Amélioration des systèmes de traduction par analyse linguistique et thématique : application à la traduction depuis l'arabe

Souhir Gahbiche-Braham

UNIVERSITÉ PARIS SUD
ÉCOLE DOCTORALE D'INFORMATIQUE

THÈSE

présentée pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE PARIS SUD

Spécialité : Informatique

par

Souhir GAHBICHE-BRAHAM

Amélioration des systèmes de traduction par analyse linguistique et thématique

Application à la traduction depuis l'arabe

soutenue publiquement le 30 Septembre 2013

Directeur de thèse : François YVON
Co-directrice de thèse : Hélène BONNEAU-MAYNARD

Jury:

Président:	Pierre ZWEIGENBAUM	Directeur de Recherches, LIMSI-CNRS
Rapporteurs:	Emmanuel MORIN	Professeur, Université de Nantes
	Kamel SMAÏLI	Professeur, Université Nancy 2
Examinateur:	Laurent BESACIER	Professeur, Université J. Fourier
Invité:	Charles HUOT	Directeur de développement, TEMIS
Directeur:	François YVON	Professeur, Université Paris Sud
Co-directrice:	Hélène MAYNARD	Maître de Conf. HDR, Université Paris Sud

À ma fille Nour,

Remerciements

À l'issue de cette thèse, je souhaite remercier toutes les personnes ayant contribué à la mise en oeuvre de ces travaux.

En premier lieu, je tiens à remercier mon directeur de thèse, François Yvon pour la confiance qu'il m'a accordé en me confiant ces travaux de recherche, de m'avoir permis de découvrir le monde de la traduction automatique et de m'avoir toujours orienté et conseillé pendant cette thèse. Je tiens également à remercier Hélène Bonneau-Maynard, ma co-directrice de thèse pour sa confiance, son encadrement, ses idées et sa bonne humeur. Leur disponibilité, leurs conseils, leurs encouragements et leur suivi régulier m'ont permis de mener à terme ces travaux de recherche. J'ai été également extrêmement sensible à leurs qualités humaines d'écoute et de compréhension. C'était vraiment un plaisir d'avoir travaillé avec vous !

Je remercie mes rapporteurs Emmanuel Morin et Kamel Smaïli d'avoir accepté de relire ce manuscrit et pour les remarques intéressantes et les suggestions qu'ils ont proposées. Merci également à Laurent Besacier, Pierre Zweigenbaum et Charles Huot d'avoir accepté de faire partie du jury et pour l'intérêt qu'ils ont porté à mes travaux.

Je remercie tous les membres de l'équipe TLP pour la bonne ambiance dans laquelle j'ai pu travailler et pour les nombreuses discussions intéressantes que j'ai eues avec chacun d'entre eux ; sans oublier les anciens membres de TLP. Merci à tous !

Je souhaite aussi remercier les partenaires du projet SAMAR avec qui j'ai eu l'occassion de travailler : Leila Zighem, Wigdan Mekki, Sylvie Guillemin-Lanne, Cécile Woehrling, Hacene Cherfi, Jérôme Mainka, Dominique Ferrandini, Denis Teyssou, Nadine Lucas, Samir Matrouf, Yves Lepage, Ayadi Chabir, Fathi Debili, etc. La participation au projet SAMAR était une bonne opportunité enrichissante et m'a permis d'avoir une vue sur le monde industriel.

J'aimerais aussi remercier tous les membres du LIMSI pour les nombreuses discussions, les pauses, les cafés et la bonne humeur. Je voudrais remercier en particulier Chahnez pour son soutien, ses encouragements et ses conseils tout au long de ma thèse. Je remercie Houda, Asma et Asma, mes compatriotes de thèse, docteures et futur docteure. Une pensée à Kaouthar pour toutes les discussions, les histoires et les plaisanteries que nous continuons à partager malgré la distance.

Je remercie tous mes amis et ma famille. Une grande reconnaissance à mes parents, qui sans eux, je n'aurais jamais pû être ce que je suis : Merci pour tous les efforts et sacrifices que vous avez fait pour nous et que vous faites encore et toujours ! Merci pour tout ! Sans oublier ma soeur Sarra et mon frère Amine qui malgré la distance étaient toujours présents ;-) Un grand merci à mes parents, mes grands-parents, mes beaux-parents, mes beaux-frères et TOUTE ma famille pour leur soutien !

Enfin, mais non des moindres, le mot merci ne suffit pas pour remercier mon mari Walid pour son soutien, sa compréhension et ses encouragements durant ces années de thèse. Merci d'avoir été patient et présent. Finalement, il m'est impossible de ne pas remercier ma petite fille Nour, qui est venue au monde en pleine thèse et qui a su supporter toutes les contraintes qui lui étaient imposées dès sa naissance. Merci pour le plaisir et la joie que tu as fait entrer dans notre vie !

Un dernier merci à toutes les personnes qui ont participé à la réalisation de ce travail de près ou de loin ainsi que toutes les personnes que j'ai pu oublier.

Résumé

La traduction automatique des documents est considérée comme l'une des tâches les plus difficiles en traitement automatique des langues et de la parole. Les particularités linguistiques de certaines langues, comme la langue arabe, rendent la tâche de traduction automatique plus difficile. Notre objectif dans cette thèse est d'améliorer les systèmes de traduction de l'arabe vers le français et vers l'anglais. Nous proposons donc une étude détaillée sur ces systèmes. Les principales recherches portent à la fois sur la construction de corpus parallèles, le prétraitement de l'arabe, la détection des entités nommées et sur l'adaptation des modèles de traduction et de langue. Tout d'abord, un corpus comparable journalistique a été exploré pour en extraire automatiquement un corpus parallèle. Ensuite, différentes approches d'adaptation du modèle de traduction sont exploitées, soit en utilisant le corpus parallèle extrait automatiquement soit en utilisant un corpus parallèle construit automatiquement. Nous démontrons que l'adaptation des données du système de traduction permet d'améliorer la traduction.

Un texte en arabe doit être prétraité avant de le traduire et ceci à cause du caractère agglutinatif de la langue arabe. Nous présentons notre outil de segmentation de l'arabe, SAPA (Segmentor and Part-of-speech tagger for Arabic), indépendant de toute ressource externe et permettant de réduire les temps de calcul. Cet outil permet de prédire simultanément l'étiquette morpho-syntaxique ainsi que les proclitiques (conjonctions, prépositions, etc.) pour chaque mot, ensuite de séparer les proclitiques du lemme (ou mot de base). Nous décrivons également dans cette thèse notre outil de détection des entités nommées, NERAr (Named Entity Recognition for Arabic), et nous examions l'impact de l'intégration de la détection des entités nommées dans la tâche de prétraitement et la pré-traduction de ces entités nommées en utilisant des dictionnaires bilingues. Nous présentons par la suite plusieurs méthodes pour l'adaptation thématique des modèles de traduction et de langue expérimentées sur une application réelle contenant un corpus constitué d'un ensemble de phrases multicatégoriques.

Finalement, les systèmes de traduction améliorés arabe-français et arabe-anglais sont intégrés dans une plateforme d'analyse multimédia et montrent une amélioration des performances par rapport aux systèmes de traduction de base.

Mots-clés : Traitement automatique des langues, Traduction automatique de l'arabe, Prétraitement de l'arabe, Détection des entités nommées, Adaptation.

Abstract

Improvements for Machine Translation Systems Using Linguistic and Thematic Analysis : an Application to the Translation from Arabic

Machine Translation is one of the most difficult tasks in natural language and speech processing. The linguistic peculiarities of some languages makes the machine translation task more difficult. In this thesis, we present a detailed study of machine translation systems from arabic to french and to english. Arabic texts needs to be preprocessed before machine translation and this because of the agglutinative character of arabic language. Our principle researches carry on building parallel corpora, arabic preprocessing and adapting translation and language models. We propose a method for automatic extraction of parallel news corpora from a comparable corpora. Two approaches for translation model adaptation are explored using whether parallel corpora extracted automatically or parallel corpora constructed automatically. We demonstrate that adapting data used to build machine translation system improves translation. A prepocessing tool for arabic, SAPA (Segmentor and Part-of-speech tagger for Arabic), much faster than the state of the art tools and totally independant of any other external resource was developed. This tool predicts simultaneously morphosyntactic tags and proclitics (conjunctions, prepositions, etc.) for every word, then splits off words into lemma and proclitics. We describe also in this thesis, our named entity recognition tool for arabic, NERAr, and we focus on the impact of integrating named entity recognition in the preprocessing task. We used bilingual dictionaries to propose translations of the detected named entities. We present then many approaches to adapt thematically translation and language models using a corpora made of a set of multicategoric sentences.

Finally, improved machine translation systems from arabic to french and english are integrated in a multimedia platform analysis and shows improvements compared to basic machine translation systems.

Keywords : Natural Language Processing, Arabic Machine Translation, Arabic Preprocessing, Named Entity Recognition, Adaptation.

Sommaire

Introduction

Le Traitement automatique du langage naturel (TALN) ou des langues (TAL) regroupe à la fois la linguistique, l'informatique et l'intelligence artificielle. Cette discipline est apparue au début des années cinquante (Cori et Léon, 2002) aux États-Unis et est devenue un axe de recherche essentiel pour analyser et traduire la grande masse d'information qui évolue sans cesse. Cependant les enjeux cognitifs du traitement automatique des langues sont importants et varient selon les applications. De nos jours, il existe plusieurs applications de traitements des langues telles que la reconnaissance de l'écriture manuscrite, le résumé automatique, le traitement de la parole ou l'annotation sémantique, etc.

La langue est définie par le dictionnaire *Larousse* comme étant *un système de signes vocaux, éventuellement graphiques, propre à une communauté d'individus, qui l'utilisent pour s'exprimer et communiquer entre eux*. Le nombre de langues vivantes parlées dans le monde s'élève à 6.909 en 2009 (Lebert, 2010). Une langue naturelle est une langue "normale" utilisée oralement par des personnes dont elle est la langue maternelle. Dès la naissance, un être humain peut apprendre plusieurs langues simultanément (Dewaele, 2004). Dans une étude réalisée par Kahn (2011) sur 221 locuteurs, 35 % des locuteurs parlent une seule langue, 55 % des locuteurs parlent deux langues et 10 % parlent trois langues. La traduction d'une langue à une autre nécessite une connaissance préalable des deux langues. Aujourd'hui, cette contrainte n'est plus obligatoire grâce à la traduction automatique.

Bien que la traduction automatique soit loin d'être parfaite, elle suscite aujourd'hui un grand intérêt dans le domaine de la recherche. La traduction automatique des documents est considérée comme l'une des tâches les plus difficiles en traitement automatique des langues et de la parole. Ce domaine de recherche a été récemment renouvelé par l'apparition des techniques à base de corpus grâce à la disponibilité des textes bilingues parallèles. Ces textes parallèles représentent des textes dans une langue source et leurs traductions en langue cible.

Les systèmes de traduction statistiques reposent sur l'analyse statistique de corpus bilingues pour former des modèles stochastiques de correspondance entre une langue source et une langue cible. Ces modèles sont généralement formés à partir de corpus parallèles, c'est à dire à partir des exemples de textes sources alignés avec leurs traductions. Un système de traduction automatique apprend à traduire donc en se basant sur des exemples de traductions extraits de corpus bilingues. Le système de traduction recombine les fragments de ces exemples pour traduire une nouvelle phrase.

Les corpus bilingues peuvent être rares pour certaines paires de langues et sont limités à des domaines spécifiques. L'exploitation d'autres types de corpus – plus facilement accessibles – afin de construire des corpus bilingues suscite toujours de l'intérêt et c'est un domaine extrêmement actif aujourd'hui grâce notamment au nombre croissant de corpus disponibles (Rapp et al., 2012). Nous abordons cette tâche à la suite de nombreux travaux, en explorant différents types de corpus (monolingues, comparables) et différentes techniques pour construire des corpus parallèles du domaine journalistique à partir de ces données.

Dans la même langue, le même terme peut avoir plusieurs sens différents, et ceci dépend du contexte dans lequel ce terme est utilisé. L'exemple le plus classique dans ce contexte pour

le français est celui de l'*avocat*, qui peut être le fruit ou le métier. En arabe par exemple le mot عين (Eyn) peut être traduit par *un oeil* ou par *une source (d'eau)* ou par *un espion*, etc. Un même mot peut être utilisé dans deux contextes différents avec deux sens différents.

Dans le domaine de la traduction automatique, ce problème est très connu et c'est un problème de recherche toujours d'actualité. La probabilité de traduire correctement un mot polysémique est très faible si le système de traduction utilisé n'est pas approprié au contexte du texte à traduire. Par exemple si un système de traduction est appris sur des données de type politique, la traduction d'un document sur l'industrie agroalimentaire avec ce système donnera de mauvaises performances (avocat le fruit sera traduit par l'avocat le métier). Nous nous intéressons dans ce manuscrit à étudier ce phénomène, et plus particulièrement à l'amélioration de nos systèmes de traduction automatique et à leur *adaptation* au type de données à traduire.

Contexte de mes travaux de recherche

L'augmentation continue des volumes de textes disponibles, dans de nombreuses langues, motive le développement d'outils de traduction permettant d'effectuer des recherches en un seul langage sur des collections multilingues.

Dans ce contexte, le travail de recherche dans ce manuscrit vise à construire des systèmes de traduction statistiques de l'arabe vers le français et vers l'anglais et à améliorer la qualité de ces systèmes de traduction pour des applications de recherche documentaire multilingue dans des archives textuelles et vocales de documents journalistiques en langue arabe.

Le Projet SAMAR (2009-2012)

Cette thèse a été financée principalement par le projet Cap Digital SAMAR[1], un projet FUI[2]. L'objectif de ce projet est de développer une plateforme de traitement multimédia en langue arabe. Le projet SAMAR a été initié par l'Agence France-Presse[3] (AFP) avec la volonté d'ouvrir son portail d'information à des contenus multilingues écrits en langue arabe.

Ce projet regroupe onze partenaires : des industriels et des laboratoires de recherche. L'*AFP* est le fournisseur des flux multimédia radio et télévisuels. Ces flux sont la source essentielle des corpus arabe, français et anglais. Du côté du traitement de la parole : la société *Vecsys*, spécialiste de la reconnaissance vocale et du traitement automatique de la parole est le fournisseur des transcriptions manuelles du texte à partir de la parole et l'entreprise *Vocapia* est le partenaire qui travaille sur la transcription automatique de la parole. Pour l'aspect linguistique on cite le *Laboratoire du Langage, Langues et Cultures d'Afrique Noire*, (*LLACAN*) pour l'analyse de l'arabe littéraire et dialectal ainsi que l'*Institut National des Langues et Civilisations Orientales* (l'*Inalco*) pour la validation de modèles. Pour la traduction automatique : le *Laboratoire d'Informatique pour la Mecanique et les Sciences de l'Ingénieur* (*LIMSI*) pour la modélisation de traduction et le *Groupe de REcherche en Informatique, Image, Automatique et Instrumentation de Caen*, le *Greyc* pour les alignements de textes. En ce qui concerne la fouille et la classification du texte : la société *Temis* pour l'extraction de connaissances à partir de textes, la société *Antidot* pour la recherche cross-lingue et la société *Mondeca* pour la gestion des ontologies. Finalement la société *Nuxeo* pour la gestion de contenu multimédia et l'intégration.

Les technologies actuelles ne permettent pas un traitement de l'information provenant du monde arabe. Ces contenus doivent ainsi être traduits pour pouvoir être intégrés à d'autres plateformes d'information internationale. Une intégration de ces nouvelles sources

[1] http://www.samar.fr/
[2] Fonds Unique Interministériel
[3] http://www.afp.com

d'information, sous-entend que le contenu doit être exploitable comme les autres contenus du portail d'information. Une analyse linguistique poussée des contenus permettrait donc d'indexer des informations multilingues et de les rendre accessibles via la recherche d'information en ligne.

Le public visé par ce projet est principalement l'AFP, mais également tous les médias arabes de la bordure Méditérranéenne et du Moyen-Orient.

Nous nous plaçons bien sûr dans le contexte de la traduction automatique. Précisément, le projet vise à la mise au point d'un système de traduction de l'arabe vers le français et vers l'anglais. Il faudra donc développer des méthodes d'apprentissage de modèles de traduction à partir de corpus non-parallèles (corpus comparables, corpus monolingues) qui nous permettraient de contourner la relative rareté des corpus parallèles.

La figure présente le schéma fonctionnel du projet SAMAR ainsi que les niveaux de traitement avec les modules technologiques produits chez chacun des partenaires.

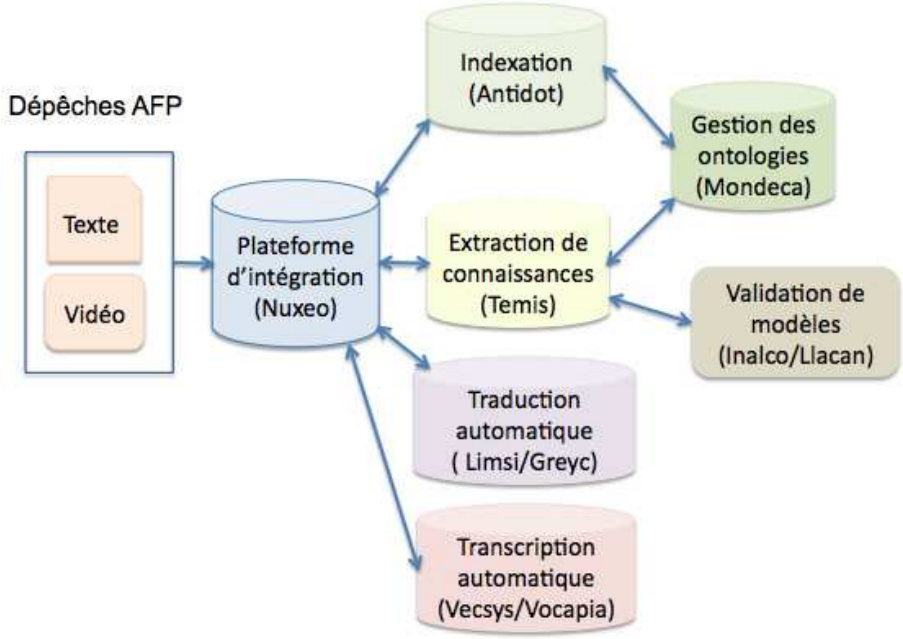

Figure 1: *Schéma fonctionnel du projet SAMAR et interaction entre les partenaires.*

Le cadre du projet SAMAR a permis de mettre en place un contexte précis pour mes recherches qui se sont focalisées essentiellement sur le traitement automatique de la langue arabe, ainsi que sur la construction et l'amélioration des systèmes de traduction automatique à partir de l'arabe.

Défis de la langue arabe

La langue arabe est une langue morphologiquement riche et complexe. C'est une langue sémitique, qui a la particularité d'avoir un vocabulaire à base de racines de mots trilitères consonantiques. À cette forme de base, peuvent s'ajouter des préfixes, suffixes, ainsi que des clitiques. Les clitiques et affixes sont agglutinés au mot de base pour former des mots de plus

en plus complexes voire des phrases, comme par exemple le mot وسيكتبها (*Et il va l'écrire*), qui est constitué de deux proclitiques و et س, un préfixe ي et un enclitique ها – sachant que le pronom *l'* refère à un nom féminin.

L'absence de voyellation de l'arabe dans les textes courants provoque une inflation de solutions d'analyses des mots composés lorsque ceux-ci sont analysés indépendamment les uns des autres. La langue arabe est modélisée par un ensemble de règles linguistiques et grammaticales complexes (Kouloughli, 1994).

Attia (2008) présente une étude détaillée de la langue arabe, sa morphologie et ses particularités. Plusieurs niveaux d'ambiguïté existent dans la langue arabe (Attia, 2008; Farghaly et Shaalan, 2009), ce qui rend son traitement automatique, une tâche très difficile. Zbib et Soudi (2012) présentent quelques particularités de la langue arabe.

Plan du manuscrit

Ce manuscrit comprend deux parties principales. La première partie théorique est constituée de deux chapitres. Le premier chapitre présente la traduction automatique statistique, alors que le deuxième chapitre présente la langue arabe et sa morphologie.

Dans le chapitre 1, nous décrivons brièvement la traduction automatique statistique ainsi que les modèles probabilistes qui constituent un système de traduction automatique statistique. Le décodage est également présenté et plus particulièrement le fonctionnement d'un système de traduction statistique à base de segments. Nous présentons de même les principaux types d'évaluation de la traduction automatique.

Le chapitre 2 est dédié à la langue arabe et sa morphologie. Nous présentons les deux formes principales de l'arabe : l'arabe moderne standard (MSA) et l'arabe dialectal. La morphologie de la langue arabe ainsi que ses caractéristiques principales sont présentés dans ce chapitre. Par la suite, nous décrivons les principales ressources numériques existantes et disponibles pour le traitement automatique de l'arabe. Finalement, nous passons en revue les travaux effectués sur le traitement automatique de l'arabe, en particulier le prétraitement de la langue arabe ainsi que la détection des entités nommées en arabe.

La deuxième partie présente essentiellement nos contributions, et est constituée de cinq chapitres. Chaque chapitre permet de répondre à une parmi les questions que nous nous sommes posées tout au long de cette thèse.

Comme nous l'avons déjà mentionné, la constitution et l'amélioration des systèmes de traduction arabe-français et arabe-anglais est la principale tâche à laquelle nous nous sommes intéressés et que nous avons traité dans ce manuscrit. Des corpus comparables en arabe, français et anglais sont fournis au LIMSI dans le cadre du projet SAMAR. Ces données comparables doivent être utilisées pour construire un système de traduction arabe-français ainsi qu'un système de traduction arabe-anglais. La première question que nous nous sommes posées était donc :

> **Q1:** Comment peut-on utiliser un corpus comparable pour adapter un système de traduction automatique statistique?

Le chapitre 3 répond à cette question en décrivant notre méthode d'extraction de corpus parallèle à partir d'un corpus comparable. D'une part, l'approche d'extraction de corpus a été décrite en détails. D'autre part, on présente deux approches différentes pour adapter le modèle de traduction au type des données à traduire. La première méthode consiste à utiliser le nouveau corpus parallèle extrait automatiquement alors que la deuxième approche consiste à utiliser un corpus parallèle traduit automatiquement. Une série d'expériences a été effectuée et montre que les deux approches d'adaptation améliorent les performances de traduction jusqu'à six points BLEU.

Lors de la construction d'un système de traduction, plusieurs étapes sont effectuées : le prétraitement des données en langues source et cible, les alignements, la construction des modèles et l'optimisation. Nous avons particulièrement remarqué que le prétraitement de la langue arabe est très lent et nécessite l'utilisation de plusieurs machines (en parallèle) lorsque nous avons beaucoup de données à traiter (le nombre de phrases dépasse quelques milliers de phrases, à partir de 5 mille phrases). Ce qui nous a mené à la deuxième question :

Q2: Comment peut-on améliorer les vitesses de prétraitement de l'arabe?

Le chapitre 4 est dédié à une présentation détaillée de notre nouvel outil de prétraitement de la langue arabe basé sur les CRF. Cet outil utilise un modèle qui prédit simultanément les étiquettes morphosyntaxiques ainsi que les proclitiques existants pour chaque mot. Une comparaison de notre outil à deux outils de segmentation les plus utilisés a été réalisée. Les résultats montrent que notre outil est aussi performant que les autres outils au niveau de la segmentation, de l'étiquetage morphosyntaxique, ainsi que pour la tâche de prétraitement pour la traduction automatique. La particularité de notre outil est qu'il est beaucoup plus rapide que les deux autres approches et qu'il est indépendant de toute autre ressource : il n'utilise ni une base de données pour l'analyse morphologique ni un outil de désambiguïsation.

L'étude détaillée des résultats des traductions, nous a permis de noter que parmi les mots inconnus dans la sortie de traduction, un quart de ces mots sont des entités nommées. Nous nous sommes intéressés à essayer de résoudre ce problème, ce qui constitue la troisième question :

Q3: Comment peut-on améliorer la traduction des entités nommées?

Le chapitre 5 présente un nouvel outil de détection des entités nommées en arabe à base de CRF. Cet outil a été évalué et comparé à l'état de l'art. Une adaptation du détecteur des entités nommées a été effectuée. Le prétraitement de l'arabe contient donc une nouvelle étape pour traiter les entités nommées. Un ensemble d'expériences de traduction automatique dans lesquelles nous avons combiné la segmentation avec la détection des entités nommées a été ensuite effectué. Finalement, une évaluation de l'impact du prétraitement des entités nommées sur la traduction automatique ainsi qu'une analyse détaillée des résultats ont été réalisées.

Le corpus utilisé pour nos expériences est constitué de dépêches AFP produites par le bureau arabe de l'Agence France Presse pouvant avoir différentes catégories (culture, politique, finances, santé, etc.). Nous avons pensé à adapter le système de traduction selon la catégorie des données à traduire. Ceci nous a mené à se poser la quatrième question :

Q4: Comment peut-on adapter nos systèmes de traduction à différents domaines et est ce que l'adaptation améliore la traduction automatique?

Dans le chapitre 6, nous étudions différents types d'adaptation de modèles de traduction et de modèles de langue en utilisant les dépêches de l'AFP. Une évaluation et une comparaison des différents systèmes de traduction spécifiques à des catégories de l'AFP ont été réalisées. Deux approches de classification ont été proposées : une classification a priori et une classification automatique basée sur la classification naïve bayésienne en utilisant l'algorithme des k-moyennes.

Cette thèse a été initiée principalement par le projet SAMAR, qui a pour objectif de construire une plateforme d'analyse multimédia pour la langue arabe. Le dernier chapitre répond donc à la question suivante :

Q5: Comment peut-on intégrer nos travaux de recherche dans une application réelle?

Ce dernier chapitre a pour objectif de présenter comment nos travaux sont intégrés dans le cadre d'une application réelle. On décrit les travaux que nous avons effectué et livré pour le projet SAMAR. Particulièrement, on présente la chaîne de traitement complète pour traduire une dépêche AFP sur la plateforme SAMAR. On présente également quelques résultats d'évaluation ainsi que des résultats obtenus dans le cadre d'autres évaluations (Quaero).

Le manuscrit est clos par une conclusion générale et quelques perspectives pour poursuivre ces travaux.

Part I

Traitement Automatique de la langue Arabe

CHAPITRE 1

Traduction Automatique Statistique

La traduction automatique est l'une des tâches les plus intéressantes et les plus difficiles du traitement automatique des langues (TAL). Les premiers travaux sur la traduction automatique ont débuté presqu'en même temps que l'apparition des premiers ordinateurs. Le besoin de traduction est apparu essentiellement pour des raisons militaires. À cette époque, les travaux sur la traduction automatique se sont inspirés des travaux en cryptographie dans un contexte de guerre froide. *Warren Weaver* fut l'un des pionniers de la traduction automatique, il a lancé les premiers travaux dans ce domaine juste après la seconde guerre mondiale en 1947.

> ...
> Also knowing nothing official about, but having guessed and inferred considerable about, powerful new mechanized methods in cryptography - methods which I believe succeed even when one does not know what language has been coded - one naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say "This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode."
>
> *Warren Weaver, 1947*

Le premier système de traduction automatique fut présenté par IBM en 1954; il traduisait des phrases du russe vers l'anglais (Hutchins, 1998).

La diffusion de l'usage des ordinateurs, d'abord pour des usages professionnels, puis, de manière concomitante à l'émergence d'Internet, dans des usages privés ou de loisir ont aidé à la progression des travaux sur les systèmes de traduction automatique. Les systèmes à base de règles (RBMT, *Rule Based Machine Translation*) sont apparus en même temps que les premières utilisations d'Internet au début des années 1970. Ces systèmes ont besoin de beaucoup de connaissances linguistiques, de dictionnaires ainsi que d'un ensemble de règles de transfert d'une langue à l'autre; ces règles sont dépendantes de la paire de langue considérée. Systran[1] est l'un des systèmes de traduction à base de règles. Vers les années 1980, les systèmes à base d'exemples (EBMT, *Example-Based Machine Translation*) ont paru. Les systèmes EBMT (Nagao, 1984) fonctionnent en se basant sur un ensemble d'exemples de traduction (corpus parallèle) qui est parcouru lors de la traduction. Si la phrase à traduire y existe alors sa traduction est

[1] www.systran.fr

extraite, sinon le système détecte, rassemble et adapte les fragments d'exemples à traduire à partir des exemples de traduction.

La facilité d'accès à des traductions de textes écrits et l'augmentation de puissance de calcul des machines ont ouvert la voie pour le développement des approches de traduction automatique statistique. La construction d'un système de traduction statistique est beaucoup plus rapide qu'un système à base de règles.

La traduction automatique statistique (SMT, *Statistical Machine Translation*) a été introduite par Brown et al. (1990) dans les années 1990. C'est une approche à base de corpus, et elle est caractérisée par l'utilisation de méthodes d'apprentissage automatique.

Au fil des temps, la traduction automatique a fait beaucoup de progrès. Une vue d'ensemble et un historique de la traduction automatique ont été présentés par Dorr, Jordan et Benoit (1998) et par Hutchins (2001). Plus récemment, un compte rendu détaillé et complet des différents stades de développement de la traduction automatique depuis sa création a été donné par Hutchins (2010). Aujourd'hui, il existe un grand nombre de systèmes de traduction automatique permettant de traduire la plupart des langues. La traduction automatique devient un besoin d'utilisation de plus en plus répandu et les systèmes de traduction en ligne sont de plus en plus utilisés. La différence entre ces systèmes de traduction réside dans les performances de chacun d'eux selon le type de données à traduire. En effet, les traductions générées par les machines sont loin d'être parfaites et les recherches sur l'amélioration de ces systèmes de traduction automatique sont toujours d'actualité.

Les travaux effectués tout au long de cette thèse ainsi que les systèmes de traduction utilisés pour les expériences sont des systèmes de traduction à base de segments. Ce chapitre présente donc brièvement la traduction automatique statistique et en particulier les principes de la traduction automatique à base de segments. Beaucoup de références existent dans la littérature et décrivent les systèmes de traduction automatiques statistiques en détails. On se limite donc dans ce chapitre à présenter seulement les grandes lignes importantes qui serviront de base pour les travaux effectués dans cette thèse. Dans la section 1.1, on présente ce qu'est un système de traduction automatique statistique. Le modèle de traduction et le modèle de langue sont les principaux modèles qui constituent un système de traduction statistique et sont présentés dans la section 1.2. Le décodage ainsi que le fonctionnement d'un système statistique sont décrits dans la section 1.3. Finalement la section 1.4 présente les deux principales méthodes d'évaluation de la traduction automatique. Ce chapitre est clôturé par une conclusion.

1.1 La traduction automatique statistique

C'est l'approche de traduction automatique dominante à l'état de l'art dans le monde de la recherche. On peut se référer à Knight et Marcu (2005) ou Lopez (2008) pour avoir un aperçu des travaux sur la traduction automatique statistique. Un tutoriel complet sur les systèmes de traduction automatique statistique, qui décrit entre autres les méthodes de construction d'un système de traduction automatique statistique, a été réalisé par Koehn (2010). De même Allauzen et Yvon (2012) présentent les modèles probabilistes qui constituent un système de traduction statistique ainsi qu'une description détaillée du fonctionnement de ce dernier.

Un système de traduction statistique apprend à traduire en se basant sur des exemples de traductions, sous la forme de textes parallèles, c'est-à-dire d'ensembles de paires de phrases qui sont des traductions les unes des autres. Des correspondances sont tout d'abord définies entre chaque mot -ou ensemble de mots- en langue source et sa traduction en langue cible; ce processus est appelé l'alignement. Ces alignements font partie des données d'entrée au système au moment de l'apprentissage. Une fois que l'apprentissage a eu lieu, le système est capable de traduire une nouvelle phrase - qui n'a pas été parmi les exemples de traduction à l'apprentissage - en recombinant des morceaux de ces exemples.

