus de tendre à y inclure des expressions en langues étrangères, un problème connu sous le nom de *code switching* ou *code mixing* dans la littérature (Barman *et al.*, 2014).

Malgré toutes les difficultés engendrées par le traitement des données issues des médias sociaux, plusieurs outils en TALN sont nécessaires pour leur analyse et leur traitement. En particulier, les traducteurs automatiques facilitent la communication entre les internautes du monde entier et élargissent le flux de données échangé entre eux.

Dans la littérature, on distingue plusieurs types d'approches en traduction automatique : l'approche basée sur l'exemple, l'approche basée sur les règles, l'approche statistique et l'approche hybride. Les approches basées sur les règles et les approches hybrides sont de moins en moins utilisées, en raison des difficultés d'utilisation et des efforts humains considérables qu'elles exigent. Pour cette raison, les recherches actuelles dans le domaine de la traduction automatique tendent à choisir les approches automatiques basées sur l'apprentissage machine, comme les approches statistiques.

Cependant, l'élaboration d'un système de traduction automatique basé sur une méthode statistique est confrontée à plusieurs obstacles, dont le plus important est

la non-disponibilité d'un corpus parallèle [4] tiré des médias sociaux, en particulier pour les *tweets*. Ce corpus est important pour l'étape d'entraînement. À notre connaissance, ce type de corpus parallèle n'est pas disponible pour le public et les travaux qui ont tenté de collecter des données pour former un tel type de corpus sont limités (Jehl *et al.*, 2012; Ling, 2015). Ainsi, le problème de la traduction automatique des *tweets* de l'arabe vers l'anglais peut être considéré comme un problème d'adaptation de domaine pour la traduction automatique statistique. L'adaptaion de domaine sert à adapter le plus possible le modèle de traduction entraîné sur des données en ASM selon la nature du texte à traduire, ici les *tweets*. Autrement dit, il s'agit de réajuster les poids du modèle de traduction.

Par la présente recherche qui s'inscrit dans le domaine du TALN, notre objectif est de développer et d'évaluer un système de traduction automatique statistique de l'arabe vers anglais pour les *tweets*. Ce projet nous oblige à aborder de front plusieurs défis et soulève plusieurs questions :

— Quelles sont les caractéristiques linguistiques de la langue arabe telle qu'utilisée dans les médias sociaux ? À quel degré cette langue est-elle proche de l'ASM ?

— Quels sont les prétraitements utiles en traduction automatique statistique lorsqu'on traite une langue source à morphologie riche et complexe comme l'arabe ?

— Est-ce que l'utilisation de corpus parallèles d'un domaine général pour la traduction des *tweets* va mener à une traduction automatique de qualité et à un traducteur automatique ayant une bonne performance ?

---

4. Un corpus parallèle est une collection de textes bilingues qui sont généralement alignés au niveau de la phrase ou des paragraphes, c'est-à-dire des textes dans la langue source avec leurs traductions dans la langue cible.

Objectifs projetés

Dans la présente recherche, nous étudions la faisabilité d'un traducteur automatique statistique pour les *tweets* de l'arabe vers l'anglais. Parce qu'elle est morphologiquement riche et complexe, la langue arabe n'a jamais cessé d'être l'une des langues les plus étudiées en TALN (Mohammad *et al.*, 2016). Par ailleus, le prétraitement de textes issus des microblogues comme les *tweets* nécessite une étude linguistique et orthographique profonde.

Pour ces raisons, des étapes de prétraitement sont nécessaires, sur lesquelles nous insistons davantage pour la langue arabe que pour la langue anglaise. En effet, pour les applications de TAS, la langue source, l'arabe, nécessite des étapes de prétraitement et de normalisation différentes que la langue cible, l'anglais. Prendre en compte ce paramètre a un effet majeur sur la performance de la machine de traduction automatique. Plusieurs chercheurs ont étudié l'effet de ces prétraitements et ont prouvé leur efficacité pour améliorer la performance et la fiabilité des traducteurs automatiques (Hassan et Darwish, 2014).

L'adaptation de la traduction automatique pour le domaine des *tweets* est le but premier de la présente démarche de recherche. Pour y arriver, nous allons entrainer notre système avec un modèle de langue des *tweets* en langue cible, l'anglais. L'apprentissage du traducteur automatique repose ainsi sur un corpus parallèle du domaine général et sur un modèle de langue dans le domaine spécialisé des *tweets* en anglais. Enfin nous évaluons notre système de traduction automatique en utilisant un corpus parallèle de *tweets* créé manuellement.

Organisation du mémoire

Ce mémoire est constitué de cinq chapitres, qui se présentent comme suit :

— Le premier chapitre fait l'objet de l'état de l'art de la traduction automatique statistique. Nous y abordons les différents concepts concernés par cette démarche et nous en donnons la définition, pour ensuite dresser le portrait de différentes méthodes d'évaluation des systèmes de TAS.

— Le deuxième chapitre est consacré à la présentation de la langue arabe, ses particularités et ses différentes expressions : la langue arabe standard moderne (ASM) et l'arabe dialectal. Nous y illustrons les différentes caractéristiques morphologiques de cette langue et ce qui la différencie des autres. Enfin, nous abordons quelques problèmes majeurs soulevés par la traduction de cette langue pour les applications en TALN ainsi que pour le domaine des médias sociaux.

— Le troisième chapitre présente une revue de littérature des travaux réalisés sur le prétraitement et la traduction automatique de l'arabe ainsi que sur la traduction des textes tirés des médias sociaux de cette langue vers d'autres langues. Nous présentons également certains travaux concernant le prétraitement des données issues des médias sociaux, par exemple les microblogues.

— Le quatrième chapitre présente la méthodologie proposée pour la traduction automatique des *tweets* de l'arabe vers l'anglais. Dans un premier temps, nous présentons notre méthode générale pour leur traduction. Ensuite, nous présentons les étapes de la collecte et du prétraitement des corpus d'apprentissage des *tweets* pour les deux langues.

— Au cinquième chapitre, nous présentons les données et les outils utilisés et nous discutons les résultats obtenus.

— En conclusion, nous résumons notre démarche et proposons des perspectives pour de futurs travaux de recherche dans ce domaine.

CHAPITRE I

LES CONCEPTS DE BASE DE LA TRADUCTION AUTOMATIQUE STATISTIQUE

Dans ce chapitre, nous présentons un bref historique de la traduction automatique, puis nous décrivons les différentes architectures linguistique des systèmes de traduction automatique. Ensuite, nous présentons les concepts de base de la traduction automatique statistique (TAS). Nous commençons par exposer les différents modèles utilisés par les systèmes de traduction statistique : les modèles de langue et les modèles de traduction. Ensuite, nous décrivons les concepts d'alignement et de décodage. Enfin, nous présentons différentes métriques permettant d'évaluer la qualité de la traduction automatique (TA).

1.1 Bref historique de la traduction automatique

La traduction automatique est l'une des tâches les plus intéressantes et les plus difficiles du traitement automatique de langage naturel (TALN). Les premiers travaux sur la TA ont débuté presque en même temps que l'apparition des premiers calculateurs électroniques. Au début, les recherches dans ce domaine ont été développées pour des raisons militaires, pendant la guerre froide. À cette époque, les travaux en traduction ont été influencés par les travaux en cryptographie. C'est

après la Seconde Guerre mondiale, en 1947, que Warren Weaver [5] a lancé les premiers travaux en traduction automatique. Le premier traducteur traduisait des phrases du russe vers l'anglais (Hutchins, 1997).

Les premiers systèmes de traduction ont été basés sur une approche fondée sur les règles, appelée «système de traduction à base de règles» ou *"Rule-based Machine Translation (RBMT)"*. Ces systèmes se fondaient sur des règles de transfert d'une langue à l'autre et des dictionnaires. Au fil du temps, les dictionnaires ont pris de l'ampleur et le nombre de règles est devenu plus important. La construction de ces ressources a nécessité l'investissement d'efforts humains considérables et le développement des grammaires formelles (Afli, 2014).

Pour cette raison et pour pallier les difficultés des approches à base de règles, vers le début des années 1990, une nouvelle famille d'approches basées sur les données (ou *"corpus-based approches"*) a vu le jour. Une première approche recensée dans cette famille est celle basée sur l'exemple (*"Example-based Machine Translation (EBMT)"*). Les systèmes de traduction adoptant cette approche fonctionnent sur la base d'un corpus parallèle composé de textes bilingues, contenant des phrases dans la langue source et leur traduction dans la langue cible.

L'amélioration des performances des ordinateurs et la fluidité du transfert de données textuelles ont ouvert la porte à l'apparition d'une deuxième approche basée sur les données, soit la « traduction automatique statistique » (TAS) (ou *"Statistical Machine Translation (SMT)"*) (Brown *et al.*, 1990). Cette approche est caractérisée par l'utilisation de méthodes d'apprentissage automatiques et de

---

5. Warren Weaver est un scientifique américain, mathématicien et administrateur de la recherche. Il est principalement connu comme un des pionniers de la TA et comme une importante figure de la promotion des sciences aux États-Unis, à travers la Fondation Rockefeller (Haithem, 2010).

données textuelles parallèles volumineuses. Depuis ce temps, l'approche statistique est devenue la méthode de traduction la plus utilisée par la communauté des chercheurs en traduction automatique (Afli, 2014). Dans notre travail, nous avons opté pour cette approche et nous la détaillons avec ses différentes composantes à la section 1.2.

1.1.1 L'architecture linguistique des systèmes de traduction automatique

L'architecture linguistique d'un système de traduction automatique est presque la même pour toutes les approches, même les plus récentes. La figure 1.1, qui représente le triangle de Vauquois (Vauquois et Boitet, 1985), synthétise les trois types de base d'architectures linguistiques : (1) les systèmes directs, (2) les systèmes à transfert et (3), les systèmes à pivot. Le plus bas niveau du triangle de

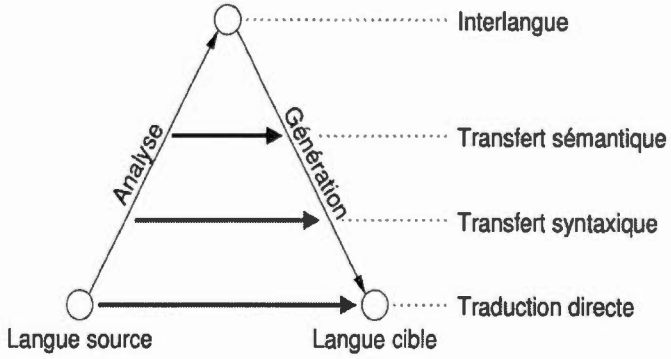

Figure 1.1 Le triangle de Vauquois des différentes architectures linguistiques (Afli, 2014)

Vauquois représente les systèmes de première génération apparus dans les années 1950, appelés *systèmes directs*. Ces systèmes traduisent, unidirectionnellement des textes en entrée (source) en des textes en sortie (cible). Les systèmes de traduction de deuxième génération, appelés *systèmes à transfert*, sont plus complexes : ils ajoutent une étape d'analyse sémantique (2^{me} niveau du triangle) et/ou une

étape d'analyse syntaxique (3^{me} niveau du triangle) pour les données à traduire. Ces systèmes sont plus utilisés que les systèmes directs. Enfin, le plus haut niveau du triangle représente les systèmes d'une autre génération, appelés *systèmes à pivot*. L'avantage de ces systèmes est que les modules d'analyse et de génération sont réutilisables pour la création d'un autre système de traduction pour un autre couple de langues (Afli, 2014).

1.2 Les modèles de la traduction statistique

Les modèles de traduction statistique sont basés sur une théorie mathématique de distribution et d'estimation probabiliste développée par Peter F.Brown et ses collègues à IBM (Brown *et al.*, 1990). En utilisant ces modèles et à partir d'un corpus parallèle annoté, il est possible d'apprendre des relations statistiques entre deux langues données. Les modèles de TA ont été inspirés des modèles présentés par Shannon (1949), appelés *modèles par «canal bruité»*, qui sont illustrés à la figure 1.2. Suivant ce modèle, on suppose qu'il existe deux personnes, un émetteur (*transmitter*) et un récepteur (*receiver*) qui veulent communiquer via un canal bruité. Si l'émetteur envoie une phrase (S), cette phrase sera transmise via ce canal bruité, puis elle sera acheminée vers le récepteur comme une autre phrase (T), qui est la traduction de (S). Pour les systèmes de traduction statistique, on suppose que pour chaque phrase source s, il existe une phrase traduite dans la langue cible t. Ainsi pour la paire de phrase (s, t), la probabilité que la traduction de la phrase dans la langue source s vers une autre phrase t dans la langue cible est notée par $P(s|t)$. En nous basant sur le théorème de Bayes sur la paire de phrases (s, t), la probabilité $P(s|t)$ est calculée suivant l'équation (1.1) :

$$P(s|t) = \frac{P(t|s)P(s)}{P(t)} \tag{1.1}$$

Généralement, le but de la traduction statistique est de trouver la meilleure traduction possible t^*, c'est-à-dire la valeur qui maximise $P(s|t)$ pour la traduction

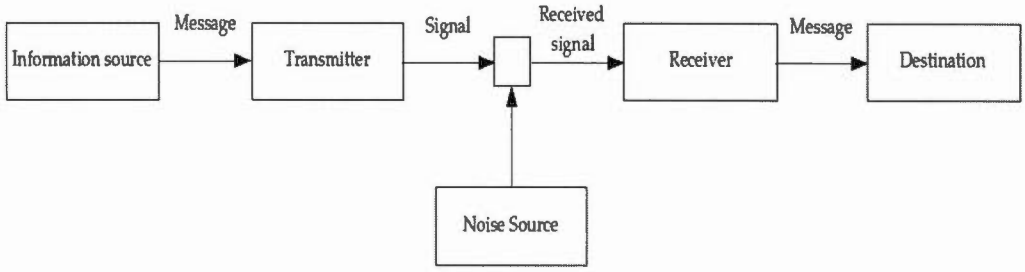

Figure 1.2 Modèle du canal bruité (Shannon, 1949)

de la phrase s en t, parmi toutes les traductions t possibles dans la langue cible. D'une façon formelle, t^* est calculé suivant l'équation (1.2).

$$t^* = arg\ \max_t P(t|s) = arg\ \max_t P(s|t)P(t) \tag{1.2}$$

L'équation (1.2) est une équation fondamentale en TAS. On suppose que s et t sont indépendantes et en utilisant le produit $P(s|t)P(s)$, on arrive à générer cette équation. En effet, on désigne par $P(t)$ le modèle de langue de la langue cible et par $P(s|t)$ le modèles de traductions. Ces deux modèles sont fondamentaux dans les systèmes de traductions statistiques et sont appris d'une manière empirique à partir des corpus utilisés. Aux sections 1.2.1 et 1.2.2, nous présentons en détail ces deux modèles.

1.2.1 Le modèle de langue

Les modèles de langues sont très utilisés dans le domaine du TALN, et ce, pour plusieurs applications, telles que la reconnaissance automatique de la parole, la correction orthographique, la traduction automatique, etc. Ainsi, le modèle de langue constitue un modèle de base pour les systèmes de traduction statistique. Son rôle est d'estimer la probabilité qu'apparaisse un segment ou une suite de

mots. Plus les mots d'une phrase sont conformes au modèle de langue, plus la probabilité de cette phrase est élevée (Brown *et al.*, 1990).

En effet, les connaissances du modèle de langue sont extraites d'un corpus monolingue dans la langue cible du traducteur. Plus le corpus à partir duquel il a été créé est grand, plus le modèle de langue est complet et couvre l'ensemble du vocabulaire de la langue cible. Pour la traduction statistique, il est préférable que le modèle de langue soit du même domaine que les corpus de test et d'entraînement, pour que le système de traduction puisse retrouver le mot dans le corpus d'entraînement (Langlais *et al.*, 2006). D'un point de vue mathématique, le modèle de langue spécifie une distribution $P(t)$ sur les chaînes t^I de la langue modélisée (chaîne cible). Si on considère que t^I est une suite de I mots, $t^I = w_1...w_I$, alors l'expression mathématique de t^I est la suivante :

$$P(t^I) = \prod_{i=1}^{I} P(w_i|w_1...w_{i-1}) \tag{1.3}$$

Pour simplifier cette modélisation, nous supposons que les mots w_i de t^I ne dépendent que des $(i-1)$ mots précédents. On peut donc formuler la probabilité $P(t^I)$ comme suit :

$$P(t^I) = P(w_1)P(w_2|w_1)P(w_3|w_1w_2)...P(w_I|w_1w_2...w_{I-2}w_{I-1}) \tag{1.4}$$

On appelle ce modèle de langue le modèle n-grammes. Ce modèle établit des prédictions sur la base d'une fenêtre de taille fixe, contenant n mots.

Pour chaque séquence, une probabilité est calculée en prenant en compte les $(n-1)$ mots qui précèdent le mot courant pour chaque position dans la phrase cible. Cette probabilité représente la dépendance de chaque mot par rapport aux $(n-1)$ mots qui le précèdent, comme l'indique l'équation (1.4). Statistiquement parlant, si la séquence de mots à traduire n'existe pas dans le modèle de langue, une probabilité nulle lui est attribuée.

1.2.2 Le modèle de traduction

Le modèle de traduction sert pour sa part à modéliser le processus de génération d'une phrase source en phrase cible. Le problème principal de ce modèle est le calcul de $P(s^J|t^I)$, soit la probabilité que la phrase t^I en langue source soit traduite en la phrase s^J en langue cible. Ce modèle est appris à partir de corpus parallèles, c'est-à-dire d'un corpus bilingue aligné, où chaque phrase de la langue cible correspond à une autre phrase de la langue source.

En effet, pour un meilleur apprentissage, le corpus parallèle doit être bien aligné, c'est-à-dire qu'à chaque phrase de la langue source sur la ligne i correspond une traduction dans la langue cible sur la même ligne i. Généralement, les données parallèles sont insuffisantes pour apprendre phrase par phrase la probabilité $P(s^J|t^I)$. Il est donc nécessaire de décomposer les phrases t^I et s^J en plus petites unités. Les unités peuvent être des segments (groupes de mots) ou des mots. On distingue ainsi deux types de modèles de traduction : celui à base de mots et celui à base de segments. Supposons que s^J se décompose en $s_1...s_M$. On aura donc $s^J = s_1 s_2 s_3 ... s_M$. De même pour t^I ; $t^I = t_1 t_2 t_3 ... t_N$. Ainsi, il est très important d'apprendre les alignements entre les différentes composantes des phrases du corpus parallèle correspondant.