Un système de traduction automatique statistique est basé sur des scores numériques. C'est un système à base de modèles statistiques construits automatiquement à partir de corpus monolingues et bilingues. Deux modèles probabilistes constituent essentiellement un système de traduction : un modèle de traduction et un modèle de langue. Si on considère que s est la langue source et t est la langue cible, le modèle de traduction probabiliste appris à partir de données bilingues est représenté par $P(s|t)$, et le modèle de langue probabiliste appris à partir de données monolingues en langue cible est représenté par $P(t)$. Généralement, un système statistique est plus performant lorsque la quantité de données augmente.

La traduction se définit par la recherche d'un texte en langue cible t ayant la plus forte probabilité $P(t|s)$ d'être la traduction d'un texte en langue source s. Le problème de traduction est reformulé par le Théorème de Bayes (équation 1.1) :

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

Sachant que $P(s)$ est indépendante du texte en langue cible t, la traduction la plus probable t^* est alors obtenue en maximisant la probabilité de $P(t|s)$. Le problème de la recherche de la traduction optimale se reformule alors selon l'équation 1.2 :

$$t^* = argmax_t P(t|s) = argmax_t P(t,s) = argmax_t P(s|t)P(t) \tag{1.2}$$

Dans ce qui suit, on présentera les modèles probabilistes qui sont au coeur d'un système de traduction automatique statistique.

1.2 Les modèles probabilistes

Dans cette section, on donne un aperçu sur les deux principaux modèles probabilistes qui constituent un système de traduction.

1.2.1 Les modèles de traduction

Cette partie du chapitre présente les modèles de traduction à base de mots et les modèles de traduction à base de segments.

1.2.1.1 Les modèles de traduction à base de mots (word-based SMT)

Les premiers modèles de traduction automatique statistique sont des modèles à base de mots (Brown et al., 1990; Brown et al., 1993). Les modèles de traduction à base de mots reposent sur un modèle de traduction qui traite les mots un par un. L'unité de traduction pour ces modèles est le mot.

L'un des problèmes majeurs des modèles de traduction à base de mots est le non déterminisme des appariemments mot à mot. Si on ne considère pas le contexte du texte à traduire, le modèle peut générer des confusions, par exemple le mot *livre* peut être l'unité monétaire de certains pays - traduite en anglais par *pound* - ou le *livre* qui représente l'ouvrage - traduit en anglais par *book*.

Un autre problème pour les modèles à base de mots est qu'un ensemble de mots en langue source peut être traduit par un seul mot en langue cible ou vice-versa. Un mot en langue source peut s'aligner avec plusieurs mots dans la langue cible, ou à aucun mot. Les alignements consistent à faire correspondre chaque mot ou segment en langue source, avec sa traduction en langue cible. La figure 1.1 montre un exemple d'alignement des mots dans une phrase en français et sa traduction en anglais.

Dans une phrase en langue source et sa traduction en langue cible, l'ordre des mots n'est pas toujours le même. La figure 1.2 présente un exemple d'une phrase en arabe et sa

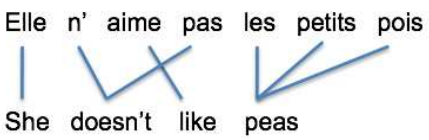

Figure 1.1: *Exemple d'alignement mot-à-mot entre une phrase en français et sa traduction en anglais.*

traduction en français : *ne* et *pas* sont alignés avec *don't*; de même *les*, *petits* et *pois* sont alignés avec *peas*.

On note, sur la figure 1.2, que la phrase en arabe a la forme *Verbe-Sujet-Complément d'Objet* alors qu'en français la phrase a la forme *Sujet-Verbe-Complément d'Objet* : l'ordre des mots n'est pas forcément le même pour une phrase en langue source et sa traduction dans une langue cible. La tâche de réordonner les mots après la traduction afin de les mettre dans le bon ordre est appelée le réordonnancement.

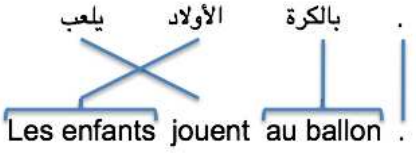

Figure 1.2: *Exemple d'alignement d'une phrase en arabe et sa traduction en français. Il faut noter que la phrase en arabe est écrite de gauche à droite.*

Brown et al. (1993) ont défini 5 modèles d'alignements à base de mots. Le modèle IBM1 considère que tous les liens d'alignement sont équiprobables, et se base sur les probabilités de traduction lexicale. Toutes les positions de réordonnancement sont possibles. Dans le modèle IBM2, le réordonnancement est basé sur les positions relatives des mots alignés entre la source et la cible. Il introduit une dépendance entre la valeur d'alignement et la position du mot aligné dans la phrase. IBM3 introduit la notion de traduction d'un mot par plusieurs mots, et intègre la notion du mot *null*. IBM4 introduit le concept de dépendance entre les traductions : la position de la traduction d'un mot dans une phrase est définie en fonction de la position de la traduction du mot précédent. IBM5 est une version améliorée du modèle IBM4.

Les modèles à base de mots présentent quelques inconvénients. Les modèles à base de mots ont un modèle de réordonnancement faible. Un autre inconvénient des modèles à base de mots c'est que les appariements mot-à-mot des modèles à base de mots sont non-déterministes et génèrent des problèmes d'ambiguïté lexicale. Un mot peut être traduit par un ou plusieurs mots, donc l'unité de traduction qui est le mot n'est pas la meilleure solution : traduire un groupe de mots aide à la désambiguïsation et améliore la traduction.

Tous ces inconvénients ont motivé les recherches afin de proposer les modèles de traduction à base de segments.

1.2.1.2 Les modèles de traduction à base de segments (*phrase-based SMT*)

Les modèles de traduction à base de segments (*phrase based*) représentent les modèles les plus utilisés à l'état de l'art (Koehn, Och et Marcu, 2003; Zens, Och et Ney, 2002).

Le modèle de traduction (ou table de traduction) est appris à partir d'un corpus parallèle. Le texte bilingue est d'abord aligné, à l'aide d'un outil d'alignement (Och et Ney, 2003; Gao et

Vogel, 2008) pour avoir, pour chaque segment de la phrase source, la traduction qui lui correspond. L'alignement se fait dans les deux sens source-cible et cible-source, donc en plus des alignements mot-à-mot et mot-à-plusieurs – utilisés dans les modèles à base de mots – les alignements plusieurs-à-plusieurs seront considérés. Une fois les alignements effectués, ils sont symétrisés afin de trouver une intersection et une union de ces alignements. Cette étape de symétrisation représente l'alignement final qui sera utilisé pour l'apprentissage. La figure 1.3 présente une grille d'alignement d'une phrase en espagnol avec sa traduction en anglais (Knight et Koehn, 2004). Cette grille présente un alignement mot-à-mot. Plusieurs alignements de segments peuvent être extraits à partir de cette figure. L'alignement peut s'effectuer par des unités constitués par des groupes de mots, appelés segments. À partir de la figure 1.3, on peut constituer les paires de segments suivants : (Maria no daba, Mary did not slap), (no daba, did not slap) ou encore (bofetada a la bruja, slap the witch).

Figure 1.3: *Exemple d'alignement d'une phrase, extrait de (Knight et Koehn, 2004). Les alignements dans cette grille sont : (Maria, Mary), (no, did not), (daba una bofetada, slap), (a la, the), (bruja, witch), (verde, green).*

Une fois les alignements obtenus, un score de probabilité est calculé pour chaque segment. Chaque segment en langue source peut avoir plusieurs hypothèses de traduction en langue cible. Sachant que s représente le segment en langue source et t représente le segment en langue cible, on note par $c(s,t)$ le nombre de paires de segments dans lesquels apparaît un segment donné sur l'ensemble du corpus. La probabilité de t sachant s est calculée selon l'équation 1.3 :

$$P(t|s) = \frac{c(s,t)}{\sum_{t_i} c(s,t_i)} \tag{1.3}$$

La probabilité qu'un mot ou un groupe de mots dans la langue source soient traduits par un autre dans la langue cible est donnée par le modèle de traduction $P(t|s)$.

1.2.2 Le modèle de langue

La modélisation du langage consiste à construire un modèle probabiliste pour les séquences grammaticales d'une langue donnée, permettant de donner un sens à la probabilité d'un mot dans son contexte, et donc de trouver le mot le plus probable sachant ceux qui le précèdent. Le modèle de langue est estimé sur un corpus monolingue. Le but du modèle de langue est de calculer la probabilité $P(t_1^I)$ d'une séquence de mots $t_1^I = t_1 ..., t_i, ..., t_I$. Cette probabilité

est définie par la formule 1.4 :

$$P(t_1^I) = P(t_1) \prod_{i=2}^{I} P(t_i|h_i) \tag{1.4}$$

avec $h_i = t_1, ..., t_{i-1}$ l'historique du mot t_i.

Les modèles de langue représentent une partie fondamentale de la traduction automatique statistique. Ils influent sur le choix des mots. Si un mot n'existe pas dans le modèle de langue (parcequ'il n'a pas été observé dans le corpus d'entraînement), alors ce mot ne sera pas choisi par le système de traduction. Pour chaque séquence, dans le modèle de langue, un score lui est attribué. Ce score correspond à la probabilité que cette séquence apparaîsse dans un texte.

Un modèle de langue n-gramme établit ses prédictions sur la base de fenêtres de taille fixe contenant n mots; généralement le n varie de 1 à 5. Pour chaque séquence, un score de probabilité est calculé en prenant compte des n-1 mots qui précèdent le mot courant pour chaque position dans la phrase cible : ce score représente la dépendance de chaque mot par rapport aux n-1 mots qui le précèdent. Pour chaque séquence de mots, un score de probabilité indiquant la qualité syntaxique et lexicale lui est attribué.

Souvent, des séquences de mots sont correctes mais n'existent pas dans le modèle de langue et donc une probabilité nulle leur est attribuée. Des méthodes, appelées méthodes de lissage (*smoothing*) ont été développées afin de résoudre ce problème et ne pas attribuer des valeurs nulles aux séquences qui n'apparaissent pas dans le corpus d'apprentissage. Parmi ces approches de lissage, on cite le lissage Kneser-Ney (Kneser et Ney, 1995), voir aussi (Chen et Goodman, 1996) pour une étude comparative des principales méthodes de lissage.

En pratique, les outils les plus utilisés pour la construction de modèles de langue : SRILM (Stolcke, 2002) ou aussi IRSTLM (Federico et al., 2007). D'autres outils sont utilisés pour l'estimation du modèle de langue comme KenLM (Heafield, 2011). Cette binarisation sert à avoir des modèles de langues de moindre taille et donc permet de traduire plus rapidement.

1.3 Le décodage

Une fois que le modèle de langue et le modèle de traduction sont construits, ces deux modèles sont combinés pour trouver, pour chaque phrase source, la meilleure hypothèse de traduction. La recherche de la meilleure hypothèse de traduction consiste à choisir la phrase cible qui maximise la probabilité conditionnelle $P(t \mid s)$ définie dans l'équation 1.2. Ceci revient à choisir l'hypothèse la mieux évaluée par ces modèles.

La figue 1.4 représente une vue d'ensemble d'un système de traduction automatique dans laquelle un modèle de traduction est combiné avec un modèle de langue. La recherche de la meilleure hypothèse de traduction revient donc à résoudre un problème d'optimisation combinatoire. Les modèles de traduction sont très limités dans les aspects linguistiques. La segmentation des phrases en des mots, ou des segments, ainsi que le réordonnancement peuvent causer des problèmes d'ambiguïté. Le décodage des modèles de traduction automatique est donc un problème NP-complet (Knight, 1999).

Afin de résoudre ce problème de complexité, il est nécessaire de réduire l'espace de recherche afin d'avoir des solutions efficaces.

Plusieurs travaux ont étudié ce problème, Wang et Waibel (1997) résolvent ce problème en implémentant un algorithme de recherche de la meilleure hypothèse à base de piles. D'autres utilisent les transducteurs à états finis pondérés pour implémenter un modèle d'alignement (Kumar et Byrne, 2003). Le problème de décodage est transformé en un problème de programmation linéaire par Germann et al. (2001) qui implémentent un algorithme de recherche par faisceau. L'approche de Crego et Mariño (2006) consiste à apprendre des règles de réordonnancement en observant des alignements en source et en cible puis effectuer un

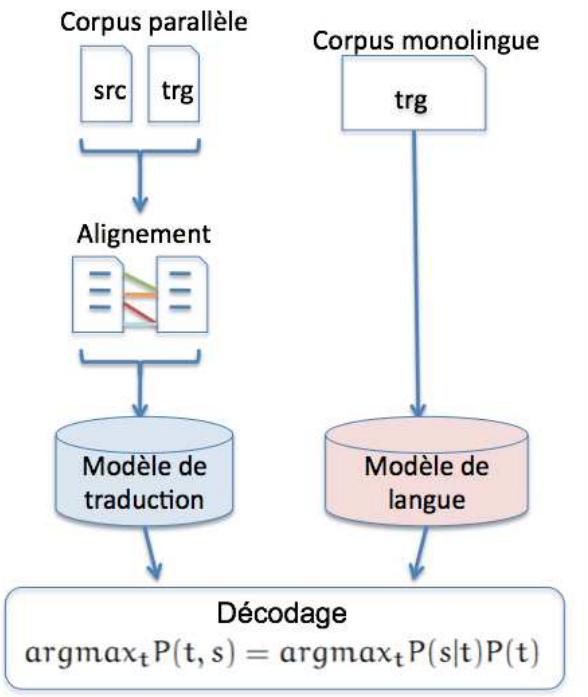

Figure 1.4: *Système de traduction automatique.*

décodage monotone. Des algorithmes de programmation dynamique avec élagage ont été implémentés par Koehn (2004) et par Moore et Quirk (2007).

Kumar, Deng et Byrne (2006) proposent une implémentation de la traduction automatique à base de segments sous la forme d'une cascade de transducteurs finis qui s'effectue en trois étapes : une étape de segmentation dans laquelle la phrase source est segmentée en plusieurs segments contigües; une étape de traduction lexicale dans laquelle chaque segment source est traduit, et finalement une étape de réordonnancement dans laquelle les segments en langue cible sont permutés dans leur ordre final.

1.3.1 Fonctionnement des systèmes de traduction à base de segments

Dans la pratique, le décodage – dans le cas des modèles à base de segments – consiste tout d'abord à traduire les groupes de mots.

Koehn (2010) définit la formule de traduction d'un système de traduction à base de segments par l'équation 1.5 :

$$t_{best} = \text{argmax} \prod_{i=1}^{I} \phi(s_i|t_i) d(start_i - end_{i-1} - 1) p_{LM}(t) \tag{1.5}$$

où $\forall$ i = 1...I, s_i est la phrase source, t_i est la phrase cible, $start_i$ et end_i sont les positions des phrases. $\phi(s_i|t_i)$ est la probabilité donnée par le modèle de traduction, d est donnée par le modèle de réordonnancement et p_{LM} est donnée par le modèle de langue.

Les modèles de langue, de traduction et de réordonnancement sont tous considérés par le décodeur comme des traits avec des poids appropriés. Chacun des trois modèles délivre un coût qui sera combiné avec les autres pour trouver la meilleure traduction (Allauzen et Yvon,

2012). La combinaison linéaire des trois modèles donne lieu à la formule 1.6 :

$$c(s,t,a) = \lambda_p c_p(s,t,a) + \lambda_r c_r(s,t,a) + \lambda_l c_l(t) \tag{1.6}$$

comme précédemment s représente la source, t représente la cible et a représente les alignements. $c_p(s,t,a)$ représente le coût donné par le modèle de traduction, $c_r(s,t,a)$ représente le coût donné par le modèle de réordonnancement et $c_l(t)$ représente le coût donné par le modèle de langue. Les λ_x représentent les paramètres associés à chacun des modèles. Afin de réduire le coût de cette fonction, il est nécessaire de régler les paramètres λ_x.

L'optimisation – ou réglage de paramètres – est effectuée sur un corpus de développement en langue source et sa traduction en référence. L'optimisation consiste à trouver la combinaison qui donne la meilleure hypothèse de traduction. À chaque itération, les valeurs λ_x changent et la traduction du corpus de développement est évaluée par rapport à la référence. Plusieurs algorithmes d'optimisation ont été proposés pour résoudre ce problème. Parmi les plus utilisés pour la traduction, on trouve MERT, *Minimum Error Rate Training* (Och, 2003) ainsi que des versions améliorées de ce dernier comme par exemple (Macherey et al., 2008; Foster et Kuhn, 2009).

Lorsque MERT est lancé, il utilise un premier jeu de λ_x et récupère les n meilleures hypothèses. MERT essaies de trouver une combinaison des λ_x qui permet de maximiser BLEU (Papineni et al., 2002). La forme particulière de la fonction donnant le score BLEU (voir section 1.4) en fonction des λ_x est connue par MERT : c'est une fonction qui évalue uniquement les morceaux de traductions. À chaque itération, les nouveaux poids λ_x sont comparés aux anciens, s'ils sont différents, le décodage est relancé afin de produire une nouvelle traduction. Le processus s'arrête quand il converge.

MIRA, *Margin Infused Relaxed Algorithm* (Crammer et Singer, 2003), est une version générique. Cet algorithme n'était pas destiné à être utilisé pour la traduction, il a été proposée au début pour l'apprentissage automatique ensuite a été appliqué à la traduction, tandis que MERT a été proposé à l'origine pour la traduction. MIRA a été proposé pour l'optimisation des systèmes de traduction automatique lorsque l'on souhaite multiplier le nombre de caractéristiques utilisées pour évaluer les traductions (Watanabe et al., 2007; Chiang, Knight et Wang, 2009).

1.3.2 Moses

Moses[2] est un outil de traduction automatique statistique à base de segments. Il permet de construire automatiquement un système de traduction automatique pour n'importe quelle paire de langues. C'est un outil libre développé collaborativement et décrit dans Koehn et al. (2007). Il implémente un algorithme de recherche par faisceau (*beam search*) Germann et al. (2001).

Pour décoder, Moses a besoin (i) d'un modèle de traduction construit à partir d'un corpus parallèle aligné, (ii) d'un modèle de langue construit à partir d'un corpus monolingue en langue cible, et (iii) d'un modèle de réordonnancement.

Pour construire ces modèles, tout d'abord, le corpus bilingue est tokenisé[3], c'est-à-dire la ponctuation est séparée des mots, les majuscules sont transformées en minuscules, et si la langue traitée est morphologiquement complexe, une analyse morphologique du texte est effectuée. La tokenisation s'effectue à l'aide d'outils spécifiques pour chaque langue traitée. Ensuite, le corpus bilingue est aligné : pour chaque phrase, des correspondances sont effectuées. Les alignements sont effectués en utilisant l'outil d'alignement mgiza++[4]

[2]Moses est téléchargeable sur `http://www.statmt.org/moses/`

[3]La tokenisation consiste à transformer le texte brut en un texte où toutes les unités et sous unités sont séparées par des espaces. Ce texte sera utilisé par les modèles de langue et de traduction.

[4]Téléchargeable sur http://sourceforge.net/projects/mgizapp/

(Gao et Vogel, 2008). Une fois les alignements obtenus, le modèle de traduction et le modèle de réordonnancement sont crées avec Moses.

Avant le décodage, il est nécessaire d'optimiser les paramètres à l'aide d'un corpus de développement. L'algorithme MERT permet d'estimer les poids des paramètres afin de trouver la meilleure combinaison de poids qui conduisent à trouver la meilleure traduction du corpus de développement. Pour chaque combinaison des poids, la traduction obtenue est évaluée par rapport au corpus de développement de référence.

Différentes versions de Moses ont été proposées afin de rajouter, modifier ou corriger des paramètres. Moses est le décodeur qui a été utilisé pour construire les systèmes de traduction automatique utilisés dans l'ensemble des expériences effectuées et présentées dans ce manuscrit. Tout au long de ces travaux, nous avons utilisé différentes version de Moses afin d'avoir des versions plus stables et de pouvoir utiliser de nouveaux paramètres.

1.4 Évaluation de la traduction automatique

Un texte généré par un processus automatique doit être évalué. Il existe deux types principaux d'évaluation : l'évaluation manuelle et l'évaluation automatique.

L'évaluation manuelle – ou subjective – est effectuée par des humains, qui jugent de la qualité de traduction en se basant sur un ensemble de critères bien définis telle que la fluidité, la fidélité et l'informativité du texte traduit. Au début, l'évaluation de la qualité de la traduction automatique se faisait exclusivement par des humains. Ce type d'évaluation nécessite l'intervention d'experts bilingues. Chaque expert doit évaluer une grande quantité de traductions et chaque traduction doit être évaluée par plusieurs experts afin de s'assurer de la fiabilité des résultats. Tous ces travaux demandent beaucoup de travail manuel et beaucoup de temps.

La quantité de données à évaluer est devenue de plus en plus importante. Aujourd'hui, les chercheurs évaluent et comparent plusieurs systèmes de traduction quotidiennement. Il était donc nécessaire de mettre au point des techniques d'évaluation automatique. C'est le seul type d'évaluation possible qui permet de donner des résultats instantanés à faible coût. De plus, les évaluations automatiques sont reproductibles, ce qui n'est pas le cas des évaluations manuelles : un même expert peut juger la même phrase différemment. Une évaluation automatique permet donc de comparer deux systèmes en utilisant exactement un même corpus de référence. Toutefois, l'évaluation manuelle est toujours considérée le moyen le plus fiable et le plus sûr. Ce type d'évaluation est utilisée dans certaines compagnes d'évaluation (Callison-Burch et al., 2009).

L'évaluation automatique est effectuée à l'aide de métriques automatiques. Ces dernières comparent la traduction générée automatiquement avec une traduction de référence, c'est-à-dire une traduction réalisée par des humains. BLEU (*BiLingual Evaluation Understudy*) proposée par Papineni et al. (2002) est la métrique la plus utilisée, et se calcule en comparant la traduction automatique à une ou plusieurs références. Le calcul est basé sur une comparaison de courtes séquences de mots (n-grammes) pour chaque phrase.

BLEU tient compte non seulement des correspondances de mots simples entre la phrase traduite automatiquement et la phrase de référence, mais aussi des correspondances entre les n-grammes. Ce concept permet de recompenser les phrases où l'ordre des mots est proche de l'ordre des mots dans la référence. BLEU est une mesure de précision, dont le principe est de calculer le degré de similitude entre une traduction automatique et une ou plusieurs références en se basant sur la précision n-gramme: si une traduction automatique est identique à une des références, alors le score BLEU est égal à 100. Par contre, si aucun des n-grammes de la traduction n'est présent dans aucune référence, alors le score BLEU est égal

à 0. Le score BLEU est calculé par la formule suivante :

$$\text{BLEU} = \text{BP} \,.\, exp(\sum_{n=1}^{N} w_n \log \text{précision}_n) \tag{1.7}$$

$$\text{où BP } (\textit{Brevity Penalty}) = \begin{cases} 1 & \text{si } c > e \\ e^{(1-\frac{r}{c})} & \text{si } c \leqslant e \end{cases}$$

sachant que BP est une pénalité calculée pour défavoriser les hypoyhèses de traduction courtes par rapport aux références ; les phrases courtes seraient trop avantagées sans cette précision. c est la longueur de l'hypothèse de traduction, r représente la longueur du corpus de référence, et w_n représentent les poids des différentes précisions des n-grammes. Cette précision représente le nombre de n-grammes de l'hypothèse de traduction présents également dans une ou plusieurs références, divisé par le nombre de n-grammes total de l'hypothèse de traduction.

Une autre métrique intéressante, METEOR (*Metric for Evaluation of Translation with Explicit Ordering*) proposée par Banerjee et Lavie (2005), Lavie et Agarwal (2007) est basée sur des alignements entre les unigrammes d'une hypothèse de traduction et ceux d'une traduction en référence. L'alignement s'effectue entre les mots d'une même forme orthographique, les mots de même racine ainsi que les synonymes grâce aux ressources linguistiques qui sont utilisées. Des tables de paraphrases peuvent être également utilisées lors de l'évaluation avec METEOR. Une fois que l'alignement est établi, le score METEOR est calculé comme suit:

$$\text{METEOR} = (1 - P_{en}) \frac{P \text{ X } R}{\alpha P + (1-\alpha)\ R} \tag{1.8}$$

où

- α est un facteur de pondération
- R est le rappel, qui représente le nombre de correspondances entre unigrammes divisé par la taille de traduction référence
- P est la précision, qui représente le nombre de correspondances entre unigrammes établies par l'alignement divisé par la taille de l'hypothèse de traduction et de R
- P_{en} est un coefficient de pénalité qui permet de réduire le score des traductions n'ayant pas de correspondances d'ordre plus élevé que l'unigramme.

$$P_{en} = \gamma \text{ X } (\frac{\text{Nb(chunks)}}{\text{Nb(unigrammes alignés}})^{\beta} \tag{1.9}$$

  - *Nb(chunks)* est le nombre de de sous-ensembles de mots contigus de l'hypothèse de traduction alignés avec un sous-ensemble de mots contigus de la référence.
  - *Nb(unigrammes alignés)* représente le nombre d'unigrammes de l'hypothèse de traduction mis en correspondance avec des mots de la référence.
  - α, β et γ sont des paramètres de METEOR qui nécessitent une phase d'optimisation. Dans le cas où seulement les unigrammes sont alignés, le facteur de pénalité vaut γ.

D'autres métriques sont aussi utilisées comme TER (*Translation Error Rate*) proposée par Snover et al. (2006a) et qui permet de mesurer le nombre minimum d'opérations qui doivent

être apportées à une hypothèse de traduction pour que celle-ci soit identique à l'une des références correspondantes. Le score TER est défini par :

$$\text{TER} = \frac{\text{N op}}{\text{N}_{ref}} \tag{1.10}$$

où N op est le nombre minimum d'opérations, et N_{ref} est la taille moyenne en mots des références.

Une description plus détaillée des mesures d'évaluation automatique sont présentés par Lavecchia (2010). L'évaluation manuelle reste le type d'évaluation le plus efficace et le plus sûr. Cependant, l'évaluation automatique est le type d'évaluation le plus utilisé puisqu'elle nécessite moins de ressources (humaines) et elle est plus rapide (instantanée). Toutefois la constitution manuelle des corpus de référence est nécessaire. Pour pouvoir évaluer un texte automatiquement, il est indispensable d'avoir un corpus de référence constitué ou vérifié manuellement.

1.5 Conclusion

Les travaux de recherche sur la traduction automatique et l'amélioration des systèmes de traduction est toujours un sujet d'actualité. En témoignent, le grand nombre de conférences nationales et internationales liées au Traitement Automatique de Langues, mais aussi les journaux de TAL et de traduction automatique. Ce chapitre donne un aperçu sur l'évolution historique de la traduction automatique depuis son apparition.

Dans ce chapitre, nous avons tout d'abord présenté les systèmes de traduction automatique statistique. Un système de traduction automatique statistique est constitué essentiellement d'un modèle de traduction et d'un modèle de langue. Différents types de modèles de traduction existent comme les modèles hiérarchiques (Chiang, 2005) ou les modèles à base de syntaxe (Yamada et Knight, 2001). Dans ce chapitre, nous avons présenté deux parmi ces modèles de traduction utilisés pour la traduction statistique : les modèles à base de mots et les modèles à base de segments. Les travaux effectués dans cette thèse se déroulent dans le cadre de la traduction automatique statistique à base de segments. Nous avons donc présenté le principe de fonctionnement d'un système de traduction à base de segments, et plus particulièrement le décodeur *Moses* qui a été utilisé pour la construction des systèmes de traduction automatique utilisés dans l'ensemble des expériences présentées dans ce manuscrit.

Finalement, nous avons présenté les principales méthodes utilisées pour l'évaluation de la traduction automatique.

CHAPITRE 2

La langue arabe et sa morphologie

L'arabe est la langue sémitique contemporaine la plus parlée de nos jours avec plus de 300 millions de locuteurs (Habash, 2010). C'est la cinquième langue parlée dans le monde (Chung, 2008; Lewis, F. Simons et D. Fennig, 2013). Dans la plupart des langues, comme le français ou l'anglais, la langue écrite et parlée est la même. La particularité de la langue arabe est que l'arabe écrit est différent de l'arabe parlé.

La langue arabe se présente sous deux formes principales : l'arabe littéraire et l'arabe dialectal. L'arabe littéraire est la langue officielle du monde arabe, tandis que l'arabe dialectal – spécifique pour chaque pays – est la vraie langue parlée dans le monde arabe.

D'après Farghaly et Shaalan (2009), l'arabe littéraire se répartit en deux catégories : l'arabe classique et l'arabe moderne standard (MSA). L'arabe classique est utilisé dans les prières et les textes religieux, et constitue la base de l'arabe moderne standard. L'arabe moderne standard est une forme plus récente de l'arabe classique; elle est utilisée dans les médias, les journaux, les salles de classe et l'administration.

La langue arabe, originaire de la péninsule arabique (Arabie Saoudite, Yémen, Oman, etc.), doit sa véritable expansion à la diffusion de l'islam. Auparavant, la langue arabe était utilisée comme moyen d'expression et d'échange dans la poésie, la prose et les histoires orales. La même langue parlée est écrite mais avec des accents différents et des variations linguistiques mineures au niveau de l'écrit (Kouloughli, 2007). Elle était à cette époque dans une forme assez proche de l'arabe moderne standard.

C'est avec l'avènement de l'islam que la langue arabe a connu un véritable essor. La langue arabe classique est née lorsque l'islam fut révélé, en 610, par la révélation du Coran formulé en arabe. Cette époque était appelée par certains historiens et linguistes, la première métamorphose de la langue arabe. La langue arabe est devenue une langue officielle du monde musulman en 685 quand le calife *Oumeya Abd Al Malik Ibn Marwan* arriva à la capitale du monde musulman, Damas, avec pour objectif de centraliser son pouvoir politique : il a imposé donc l'arabe comme unique langue officielle. Le calife entreprend des réformes de l'écriture par la suite et prend de grandes décisions concernant les signes écrits. À partir du VIII^{e} siècle une codification de la grammaire fixa la langue dans sa forme classique définitive et facilita la propagation de la langue par l'enseignement partout où l'islam a pu pénétrer (Djili, 2011). C'est à cette époque que les premiers traités et dictionnaires sont apparus.

Entre le VIII^{e} et le X^{e} siècle, les sciences et techniques islamiques se sont développées.

Dans la maison de la sagesse[1] à Baghdad, des manuscrits grecs, de philosophie et de sciences, furent traduits en arabe : c'est la seconde métamorphose de la langue arabe. Farghaly (2010) présente un compte rendu détaillé et complet sur l'apparition de la langue arabe ainsi que sa structure. L'arabe est une langue de civilisation qui a duré plus de quatorze siècles, et était arrivée jusqu'à l'occident entre le VIII^{e} et le X^{e} siècles.

La langue Arabe s'est étendue sur plusieurs continents à des peuples non arabes, et est devenue la langue officielle de plusieurs pays. La figure 2.1 montre les pays du monde qui ont actuellement pour langue officielle la langue arabe.

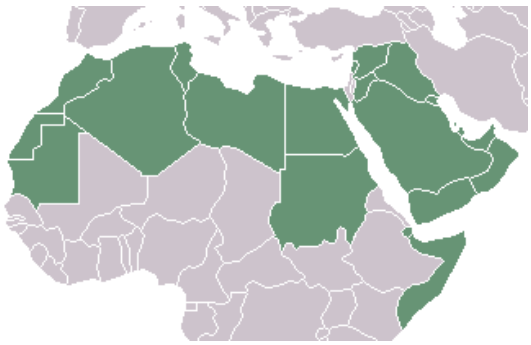

Figure 2.1: *Le monde arabe*[2].

Les travaux effectués dans cette thèse concernent essentiellement le traitement automatique de l'arabe. Ce chapitre est consacré donc à la définition et à la présentation de la langue arabe et de ses spécificités. La section 2.1 présente la langue arabe sous ses deux formes utilisées : l'arabe dialectal, et l'arabe moderne standard. On présentera également les ressources numériques de l'arabe disponibles et pouvant être utilisées pour le traitement automatique des langues dans la section 2.2. La section 2.3 dédiée au traitement automatique de l'arabe, donne un état de l'art et un compte rendu des travaux existants sur le traitement automatique de l'arabe et en particulier sur la traduction automatique de l'arabe et la détection des entités nommées.

2.1 La langue arabe

L'arabe moderne standard est fondé syntaxiquement, morphologiquement et phonologiquement sur l'arabe classique. Lexicalement, l'arabe moderne standard est plus récent (Habash, 2010). L'arabe moderne standard (MSA), ou arabe formel, est la forme de l'arabe utilisée formellement dans les articles de presse, les documents administratifs ou aussi la littérature. C'est la langue enseignée à l'école et commune à tous les pays qui ont pour langue officielle la langue arabe. L'arabe dialectal ou parlé est la langue parlée utilisée dans la vie quotidienne pour communiquer. Elle est apprise comme langue maternelle par les locuteurs arabophones.

[1] La maison de la sagesse était un lieu de collecte, de diffusion, de copie et de traduction de la littérature.

[2] http://fr.wikipedia.org/wiki/Monde_arabe

2.1.1 L'arabe dialectal

كانت العرب وان جمع جميعها اسم انهم عرب، فهم مختلفو الالسن بالبيان متباينو المنطق والكلام.

محمد بن جرير الطبري

"Les Arabes, même étant compris sous la même appellation d'Arabes, parlaient un langage différent et avaient une logique différente les uns par rapport aux autres.",

Mohamed Ibn Jarir Al-Tabari

Dans le monde arabe, chaque pays dont la langue principale est l'arabe a son propre dialecte. Dans un même pays, le dialecte peut être différent d'une région à une autre avec de petites variations et quelques mots différents. Si deux pays arabes sont voisins, alors plus on se rapproche des frontières, plus le dialecte se rapproche du dialecte du pays voisin. Dans la région de Annaba qui se situe au Nord-Est de l'Algérie, par exemple, le dialecte de Annaba est plus proche du tunisien que de l'algérien (Meftouh, Bouchemal et Smaïli, 2012). Toutefois, les deux pays partagent exactement la même langue écrite qui est l'arabe moderne standard.

Plusieurs classifications des dialectes ont été proposées pour le monde arabe. Habash (2010) a classifié les dialectes comme suit:

- L'égyptien concerne les dialectes égyptien et soudanais.
- Le levantin inclut les dialectes libanais, syrien, jordanien, palestinien et israélien.
- Le golfe inclut les dialectes du Koweit, les Émirates, le Bahrein, l'Arabie Saoudite et le Qatar.
- L'afrique du nord inclut les dialectes marocain, algérien, tunisien, lybien et mauritanien.
- L'irakien
- Le yéménite
- Le maltais est la seule variante de l'arabe qui est considérée comme une langue séparée et qui est écrite avec les caractères romains.

Le dialecte égyptien est compréhensible presque par tous les arabes puisque l'Égypte est le plus grand producteur arabe de films, de séries et de chansons. Les séries et chansons égyptiennens passent très souvent dans les chaînes télévisées maghrébines. Un maghrébin comprend donc parfaitement l'égyptien, mais en général un égyptien ne comprend pas les dialectes du Maghreb. Il pourra peut être comprendre seulement quelques mots qui dérivent de l'arabe littéraire.

Dans le dialecte arabe, il y a des mots qui dérivent d'autres langues. Les dialectes arabes varient d'un pays à un autre sur plusieurs dimensions : ils dépendent surtout de l'histoire de chaque pays et de son emplacement géographique. Prenons par exemple, la Tunisie, voisine de l'Italie qui était mise sous protectorat français et placée sous souveraineté de l'Empire ottoman. En dialecte tunisien, le mot *piscine* dérive du français et est dit *piscine* en tunisien. De même pour le mot *machine* qui dérive de l'italien et est dit *mekina* en dialecte tunisien, ou

aussi les mots *violoniste* ou *tailleur* qui dérivent du turc et sont dits respectivement *kemanji* ou *tarzi* en tunisien. Le dialecte tunisien comprend également plusieurs termes qui dérivent du berbère comme par exemple *fakrun* pour dire *tortue*.

Le dialecte, comme toute autre langue, se développe et s'adapte à chaque époque. On a donc souvent de nouveaux termes qui apparaissent et qui peuvent dériver d'autres langues; on trouve dans le dialecte tunisien par exemple des verbes français qui ont été adaptés à la langue arabe comme par exemple *Il démarre* est dit en tunisien *ydémarri*. De même, en dialecte algérien, *Essaie* est dit *Sayyi*.

Avec l'apparition d'Internet et des nouvelles technologies, le dialecte devient de plus en plus une langue écrite dans les forums, les SMS, le chat ou aussi dans les messages électroniques. Le dialecte peut être écrit soit en caractères arabes, ou aussi en caractères latins (arabe translitteré) si l'utilisateur n'a pas l'habitude d'utiliser le clavier arabe ou a des problèmes d'encodage lors de l'utilisation des caractères arabes (Diab et Habash, 2012). Toutefois, même s'il est écrit, le dialecte reste de l'arabe informel.

En arabe dialectal, il existe des graphies qui ne font pas partie de l'alphabet arabe. Ces graphies peuvent faire partie de noms propres de villes ou de personnes, comme par exemple la lettre *g* en dialecte tunisien ou algérien.

2.1.2 L'arabe moderne standard

L'arabe moderne standard – appelé aussi arabe formel – est la forme écrite de l'arabe. L'arabe s'écrit de droite à gauche et les lettres sont liées entre elles. Il n'existe pas de distinction entre majuscules et minuscules en alphabet arabe.

L'alphabet arabe est constitué de 28 graphies. Chaque lettre en arabe peut avoir quatre manières différentes de s'écrire selon la position de la lettre dans le mot. La forme de la lettre varie selon qu'elle apparaît au début, au milieu ou à la fin du mot, ou aussi si elle est isolée. Le tableau 2.1 montre les quatre formes différentes de la lettre ب (*b*).

Début	Milieu	Fin	Isolée
بـ	ـبـ	ـب	ب

Tableau 2.1: *Les quatre formes d'écriture de la lettre* ب.

Généralement, on utilise le terme de consonnes pour désigner les lettres de base de l'alphabet arabe (au nombre de trois dans la plupart des cas, voir section 2.1.2.3), et de voyelles pour désigner les diacritiques. D'une manière générale les voyelles ne sont pas écrites. En arabe moderne standard, le texte peut être voyellé, c'est-à-dire que des diacritiques peuvent être ajoutées au dessus et au dessous des lettres, comme par exemple كتب devient كَتَبَ après voyellation.

2.1.2.1 La voyellation et la diacritisation

Les notions de voyelles et consonnes n'ont pas vraiment de signification pour la langue arabe, mais sont néanmoins généralement utilisées pour en faciliter la compréhension.

Les voyelles sont des signes diacritiques qu'on ajoute aux lettres arabes (comme l'accent par exemple en français comme é ou ù, ...). Toutes les lettres en arabe peuvent être diacritisées. L'utilisation des diacritiques est optionnelle, mais parfois nécessaire pour lever une ambiguïté sémantique et/ou syntaxique. Par exemple la voyellation de la phrase كتب الولد permet de savoir si le sens de la phrase est كَتَبَ الوَلَدُ (Le garçon a écrit) ou كُتُبُ الوَلَدِ (Les livres du garçon). Debili, Ben Tahar et Souissi (2008); Debili, Ben Tahar et Souissi (2007) calculent le coût moyen du caractère exprimé en nombre de frappes, et montrent que la voyellation est coûteuse (1,43 pour l'arabe voyellé contre 1,03 pour l'arabe non voyellé). Généralement les textes écrits dans les journaux, documents administratifs, livres, magazines, etc. ne sont pas voyellés. Une personne connaissant la langue arabe et le sens des mots peut lire facilement un texte non voyellé : c'est en comprenant le contexte que l'on arrive à lire un texte non voyellé.

La voyellation ou diacritisation complète est utilisée dans les livres de débutants pour apprendre l'arabe et aussi dans les textes religieux. Les différents types de diacritiques sont :

- Les voyelles courtes : ـَ (a), ـُ (u), ـِ (i) ;
- Les voyelles longues : ا (A), و (U), ي (I) ;
- Le tanwiin, pour les mots indéfinis c'est le dédoublement d'une voyelle courte en fin de mot : ـً (an), ـٌ (un), ـٍ (in) ;
- Les signes de syllabation :
  - Le sukun (ـْ) est placé au dessus d'une lettre en arabe pour indiquer que cette lettre ne porte aucune voyelle.
  - Le chadda (ـّ) indique le dédoublement d'une lettre : le chadda est placé sur la lettre en question au lieu de dédoubler cette lettre.

La même forme d'un mot peut avoir plusieurs sens différents si la voyellation change. Par exemple دَرَسَ (il a étudié) est différent de دَرْسٌ (une leçon).

2.1.2.2 La structure des phrases en arabe

En langue arabe, il y a deux types de phrases : les phrases verbales et les phrases nominales. Les phrases verbales servent à indiquer un évènement ou une action. Elles commencent par un verbe suivi d'un sujet et d'un complément; ce dernier est optionnel. La phrase verbale est celle que l'on rencontre le plus souvent dans l'expression courante.

Les phrases nominales en arabe ne contiennent pas de verbe, il est sous entendu. En arabe, le verbe *être* est implicite, les verbes ne sont pas obligatoires pour construire une phrase. Les phrases nominales sont constituées d'un sujet et d'un attribut (adjectif qualificatif, complément circonstanciel, ...). Les phrases nominales sont affirmatives, par exemple الولد

في المدرسة est traduite par *L'enfant est à l'école*, mais le verbe est absent de la phrase en arabe (*L'enfant à l'école*).

L'utilisation des phrases verbales, comme les phrases nominales, dépend du contexte dans lequel la phrase est exprimée. Dans le style journalistique, les phrases verbales sont les plus utilisées.

2.1.2.3 La morphologie de l'arabe

Les mots en arabe se représentent sous forme d'agglutinations de proclitiques et d'enclitiques autour d'une forme de base. Le lexique arabe comprend trois catégories de mots : verbes, noms et particules (adverbes, conjonctions, prépositions). Plusieurs références existent sur la morphologie de l'arabe. Une description détaillée de la grammaire et la morphologie de l'arabe a été, par exemple, réalisée par Kouloughli (1994) et par Watson (2002). Plus récemment, une étude approfondie de la morphologie et de la syntaxe arabe a été réalisée par Attia (2012) qui aborde les différentes stratégies de développement de l'analyse morphologique en arabe. Il élabore une description des principales structures syntaxiques de la langue arabe, l'ordre des mots, l'accord, les dépendances, ainsi que le problème redoutable de désambiguïsation syntaxique en identifiant les sources d'ambiguïtés.

Un mot est constitué d'une base - ou lemme, racine - (verbe, nom) à laquelle s'agglutinent des affixes (préfixes et suffixes) et des clitiques (proclitiques et enclitiques). Attia (2008) propose une pyramide d'ambiguïté (figure 2.2) qui suppose que le caractère riche et complexe des inflexions de l'arabe et l'agglutination des affixes et des clitiques permet de réduire l'ambiguïté plutôt que de l'augmenter. La figure 2.2 montre que le taux d'ambiguïté baisse, en moyenne, si le mot contient plus de morphèmes.

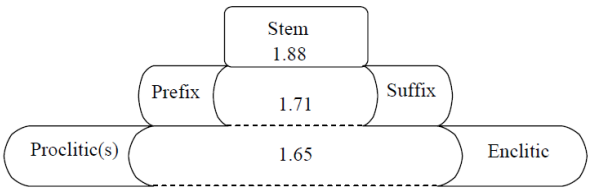

Figure 2.2: *Pyramide d'ambiguïté par Attia (2008).*

Les racines de mots sont souvent ambigües comme par exemple la racine كتب (ktb) peut être كَتَبَ (il a écrit) ou كُتُبُ (des livres) ou كُتِبَ (il a été écrit). Par contre lorsque des clitiques lui sont ajoutées, l'ambiguïté est réduite comme par exemple يكتبه ne peut être que *Il l'écrit*.

Un mot en arabe peut former une phrase à lui seul. Les mots en arabe ont généralement la forme suivante :

$$\text{PROC} + \text{PRE} + [\text{BASE}] + \text{SUF} + \text{ENC}$$

BASE : représente la forme de base pour chaque mot. En langue arabe la majorité des racines sont des mots trilitères, c'est-à-dire constitués de trois consonnes. Il existe aussi des racines bilitères et quadrilitères. La racine renvoie en principe à une même notion, par exemple pour كتب la notion d'*écrire*. La racine est habillée de différentes façons avec des voyelles, ajouts des lettres, etc. pour former des mots suivant des modèles appelés schèmes

(صيغة/ وزن). Ces schèmes sont nominaux ou verbaux.

D'après Neyreneuf et Al-Hakkak (2001), il y a neuf façons de dériver un verbe. À partir de cette racine, il est possible de former des verbes, noms, adjectifs, etc.

PROC : représente les proclitiques. Un proclitique représente un mot rattaché au mot suivant, alors qu'il peut être détaché, comme par exemple les conjonctions de coordination ou les prépositions. D'après Habash (2010), il existe quatre niveaux de clitisation qui peuvent être appliqués à un mot en respectant un ordre bien déterminé comme suit:

$$QST + [CNJ + [PRT + [DET + PRE + [BASE] + SUF + ENC]]]$$

Au niveau le plus profond, on a *DET* qui représente l'article défini ال (*Al*). Ce dernier ne s'attache jamais à un verbe ou à une préposition.

Ensuite, *PRT* représente la classe des particules. Il existe trois types de particules :

- س (*s*) représente la marque du futur et ne s'applique qu'aux verbes,
- ب (*b*, avec) et ك (*k*, comme) représentent des prépositions qui sont généralement attachées à des noms, adjectifs et quelques autres variantes mais jamais aux verbes
- ل (*l*, pour) qui s'attache aux verbes et aux noms.

Au niveau suivant, le proclitique *CNJ* représente les conjonctions de coordination و (*w*, et) et ف (*f*, alors), qui peuvent être rattachées à n'importe quel mot, indépendamment de sa fonction dans la phrase.

Finalement, le clitique *QST* est réservé pour les formes interrogatives. La lettre ا (Â, est ce que) est attachée au premier mot dans n'importe quelle phrase qui sera transformée en question. Cette dernière est moins fréquente que les autres proclitiques.

ENC : représente les enclitiques, ils sont attachés à la fin de la forme de base. Ces enclitiques sont des pronoms. Les enclitiques pronominaux peuvent être attachés aux noms (comme possessifs) ou aux verbes et prépositions (comme objets). En arabe, on distingue en particulier le dual du pluriel. Il existe un pronom personnel spécifique pour le dual et les verbes peuvent être conjugués au dual (masculin, féminin) comme au singulier (masculin, féminin) et au pluriel (masculin, féminin).

PRE : représente les préfixes. Selon le dictionnaire *Larousse*, *un préfixe est un affixe qui se place devant un mot base pour constituer avec lui un nouveau mot, appelé dérivé – comme par exemple le mot "refaire" constitué du préfixe re et du verbe faire*. Les préfixes font partie du mot et peuvent être soit des préfixes verbaux comme dans le mot يكتب (il écrit), ou des préfixes nominaux comme dans مكتب (un bureau). Ces deux mots sont engendrés d'un même concept sémantique qui est l'écriture (كتب). Pour le premier exemple, le préfixe ي refère à la marque de la troisième personne du singulier (il) à l'inaccompli (المضارع). Dans le deuxième exemple, le préfixe م ajouté au mot de base كتب indique le nom de lieu.

SUF : représente les suffixes, ils font également partie du mot et expriment généralement la marque du pluriel et du féminin.

Par exemple le mot اتصدقونها, traduit en français par *" Est ce que vous la croyez ?"* peut être segmenté de la façon suivante :

ا	ت	صدق	ون	ها
PROC	PRE	BASE	SUF	ENC
Est ce que	-	croire	vous	la

Dans cet exemple, le proclitique est constitué d'une conjonction interrogative, le préfixe représente un préfixe verbal du temps de l'inaccompli, le suffixe exprime la marque du pluriel et l'enclitique représente le pronom complément de verbe. Le mot اتصدقونها est traduit par une phrase en français.

Nguyen et Vogel (2008) analysent un corpus parallèle anglais-arabe, la partie anglais contient 6,2M tokens[3] et 68K de types de mots[4] alors que les 5,2M de tokens en arabe correspondent à 155K types de mots. Si on considère donc un texte constitué de mots agglutinés, ce texte est constitué d'un vocabulaire très grand. Par contre, si le texte est segmenté et les clitiques sont séparés du mot de base, ceci permet de réduire énormément le vocabulaire.

2.1.3 Autres caractéristiques de la langue arabe

Une autre caractéristique de la langue arabe est qu'elle est une langue *pro-drop* car elle permet l'omission des pronoms sujets. Le terme *pro-drop* a été introduit par Chomsky (1981) et provient de l'expression *pronoun dropping*. Ce pronom peut être soit un véritable pronom suffixe, soit il est sous-entendu (on dit qu'il est *caché* ضمير مستتر), ce qui est le cas pour la troisième personne du singulier lorsque le sujet n'est pas énoncé sous la forme d'un nom après le verbe (Neyreneuf et Al-Hakkak, 2001).

Considérons l'exemple donné par Attia (2008), اكلت الدجاجة (la poule a mangé). C'est une phrase verbale constituée d'un verbe suivi d'un nom. Le verbe est constitué de la racine اكل (manger) suivie d'un suffixe ت qui représente la marque de féminin. Si on considère que le deuxième mot qui apparaît après le verbe est le sujet, alors la phrase est traduite par *La poule a mangé*. La marque de féminin attachée au verbe fait référence donc à *la poule*. Par contre, si on considère que le deuxième mot est le complément d'objet, alors la phrase devient *Elle a mangé le poulet*. Dans le dernier cas, la phrase en arabe ne contient pas de sujet (qui est le pronom personnel *Elle*), il est sous entendu. La marque de féminin attachée au verbe fait référence dans ce cas à *elle*. Il est à noter que pour les deux phrases, la diacritisation est exactement la même, et qu'il est difficile de traduire cette phrase sans savoir le contexte dans lequel elle apparaît.

2.2 Les ressources numériques sur l'arabe

Au début de cette thèse, une période a été consacrée à étudier les ressources numériques disponibles sur l'arabe, qui sont nécessaires à son traitement automatique.

[3]Les tokens représentent l'ensemble des unités utilisées par les modèles de langue et de traduction. Ces unités représentent les mots, les proclitiques (sous unités) ainsi que les signes de ponctuation.

[4]Un type de mot représente le nombre de mots disctincts.

Afin de pouvoir traiter la langue arabe automatiquement, il est important d'avoir des données en arabe, ce qui n'est pas toujours une tâche facile.

Les corpus les plus intéressants pour le traitement des langues, et plus spécifiquement pour la traduction automatique, sont les corpus parallèles, c'est à dire des données en langue source et leur traduction en langue cible. Comme déjà mentionné dans le chapitre 1, ces corpus servent à entraîner et construire des systèmes de traduction automatique. Les corpus parallèles sont très rares et sont difficiles à trouver pour certaines paires de langues : la taille est souvent limitée, et le domaine n'est pas toujours approprié au type de données à traduire.

Plusieurs agences sont spécialistes dans la collecte et la distribution des ressources linguistiques. Les catalogues les plus connus dans la constitution des ressources linguistiques sont le LDC (Linguistic Data Consortium) et l'ELRA (European Language Resources Association).

2.2.1 Le Linguistic Data Consortium (LDC)

Le Linguistic Data Consortium[5] est un consortium ouvert d'universités, librairies, syndicats et de laboratoires de recherche. Le LDC produit des ressources linguistiques les plus importantes pour la langue arabe (Bies, DiPersio et Maamouri, 2012).

Le LDC collecte beaucoup de données en arabe *écrit* à partir de plusieurs sources incluant les flux de journaux, le web (blogs, newsgoups, email) dans plusieurs domaines. Le LDC produit également des traductions (manuelles et automatiques) pour des textes en arabe afin de construire des corpus parallèles arabe-anglais.

Son catalogue inclut aussi une grande quantité de données monolingues, appelée le gigaword. Des données provenant de certaines campagnes d'évaluation - comme celles organisées par NIST - font partie aussi du LDC, ainsi que des données conçues dans le cadre du projet GALE (Global Autonomous Language Exploitation).

Le LDC produit également des corpus audios, issus principalement à partir d'enregistrements d'émissions de radio, et produit des programmes télévisés. Parmi ces corpus, il y a des corpus monolingues en arabe moderne standard et quelques uns en dialecte. Ces corpus sont souvent transcrits.

Parmi les ressources du LDC les plus utilisées pour la langue arabe, citons en particulier BAMA et l'Arabic Treebank.

2.2.1.1 BAMA

BAMA -ou *Buckwalter Arabic morphological analyzer*- (Buckwalter, 2002; Buckwalter, 2004) est un analyseur morphologique de l'arabe qui a été développé par *Tim Buckwalter*. BAMA utilise une approche où les règles morphologiques et orthographiques sont intégrées directement dans le lexique. L'analyseur morphologique est représenté sous forme d'une grande base de données dans laquelle chaque mot en arabe est présenté avec toutes ses formes dérivées possibles *'préfixe-racine-suffixe'*. Chaque forme est donnée avec la version voyellée, l'ensemble des morphèmes (lemme, préfixes et suffixes) constituants de chaque mot, et toutes les étiquettes morphosyntaxiques (ou grammaticales) de chaque composante du mot.

Pour chaque mot, BAMA fournit un ensemble de toutes les segmentations possibles. Cet ensemble comprend un lemme sous la forme d'un identifiant unique, ainsi que pour chaque solution, l'ensemble des morphèmes constituants de chaque mot, leurs étiquettes grammaticales et la traduction correspondante en anglais. Toutes les propositions de segmentations sont proposées avec la translittération Buckwalter[6] (voir section 2.3.1.1). La figure 2.3 montre un exemple d'analyse morphologique du mot فِي (dans) par BAMA.

SAMA -ou *Standard Arabic Morphological Analyzer*- (Maamouri et al., 2010a) est une évolution de BAMA.

[5] http://www.ldc.upenn.edu/
[6] http://www.qamus.org/transliteration.htm

INPUT STRING:	في
LOOK-UP WORD:	fy
Comment:	
INDEX:	P1W4
SOLUTION 1:	(fiy) [fiy_1] fiy/PREP
(GLOSS):	in
SOLUTION 2:	(fiy~a) [fiy_1] fiy/PREP+ya/PRON_1S
(GLOSS):	in + me
SOLUTION 3:	(fiy) [fiy_2] Viy/ABBREV
(GLOSS):	V.
SOLUTION 4:	(fay) [y_1] fa/CONJ+y/ABBREV
(GLOSS):	and/so + Y./10th
SOLUTION 5:	(fy) [DEFAULT] fy/NOUN_PROP
(GLOSS):	NOT_IN_LEXICON
SOLUTION 6:	(fY) [DEFAULT] fY/NOUN_PROP
(GLOSS):	NOT_IN_LEXICON

Figure 2.3: *Analyse morphologique du mot* في *par BAMA*

2.2.1.2 L'Arabic Treebank

L'Arabic Treebank – ou ATB – est un corpus en arabe annoté manuellement. C'est une ressource très précieuse pour le traitement automatique de l'arabe. Nous avons utilisé dans certains travaux au cours de cette thèse le *Penn Arabic Treebank (PATB)* distribué par le LDC (Maamouri et al., 2005a et Maamouri et al., 2005b).

Dans l'ATB, des corpus sont analysés morphologiquement et annotés manuellement au niveau syntaxique. Au niveau symbolique, les corpus sont annotés par des étiquettes morphosyntaxiques et des informations morphologiques. Chaque mot est translittéré (cf. section 2.3.1.1) avec la translittération de Buckwalter. Pour chaque mot, on trouve toutes les propositions de voyellation et de segmentation possibles selon BAMA. L'analyse correcte est annotée dans l'ATB avec une *. La figure 2.4 montre toutes les possibilités d'analyse du mot *why* وهي (et elle) proposées dans l'ATB. Toutes les possibilités d'analyses sont proposées à partir de la base de données de BAMA. La segmentation correcte est annotée par une *.

L'ATB est souvent utilisé pour entraîner et construire des classifieurs, des annotateurs, des parseurs ou pour évaluer des données. Il existe d'autres Treebanks distribués par le LDC, comme le PADT -ou Prague Arabic Dependency Treebank- (Hajic et al., 2004) ou le Penn Treebank qui a inspiré tous ces travaux.

2.2.2 L'European Language Resources Association (ELRA)

ELRA est aussi une organisation de collecte de ressources linguistiques. Les ressources de l'ELRA sont évaluées par l'agence de distribution ELDA (Evaluations and Language resources Distribution Agency)[7].

Le catalogue ELRA contient beaucoup de ressources en arabe, mais il n'est pas aussi riche que le LDC au niveau des corpus. Parmi les ressources que nous avons utilisé dans certains travaux de cette thèse et qui proviennent d'ELRA, le corpus Arcade II (Chiao et al., 2006) constitué de données provenant du journal 'le Monde Diplomatique' annoté en entités nommées en arabe et en français.

[7] http://www.elda.org/

INPUT STRING:	وهي
LOOK-UP WORD:	why
Comment:	
INDEX:	P4W42
SOLUTION 1:	(wahiya) [wahiy-a_1] wahiy/PV+a/PVSUFF_SUBJ:3MS
(GLOSS):	be frail/be fragile + he/it [verb]
* SOLUTION 2:	(wahiya) [hiya_1] wa/CONJ+hiya/PRON_3FS
(GLOSS):	and + it/they/she
SOLUTION 3:	(why) [DEFAULT] why/NOUN_PROP
(GLOSS):	NOT_IN_LEXICON
SOLUTION 4:	(wahy) [DEFAULT] wa/CONJ+hy/NOUN_PROP
(GLOSS):	and + NOT_IN_LEXICON
SOLUTION 5:	(wahaY) [wahaY-i_1] wahaY/PV+(null)/PVSUFF_SUBJ:3MS
(GLOSS):	be frail/be fragile + he/it [verb]
SOLUTION 6:	(whY) [DEFAULT] whY/NOUN_PROP
(GLOSS):	NOT_IN_LEXICON
SOLUTION 7:	(wahY) [DEFAULT] wa/CONJ+hY/NOUN_PROP
(GLOSS):	and + NOT_IN_LEXICON

Figure 2.4: *Un exemple d'analyse du mot why (وهي, et elle) extrait de l'ATB. Les différentes possibilités de segmentation sont proposées par BAMA. La segmentation correcte est marquée par *.*

2.2.3 Autres ressources

Des communautés de traducteurs volontaires existent comme Meedan[8], qui est constitué d'une communauté de personnes volontaires qui traduisent des articles de presse sur Internet ainsi que les commentaires des internautes de l'anglais vers l'arabe; et permettent donc de récupérer des données bilingues à partir d'Internet. Des associations de traduction[9] existent aussi et font appel à des personnes volontaires pour traduire.

Wikipedia aussi est considérée comme une ressource riche. On peut y trouver parfois le même article en plusieurs langues différentes, mais des articles en arabe traduits vers le français ou vers l'anglais dans le domaine journalistique sont rares.

Il est possible de collecter des corpus monolingues gratuitement sur le web. Plusieurs articles de journaux en arabe : Aljazeera, Alhayat, ... peuvent être récupérés à travers le web.

D'autres types de corpus en langue arabe sont intéressants, comme le corpus ANER[10], qui est un corpus monolingue en arabe annoté manuellement en entités nommées. Il a été utilisé dans plusieurs travaux sur la reconnaissance des entités nommées en arabe, et entre autres pour construire un système de détection des entités nommées au cours de nos travaux. Il a été constitué manuellement par Benajiba, Rosso et Benedí (2007) et a été déposé sur Internet.

Des dictionnaires peuvent être téléchargés à partir d'Internet, notamment des dictionnaires de noms propres comme ceux disponibles avec le corpus ANER[11]. *Geonames*[12] est également une base de données géographique qui contient les noms propres de localisations dans plusieurs langues, et qui permet de construire des dictionnaires de noms propres de localisations. D'autres types de dictionnaires existent telle que les dictionnaires en ligne

[8] http://news.meedan.net/
[9] http://www.therosettafoundation.org/index.php/fr/ourmission
[10] http://lingpipe-blog.com/2009/07/28/arabic-named-entity-recognition-with-the-aner-corpus/
[11] Anergazetteers téléchargeables à partir du site http://users.dsic.upv.es/~ybenajiba/
[12] http://www.geonames.org/

comme Ajeeb[13] développé par la compagnie *Sakhr*[14]. Le problème des données disponibles sur le web, c'est qu'elles sont très difficiles à collecter, et très bruitées.

Des collaborations ont été également réalisées pour annoter le Coran (Dukes, Atwell et Habash, 2013). Le corpus du Coran annoté est disponible gratuitement sur le web[15].

2.3 Le traitement automatique de l'Arabe : État de l'art

Les premières recherches sur le traitement automatique de l'arabe ont commencé vers les années 1970 (Cohen, 1970) et concernaient notamment le lexique et la morphologie. Avec Internet, la diffusion de la langue arabe et la disponibilité des moyens de traitement de textes arabes, les travaux de recherche ont abordé des problématiques plus variées comme la traduction automatique. Les outils d'apprentissage sont de plus en plus disponibles et permettent de développer facilement des outils de traduction et d'extraction d'information.

Dans cette thèse, nous nous intéressons au traitement automatique de la langue arabe et plus particulièrement à des applications de traitement de texte, essentiellement la traduction automatique, ainsi qu'à des applications d'extraction d'information et particulièrement la détection des entités nommées.

Comme le montre la section 2.1.2.3, les mots en arabe sont complexes et ceci est dû au phénomène d'agglutination. La séparation des proclitiques de la forme de base – appelé aussi analyse morphologique ou segmentation – permet de réduire le vocabulaire et donc d'améliorer son traitement automatique. En langue arabe, le prétraitement est effectué en trois étapes. La translittération est la première étape de prétraitement du texte arabe (voir section 2.3.1.1). Les deuxième et troisième étapes de prétraitement du texte arabe concernent l'analyse morphosyntaxique et l'analyse morphologique. L'ordre dans lequel sont effectués ces deux traitements n'est pas forcément rigide et chaque outil implémente une manière particulière d'articuler ces traitements. La détection des entités nommées peut également faire partie des prétraitements du texte avant la traduction.

La section 2.3.1 présentera et définira ce que c'est la translittération du texte arabe ainsi qu'un compte rendu des travaux effectués sur les prétraitements de l'arabe. L'enrichissement des systèmes de traduction par des connaissances linguistiques, comme les entités nommées, permet d'améliorer la traduction de l'arabe. Nous présenterons donc un échantillon de travaux sur la détection des entités nommées en arabe dans la section 2.3.2. Quelques travaux sur la traduction automatique de l'arabe seront présentés dans la section 2.3.3.

2.3.1 Le prétraitement de l'arabe

2.3.1.1 La translittération

La translittération est une *opération qui consiste à transcrire, lettre à lettre, chaque graphème d'un système d'écriture correspondant à un graphème d'un autre système, sans qu'on se préoccupe de la prononciation*, d'après le dictionnaire *Larousse*.