Dans la partie qui suit, nous expliquons plus en détail la notion d'alignement pour ces deux modèles de traduction.

1.3 La notion d'alignement

La plupart des modèles de traduction statistique existants utilisent une variable cachée A, appelée *alignement*. Cette variable décrit une correspondance entre les mots ou groupes de mots (segments) d'une phrase et ceux de leur traduction parmi les traductions possibles.

Supposons toujours le modèle de traduction statistique $P(s|t)$, calculant la probabilité que la phrase $t = t_1 t_2 ... t_N$ soit une traduction de la phrase $s = s_1 s_2 ... s_M$. La probabilité de traduction est estimée par la somme des alignements possibles entre s et t. Nous aurons ainsi $P(s|t) = \sum_{a \in A} P(s, a|t)$ avec : (1) a est l'alignement possible, (2)$A = (N+1)^M$, (3)N est le nombre de mots ou de segments dans la phrase source et (4) M est le nombre de mots ou de segments dans la phrase cible.

Dans ce qui suit, nous détaillons les deux modèles d'alignement : l'alignement par mot (*"word-based"*) et l'alignement par segment (*"phrase-based"*).

1.3.1 Le modèle de traduction à base de mots

En 1990, IBM a proposé les premiers modèles probabilistes pour la traduction automatique (Brown *et al.*, 1990). Ces modèles sont à base de mots, c'est-à-dire que l'unité de traduction qui apparaît dans les lois de probabilité est le mot. Autrement dit, ils reposent sur des modèles qui traitent la traduction d'une phrase mot par mot.

Les figures 1.3 et 1.4 illustrent deux exemples d'alignement pour les couples de langue anglais-français et arabe-français.

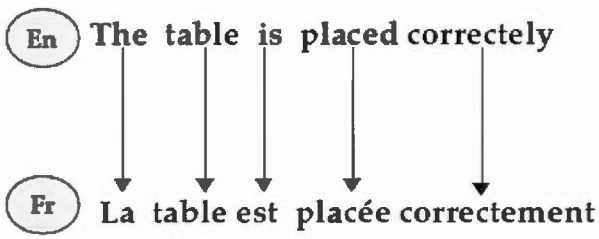

Figure 1.3 Exemple d'alignement à base de mots d'une phrase en anglais et sa traduction en français

On note que pour la figure 1.4, la phrase en arabe prend la forme *verbe-sujet-complément d'objet*, alors qu'en français, la phrase prend plutôt la forme *sujet-verbe-complément d'objet*. La phrase cible et la phrase source ne présentent pas forcément le même ordre de mots. Dans l'étape de réordonnancement, après l'étape de traduction, les mots seront remis dans l'ordre normal de la langue cible (Gahbiche-Braham, 2013).

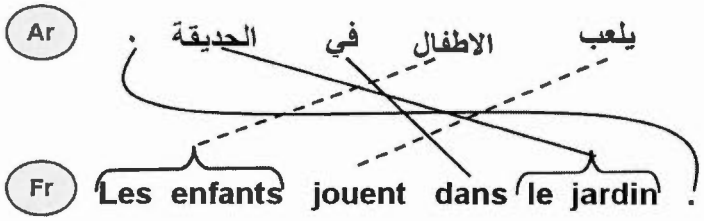

Figure 1.4 Exemple d'alignement d'une phrase en arabe et sa traduction en français

Les modèles de traduction à base de mots, ou *«word-based»*, ont été proposés par IBM dans les années 90 (Brown *et al.*, 1990, 1993). Ils ont ainsi proposé cinq modèles appelés les «modèles IBM», dont le but est d'évaluer la probabilité de traduction $P(s|t)$ d'une phrase source en phrase cible.

Le modèle IBM-1 défini par Brown *et al.* (1990) ne prend pas en considération l'ordre des mots dans la phrase. La distribution des mots est supposée uniforme, donc les alignements sont équiprobables. Autrement dit, il n'y a pas d'alignement assuré entre les mots sources et les mots cibles. Ce modèle repose sur une seule loi de probabilité. Afin de combler les défaillances de ce premier modèle, le modèle IBM-2 a été proposé. Ce dernier prend en considération l'ordre des mots ; il intègre un modèle de réordonnancement qui représente la distance entre un mot de la phrase source et le mot de la phrase cible. Ensuite le modèle IBM-3 a été proposé, qui introduit la notion de fertilité, c'est-à-dire que plusieurs mots la phrase source

peuvent être aligné avec une mot dans la phrase cible. Enfin, les modèles IBM-4 et IBM-5 sont semblables au modèle IBM-3, mais ils utilisent une modélisation un peu plus complexe pour le réordonnancement des mots (Afli, 2014).

Le modèle de traduction à base de mots présente plusieurs problèmes, entre autres celui de non-détermination des appariements mot à mot. Ce modèle suppose que les mots sont indépendants les uns des autres et ne permet donc pas de prendre en considération le contexte du texte à traduire. Du fait, le modèle peut générer de la confusion, par exemple dans les cas de polysémie. Ainsi, le mot «livre», qui définit à la fois un objet et une unité de mesure, pourrait être traduit par *"pound"* ou par *"book"* (Gahbiche-Braham, 2013).

Un autre problème du modèle de traduction à base de mots est le principe de fertilité. En effet, ce problème empêche qu'un mot dans la langue source soit aligné avec plusieurs mots en langue cible. Comme illustré à la figure 1.4, pour les langues morphologiquement riches comme l'arabe, il est très probable qu'un mot en arabe soit aligné avec plusieurs mots en français ou en anglais.

En plus, bien que les modèles d'alignement à base de mots autorisent qu'un mot cible soit aligné avec plusieurs mots sources (traduction m à 1), ils n'autorisent pas que plusieurs mots cibles soient alignés avec un mot source (traduction 1 à n) ou plusieurs mots sources (traduction n à m) (voir la figure 1.5). Or, pour prendre en compte toutes les complexités de la langue source, il est nécessaire de pouvoir générer ce dernier type d'alignement (n à m). Pour cette raison, il est indispensable de considérer pour l'alignement non seulement les mots, mais aussi les groupes de mots (Afli, 2014).

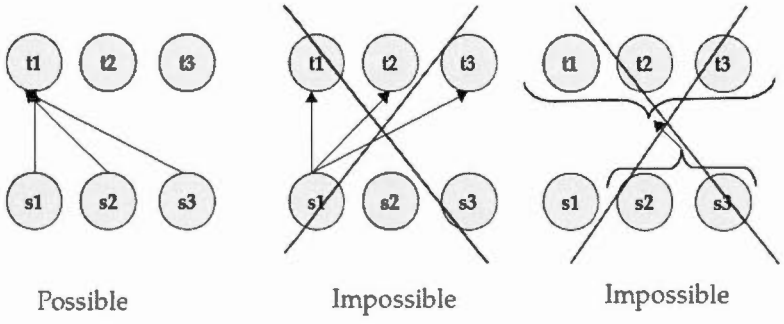

Figure 1.5 Possibilités d'alignement en mots pour les modèles IBM

1.3.2 Le modèle de traduction à base de segments

Les modèles de traduction à base de segments (Koehn *et al.*, 2003) sont les plus utilisés dans les travaux de traduction statistique automatique. Les systèmes qui utilisent ces modèles sont appelés systèmes de traduction à base de segment (*"Phrase-based-SMT :PBSMT"*).

Les modèles de traduction à base de segments sont appris à partir d'un corpus parallèle pour deux langues. Le corpus parallèle doit être bien aligné avant l'apprentissage ; l'alignement est élaboré à l'aide d'un outil d'alignement permettant que chaque segment dans le corpus source soit aligné avec un segment dans le corpus cible (Och et Ney, 2003; Gao et Vogel, 2008). En plus des alignements mot-à-mot et mot-à-plusieurs mots offerts par les modèles à base de mots, des alignements plusieurs-à-plusieurs mot sont offerts par les modèles de traduction à base de segments. À l'étape de l'alignement, les segments sont symétrisés afin de trouver des interactions de ces alignements. Cette étape de symétrisation représente l'alignement final qui sera utilisé pour l'apprentissage. Une fois que les alignements sont obtenus, un score de probabilité est calculé pour tous les segments. Chaque segment en langue source peut avoir plusieurs hypothèses de traduction

dans la langue cible. On note par s le segment en langue source, par t le segment en langue cible et par $c(s,t)$ le nombre de segments dans lesquels apparaît un segment donné dans l'ensemble du corpus. La probabilité qu'un mot ou qu'un groupe de mots dans la langue source soit traduit par un autre dans la langue cible est donnée par le modèle de traduction $P(t|s)$. Mathématiquement parlant, $P(t|s)$ est calculée comme suit :

$$P(t|s) = \frac{c(s,t)}{\sum_{t_i} c(s,t_i)} \tag{1.5}$$

1.4 Le décodage

Dans le domaine de la traduction automatique statistique, le terme *décodage* désigne le processus de traduction qui consiste à transformer une phrase source en phrase cible. Ce terme est inspiré de l'ancien cryptographe Warren Weaver, qui considérait une phrase en russe en tant qu'une phrase anglaise chiffrée. Une fois construits le modèle de langue et le modèle de traduction, ils sont ensuite combinés afin de trouver, pour chaque phrase source, la meilleure traduction possible. Autrement dit, il s'agit de choisir la phrase cible qui maximise la probabilité conditionnelle $P(t|s)$, tel qu'indiqué par l'équation (1.2).

Il est bien connu dans le domaine de la recherche en informatique que le problème de décodage est un problème NP-complet (Knight et Marcu, 2005). Pour résoudre ce problème, il est très important de réduire l'espace de recherche si on veut obtenir des solutions efficaces. Wang et Waibel (1997) ont proposé un algorithme de recherche de la meilleure hypothèse à base de piles. D'autres ont utilisé des transducteurs à états finis pondérés pour implémenter un modèle d'alignement (Zhang *et al.*, 2012). En 2001, Germann *et al.* (2001) ont transformé le problème du décodage en un problème de programmation linéaire, en implémentant un algorithme de recherche par faisceau. Aussi, des algorithmes de programmation

dynamique avec élagage ont été implémentés par Koehn (2004) et par Quirk et Moore (2007).

Il existe plusieurs décodeurs utilisés par la communauté des chercheurs en traduction automatique, parmi lequels Pharaoh[6] (Koehn, 2004), Portage (Johnson *et al.*, 2006) et Moses[7] (Koehn *et al.*, 2007). Pour la présente recherche, nous avons mis en œuvre un système de TA pour les *tweets* reposant sur le décodeur Moses.

1.5 L'évaluation de la qualité des traductions

Une fois la traduction réalisée, il est nécessaire d'évaluer la qualité de la traduction produite. Il existe ainsi deux types d'évaluations : l'évaluation manuelle et l'évaluation automatique.

1.5.1 L'évaluation manuelle

L'évaluation manuelle, dite aussi «subjective», est réalisée par les humains. On demande ainsi à plusieurs participants d'évaluer la traduction selon des critères de qualité bien précis. Ces critères couvrent, par exemple, la fluidité, la fidélité au sens du texte et les corrections grammaticales. Ce type d'évaluation exige l'intervention d'experts bilingues, qui doivent évaluer une très grande quantité de traductions. De plus, chaque traduction doit être évaluée plusieurs fois par différents experts pour s'assurer de la fiabilité des résultats. Toutes ces étapes demandent un travail manuel fastidieux et beaucoup de temps. S'il y a plusieurs traductions pour le même texte, le degré de complexité de l'évaluation manuelle augmente d'autant plus. Considérant que les chercheurs ont souvent besoin d'évaluer et de comparer

---

6. http ://www.isi.edu/licensed-sw/pharaoh/

7. Disponible sous licence LGPL depuis http ://www.statmt.org/moses

plusieurs systèmes de traduction automatique, il est absolument nécessaire de trouver des méthodes d'évaluation automatique capables de faire gagner un temps précieux aux chercheurs (Gahbiche-Braham, 2013).

1.5.2 L'évaluation automatique

L'évaluation automatique est le seul type d'évaluation qui permet de produire des résultats instantanés à faible coût. Qui plus est, ces évaluations sont reproductibles, ce qui n'est pas le cas des évaluations manuelles où le même expert peut évaluer une traduction de différentes manières. Une évaluation automatique permet donc de comparer deux systèmes en utilisant le même corpus de référence. On distingue plusieurs métriques d'évaluation automatique. Cependant, la tâche manuelle a toujours sa place, puisque ces métriques comparent la traduction produite par les systèmes de traduction avec les traductions de référence, qui sont réalisées par des humains.

Nous présentons dans ce qui suit une liste non excustive des métriques d'évaluation automatique qui mesurent la qualité de la traduction produite par les systèmes de traduction automatique.

La métrique BLEU

La métrique BLEU (*«BiLingual Evaluation Understudy»*) proposée par Papineni *et al.* (2002) est la plus utilisée par la communauté scientifique en traduction automatique. Elle compare la traduction produite avec un ou plusieurs fichiers de référence. Le calcul est basé sur une comparaison de courtes séquences de mots, n-grammes, pour chaque phrase du texte traduit et du texte de référence. En effet, le BLEU score tient compte des correspondances entre les n-grammes et entre les mots simples de la phrase traduite et de la phrase référence. Le BLEU récompense la phrase ou l'ordre des mots est le plus proche de l'ordre des mots dans la phrase

de référence. Autrement dit, c'est une mesure de précision qui calcule le degré de similitude entre une traduction et sa référence, en se basant sur la précision n-grammes. Le score BLEU est calculé en pourcentage. Si la traduction produite par le traducteur est identique à la traduction de référence, le score est égal à 100. Dans le cas contraire, où aucune phrase traduite n'existe dans la traduction référence, le score est égal à 0. Les mesures BLEU peuvent être comprises entre 0 et 1. Cette métrique est calculée suivant la formule suivante :

$$BLEU = BP.exp\left(\sum_{n=1}^{N} w_n \log \text{précisions}_n\right) \tag{1.6}$$

$$\text{Où } BP(BrevityPenalty) = \begin{cases} 1 & si\ c > e \\ e^{(1-\frac{r}{c})} & si\ c \leqslant e \end{cases}$$

Ici, c est la longueur de la traduction candidate, r est la longueur de la traduction référence et w_n représente les poids des différentes précisions des n-grammes. La variable BP est une pénalité calculée pour défavoriser les hypothèses de traduction courtes par rapport aux références.

La métrique du NIST

La métrique du NIST (National Institue of Standards and Technology) proposée en 2002 par Doddington (2002), adapte légèrement le score BLEU. Cette métrique est considérée dans plusieurs conférences scientifiques. La différence avec le BLEU est que le score NIST donne plus d'importance aux n-grammes rares. Ainsi, les n-grammes sont pondérés selon leur quantité et par leur fréquence. L'expression de pénalité BP est un peu différente de celle de la métrique BLEU. De plus, le score NIST, prend en compte les précisions des 1-grammes jusqu'à 5-grammes.

La métrique METEOR

Une autre métrique, METEOR (*Metric for Evaluation of Translation with Explicit Ordering*) (Banerjee et Lavie, 2005; Lavie et Agarwal, 2007), introduit plusieurs concepts intéressants. Elle est basée sur des alignements entre les uni-grammes de la traduction automatique et ceux de la traduction de référence. Elle utilise un algorithme itératif qui aligne, lors d'une première étape, les mots strictement identiques et tente, lors d'une seconde étape, d'aligner les mots restants. D'autre part, le score METEOR intègre une pénalité dont le but est de favoriser une traduction qui présente de longs segments consécutifs alignés avec la référence.

Le score OOV

Le score OOV calcule le pourcentage des mots qui n'ont pas été traduits par le système de traduction automatique en question. En TA, généralement, ces mots sont conservés sous leur forme initiale. Ce score dépend principalement de la qualité et la quantité de données utilisées pour l'entraînement de la table de traduction. Plus le score OOV est petit, meilleur il est.

Dans ce mémoire, nous mesurons les traductions et la qualité des systèmes en utilisant le score BLEU et la mesure OOV, qui sont les métriques les plus communément employées par la communauté des chercheurs en traduction automatique.

Les travaux de recherche sur la TAS et l'amélioration des systèmes de TA sont toujours d'actualité. Dans ce chapitre, nous avons fait le survol des notions de base de la TAS. Nous avons présenté d'abord les deux principaux modèles pour la traduction statistique : le modèle de traduction et le modèle de langue. Ensuite, nous avons expliqué les notions d'alignement par mot ainsi que par segment. Nous avons aussi présenté brièvement la notion de décodage, qui consiste à maximiser l'équation de Bayes représentée par l'équation (1.2). Enfin, nous avons clôturé

le chapitre en présentant les métriques les plus utilisées pour l'évaluation des systèmes de traduction automatique.

Dans le prochain chpitre nous présentons les caractéristiques linguistiques de la langue arabe que nous avons traité dans notre présent travail.

CHAPITRE II

LA LANGUE ARABE, LE TALN ET LES MÉDIAS SOCIAUX

La langue arabe est une langue sémitique parlée par plus que 300 millions de locuteurs (Habash, 2010). Elle est la cinquième langue parlée dans le monde [8]. C'est à partir du 8^{eme} siècle qu'une codification de la grammaire de la langue arabe a été proposée pour la fixer dans sa forme classique définitive et faciliter sa propagation par l'enseignement partout où l'islam s'est installé. C'est à cette époque que les premiers traités et dictionnaires de la langue arabe sont apparus. Entre le 8^{eme} et le 10^{eme} siècle, les sciences et techniques islamiques se sont développées (Gahbiche-Braham, 2013). L'alphabet arabe est constitué de 28 lettres qui sont représentées au tableau 2.1.

Dans nos travaux, nous nous sommes intéressés au traitement de la langue arabe dans des textes extraits du microblogue Twitter. Ainsi, ce chapitre est consacré à la représentation de la langue arabe et de ses spécificités dans le domaine du TALN. Nous présentons dans la première section les deux formes de la langue arabe (la forme dialectale et la forme standard) et nous détaillons sa morphologie. Nous définissons ensuite le phénomène des médias sociaux en général, et le microblogue Twitter en particulier. Nous présentons enfin les caractéristiques linguistiques de la langue arabe détectées dans les médias sociaux, y compris Twitter.

---

8. http ://www.ethnologue.com/statistics/size

Tableau 2.1 L'alphabet arabe

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

Dans ce chapitre et le reste du mémoire, plusieurs exemples sont écrits en caractères arabes. Pour faciliter leur compréhension, nous mettons entre deux accolades et après chaque mot ou phrase en arabe une translittération suivant la norme de Buckwalter[9]. La translittération Buckwalter est souvent utilisée dans les articles scientifiques qui traitent la langue arabe afin de faciliter la lecture et l'écriture de cette langue par les chercheurs qui ne maîtrisent pas l'arabe en utilisant des

9. http ://www.qamus.org/transliteration.htm

lettres latines, des symboles et des chiffres (voir par exemple (Habash *et al.*, 2009) et (Habash et Sadat, 2006)).

2.1 Les variétés de la langue arabe

La langue arabe se distingue des autres langues par le fait que la langue écrite est différente de la langue parlée. On distingue généralement l'arabe dialectal et l'arabe standard moderne (ASM). En effet, l'ASM est le langage officiel utilisé par toutes les populations de culture arabe. Il est utilisé dans les manuels scolaires, les médias, les journaux, etc. Par contre, l'arabe dialectal diffère d'un pays à l'autre (Gahbiche-Braham, 2013).

2.1.1 L'arabe dialectal

Malgré que tous les pays de culture arabe aient en commun l'usage de l'ASM pour l'éducation, les manuels officiels et les médias (Habash, 2010), chaque pays a son dialecte propre. En effet, les dialectes arabes varient d'un pays à un autre ; ils dépendent surtout de l'histoire de chaque pays et de son emplacement géographique. L'arabe dialectal varie même d'une région à l'autre à l'intérieur d'un même pays, présentant parfois des mots et des prononciations différentes. Selon (Habash, 2010), on distingue les dialectes arabes suivants : l'*égyptien* (dialectes égyptiens et soudanais), le *levantin* (dialectes libanais, syrien, jordanien et palestinien), l'*arabe des pays du Golfe* (dialectes du Koweït, des Émirats Arabes Unis, du Bahrein, de l'Arabie Saoudite et du Qatar), le *maghrébin* (dialectes marocain, algérien, tunisien, libyen et mauritanien), l'*irakien*, le *yéménite* et le *le maltais*, qui diffère des autres dialectes par son utilisation des caractères romains (Sadat *et al.*, 2014a).

Face à l'évolution des moyens de communication et à l'émergence rapide des médias sociaux, l'arabe dialectique est de plus en plus mis à l'écrit dans les com-

munications, les SMS, les forums virtuels, etc. Récemment, l'arabe dialectique a davantage été mis à l'écrit à l'aide des caractères latins (arabe translitéré), surtout si l'utilisateur n'a pas l'habitude du clavier arabe (Diab et Habash, 2007). Cependant, le dialecte reste de l'arabe informel, même s'il est écrit avec des lettres arabes. Aussi, les dialectes arabes sont influencés par des langues comme le français ou l'anglais. Cette influence s'inscrit dans un contexte historique post-colonisation.

2.1.2 L'arabe standard moderne

L'arabe standard moderne (ASM), appelé aussi arabe contemporain ou formel, est la forme écrite de l'arabe. Contrairement à l'anglais ou au français, l'arabe s'écrit de droite à gauche. Aussi, il n'existe pas de distinction entre les lettres majuscules et les lettres minuscules dans l'alphabet arabe.

En effet, les lettres arabes changent de forme en fonction de leur position dans le mot (début, milieu, fin, isolée). Le tableau 2.2 montre la lettre « ت [t] » et ses différentes représentations dans le mot.

Tableau 2.2 Différentes graphies de la lettre [t] selon sa position dans un mot en arabe

Début	Milieu	Fin	Isolée
تـ	ـتـ	ـت ou ـة	ت ou ة

2.1.3 La voyellation de l'arabe standard moderne

Les voyelles (ou diacritiques) sont rarement notées en arabe. Elles le sont dans les ouvrages didactiques ou religieux comme le Coran, pour lever des ambiguïtés sémantiques et/ou syntaxiques. Par exemple, la voyellation de la phrase « كتب الولد [ktb Alwld] » permet de dire si le sens de la phrase est « كَتَبَ الولَدُ [ka-

taba Alwladu] (en français : le garçon a écrit) » ou « كُتُبُ الوَلَدِ [kkutubu Alwldi] (en français : les livres du garçon) » (Gahbiche-Braham, 2013).

En général, dans les textes écrits comme les journaux, les documents administratifs, les livres et les magazines, l'arabe n'est pas voyellé. Les signes diacritiques peuvent être placés au-dessus ou au-dessous des lettres. Voici les différents types de signes de diacritiques :

— les voyelles brèves sont marquées à l'aide de signes diacritiques placés au dessus ou au-dessous de la lettre ;

— les voyelles longues sont marquées à l'aide de l'une des trois lettres consonnes « *ALEF* », « *WAW* » ou « *YEH* » placée après la lettre à voyeller ;

— Les signes de syllabation, qui sont le « *sukun* » et le « *chadda* » au dessus des lettres ;

— Le « *tanwiin* », qui est le dédoublement d'une voyelle courte en fin de mot.

2.1.4 La structure des phrases en arabe

En arabe, il y a deux types de phrases : la phrase verbale et la phrase nominale. La phrase verbale est constituée d'un verbe suivi d'un sujet. Optionnellement, elle peut contenir aussi un complément. Elle est la plus utilisée dans l'expression courante en arabe et sert à indiquer une action ou un évènement.

La phrase nominale ne contient pas de verbe. En arabe, le verbe n'est pas obligatoire pour construire une phrase. La phrase nominale est constituée d'un sujet et d'un attribut, qui peut être un adjectif qualificatif, un complément circonstanciel, un complément d'objet, etc. Par exemple, la phrase « الطقس جميل [AlTqs jmyl] (en français : il fait beau) », est une phrase nominale qui ne contient pas de verbe. Elle est constituée d'un sujet et d'un adjectif.

2.2 La morphologie de la langue arabe

Dans le lexique de la langue arabe, on distingue trois catégories principales de mots : les verbes, les noms et les particules. Les noms et les verbes sont généralement dérivés d'une racine à trois consonnes. En appliquant différents schémes sur la racine, on peut générer plusieurs mots sémantiquement différents. Pour cette raison, l'arabe est considéré comme une langue à racine réelle (Ghoul, 2011).

Tableau 2.3 Exemple de schèmes appliqués sur un mot en arabe (Ghoul, 2011).

Schème	Mot	Traduction en français
فَعل	عَمل [Eaml]	travail
فاعِل	عامِل [EAmil]	ouvrier
فَعَلَ	عَمَلَ [Eamala]	a travaillé
مَفعل	مَعمل [maEml]	atelier
فُعِلَ	عُمِلَ [Eumila]	a été travaillé
مَفعول	مَعمول [maEmwl]	applicable

Le tableau 2.3 donne quelques dérivations du verbe « عمل [Eml] (en français : travailler) » par l'application de différents schèmes sur la même racine. Un schème en arabe est représenté par les adjonctions et les manipulations de la racine pour obtenir différents mots. Pour l'exemple décrit au tableau 2.3, à partir de la racine et en appliquant différents schèmes, on génère une famille de mots de sens différents.

En effet, les mots en arabe sont des mots agglutinés qui se composent d'enclitiques (ou postfixes), de suffixes, de préfixes et de proclitiques (ou antéfixes) qui sont ajoutés autour de la forme de base constituant la racine du mot (Habash et Sadat, 2012). Une liste exhaustive des différents affixes de la langue arabe (préfixe,

proclitique, suffixe et enclitique) a été présenté dans les travaux de Kadri et Nie (2006).

Sachant que la phrase en arabe s'écrit de droite à gauche, on peut segmenter un mot en arabe comme présenté dans la figure 2.4.

Tableau 2.4 Structure générale d'un mot en arabe

Enclitique	Suffixe	**Racine**	Préfixe	Proclitique

Nous présentons les différents segments d'un mot arabe comme suit :

— La *racine* est la forme de base pour les mots. Elle est généralement formée par trois consonnes (trilitères), mais certaines sont formées de deux consonnes (bilitères) ou de quatre consonnes (quadrilitères). En ajoutant la voyellation ou les affixes à la racine, nous obtenons plusieurs mots de différents sens. Les mots formés suivent des schèmes différents (Attia, 2008).

— Le *préfixe* peut être de type verbal ou nominal. Prenons l'exemple du mot « كتب [ktb] (en français : écris) » : si on y ajoute le préfixe « ي [y] », on aura le verbe conjugué « يكتب [yktb] (en français : il écrit) », tandis qu'avec l'ajout du préfixe nominal « م [m] », on obtiendra le mot « مكتب [mktb] (en français : bureau) » indiquant le nom d'un lieu.

— Le *proclitique* peut être attaché ou détaché du mot qui le suit. Ce peut être une préposition ou une conjonction.

— Le *suffixe* exprime généralement la marque du pluriel et du féminin.

— L'*enclitique* est un pronom personnel attaché à la fin de la forme de base. À ce propos, un trait particulier de la langue arabe est qu'elle contient des pronoms qui expriment à la fois le masculin et le féminin.

Prenons l'exemple de segmentation du mot «أتكلموننا [>tklmwnnA] (en français la phrase : Est-ce que vous parlez à nous ?)».

Tableau 2.5 Exemple de segmentation d'un mot en arabe

نا	ون	كلم	ت	أ
Enclitique	Suffixe	Racine	Préfixe	Proclitique
nous	vous	parler	-	Est-ce que

Cet exemple montre bien la richesse morphologique de la langue arabe et la nécessité d'une étape de segmentation des mots arabe dans les applications en TALN, comme la traduction automatique depuis et vers l'arabe.

2.3 Catégories d'un mot en arabe

Pour la langue arabe, nous distinguons trois catégories principales de mots : le verbe, le nom et la particule (Ghoul, 2011).

Le verbe

Comme mentionné précédemment, la plupart des mots en arabe, y compris les verbes, sont dérivés d'une racine de deux, trois ou quatre consonnes. La conjugaison des verbes en arabe dépend des traits flexionnels suivants :

- L'aspect temps : l'accompli (qui correspond au passé en français) et l'inaccompli (qui correspond au présent ou le futur en français).
- Le nombre : le sujet peut être singulier, pluriel ou duel.
- Le genre : le sujet peut être masculin ou féminin.
- La personne : première, deuxième ou troisième (comme en français).
- Le mode : actif ou passif.

La position du verbe dans une phrase en arabe peut être au tout début de la phrase, au milieu ou aussi à la fin de la phrase.

Le nom

Il existe deux familles de noms en arabe : les noms verbaux qui sont dérivés d'une racine, et les noms qui ne le sont pas, comme les noms propres et les noms communs. Les noms sont très variés et s'écrivent selon différentes règles selon les cas, en ajoutant des morphèmes spécifiques qui sont de quatre familles :

— Le féminin singulier : pour transformer un nom masculin en féminin, on ajoute la lettre « ة [t] » à la fin du mot. Par exemple, le mot masculin « واسع [wAsE] (en français : large) » devient « واسعة [wAsEp] (en français : large) » au féminin.

— Le féminin pluriel : pour obtenir le nom féminin pluriel, on ajoute les deux lettres « ات [At] » à la fin du mot. Par exemple, « معلم [mElm] (en français : enseignant) » devient « معلمات [mElmAt] (en français : enseignantes) ».

— Le masculin pluriel : selon la position du mot dans la phrase (avant ou après le verbe), on ajoute les lettres « ين [yn] » et « ون [wn] » à la fin du mot. Par exemple, « معلم [mElm] (en français : enseignant) » devient « معلّمون [mEl mwn] » ou « معلّمين [mEl myn] (en français : enseignants) »).

— Le pluriel irrégulier : ce type suit des règles très complexes. Les lettres qu'on ajoute à la racine peuvent être placées au début, au milieu ou à la fin du mot. Par exemple, « طِفل [Tifl] (en français : un enfant) » devient « أطفال [>TfAl] (en français : des enfants) ». Ou encore, le mot « فصل [fSl] (en français : une saison) » devient « فُصول [fuSwl] (en français : des saisons) ».

La particule

Les particules jouent un rôle très important dans l'interprétation de la phrase en arabe. En effet, elles représentent les mots qui expriment des faits ou des choses en relation avec le lieu (particules spatiales comme par exemple « حيث [Hyv] (en

français : où) ») ou le temps (particules temporelles comme par exemple « بعد [bEd] (en français : après) », «قبل [qbl] (en français : avant)» et «مُنذُ [mun*u] (en français : pendant)»). De plus, elles assurent la cohérence et l'enchaînement du texte en arabe. Ce sont principalement des conjonctions de coordination et de subordination. On en distingue plusieurs types, tels que l'introduction, l'explication et la conséquence (Ghoul, 2011).

Les particules peuvent aussi exprimer des pronoms relatifs, par exemple « الذي [Al*y] (en français : cette) » et « الّذين [Al *yn] (en français : ceux) ».

2.4 La langue arabe et les médias sociaux

Ces dernières années, avec les évènements socio-politiques qui ont marqué le monde arabe, les gens qui maîtrisent cette langue sont devenus de plus en plus actifs sur les réseaux sociaux, pour exprimer leurs points de vue ou pour chercher et lire des nouvelles. Les données publiées en arabe ont contribué à plusieurs travaux de recherche.

La richesse de la structure des mots arabes peut engendrer trois types d'ambiguïtés au niveau de son traitement automatique : l'ambiguïté morphologique, l'ambiguïté syntaxique et l'ambiguïté orthographique. Comme présenté dans les parties précedentes de ce chapitre, la langue arabe est une langue naturelle qui présente plusieurs défis pour les applications de TALN en général et de TA en particulier. Parmi les défis, les mots en arabe peuvent avoir plusieurs sens dépendemment de la voyellation accordé au lettres arabes dans une phrase. Aussi, on remarque que dans les phrases il est probable que le sujet soit abscent dans une phrase, comme par exemple la phrase « كلنا أبطال » [klnA AbTl] (en français : Nous sommes tous les héros). Face à une telle langue morphologiquement riche le sphère des problèmes à traiter dans les applications de TALN sont plus large

que d'autre langues comme l'anglais ou le français. En effet, le sphère de ces problèmes s'élargit beaucoup avec l'arabe tel qu'écrit dans les médias sociaux. Ces problèmes sont souvent dûs aux écritures non standard qui caractérisent ce genre de textes. En effet, les utilisateurs des médias sociaux tendent à commettre des malformations orthographiques ; ils ont tendance à écrire des mots en arabe en utilisant l'alphabet latin et des chiffres. Aussi, chaque utilisateur translittère le mot à sa façon. Ce phénomène est appelé « Arabizi » (Bies *et al.*, 2014; Darwish, 2014; Adouane *et al.*, 2016).

Ainsi, la lettre arabe « ح [H] » est souvent translitérée par le nombre « 7 », la lettre « ق [q] » par le nombre « 9 », etc. Par exemple, les utilisateurs de Twitter écrivent le nom propre « احمد [AHmd] » comme étant *ahmed*, *ahmad*, *ahmd*, *a7mad*, *a7med*, *a7mmd*, *a7md* ou *ahmmd*. Ils écrivent aussi le nom propre « أشرف [>$rf] » comme étant *ashraf*, *ashref*, *ashrf*, *shrf*, *achraf*, ou *aschraf* (Mubarak et Abdelali, 2016).

Les utilisateurs des médias sociaux s'expriment en alternant entre leur dialecte et d'autres langues. Ce phénomène est appelé *code switching* et il a été abordé par (Barman *et al.*, 2014). Ce problème nuit beaucoup aux applications qui traitent une seule langue.

Par ailleurs, les noms des utilisateurs sur Twitter ne sont pas écrits de la même manière et ils sont souvent composés de deux mots ou plus, séparés par un caractère spécial, comme le trait de souligement (_). Les noms d'utilisateurs ou *usernames* commencent toujours par le symbole (@). Leur traitement présente plusieurs défis pour les applications en TALN (Mubarak et Abdelali, 2016).

L'utilisation de l'arabe dialectal dans les médias sociaux est un problème reconnu des chercheurs en TALN, surtout que les outils et les ressources linguistiques pour traiter les dialectes arabes sont peu disponibles. Le traitement d'un dialecte arabe est plus difficile au niveau de la translittération. Par exemple, la phrase en

arabe moderne standard « لا يلعب [lA ylEb] (en français : il ne joue pas) » peut être écrite en dialecte égyptien « مابيلعبش [mAbylEb$] », « مايلعبش [mAylEb$] », « ميلعبش [mylEb$] », « مابلعبش [mAblEb$] », ce qui est translittéré par les utilisateurs de médias sociaux de différentes façons : *mayel3absh, mabyelaabsh, mabyel3absh, etc* (Darwish, 2014).

Nous avons présenté dans ce chapitre les caractéristiques de la langue arabe et les défis que pose son traitement dans le domaine du TALN. Ces défis sont plus nombreux pour l'arabe tel qu'employé dans les médias sociaux. Dans ce travail, nous nous sommes concentrés sur les phases de prétraitement pour le couple de langues arabe/anglais. Au prochain chapitre, nous présentons une revue de littérature des travaux qui traitent des problèmes de prétraitement et de traduction des textes issus des médias sociaux en général et de la langue arabe en particulier.

CHAPITRE III

ÉTAT DE L'ART

La traduction automatique depuis et vers l'arabe a fait l'objet de plusieurs travaux de recherche. Ces travaux se sont principalement focalisés sur le prétraitement de cette langue dans le but d'améliorer la qualité des systèmes de traduction automatique. Comme nous l'avons vu, de nos jours, la traduction n'est plus concernée seulement par l'arabe standard moderne (ASM), mais aussi par d'autres expressions de cette langue, en particulier l'arabe dialectal, qui est le plus utilisé sur les médias sociaux (Twitter, Facebook, etc.). C'est pour cette raison que les travaux portant sur la traduction automatique s'intéressent de plus en plus à cette forme de la langue arabe.