Afin de traduire les noms propres, ces derniers sont souvent translittérés d'une langue source vers une langue cible : ils sont réécrits avec les caractères alphabétiques de la langue cible les plus proches des caractères lui correspondant en langue source.

Dans le traitement automatique de l'arabe, un grand nombre de chercheurs utilisent dans leurs travaux, une romanisation des caractères arabes. Cette romanisation consiste à transformer les caractères arabes en caractères latins et donc à chaque caractère en arabe lui faire correspondre un caractère en latin (Knight et Graehl, 1998).

[13] http://dictionary.ajeeb.com/
[14] http://www.sakhr.com/
[15] http://corpus.quran.com/

Une autre raison pour laquelle le texte arabe est translittéré dans la plupart des travaux, est que dans beaucoup de travaux, les chercheurs utilisent l'analyseur morphologique BAMA pour la segmentation du texte.

L'outil de translittération de l'arabe le plus connu et le plus utilisé dans les travaux sur la langue arabe est celui de Buckwalter [6]. L'avantage de Buckwalter est qu'il utilise les caractères ASCII, qui sont facilement reproduits sans avoir besoin d'utiliser des encodages spéciaux ou différents. Les encodages les plus utilisés pour l'arabe sont Unicode, CP-1256 et ISO-8859 (Habash, 2010). En particulier, pour l'arabe, il y a aussi des caractères utilisés pour indiquer qu'une connexion cursive ou une ligature doit être faite entre deux caractères, telle que par exemple lam-alif; et ces caractères sont difficiles à détecter (invisibles, ce sont des caractères de contrôle).

D'autres méthodes et algorithmes de translittération de l'arabe ont été proposés comme le translittérateur développé par Kashani, Popowich et Sarkar (2006) et celui proposé par Al-Onaizan et Knight (2002a) qui translittèrent de l'arabe vers l'anglais. Sherif et Grzegorz (2007) ont également proposé un algorithme de translittération. Ces travaux avaient pour objectif essentiellement l'amélioration de la traduction des noms propres.

2.3.1.2 L'analyse morphologique et morphosyntaxique

Comme il a été déjà mentionné, l'arabe est une langue morphologiquement riche, l'analyse morphologique est une tâche importante qui permet à la fois de réduire le vocabulaire ainsi qu'à faciliter et améliorer les alignements en essayant d'avoir le même nombre de mots en source et en cible.

La segmentation est un processus résultant généralement d'une analyse du texte qui consiste essentiellement en une analyse morphosyntaxique et une analyse morphologique. Souvent, ces deux tâches sont liées.

L'analyse morphosyntaxique consiste à détecter pour chaque mot du texte sa fonction grammaticale dans la phrase et l'étiqueter. Des outils d'analyse morphosyntaxique sont disponibles sur Internet comme Stanford Tagger[16].

Darwish (2002) présente l'un des premiers travaux sur l'analyse morphologique de l'arabe où il présente une approche qui permet de construire rapidement un analyseur morphologique. L'analyseur produit la racine éventuelle pour chaque mot en arabe. Il est basé sur des règles dérivées automatiquement et statistiquement. Des approches de plus en plus performantes sont apparues par la suite.

Lee (2004) utilise les étiquettes morphosyntaxiques du texte en arabe -segmenté- et qui sont alignées avec les étiquettes morphosyntaxiques du texte en anglais afin de décider de garder ou pas les segmentations.

AMIRA développée par Diab (2009) implémente une approche différente, où la séparation des clitiques est effectuée indépendamment de l'étiquetage morphosyntaxique. Marsi, Bosch et Soudi (2005) utilisent l'apprentissage basé sur la mémoire (*memory-based learning*) pour l'analyse morphologique et l'étiquetage de l'arabe. Ils utilisent le k-plus-proche voisin et montrent que l'étiquetage morphosyntaxique peut être utilisé pour choisir l'analyse morphologique la plus appropriée.

Dans les travaux de Kulick (2010), la segmentation et l'étiquetage morphosyntaxique sont effectués simultanément en utilisant un classifieur, et sans utiliser d'analyseur morphologique. Plus récemment Kulick (2011) a fait une extension de l'approche, en distinguant entre les "tokens" classe ouvert (telle que nom, verbe, nom propre, etc.) et les "tokens" classe fermée (telle que préposition, pronom relatif, etc.), qui diffèrent dans leurs affixations morphologiques possibles et leur fréquences. Une liste de noms propres extraites de la base de données SAMA-v3.1 (Maamouri et al., 2010b) était utilisée comme trait pour l'aide à la classification.

[16] http://nlp.stanford.edu/software/tagger.shtml

MADA développé par Habash, Rambow et Roth (2009) est l'outil d'analyse morphologique et de désambiguïsation pour l'arabe le plus utilisé. Cet outil effectue une analyse morphosyntaxique et choisit une proposition de segmentation parmi les propositions de mots segmentés proposés par BAMA (Habash et Rambow, 2005). D'autres outils de segmentation de l'arabe ont été développés initialement pour le prétraitement d'autres langues, ensuite ils ont été adaptés pour la langue arabe comme MorphTagger (Mansour, 2010) qui a été conçu d'abord pour l'étiquetage morphosyntaxique de l'hébreu (Mansour, Sima'an et Winter, 2007) et adapté par la suite pour l'étiquetage et la segmentation de l'arabe. MorphTagger utilise également l'analyseur morphologique BAMA. Les outils MADA et MorphTagger sont présentés plus en détail dans le chapitre 4 de ce manuscrit.

El Isbihani et al. (2006) proposent trois méthodes de segmentation de la langue arabe : une méthode à base d'apprentissage supervisé, une méthode basée sur les fréquences, et une méthode fondée sur les automates à états finis. Ils montrent que la dernière approche donne les meilleurs résultats et est adaptable à différentes tâches.

Parmi les approches à base de règles, on cite G-LexAr (Debili, Achour et Souissi, 2002) qui est un analyseur morphologique de l'arabe à base de règles. Il effectue à la fois voyellation, lemmatisation et segmentation d'un texte en arabe.

Des travaux ont été également effectués pour la segmentation de l'arabe dialectal comme ceux de Habash et Rambow (2006) ou aussi ceux de Mohamed, Mohit et Oflazer (2012).

Une approche à base du Naïve Bayes propose une désambiguïsation des mots traduits de l'arabe vers l'anglais en utilisant des schémas de correspondances dans un corpus parallèle (Ahmed et Nürnberger, 2008). Shah et al. (2010) proposent un modèle d'analyse lexicale utilisé dans plusieurs tâches entre autres l'annotation manuelle du texte en arabe.

L'ordre standard en arabe moderne standard est l'ordre Verbe-Sujet-Objet (Wright, 1988). El Kassas et Kahane (2004) utilisent un arbre de dépendance afin de présenter la structure syntaxique des phrases en arabe.

2.3.1.3 Discussion

Un des problèmes majeurs de la langue arabe est la non voyellation. L'absence des voyelles génère une certaine ambiguïté du sens du mot d'une part, et à la difficulté à identifier sa fonction dans la phrase d'autre part (voir section 2.1.2.1). Cette ambiguïté peut être en partie levée par la voyellation. L'étiquetage grammatical permet également de prédire la voyellation, puisque la terminaison des mots peut être révélatrice de la fonction du mot dans la phrase et permet de distinguer le mode des verbes, la fonction des noms, ... , par exemple la damma (ـُ (u)) pour les noms sujets et les verbes à l'inaccompli et au futur, la fatha (ـَ (a)) pour les noms objets et les verbes au subjonctif, la kasra (ـِ (i)) pour les noms au cas indirect (Baloul et Mareüil, 2002).

Certains outils d'analyse morphologique déterminent la fonction du mot dans la phrase à l'aide des mots voyellés. Il est alors préférable d'avoir des mots diacritisés avant d'effectuer l'étiquetage morphosyntaxique puisque la voyellation aide à l'étiquetage morphosyntaxique (Baloul et Mareüil, 2002). Par contre, cette méthode est lente, puisqu'il faut remettre les

diacritiques pour détecter les informations lexicales et ensuite remettre chaque mot à son format original de nouveau. De plus, cette méthode peut être une source d'erreurs.

Le traitement de l'arabe est le centre d'intérêt de beaucoup de travaux aujourd'hui et c'est un sujet toujours d'actualité. En témoignent les conférences spécifiques à la langue arabe comme CITALA[17] ou des projets comme MEDAR[18] qui a pour objectif d'ouvrir la voie à des ressources résultant d'un effort de collaboration pour créer des ressources en langue arabe dans la région méditerranéenne. Le projet GALE (Global Autonomous Language Exploitation) également vise à traduire de gros volumes de données (écrites et orales) vers l'arabe.

Beaucoup d'outils de traitement automatique de l'arabe sont de plus en plus disponibles comme par exemple les outils de translittération en ligne (ta3reeb[19] développé par google, yamli[20], ou encore maren[21] de Microsoft, etc.).

Il est à noter que l'alphabet arabe a été adapté à certaines langues comme le perse, le pashto, ou le pakistanais. L'alphabet de ces langues contient entre autres des lettres en arabe, mais ce n'est pas la langue arabe.

2.3.2 La détection des entités nommées en langue arabe

La détection des entités nommées (EN) consiste à repérer dans un texte donné, les noms propres telle que les noms de personnes, noms de localisations ou aussi noms d'organisations, etc.

L'analyse automatique des mots arabes est compliquée d'une part par l'existence de nombreuses variantes orthographiques, notamment sur les noms propres, et d'autre part par l'absence de voyellation dans les textes écrits (Habash, 2010) qui engendre de nombreuses ambiguïtés, levées par la voyellation (Debili et Souissi, 1998). Ces facteurs tendent à multiplier les formes inconnues dans les textes, dont il faudra décider (ou non) de l'appartenance à la catégorie EN.

L'arabe se caractérise également par le caractère lacunaire des ressources dictionnairiques et surtout par l'absence de distinction majuscule/minuscule qui est un indicateur très utile pour identifier les noms propres dans les langues utilisant l'alphabet latin. Pour ces raisons, la détection des EN en langue arabe représente de nombreux défis intéressants.

Les premiers travaux sur la reconnaissance des EN pour l'arabe datent de 1998 et reposent sur des méthodes à base de règles (Maloney et Niv, 1998). Plus récemment, d'autres outils de détection d'entités nommées ont été proposés par Shaalan et Raza (2009) qui utilisent une approche à base de règles, des dictionnaires et une grammaire locale, et par Zaghouani et al. (2010) qui adaptent un outil de détection des entités nommées à l'arabe. Samy, Moreno et M. Guirao (2005) utilisent un corpus parallèle pour extraire des EN en arabe. Ils utilisent un étiqueteur à base de règles enrichies avec un lexique monolingue espagnol pour extraire les EN en espagnol qui sont, par la suite, translittérées vers l'arabe. Zitouni et al. (2005) s'appuient sur l'utilisation des techniques d'apprentissage automatique (des *Maximum Entropy Markov Models*) en considérant des jeux de descripteurs idoines, et parviennent à de très bons résultats.

Ces travaux ont été prolongés en particulier par Benajiba et ses co-auteurs, et ont donné lieu notamment à la construction du corpus ANER (Benajiba, Rosso et Benedí, 2007). Dans une première approche, Benajiba et Rosso (2007) explorent un étiquetage fondé sur le maximum d'entropie. Cette approche est étendue ensuite en décomposant la prédiction en deux temps: d'abord les frontières de l'EN en introduisant des catégories morphosyntaxiques (POS), puis la détermination de son type. Une seconde approche, fondée sur l'utilisation des champs

[17] www.citala.org
[18] http://www.medar.info
[19] http://www.google.com/intl/ar/inputtools/cloud/try/
[20] http://www.yamli.com/fr/
[21] http://afkar.microsoft.com/en/maren/

markoviens conditionnels -ou CRF- (Benajiba et Rosso, 2008) a permis d'explorer l'intégration de l'ensemble des traits dans un modèle unique, amenant à de meilleures performances. Benajiba, Diab et Rosso (2008) montrent également l'efficacité d'un prétraitement des textes pour séparer les différents constituants du mot (proclitiques, lemme, et enclitiques). Abdul Hamid et Darwish (2010) proposent d'intégrer des traits intra-mot (notamment n-grammes de caractères) dans une modélisation CRF. Cette approche permet de capturer implicitement les caractéristiques morphosyntaxiques, introduites explicitement dans les expériences de Benajiba et Rosso (2008).

Plusieurs travaux sur la détection des entités nommées avaient pour objectif l'amélioration de la traduction automatique. Dans certaines approches, des translittérations des noms propres sont proposées (Kashani, Popowich et Sarkar (2006); Al-Onaizan et Knight (2002a); Hermjakob, Knight et Daumé III, 2008) et dans d'autres travaux des dictionnaires sont utilisés (Halek, Rosa et Tamchyna, 2011). Babych et Hartley (2003) montrent que la reconnaissance des entités nommées en anglais améliore la qualité de la traduction vers le russe et le français.

2.3.3 La traduction automatique *depuis* et *vers l'Arabe*

Les particularités linguistiques de certaines langues rendent la tâche de traduction automatique plus difficile. La complexité de la morphologie de quelques langues a incité les chercheurs à étudier l'apport de la segmentation du texte sur plusieurs langues comme le tchèque (Goldwater et Mc Closky, 2005) ou l'allemand (Durgar El-Kahlout et Yvon, 2010; Nießen et Ney, 2000). L'impact de la segmentation des mots en des racines et suffixes a été étudié par Popovic et Ney (2004) pour l'espagnol, le catalan et le serbe. Pour certaines langues, comme le chinois, l'unité d'analyse est le caractère (Jing et al., 2003).

D'après Attia (2008); Farghaly et Shaalan (2009) plusieurs niveaux d'ambiguïté posent un défi de taille pour le traitement automatique de l'arabe. La richesse de la morphologie de la langue arabe est donc l'une des principales raisons pour lesquelles plusieurs études se sont focalisées sur l'impact de la segmentation du texte en arabe pour sa traduction automatique.

Un mot en arabe peut se traduire par une phrase en français, ou en anglais. Selon Zbib et Soudi (2012), le nombre de types de mots dans un corpus en arabe sera 20 à 25 % plus important que le nombre de types de mots dans un corpus lui correspondant en anglais dans la même taille. Alotaiby, Alkharashi et Foda (2009) montrent dans une étude sur les corpus *Arabic Gigaword Third Edition* (Graff, 2007) et *English Gigaword Third Edition* (Graff et al., 2007) que le nombre de types de mots dans le corpus arabe est 76 % plus important que le nombre de types de mots dans le corpus anglais.

La segmentation permet donc (i) de réduire le nombre de mots inconnus ainsi que la taille du vocabulaire, et (ii) d'améliorer la qualité de l'alignement des mots en réduisant l'écart du nombre de mots entre la langue source et la langue cible. La segmentation du texte en arabe présente les mêmes avantages que l'arabe soit la langue à traduire ou la langue traduite.

Habash et Sadat (2012) et Al-Haj et Lavie (2012) comparent différentes segmentations possibles et montrent que les modèles de traduction automatique statistique donnent de meilleures performances si le texte en arabe est segmenté, et ceci que ce soit pour le cas où l'arabe est la langue source (Habash et Sadat, 2012) ou la langue cible (Al-Haj et Lavie, 2012). Zollmann, Venugopal et Vogel (2006) ainsi que Larkey, Ballesteros et Connell (2002) montrent également qu'il est préférable de segmenter le texte en arabe avant de le traduire.

Plusieurs projets s'intéressent à la traduction de données (texte, parole) depuis ou vers l'arabe. Meedan[22] vise à traduire tout ce qui est posté sur le site vers l'arabe et l'anglais, que ce soit les titres, les commentaires, ou les articles partagés. L'objectif du projet GALE (Global Autonomous Language Exploitation)[23] est de rendre des langues étrangères (arabe et chinois) texte et parole accessibles en anglais, en particulier dans les milieux militaires

[22]http://news.meedan.net/
[23]http://projects.ldc.upenn.edu/gale/

(Soltau et al., 2009). La reconnaissance automatique de la parole (ASR) est une composante essentielle dans GALE. Des données en dialecte arabe sont transcrites et traduites vers l'anglais, essentiellement des dialectes égyptiens. Le projet SAMAR qui fait l'objet de cette thèse a pour objectif de construire une plateforme de traitement multimédia en langue arabe, et traduire les dépêches AFP de l'arabe vers le français et l'anglais.

Les travaux sur la traduction vers l'arabe sont moins nombreux par rapport aux travaux concernant la traduction automatique depuis l'arabe. Néanmoins la traduction automatique vers l'arabe est un problème difficile. Lors de l'entraînement d'un système de traduction, des prétraitements sont nécessaires afin de réduire l'écart entre la morphologie et la syntaxe de l'arabe et la langue source (souvent l'anglais). Les mots en arabe sont donc segmentés afin de se rapprocher au plus de l'anglais. La recombinaison après la traduction n'est pas triviale : ceci est dû aux problèmes morphologiques et d'ambiguïté qui apparaîssent. Par exemple la recombinaison du proclitique *l* suivi de Al entraîne une altération du *A*, donc le segment *l* + *Almktb* (pour le bureau) devient *llmktb*. Il y a aussi les problèmes d'ambiguïtés lexicales comme par exemple le *y* (YAA) est souvent écrit *Y* (ALIF MAKSURA). El Kholy et Habash (2012); Zbib et Badr (2012) proposent des schémas de recombinaison afin de remédier à ces problèmes.

2.4 Conclusion

Plusieurs travaux se sont intéressés au traitement de l'arabe. Carpuat, Marton et Habash (2012) améliorent la traduction de l'arabe vers l'anglais en utilisant une méthode de réordonnancement des phrases verbales (VS) – en arabe – en phrases nominales (SV) lors de l'étape d'alignement. Cette stratégie améliore les scores BLEU jusqu'à 0,6 points BLEU et réduit le taux d'erreur (TER) jusqu'à 0,7 points.

La langue arabe est une langue ayant un vocabulaire très riche. Parfois, il est difficile de trouver une traduction exacte pour certains mots, et certains mots sont souvent traduits de l'arabe vers une autre langue (comme le français par exemple) par un ensemble de mots ou même une phrase. La traitement automatique et en particulier la traduction automatique de l'arabe présentent donc plusieurs défis. La langue arabe est caractérisée par sa forme morphologiquement complexe qui amplifie énormément le vocabulaire et rend la tâche de traduction automatique plus complexe.

Actuellement, de plus en plus de recherches se focalisent sur le traitement de cette langue morphologiquement riche et complexe. Le traitement automatique de l'arabe moderne standard est différent du traitement automatique de l'arabe dialectal. Récemment, des outils de traduction automatique de l'arabe dialectal vers l'arabe moderne standard ont été implémentés (Salloum et Habash, 2012).

L'arabe parlé cause un sérieux défi pour la traduction automatique et ceci est dû surtout à la rareté des données dialectales. Il existe plusieurs dialectes différents en arabe et plusieurs variantes régionales issues d'un seul dialecte. Bien que la traduction du dialecte soit un problème difficile, plusieurs travaux se sont focalisés sur cette tâche. La traduction de la parole nécessite au préalable une transcription de l'oral. Ceci peut être effectué par des transcripteurs – ce qui est très coûteux – ou en utilisant des outils de reconnaissance de la parole et de transcription automatique (Lamel, Messaoudi et Gauvain, 2007; Messaoudi, Gauvain et Lamel, 2006). Avec la facilité d'accès à Internet, le dialecte peut être collecté à partir du web (forums, chat, ...). Zbib et al. (2012) constituent un corpus parallèle en collectant des données dialectales sur le web (forums et facebook) et en les traduisant en utilisant Amazon Mechanichal Turk[24]. L'utilisation de données dialectales pour traduire le dialecte égyptien montre une amélioration des performances de traduction.

[24] http://www.mturk.com/

Les travaux présentés dans ce manuscrit concernent l'amélioration et l'enrichissement des systèmes de traduction automatique statistique de l'arabe écrit vers le français et vers l'anglais. Dans la suite de nos travaux nous nous intéressons principalement au traitement automatique de l'arabe moderne standard. Plus particulièrement, nous avons travaillé sur le prétraitement de l'arabe qui représente une partie intéressante de nos travaux.

Dans ce chapitre, nous avons donné un bref aperçu sur l'origine de la langue arabe. Nous avons présenté les deux formes principales de la langue arabe. Un échantillon des principales ressources existantes pour le traitement automatique de l'arabe a été fourni. Finalement, nous avons présenté un état de l'art des travaux existants sur le traitement automatique de l'arabe et plus particulièrement les travaux existants sur les prétraitements de l'arabe, la traduction automatique de l'arabe et la détection des entités nommées en arabe. Ces deux traitements font partie des analyses effectuées avant de traduire un texte en arabe et seront étudiés plus en détails dans la deuxième partie du manuscrit.

Part II

Enrichissement et amélioration des systèmes de traduction

CHAPITRE 3

Exploration d'un corpus comparable pour l'adaptation du modèle de traduction

En traduction automatique statistique, les systèmes de traduction sont constitués à partir de corpus parallèles, constitués d'un ensemble de textes en langue source alignés avec leur traduction en langue cible. Pour certaines paires de langues, les corpus parallèles sont seulement disponibles pour un nombre limité de domaines comme les débats politiques ou les textes réglementaires.

Une autre ressource peut être très utile pour les travaux en traduction automatique statistique et particulièrement pour les langues qui sont très limitées en ressources, les corpus comparables. Ces derniers sont constitués de paires de corpus qui contiennent des textes de types similaires, ayant été construits dans la même période et traitant les mêmes thèmes, mais ne sont pas des traductions exactes les uns des autres.

D'après Fung et Cheung (2004), il existe plusieurs types de corpus comparables : des corpus comparables constitués d'un ensemble de textes parallèles, et des corpus comparables constitués à partir de textes différents. Les approches d'exploitation de corpus comparables doivent donc être adaptées aux particularités des données auxquelles elles sont appliquées.

Les corpus comparables représentent une source d'information utile à partir de laquelle il est possible d'extraire des dictionnaires bilingues (Fung et Yee, 1998; Rapp, 1995) ou aussi pour l'apprentissage des termes multilingues (Langé, 1995; Smadja, McKeown et Hatzivassiloglou, 1996). Plus récemment, des travaux ont montré que les corpus comparables peuvent être la source de création de données parallèles telle que des phrases, des mots, des fragments (Bouamor, Semmar et Zweigenbaum, 2013), ou aussi des paragraphes (Li et Gaussier, 2010).

Des travaux s'intéressent à l'extraction de lexiques bilingues à partir de corpus comparables. Morin et Daille (2004) extraient les termes complexes dans chaque langue et les alignent au niveau mot en utilisant une méthode statistique exploitant le contexte des termes. Une approche d'extraction de termes médicaux et leurs traductions à partir d'un corpus comparable est proposée par Morin et al. (2010). Morin et Prochasson (2011) présentent une approche pour l'extraction de lexiques bilingues spécialisé à partir d'un corpus comparable et montrent que l'utilisation d'un petit lexique bilingue spécialisé – induit à partir des phrases parallèles incluses dans le corpus comparable – améliore la qualité d'extraction de lexique bilingue. Hazem, Morin et Saldarriaga (2011) proposent une approche d'extraction de traductions de mots à partir de corpus comparables inspirée des métamoteurs de recherche

d'information. Récemment, Hazem et Morin (2012) présentent une approche – basée sur les systèmes Question-Réponse – pour l'extraction de lexiques bilingues à partir de corpus comparable. Le mot est considéré comme une partie d'une question, sa traduction est recherchée dans la réponse à cette question dans la langue cible.

Comme déjà mentionné, l'extraction de corpus parallèles à partir d'un corpus comparable constitue un sujet de recherche très intéressant. Dans l'approche de Munteanu et Marcu (2005) par exemple, des méthodes d'extraction d'information sont combinées afin d'identifier des documents parallèles : un dictionnaire bilingue existant est tout d'abord utilisé pour traduire chaque document en langue source vers la langue cible. Pour chaque paire de documents, des paires de phrases et de segments parallèles sont extraites en utilisant un lexique de traduction et un classifieur à maximum d'entropie. Un échantillon de travaux récents sur les corpus comparables est rapporté dans Rapp et al. (2012).

D'autres approches d'amélioration de systèmes de traduction statistique avec des corpus comparables ou monolingues ont été proposées. Schwenk (2008); Ueffing, Haffari et Sarkar (2008) proposent une approche d'auto-apprentissage sur des données *du-domaine* afin d'adapter et améliorer le système de base entraîné sur des données *hors-domaine*.

Dans ce chapitre, nous présentons une nouvelle approche d'extraction de corpus parallèle à partir d'un corpus comparable bruité constitué de données journalistiques en arabe et en français. Nous présentons un ensemble d'expériences dont le but est d'utiliser le corpus comparable de deux manières différentes. La première approche consiste à extraire des phrases parallèles *du-domaine* à partir du corpus comparable, et la deuxième approche consiste à adapter le modèle de traduction et comparer différentes méthodes d'adaptation en utilisant des données du domaine. Ces deux approches visent à adapter les modèles de traduction au type de données à traduire.

L'objectif de ces expériences est de comparer différentes méthodes pour construire un système de traduction automatique spécifique aux données journalistiques en utilisant (i) un système de traduction de base entraîné sur des données *hors domaine* et (ii) un corpus comparable. Comme nous l'avons déjà mentionné, vu le grand nombre de documents parallèles, la meilleure solution est de construire un nouveau corpus parallèle *du domaine*.

Dans la section suivante, nous présentons un aperçu des travaux réalisés sur l'extraction de corpus parallèles et l'adaptation à partir de corpus comparables. La section 3.2 présente les motivations pour ces travaux suivie par une description détaillée de nos approches d'extraction de corpus parallèle et d'adaptation de modèles de traduction dans les sections 3.3 et 3.4. Dans la section 3.5, on présente les expériences effectuées, ainsi que les résultats que nous avons obtenus et l'évaluation de nos approches d'adaptation. Finalement on conclut le chapitre dans la section 3.6.

3.1 État de l'art

Dans les systèmes de traduction automatique statistique, l'utilisation des corpus comparables est répartie en deux catégories : d'une part, des approches dont le but est d'extraire des fragments parallèles et d'autre part, des approches dont le but est d'adapter des ressources existantes à un nouveau domaine en utilisant des corpus comparables.

3.1.1 Extraction de segments parallèles

Comme indiqué précédemment, les corpus parallèles sont des ressources très rares et très limitées dans certains domaines (médecine, agriculture, etc.) et c'est pour cette raison que plusieurs travaux se sont intéressés à explorer les corpus comparables afin d'en extraire des corpus parallèles.

Dans la plupart des travaux d'extraction automatique de segments parallèles, le processus est constitué de deux étapes. Tillmann et Xu (2009) identifient tout d'abord un ensemble de

paires de textes candidats, ensuite les phrases parallèles sont identifiées en se basant sur des scores de similarité.

Les travaux de Zhao et Vogel (2002) se sont focalisés sur l'extraction de phrases parallèles à partir d'un ensemble de textes comparables chinois/anglais. Les similarités entre les phrases sont extraites à partir d'un modèle d'alignement probabiliste. L'identification des phrases parallèles s'effectue en se basant sur le rapport de leur longueur, ainsi que le score du modèle IBM 1 pour les alignements mot-à-mot.

Resnik et Smith (2003) proposent une approche pour extraire des corpus parallèles à partir de documents collectés à partir d'Internet. Les correspondances entre les documents et les phrases sont principalement détectées en se basant sur la similarité des adresses web ou des pages web ayant la même structure interne.

Une approche d'extraction de corpus parallèles en utilisant des techniques d'extraction d'information – *Cross-Lingual Information Retrieval* ou aussi *CLIR* – a été proposée également par Munteanu et Marcu (2005). Les scores de similarité sont calculés en utilisant un modèle de régression logistique dont l'objectif est de détecter les phrases parallèles. Chaque phrase dans le document source est mise en correspondance avec plusieurs phrases dans la langue cible. Toutes les paires de phrases candidates sont classifiées par la suite pour chaque paire de documents. Smith, Quirk et Toutanova (2010) montrent dans des travaux plus récents, des améliorations des méthodes de détection de phrases parallèles à partir de données extraites de Wikipedia. L'extraction est effectuée, en utilisant des modèles basés sur les champs markoviens conditionnels (CRF), avec un ensemble de paramètres plus riche et des dépendances entre les phrases. Kumano et Tokunaga (2007); Munteanu et Marcu (2006) traitent également la question de fouille de fragments parallèles à partir d'un corpus comparable.

L'approche proposée par Abdul-Rauf et Schwenk (2009) consiste à utiliser un système de traduction existant pour construire la partie cible du corpus, et ensuite, comparer les phrases existantes avec la partie cible générée automatiquement, en utilisant des mesures de distance (WER, TER). Cette approche a été généralisée sur plus de données par Uszkoreit et al. (2010) qui montrent les avantages d'utilisation d'une seule langue pour adopter des techniques efficaces de détection de parallélisme (Broder, 2000).

3.1.2 Exploration des corpus comparables pour l'adaptation

Les corpus comparables représentent une ressource très utile pour *adapter* et *spécialiser* des ressources existantes (dictionnaires, modèles de traduction, modèles de langue) à des domaines spécifiques. Dans ce chapitre, on s'intéresse seulement à l'adaptation de modèles de traduction. Une étude détaillée des différentes techniques d'adaptation de modèles de langue est proposée par Bellagarda (2001). Zhao, Eck et Vogel (2004) par exemple explorent différentes techniques d'adaptation de modèle de langue, en utilisant des modèles de langue spécifiques. L'adaptation de modèles de langues sera étudiée plus en détails dans le chapitre 6 de ce manuscrit où on explore les différentes thématiques des dépêches AFP.

Une amélioration du modèle de traduction a été proposée par Snover, Dorr et Schwartz (2008) en ajoutant de nouvelles règles de traduction. Ces règles associent chaque phrase dans un document source avec les phrases cibles les plus fréquentes dans un corpus comparable spécialement conçu pour ce document. La correspondance entre les documents s'effectue à l'aide des scores de probabilités.

Cettolo, Federico et Bertoldi (2010) utilisent un modèle de traduction de référence pour obtenir plutôt des alignements de segments entre la phrase source et la phrase cible dans un corpus comparable. Ces alignements sont raffinés avant que les nouvelles phrases soient collectées. La figure 3.1 illustre le processus pour construire un modèle de traduction adapté en utilisant un corpus parallèle extrait automatiquement à partir d'un corpus comparable.