Dans ce chapitre, nous ferons un survol des travaux réalisés sur la traduction automatique en langue arabe. Ensuite, nous présenterons quelques travaux portant plus précisément sur la traduction des dialectes arabes, pour enfin traiter de travaux qui s'intéressent à la traduction et au prétraitement de contenus issus des médias sociaux.

3.1 Revue de littérature

La traduction automatique de l'arabe et en général celle des langues présentant une morphologie riche et complexe est une tâche très difficile. Cette particularité

a incité les chercheurs à étudier l'apport du prétraitement des textes avant la traduction, à savoir les étapes de segmentation et de normalisation. La segmentation consiste à séparer la racine des mots des affixes qui l'entoure. Au chapitre précédent nous avons présenté phénomène pour la langue arabe. La traduction automatique de l'arabe et en général celle des langues présentant une morphologie riche et complexe est une tâche très difficile. Cette particularité a incité les chercheurs à étudier l'apport du prétraitement des textes avant la traduction, à savoir les étapes de segmentation et de normalisation. La segmentation consiste à séparer la racine des mots des affixes qui l'entoure. Au chapitre précédent nous avons présenté phénomène pour la langue arabe.

Les travaux de recherche qui rendent compte du prétraitement de langues sources à morphologie complexe comme l'arabe (Habash et Sadat, 2012), le tchèque (Goldwater et McClosky, 2005) ou l'allemand (El-Kahlout et Yvon, 2010), indiquent que cette méthode contribue à l'amélioration de la performance des systèmes de traduction automatique.

En particulier, l'étape de segmentation des mots est très importante, car elle permet, d'une part, de réduire le nombre de mots inconnus et la taille du vocabulaire à traiter et, d'autre part, d'améliorer la qualité de l'alignement. Lee (2004); Habash et Sadat (2012); Al-Haj et Lavie (2012) ont prouvé par leurs travaux que la segmentation des mots pour les phrases dans la langue source pour la traduction automatique statistique donne de meilleurs résultats.

Habash et Sadat (2006) ont distingué différents niveaux de segmentation des mots arabes, à partir desquels ils ont défini onze schèmes de prétraitement (S1, ON, D1, D2, D3, WA, TB, MR, L1, L2 et EN). Aussi, selon Larkey *et al.* (2002), il est préférable que la segmentation des textes en arabe, soit l'étape de prétraitement, ait lieu avant l'étape de la traduction.

Attia (2008) a implémenté un segmenteur pour l'arabe à base de règles, qui a été utilisé plus tard pour la traduction automatique. L'avantage de cette implémentation est que les règles peuvent être facilement améliorées et modifiées. Plusieurs schèmes de segmentation ont été ensuite testés à différents niveaux linguistiques.

D'autres défis de la traduction automatique de l'arabe ont été relevés dans la littérature, entre autres la structure syntaxique de la phrase (Hebresha et Ab Aziz, 2013; Hassan et Darwish, 2014). En effet, les phrases en arabe sont généralement des phrases verbales dont la structure est *verbe+sujet+objet* (VSO), alors que la phrase dans les langues latines comme le français ou l'anglais présente une structure *sujet+verbe+bbjet* (SVO).

Bisazza et Federico (2010) ont déjà proposé une technique qui identifie automatiquement les verbes en arabe à clause initiale et les élimine ensuite du corpus parallèle. Cette méthode a été appliquée pour le pré-traitement des données d'entraînement et le calcul des statistiques sur le mouvement des verbes dans le corpus. Ces statistiques ont été employées à l'étape de réordonnancement des verbes dans les phrases verbales pour la traduction de l'arabe vers l'anglais.

Dans le même cadre, Carpuat *et al.* (2012) ont proposé une méthode de ré-ordonnancement des phrases verbales en arabe en phrases nominales lors de l'étape d'alignement, pour le développement d'une application en traduction automatique statistique. Les auteurs n'ont pas pu résoudre le problème de ré-ordonnancement pour tous les verbes. Malgré cela, la traduction a été légèrement améliorée.

D'autres travaux ont traité la traduction de l'arabe, en particulier de l'arabe vers l'anglais. Abuelyaman *et al.* (2014) ont proposé une méthode de traduction pour cette paire de langues (arabe vers l'anglais). Dans leurs travaux, ces auteurs insistent sur les défis à relever pour la traduction automatique des deux langues traitées, tout en présentant les différences linguistiques entre l'arabe et l'anglais.

Parmi les différences citées, on trouve la structure de la phrase, les mots qui comportent plusieurs sens pour une même graphie, les lettres ambiguës, etc. Dans la même veine, Hailat *et al.* (2013) insistent sur la traduction de l'arabe vers l'anglais et de l'anglais vers l'arabe. Ils ont évalué l'efficacité de deux traducteurs populaires, Google Translate et Babylon pour ces deux paires de langues, en comparant leurs traductions avec des traductions humaines.

Dernièrement, la traduction neuronale est de plus en plus utilisé. C'est dans ce cadre que Almahairi *et al.* (2016) ont testé la traduction d'un corpus du domaine différent de celui du corpus d'entraînement avec deux types de systèmes de traduction : un système statistique à base de segments et un système basé sur les réseaux de neurones. Les résultats du score BLEU avec les deux systèmes ont donné des résultats de traduction beaucoup plus faibles pour toutes les expériences que ceux obtenus lorsque le corpus test est du même domaine que le corpus d'entraînement.

Récemment, et avec la facilité d'accès à Internet, les recherches ne sont plus basées uniquement sur l'ASM, mais aussi sur l'arabe dialectal, qui est le plus utilisé sur les médias sociaux. En effet, le traitement automatique de l'ASM est différent du traitement automatique de l'arabe dialectal. À ce titre, Salloum et Habash (2012) ont proposé un système de traduction automatique du dialecte arabe vers l'arabe moderne standard. Aussi, Sadat *et al.* (2014b), dans le cadre du projet ASMAT (*Arabic Social Media Analysis Tools*), ont construit un système de traduction hybride (statistique et à base de règles) pour la traduction du dialecte tunisien vers l'arabe standard et vers le français.

Après le printemps arabe[10], les chercheurs ont pris davantage conscience de la fluidité des données en dialecte arabe sur les médias sociaux. Zbib *et al.* (2012)

---

10. http ://www.ledevoir.com/international/actualites-internationales/315749/le-printemps-arabe

ont développé un système de traduction statistique de dialectes arabes (égyptien et levantin) vers l'anglais. Pour ce faire, il était très important de construire deux corpus parallèles égyptien-anglais et levantin-anglais de 1.1 millions de mots. Les données en dialecte ont été collectées sur des forums de discussion et sur Facebook et ont ensuite été traduites en utilisant le traducteur libre Amazon Mechanichal Turk [11]. Le score BLEU (Papineni *et al.*, 2002) mesuré sur les données traduites s'est révélé plus élevé que pour le système de traduction de l'arabe standard vers l'anglais.

Sajjad *et al.* (2013) ont proposé un système de traduction automatique statistique du dialecte égyptien vers l'anglais. Contrairement aux chercheurs précédemment cités, ils ont utilisé l'ASM comme langue pivot. Ils ont ainsi appliqué un modèle de transformation automatique qui rapproche le dialecte de l'arabe standard, puis ils ont évalué leur système de traduction automatique du dialecte depuis et vers l'anglais.

Toutefois, la non disponibilité de corpus parallèles des médias sociaux reste toujours un défi pour les applications en TA, peu importe la langue traitée. Ce manque de ressources pose des problèmes, que ce soit pour la couverture du vocabulaire ou pour le choix des structures syntaxiques des phrases à traduire. De plus, les textes issus de domaines spécialisés, comme ceux tirés des microblogues, sont en général plus difficiles à traduire automatiquement (Langlais *et al.*, 2006).

Dans ses travaux, afin de pallier le problème de manque de corpus parallèle, Jehl (2010) a proposé une méthode de traduction des *tweets* qui utilise un corpus parallèle hors domaine et un grand modèle de langue des *tweets* dans la langue cible. Le travail a commencé par la création d'un petit corpus parallèle anglais-allemand de 1000 *tweets* et l'extraction de leurs caractéristiques linguistiques,

11. http ://www.mturk.com/

afin de déterminer les défis de la traduction automatique statistique pour ce type de textes. Les expériences réalisées ont prouvé que le score BLEU est meilleur pour un domaine restreint que pour un domaine général. Aussi, ce score varie selon la nature des mots utilisés dans les *tweets*, soit des mots spécifiques au domaine traité (*smiles*, expressions de «chat», abréviations, etc.), des mots en langue étrangère, des noms propres ou des expressions particulières de Twitter, comme les *retweets*, les noms d'utilisateurs (*usernames*) et les *hashtags*. Les mots inconnus qui ne sont couverts ni par le corpus parallèle, ni par le modèle de langue, ont biaisé les mesures du score BLEU. En effet, toutes les expériences de Jehl (2010) étaient basées sur le système de traduction statistique Moses (Koehn *et al.*, 2007). Ils ont ainsi prouvé que la longueur des *tweets* (qui ne dépassent pas 140 mots) facilite beaucoup la tâche des systèmes de traduction statistique, alors que les caractéristiques linguistiques des données qui sont issues de Twitter, ont fait baisser la performance de leurs système de traduction.

En effet, plusieurs autres chercheurs ont étudié et analysé la masse de données bilingues publiée sur Twitter. Carrera *et al.* (2009) ont proposé un système de traduction hybride (statistique et à base de règles) destiné au traitement des données des médias sociaux. Après leurs études des *tweets*, Jehl *et al.* (2012) ont remarqué que les abonnés de ce média s'expriment en utilisant deux langues ou plus. Ainsi, les *tweets* publiés le sont souvent en plus d'une langue. En particulier, les *tweets* relatifs au printemps arabe sont souvent publiés à la fois en arabe et en anglais. Ces auteurs ont donc proposé une méthode de traduction automatique des *tweets* basée sur la recherche d'information translinguistique (*Cross-Lingual Information Retrieval [CLIR]*). L'idée principale était de lancer des requêtes, qui sont des *tweets* en anglais, pour chercher leur traduction en arabe. Le but de cette approche était de réduire le nombre de mots non traduits (les mots OOV) pour les applications de TAS des *tweets*. Ce sont généralement des mots hors vocabulaire,

qui figurent dans la table de traduction produite par le système de traduction et qui nécessitent un prétraitement particulier.

Ling (2015) a aussi envoyé les données bilingues qui existent dans les microblogues. En premier lieu, il a procédé par une extraction d'un corpus parallèle des microblogues Twitter et Weibo[12]. En deuxième lieu, il a utilisé ces données pour entraîner un système de traduction statistique à base de segments en utilisant le décodeur Moses (Koehn *et al.*, 2007). Le meilleur résultat du BLEU (Papineni *et al.*, 2002) a été de 21.5, en utilisant l'ensemble des tests formés des données NIST 2012 (chinois-anglais) et des données issues du microblogue Weibo.

Gotti *et al.* (2014) ont proposé une méthode de traduction statistique des *tweets* publiés par les agences et les organisations gouvernementales canadiennes. Selon ces chercheurs, un tel système serait très utile pour les 150 agences canadiennes qui possèdent un site sur les réseaux sociaux comme Twitter et dont les abonnés ne sont pas tous des Canadiens. Pour le ré-ordonnancement des poids de leur système de traduction, ils ont utilisé un corpus hors-domaine. Les résultats obtenus ont montré que 37% des mots hors vocabulaire sont des *hashtags*. Ces mots sont généralement des expressions composées de plusieurs mots séparés par un caractère spécial et qui commencent par un «#». Selon leur étude, 20% des *hashtags* du corpus utilisé nécessitent une segmentation avant de passer à l'étape de traduction.

Toral *et al.* (2015) ont essayé de développer un système de traduction statistique plus performant pour les *tweets*. Leur but était de déterminer la meilleure combinaison d'outils, de techniques et de ressources disponibles pour la traduction statistique de ces données. Les paires de langues étudiées étaient l'espagnol du/vers le catalan, l'espagnol du/vers le basque et l'espagnol du/vers le portugais. Ils ont

12. http ://www.weibo.com/

ainsi détaillé les résultats de traductions réalisées avec les outils de traductions en licence libre que sont Moses (Koehn *et al.*, 2007), cedec (Dyer *et al.*, 2010) et le système de traduction à base de règles Apertium (Forcada *et al.*, 2011), uniquement pour les paires de langues espagnol/catalan, espagnol/portugais et du basque vers l'espagnol. Enfin. l'outil Matxim (Mayor *et al.*, 2011) a été utilisé pour la paire de langues espagnol/basque. Pour comparer les sorties des différents systèmes de traduction, les métriques BLEU (Papineni *et al.*, 2002) et TER (Snover *et al.*, 2006) ont été utilisées. Les conclusions de ces chercheurs sont que la combinaison d'outils de traduction donne de meilleurs résultats pour certaines paires de langues, mais pas pour d'autres.

Récemment, la traduction des données issues des médias sociaux a prouvé son importance pour les tâches d'analyse des sentiments. Salameh *et al.* (2015); Refaee et Rieser (2015); Mohammad *et al.* (2016) ont étudié l'effet de la traduction automatique de ces données pour une application d'analyse des sentiments exprimés en arabe sur les médias sociaux. Les auteurs ont profité des ressources et des outils en analyse des sentiments pour la langue anglaise, en traduisant d'abord leurs données de l'arabe vers l'anglais, puis en procédant en second lieu à l'analyse des sentiments des données traduites. Les résultats issus de leur méthode sont généralement meilleurs que les autres méthodes d'analyse des sentiments des données en arabe issues des médias sociaux, sans passer par l'étape de traduction.

La demande pour des traducteurs automatiques de textes issus des médias sociaux ne cesse de s'accroître, surtout face à la propagation croissante des contenus générés par les utilisateurs dans différentes langues. Cependant, la qualité des textes publiés pose plusieurs défis pour le domaine du TALN (Liu *et al.*, 2012). À ce sujet, plusieurs travaux ont insisté sur l'effet de la normalisation de ces données sur la traduction automatique. La normalisation revient à «standardiser» les mots hors vocabulaire pour qu'ils puissent être reconnus et traités par les outils de TALN.

Wang et Ng Hwee (2013) ont proposé un décodeur pour la normalisation des textes des microblogues pour l'anglais et le chinois. Ils ont traité les problèmes causés par la ponctuation fautive ainsi que celui des mots manquants. Pour toutes les expériences réalisées, le système de traduction après la normalisation a présenté un score BLEU plus élevé que celui du système de base. Plusieurs autres travaux de normalisation ont été cités par Ling (2015).

Dans la littérature, il existe aussi plusieurs travaux qui ont traité des défaillances de la langue arabe telle qu'utilisée dans les médias sociaux. Ces travaux ont abordé, entre autres, le problème des mots arabes écrits en caractères latins (Darwish, 2014). Pour certaines applications de traitement des langues naturelles, il est nécessaire de fixer une liste de mots vides, ce qui n'est pas évident pour des données bruitées issues des médias sociaux. Ce problème a été traité par Medhat *et al.* (2014).

Dans ce chapitre, nous avons présenté les différents travaux de traduction automatique de la langue arabe avec ses différentes variétés. Nous avons recensé différents travaux de traduction de textes issus de microblogues comme Twitter. Enfin, nous avons clôturé notre survol de la littérature en présentant quelques travaux réalisés sur la normalisation des données issues des médias sociaux. Au prochain chapitre, nous proposons notre méthode pour la traduction automatique des *tweets* de l'arabe vers l'anglais.

CHAPITRE IV

MÉTHODOLOGIE

Au présent chapitre, nous présentons notre approche pour la traduction des *tweets* de l'arabe vers l'anglais. Nous détaillons en premier lieu les étapes de notre méthodologie. Ensuite, nous présentons les étapes de notre collecte de données dans des corpus de *tweets* ainsi que les outils utilisés pour la réaliser. Nous détaillons aussi brièvement les caractéristiques des textes publiés sur Twitter par rapport aux autres microblogues. Enfin, nous insistons sur l'importance de l'étape de prétraitement des données pour les deux langues traitées et nous détaillons les différentes stratégies adoptées pour l'entrainement des modèles de traduction.

4.1 Étapes de traduction des *tweets* arabes vers l'anglais

Les méthodes statistiques ont prouvé leur robustesse grâce à la disponibilité d'une grande masse de données bilingues et monolingues. Ces données sont indispensables pour l'entraînement des modèles de traduction et des modèles de langues (Brown *et al.*, 1990). Il est aussi préférable que les corpus d'entraînement soient du même domaine que les données de test ; on entend par *corpus test* les données à traduire dans la langue source. Ainsi, si le corpus parallèle utilisé à l'étape d'entraînement est du domaine des nouvelles politiques (*news*), nous aurons une meilleure qualité de traduction en utilisant un corpus issu de ce domaine (Langlais *et al.*, 2006).

Dans nos travaux, pour la réalisation d'un traducteur de *tweets*, nous avons besoin d'un grand corpus parallèle pour ce type de données. Ainsi, des efforts considérables ont été investis dans plusieurs travaux recensés par la littérature afin de collecter ce type de données. Cependant, ces corpus ne sont pas disponibles pour le public et leur collecte est confrontée à plusieurs problèmes en ce qui concerne la recherche d'information (Jehl *et al.*, 2012; Ling, 2015). Face à cet obstacle, notre méthodologie revient donc à adopter une méthode de traduction statistique pour le domaine des microblogues et en particulier pour les *tweets*, mais en utilisant un corpus parallèle en ASM et non pas pour les *tweets*. Cette méthodologie est inspirée de (Jehl, 2010), qui a essayé dans ses travaux d'adapter un système de traduction classique pour la traduction des *tweets* pour la paire de langues allemand/anglais.

Les principaux modèles sur lesquels repose notre système de traduction pour les *tweets* sont le modèle de traduction et le modèle de langue. Le modèle de traduction a été entraîné à partir d'un grand corpus parallèle arabe/anglais pour un domaine général ou hors domaine. Ensuite, nous avons tenté d'adapter notre système avec l'utilisation d'un modèle de langue pour les *tweets* en anglais ; le modèle de langue est donc du même domaine que les données à traduire. Enfin, pour le décodage, nous avons utilisé un petit corpus parallèle de *tweets* arabe/anglais que nous avons collecté et préparé. La figure 4.1 donne un aperçu général de notre méthodologie.

Pour les applications en traduction automatique statistique (TAS), la collection et le prétraitement des données sont des tâches très importantes. La qualité des données influence beaucoup la performance des systèmes de TA.

Aux sections suivantes, nous détaillons les étapes dont nous avons convenu pour collecter et préparer nos données.

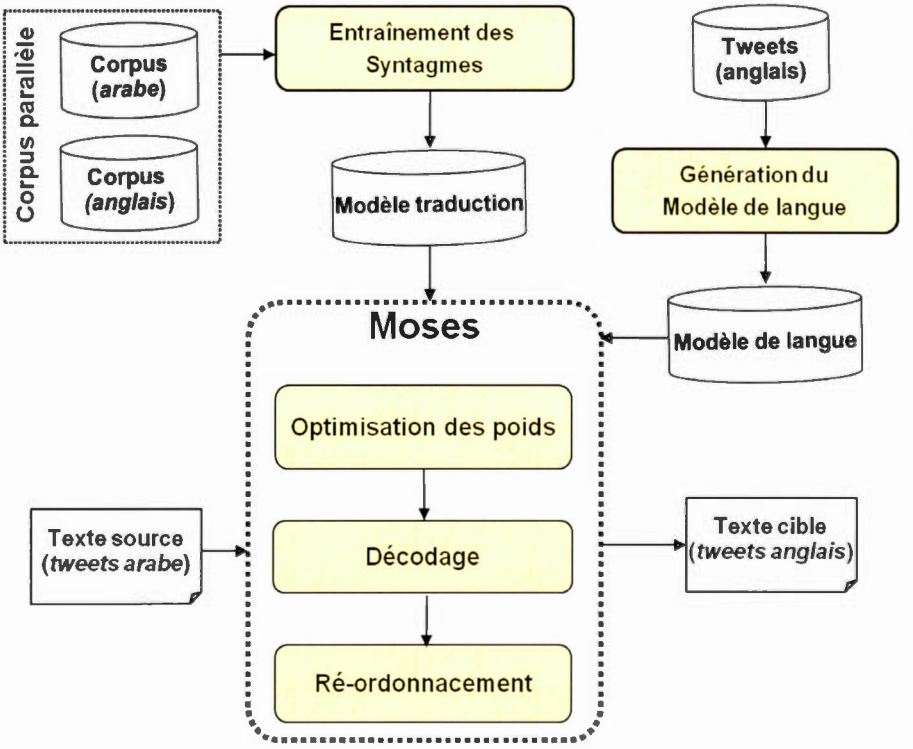

Figure 4.1 Les différentes étapes de la traduction automatique statistique pour les *tweets* de l'arabe vers l'anglais

4.2 La collecte des données

Si les médias sociaux sont marqués par la diversité, le microblogue Twitter est parmi les plus fréquentés par jour en termes de nombre de personnes actives. En plus, l'accessibilité des *tweets* publiés par les abonnés au public en temps réel a fait de ce média social une source de donnée très importante pour les chercheurs qui s'intéressent à l'analyse et au traitement des contenus de microblogues (Farzindar et Roche, 2013).

Pour nos travaux, nous avons besoin d'un grand corpus de *tweets* en anglais pour la construction du modèle de langue, et d'un petit corpus en arabe pour tester le système de traduction. Nous avons donc opté pour une méthode de collecte

automatique. Pour ce faire, plusieurs interfaces de programmation pour Twitter APIs et des librairies sont disponibles sous licence libre. Nous avons donc implémenté un programme Java en utilisant la librairie Twitter4j [13] afin de collecter nos données. Nous avons choisi le *STREAMING APIs* [14] pour extraire les *tweets* et ainsi construire un corpus d'entraînement et de test. Nous avons lancé des requêtes contenant des mots-clés portant sur l'actualité des pays arabes, par exemple les mots (سوريا ، حلب، بشار، ثورة، الربيع العربي، الجزيرة) [*swryA, Hlb, b$Ar, vwrp, AlrbyE AlErby, Aljzyrp*] (en français : Syrie, Halab, Bachar, révolution, printemps arabe, Aljazeera). Les mots-clés qui ont été utilisés pour la cueillette des *tweets* en arabe ont été ensuite traduits en anglais pour collecter les *tweets* en anglais qui correspondent aux mots-clés (*Syria, Halab, Bachar, revolution, arab spring, AlJazeera*). La collecte été durant la période de septembre à novembre 2015. Le nombre de *tweets* collectés est représenté au tableau 4.1.

Tableau 4.1 Nombre de *tweets* collectés

Langue	Nombre des *tweets*
anglais	255 602
arabe	1 930

Notons que les deux corpus ont pris beaucoup de temps à construire et à filtrer. La tâche de collecte et de préparation de données est la plus lourde en termes de temps pour les applications en TALN, en particulier en TA. Généralement, les corpus de *tweets* pour les deux langues traitées, surtout pour l'arabe, sont payants et/ou non disponibles pour le public. Des efforts considérables ont donc été investis pour les préparer manuellement.

13. http://twitter4j.org/en/index.html

14. https://dev.twitter.com/streaming/public

4.3 Le prétraitement des données

Les utilisateurs de Twitter et des microblogues en général utilisent un style d'écriture incompréhensible par la machine et commettent souvent des fautes d'orthographe et de grammaire. À la place des mots normalisés, ils utilisent des abréviations (par exemple : " *just4u*" au lieu de "*just for you*") et étirent les mots (par exemple : "*haaaaappy*" au lieu de "*happy*") pour exprimer leurs sentiments. Ces ambiguïtés orthographiques existent aussi du côté de la langue arabe. De plus, les *tweets* suivent une forme standard et emploient des termes spéciaux qui les distinguent des autres microblogues :

- Le nom d'utilisateur identifie chaque utilisateur et commence par (@ ; @UserName). Chaque utilisateur peut mentionner un autre utilisateur dans son *tweet* en écrivant son nom précédé par (@).
- Le *hashtag* est précédé par le caractère (#). Il met en relief les mots-clés spécifiques à un évènement. En cliquant sur un *hashtag* donné, les *tweets* qui contiennent le même hashtag apparaissent.
- Le *retweet* commence par (RT ; @username) et indique la réémission du *tweet* publié par un même *username.*
- L'URL est inséré à la fin du *tweet* pour indiquer son origine. Les URLs sont réduits par Twitter à l'aide d'un service de réduction, afin qu'ils ne dépassent pas la taille maximale permise pour un *tweet*, soit 140 caractères.

Toutes ces caractéristiques linguistiques et orthographiques des *tweets* nous obligent à procéder à une étape de prétraitement, qui est détaillée dans la section suivante.

4.3.1 Le processus de prétraitement des *tweets* en anglais

Face aux fautes d'orthographe commises par les utilisateurs, nous proposons une méthodologie permettant de normaliser le plus possible les mots non standards (ou

mots vides) pour les *tweets* anglais. Les étapes générales de notre méthodologie sont décrites à la figure 4.2.

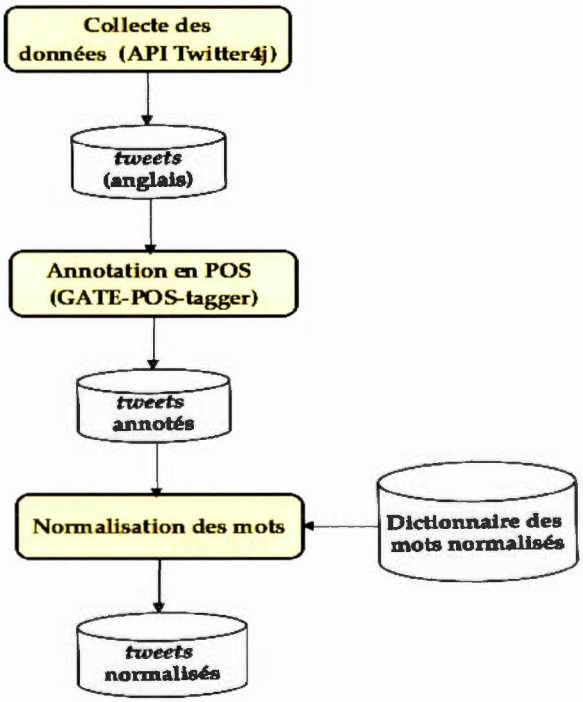

Figure 4.2 Processus de normalisation automatique des mots non standards pour le modèle de langue

Après la collecte des *tweets*, nous avons procédé à l'étape d'annotation syntaxique. En effet, dans la procédure appliquée, nous avons éliminé les *hashtags*, les *user-names*, les *retweets* et les URLs de la normalisation. Seul le texte du *tweet* écrit par l'utilisateur a été traité par notre processus de prétraitement, dans le but de construire un système de TAS.

Pour ce faire et afin de bien repérer ces champs spéciaux à Twitter, nous avons eu recours à une annotation syntaxique des *tweets* par catégorie grammaticale (*part-of-speech* [POS]). Nous avons ainsi annoté notre corpus avec l'outil *General*

Architecture for Text Engineering (GATE) pour l'annotation partielle des *tweets* (*Gate Twitter Part-of-Speech Tagger*) [15]. Cet outil a atteint les 97.5% de précision pour l'annotation de données bruités comme les *tweets* (Derczynski *et al.*, 2013). Pour cette raison, nous l'avons choisi parmi les annotateurs disponibles. Le tableau 4.2 présente l'exemple d'un *tweet* annoté avec l'outil GATE.

Tableau 4.2 Exemple d'annotation d'un *tweet* en POS

Sans annotation en POS	How exciting! RT @BunchesUK : Hello! What's happening in your world? We're all gearing up for #Valentines with bouquets flying out the door.
Avec annotation en POS	How_**WRB** exciting!_**JJ** RT_**RT** @BunchesUK :_**USR** Hello!_**NNP** What's_**NNP** happening_**VBG** in_**IN** your_**PRP$** world?_**NN** We're_**NNP** all_**DT** gearing_**VBG** up_**RP** for_**IN** #Valentines_**HT** with_**IN** bouquets_**NNS** flying_**VBG** out_**IN** the_**DT** door._**NN**

L'outil *Gate Twitter Part-of-Speech Tagger* accorde pour sa part des étiquettes tirées de la liste des étiquettes syntaxiques proposées par le projet Penn Treebank [16], en plus de quatre nouvelles étiquettes : (1) *"URL"* pour les hyperliens, (2) *"HT"* pour les *hashtags*, (3) *"USR"* pour les *username* et (4) *"RT"* pour les *retweets*.

15. https ://gate.ac.uk/wiki/twitter-postagger.html

16. https ://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Après l'étape d'annotation en POS, nous avons divisé nos *tweets* en segments selon les annotations proposées. Nous n'avons considéré dans ce travail que les mots annotés par des étiquettes syntaxiques autres que URL, HT, USR et RT. Chaque mot annoté a ensuite été séparé de son étiquette syntaxique

Ensuite a lieu l'étape de normalisation des mots. Pour cette étape, nous nous sommes basés sur un dictionnaire de mots non normalisés (ou mots vides) et leurs correspondances normalisés en anglais [17]. Le dictionnaire contient 44 983 couples de mots. Une fois l'étape de normalisation lexicale faite, nous avons obtenu un corpus de *tweets* en anglais prêt pour la construction du modèle de langue en utilisant l'outil SRILM (Stolcke *et al.*, 2002).

4.3.2 Le processus de prétraitement des *tweets* en arabe

Nous avons utilisé un petit corpus de *tweets* arabes à la phase d'évaluation du système de traduction automatique. Suite aux étapes de prétraitement et de normalisation, le corpus a ensuite été traduit en anglais par un expert en TA. La qualité de ce corpus influence beaucoup les scores obtenus.

Dans un premier temps, nous avons testé les données bruitées sans normalisation. Ensuite, nous avons étudié notre corpus de plus près et nous avons traité les erreurs lexicales commises par les utilisateurs, à savoir l'étirement des mots, les translittérations, les abréviations, etc. Ensuite nous avons eu recours aux étapes classiques du traitement de l'ASM, à savoir la normalisation des lettres « Hamza

---

17. Le dictionnaire contient des couples de mots : le mot non normalisé et sa correspondance normalisée. Ce dictionnaire a été proposé lors d'une compétition organisée dans le cadre de la conférence W-NUT-2015 (ACL 2015 Workshop on Noisy User-generated Text [W-NUT]) et utilisé dans les travaux de Han *et al.* (2013); Pennell et Liu (2014)). Les données sont disponibles en ligne : http ://noisy-text.github.io/2015/norm-shared-task.html (dernière consultation 11-1-2017).

[a] » et « Ya [y]» et la segmentation des mots par l'analyseur morphologique MADA (Habash *et al.*, 2009). Les dernières étapes de prétraitement classique de la langue arabe (segmentation, normalisation des lettres « Hamza [a] » et « Ya [y]») ont aussi été appliquées à la partie source (en arabe) du corpus parallèle utilisé pour l'entraînement du modèle de traduction. Ces étapes sont décrites à la figure 4.3.

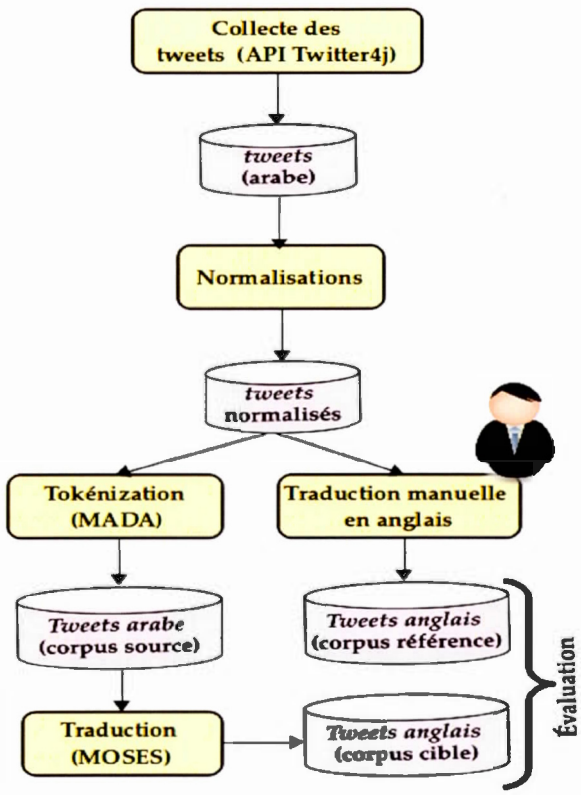

Figure 4.3 Processus de prétraitement pour le corpus de test

Dans nos expériences, nous avons étudié l'effet de ces prétraitements sur la performance de notre système de traduction. Nous rapportons les résultats obtenus au chapitre suivant.

En effet, le manque d'outils et de ressources linguistiques en TALN qui traitent les données des microblogues et des médias sociaux en arabe et le besoin de traducteurs automatiques pour ce type de donnée, nous a conduite à développer un traducteur pour les *tweets* arabes. Le plus grand obstacle pour y arriver est le manque de données issues des microblogues en arabe, qu'il s'agisse du corpus test à traduire ou bien des corpus parallèles des *tweets*. Dans ce travail de recherche, tout ce que nous avons été en mesure de collecter est un corpus de test de l'ordre de 551 *tweets* arabes. Pour ces raisons, la stratégie adoptée a été de rapprocher le plus possible le langage utilisé dans les *tweets* de l'arabe. L'objectif est que le langage utilisé dans le corpus à traduire soit le plus proche possible des données d'entraînement, ce qui va influencer positivement la qualité du traducteur. Rappelons que suivant (Langlais *et al.*, 2006), les traducteurs automatiques offrent une meilleure qualité de traduction si les données de test et d'entraînement relèvent du même domaine, ce qui n'est pas le cas dans notre travail.

4.4 Le prétraitement de l'arabe standard moderne

Les étapes de prétraitement présentées dans cette section sont appliquées aux corpus d'entraînement et de test après la normalisation des mots bruités. Ces étapes de prétraitement sont décrites comme suit :

La normalisation

Comme indiqué dans la partie précédente, nous visons par cette étape la normalisation des deux lettres « Hamza [a] » et « Ya [y]». En effet le « Hamza [a] » en arabe peut être représenté sous plusieurs formes selon la position de cette lettre dans le mot. Nous distinguons donc les lettres (« إ ، آ ، أ » [>, |, <]) pour le représenter. La normalisation que nous proposons consiste à changer les différentes formes orthographiques de cette lettre en une seule lettre : le « ا » [A]. Aussi,

la lettre « Ya [y] » est représentée par une des deux lettres (« ى ،ي » [Y, y]) ; toutes les lettres « ي » [y] seront donc normalisées en « ى » [Y] pour diminuer le degré d'ambiguïté orthographique.

La segmentation des mots

La segmentation consiste à séparer les clitiques de la racine du mot. Habash et Sadat (2006) ont distingué onze niveaux de segmentation des mots arabes qui sont représentés par les schèmes de prétraitement suivants : S1, ON, D1, D2, D3, WA, TB, MR, L1, L2 et EN. Nous avons utilisé le type de segmentation D2, qui revient à séparer les proclitiques « و » [w], « ف » [f], « ل » [l], « ك » [k], « ب » [b] et « س » [s] de la forme de base. D'après Habash et Sadat (2006), D2 est le type de segmentation le plus efficace pour la construction d'un système de TAS de l'arabe vers l'anglais.

Pour notre étude, nous avons utilisé le segmenteur *Morphological Analysis and Disambiguation for Arabie (MADA)* (Habash *et al.*, 2009). L'exemple présenté dans la figure 4.4 illustre la segmentation des mots d'une phrase en arabe avec le schéma D2 en utilisant MADA.

Phrase arabe
و تمت عملية النقل بحماية 120 شرطيا و جنديا في اطار قافلة سيارات بينما حلقت مروحية للشرطة فوق المنطقة
[wtmt Emlyp Alnql bHmAyp 120 $rTyA w jndyA fy ATAr qAflp syArAt bynmA Hlqt mrwHyp ll$rTp fwq AlmnTqp]

Traduction en anglais :
The transfer operation was conducted in a car convoy under the protection of 120 policemen and soldiers while a police helicopter hovered over the area.