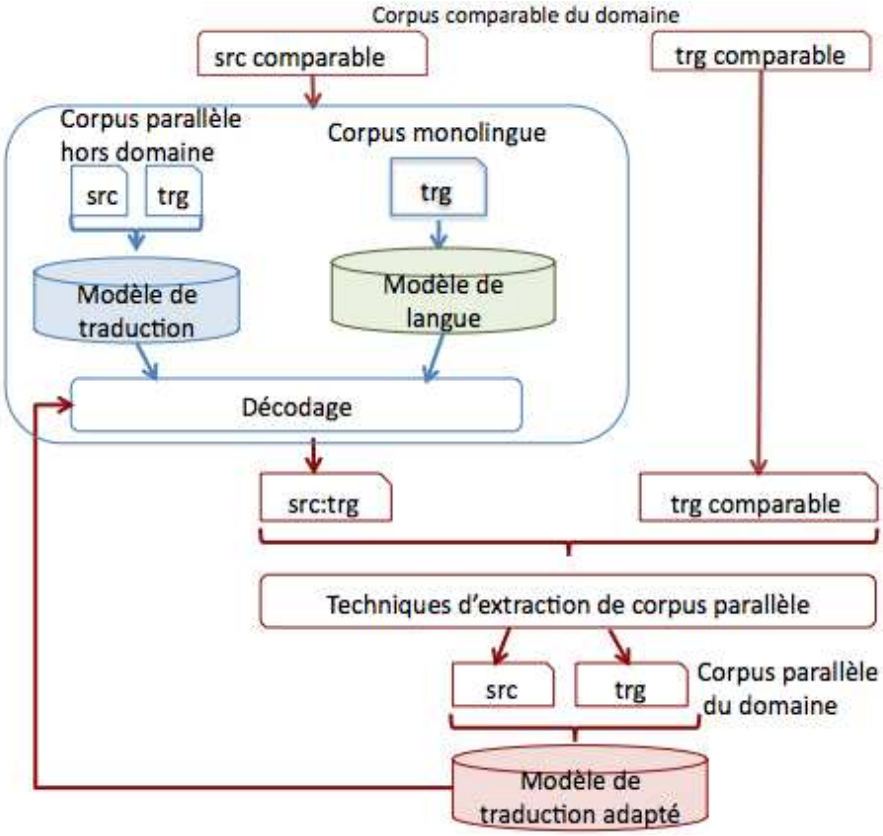

Figure 3.1: *Techniques d'extraction de corpus parallèle à partir d'un corpus comparable.*

3.2 Motivations

Comme mentionné auparavant, les corpus monolingues et les corpus comparables sont beaucoup plus faciles à récupérer que les corpus parallèles. Dans le cadre du projet SAMAR, nous devons construire un système de traduction arabe-français et un système de traduction arabe-anglais. Ces systèmes de traduction vont servir de plateforme de traduction de dépêches de l'AFP provenant principalement du Moyen-Orient.

Pour la paire de langues arabe-français, nous avons (i) des données parallèles arabe-français constituées principalement de données provenant de l'ONU[1], et (ii) un corpus comparable constitué à partir de dépêches AFP. De plus, nous savons que dans le corpus comparable arabe-français, un grand nombre des dépêches en français ont été traduites par des journalistes vers l'arabe. Dans certains cas la traduction n'est pas littérale et dans d'autres cas elle implique également une réécriture. Le corpus comparable dont on dispose, est constitué donc d'une sous partie de corpus parallèle, mais on ne sait pas quelles sont les dépêches qui sont les traductions les unes des autres.

Pour la partie arabe-anglais, nous avons (i) un corpus parallèle arabe-anglais non distribuable, et (ii) un corpus comparable arabe-anglais constitué à partir de dépêches AFP. Par contre, pour cette paire de langues, le corpus comparable ne contient pas un sous ensemble de corpus parallèle.

La technique d'extraction de corpus parallèle peut être appliquée dans notre cas pour la paire de langues arabe-français. Cela est d'autant plus facile puisque toutes les dépêches sont

[1]Organisation des Nations Unies

horodatées, ce qui permet de repérer facilement les textes parallèles candidats. Pour traduire des dépêches AFP, il est préférable d'utiliser un système de traduction adapté à ce type de données. Dans les deux cas, nous appliquerons une stratégie de *bootstrapping* (amorçage) en utilisant comme référence un système constitué de données *hors-domaine*. Nous allons donc décrire nos approches d'extraction de corpus parallèle à partir d'un corpus comparable et d'adaptation du modèle de traduction.

Dans ce chapitre, toutes les expériences effectuées ont été réalisées sur la paire de langues arabe-français puisque nous disposons d'un corpus de développement et de test constitués à partir de dépêches AFP provenant du Moyen-Orient seulement pour cette paire de langues.

3.3 Approche d'extraction de corpus parallèle

Dans cette section, nous présentons notre approche d'extraction de corpus parallèle. Cette approche consiste à présenter notre méthode d'extraction d'un corpus parallèle à partir d'un corpus comparable *du domaine*. En ce qui concerne cette approche, il est nécessaire d'avoir principalement un système de traduction de base *hors domaine* et un corpus comparable *du domaine*. Ce sont les deux principales ressources nécessaires pour l'extraction de corpus.

Comme le montre la figure 3.2, notre approche d'extraction d'un corpus parallèle *du domaine* à partir d'un corpus comparable *du domaine* est constituée essentiellement de trois étapes principales.

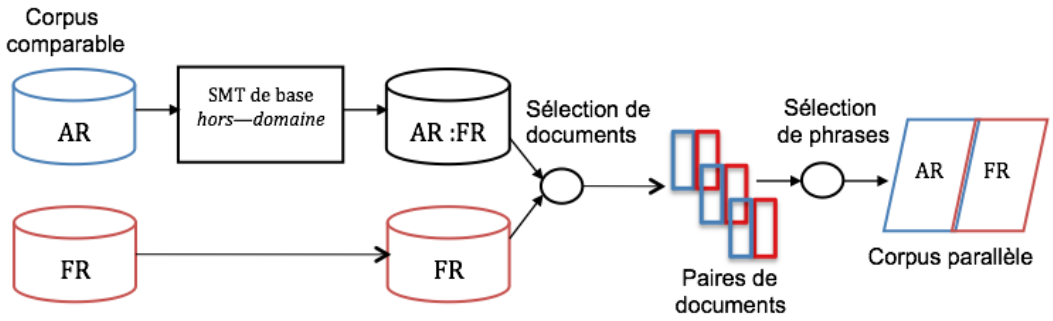

Figure 3.2: *Processus d'extraction de corpus parallèle à partir d'un corpus comparable.*

Les trois étapes sont réparties comme suit :

1. **Traduction**: traduire le côté source du corpus comparable;

2. **Sélection de paires de documents** : choisir, dans le corpus comparable, les documents qui sont similaires aux sorties de traduction;

3. **Sélection de paires de phrases** : choisir les paires de phrases parallèles à partir des paires de documents choisis.

L'idée principale consiste dans le fait que le calcul des scores de similarité entre des documents ayant la même langue permet d'utiliser des méthodes de comparaison efficaces. Cette méthode est plus simple que de devoir définir des scores de similarité complexes en utilisant des modèles d'alignement.

L'étape de *traduction* (1) consiste à traduire les documents de la partie source du corpus comparable, en arabe, en utilisant le système de traduction de base *hors domaine*. Ce dernier a été entraîné sur des données parallèles *hors domaine*. Cette étape est différente de l'approche de Munteanu et Marcu (2005) où un dictionnaire bilingue est utilisé – à la place d'un système de traduction – pour traduire les documents.

Avant de traduire les dépêches AFP en arabe, des prétraitements ont été réalisés sur ces données. Nous avons remarqué que dans les dépêches en arabe – qui sont à l'origine des traductions manuelles de dépêches en français – certaines phrases ont été regroupées sur la même ligne. Une méthode pour séparer les phrases a été donc développée afin d'avoir une phrase complète par ligne. Le classifieur Wapiti[2] (Lavergne, Cappé et Yvon, 2010) – développé au LIMSI et fondé sur les champs markoviens conditionnels (ou CRF Lafferty, McCallum et Pereira, 2001) – a été utilisé pour la classification des phrases. La séparation des phrases permet d'augmenter les chances d'avoir plus de phrases à aligner entre la langue source et la langue cible. La structure des phrases traduites se rapproche donc de la structure des phrases extraites des dépêches en français.

Une fois que les documents sont traduits, l'étape suivante est la *sélection de paires de documents* (2) qui consiste à apparier les documents traduits automatiquement (AR:FR) avec les documents originaux dans la langue source (FR). Pour chaque document traduit, un score de similarité est calculé entre ce document et tous les documents dans la langue cible datés du même jour. Les scores de similarité sont calculés en se basant sur le coefficient Dice (Dice, 1945; Lin, 1998). Ce coefficient calcule le nombre de mots en commun dans les deux documents, normalisé par la longueur du document traduit (AR:FR).

Le nombre des documents à comparer est limité grâce à des connaissances préalables telles que la date de publication des dépêches. Pour chaque document en langue source, le document en langue cible ayant obtenu le meilleur score est choisi comme étant un candidat potentiel de document similaire. Une fois les documents candidats appariés, seules les paires de documents ayant obtenu un score de similarité supérieur à un seuil T_d sont conservées. Les paires de documents sont filtrées afin d'éviter les fausses correspondances entre documents. Dans les expériences qui seront présentées par la suite, le seuil a été fixé de manière à favoriser la précision sur le rappel.

Un ensemble de paires de documents similaires en source et en cible sont finalement sélectionnées. Ces paires de documents peuvent être des traductions exactes les uns des autres. Cependant, dans la plupart des cas, les documents sont des traductions bruitées et seulement un sous ensemble des phrases de ces documents sont des traductions exactes.

La dernière étape de *sélection de phrases* (3) consiste à effectuer un alignement au niveau des phrases de chaque paire de documents. Le but de cette étape est de garder seulement les phrases parallèles. L'alignement des phrases est effectué avec l'outil d'alignement de phrases hunalign[3] (Varga et al., 2005). Hunalign fournit également des scores de confiance pour les alignements entre phrase(s) source(s) et phrase(s) cible(s).

Comme pour l'étape de sélection de documents, seules les paires de phrases qui ont un score d'alignement supérieur à un seuil de confiance T_s sont choisies. Le seuil de confiance T_s est choisi également de manière à favoriser la précision sur le rappel. Finalement, un alignement 1:1 est retenu, permettant de construire un petit corpus parallèle adapté au type de données à traduire. Notre approche est assez différente de l'approche de Munteanu et Marcu (2005) où l'étape de sélection de phrases est effectuée en utilisant un classifieur d'entropie maximale (*Maximum Entropy*) et l'évaluation du degré de parallélisme des phrases extraites est effectuée avec des mesures comme le taux d'erreur mot (WER) ou le taux d'édition de la traduction (TER).

[2] http://wapiti.limsi.fr
[3] Téléchargeable depuis http://mokk.bme.hu/resources/hunalign/

3.4 Approche d'adaptation des modèles de traduction

Dans cette section, on présente notre approche d'adaptation de système de traduction aux données à traduire en utilisant un corpus comparable. Cette approche consiste à adapter les modèles de traduction avec différents corpus parallèles soit (i) extraits automatiquement ou (ii) construits automatiquement.

Afin de pouvoir exploiter notre corpus comparable, nous avons produit des traductions vers le français de toutes les dépêches en arabe. Ceci signifie que nous avons donc trois corpus parallèles :

- Le corpus d'entraînement de base, ayant une très grande taille, et constitué d'une centaine de millions de mots, qui produit des traductions de qualité acceptable, mais qui est *hors domaine*;
- Le corpus *du domaine* extrait, qui est beaucoup plus petit et potentiellement très bruité;
- Le corpus *du domaine* traduit, qui est de taille moyenne, et qui a une qualité beaucoup plus mauvaise que les autres, puisqu'il contient des erreurs de traductions.

Compte tenu de ces trois corpus, différentes méthodes d'adaptation des modèles de traductions ont été étudiées. La première approche consiste à concaténer le corpus d'entraînement de base (*hors domaine*) et le corpus *du domaine* (soit le corpus **extrait** ou le corpus **traduit**) pour entraîner un nouveau modèle de traduction. Étant donné la différence de taille entre les deux corpus, cette approche peut donner l'impression que le modèle de traduction *hors domaine* est favorisé par rapport au modèle *du domaine*.

La deuxième approche consiste à entraîner séparément des modèles de traduction : d'une part, le modèle de base, et d'autre part le modèle avec les données *du domaine*. La combinaison des deux modèles est ensuite optimisée avec MERT (Och, 2003). Ceci réduit la taille du modèle de traduction mais par contre augmente le nombre de poids qui doivent être optimisés, et donc MERT peut devenir moins stable.

Une dernière approche est également considérée, qui consiste à utiliser seulement les données *du domaine* pour entraîner le modèle de traduction. Dans ce cas, l'approche consiste à étudier l'impact de la petite taille des données *hors domaine*.

Une étude comparative des trois approches, utilisant les trois corpus, est étudiée dans la section suivante.

3.5 Expériences et résultats

Cette section présente les données que nous avons utilisées, le système de traduction de base, les expériences effectuées pour développer la méthode d'extraction de corpus parallèles ainsi que les expériences réalisées pour l'adaptation du modèle de traduction avec les différentes approches présentées dans la section 3.4. Une comparaison et une évaluation de ces approches ont été effectuées.

3.5.1 Contexte et données

Les expériences se sont déroulées dans le contexte du projet SAMAR, décrit dans l'introduction de ce manuscrit. Chaque jour, l'AFP produit environ 250 dépêches en langue arabe, 800 dépêches en français et en anglais. Ces dépêches proviennent directement de l'AFP et sont stockées sur les disques du LIMSI.

L'ensemble des dépêches utilisées dans ce chapitre ont été récupérées entre décembre 2009 et décembre 2010. Ce corpus comparable est constitué de 75,975 dépêches en arabe et 288,934 dépêches en français (environ 1 million de phrases pour la partie arabe et 5 millions de phrases pour le français).

La particularité de ce corpus comparable est que beaucoup de dépêches en arabe sont des traductions de dépêches en français. Ces traductions peuvent être partiellement fidèles. Lorsqu'un journaliste traduit une dépêche, il est libre de réorganiser la structure du contenu d'un document ou d'étendre ce dernier. La figure 3.3 montre une paire de phrases extraites d'une paire de dépêches dont une a été traduite manuellement par un journaliste. Les deux dépêches sont donc des traductions l'une de l'autre. La figure contient un exemple d'une phrase extraite de la dépêche en arabe et sa traduction extraite de la dépêche qui lui correspond en français. On présente également dans cette figure la traduction exacte de la phrase en arabe vers le français.

Dépêche AR :	واضاف نحن لا مانع لدينا من استئناف المفاوضات غير المباشرة حول الصفقة من النقطة التي انتهت اليها والتي حاول ان يفشلها نتانياهو.
Traduction exacte:	*Et il a ajouté, nous à Hamas, on n'a pas de problème de reprendre les négociations indirectes autour de l'affaire au point où elles s'étaient arrêtées et que Netanyahu a essayé de les faire échouer.*
Dépêche FR :	Le porte-parole a réeaffirmé que le Hamas était prêt à reprendre les tractations au point où elles s'étaient arrêtées.

Figure 3.3: *Un exemple d'une traduction approximative extraite d'une paire de documents arabe-français similaires.*

On remarque que la traduction de la phrase dans cet exemple n'est pas exacte. La phrase en arabe contient plus d'informations qui ont été ajoutées par le journaliste.

Dans nos expériences, le corpus comparable *du domaine* est constitué d'un ensemble de documents en arabe et en français qui peuvent être parallèles, partiellement parallèles, ou pas du tout parallèles. Il n'y a pas de lien explicite entre les parties arabe et français du corpus comparable qui indiquerait le degré de parallélisme entre deux dépêches. La date de parution de la dépêche peut être un indicateur utile pour l'extraction de phrases parallèles.

3.5.2 Système de traduction de base

Le système de traduction de base *hors domaine* a été entraîné sur un corpus de 7,6 millions de phrases parallèles (voir tableau 3.1).

	Taille
ONU	7,4 M
WHO	221,7 K
Project Syndicate	18,6 K
Total	7,6 M

Tableau 3.1: *La répartition des données du corpus d'entraînement hors-domaine.*

Ce corpus a été récupéré à partir de sources accessibles au public sur le web : la base de données des documents des Nations Unies (ONU), le site web de l'Organisation Mondiale de la Santé (*World Health Organization*, WHO) et le site web du Project Syndicate. Les données de l'ONU représentent de loin la plus grande partie de ce corpus, sachant que seuls les documents du Project Syndicate peuvent être considérés comme étant *du domaine*.

Le tableau 3.2 présente le nombre de tokens et la taille des vocabulaires pour chacun des trois corpus dont on dispose : le corpus d'entraînement de base (base), le corpus extrait automatiquement (extrait) et le corpus traduit automatiquement (traduit).

	ar		fr	
Corpus	**#tokens**	**voc**	**#tokens**	**voc**
base	162M	369K	186M	307K
extrait	3,6M	72K	4,0M	74K
traduit	20,8M	217 K	22,1M	181K

Tableau 3.2: *Statistiques sur les corpus : le nombre total des tokens dans les parties arabes et français, ainsi que la taille des vocabulaires arabe et français. Les nombres sont donnés pour des données prétraitées.*

Les *tokens* représentent les unités séparées par des espaces dans le corpus. Ils peuvent être des mots ou aussi des marques de ponctuation. Les mots en arabe sont souvent composés, et les sous mots représentent donc des tokens (lorsque les données sont prétraitées).

En ce qui concerne le modèle de langue utilisé pour l'ensemble des expériences, c'est un modèle de langue 4-gramme construit à partir d'un corpus de 2,4 milliards de mots issus principalement du corpus français Gigaword constitué essentiellement de données journalistiques.

Avant de construire un système de traduction et d'aligner le texte parallèle, il est nécessaire de prétraiter le texte parallèle afin d'améliorer la traduction (Fournier, 2008) surtout si les langues traitées ont une morphologie complexe. Comme déjà mentionné dans le chapitre 2, le prétraitement de l'arabe est une tâche qui s'avère nécessaire pour faire face à la rareté des données et réduire le vocabulaire morphologiquement riche et complexe.

La chaîne de prétraitement que nous avons mise en place pour la partie arabe pour l'ensemble de ces expériences était constituée de plusieurs étapes. En premier lieu, toutes les données en arabe sont translittérées avec Buckwalter (Buckwalter, 2002). Une analyse morphologique du texte en arabe est ensuite effectuée avec l'outil d'analyse morphologique et de désambiguïsation MADA (Habash, Rambow et Roth, 2009). MADA présente plusieurs types de segmentations. Nous avons utilisé le type de segmentation MADA-D2, puisque d'après Habash et Sadat (2006), c'est le type de segmentation le plus efficace lorsqu'on a un grand corpus. MADA-D2 consiste à séparer les proclitiques و (w), ف (f), ل (l), ك (k), ب (b) et س (s) de la forme de base.

Pour la partie française du corpus, la tokenisation consiste à séparer les mots des ponctuations et à remplacer les majuscules par des minuscules, sauf pour les noms propres.

Les données parallèles prétraitées en arabe et en français ont été alignées avec MGiza++[4] (Gao et Vogel, 2008). Le décodeur Moses (Koehn et al., 2007) a été ensuite utilisé pour symmétriser les alignements en utilisant l'heuristique *grow-diag-final-and* et extraire les phrases avec une longueur maximale de 7 mots. Un modèle de distorsion lexicale a été entraîné pour les phrases en arabe et en français. Les poids des paramètres ont été optimisés par MERT (Och, 2003) sur un corpus de développement AFP (ce corpus sera présenté à la fin de la section 3.5.3).

[4] http://geek.kyloo.net/software/doku.php/mgiza:overview

3.5.3 Extraction du corpus parallèle *du domaine*

Afin d'extraire le corpus parallèle à partir de notre corpus comparable *du domaine*, nous avons utilisé la méthode décrite dans la section 3.3. En premier lieu, les documents en arabe sont traduits vers le français en utilisant le système de traduction de base. Ensuite, pour choisir les paires de documents candidats, chaque document traduit (ar:fr) est comparé à tous les documents en français qui sont parus le même jour que le document en arabe. Afin de fixer les meilleurs seuils pour la sélection des documents, nous avons varié les valeurs des seuils T_d et T_s, et nous avons calculé à chaque fois le nombre de documents et de phrases extraits comme le montre le tableau 3.3.

Système	T_d	T_s	#documents	#phrases extraites	BLEU
S1	0,5	0,7	24 354	158 230	29,2
S2	0,5	0,5	24 354	209 910	30,7
S3	0,3	0,5	43 279	257 524	32,4

Tableau 3.3: *Nombre de phrases extraites en fonction des valeurs de seuil* T_d *et* T_s

En augmentant la valeur des seuils, le nombre des paires de documents et des paires de phrases choisies baisse. En réduisant les seuils pour l'extraction des phrases parallèles extraites automatiquement, le bruit est également réduit. Le score BLEU, évalué sur chacun des trois systèmes entraînés sur les données extraites automatiquement, augmente à chaque fois qu'il y a des données supplémentaires. D'après Goutte, Carpuat et Foster (2012), les systèmes de traduction construits à partir de corpus parallèles bruités n'affectent les scores de traduction que lorsque le taux des phrases bruitées dépasse 30 % des données.

Les seuils pour la sélection des documents et des phrases parallèles ont été donc fixés respectivement à 0,5 et à 0,7. Le score BLEU donné par le système S1 est inférieur aux scores BLEU des systèmes S2 et S3 puisqu'il contient moins de données; mais le contenu du corpus est moins bruité et contient moins de phrases non-alignées, ce qui justifie notre choix pour les seuils T_s et T_d. La figure 3.4 présente le pourcentage des phrases extraites pour $T_s = 0,5$ et $T_s = 0,7$.

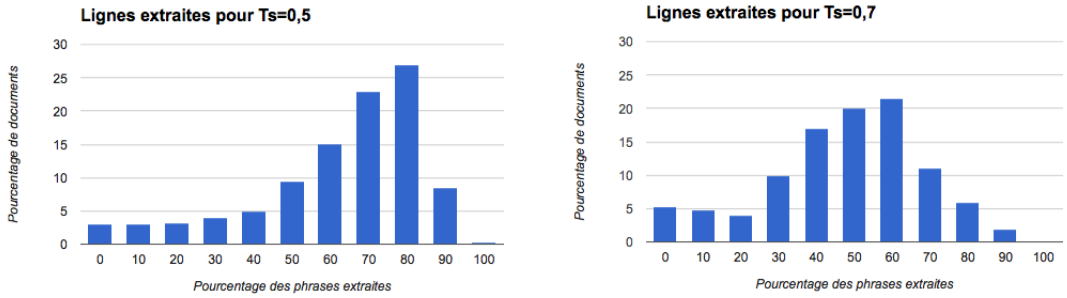

Figure 3.4: *Pourcentage des documents en fonction des phrases extraites pour* $T_s = 0,5$ *et* $T_s = 0,7$

La figure montre, pour les documents à partir desquels des phrases parallèles ont été extraites, le pourcentage des phrases extraites. Initialement, le nombre de phrases traduites représente 1 million de phrases, à partir desquelles nous n'avons gardé que 158 milles phrases. D'après les histogrammes, on voit pour $T_s = 0,5$ et pour $T_s = 0,7$ le pourcentage de documents en fonction du pourcentage de phrases parallèles extraites.

Pour une paire de documents similaires, le pourcentage moyen des phrases choisies est de 43 % environ pour $T_s = 0,7$. L'étape de sélection de documents permet de choisir un ensemble de documents contenant autour de 35 % du nombre total de phrases initialement présentes dans la partie arabe du corpus comparable. Ce pourcentage baisse jusqu'à 15 % après l'étape d'alignement. Le corpus parallèle *du domaine* résultant est ainsi constitué d'un ensemble de 156 000 paires de phrases parallèles.

Les données collectées durant la période du dernier mois (décembre 2010) a été isolée du corpus résultant. Ces données ont été utilisées pour choisir aléatoirement un corpus de développement et un corpus de test constitués d'à peu près 1 000 lignes chacun.

Les corpus de développement et de test ont été vérifiés manuellement afin d'évaluer la précision de l'approche. 98,2 % des phrases ont été correctement appariées. Le tableau 3.2 compare les caractéristiques principales des trois corpus utilisés pour l'entraînement.

3.5.4 Résultats des traductions

Les résultats de traduction obtenus sur le corpus de test sont évalués avec la métrique BLEU (Papineni et al., 2002). Le tableau 3.4 montre les résultats obtenus avec, pour chaque expérience, la taille de la table de traduction correspondante.

Les différentes approches d'adaptation décrites dans la section 3.4 ont été expérimentées avec les corpus **extrait** et **traduit** comme étant des corpus d'adaptation (voir section 3.3).

Système	**#Paires de segments**	**BLEU**
Base	312,4M	24,0
Extrait	10M	29,2
Extrait + Base (1 table)	321,6M	29,0
Extrait + Base (2 tables)	312,4M + 10M	**30,1**
Traduit	39M	26,7
Extrait + Traduit (2 tables)	10M + 39M	28,2

Tableau 3.4: *Comparaison et évaluation des différents modèles adaptés avec le système de base, de l'arabe vers le français sur un corpus de test de 1 000 phrases de l'AFP*

On note que l'adaptation du modèle de traduction aux données journalistiques est très efficace. En comparant par rapport au système de base tous les systèmes adaptés obtiennent de meilleurs résultats (de 2 à 6 points BLEU d'amélioration). Le système **extrait** améliore la traduction de 5 points BLEU par rapport au système de base. Cette amélioration est obtenue malgré la grande différence de taille des deux corpus : le corpus d'entraînement du système *extrait* est beaucoup plus petit (3,6 millions de tokens pour le système *extrait* contre 162 millions de tokens pour le système de base). Ce résultat confirme indirectement la précision de notre méthodologie.

La construction d'un modèle de traduction en concaténant le corpus du système de base avec le corpus *extrait* ne donne pas de meilleurs résultats que le système **extrait** seul. Ceci peut être dû au fait que la grande quantité du corpus *hors domaine* domine dans l'ensemble du corpus, et que les données *du domaine* sont trop faiblement représentées. Toutefois, une interpolation log-linéaire des deux modèles de traduction, apporte une petite amélioration de 0,8 points BLEU. L'interpolation log-linéaire consiste à donner au système de traduction deux modèles de traduction. Dans notre cas, nous avons privilégié le premier modèle de traduction extrait par rapport au modèle de traduction de base en utilisant le mode *either* dans Moses et l'option *decoding-graph-backoff*.

Un autre résultat intéressant à discuter est le résultat obtenu avec le système entraîné avec seulement le corpus **traduit**. En effet, les données parallèles artificielles – dont la partie source a été construite automatiquement – permettent d'améliorer le score de traduction de 2,7

points BLEU par rapport au système de base. Il faut noter que les traductions automatiques de ce système n'ont pas été filtrées. Ce dernier résultat montre l'importance d'avoir des données correspondant au domaine des données à traduire, même si le système de traduction *du domaine* est de moins bonne qualité qu'un système de traduction *hors domaine* (Cettolo, Federico et Bertoldi, 2010).

Dans la dernière expérience, les données *du domaine* ont été utilisées ensemble, c'est à dire les données **extrait** et **traduit**. Deux tables de traduction ont été entraînées pour chaque corpus séparément. Toutefois, cette combinaison ne permet pas d'améliorer les résultats du système **extrait**. Le filtrage des traductions automatiques peut être une solution : le score donné par le système de traduction pour chaque phrase à la sortie de traduction peut être utilisé pour ne garder que les meilleures phrases.

Une observation rapide des traductions fournies par le système de base et le système **extrait** montre que les sorties de traductions sont très différentes. La figure 3.5 montre deux exemples typiques.

Référence :	Le ministre russe des Affaires étrangères, Sergueï Lavrov *a prévenu* mercredi [...]
Base :	*Pronostiquait* Ministre des affaires étrangères russe, Sergei Lavrov mercredi [...]
Extrait :	Le ministre russe des Affaires étrangères, Sergueï Lavrov *a averti* mercredi [...]
Référence :	Le porte-parole de Mme Clinton, *Philip Crowley*, a toutefois reconnu [...]
Base :	Pour *ukun FILIP Cruau* porte-parole de Clinton a reconnu ...
Extrait :	Mais *Philip Crowley*, le porte-parole de Mme Clinton a reconnu [...]

Figure 3.5: *Comparaison de traductions automatiques de deux phrases en utilisant le système de traduction de* **base** *et le système de traduction* **extrait**.

Le premier exemple illustre les différents styles d'écriture dans les deux types de corpus. Dans les données de l'ONU (où le texte est souvent sous forme de dialogue et de narration, *"Pronostiquait"* suivi par le sujet *"ministre russe des Affaires étrangères"*) alors que dans le style journalistique, on a plutôt un style de textes où on présente une suite d'évènements *"Le ministre russe des affaires étrangères a prévenu"* ou *"a averti"* – qui sont sémantiquement équivalents. Le deuxième exemple montre l'impact de l'adaptation pour la correction de la traduction des mots, et plus particulièrement les entités nommées comme (*"Philip Crowley"*) qui a été traduit incorrectement par le système de traduction de base par (*"ukun FILIP Cruau"*). Ce problème sera traité plus en détail dans le chapitre 5 dans lequel on présente des travaux sur le traitement automatique des entités nommées.

3.6 Conclusion

Nous avons présenté dans ce chapitre une étude empirique des deux méthodes différentes (i) extraction de corpus parallèle à partir d'un corpus comparable et (ii) utilisation de données *du domaine* pour adapter un système de traduction automatique de base.

L'approche *traduit* est limitée, et ne permet pas au système d'apprendre les traductions de mots inconnus qui font partie de la nouvelle source de données. Seuls les mots qui se trouvent dans les modèles de traduction vont apparaitre dans le nouveau corpus créé automatiquement. Cette approche est limitée pour créer par exemple de nouvelles expressions, qui n'existent pas dans le modèle de traduction, telle que "A fish out of water" qui doit être traduite par "Quelqu'un dans une situation qui ne lui convient pas". Les expériences de l'auto-apprentissage montrent que – bien que cette approche soit limitée – le modèle de langue adapté (construit à partir de données journalistiques) influe sur la distribution des segments dans le modèle de traduction, et donc sur le choix de la meilleure hypothèse de traduction.

En utilisant notre nouveau système adapté en auto-apprentissage, nous avons étendu le corpus parallèle : les données en arabe – prétraitées avec notre nouvel outil de prétraitement SAPA qui sera présenté dans le chapitre 4 – sont traduites vers le français, et ensuite comparées à la partie du corpus comparable en français pour extraire les paires de phrases parallèles. Un nouveau système de traduction automatique a été donc construit. Le corpus parallèle extrait automatiquement a été amélioré et étendu à la suite de ces travaux. Dans le chapitre 7, on présente les résultats du nouveau système de traduction amélioré construit à partir du corpus parallèle étendu.

Les recherches sur ce thème sont toujours d'actualité. Aker, Feng et Gaizauskas (2012) par exemple proposent une méthode d'extraction de corpus parallèle à partir d'un corpus comparable en utilisant les Séparateurs à Vaste Marge (SVM) et montrent que l'ajout du nouveau corpus extrait automatiquement améliore les performances de traduction.

Afin de poursuivre ces travaux, nous avons l'intention d'étudier l'évolution des résultats de traduction comme par exemple la qualité de la précision/rappel du corpus extrait, et la qualité des données traduites automatiquement. Nous avons dans ce chapitre travaillé sur l'adaptation du modèle de traduction particulièrement. Des travaux sur l'adaptation du modèle de langue seront présentés dans le chapitre 6 dans lequel on exploite les différentes thématiques des dépêches de l'AFP.

CHAPITRE 4

Amélioration des pré-traitements de l'arabe

La langue arabe est une langue sémitique, qui a la particularité d'avoir un vocabulaire à base de racines de mots trilitères consonantiques comme il a été détaillé dans le chapitre 2. À cette forme de base, peuvent s'ajouter des préfixes, suffixes, ainsi que des clitiques. Les clitiques et affixes sont agglutinés au mot de base pour former des mots de plus en plus complexes qui peuvent même former des phrases. Prenons par exemple le mot استكتبينها (*Est ce que tu vas l'écrire ?*) – sachant que le *tu* dans la phrase fait référence à une fille –, est constitué de deux proclitiques ا et س, un préfixe ت, un suffixe ن, un proclitique ها et la racine de base كتب.

Cette forme morphologique complexe augmente la taille du vocabulaire de la langue arabe comme le montre la figure 4.1. On observe que lorsque le nombre de phrases est 1 000,

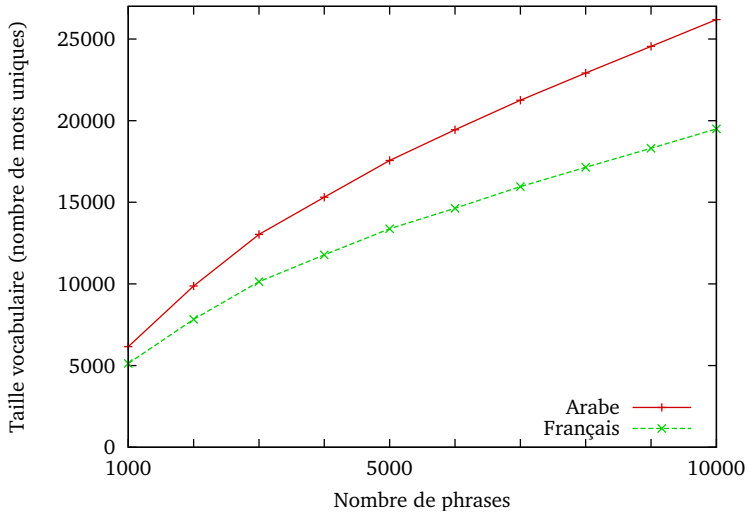

Figure 4.1: *Croissance du vocabulaire dans un corpus parallèle arabe-français*

la différence entre le nombre de mots uniques en français et en arabe est faible. Au fur et à mesure que le nombre de phrases croît, la différence entre le vocabulaire français et arabe

devient de plus en plus importante. Dans le cas du traitement automatique des langues, et particulièrement si on utilise des méthodes d'apprentissage, le nombre de formes différentes doit être réduit en appliquant des techniques de normalisation. Ceci permet d'avoir moins de mots inconnus.

Plusieurs niveaux d'ambiguïté posent un défi de taille pour le traitement automatique de l'arabe d'après Farghaly et Shaalan (2009) et Attia (2008). L'arabe est une langue morphologiquement riche et complexe et c'est l'une des principales raisons pour laquelle plusieurs études se sont focalisés sur l'impact de la segmentation du texte en arabe pour la traduction automatique. Une vue d'ensemble de certains travaux sur l'analyse morphologique est présentée par Al-Sughaiyer et Al-Kharashi (2004).La langue arabe est une langue qui nécessite des traitements avant de la traduire et ceci à cause du caractère agglutinatif de sa morphologique. Nous avons présenté dans le chapitre 2 la langue arabe et sa morphologie ainsi qu'un compte rendu des travaux sur la traduction automatique de l'arabe et les prétraitements de l'arabe.

Dans ce chapitre, on présente une nouvelle approche pour effectuer des prétraitements de l'arabe qui a la particularité d'être indépendante de toute autre ressource tel qu'une base de données morphologique ou les outils de désambiguïsation. Ensuite, on compare notre approche aux deux outils de segmentation de l'arabe les plus utilisés, MADA et MorphTagger. Dans la section suivante, on présente un aperçu des travaux effectués sur le prétraitement de l'arabe. Dans le chapitre 2, un état de l'art sur les prétraitements de l'arabe a été présenté, la section 4.1 décrit donc plus particulièrement les deux approches auxquelles nous avons comparé notre approche, en détails. La section 4.2 décrit les motivations à l'origine de ces travaux. Dans la section 4.3, notre approche de segmentation et d'étiquetage morphosyntaxique à base de CRF est présentée, suivie dans la section 4.4 par l'évaluation de cette approche et sa comparaison avec MADA et MorphTagger. Ce chapitre est finalement clôturé par une conclusion (section 4.5).

4.1 État de l'art

Comme indiqué dans le chapitre 2, l'arabe est une langue morphologiquement complexe, et il est nécessaire de prétraiter le texte en arabe avant de le traduire. Plusieurs travaux se sont intéressés à l'analyse morphologique de l'arabe comme l'outil AMIRA (Diab, 2009) basé sur l'apprentissage supervisé et qui utilise les Séparateurs à Vaste Marge (SVM) pour la classification, ou encore l'approche de Marsi, Bosch et Soudi (2005) qui utilisent le k-plus-proche voisin et montrent que l'étiquetage morphosyntaxique peut être utilisé pour choisir l'analyse morphologique la plus appropriée. L'analyse morphologique ainsi que l'étiquetage morphosyntaxique sont effectués en utilisant l'apprentissage basé sur la mémoire. El Isbihani et al. (2006) proposent une méthode de segmentation fondée sur les automates à états finis. Kulick (2010) propose une approche dans laquelle il utilise des expressions régulières qui englobent toutes les possibilités de tokenisations et d'étiquettes morphosyntaxiques. Mohamed et Kübler (2010) montrent, que pour la tâche d'étiquetage morphosyntaxique, la segmentation du texte en arabe améliore la précision lorsque le nombre de mots inconnus – par le système d'apprentissage – est important. Des travaux sur la segmentation de l'arabe dialectal ont été réalisés par Habash et Rambow (2006) ou aussi Mohamed, Mohit et Oflazer (2012) par exemple.

MADA (Habash, Rambow et Roth, 2009) est l'outil d'analyse morphologique et de désambiguisation de l'arabe le plus utilisé et le plus connu. Il effectue également une analyse morphosyntaxique qui aide à l'analyse morphologique. MADA implémente un processus à deux niveaux pour sélectionner la meilleure décomposition/analyse possible parmi les analyses proposées par l'analyseur morphologique BAMA (Buckwalter, 2004). La section 4.1.1 présente le fonctionnement détaillé de MADA.

Mansour (2010) présente MorphTagger, un outil de segmentation basé sur les modèles de Markov cachés (Hidden-Markov-Model) pour l'Arabe et le compare à MADA (Habash et

Rambow, 2005; Habash, Rambow et Roth, 2009) et à l'approche de El Isbihani et al. (2006). MorphTagger a été conçu d'abord pour l'étiquetage morphosyntaxique de l'hébreu (Barhaim et Winter, 2005) ensuite il a été adapté pour l'étiquetage et la segmentation de l'Arabe (Mansour, Sima'an et Winter, 2007). MorphTagger est présenté en détails dans la section 4.1.2

MADA et MorphTagger utilisent l'analyseur morphologique BAMA (Buckwalter, 2004) qui donne pour chaque mot, l'ensemble des possibilités de segmentation de ce mot. *Tim Buckwalter* a développé cette base de données qui contient un lexique de racines de mots en arabe, de préfixes et de suffixes ainsi que les différentes possibilités de combinaison de ces composantes.

Dans ce qui suit, nous allons présenter une description détaillée de MADA dans la section 4.1.1, et de MorphTagger dans la section 4.1.2.

4.1.1 MADA

C'est l'un de nos principaux points de comparaison. Dans cette section, on présente le fonctionnement détaillé de MADA ainsi que les ressources qui lui sont associées. MADA (Morphological Analysis and Disambiguation for Arabic) est un outil d'analyse morphologique et de désambiguïsation qui accomplit la suite d'étapes suivante.

En premier lieu, MADA translittère le texte en arabe en utilisant l'encodage de Buckwalter. Ensuite, un outil interne d'analyse morphologique et de génération Almorgeana[1] est utilisé pour produire une liste d'analyses morphologiques potentielles de chaque mot. Ces analyses sont générées sans tenir compte du contexte. Toutes les segmentations possibles du mot sous la forme préfixe-racine-suffixe sont engendrées et seule la compatibilité bilatérale est vérifiée en respectant les règles définies par la base de données BAMA.

La tâche de MADA consiste à déterminer quelle est l'analyse la plus probable de ce mot étant donné son contexte. Une fois cette décision prise, tous les paramètres de cette analyse sont supposés être les caractéristiques appropriées pour ce mot.

Afin de choisir l'analyse appropriée, MADA calcule des scores pour les analyses proposées et utilise pour celà 19 paramètres : 14 prédits par des modèles SVM, 2 paramètres prédits avec l'outil SRILM, 1 paramètre prédit à partir du modèle unigramme, et 2 heuristiques supplémentaires. MADA utilise donc 17 modèles prédits et quelques règles déterministes afin de lever l'ambiguïté sur les analyses.

Tout d'abord, le désambigüiseur de SRILM[2] (Stolcke, 2002) est utilisé pour prédire deux paramètres : le lemme (lex) et la diacritisation (diac). Ensuite, MADA utilise 14 modèles SVM, chacun prédisant l'une des caractéristiques morphologiques en utilisant le classifieur SVMTool (Giménez et Màrquez, 2004): asp, cas, enco, gen, mod, num, per, pos, prc0, prc1, prc2, prc3, stt et vox.

Ceux 16 sont les seuls modèles créés pour prédire des paramètres. Un modèle supplémentaire unigramme est également utilisé pour modéliser la probabilité de l'analyse globale (plutôt que de ses composants).

La décision de l'analyse proposée par MADA est fondée uniquement sur les 17 modèles et les règles déterministes. Une fois la sélection de l'analyse est effectuée, les paramètres bw, gloss, rat, source, stem, et stemcat sont des résultats de cette analyse.

Bw fournit l'étiquette complète générée par BAMA/SAMA. *Gloss* est la traduction anglaise extraite de la base de données. *Source* indique la référence à partir de laquelle l'analyse est extraite (base de données, variation orthographique). *Stem* est le lemme du mot, et *stemcat* est la catégorie du lemme (utilisé en interne par ALMOR). *Rat* est synonyme de rationalité. C'est une information qui pourrait enrichir la base de données ALMOR. Pour l'instant elle n'est pas encore utilisée par MADA.

La figure 4.2 présente un exemple d'analyse du mot رئيس (*président*). La figure montre

[1] Almorgeana ou ALMOR utilise la base de données BAMA (Buckwalter, 2004).
[2] http://www.speech.sri.com/projects/srilm.

```
;;WORD r}ys
diac:ra}iys lex:ra}iys_1 bw:+ra}iys/NOUN+ gloss:president;head;chairman pos:noun prc3:0 prc2:0 prc1:0 prc0:0
     per:na asp:na vox:na mod:na gen:m num:s stt:i cas:u enc0:0 rat:y source:lex stem:ra}iys stemcat:N/ap
diac:ra}iysK lex:ra}iys_1 bw:+ra}iys/NOUN+K/CASE_INDEF_GEN gloss:president;head;chairman pos:noun prc3:0 prc2:0
     prc1:0 prc0:0 per:na asp:na vox:na mod:na gen:m num:s stt:i cas:g enc0:0 rat:y source:lex stem:ra}iys stemcat:N/ap
diac:ra}iysN lex:ra}iys_1 bw:+ra}iys/NOUN+N/CASE_INDEF_NOM gloss:president;head;chairman pos:noun prc3:0 prc2:0
     prc1:0 prc0:0 per:na asp:na vox:na mod:na gen:m num:s stt:i cas:n enc0:0 rat:y source:lex stem:ra}iys stemcat:N/ap
diac:ra}iysa lex:ra}iys_1 bw:+ra}iys/NOUN+a/CASE_DEF_ACC gloss:president;head;chairman pos:noun prc3:0 prc2:0
     prc1:0 prc0:0 per:na asp:na vox:na mod:na gen:m num:s stt:c cas:a enc0:0 rat:y source:lex stem:ra}iys stemcat:N/ap
diac:ra}iysi lex:ra}iys_1 bw:+ra}iys/NOUN+i/CASE_DEF_GEN gloss:president;head;chairman pos:noun prc3:0 prc2:0
     prc1:0 prc0:0 per:na asp:na vox:na mod:na gen:m num:s stt:c cas:g enc0:0 rat:y source:lex stem:ra}iys stemcat:N/ap
diac:ra}iysu lex:ra}iys_1 bw:+ra}iys/NOUN+u/CASE_DEF_NOM gloss:president;head;chairman pos:noun prc3:0 prc2:0
     prc1:0 prc0:0 per:na asp:na vox:na mod:na gen:m num:s stt:c cas:n enc0:0 rat:y source:lex stem:ra}iys stemcat:N/ap
```

Figure 4.2: *Exemple d'analyse du mot* رئيس *(r{ys)*

que pour chaque analyse, MADA propose 22 valeurs : 6 extraits de la base de données comme étant un résultat de l'analyse (*bw*, *gloss*, *rat*, *source*, *stem* et *stemcat*) et 16 prédits par les modèles de prédictions (*diac*, *lex*, *pos*, les proclitiques *prc3*, *prc2*, *prc1*, *prco*, si le mot contient ou pas un pronom personnel (*per*), aspect (*asp*), voix (*vox*), mode (*mod*), genre (*gen*), singulier ou pluriel (*num*), état défini ou indéfini (*stt*), *cas*, la valeur de l'enclitique s'il existe (*enco*)). En arabe, il existe trois aspects : le passé (الماضي), l'inaccompli (المضارع) et l'impératif (الامر). Il existe également trois valeurs possibles pour le paramètre *cas* : nominatif (مرفوع), accusatif (منصوب), et génitif (مجرور). Les paramètres *per*, *asp*, *mod* et *vox* sont appliqués seulement aux verbes, les paramètres *cas* et *stt* s'appliquent seulement aux noms et adjectifs, et les paramètres *gen* et *num* s'appliquent à la fois aux verbes et aux noms et adjectifs. Ces huit paramètres sont des paramètres inflexionnels, qui apparaîssent agglutinés au lemme.

D'après la figure 4.2, on note que les propositions d'analyses sont classifiées selon un score calculé par MADA, et que la meilleure analyse est marquée par une $*$.

L'ensemble des prédictions sont comparés avec la liste des analyses potentielles complètes. Ces dernières sont ensuite ordonnés et l'analyse morphologique qui a le score le plus élevé est finalement choisie, en fournissant les valeurs prédites de tous les paramètres morphologiques. La figure 4.3 illustre le processus de calcul des analyses avec MADA pour extraire la meilleure segmentation.

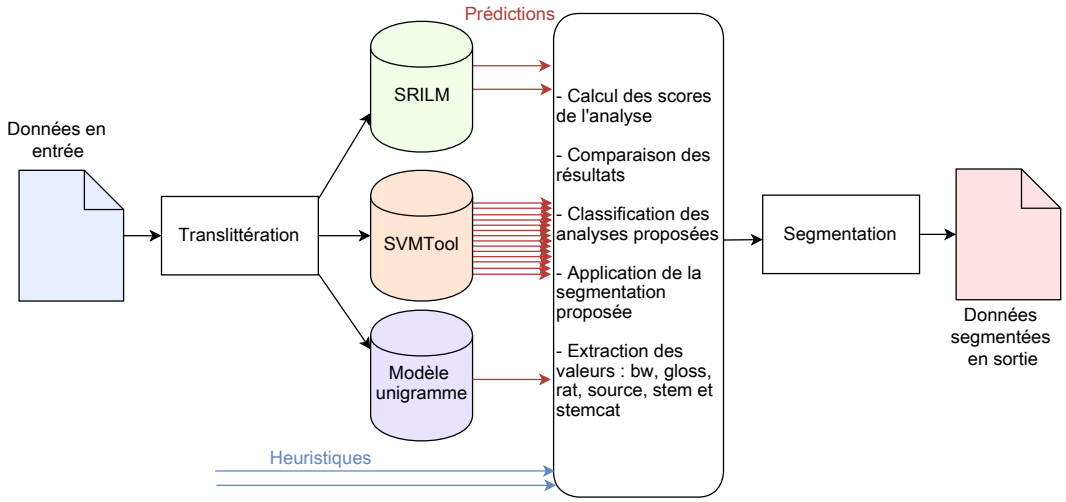

Figure 4.3: *Processus de calcul des analyses proposées par MADA.*

En se basant sur l'analyse morphologique choisie, les mots peuvent être segmentés selon des schémas de segmentation prédéfinis et déterministes. Habash et Sadat (2012) présentent les différents schémas de segmentation proposés par MADA.

Dans le contexte spécifique des applications de traduction automatique, l'utilisation de MADA ainsi que sa description morphologique très précise, est probablement une analyse très excessive, surtout si tout ce dont on a besoin est la séparation des différents morphèmes qui constituent un mot. Le coût de calcul associé est en effet assez élevé.

MADA nécessite d'exécuter, pour chaque token, plusieurs classifieurs SVM avant de combiner les résultats et décider de la segmentation de chaque mot. Le prétraitement d'un corpus de grande taille est possible seulement en répartissant le traitement sur plusieurs machines et ceci nécessite généralement un temps de traitement très important pour la construction totale du système.

4.1.2 MorphTagger

MorphTagger est un segmenteur de l'arabe à base de modèles de Markov cachés (HMM). Cet outil a été également conçu pour effectuer des analyses morphologiques dans le cadre des applications de traduction automatique. Il est plus rapide que MADA et il constitue notre deuxième point de comparaison dans ce chapitre.

MorphTagger a été conçu initialement pour la tâche d'analyse morphosyntaxique de l'hébreu, et ensuite adapté à la langue arabe (Mansour, Sima'an et Winter, 2007). L'étape de segmentation ainsi que quelques règles de normalisation ont été ajoutées à l'outil. L'architecture de MorphTagger est similaire à celle de MADA étant donné qu'il utilise la base de données BAMA ainsi que l'outil SRILM pour la désambiguïsation.

Tout d'abord, le texte en arabe passe à travers l'analyseur morphologique BAMA, qui produit en sortie pour chaque mot, toutes les analyses possibles ainsi que leurs étiquettes morphosyntaxiques. MorphTagger produit en sortie la séquence d'étiquettes la plus probable en fonction du modèle. Par la suite, le choix de l'analyse correcte est effectué en choisissant le morphème le plus probable tout en tenant compte de l'étiquette morphosyntaxique. L'outil SRILM est utilisé pour la désambiguïsation.

Une fois l'analyse morphologique effectuée, les prépositions (à part le déterminant) ainsi que les pronoms possessifs et les pronoms objet sont séparés en respectant un ensemble de règles. Le segmenteur applique également quelques étapes de normalisation, dont les plus importantes sont (i) Alif maksura reprend sa forme originale lorsqu'un mot, dont le suffixe a été séparé, se termine par Alif maksura (yX $\rightarrow$ Y+X); (ii) la marque de féminin reprend sa forme originale lorsqu'un nom est séparé de son suffixe (tX $\rightarrow$ p+X) et (iii) l'article défini *Al* reprend sa forme originale, après la segmentation, lorsqu'il est précédé par le préfixe *l* (llX $\rightarrow$ l+Al+X). La figure 4.4 est extraite de Mansour (2010) et montre l'architecture du segmenteur MorphTagger.

Mansour, Sima'an et Winter (2007) comparent MorphTagger à MADA et au segmenteur à base de transducteurs à états finis introduit par El Isbihani et al. (2006) et montrent que MorphTagger donne de meilleurs résultats de traduction sur différentes conditions de traductions et différents ensembles de tests.

4.2 Motivations

Lors de nos travaux précédents, nous avons remarqué que le processus de prétraitement est lent et nécessite beaucoup de ressources. De plus, l'installation de ces ressources est un peu lente et nécessite une installation préalable d'autres outils d'analyse morphologique et de désambiguïsation.

D'autre part, dans le cadre du projet SAMAR, nous devons fournir la chaîne de prétraitement de l'arabe complète pour la tâche de traduction. L'outil de prétraitement de l'arabe

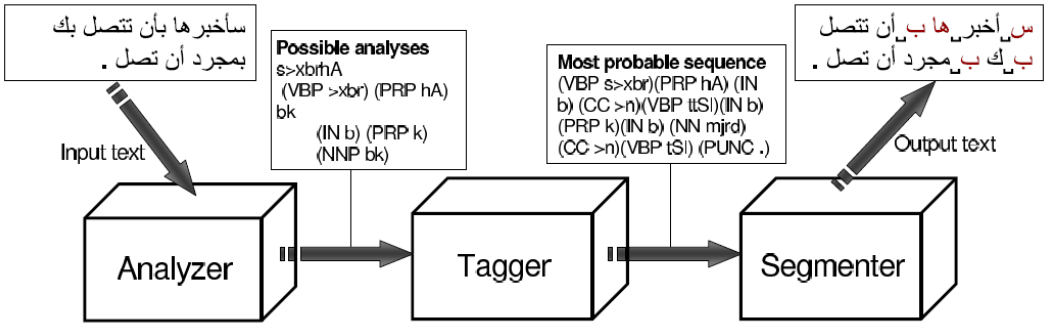

Figure 4.4: *L'architecture du segmenteur MorphTagger (Mansour, 2010)*

développé par le partenaire du projet concerné est trop lent, et nécessite l'utilisation exclusive du système d'exploitation Windows.

Ces deux principales raisons nous ont motivé à développer un outil de prétraitement de l'arabe indépendant de toute ressource externe, et beaucoup plus rapide pour le traitement du texte.

L'approche de prétraitement de l'arabe que nous avons proposé a été inspirée de MADA. Nous avons choisi de séparer les mêmes proclitiques que MADA-D2. Notre approche a été comparée principalement à MADA et MorphTagger.

4.3 Approche

Afin de développer notre outil de prétraitement de l'arabe, nous avons utilisé l'outil Wapiti (Lavergne, Cappé et Yvon, 2010) développé au LIMSI. Wapiti est fondé sur les champs markoviens conditionnels (ou CRF Lafferty, McCallum et Pereira, 2001), et permet de construire des modèles intégrant un très grand nombre de descripteurs.

Notre objectif est de segmenter le texte en arabe afin de séparer les proclitiques et reproduire la segmentation de MADA-D2 comme présenté par Habash et Sadat (2006). L'objectif de notre approche est de reproduire la segmentation dans un délai beaucoup plus court et en utilisant le moins de ressources possibles. L'analyse morphologique n'est qu'une étape de prétraitement qui doit être appliquée sur des textes parallèles de grande taille. C'est une parmi plusieurs étapes pour construire un système de traduction.

Dans ce cadre, notre objectif est de prédire seulement les paramètres qui sont informatifs pour la segmentation de mot. L'étiquette morphosyntaxique du mot de base est également une information intéressante à connaître. Ceci nécessite donc d'encoder notre étiquette composée à prédire par les CRF comme suit : POS+pr1+pr2+pr3 , où POS représente l'étiquette morphosyntaxique du mot de base, et les étiquettes pr1, pr2 et pr3 indiquent respectivement l'absence/présence et les types possibles de proclitiques (voir Tableau 4.1). Le premier proclitique, pr1, concerne les conjonctions de coordination و (*w*) et ف (*f*); pr2 concerne les prépositions ب (*b*), ل (*l*) ,ك (*k*) et س (*s*) et l'étiquette pr3 indique si un mot

contient le proclitique ال (*Al*) ou non. La valeur de pr1 est *CONJ* pour les conjonctions de coordination ou *none*; pr2 est représenté par *PREP* si le proclitique correspond à une préposition, à *SUB* si le proclitique correspond à une subordination, à *FUT* pour indiquer la marque du futur, ou *none*; et pr3 peut être *DET* ou *none*.

Proclitique	**Étiquette/Valeur**
pr1	CONJ/*w*+, *f*+ ou *none*
pr2	PREP/*b*+, *l*+, *k*+ ou SUB/l+ ou FUT/*s*+ ou *none*
pr3	DET/*Al*+ ou *none*

Tableau 4.1: *Proclitiques, étiquettes et valeurs*

En utilisant cette structure, le mot وللإنتخابات (*et pour les votes*) par exemple va être étiqueté comme suit NOUN+CONJ+PREP+DET.

En ce qui concerne les catégories syntaxiques, nous avons utilisé une liste des 24 étiquettes morphosyntaxiques principales de l'Arabic Treebank (Maamouri et al., 2005a; Maamouri et al., 2005b). Cependant, il faut noter que certains préfixes ne sont combinés qu'avec des mots ayant une fonction grammaticale bien déterminée. Par exemple, le préfixe س (*s*+) ne peut être associé qu'aux verbes afin d'indiquer la marque du futur. Cela signifie que le nombre d'étiquettes possibles est bien inférieur au nombre total de toutes les combinaisons possibles d'étiquettes (384).

La figure 4.5 montre les étapes de pré-traitement du mot وللإنتخابات. Selon notre approche, les textes en arabe sont tout d'abord translittérés en utilisant l'encodage de Buckwater[3]. Ensuite, la prédiction de segmentation est effectuée, suivie d'une étape de normalisation et finalement par l'étape de segmentation qui est basée sur un ensemble de règles.

Figure 4.5: *Les étapes de pré-traitement.*

La section 4.3.1 présente les paramètres utilisés pour la création des modèles de prédiction des étiquettes morphosyntaxiques et de la segmentation. La normalisation appliquée lors de nos pré-traitements est présentée dans la section 4.3.2. La section 4.3.3 est dédiée à présenter les règles de segmentation qui sont appliquées lors de la segmentation.

4.3.1 Modèles et sélection de paramètres

Le modèle de prédiction des étiquettes morphosyntaxiques et de la segmentation a été entraîné sur un ensemble de descripteurs des séquences de mots à l'entrée. Dans Wapiti, ces traits sont décrits par des modèles génériques qui testent simultanément les unigrammes et bigrammes d'étiquettes ainsi que d'autres traits directement observables sur les séquences de mots.

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

Dans nos expériences, ces tests à l'entrée sont définis comme suit : (1) tests unigrammes, qui évaluent la présence/absence de mots dans une fenêtre de taille 7 mots autour du mot considéré, (2) tests bigrammes qui évaluent la présence/absence de bigrammes de mots dans une fenêtre glissante de taille 5 mots, et (3) des tests qui évaluent la présence/absence de trigrammes de mots sur des fenêtres glissantes de taille 3 mots. Nous avons également utilisé comme descripteurs (4) des tests sur les préfixes et suffixes, qui évaluent jusqu'au 5 premiers (respectivement derniers) caractères avec une fenêtre glissante de 3 mots. Finalement (5) des tests sur les ponctuations et les nombres pour la présence ou absence de marques de ponctuation et des nombres dans des fenêtres glissantes de 5 mots.

4.3.2 Normalisation

L'étape suivante concerne la normalisation de quatre caractères arabes. Cette étape a été inspirée de la normalisation effectuée dans MADA (Habash, 2010). Le problème de normalisation est apparu à cause des incohérences dans l'utilisation des diacritiques et de certaines lettres (Farghaly et Shaalan, 2009). Certaines lettres en langue arabe ont la même forme et sont différenciées seulement en ajoutant certaines marques comme la hamza ou les points.

Les différentes façons d'écrire le Alif إ, أ, آ, ٱ et ا sont normalisées par ا (dans les textes translittérés, respectivement <, > , | , { et *A* sont remplacés par *A*). Le Yaa Maqsura ى (*Y*) et Yaa ي (*y*) sont normalisés par ي (*y*) , Le Taa Marbuta ة (*p*) devient ه (*h*) et les différentes formes de Hamza ؤ (&), ئ (}), ء (') sont normalisées par ء (').

4.3.3 Règles de segmentation

La dernière étape est la segmentation. Elle est effectuée en appliquant un ensemble de règles sur ces prédictions. Cet ensemble de règles est réduit à quatre règles principales. La première règle consiste à segmenter en deux parties un mot contenant seulement un proclitique pr1 suivi d'un mot plein. La deuxième règle vérifie si un mot qui contient deux proclitiques pr1 et pr2 n'est pas une préposition, alors il est segmenté en trois parties. La troisième règle vérifie si un mot contenant un seul proclitique n'est pas une préposition, alors il est segmenté en deux parties.

La dernière règle concerne le proclitique pr3 : chaque mot est segmenté en respectant les règles précédentes, avec une autre condition qui consiste à changer les mots contenant le proclitique *l* suivi par l'article défini *Al*. Cette règle supplémentaire consiste à ajouter le caractère *A* à l'article défini *Al* qui a dû être supprimé à cause du phénomène d'agglutination et afin de respecter les règles morphosyntaxiques de l'arabe. Par exemple, si un mot commence avec *ll* ou *wll* ou *fll*, après segmentation, le mot devient *l+ Al*, *w+ l+ Al* ou *f+ l+ Al*.

4.4 Expériences et résultats

Dans cette section on présente les données que nous avons utilisé pour entraîner les modèles de prédiction d'étiquetage morphosyntaxique et de segmentation (section 4.4.1) ainsi que les

expériences sur l'étiquetage morphosyntaxique (section 4.4.2). Des analyses des résultats sur la prédiction de segmentation sont présentés dans la section 4.4.3 ainsi que des expériences sur la traduction automatique (section 4.4.4).

4.4.1 Données

Pour entraîner nos modèles de prédiction d'étiquettes morphosyntaxiques et de segmentation, nous avons utilisé l'Arabic Treebank (ATB) constitué de 498 339 tokens (18 826 phrases).

Chaque token, tel qu'il apparaît dans sa forme originale dans une dépêche est accompagné dans l'Arabic Treebank par sa translittération en Buckwalter, par toutes les possibilités de diacritisation de ce mot, par les séquences d'étiquettes morphosyntaxiques ainsi que par les analyses morphologiques possibles de ce mot. L'analyse morphologique correcte est indiquée dans l'ATB par une * (voir chapitre 2, figure 2.4).

Selon l'Arabic Treebank, environ 17 % des tokens doivent être segmentés. Il faut noter que le nombre des combinaisons observées POS+pr1+pr2+pr3 dans tout l'Arabic Treebank est seulement 88, c'est-à-dire bien moins que le nombre total d'étiquettes composées possibles (384).

Pour la tâche de segmentation, nous avons lancé une série d'expériences avec des ensembles de descripteurs de complexité croissante. Nous avons, en premier lieu, évalué l'étiqueteur morphosyntaxique (section 4.4.2). Ensuite, nous avons étendu le jeu de descripteurs pour introduire la prédiction de proclitiques. Différentes approches ont été explorées et décrites dans la section 4.4.3. Tous ces modèles ont été évalués avec une validation croisée, en utilisant des ensembles de 2 000 phrases.

Pour la tâche de traduction, nous avons évalué l'influence du changement des outils de prétraitement. Les tests ont été évalués sur des dépêches AFP arabe-français, délivrées dans le cadre du projet SAMAR. Les systèmes de traduction ont été entraînés sur 145 mille paires de phrases extraites du corpus comparable (chapitre 3). Le texte original contient 3,3 millions de tokens, correspondant à 106 mille types, tandis que le nombre de tokens dans le texte prétraité passe à 3,6 millions de tokens, comprenant seulement 75 mille types. La partie français du corpus parallèle contient 4,2 millions de tokens, dont 61 mille types. L'évaluation de la traduction a été mesurée avec BLEU (Papineni et al., 2002), METEOR (Banerjee et Lavie, 2005) et HTER (Snover et al., 2006b).

4.4.2 Étiquetage morphosyntaxique (POS)

Nous avons commencé par entraîner un *petit* modèle de prédiction des étiquettes morphosyntaxiques. Ce dernier contient seulement des tests unigrammes sur une fenêtre de trois mots et des tests limités sur les préfixes et les suffixes. Ce système atteint un taux d'erreur de 5,34 %.

La fenêtre contextuelle a été par la suite étendue en incluant les trois mots précédents et suivants, et en effectuant des tests bigrammes et trigrammes sur l'observation. Une amélioration de plus d'un point a été obtenue avec les nouveaux traits. La figure 4.6 montre le taux d'erreur à chaque fois qu'on ajoute de nouveaux traits. On note qu'à chaque fois que de nouveaux traits sont ajoutés, le taux d'erreur sur la détection des étiquettes morphosyntaxiques baisse.

Les meilleurs résultats à ce jour ont été obtenus avec un système comprenant initialement plus de 210 millions de traits, à partir duquel les termes de régularisation ℓ_1 choisissent seulement 2,4 millions de traits.

Ce système donne un taux d'erreur de **4,2 %**, qui a été comparé à MADA. Ce dernier a un taux d'erreur de 3,77 %[4] pour l'étiquetage morphosyntaxique. Il faut souligner que pour notre approche ces résultats ont été obtenus sans utiliser BAMA. Cela signifie que notre

[4]Nous avons réduit les 34 étiquettes morphosyntaxiques de MADA à 24 pour pouvoir comparer les deux segmenteurs.