Segmentation aevc MADA:
و+ تمت عملية النقل ب+حماية 120 شرطيا و +جنديا في اطار قافلة سيارات بينما حلقت مروحية ل+الشرطة فوق المنطقة
[w+tmt Emlyp Alnql b+HmAyp 120 $rTyA w+jndyA fy ATAr qAflp syArAt bynmA Hlqt mrwHyp l+l$rTp fwq AlmnTqp]

Figure 4.4 Exemple de segmentation avec MADA

Dans ce chapitre nous avons présenté notre méthodologie pour le prétraitement des *tweets* arabes et anglais dans le but de construire un système de TAS pour les *tweets*, de l'arabe vers l'anglais. Les données utilisées ont été collectées avec l'APIs de Twitter. Nous avons ensuite détaillé les étapes suivies pour le prétraitement de ces *tweets* pour l'arabe et l'anglais. Finalement nous avons mis l'accent sur la normalisation ainsi que sur la segmentation des mots arabes en utilisant l'analyseur morphologique MADA.

CHAPITRE V

ÉVALUATION DE LA MÉTHODE PROPOSÉE

Dans ce chapitre, nous présentons les différentes évaluations que nous avons fait pour notre système de traduction automatique statistique (TAS) des *tweets* arabes vers l'anglais.

5.1 Les outils linguistiques utilisés

Afin d'évaluer notre méthode, nous avons entraîné plusieurs systèmes de traduction automatique probabilistes. Pour ce faire, nous avons utilisé les outils disponibles en licence libre, qui sont indispensables pour la création du modèle de langue, l'apprentissage du modèle de traduction et le décodage. Nous présentons ces outils aux sections suivantes.

5.1.1 Le *SRI Language Modeling Toolkit (SRILM)*

Le *SRI Language Modeling Toolkit (SRILM)*[18] (Stolcke *et al.*, 2002) est une boite à outils utilisée pour la construction des modèles de langues statistiques pour des applications en reconnaissance vocale, en segmentation et étiquetage statistique et en traduction. Cet outil automatique se base sur les modèles statistiques n-

18. http ://www.speech.sri.com/projects/srilm/

grammes pour la création des modèles. Pour nos travaux, nous avons créé des modèles 3-grammes pour les modèles de langues.

5.1.2 La librairie MGIZA++

La librairie MGIZA++ [19] (Gao et Vogel, 2008) sert à aligner les corpus parallèles utilisés par les systèmes de traduction. Cette librairie implémente les modèles IBM 1-5 (Och et Ney, 2003) pour l'alignement parallèle des segments du système de traduction à base de segments. Ensuite, le décodeur Moses (Koehn *et al.*, 2007) est utilisé pour symétriser les alignements en se basant sur l'heuristique *grow-diag-final-and.*

5.1.3 Le décodeur Moses

Moses [20] (Koehn *et al.*, 2007) offre l'ensemble des outils nécessaires pour la construction de systèmes de traduction statistique basée sur les segments. Il existe d'autres systèmes en licence libre, comme Pharaoh (Koehn, 2004) (le prédécesseur de Moses), Portage (Johnson *et al.*, 2006) et Cdec (Dyer *et al.*, 2010).

Moses offre des algorithmes d'apprentissage et de décodage très performants et présente plusieurs caractéristiques intéressantes. Les principaux modules de Moses que nous avons utilisés sont les suivants :

- Train-Moses : offre des scripts pour préparer les données d'apprentissage et entraîner le modèle de langue. Il offre aussi des outils d'évaluation et d'analyse des résultats.
- MERT-Moses : contient un outil utilisé pour régler les paramètres par minimisation du taux d'erreur (*tuning*) (Bertoldi *et al.*, 2009).

---

19. https ://github.com/moses-smt/mgiza

20. http ://www.statmt.org/moses/ ?n=Moses.Overview

- Moses-cmd : contient tous les outils et les scripts pour le décodage.

5.2 Création du système de traduction automatique

De manière générale, pour créer un système de traduction, la boite à outils Moses effectue les neuf étapes suivantes :

1. préparation du corpus parallèle ;
2. alignement des mots avec MGIZA++ ;
3. fusion des alignements dans les deux sens de la traduction ;
4. apprentissage du modèle de traduction ;
5. extraction des segments ;
6. attribution des probabilités à chaque segment ;
7. apprentissage du modèle avec ré-ordonnancement des poids ;
8. apprentissage du modèle générique ;
9. création du fichier de configuration (moses.ini).

Pour la réalisation de nos expériences et l'évaluation de notre méthodologie, nous avons créé différents systèmes de traduction automatique. Les premiers systèmes sont des systèmes de base, qui ne passent pas par l'étape de prétraitement des *tweets*. Ces derniers systèmes ont servi pour une étude comparative avec les autres systèmes entraînés après le prétraitement des *tweets*.

En plus, et afin d'évaluer notre méthodologie qui se base sur un modèle de langue pour les *tweets*, nous avons introduit différents modèles de langues pour les systèmes, qui ont été créés comme suit :

(i) modèle de langue pour les tweets ;

(ii) modèle de langue en MSA ;

(iii) deux modèles de langues (tweets et MSA).

Nous avons étudié par la suite l'effet de chaque modèle de langue sur les résultats obtenus. Enfin, nous avons évalué les différents systèmes avec le score BLEU et le taux de mots non traduits (le taux de OOV).

Après les étapes de préparation du corpus parallèle et d'entraînement des deux modèles de langues et de traduction, nous sommes passés à l'étape de l'optimisation des poids. Enfin, nous avons généré le corpus traduit ; c'est l'étape de décodage. Les corpus sur lesquels se base notre travail relèvent donc de quatre catégories : le corpus d'entraînement (TRAIN), le corpus d'optimisation des poids (TUNING), le corpus monolingue dans la langue cible utilisé pour générer le modèle de langue (LM) et enfin le corpus test (TEST).

Dans ce qui suit, nous présentons en détail les différentes étapes et les ressources utilisées.

5.3 Préparation des données

Corpus parallèle

Il aurait été préférable que le corpus parallèle soit du domaine des microblogs et de grande taille. Cependant, la non-disponibilité d'une telle ressource nous a obligé à utiliser un corpus en ASM. Nous avons ainsi choisi nos donnée d'entrainement à partir de corpus parallèle de l'Organisation des Nations Unies (*United Nations [UN]*) qui est disponible dans six langues différentes, dont l'arabe (ASM) et l'anglais [21]. La taille du corpus que nous avons utilisé est de l'ordre de 600 000 phrases alignées. Le tableau 5.1 donne plus de détail sur cette ressource.

---

21. Les corpus UN bruts sont disponibles en ligne via le site de documentation de l'UN http ://www.un.org/en/documents/ods/

La raison pour laquelle nous nous sommes limités à cette taille de corpus, qu'elle est la seule ressource bien alignée phrase par phrase [22], dont nous disposons. En effet, le corpus UN qui est disponible en ligne est brute et demande une étape d'alignement automatique et des vérifications manuelles de l'alignement réalisé, ce qui demande beaucoup de temps. Malheureusement, pour les applications en TA, les textes parallèles librement disponibles, qui sont de bonnes qualité et bien alignés phrase par phrase, sont des ressources rares et très chères ; la taille est souvent limitée et le domaine est rarement approprié. Le domaine du corpus de l'UN n'est pas du même domaine que celui des *tweets*, mais nous l'avons utilisé quand même pour l'entraînement dans le cadre de notre recherche, puisqu'il est la seule ressource dont nous disposons. De plus, les corpus parallèles des *tweets* de l'arabe vers l'anglais ne sont pas disponibles. Ceci, représente un grand défi pour les langues peu dotées, comme les *tweets* pour les applications de TAS.

Tableau 5.1 Taille du corpus d'entrainement

	Taille en Mo	Nombre de mots
Arabe (source)	165,5	16 866 817
Anglais (cible)	113,5	19 408 007

En premier lieu, nous avons nettoyé le corpus parallèle en éliminant les phrases qui dépassent les 100 mots. Pour ce faire, nous nous sommes basés sur le manuel d'utilisation de Moses [23]. Ensuite nous avons vérifié qu'il n'existe pas de lignes vides dans le corpus parallèle, ce qui aurait nui à l'étape d'entrainement par Moses et aussi à l'étape d'alignement entre les segments. Nous sommes passés par la suite

22. L'alignement du corpus parallèle consiste à vérifier que la phrase en arabe à la ligne (i) correspond bien à la traduction en anglais dans la même ligne (i).

23. http ://www.statmt.org/moses/ ?n=FactoredTraining.PrepareTraining

à l'étape de segmentation et de casse minuscule de la partie cible (anglais) du corpus parallèle avec le script offert par Moses. En nous basant sur ce corpus, nous avons entraîné trois systèmes de base, des systèmes n'incluant pas de prétraitement de la langue arabe.

L'étape suivante inclus l'entraînement des systèmes de traduction après les prétraitements nécessaires pour la partie source du corpus en langue arabe. Nous avons commencé par la normalisation des lettres « Hamza [a] » et « Ya [y]», ainsi que par la segmentation des mots avec l'outil MADA (Habash *et al.*, 2009). Comme présenté au chapitre précédent, cette étape revient à segmenter les mots en séparant les particules « و » [w], « ف » [f], « ل » [l], « ك » [k], « ب » [b] et « س » [s] de la forme de base, en suivant le schème D2 (Habash et Sadat, 2006). Nous avons donc obtenu un corpus plus important en termes de mots, qui est présenté au tableau 5.2. Nous avons ensuite utilisé ce corpus pour construire les autres systèmes.

Tableau 5.2 Taille du corpus en termes de mots après la tokénisation

	Avant tokénisation	Après tokénisation
Nombre de mots	16 866 817	18 791 118

Pour lancer la commande d'entraînement, il faut utiliser le script *train-model.perl* offert par Moses. Aussi, Moses offre la possibilité d'introduire plusieurs modèles de langue en parallèle dans le système de TAS. À la fin de cette étape, nous avons obtenu la table de traduction (*phrase table*) contenant les alignements de tous les segments. Cette table est très volumineuse et a été enregistrée sous la forme d'un fichier compressé dans le même répertoire que le fichier de configuration du système *moses.ini*. Un extrait de cette table est présenté en annexe.

Corpus pour les modèles de langue

Pour les systèmes entraînés, nous avons utilisé deux types de modèles de langues : un premier modèle de langage (*Language Model [LM]*) des *tweets* en anglais, et le deuxième issu de la partie cible du corpus d'entrainement. Pour le premier corpus, nous avons collecté les *tweets* via le STREAMING APIs de Twitter, puis nous avons fait les prétraitements nécessaires et la normalisation des mots hors vocabulaire. Notre corpus est formé de 255 602 *tweets* anglais. Ensuite, nous avons utilisé l'outil SRILM pour générer les modèles de langues.

Corpus test

Pour l'étape de décodage, nous avons utilisé un corpus parallèle que nous avons créé manuellement. Ce corpus contient 551 *tweets* en arabe et leur traduction en angalais. Nous avons tout d'abord testé la qualité de la traduction des systèmes de base (sans faire les prétraitements), puis nous avons effectué les prétraitements et les normalisations nécessaires.

5.4 Optimisation des poids des traductions

Pour optimiser les poids dans la table de traduction, nous avons besoin d'un petit corpus parallèle de développement, différent du corpus d'entraînement. Nous avons donc utilisé dans un premier temps une partie du grand corpus parallèle de l'UN, de l'ordre de 1000 lignes (arabe-anglais).

Pour la phase de décodage, nous avons traduit la partie source de notre corpus de test. Le fichier ainsi produit par le système de traduction automatique a été évalué et comparé avec le fichier de référence du corpus test (la partie cible du corpus test).

5.5 Expérimentations et évaluation

Systèmes de base

Comme première expérience, nous avons testé notre système de traduction sans aucun prétraitement. Dans notre méthodologie, nous avons proposé d'introduire un modèle de langue pour les *tweets* afin d'améliorer la performance de leur traduction automatique. Nous avons par la suite testé plusieurs systèmes avec différentes combinaisons de modèles de langues, afin d'évaluer la méthodologie proposée. Ensuite, pour mesurer la qualité des traductions produites par les systèmes entraînés, nous avons calculé le score BLEU à l'aide du script *multi-bleu.perl* de Moses. Ce score varie entre 0 et 100% ; plus le score est élevé, plus le système de traduction est performant. Nous avons également observé le taux de mots non traduits (OOV) avec le script *nontranslated_words.pl.* Il est très important que ce taux soit le plus bas possible pour un bon système de traduction.

Les résultats obtenus sont détaillés au tableau 5.3 :

Tableau 5.3 Évaluation de la traduction des *tweets* avant prétraitement

	BLEU	Taux OOV
Système 1 (LM tweets)	2.39	28.74
Système 2 (LM tweets+LM MSA)	2.49	27.75
Système 3 (LM MSA)	6.31	24.53

Le système entraîné avec un modèle de langue du même domaine que le corpus parallèle s'est avéré le plus performant, avec un score BLEU de 6.31 et le taux de mots non traduits le plus faible (Système 3). Ce résultat nous montre que le modèle de langue ne peut aider dans le choix des meilleures phrases traduites tant que les mots de ces phrases ne figurent pas dans le corpus d'entraînement.

En revanche, en examinant de près la liste des mots non traduits, nous avons remarqué les faits suivants :

— La présence de noms propres (comme : « بوتين » [bwtyn], « درعا [drEA] », « الجزيرة » [Aljzyrp] , « تويتر [twytr] » , « جوجل [jwjl] » , «حماه [HmAh] », «حمص [HmS] », «داعش [dAE$] », «روسيا [rwsyA] » , «صبرة [Sbrp] ») et de mots composés de deux mots ou plus et séparés par des tirets, comme : « تنظيم_الدولة [tnZym_Aldwlp] », « الربيع_العربي [AlrbyE_AlErby]». Ces mots sont en relation avec les mots-clés que nous avons utilisés pour collecter le corpus de test. Ce sont des mots relatifs à l'actualité des pays arabes, probablement des *hashtags*, que le corpus parallèle ne couvre pas.

— L'existence des mots agglutinés dont les particules sont collées aux mots. Ces mots n'ont pas subi l'étape de segmentation.

— La présence de plusieurs mots non normalisés dont on remarque souvent la présence dans les textes des médias sociaux, à savoir les mots étirés (comme par exemple :« تكلمنـــــــي [tklmn_____y] », « دبـــــــي [db_____y] ») et des mots avec une lettre qui se répète exprimant une émotion (comme : « كبررررر [kbrrrrrr] », « لايمككننا [lAymkknnA] », « هههههههههههههههههههههههههه [hhhhhhhhhhhhhhhhhhhhhhhhh] »).

— L'utilisation de mots en arabe dialectal et non en ASM, en particulier le dialecte levantin. La présence de ce type de mot est facilement remarquable dans le corpus test.

Systèmes de traduction après les prétraitements

Après l'étude des problèmes qui ont mené à un score BLEU faible et à une mauvaise qualité de la traduction automatique, nous avons procédé aux prétraitements nécessaires et au traitement des mots hors vocabulaires. Ensuite, nous avons normalisé les lettres « Hamza [a] » et « Ya [y]». Enfin, nous avons procédé à la seg-

mentation des mots de leurs particules, avec MADA. Ces deux dernières étapes ont été réalisées sur les corpus parallèles d'entraînement et de développement et sur le corpus test. Les résultats obtenus sont reportés au tableau 5.4.

Tableau 5.4 Évaluation de la traduction automatique des *tweets* après prétraitement

	BLEU	Taux OOV
Système 4 (LM *tweets*)	6.09	11.18
Système 5 (LM Tweets+LM MSA)	3.50	12.11
Système 6 (LM MSA)	8.39	9.95

Les scores ont augmenté de +2 points pour la majorité des systèmes entraînés, avec le système 6 présentant le meilleur score BLEU (8.39). Plusieurs mots issus des microblogues, des mots du domaine des *tweets*, n'ont pas pu être traduits. En fait, les noms propres qui ont été repérés pour les systèmes de base figurent encore dans la liste des mots non traduits.

Malgré qu'il subsiste plusieurs mots non traduits, le taux de OOV n'est pas trop élevé par rapport au score BLEU trouvé. Nous avons donc examiné le fichier traduit de près et nous avons remarqué que la majorité des mots dans une phrase ont été traduits en anglais, mais dans des mots qui ne correspondent pas aux mots présents dans le fichier référence. Autrement dit, le sens de la phrase traduite avec notre système n'est pas le même que dans le fichier référence. Prenons l'exemple d'une phrase traduite avec le système 6, soit celui qui présente le meilleur score BLEU : Dans cet exemple, nous remarquons que tous les mots de la phrase ont été traduits ainsi que les noms propres, mais que le sens de la phrase ne correspond pas à celui de la phrase issue du fichier de référence. Si on calcule la correspondance entre la phrase traduite et celle de référence en termes de n-grammes similaires, à l'image du principe appliqué pour établir le score BLEU, les résultats ne seront

Phrase source :	علاقات متميزة راسخة تجمعنا ب+ الأشقاء في سوريا
Écriture Buckwalter :	[ElAqAt mtmyzp rAsxp tjmEnA b+ Al>$qA' fy swryA]
Phrase traduite :	relations firm distinct link in Syria brothers
Phrase référence :	Distinct and well established relations with our brothers in Syria.

certainement pas concluants. De plus, l'ordre des mots dans la phrase en arabe n'est pas respecté dans la traduction, ce qui est dû au fait que notre système de traduction se base sur une approche statistique. En effet, les mots ont été traduits en suivant une distribution probabiliste dans le corpus d'entraînement. L'alignement de la phrase arabe avec celle en anglais présente un grand défi, puisque la phrase en arabe est généralement de type verbal (VSO), tandis que la phrase en anglais est de type nominal (SVO).

Pour toutes les expériences réalisées, nous nous sommes basés sur un petit corpus extrait du grand corpus de l'UN pour l'étape de ré-ordonnancement.

Utilisant une nouvelle stratégie, nous avons essayé de réajuster les poids de ces trois systèmes, entraînés après prétraitement, en utilisant un corpus de développement composé de *tweets*. Ce corpus a été utilisé par Refaee et Rieser (2015). Les résultats sont présentés au tableau 5.5.