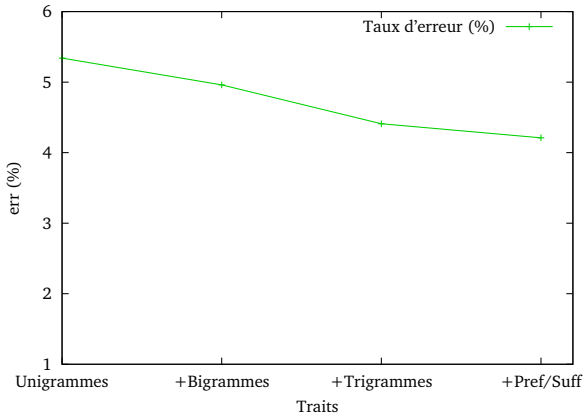

Figure 4.6: *Taux d'erreur sur la détection des étiquettes morphosyntaxiques pour des modèles de complexité croissante.*

système traite un problème plus complexe que celui que traite MADA, puisqu'on n'utilise aucun analyseur morphologique ni désambigüiseur préalable.

4.4.3 Prédiction de segmentation

Trois types de segmentation ont été comparés. Dans le premier type (SEG), les proclitiques sont prédits sans utiliser aucun trait morphosyntaxique. Dans le deuxième type (POS-puis-SEG), les étiquettes morphosyntaxiques sont prédites en premier lieu sur les unités non segmentées, et ensuite, utilisées comme un trait supplémentaire pour prédire les proclitiques. Dans le dernier type de segmentation (POS+SEG), les étiquettes morphosyntaxiques et les proclitiques sont prédits simultanément.

4.4.3.1 Résultats

Le tableau 4.2 présente le nombre de traits selon le type de segmentation ainsi que le nombre des traits actifs sélectionnés par Wapiti pour chaque modèle. Lors de la construction des modèles de prédiction pour les tâches POS+SEG, nous avons réduit le nombre de paramètres puisque la combinaison des paramètres utilisés pour les tâches POS et SEG implique l'utilisation d'un très grand nombre de paramètres – difficile à gérer techniquement. Les fenêtres glissantes pour les préfixes et suffixes ont été donc réduites à un mot pour le type de segmentation POS+SEG, et en conséquence pour le type de segmentation POS-puis-SEG. Pour cette dernière tâche, le nombre de traits représente la somme des traits actifs pour l'étiquetage morphosyntaxique (POS) et pour la prédiction des étiquettes de segmentation.

Par conséquent, le nombre total de traits pour le type de segmentation POS+SEG est de 1 511 millions, engendrant plus de 4 millions de traits actifs. En utilisant le même motif (*pattern*), le nombre total de traits possibles est de 245 millions pour le type de segmentation POS-puis-SEG, avec 3 millions de paramètres actifs (plus de traits que pour le type de segmentation SEG).

Dans le tableau 4.3, on compare les performances des trois types de segmentation. Les taux d'erreurs sont calculés pour (i) la prédiction de l'étiquette composée pr1+pr2+pr3, (ii) la prédiction de chaque proclitique indépendamment et (iii) la segmentation. Ce dernier score évalue seulement la sortie de segmentation, qui peut être correcte même si la catégorie du proclitique prédite est erronée.

Même pour le modèle le plus petit (SEG), le taux d'erreur de prédiction des proclitiques pr1+pr2+pr3 est inférieur à 1 % (seulement 0,78 % pour la prédiction jointe des 3 proclitiques).

Type de segmentation	#traits	#traits actifs
POS	195,6M	2 262,4K
SEG	64,3M	574K
POS-puis-SEG	245,0M	3 024,4K
POS+SEG	1 511,5M	4 168,9K

Tableau 4.2: *Nombre total de traits et de traits actifs pour chaque type de segmentation.*

Type de segmentation	SEG	POS-puis-SEG	POS+SEG
pr1+pr2+pr3	0,78 %	0,64 %	**0,60** %
pr1	0,22 %	**0,18** %	**0,18** %
pr2	0,46 %	0,35 %	**0,34** %
pr3	0,13 %	0,13 %	**0,11** %
POS	-	4,20 %	**3,72** %
Après segmentation	0,55 %	0,42 %	**0,40** %

Tableau 4.3: *Taux d'erreur de la segmentation pour les différents types de segmentation*

Lorsque l'étiquette morphosyntaxique POS est prédite en premier lieu (POS-puis-SEG), le taux d'erreur pour la prédiction des proclitiques et pour la segmentation baisse. Finalement, le type de segmentation POS+SEG donne les meilleurs résultats, avec un taux d'erreur de 0,6 % comparé à 0,64 % pour le type de segmentation POS-puis-SEG.

La prédiction jointe des proclitiques et des étiquettes morphosyntaxiques permet d'améliorer la prédiction des étiquettes morphosyntaxiques. Un taux d'erreur de 3,72 % est obtenu contre 4,2 % si l'étiquette morphosyntaxique est prédite indépendamment des proclitiques. On peut observer que l'étiquette morphosyntaxique est un trait important, puisqu'elle permet une amélioration de 0,18 points pour la prédiction des proclitiques pr1+pr2+pr3.

Finalement, le taux d'erreur de segmentation est de 0,55 % pour le type de segmentation SEG et il baisse jusqu'à 0,40 % pour le type de segmentation POS+SEG.

MADA D2 est le type de segmentation de MADA qui effectue la segmentation la plus proche de celle que nous effectuons (Habash et Sadat, 2006). Nous avons donc comparé notre segmentation à celle de MADA D2 et nous avons observé que notre approche nous permet d'avoir une meilleure segmentation : le taux d'erreur du type de segmentation POS+SEG est de 0,40 %, légèrement mieux que 0,57 %, le taux d'erreur obtenu par MADA D2 sur les mêmes données. Une telle comparaison n'est pas possible avec MorphTagger, puisqu'il sépare les proclitiques ainsi que les pronoms possessifs et les pronoms objet alors que dans notre référence seuls les proclitiques sont séparés.

4.4.3.2 Analyse d'erreur et segmentation

Le tableau 4.4 présente plus de détails concernant les erreurs générées par la segmentation effectuée après la prédiction conjointe des proclitiques et des étiquettes morphosyntaxiques (POS+SEG). La plupart des erreurs de segmentation générées par pr1 et pr2 sont des mots qui n'ont pas été segmentés (seg- dans le tableau) alors que pour pr3, la plupart des erreurs sont des sur-segmentations (seg+ dans le tableau). Ces sur-segmentations du proclitique pr3 sont dues d'une part au fait que l'article défini *Al* peut être attaché à des noms. Et d'autre part, au fait qu'ils peuvent faire partie d'un nom propre en arabe. Pour cette raison, les noms propres contenant l'article défini *Al* peuvent être segmentés par erreur. En se basant sur ces

résultats, nous avons donc choisi d'effectuer seulement la segmentation des proclitiques pr1 et pr2 s'ils existent.

Type de segmentation	pr1 (%)		pr2 (%)		pr3 (%)	
	seg+	seg-	seg+	seg-	seg+	seg-
POS+SEG	39,5	60,5	28,9	65,3	76,8	23,2

Tableau 4.4: *Détails sur les erreurs de segmentation*

Concernant le proclitique pr2, on note qu'à peu près 94,2 % des erreurs concernent les erreurs de segmentation (seg+ et seg-) et environ 6 % des erreurs résultent d'une attribution erronée d'étiquettes par exemple PREP au lieu de SUB, etc.

Comme mentionné dans la section 4.3.3, la détection de l'article défini *Al* est très utile pour la segmentation, et plus particulièrement pour les mots contenant le proclitique *l* suivi par *Al*. Par exemple, le mot للوطن (*llwtn*, qui signifie *pour la patrie*) est marqué par NOUN+none+PREP+DET et devient après la segmentation *l+ Alwtn* selon les règles que nous avons définies. Si l'article défini *Al* n'est pas détecté, ce mot sera segmenté à tort par *l+lwtn* (*pour pour patrie*).

Si on analyse plus minutieusement les erreurs, on note que 6,70 % des erreurs sont liées à une sur-segmentation des entités nommées. Par exemple, le nom propre بلقاضي (*blqADy*) a été étiqueté par noun+none+PREP+none au lieu de noun_prop+none+none+none et donc le mot a été segmenté par erreur et devient *b+ lqADy* puisqu'il commence par la lettre *b*. On note également que le mot *qADy/AlqADy* en arabe signie *juge*, ce qui peut affecter la prédiction des étiquettes morphosyntaxiques.

Le nombre de mots uniques dans le corpus parallèle de la figure 4.1 a été recalculé après la segmentation et il est présenté par la figure 4.7. On note que le vocabulaire a été réduit

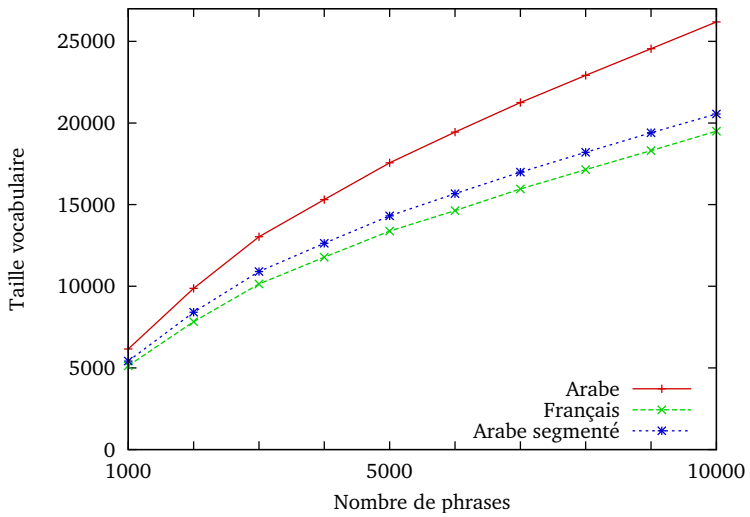

Figure 4.7: *Impact de la segmentation sur la croissance du vocabulaire*

grâce à la segmentation. Dans cette figure le type de segmentation POS+SEG a été effectué.

4.4.4 Expériences de traduction

La dernière série d'expériences a pour objectif d'évaluer l'effet de la segmentation sur les performances de la traduction de l'arabe vers le français.

Plusieurs systèmes de traduction automatique ont été entraînés en utilisant les mêmes données. La partie arabe du corpus parallèle a été prétraitée avec différents outils de prétraitements (MADA, MorphTagger, notre segmenteur). Les données prétraitées en arabe et en français ont été alignées avec MGiza++ (Gao et Vogel, 2008). Le décodeur Moses (Koehn et al., 2007) a été utilisé pour symmétriser les alignements et extraire les phrases avec une longueur maximale de 7 mots. Les poids des paramètres sont fixés après optimisation avec MERT (Och, 2003).

Deux types de segmentation ont été testés pour MADA : MADA-D2 puisqu'il sépare les mêmes proclitiques que ceux dans notre approche (Habash et Sadat, 2006), et MADA-TB puisque c'est le type de segmentation recommandé pour les applications de traduction automatique (El Kholy et Habash, 2012). Un récapitulatif des scores BLEU, METEOR et TER est présenté dans le tableau 4.5.

	BLEU	METEOR	TER
MADA D2	32,8	54,2	60,3
MADA TB	32,9	54,2	59,1
Morphtagger	33,2	54,5	58,8
SEG	32,8	53,4	59,6
POS-puis-SEG	33,1	53,7	59,4
POS+SEG	33,3	54,0	59,1

Tableau 4.5: *Impact du changement de l'outil de prétraitement de l'arabe sur la traduction automatique d'un corpus de test de l'arabe vers le français*

On note que notre outil de segmentation permet d'avoir les mêmes performances que MADA et MorphTagger pour la traduction automatique (32,8 à 33,3 points BLEU en comparaison avec 32,8 et 32,9 pour MADA et 33,2 pour MorphTagger). L'amélioration de la segmentation en utilisant l'information morphosyntaxique (POS) permet une légère amélioration sur les résultats de traduction. Le type de segmentation POS+SEG donne le meilleur score sur les trois types de segmentation développés dans notre approche et pour les trois métriques.

D'après le tableau 4.3, la prédiction jointe de l'étiquetage morphosyntaxique et la segmentation est légèrement meilleure que la prédiction séparée. Ceci est vrai aussi pour la tâche de traduction automatique. Notre outil de prétraitement avec le type de segmentation POS+SEG réussit à améliorer respectivement le score BLEU de 0,5 et de 0,4 par rapport à MADA D2 et MADA TB et donne les mêmes performances que MorphTagger.

	vitesse (m/s)
MADA	90
MorphTagger	2020
POS+SEG	2960

Tableau 4.6: *Vitesse de prétraitement calculée en mots par secondes*

Le tableau 4.6 présente une comparaison de la vitesse de traitement des différents segmenteurs. La vitesse est calculée en mots par seconde (m/s)[5]. La différence entre les vitesses

[5]Pour ces expériences nous avons utilisé des serveurs 8 x 2.3Hz Xeon HT CPU.

de traitements est très importante : notre segmenteur est environ 30 fois plus rapide que MADA et 30 % plus rapide que MorphTagger.

Comme le montrent les tableaux 4.5 et 4.6, notre outil de segmentation est aussi performant que MADA et MorphTagger pour le prétraitement des données. L'avantage de notre outil est qu'il est considérablement plus rapide que ses deux concurrents et ne nécessite l'installation d'aucune autre ressource supplémentaire. En effet, MADA ainsi que MorphTagger nécessitent l'installation de l'outil SRILM et d'avoir accès à la base de données BAMA. L'utilisation de BAMA pour l'analyse morphologique est un choix raisonnable. Par contre, la segmentation peut être effectuée beaucoup plus rapidement sans avoir besoin d'utiliser BAMA. MADA utilise également SVMTool pour prédire quelques paramètres (section 4.1.1), et c'est ce qui le rend extrêmement lent.

En comparaison, pour utiliser note outil de prétraitement – que nous avons appelé SAPA, pour *Segmentor and Part-of-speech tagger for Arabic* – nous avons besoin juste d'installer l'outil Wapiti ainsi que les modèles entraînés et les scripts de normalisation définissant les règles de segmentation.

Le modèle de prédiction des étiquettes morphosyntaxiques (POS), ainsi que celui de prédiction des étiquettes morphosyntaxiques et des proclitiques (POS+SEG) sont téléchargeables sur Internet[6]. SAPA est également téléchargeable sur Internet[7].

4.5 Conclusion

Dans ce chapitre, nous avons proposé une approche de prétraitement de l'arabe et nous avons développé un outil de prétraitement pour l'arabe : SAPA[7], pour *Segmentor and Part-of-speech tagger for Arabic*. Pour cela, nous avons utilisé Wapiti, basé sur les CRF et capable d'utiliser un grand nombre de descripteurs. Nous avons évalué notre outil de prétraitement indépendamment, ainsi que pour la tâche de traduction automatique, et nous l'avons comparé aux outils de prétraitement de l'arabe MADA et MorphTagger.

Les premiers résultats montrent que le modèle que nous avons entraîné avec Wapiti est aussi bon que MADA lorsqu'on l'utilise pour la simple tâche d'étiquetage morphosyntaxique. Le taux d'erreur pour la segmentation de proclitiques est également très faible. En outre, nous avons remarqué que la prédiction simultanée de l'étiquette morphosyntaxique ainsi que l'absence/présence de proclitiques améliore les performances : le taux d'erreur de l'étiquetage morphosyntaxique est réduit de 4,2 % à 3,72 %.

Nous avons également évalué cet outil sur une chaîne complète de traduction. Les résultats obtenus sont aussi bons que les résultats obtenus avec d'autres chaînes de prétraitements. Comparé à d'autres solutions existantes, SAPA est (i) capable de traiter plusieurs milliers de mots par seconde et (ii) il est totalement indépendant de toute autre ressource, telle qu'un analyseur morphologique ou désambigüiseur. Les résultats obtenus à ce jour sont prometteurs et suggèrent plusieurs perspectives pour améliorer encore la chaîne de prétraitement arabe. En particulier, nous poursuivons ces travaux dans le chapitre 5, la chaîne de prétraitement est complétée par la reconnaissance d'entités nommées en utilisant Wapiti et, là encore, différents scénarios peuvent être envisagés pour réaliser cette série de tâches.

[6] http://wapiti.limsi.fr/
[7] SAPA est téléchargeable à partir de https://github.com/SouhirG/SAPA

CHAPITRE 5

Traitement automatique des entités nommées en arabe : détection et application à la traduction

La détection des Entités Nommées (EN) est un élément essentiel pour de nombreuses tâches du TAL, qu'elles soient mono ou multilingues, comme la recherche d'information ou la traduction automatique. En témoignent les ateliers "Named Entities" organisés par l'ACL ainsi que, notamment, les nombreuses campagnes d'évaluation internationales (MUC, CoNLL, ACE) ou nationales (ESTER) organisées au cours des vingt dernières années.

Dans ce chapitre, nous nous intéressons au traitement des EN dans un contexte de traduction automatique statistique depuis l'arabe vers le français, dans lequel le traitement des EN pose des problèmes particuliers. Comme il a été présenté dans le premier chapitre de ce manuscrit, un système de traduction statistique apprend à traduire en se basant sur des exemples de traductions sous forme de textes parallèles. Il est fréquent qu'une EN à traduire ne soit pas présente dans les corpus parallèles qui servent à entraîner les systèmes de traduction. Comme tout mot hors-vocabulaire, elle ne pourra donc pas être traduite correctement.

D'après une étude réalisée par Habash (2008) sur des corpus extraits de journaux, environ 40 % des mots hors-vocabulaire sont des noms propres. Il existe donc d'autres causes à la prolifération des mots hors-vocabulaire, en particulier la complexité morphologique de l'arabe représente une autre cause importante de ce phénomène (Heintz, 2008). La fréquence des formes morphologiquement complexes impose également d'effectuer des analyses préalables afin de simplifier le vocabulaire (cf. chapitre 4).

Le traitement des mots hors-vocabulaire ne peut donc pas être général et doit être adapté selon les types de mots. La stratégie par défaut consiste à recopier la forme inconnue dans la sortie en langue cible. Cette stratégie est opérante lorsque les langues source et cible ont le même alphabet; dans le cas de l'arabe, elle s'avère inappropriée. Pour améliorer la sortie, il sera ainsi nécessaire de consulter des dictionnaires (Daumé III et Jagarlamudi, 2011), ou encore de produire une forme translittérée en alphabet latin (Hermjakob, Knight et Daumé III, 2008; Zhang et al., 2011; Zhang, 2012). Ce traitement ne semble pourtant pas utile pour d'autres types de mots, tels que les verbes ou les adjectifs (Hermjakob, Knight et Daumé III, 2008). La prédétection des EN, qui permet de leur appliquer des traitements différentiels, apparaît donc comme un traitement potentiellement utile à la traduction.

L'étiquetage en EN en langue arabe représente de nombreux défis intéressants : l'arabe se

caractérise par le manque de ressources dictionnairiques et surtout par l'absence de distinction majuscule/minuscule qui est un indicateur très utile pour identifier les noms propres dans les langues utilisant l'alphabet latin.

En langue arabe, la traduction automatique des noms propres n'est pas une tâche simple et ceci est dû essentiellement à deux problèmes principaux : l'ambiguïté et la variabilité. D'une part, les noms propres de personnes ont souvent une signification, ce qui peut engendrer des confusions lors de la traduction. En Afrique du Nord, il est fréquent de trouver des noms propres contenant بن (Ben, fils de). On peut également trouver des noms propres comme محمد التونسي (Mohamed Altounsi), qui peut être traduit par *Mohamed le tunisien*. عبد (Abd) signifie *serviteur de*, peut être associé à un nom décrivant Dieu[1].

D'autre part, il y a aussi une grande variabilité pour écrire les noms propres et ceci est dû surtout aux variations dialectales et régionales; par exemple le prénom جميلة est translittéré par *Jamila* en tunisien, *Djamila* en algérien, *Gamila* en égyptien.

De plus, il est courant dans les pays du monde arabe d'utiliser des *dénominations* ou des *surnoms* à la place des prénoms de personnes. Dans certains pays du Moyen-Orient par exemple, au lieu de l'appeler par son prénom, la mère d'Ali est appelée *Oum Ali* et le père d'Ali est appelé *Abu Ali*[2]. On trouve également certaines dénominations qui précèdent les prénoms de personnes. Ces dénominations varient souvent selon la fonction de la personne en question comme par exemple *Docteur Ali* (au lieu d'Ali ou de Monsieur Ali), si Ali a obtenu un diplôme de doctorat, ou *Professeur Ali* si Ali est un professeur. Le terme *Cheikh*[3] remplace Monsieur (Cheikha remplace Madame) et est utilisé pour dénommer quelqu'un qui a de l'expérience ou qui appartient à un certain niveau social. Le terme Hadj[4] est courant également et remplace Monsieur (Hadja pour remplacer Madame). La dénomination complique donc la tâche de traduction dans certains cas.

Toutes ces formes de noms propres peuvent générer des ambiguïtés lors de la traduction.

Les noms de certains lieux peuvent aussi générer des conflits lors de la traduction, comme par exemple, تونس (Tunis) peut être traduite en français par la ville *Tunis* ou *la Tunisie*, au sens de Tunis, capitale de la Tunisie. En ce qui concerne les noms d'organisations, ils sont plus difficile à reconnaître en arabe puisqu'ils prennent le plus souvent la forme de groupes nominaux complexes. Pour les langues latines, comme le français ou l'anglais, on sait que *L'Organisation Mondiale de la Santé* par exemple est un nom d'organisation facile à reconnaître puisque tous les mots pleins qui le constituent commencent par des majuscules, alors qu'en arabe ce n'est pas le cas. Plus une EN est longue en arabe, plus il est difficile de déterminer ses limites (le nombre de mots la constituant).

Nous proposons dans ce chapitre une étude complète sur les EN en arabe. Un système de détection d'EN pour l'arabe a été développé et combiné avec notre outil de segmentation de l'arabe présenté dans le chapitre 4. Les EN sur lesquelles nous travaillons correspondent à trois grandes classes non-structurées : lieux, personnes et organisations (Rosset et al., 2012) puisque notre corpus de base est annoté seulement pour ces trois catégories. Une approche d'adaptation par auto-apprentissage a été proposée. Finalement, une méthode de traduction des EN à l'aide de dictionnaires bilingues est proposée. À notre connaissance, il n'y a pas eu de travaux précédents qui présentent une étude complète sur les EN en arabe, depuis le prétraitement jusqu'à la traduction. Nos travaux sont à rapprocher de ceux de Hálek et al.

[1]Il existe 99 noms décrivant Allah. Ils peuvent tous être des noms propres, comme par exemple كريم (Karim) traduit par 'généreux', ou aussi عبد الكريم (Abdelkarim).

[2]Dans ce cas, Ali est généralement le fils aîné.

[3]Le terme Cheikh est traduit littéralement par vieux. Dans ce cas, ce terme indique le respect vis à vis de la personne concernée et que cette personne appartient à un certain niveau social.

[4]Le terme Hadj est à l'origine attribué à quelqu'un qui vient d'avoir accompli son pèlerinage à la Mecque.

(2011) qui détectent les EN en tchèque pour ensuite proposer au décodeur des traductions de ces entités nommées extraites de Wikipedia.

La suite du chapitre est organisée comme suit. Un échantillon sur les travaux existants sur la détection et la traduction des EN est présenté dans la section 5.1. La section 5.2 présente le contexte et les motivations à l'origine de ces travaux. Les données utilisées pour l'ensemble de nos expériences sont présentées dans la section 5.3. Notre approche de détection des entités nommées ainsi que le système de base de détection des entités nommées (NERAr) et les expériences d'adaptation sont décrits dans la section 5.4. La section 5.5 présente les expériences effectuées sur la pré-traduction des entités nommées en utilisant des dictionnaires ainsi qu'une analyse des résultats obtenus. Finalement, la section 5.6 conclut ces travaux et donne quelques perspectives.

5.1 État de l'art

Dans cette section on présente principalement quelques travaux effectués sur la détection des EN. Des travaux sur l'adaptation sont également présentés ainsi que des travaux qui traitent le problème de traduction des EN.

5.1.1 Étiquetage en EN pour l'arabe

Les premiers travaux sur la reconnaissance des EN pour l'arabe datent de 1998 et reposent sur des méthodes à base de règles (Maloney et Niv, 1998). Des travaux plus récents également sur les EN ont été réalisés par Shaalan et Raza (2009) ou par Zaghouani et al. (2010). Samy, Moreno et M. Guirao (2005) utilisent un corpus parallèle aligné pour extraire des EN en arabe. Ils extraient les EN en espagnol à l'aide d'un étiqueteur à base de règles enrichies avec un lexique monolingue espagnol, ensuite cherchent, en utilisant les alignements, pour chaque entitée annotée en espagnol la traduction lui correspondant. Une fois les correspondances effectuées, les entités nommées sont translittérées vers l'arabe. L'avantage principal de cette approche est que le corpus parallèle joue un double rôle comme étant une ressource et une cible en même temps. Che et al. (2013) utilisent des données parallèles chinois-anglais pour améliorer les performances d'un détecteur d'EN en chinois en utilisant des alignements entre mots. Les expériences de Zitouni et al. (2005) reposent sur l'utilisation des techniques d'apprentissage automatique (des *Maximum Entropy Markov Models*) en considérant des jeux de descripteurs idoines, et parviennent à de très bons résultats sur les données de la campagne ACE 2004.

Ces travaux ont été prolongés en particulier par Benajiba et ses co-auteurs, et ont donné lieu notamment à la construction du corpus ANER (voir section 5.3.1). Dans une première approche, Benajiba et Rosso (2007) explorent un étiquetage fondé sur le maximum d'entropie. Cette approche est étendue ensuite en décomposant la prédiction en deux temps : d'abord les frontières de l'EN en introduisant des catégories morpho-syntaxiques (POS), puis à la détermination de son type. Une seconde approche, fondée sur l'utilisation des CRF (Benajiba et Rosso, 2008) a permis d'explorer l'intégration de l'ensemble des traits dans un modèle unique, amenant à de meilleures performances, essentiellement en termes de rappel. Benajiba, Diab et Rosso (2008) montrent également l'efficacité d'un pré-traitement des textes pour séparer les différents constituants du mot (proclitiques, lemme, et enclitiques). Abdul Hamid et Darwish (2010) intègrent des traits intra-mot (notamment n-grammes de caractères) dans une modélisation CRF. Cette approche permet de capturer implicitement les caractéristiques morphosyntaxiques, qui sont introduites explicitement par l'analyse préalable réalisée dans les expériences de Benajiba et Rosso (2008).

Al-Jumaily et al. (2012) présentent, dans des travaux plus récents, un système de détection des entités nommées, qui permet de détecter, en temps réel, l'apparition de certaines entités

nommées et des événements dans des dépêches. Ce système peut être utilisé pour les applications web.

Adaptation et combinaison de systèmes

En apprentissage automatique, l'adaptation consiste à développer un système de traitement pour un domaine cible à partir de données et/ou d'un système de traitement développé pour un domaine source. D'un point de vue statistique, cela implique que les distributions des exemples observés sont différentes au moment de l'apprentissage et au moment du test.

Cette problématique a fait l'objet de multiples propositions en modélisation statistique des langues comme par exemple l'étude de Bellagarda (2001) pour les modèles statistiques de langue, ou encore les états de l'art produits par Daume III et Marcu (2006) et par Blitzer (2008) : combinaison linéaire de systèmes entraînés sur la source et la cible, utilisation de pondérations différentielles pour les exemples de la source et de la cible (Jiang et Zhai, 2007), utilisation de descripteurs spécifiques pour les exemples source et cible (Daume III, 2007), etc. On se reportera, par exemple, à Daume III et al. (2010) pour un échantillon de travaux récents. Dans un cadre non supervisé, la stratégie la plus commune est l'auto-apprentissage (*self-training*) consistant à engendrer automatiquement des données d'apprentissage pour le domaine cible à partir du système source (Mihalcea, 2004).

Concernant le repérage des EN, le problème de l'adaptation se pose avec une acuité particulière, due au fait que les EN (i) sont souvent associées avec un thème particulier et (ii) ont également des distributions d'occurrences très variables dans le temps. Cette problématique est étudiée en particulier par Béchet, Sagot et Stern (2011) qui (i) combinent deux approches d'étiquetage en EN pour le français : une approche symbolique avec une approche probabiliste et (ii) adaptent le système probabiliste fondé sur un processus discriminant à base de CRF, au domaine des données de test.

Plusieurs travaux se sont focalisés sur l'adaptation du vocabulaire dans le texte, comme ceux de Allauzen (2003) qui portent sur l'adaptation du vocabulaire et du modèle de langue d'un système de transcription automatique.

5.1.2 Traduction des EN

Plusieurs travaux ont été effectués pour évaluer et améliorer la traduction des entités nommées. Santanu et al. (2010) translittèrent les EN dans la langue source (anglais) vers la langue cible (bengale), afin de les aligner. Ils montrent que les alignements des EN améliorent les performances de la traduction automatique jusqu'à 4,6 points BLEU.

Al-Onaizan et Knight (2002b) présentent un algorithme qui améliore la traduction des entités nommées de l'arabe vers l'anglais, en utilisant des ressources bilingues et monolingues. Ils combinent à la fois la traduction (en utilisant des dictionnaires bilingues) avec la translittération (Al-Onaizan et Knight, 2002a) afin de trouver la meilleure traduction d'une entité nommée. Jiang et al. (2007) combinent également la translittération avec des données extraites du web et choisissent la meilleure hypothèse de traduction.

D'autres approches comme celle de Kashani et al. (2007) améliorent les performances d'un système de traduction en utilisant une méthode qui propose des translittérations des mots inconnus.

Hassan, Fahmy et Hassan (2007) améliorent la traduction des entités nommées en enrichissant les systèmes de traduction avec des entités nommées et leurs traductions extraites à partir de corpus comparables et parallèles. Ling et al. (2011) utilisent des liens web pour récupérer des traductions des entités nommées. Moore (2003) a développé une approche pour extraire des traductions d'entités nommées à partir d'un corpus parallèle. Les entités nommées sont extraites du corpus en langue source (anglais) en utilisant les majuscules comme indice.

Abdul Rauf (2012) améliore la traduction des entités nommées en utilisant des dictionnaires, constitués à la base à partir de mots inconnus. Leurs traductions sont extraites à partir de corpus comparables en utilisant la recherche d'information.

L'idée de base de notre approche est assez proche de celle de Hálek et al. (2011) qui ont effectué une étude préliminaire pour améliorer la traduction automatique des EN en tchèque avec Wikipedia. Les EN sont détectées et ensuite des traductions de ces EN sont proposées au décodeur.

5.2 Contexte et motivations

Les données utilisées dans ce chapitre sont principalement des dépêches journalistiques de l'AFP. Le système de traduction de base utilisé dans ces travaux est entraîné sur un corpus parallèle arabe-français de 265 mille phrases extraites entre décembre 2009 et juillet 2012 avec la méthode présentée dans le chapitre 3. Le mois de décembre 2010 ne fait pas partie de ces données puisqu'il a servi à extraire les corpus de test et de développement.

Nous avons présenté dans le chapitre précédent l'outil de prétraitement de l'arabe que nous avons développé (SAPA). Cet outil a la particularité d'être (i) indépendant de toute autre ressource et (ii) beaucoup plus rapide que d'autres outils. Les résultats obtenus précédemment nous ont encouragé à compléter la chaîne de prétraitement avec la détection des entités nommées.

5.2.1 Problématique

Comme nous l'avons déjà mentionné, il est fréquent qu'une EN à traduire ne soit pas présente dans les corpus parallèles qui servent à entraîner les systèmes de traduction. À partir des résultats obtenus dans le chapitre 3, nous avons en effet observé qu'environ 25 % des formes inconnues correspondent à des EN.