Tableau 5.5 Résultats après un tuning avec un corpus du domaine

	BLEU	Taux OOV
Système 4 (LM *tweets*)	10.31	7.16
Système 5 (LM *tweets*+LM MSA)	10.98	6.89
Système 6 (LM MSA)	8.96	9.61

Les scores BLEU obtenus après le réajustement des poids du modèle de traduction sont les meilleurs parmi tous les autres obtenus, c'est-à-dire plus élevés. Les taux OOV sont également mois élevés. L'optimisation des poids avec un corpus de *tweets* et l'utilisation d'un modèle de traduction de ces *tweets* est une combinaison qui a prouvé son efficacité. Elle a aidé au mieux le système statistique de traduction des *tweets*.

Pour les systèmes de traduction basés sur une approche statistique, il s'avère très important que les données d'entraînement relèvent du même domaine que le fichier à traduire, afin de garantir une bonne traduction. Nous avons continué nos expérimentations en testant la traduction du corpus NIST 2005 (MT05) avec le système 3 (système de base entraîné avec un modèle de langue et un corpus d'entraînement en ASM) et ensuite avec le système 6 (système après prétraitements entraîné avec un modèle de langue et un corpus d'entraînement en ASM). Les résultats reportés au tableau 5.6 prouvent bien notre proposition, à savoir l'importance de la correspondance entre le corpus test et le corpus d'entraînement pour la qualité de la traduction automatique.

Tableau 5.6 Résultats de traduction du MT05 (Bleu et taux de OOV)

	BLEU	Taux OOV
Sans prétraitement	10.87	15.86
Avec prétraitement	16.91	3.51

En comparant le score BLEU après prétraitement pour la traduction de MT05 (16.91) et celui de notre système pour la traduction des *tweets* (10.98), nous remarquons que la différence n'est pas trop grande. Certainement, comme nous l'avons expliqué, la non disponibilité d'un corpus parallèle de *tweets* pour l'entraînement de nos systèmes a biaisé nos résultats, mais la stratégie suivie pour adapter le

système de traduction pour les *tweets* s'est avérée bonne en général. De plus, pour l'évaluation des systèmes de traduction, il est préférable d'utiliser plus d'un fichier de référence, ce qui n'est pas évident dans le cas d'une traduction produite manuellement.

Dans notre travail, nous avons donc réalisé différentes évaluations en analysant à chaque fois soit la combinaison des modèles de langue, soit les prétraitements, soit le corpus de ré-ordonnacement des poids. Nous avons reporté tous nos résultats des différentes expériences au tableau 5.7.

Tableau 5.7 Tableau récapitulatif des résultats obtenus

	LM	Test	Prétraitements	tuning	Évaluation	
					BLEU	Taux OOV
1	*tweets*	*tweets*	-	UN	2.39	28.74
2	*tweets*	*tweets*	+	UN	6.09	9.95
3	*tweets*	*tweets*	+	*tweets*	10.31	7.16
4	UN	*tweets*	-	UN	6.31	24.53
5	UN	*tweets*	+	UN	8.39	12.11
6	UN	*tweets*	+	*tweets*	8.96	9.61
7	*tweets*+UN	*tweets*	-	UN	2.49	27.75
8	*tweets*+UN	*tweets*	+	UN	3.50	11.18
9	*tweets*+UN	*tweets*	+	*tweets*	**10.98**	6.89
10	*Grand_tweets*	*tweets*	+	*tweets*	10.58	7.49

Le meilleur système employé pour la traduction des *tweets* a atteint un score BLEU de 10.98. Ce résultat reste plus faible que celui d'un système dont le corpus test, le corpus d'entraînement et le corpus de *tuning* relèvent du même domaine. L'étape de prétraitement est très importante et elle a amélioré les scores de 1 à 4

points. De plus, nous avons normalisé les mots hors vocabulaire non traduits par les systèmes de traduction de base. Nous avons ainsi remarqué que le taux de OOV a vraiment chuté. Comme dernière étape, nous avons eu recours à l'optimisation des poids avec un corpus de *tweets*, ce qui a amélioré les scores BLEU et a diminué les taux de OOV. La meilleure combinaison issue de ces expériences est donc un système basé sur deux modèles de langue (UN et *tweets*) avec les prétraitements nécessaires pour les corpus d'entraînement, de développement et de test.

5.6 Discussion des résultats

Dans les travaux de recherche portant sur la traduction automatique de l'arabe vers l'anglais, les scores dépassent souvent 50% de précision selon la taille du corpus parallèle (Habash et Sadat, 2012). Néanmoins, pour les différents travaux réalisés portant sur la traduction des données issues des microblogues, les scores BLEU sont souvent faibles et malgré les prétraitements réalisés, les traductions de ce type de données restent de piètre qualité. Notre résultat est meilleur que celui de Ling *et al.* (2013). Ainsi, pour la traduction d'un corpus test issu des microblogues sur une machine entraînée avec des données hors domaine, les auteurs ont obtenu 8.75 de score BLEU pour la paire de langue anglais-chinoix. Dans le même esprit, Toral *et al.* (2015) ont testé plusieurs décodeurs pour traduire les données des *tweets* pour des langues sémantiquement très proches. Malgré cela, les résultats du score BLEU ont été de 22.57 pour la traduction de l'espagnol vers le basque. Cette valeur était attendue, puisque ces deux langues sont très proches. Pour leur part, Jehl *et al.* (2012) ont essayé plusieurs évaluations pour la traduction des *tweets* de l'arabe vers l'anglais et le meilleur score BLEU obtenu a été de 15.68 pour un corpus parallèle de 5 823 363 phrases, issu de la campagne d'évaluation NIST. Ce dernier résultat prouve l'importance de la taille du corpus parallèle et explique mieux les résultats que nous avons obtenus malgré les prétraitements qui ont été effectués.

Au présent chapitre, nous avons évalué la méthodologie proposée pour la traduction automatique de l'arabe vers l'anglais des *tweets*. Nous avons testé plusieurs stratégies et la meilleure a mené à un score BLEU de 10.98. Ce score est très encourageant si on le compare avec celui obtenu par d'autres travaux recensés par la littérature. Toutefois, le traitement d'une langue morphologiquement riche et complexe comme l'arabe, comprenant par surcroît des données bruitées, n'est pas une tâche facile. En effet, tous les prétraitements et les normalisations que nous avons réalisés ont prouvé leur efficacité et ont visiblement amélioré les résultats obtenus tout en baissant le taux de mots non traduits (OOV).

CONCLUSION

Dans ce travail de recherche, nous avons proposé une méthodologie pour la traduction automatique des *tweets* de l'arabe vers l'anglais en nous basant sur une approche statistique à base de segments. Nous avons utilisé pour nos travaux le décodeur le plus populaire pour la construction des systèmes de traduction statistique à base de segments (PBSMT), Moses (Koehn *et al.*, 2007).

Par le présent travail, nous avons proposé plusieurs stratégies d'adaptation des systèmes de traduction pour les *tweets*. Ces stratégies se basent sur un modèle de langue pour les *tweets* en premier lieu, ainsi que sur un petit corpus parallèle de *tweets* comme corpus de test. La première étape a été consacrée à la collecte d'un grand corpus de données monolingues pour les *tweets* en anglais via l'API de Twitter. Ces *tweets* collectés ont ensuite subi des étapes de prétraitement et de normalisation des mots hors vocabulaire. Puis nous avons généré le modèle de langue des *tweets* pour la langue cible de notre système. Comme deuxième étape, nous avons collecté des *tweets* en arabe et nous avons demandé à traducteur humain expert de traduire le corpus test de l'arabe vers l'anglais.

Les expériences pour les systèmes de base qui ne sont pas passés par les prétraitements nécessaires ont montré des scores plus faibles que les expériences suivant le prétraitement et la normalisation des données pour la partie arabe. Cette langue dont la morphologie est riche et complexe nécessite de passer par des étapes de segmentation et de normalisation. La segmentation revient à séparer les mots de leurs particules, puisque les mots en arabe sont souvent agglutinés et qu'un mot en arabe peut être exprimé par toute une phrase en anglais ou en français. Pour la

normalisation, nous avons traité les lettres « Hamza [a] » et « Ya [y]», qui s'écrivent de plusieurs façons, ce qui augmente le degré d'ambiguïté orthographique dans le corpus à traduire.

Nous avons aussi traité les mots issus des médias sociaux qui sont hors vocabulaire, comme les mots étirés, les noms composés et les entités nommées. Comme dernière stratégie, nous avons ré-ajusté les poids des systèmes entraînés après prétraitement en utilisant un corpus de développement pour les *tweets*. La valeur du score BLEU la plus élevée atteinte a été de 10.98 et le nombre de mots non traduits a diminué de 3 points, ce qui est considérable.

Nous évaluons les résultats qui ont permis de mettre en évidence l'importance des étapes de prétraitement et de normalisation pour les données utilisées. Une telle étape est capitale pour améliorer les scores, en particulier pour la langue arabe, qui est morphologiquement complexe. Le score que nous avons obtenu pour la traduction des *tweets* n'est pas très élevé par rapport aux valeurs obtenues dans le cadre de travaux portant sur la traduction de données autres que celles tirées des médias sociaux. Pour la traduction automatique statistique, plus le corpus parallèle est proche du domaine du corpus test, plus le score BLEU est élevé. Toutefois, il existe plusieurs travaux qui ont traité de la traduction des textes de microblogues pour d'autres langues et leurs résultats étaient également faibles. La nature bruitée de ces données nuit toujours aux résultats obtenus.

En effet, pour améliorer les résultats pour la traduction de ce type de donnée, et comme pistes pour des recherches futures, il serait utile de :

— Collecter un grand corpus parallèle pour les *tweets* en nous basant sur des méthodes automatiques de la recherche d'information translingue ;
— Profiter des méthodes d'apprentissage automatique pour la normalisation automatique des corpus ;

— Tester d'autres méthodes de traduction, par exemple la traduction automatique neuronale

— Proposer une méthode de normalisation textuelle supervisée pour la langue arabe, afin de rapprocher le langage des *tweets* le plus possible à l'arabe standard moderne (ASM).

— Traiter les phénomèmes « d'Arabizi » et du « code switching » dans les *tweets* arabe.

ANNEXE A

EXTRAIT DU DICTIONNAIRE DES MOTS NORMALISÉS

papaer|paper	faake|fake	cancelar|cancel
sorrryy|sorry	throwning|throwing	coookies|cookies
summin|something	sumn|some	sumone|someone
sumtimes|sometimes	sumtin|something	sumwhere|somewhere
sumwhr|somewhere	sunt|sent	suny|sunny
sup|super	supa|super	superrrr|super
suposed|supposed	supossed|supposed	supp|support
supr|super	suprise|surprise	suree|sure
surpose|suppose	sussed|successes	suttin|something
suuck|suck	suuuuuuuunnnn|sun	sux|sucks
svc|service	svcs|services	swang|swing
swap'd|swapped	swea|swear	sweatin|sweating
sweeeeeeet|sweet	sweeeeeet|sweet	sweeeet|sweet
sweeet|sweet	sweetiing|sweeting	sweetss|sweets
swimmig|swimming	swt|sweet	syaing|saying
syg|saying	sz|size	t2|to

ANNEXE B

EXTRAIT DU FICHIER D'ALIGNEMENT DU CORPUS PARALLÈLE

```
# Sentence pair (1) source length 3 target length 16 alignment score : 1.38622e-31
ذ صدر أصلا في شكل نسخة بالاستنسل تحت الرمز /48/16) ةرت ِ ندّدد .
NULL ({ 4 }) A/48/16 ({ 1 2 3 5 6 7 8 9 11 }) (Part ({ 10 12 13 14 15 }) I). ({ 16 })
# Sentence pair (2) source length 46 target length 49 alignment score : 1.18095e-80
1 - عقدت لجنة البرنامج و+ التنسيق دورة تنظيمية (الجلستان الأولى و+ الثانية) في مقر الأمم المتحدة يومي 8 نيسان / أبريل و 5 أيار / مايو 1993 , كما عقدت الجزء الأول من دورتها الثالثة و+ الثلاثين في المقر من 10 إلى 14 أيار / مايو 1993 .
NULL ({ }) 1 ({ 1 }) . ({ 2 }) The ({ }) Committee ({ 4 }) for ({ }) Programme ({ 5 }) and ({ 6 }) Coordination ({ 7 }) (CPC) ({ 3 }) held ({ }) an ({ }) organizational ({ 9 }) session ({ 8 }) (1st ({ 10 11 }) and ({ 12 }) 2nd ({ }) meetings) ({ 13 30 31 }) at ({ 14 }) United ({ }) Nations ({ 16 17 }) Headquarters ({ 15 }) on ({ }) 8 ({ 18 19 }) April ({ 20 21 22 }) and ({ 23 }) 5 ({ 24 }) May ({ 25 26 27 }) 1993 ({ 28 }) , ({ 29 }) and ({ 37 }) the ({ }) first ({ 33 }) part ({ 32 }) of ({ 34 }) its ({ }) thirty-third ({ 36 38 }) session ({ 35 }) at ({ 39 }) Headquarters ({ 40 }) from ({ 41 }) 10 ({ 42 }) to ({ 43 }) 14 ({ 44 }) May ({ 45 46 47 }) 1993 ({ 48 }) . ({ 49 })
# Sentence pair (3) source length 15 target length 22 alignment score : 1.22811e-42
و+ لقد عقدت في إطار ذلك الجزء تسع جلسات (الجلسة الثالثة إلى الجلسة الحادية عشرة) و+ عددا من الجلسات غير الرسمية .
NULL ({ }) It ({ 1 2 }) held ({ 3 4 }) nine ({ 8 }) meetings ({ 9 }) (3rd ({ 5 6 7 10 11 13 14 15 }) to ({ 12 }) 11th ({ }) meetings) ({ }) and ({ 16 }) a ({ }) number ({ 17 }) of ({ 18 }) informal ({ 20 21 }) meetings ({ 19 }) . ({ 22 })
# Sentence pair (4) source length 25 target length 22 alignment score : 3.99852e-42
2 - يرد في المرفق الأول أدناه جدول أعمال الدورة الثالثة و+ الثلاثين , و+ لقد أقرته اللجنة في جلستها الأولى .
NULL ({ }) 2 ({ 1 2 }) . ({ 22 }) The ({ 15 16 }) agenda ({ 8 9 }) for ({ }) the ({ }) thirty-third ({ 11 13 }) session ({ 10 }) , ({ 12 }) adopted ({ 17 }) by ({ }) the ({ }) Committee ({ 18 }) at ({ 19 }) its ({ }) 1st ({ 20 21 }) meeting ({ }) , ({ 14 }) is ({ }) reproduced ({ 3 }) in ({ 4 }) annex ({ 5 }) I ({ 6 }) below ({ 7 }) . ({ })
# Sentence pair (5) source length 51 target length 48 alignment score : 6.22912e-85
3 - و+ ب+ إقرار اللجنة جدول الأعمال ف+ إنها قررت , وفقا ل+ القرار الذي اتخذته في دورتها التنظيمية ل+ عام 1993 , أن تنظر في دورتها الثالثة و+ الثلاثين في تقرير وحدة التفتيش المشتركة المعنون &quot; تعاون منظومة الأمم المتحدة مع المؤسسات المالية المتعددة الأطراف .
NULL ({ 4 }) 3 ({ 1 2 }) . ({ }) In ({ 3 }) adopting ({ 5 }) the ({ }) agenda ({ 7 8 }) , ({ 9 }) the ({ }) Committee ({ 6 }) , ({ 12 }) in ({ }) accordance ({ 13 14 }) with ({ }) the ({ }) decision ({ 15 16 }) taken ({ 17 }) at ({ 18 }) its ({ }) organizational ({ 20 }) session ({ 19 }) for ({ 21 }) 1993 ({ 22 23 }) , ({ 24 }) decided ({ 11 }) to ({ }) consider ({ 25 26 }) , ({ }) at ({ 27 }) its ({ }) thirty-third ({ 29 31 }) session ({ 28 }) , ({ 30 }
```

ANNEXE C

EXTRAIT DU LA TABLE DE TRADUCTION

النفقات العامة ||| 2 56 1 ||| public spending ||| 0.185263 0.00879056 0.0357143 0.000238297 ||| 1-0 0-1 ||| 19 56 2
النفقات العامة ||| 1 56 1 ||| streamlined public expenditures ||| 1 0.127315 0.0178571 9.44624e-08 ||| 1-1 0-2 ||| 1 56 1
النفقات العامة ||| 1 56 1 ||| the overall expenditures of the ||| 1 0.0692087 0.0178571 7.134e-05 ||| 1-1 0-2 ||| 1 56 1
النفقات العامة ||| 1 56 1 ||| the overall expenditures of ||| 1 0.0692087 0.0178571 0.000233115 ||| 1-1 0-2 ||| 1 56 1
النفقات العامة ||| 2 56 1 ||| the overall expenditures ||| 0.5 0.0692087 0.0178571 0.00136335 ||| 1-1 0-2 ||| 2 56 1
النفقات العامة ||| 1 56 1 ||| the overhead expenditure ||| 1 0.100809 0.0178571 6.38548e-05 ||| 1-1 0-2 ||| 1 56 1
النفقات العامة ||| 1 56 1 ||| the overhead expenditures of ||| 1 0.0808317 0.0178571 3.3656e-06 ||| 1-0 1-1 0-2 ||| 1 56 1
النفقات العامة ||| 1 56 1 ||| the overhead expenditures ||| 1 0.0808317 0.0178571 1.96835e-05 ||| 1-0 1-1 0-2 ||| 1 56 1
النفقات العامة ||| 1 56 1 ||| the total ||| 0.000693481 1.93726e-07 0.0178571 1.10887e-05 ||| 1-0 ||| 1442 56 1
النفقات العامة ||| 1 56 1 ||| the ||| 1.95912e-06 1.93726e-07 0.0178571 0.051889 ||| 1-0 ||| 510434 56 1
النفقات الإدارية ||| 1 30 1 ||| administrative costs to the ||| 0.25 0.00288875 0.0333333 9.53205e-05 ||| 1-0 0-1 ||| 4 30 1
النفقات الإدارية ||| 1 30 1 ||| administrative costs to ||| 0.25 0.00288875 0.0333333 0.000311475 ||| 1-0 0-1 ||| 4 30 1
النفقات الإدارية ||| 3 30 1 ||| administrative costs ||| 0.025641 0.00288875 0.1 0.0127308 ||| 1-0 0-1 ||| 117 30 3
النفقات الإدارية ||| 1 30 1 ||| administrative expenditure should be ||| 1 0.0786294 0.0333333 2.00792e-06 ||| 1-0 0-1 ||| 1 30 1
النفقات الإدارية ||| 1 30 1 ||| administrative expenditure should ||| 1 0.0786294 0.0333333 0.000225875 ||| 1-0 0-1 ||| 1 30 1
النفقات الإدارية ||| 4 30 1 ||| administrative expenditure ||| 0.444444 0.0786294 0.133333 0.119045 ||| 1-0 0-1 ||| 9 30 4
النفقات الإدارية ||| 1 30 1 ||| administrative expenditures as well as ||| 0.5 0.125394 0.0333333 5.57058e-09 ||| 1-0 0-1 ||| 2 30 1
النفقات الإدارية ||| 1 30 1 ||| administrative expenditures as well ||| 0.5 0.125394 0.0333333 1.06064e-06 ||| 1-0 0-1 ||| 2 30 1
النفقات الإدارية ||| 1 30 1 ||| administrative expenditures as ||| 0.5 0.125394 0.0333333 0.00113668 ||| 1-0 0-1 ||| 2 30 1
النفقات الإدارية ||| 8 30 1 ||| administrative expenditures ||| 0.444444 0.125394 0.266667 0.216424 ||| 1-0 0-1 ||| 18 30 8
النفقات الإدارية ||| 4 30 1 ||| administrative expenses ||| 0.0487805 0.0207802 0.133333 0.0189846 ||| 1-0 0-1 ||| 82 30 4
النفقات الإدارية ||| 1 30 1 ||| administrative ||| 0.000265816 2.70298e-05 0.0333333 0.463671 ||| 1-0 ||| 3762 30 1
النفقات الإدارية ||| 1 30 1 ||| of administrative expenditures ||| 0.5 0.125394 0.0333333 0.0370056 ||| 1-1 0-2 ||| 2 30 1
النفقات الإدارية ||| 1 30 1 ||| the administrative costs ||| 0.0357143 0.00288875 0.0333333 0.00389601 ||| 1-1 0-2 ||| 28 30 1
النفقات الإدارية ||| 1 30 1 ||| the administrative expenditure ||| 0.5 0.0786294 0.0333333 0.0364312 ||| 1-1 0-2 ||| 2 30 1
النفقات في ||| 1 70 1 ||| expenditure amounted to $ ||| 0.142857 0.0120548 0.0142857 9.80176e-10 ||| 0-0 1-2 ||| 7 70 1
النفقات في ||| 1 70 1 ||| expenditure amounted to ||| 0.0769231 0.0120548 0.0142857 6.53494e-07 ||| 0-0 1-2 ||| 13 70 1
النفقات في ||| 1 70 1 ||| expenditure at the ||| 1 0.110272 0.0142857 0.00458112 ||| 0-0 1-1 ||| 1 70 1
النفقات في ||| 1 70 1 ||| expenditure at ||| 1 0.110272 0.0142857 0.0149695 ||| 0-0 1-1 ||| 1 70 1
النفقات في ||| 1 70 1 ||| expenditure for the ||| 0.0277778 0.00428672 0.0142857 0.000421613 ||| 0-0 1-1 ||| 36 70 1
النفقات في ||| 1 70 1 ||| expenditure for ||| 0.0123457 0.00428672 0.0142857 0.00137769 ||| 0-0 1-1 ||| 81 70 1
النفقات في ||| 1 70 1 ||| expenditure in the ||| 0.0625 0.18949 0.0142857 0.0380643 ||| 0-0 1-1 ||| 16 70 1

RÉFÉRENCES

Abuelyaman, E., Rahmatallah, L., Mukhtar, W. et Elagabani, M. (2014). Machine translation of Arabic language : Challenges and keys. Dans *Fifth International Conference on Intelligent Systems, Modelling and Simulation (ICSMO)*, 111–116. IEEE Transactions on Computers.

Adouane, W., Semmar, N., Johansson, R. et Bobicev, V. (2016). Automatic detection of arabicized berber and arabic varieties. *VarDial 3*, p. 63.

Afli, H. (2014). *La Traduction automatique statistique dans un contexte multimodal.* (Thèse de doctorat). Université du Maine.

Al-Haj, H. et Lavie, A. (2012). The impact of Arabic morphological segmentation on broad-coverage English-to-Arabic statistical machine translation. *Machine translation*, *26*(1-2), 3–24.

Almahairi, A., Cho, K., Habash, N. et Courville, A. (2016). First result on arabic neural machine translation. Dans *Conference on Empirical Methods in Natural Language Processing (EMNLP 2016).*

Attia, M. A. (2008). *Handling Arabic morphological and syntactic ambiguity within the LFG framework with a view to machine translation.* (Thèse de doctorat). University of Manchester.

Banerjee, S. et Lavie, A. (2005). Meteor : An automatic metric for mt evaluation with improved correlation with human judgments. Dans *The Association for Computational Linguistics workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, volume 29, 65–72.

Barman, U., Das, A., Wagner, J. et Foster, J. (2014). Code mixing : A challenge for language identification in the language of social media. Dans *First Workshop on Computational Approaches to Code Switching (EMNLP 2014)*, 13–23. Association for Computational Linguistics (ACL).

Bertoldi, N., Haddow, B. et Fouet, J.-B. (2009). Improved minimum error rate training in moses. *The Prague Bulletin of Mathematical Linguistics*, *91*, 7–16.

Bies, A., Song, Z., Maamouri, M., Grimes, S., Lee, H., Wright, J., Strassel, S., Habash, N., Eskander, R. et Rambow, O. (2014). Transliteration of arabizi into arabic orthography : Developing a parallel annotated arabizi-arabic script sms/chat corpus. Dans *Workshop on Arabic Natural Langauge Processing (ANLP)*, 93–103.

Bisazza, A. et Federico, M. (2010). Chunk-based verb reordering in vso sentences for Arabic-English statistical machine translation. Dans *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics (MATR 2010)*, 235–243. Association for Computational Linguistics (ACL).

Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L. et Roossin, P. S. (1990). A statistical approach to machine translation. *Computational linguistics*, *16*(2), 79–85.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D. et Mercer, R. L. (1993). The mathematics of statistical machine translation : Parameter estimation. *Computational linguistics*, *19*(2), 263–311.

Carpuat, M., Marton, Y. et Habash, N. (2012). Improved arabic-to-english statistical machine translation by reordering post-verbal subjects for word alignment. *Machine Translation*, *26*(1-2), 105–120.

Carrera, J., Beregovaya, O. et Yanishevsky, A. (2009). Machine translation for cross-language social media.

Darwish, K. (2014). Arabizi detection and conversion to arabic. 217–224., Doha, Qatar. Association for Computational Linguistics.

Derczynski, L., Ritter, A., Clark, S. et Bontcheva, K. (2013). Twitter part-of-speech tagging for all : Overcoming sparse and noisy data. Dans *Recent Advances in Natural Language Processing (RANLP 2013)*, 198–206.

Diab, M. et Habash, N. (2007). Arabic dialect processing tutorial. Dans *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2007)*, 5–6. Association for Computational Linguistics (ACL).

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. Dans *Proceedings of the second international conference on Human Language Technology Research*, 138–145. Morgan Kaufmann Publishers.

Dyer, C., Weese, J., Setiawan, H., Lopez, A., Ture, F., Eidelman, V., Ganitkevitch, J., Blunsom, P. et Resnik, P. (2010). cdec : A decoder, alignment, and learning framework for finite-state and context-free translation models. Dans *Proceedings of the Association for Computational Linguistics System Demonstrations*, 7–12. Association for Computational Linguistics (ACL).

El-Kahlout, I. D. et Yvon, F. (2010). The pay-offs of preprocessing for german-english statistical machine translation. Dans *The International Workshop on Spoken Language Translation (IWSLT)*, 251–258.

Farzindar, A. et Roche, M. (2013). Les défis de l'analyse des réseaux sociaux pour le traitement automatique des langues. *Traitement Automatique des Langues*, *54*(3), 7–16.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., ORegan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G. et Tyers, F. M. (2011). Apertium : a free/open-source platform for rule-based machine translation. *Machine translation*, *25*(2), 127–144.

Gahbiche-Braham, S. (2013). *Amélioration des systèmes de traduction par analyse linguistique et thématique.* (Thèse de doctorat). Université Paris Sud.

Gao, Q. et Vogel, S. (2008). Parallel implementations of word alignment tool. Dans *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, 49–57. Association for Computational Linguistics (ACL).

Germann, U., Jahr, M., Knight, K., Marcu, D. et Yamada, K. (2001). Fast decoding and optimal decoding for machine translation. Dans *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, 228–235. Association for Computational Linguistics (ACL).

Ghoul, D. (2011). Outils génériques pour l'étiquetage morphosyntaxique de la langue arabe : segmentation et corpus d'entraînement.

Goldwater, S. et McClosky, D. (2005). Improving statistical mt through morphological analysis. Dans *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 676–683. Association for Computational Linguistics (ACL).

Gotti, F., Langlais, P. et Farzindar, A. (2014). Hashtag occurrences, layout and translation : A corpus-driven analysis of tweets published by the canadian go-

vernment. Dans *Language Resources and Evaluation Conference (LREC)*, 2254–2261.

Habash, N., Rambow, O. et Roth, R. (2009). Mada+ tokan : A toolkit for arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization. Dans *Proceedings of the 2nd international conference on Arabic language resources and tools (MEDAR)*, 102–109., Cairo, Egypt.

Habash, N. et Sadat, F. (2006). Arabic preprocessing schemes for statistical machine translation. Dans *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume : Short Papers*, 49–52. Association for Computational Linguistics (ACL).

Habash, N. et Sadat, F. (2012). Challenges for arabic machine translation. *by Abdelhadi Soudi, Ali Farghaly, Günter Neumann, and Rabih Zbib. Natural language processing. Amsterdam*, 73–94.

Habash, N. Y. (2010). Introduction to arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, *3*(1), 1–187.

Hailat, T., Al-Kabi, M. N., Alsmadi, I. M. et Al-Shawakfa, E. (2013). Evaluating english to arabic machine translators. Dans *Applied Electrical Engineering and Computing Technologies (AEECT), 2013 IEEE Jordan Conference on*, 1–6. Institute of Electrical and Electronics Engineers (IEEE).

Haithem, A. (2010). *Approche mixte pour la traduction automatique statistique.* (Mémoire de maîtrise). Université Stendhal-Grenoble III.

Han, B., Cook, P. et Baldwin, T. (2013). Lexical normalization for social media text. *Association for Computing Machinery Transactions on Intelligent Systems and Technology (TIST)*, *4*(1), 5.

Hassan et Darwish (2014). Statistical machine translation. Dans *Natural Language Processing of Semitic Languages*, 199–219. Springer Publishing Company.

Hebresha, H. A. et Ab Aziz, M. J. (2013). Classical arabic english machine translation using rule-based approach. *Journal of Applied Sciences*, *13*(1), 79.

Hutchins, J. (1997). From first conception to first demonstration : the nascent years of machine translation, 1947–1954. a chronology. *Machine Translation*, *12*(3), 195–252.

Jehl, L., Hieber, F. et Riezler, S. (2012). Twitter translation using translation-based cross-lingual retrieval. Dans *Proceedings of the Seventh Workshop on Statistical Machine Translation*, 410–421. Association for Computational Linguistics (ACL).

Jehl, L. E. (2010). *Machine Translation for Twitter*. (Mémoire de maîtrise). Speech and Language Processing School of Philosophy, Psychology and Language Studies, University of Edinburgh.

Johnson, J. H., Sadat, F., Foster, G., Kuhn, R., Simard, M., Joanis, E. et Larkin, S. (2006). Portage : with smoothed phrase tables and segment choice models. Dans *The Workshop on Statistical Machine Translation*, 134–137., New York City. Association for Computational Linguistics.

Kadri, Y. et Nie, J.-Y. (2006). Effective stemming for arabic information retrieval. Dans *The Challenge of Arabic for Natural Language Processing/ Machine Translation NLP/MT*, 68–74.

Knight, K. et Marcu, D. (2005). Machine translation in the year 2004. Dans ICASSP (dir.). *ICASSP (5) (In International Conference on Acoustics, Speech, and Signal Processing)*, volume 5, 965–968. Institute of Electrical and Electronics Engineers (IEEE).

Koehn, P. (2004). Pharaoh : a beam search decoder for phrase-based statistical machine translation models. Dans *Conference of the Association for Machine Translation in the Americas*, 115–124. Springer.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R. *et al.* (2007). Moses : Open source toolkit for statistical machine translation. Dans *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, 177–180. Association for Computational Linguistics (ACL).

Koehn, P., Och, F. J. et Marcu, D. (2003). Statistical phrase-based translation. Dans *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, volume 1, 48–54. Association for Computational Linguistics (ACL).

Langlais, P., Gotti, F. et Patry, A. (2006). De la chambre des communes à la chambre d'isolement : adaptabilité d'un système de traduction basée sur les segments. Dans *Les actes de TALN*, 217–226.

Larkey, L. S., Ballesteros, L. et Connell, M. E. (2002). Improving stemming for arabic information retrieval : light stemming and co-occurrence analysis. Dans *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, 275–282. Association for Computing Machinery (ACM).

Lavie, A. et Agarwal, A. (2007). Meteor : An automatic metric for mt evaluation with high levels of correlation with human judgments. Dans *Proceedings of the Second Workshop on Statistical Machine Translation*, 228–231. Association for Computational Linguistics (ACL).

Lee, Y.-S. (2004). Morphological analysis for statistical machine translation. Dans *Proceedings of the Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL 2004)*, 57–60. Association for Computational Linguistics (ACL).

Ling, W. (2015). *Machine Translation 4 Microblogs.* (Thèse de doctorat). Carnegie Mellon University.

Ling, W., Xiang, G., Dyer, C., Black, A. W. et Trancoso, I. (2013). Microblogs as parallel corpora. Dans *Association for Computational Linguistics (ACL)*, volume 1, 176–186.

Liu, F., Weng, F. et Jiang, X. (2012). A broad-coverage normalization system for social media language. Dans *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics : Long Papers-Volume 1*, 1035–1044. Association for Computational Linguistics (ACL).

Mayor, A., Alegria, I., De Ilarraza, A. D., Labaka, G., Lersundi, M. et Sarasola, K. (2011). Matxin, an open-source rule-based machine translation system for basque. *Machine translation*, *25*(1), 53–82.

Medhat, W., Yousef, A. H. et Korashy, H. (2014). Corpora preparation and stop-word list generation for arabic data in social network. Dans *Fourteenth Conference on Language Engineering (ESOLE 2014)*, Egypt, Cairo.

Mohammad, S. M., Salameh, M. et Kiritchenko, S. (2016). How translation alters sentiment. *Journal of Artificial Intelligence Research (JAIR)*, *55*, 95–130.

Mubarak, H. et Abdelali, A. (2016). Arabic to english person name transliteration using twitter. Dans *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Slovenia. European Language Resources Association (ELRA).

Och, F. J. et Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, *29*(1), 19–51.

Papineni, K., Roukos, S., Ward, T. et Zhu, W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. Dans *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318. Association for Computational Linguistics.

Pennell, D. L. et Liu, Y. (2014). Normalization of informal text. *Computer Speech & Language*, *28*(1), 256 – 277.

Quirk, C. et Moore, R. (2007). Faster beam-search decoding for phrasal statistical machine translation. *Machine Translation Summit XI*.

Refaee, E. et Rieser, V. (2015). Benchmarking machine translated sentiment analysis for arabic tweets. Dans *Student Research Workshop (SRW-2015)*, 71–78.

Sadat, F., Kazemi, F. et Farzindar, A. (2014a). Automatic identification of arabic language varieties and dialects in social media. *The 4th International Workshop on Natural Language Processing for Social Media of (SocialNLP 2014)*.

Sadat, F., Mallek, F., Sellami, R., Boudabous, M. M. et Farzindar, A. (2014b). Collaboratively constructed linguistic resources for language vari-ants and their exploitation in nlp applications–the case of tunisian arabic and the social media. Dans *Workshop on lexical and grammatical resources for language processing*, p. 102. Citeseer.

Sajjad, H., Darwish, K. et Belinkov, Y. (2013). Translating dialectal arabic to english. Dans *the Association for Computational Linguistics (ACL)*, volume 2, 1–6.

Salameh, M., Mohammad, S. M. et Kiritchenko, S. (2015). Sentiment after translation : A case-study on arabic social media posts. Dans *Human Language Technologies : The 2015 Annual Conference of the North American Chapter of the Association for Computational Linguistics (ACL)*, 767–777. Association for Computational Linguistics.

Salloum, W. et Habash, N. (2012). Elissa : A dialectal to standard arabic machine translation system. Dans *Coling (demos)*, 385–392.

Shannon, C. E. (1949). The mathematical theory of communication. *Urbana*.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. et Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. Dans *Proceedings of association for machine translation in the Americas*, volume 200.

Stolcke, A. *et al.* (2002). Srilm-an extensible language modeling toolkit. Dans *INTERSPEECH*, volume 3, 901–904. Citeseer.

Toral, A., Wu, X., Pirinen, T., Qiu, Z., Bicici, E. et Du, J. (2015). Dublin city university at the tweetmt 2015 shared task. *Tweet Translation Workshop at the International Conference Of the Spanish Society For Natural Language (SEPLN 2015)*.

Vauquois, B. et Boitet, C. (1985). Automated translation at grenoble university. *Computational Linguistics*, *11*(1), 28–36.

Wang, L. et Ng Hwee, T. (2013). A beam-search decoder for normalization of social media text with application to machine translation. Dans *The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (HLT-NAACL)*, 471–481.

Wang, Y.-Y. et Waibel, A. (1997). Decoding algorithm in statistical machine translation. Dans *Proceedings of the eighth conference on European chapter of*

the Association for Computational Linguistics, 366–372. Association for Computational Linguistics.

Zbib, R., Malchiodi, E., Devlin, J., Stallard, D., Matsoukas, S., Schwartz, R., Makhoul, J., Zaidan, O. F. et Callison-Burch, C. (2012). Machine translation of arabic dialects. Dans *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics : Human language technologies*, 49–59. Association for Computational Linguistics.

Zhang, M., Li, H., Kumaran, A. et Liu, M. (2012). Report of news 2012 machine transliteration shared task. Dans *Proceedings of the 4th Named Entity Workshop*, 10–20. Association for Computational Linguistics.