Une première analyse menée sur une extraction automatique des EN[5] en arabe sur le corpus AFP (250 dépêches par jour) permet d'illustrer ce phénomène, particulièrement présent dans des textes traitant de l'actualité. La figure 5.1 montre l'évolution du nombre d'EN nouvelles apparaissant chaque semaine sur la période mars à juillet 2012 dans les dépêches AFP pour les trois catégories : lieux, personnes et organisations.

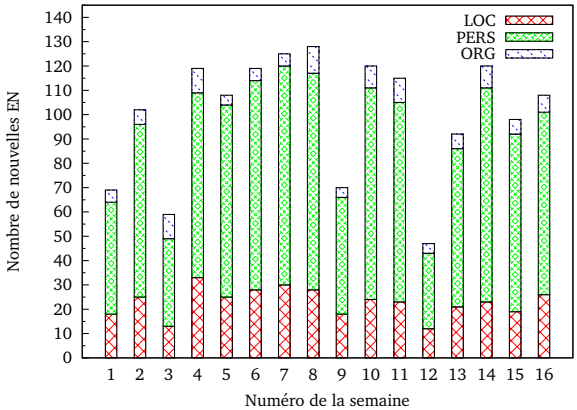

Figure 5.1: *Évolution du nombre d'EN nouvelles chaque semaine, de mars à juillet 2012.*

[5]Les EN ont été extraites avec notre détecteur d'EN qui sera présenté dans la section 5.4.

À partir d'un nombre initial de 37 446 EN distinctes (extraites sur la période décembre 2009 - mars 2012), on observe que chaque semaine, de façon assez régulière, une centaine d'EN nouvelles apparaît. Sur cette période de 16 semaines, au total 1 599 nouvelles EN sont apparues : 1 122 personnes, 366 lieux, et 111 organisations. Non connues d'un système de traduction qui aurait été entraîné sur le corpus correspondant à la période initiale, ces EN seront nécessairement mal (ou pas) traduites.

Nous nous sommes donc intéressés à mettre en place des stratégies pour la traduction des noms propres, ne nécessitant pas l'entraînement du système de traduction sur un corpus couvrant toujours plus de données.

Afin d'avoir une idée sur le pourcentage d'erreurs de traduction des EN, nous avons utilisé le corpus parallèle Arcade II (Chiao et al., 2006) annoté en arabe et en français en EN. C'est le seul moyen de rendre possible une évaluation de la traduction des EN. Une correction/vérification manuelle de ce corpus a été effectuée afin de rendre le nombre d'EN en source et en cible le même pour chaque type d'EN. Le corpus Arcade II a été traduit avec le système de traduction AFP présenté au début de la section 5.2. Le tableau 5.1 présente le pourcentage des EN traduites correctement.

	LOC	PERS	ORG	**Total**
système de base	70,38 %	51,24 %	44,63 %	55,41 %

Tableau 5.1: *Pourcentage des EN bien traduites par le système de traduction AFP (calculé par rapport à la référence).*

Une vérification automatique des EN a été réalisée pour voir le taux d'EN traduites correctement. Pour chaque ligne, toutes les EN se trouvant dans la référence ont été recherchées dans la ligne correspondante de la sortie de traduction automatique. D'après le tableau 5.1, presque la moitié (45 %) des EN n'ont pas été traduites correctement.

5.2.2 Objectifs

Ce travail se focalise principalement sur deux objectifs. Le premier est d'étudier et de comparer différentes versions d'un modèle de base de détection d'EN : (i) une première version du modèle de base, suivie d'enrichissements avec différents paramètres, (ii) une deuxième version exploitant une adaptation non-supervisée.

Le deuxième objectif consiste à tester différentes méthodes de traduction des EN à l'aide de dictionnaires bilingues de noms propres pour améliorer la traduction automatique des EN et réduire le nombre de mots inconnus qui sont mal traduits.

5.3 Description des données

Dans ces travaux, deux types de corpus ont été utilisés : des corpus monolingues pour la tâche de détection des EN et des corpus bilingues pour la tâche de traduction.

5.3.1 Corpus monolingues en arabe

Trois corpus monolingues ont été utilisés : le corpus ANER, le corpus AFP et le corpus *Gold AFP*.

Corpus ANER : Les expériences de détection des EN ont été réalisées sur le corpus ANER[6] (Benajiba, Rosso et Benedí, 2007) composé de plus de 150 mille occurrences de mots (4 871 phrases). Il peut être considéré comme le corpus de référence pour la tâche.

[6] http://users.dsic.upv.es/~ybenajiba/downloads.html

Le corpus distingue 4 types d'EN : lieu (LOC : 40 % des EN observées), personne (PERS : 32 %), organisation (ORG : 18 %) et une classe "divers" regroupant tous les autres types (MISC : 10 %)[7]. Il utilise le schéma d'annotation BIO et distingue donc 9 étiquettes. La répartition en EN est présentée dans le tableau 5.2.

	LOC	ORG	PERS	MISC
Entités nommées	4 431	2 026	3 602	1 117
Entités nommées distinctes	1 004	657	1 446	437

Tableau 5.2: *Répartition des entités nommées dans le corpus ANER*

La figure 5.2 montre un extrait d'une phrase en format BIO. L'étiquette B-X (*Begin*) indique le premier mot d'une EN de type X. L'étiquette I-X (*Inside*) indique qu'un mot fait partie d'une EN mais qui n'est pas le premier mot. L'étiquette O (*Outside*) est utilisée pour les mots qui ne sont pas des EN.

وقال	ستيفان	دوجاريك	المتحدث	باسم	الامم	المتحدة
O	B-PERS	I-PERS	O	O	B-ORG	I-ORG
Et a dit	Stéphane	Dujarric	le porte-parole	des	Nations	Unies

Figure 5.2: *Exemple d'un extrait de phrase en format BIO.*

Corpus monolingue AFP : Nous disposons dans ce cadre de ressources supplémentaires pour adapter la détection des EN:

- de données du domaine (AFP), non-annotées (130 milles phrases, 3 500K mots);
- d'un corpus de test, Gold AFP, annoté manuellement en EN (LOC, PERS et ORG) et constitué de 900 phrases issues des données de l'AFP.

Les dépêches AFP traitées dans notre application diffèrent substantiellement des données du corpus ANER, qui contient à la fois des articles de presse, des données collectées en ligne, en particulier des extraits de Wikipedia. Il existe également un décalage temporel entre la constitution du corpus ANER (2007) et les données que nous devons traiter, qui sont postérieures à 2009.

5.3.2 Corpus bilingues

Le corpus de développement utilisé pour l'optimisation est constitué de 1 000 phrases extraites de dépêches AFP (décembre 2010). Pour évaluer notre méthode nous disposons de deux corpus de test : le corpus AFP (extrait par la méthode présentée dans le chapitre 3) et le corpus Arcade II (Chiao et al., 2006).

Corpus parallèle AFP : Le corpus de test AFP adapté au modèle de traduction est constitué de 1 000 phrases extraites de dépêches datées de décembre 2010 également.

Corpus parallèle Arcade : Le corpus de test Arcade II est annoté en arabe et en français en EN. Ce corpus a été vérifié manuellement pour réduire les EN à trois types. Des alignements ont été également corrigés ainsi que certaines phrases du corpus afin d'avoir le même nombre d'EN en source et en cible. Finalement, le corpus de test utilisé pour les expériences contient 1312 lignes dans lesquelles il y a 1079 noms de lieux, 525 noms de personnes et 354 noms d'organisations.

[7]Seuls les trois premiers types sont utilisés dans nos évaluations.

5.3.3 Dictionnaires

Deux types de dictionnaires sont utilisés : les dictionnaires monolingues pour la détection des EN et les dictionnaires bilingues pour la traduction des EN.

	Sources	LOC	PERS	ORG
Dictionnaires monolingues	ANERGazet	x	x	x
	Wikipedia	x	x	x
	Geonames[8]	x	-	x
	Total	4 920	18 098	1 043
Dictionnaires bilingues	ANERGazet	x	-	-
	Wikipedia	x	x	-
	JRC[9]	x	x	-
	Geonames	x	x	x
	Total	3 810	16 470	728

Tableau 5.3: *Constitution des dictionnaires monolingues et bilingues*

5.4 Détection des entités nommées

À la suite de nombreux travaux, nous abordons cette tâche avec des outils d'apprentissage automatique et utilisons le modèle des champs markoviens conditionnels (ou CRF), avec l'implémentation Wapiti, qui permet de construire des modèles intégrant un très grand nombre de descripteurs. Cette implémentation permet (i) d'utiliser de très gros modèles incluant nominalement des centaines de millions de descripteurs, et (ii) une stratégie d'optimisation permettant de sélectionner les descripteurs les plus utiles par le biais d'une pénalité ℓ_1 (Sokolovska, Cappé et Yvon, 2009). L'utilisation de modèles statistiques pose la question de la pertinence des corpus d'apprentissage au regard des données de test. Nous traitons cette question en explorant une adaptation non-supervisée.

5.4.1 Protocole expérimental

Les expériences sont réalisées à partir de données translittérées[10] et segmentées avec l'outil de segmentation SAPA présenté dans le chapitre 4. Les scores sont calculés en utilisant l'outil d'évaluation développé pour la tâche de repérage des EN proposée dans le cadre de CoNLL 2002[11], qui calcule le rappel, la précision et la F-mesure comme suit :

$$\text{Précision} = \frac{C}{T} \qquad \text{Rappel} = \frac{C}{T*C}$$

$$F_{\beta=1} = \frac{2*\text{Précision}*\text{Rappel}}{\text{Precision}+\text{Rappel}}$$

où C représente le nombre des étiquettes correctes pour une EN, T le nombre des étiquettes trouvées, et $T * C$ le nombre des EN trouvées et correctes.

Les modèles sont évalués sur le corpus ANER par validation croisée à 10 partitions, sur des tests d'environ 25 mille mots chacun.

[8] http://download.geonames.org/export/dump/
[9] *Joint Research Center* de la Communauté européenne : http://langtech.jrc.it/JRC-Names.html
[10] http://www.qamus.org/transliteration.htm
[11] http://bredt.uib.no/download/conlleval.txt

5.4.2 Sélection de caractéristiques

Différentes versions du modèle de détection des EN ont été développées, qui incluent des jeux de descripteurs de richesse croissante. Nous décrivons ci-dessous les principales familles de descripteurs; chaque réalisation x d'un élément d'une de ces familles donne lieu à un ensemble de fonctions booléennes testant x avec chaque étiquette et avec chaque bigramme d'étiquettes possibles.

N-grammes de mots : ces caractéristiques testent tous les unigrammes, les bigrammes, les trigrammes et les quadrigrammes dans, respectivement, des fenêtres de tailles 5, 3, 4 et 5 autour du mot courant.

Préfixes et suffixes : chaque séquence d'une, deux, ou trois lettres observée à l'initiale ou à la finale d'un mot du corpus d'apprentissage donne lieu à un nouveau descripteur. L'apparition de ces préfixes et suffixes est testée dans une fenêtre de taille 5 centrée sur le mot courant. Ces traits sont très importants puisqu'en langue arabe, les mots contiennent souvent des proclitiques et enclitiques que ce trait aide à détecter. Ainsi, certains noms propres sont précédés de l'article défini *Al*, d'autres mots sont agglutinés à des pronoms personnels en fin de mot, etc.

POS-tags : ce trait concerne les étiquettes morpho-syntaxiques prédites en utilisant un modèle entraîné par Wapiti sur l'*Arabic Tree Bank*[12] (Maamouri et al., 2005a; Maamouri et al., 2005b); des détails sur cet étiqueteur[13] sont donnés dans le chapitre 4. Les tests évaluent les unigrammes d'étiquettes dans une fenêtre de taille 5, et les bigrammes d'étiquettes dans une fenêtre de taille 3.

Ponctuation et nombres : ce trait teste la présence de caractères de ponctuation et de chiffres dans le mot courant ainsi que dans les deux mots voisins.

Dictionnaires monolingues : Ce trait teste pour chaque mot $w = w_1...w_n$, si w figure dans le dictionnaire, ou s'il y figure précédé d'un ou de deux proclitiques contenant un ou deux caractères (Pref.Dict sur la figure 5.3). En langue arabe, il est possible d'avoir des EN précédées par la conjonction *et* (و) – constituée d'un seul caractère et collée au mot suivant – ou aussi précédées à la fois par une conjonction et une préposition, comme par exemple ولفرنسا (et pour la France) ou bien للجزائر (pour l'Algérie).

Les résultats de ces expériences sont reportés sur la figure 5.3, qui représente la variation globale de la précision, du rappel et de la F-mesure, ainsi que le nombre de traits actifs. On constate qu'au fur et à mesure que de nouveaux traits sont ajoutés au modèle précédent, le rappel et la F-mesure augmentent, parfois au prix d'une légère dégradation de la précision.

5.4.3 Système de détection des EN préliminaire et adaptations

Dans cette section on présente une version préliminaire du système de détection des EN ainsi que son adaptation. Dans cette section, le corpus d'entraînement pour la construction du modèle de détection des EN n'est pas segmenté[14]. Pour la suite des expériences à partir de la section 5.4.4, le corpus d'entraînement est segmenté. Le tableau 5.4 montre les résultats du système de détection des EN par validation croisée sur le corpus ANER.

Pour adapter le détecteur d'entités nommées, nous avons utilisé un corpus d'apprentissage constitué de données du domaine (AFP), ainsi qu'un corpus de test *Gold AFP* annoté manuellement pour évaluer les résultats. La section 5.4.3.1 décrit l'adaptation du système de détection des EN par auto-apprentissage, l'hybridation est présentée dans la section 5.4.3.2.

[12] http://www.ircs.upenn.edu/arabic/

[13] Les étiquettes morpho-syntaxiques sont prédites avec SAPA.

[14] Cette étude a été réalisée au cours du projet SAMAR, avant le développement de SAPA, c'est pour cette raison que les données utilisées pour ces expériences n'ont pas été segmentées.

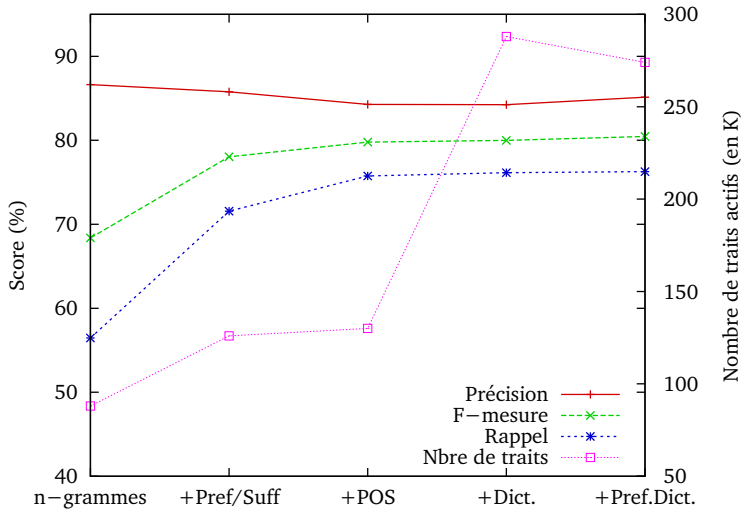

Figure 5.3: *Précision (en %), rappel (en %), F-mesure et nombre de traits actifs pour des modèles de complexité croissante (à chaque nouveau modèle, de nouveaux traits sont ajoutés).*

	Précision	Rappel	$F_{\beta=1}$
LOC	91,17 %	85,87 %	88,44
ORG	83,00 %	63,31 %	71,83
PERS	83,26 %	75,38 %	79,13
Total	86,89 %	77,56 %	81,96

Tableau 5.4: *Précision, rappel et F-mesure du modèle de base préliminaire (qui combine tous les traits) sur le corpus ANER*

5.4.3.1 Adaptation du système préliminaire par auto-apprentissage

Le système préliminaire de détection des EN, constitué à partir du corpus ANER, est utilisé pour annoter automatiquement le corpus AFP. Deux systèmes adaptés sont alors obtenus en utilisant comme corpus d'entraînement soit (i) le corpus étiqueté automatiquement seul, soit (ii) l'union des deux corpus. Le tableau 5.5 donne les résultats des trois systèmes sur les données de test AFP. On constate une baisse sensible des performances du système de base (la F-mesure passe de 81,96 sur les données de test ANER à 72,64 sur les données de test AFP). Après adaptation (AFP et ANER+AFP), on constate une amélioration de la F-mesure pour les noms de lieux et de personnes. On note que bien que le modèle AFP ait été annoté automatiquement, on arrive à avoir de meilleurs scores par rapport au modèle ANER.

5.4.3.2 Hybridation du modèle adapté de détection des entités nommées

Le processus consiste à annoter automatiquement le corpus de test par l'annotateur symbolique à base de règles de Temis[15] (Guillemin-Lanne et al., 2007), dont les performances sont données dans le tableau 5.6. Le corpus de test est ensuite annoté une seconde fois par Wapiti,

[15]Temis est l'un des partenaires du projet SAMAR. Comme déjà mentionné, cette étude a été réalisée avant le développement de SAPA, c'est pour cette raison que les données utilisées pour ces expériences n'ont pas été segmentées.

	Précision	Rappel	$F_{\beta=1}$	Modèle
LOC	89,30 %	78,85 %	83,75	ANER
ORG	50,24 %	37,72 %	43,09	
PERS	68,07 %	66,03 %	67,03	
Total	77,61 %	68,27 %	72,64	
LOC	90,81 %	77,84 %	83,83	AFP
ORG	51,01 %	35,94 %	42,17	
PERS	70,83 %	69,29 %	70,05	
Total	79,39 %	68,14 %	73,34	
LOC	91,45 %	78,18 %	84,29	ANER+AFP
ORG	51,76 %	36,65 %	42,92	
PERS	70,87 %	68,75 %	69,79	
Total	79,86 %	68,34 %	73,65	

Tableau 5.5: *Comparaison et Adaptation du système de reconnaissance d'entités nommées préliminaire sur le corpus de test AFP*

	Précision	Rappel	$F_{\beta=1}$
LOC	87,97 %	76,49 %	81,83
ORG	63,51 %	50,18 %	56,06
PERS	75,80 %	57,88 %	65,64
Total	81,03 %	67,23 %	73,49

Tableau 5.6: *Performances de l'annotateur automatique de Temis*

en considérant que les entités nommées annotées par Temis sont correctes et en n'utilisant Wapiti que pour prédire les zones qui n'ont pas été détectées comme entités nommées par l'annotateur symbolique. Les résultats sont dans le tableau 5.7.

	Modèle ANER+AFP			Modèle ANER+AFP hybride		
	Précision	Rappel	$F_{\beta=1}$	Précision	Rappel	$F_{\beta=1}$
LOC	91,45 %	78,18 %	84,29	85,81 %	82,34 %	84,04
ORG	51,76 %	36,65 %	42,92	58,49 %	55,16 %	56,78
PERS	70,87 %	68,75 %	69,79	68,02 %	72,83 %	70,34
Total	79,86 %	68,34 %	73,65	76,39 %	75,10 %	75,74

Tableau 5.7: *Comparaison de système de reconnaissance d'entités nommées adapté (ANER+AFP) avec le système de reconnaissance des entités nommées adapté et hybride*

Ces résultats montrent une amélioration par rapport aux résultats antérieurs (+2 points) en F-mesure totale ainsi qu'une perte en précision pour les lieux et les personnes. On note que le modèle hybride améliore la reconnaissance de toutes les entités nommées et surtout la reconnaissance des organisations puisque l'annotateur de Temis a de meilleures performances sur la reconnaissance des organisations.

5.4.4 Système de détection des EN de base et comparaison à l'état de l'art

Dans la section 5.4.3, nous avons présenté le système préliminaire de détection des EN sur des données non segmentées. Le système de détection des EN de base (NERAr) constitué à partir des données d'apprentissage segmentées est présenté dans la section 5.4.4.1. La section 5.4.4.2 présente les expériences effectuées pour l'adaptation de NERAr en utilisant des données extraites de dépêches de l'AFP. Les performances de NERAr sur le corpus Arcade II sont présentées dans la section 5.4.4.3.

5.4.4.1 Système de détection des entités nommées de base (NERAr)

Dans cette section, nous présentons les performances du système de détection des EN de base (NERAr pour *Named Entity Recognition for Arabic*). Ce dernier est entraîné sur des données segmentées avec SAPA et englobe toutes les caractéristiques présentées dans la section 5.4.2. Le tableau 5.8 montre les performances de NERAr sur le corpus ANER en comparaison avec les performances du système de base de Benajiba et Rosso (2008). Ce dernier utilise des données segmentées avec l'outil de segmentation de Diab (2009) pour la détection des EN.

(Benajiba et Rosso, 2008)				Système de base (NERAr)			
	Précision	Rappel	$F_{\beta=1}$		Précision	Rappel	$F_{\beta=1}$
LOC	93,03 %	86,67 %	89,74	LOC	89,75 %	89,39 %	89,57
ORG	84,23 %	53,94 %	65,76	ORG	83,37 %	66,01 %	73,68
PERS	80,41 %	67,42 %	73,35	PERS	82,49 %	78,53 %	80,46
Total	85,89 %	69,34 %	76,28[16]	Total	86,02 %	80,78 %	83,32

Tableau 5.8: *Précision, rappel et F-mesure du système de détection des EN de base (NERAr) sur le corpus ANER en comparaison avec le système de détection des EN de Benajiba et Rosso, (2008)*

On note que NERAr donne des performances comparables à l'état de l'art pour les noms de lieux, et améliore les performances pour les noms de personnes et d'organisations. La segmentation du texte avant l'entraînement du modèle de détection des EN améliore les résultats. L'évaluation du système de base sans segmentation préalable des mots du texte a donné une $F_{\beta=1}$ totale égale à 81,96 (voir tableau 5.4).

Pour vérifier l'impact de l'ajout des dictionnaires comme étant une des caractéristiques pour la détection des EN, le tableau 5.9 montre les performances du système de base sans ajout de dictionnaires. On note une dégradation de la précision pour les organisations et les noms de personnes ainsi qu'une baisse du rappel et de la F-mesure pour tous les types d'EN.

	Précision	Rappel	$F_{\beta=1}$
LOC	91,14 %	86,03 %	88,51
ORG	80,70 %	64,44 %	71,66
PERS	82,28 %	77,52 %	79,83
Total	85,99 %	78,53 %	82,09

Tableau 5.9: *Précision, rappel et F-mesure du modèle de base sans ajout de dictionnaires sur le corpus ANER*

[16]Le total reporté dans Benajiba et Rosso (2008) inclut l'EN MISC. Le total ici a été fait en calculant la moyenne.

5.4.4.2 Adaptations du système de base de détection des entités nommées

Comme précédemment, dans la section 5.4.3.1, pour adapter le détecteur d'EN, nous avons utilisé le corpus d'apprentissage constitué de données du domaine (AFP), ainsi que le corpus de test monolingue *Gold AFP* annoté manuellement.

Le système de base, constitué à partir du corpus ANER, est utilisé pour annoter automatiquement le corpus AFP comme le montre la figure 5.4. Deux systèmes adaptés sont alors

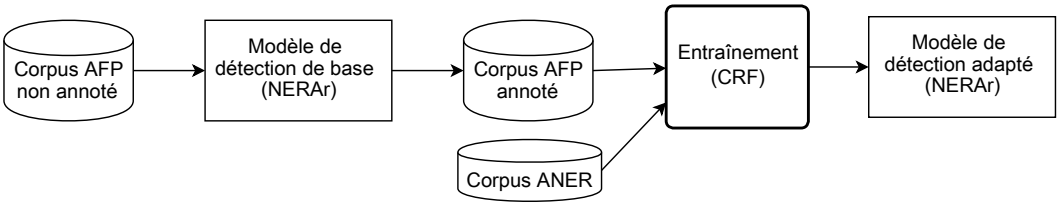

Figure 5.4: *Adaptation du système de base de détection des entités nommées par auto-apprentissage.*

obtenus en utilisant comme corpus d'entraînement soit (i) le corpus étiqueté automatiquement seul (AFP), soit (ii) l'union des deux corpus (ANER+AFP). Le tableau 5.10 donne les résultats

	Précision	Rappel	$F_{\beta=1}$	Modèle
LOC	91,30 %	80,31 %	85,46	ANER
ORG	52,20 %	33,81 %	41,04	
PERS	70,21 %	71,74 %	70,97	
Total	80,07 %	69,77 %	74,57	
LOC	90,57 %	82,11 %	86,14	AFP
ORG	55,44 %	38,08 %	45,15	
PERS	71,51 %	69,57 %	70,52	
Total	80,55 %	71,07 %	75,51	
LOC	87,20 %	85,04 %	86,10	ANER+AFP
ORG	49,79 %	41,99 %	45,56	
PERS	71,16 %	73,10 %	72,12	
Total	77,13 %	74,32 %	75,70	

Tableau 5.10: *Comparaison et Adaptation du système de reconnaissance d'entités nommées sur le corpus de test Gold AFP.*

des trois systèmes sur les données de test *Gold AFP*. Le changement de domaine entraîne une baisse sensible des performances du système de base (la F-mesure passe de 83,32 sur les données de test ANER à 74,57 sur les données de test *Gold AFP*). Après adaptation (AFP et ANER+AFP), on constate une amélioration du rappel et de la F-mesure pour tous les types d'EN. On note que bien que le modèle AFP ait été annoté automatiquement, il conduit à de meilleurs scores par rapport au modèle ANER. Les performances des modèles données dans le tableau 5.10 entraînés sur des données segmentées sont meilleures que les performances des modèles entraînés sur des données non segmentées (tableau 5.5).

Les proportions des deux corpus d'entraînement sont assez différentes (~5 milles phrases pour le corpus ANER et 130 milles phrases pour le corpus AFP). Les données AFP sont annotées automatiquement, ce qui peut engendrer du bruit. Un test supplémentaire a été donc effectué en sélectionnant seulement les 5 milles phrases qui ont le meilleur score de confiance donné par Wapiti. La F-mesure totale obtenue pour le modèle limité ANER+AFP

est de 74,90. On constate donc que plus on a de données d'adaptation, plus la détection des EN s'améliore.

Une autre expérience a été réalisée afin de vérifier l'impact des connaissances lexicales sur l'apprentissage. Un nouveau modèle de détection des entités nommées a été créé en utilisant tous les paramètres précédents sauf les informations lexicales (unigrammes, bigrammes et trigrammes de mots). Les résultats sont présentés dans le tableau 5.11. On

		Précision	Rappel	$F_{\beta=1}$	Modèle
Détect. sans aucun trait lexical	LOC	92,56 %	72,78 %	81,49	ANER
	ORG	47,80 %	30,96 %	37,58	
	PERS	72,46 %	60,05 %	65,68	
	Total	80,52 %	62,09 %	70,12	
	LOC	92,07 %	60,07 %	72,70	AFP
	ORG	64,18 %	15,30 %	24,71	
	PERS	62,54 %	60,33 %	61,41	
	Total	79,74 %	51,95 %	62,91	
	LOC	92,17 %	74,13 %	82,17	ANER+AFP
	ORG	50,60 %	29,89 %	37,58	
	PERS	67,44 %	63,59 %	65,45	
	Total	79,56 %	63,52 %	70,64	

Tableau 5.11: *Comparaison et Adaptation du système de reconnaissance d'entités nommées, entraîné sans aucun trait lexical. Évaluation sur le corpus de test Gold AFP.*

observe qu'en adaptant le détecteur d'EN, le rappel global ainsi que celui des lieux et des personnes s'améliore pour le système adapté ANER+AFP. La F-mesure pour les lieux s'améliore également. On voit que pour les organisations les performances – sauf la précision – sont moins bonnes pour le système adapté AFP par rapport au système ANER. En combinant les deux systèmes, ANER+AFP, les résultats sont améliorés pour les organisations surtout au niveau de la précision. Par contre, on observe que bien que le système n'ait aucune connaissance lexicale, il arrive quand même à reconnaitre des EN. En général, on arrive à améliorer les performances en adaptant le système de détection des EN. Une F-mesure globale de 70,64 a été obtenue pour le système adapté ANER+AFP.

5.4.4.3 Détection des EN pour le corpus Arcade II

Le corpus Arcade II annoté en EN est différent à la fois du corpus ANER et du corpus AFP. Ce corpus est très intéressant puisqu'il est annoté en EN à la fois dans les deux langues (en arabe et en français); ce qui constitue à la fois un corpus d'évaluation de la détection des EN et d'évaluation de la traduction des EN. Afin d'avoir une idée des performances du détecteur d'EN sur le corpus Arcade, – qui sera utilisé pour évaluer l'impact de la traduction des EN par la suite – nous avons évalué notre modèle de détection des EN de base sur ce corpus. Les EN détectées dans ce corpus seront par la suite prétraitées lors de la traduction automatique.

Le tableau 5.12 montre les performances du système de détection des EN de base ainsi que le système adapté (ANER+AFP) sur le corpus de test Arcade II. On constate que le modèle adapté ANER+AFP donne de meilleures performances que le modèle ANER seul sur le corpus Arcade II en termes de rappel et de F-mesure. Nous avons donc choisi le modèle adapté pour détecter les EN sur les corpus de test AFP et Arcade. Puisque le corpus Arcade est annoté en EN en source et en cible, nous avons pensé à détecter les EN en français afin d'évaluer et comparer les performances du détecteur des EN en arabe et en français.

Nous avons pour cela utilisé le détecteur d'entités nommées développé au LIMSI dans le cadre du projet Quaero (Dinarelli et Rosset, 2011) et entraîné sur les données utilisées dans

Modèle ANER				Modèle ANER+AFP			
	Précision	Rappel	$F_{\beta=1}$		Précision	Rappel	$F_{\beta=1}$
LOC	81,69 %	74,50 %	77,93	LOC	78,75 %	78,90 %	78,83
ORG	71,67 %	22,22 %	33,93	ORG	60,54 %	28,94 %	39,16
PERS	70,47 %	77,92 %	74,01	PERS	69,61 %	79,04 %	74,02
Total	77,08 %	65,38 %	70,75	Total	74,01 %	69,35 %	71,60

Tableau 5.12: *Précision, rappel et F-mesure sur le corpus Arcade II en utilisant le modèle de détection des entités nommées ANER puis ANER+AFP.*

la campagne d'évaluation ESTER2 (Galliano, Gravier et Chaubard, 2009). Les performances du détecteur d'entités nommées sur la partie français du corpus Arcade II sont données dans le tableau 5.13. On voit que la F-mesure globale du détecteur d'EN en français baisse

	Precision	Rappel	$F_{\beta=1}$
LOC	86,3 %	68,5 %	76,4
ORG	64,9 %	24,8 %	35,9
PERS	58,9 %	62,0 %	60,4
Total	74,7 %	59,0 %	65,9

Tableau 5.13: *Performances du détecteur des EN en français sur la partie français du corpus Arcade II*

par rapport à la F-mesure globale du détecteur d'EN sur l'arabe. Ceci est dû au fait que le détecteur d'EN du français a été entraîné sur des données audio. Les données audio diffèrent des données écrites sur plusieurs niveaux comme la ponctuation par exemple (la ponctuation des données transcrites est souvent incomplète). On observe également que le détecteur d'EN en français est meilleur que le détecteur d'EN de base en arabe pour les organisations. Le détecteur d'EN en arabe (cf. tableau 5.12) et le détecteur d'EN en français (cf. tableau5.13) n'ont pas les mêmes performances sur le corpus Arcade II.

5.5 Traduction automatique

Pour la traduction automatique, nous utilisons le décodeur à base de segments et *open source* Moses[17] (Koehn et al., 2007) qui intègre en particulier l'aligneur sous-phrastique MGIZA++[18] (Gao et Vogel, 2008) pour la phase d'entraînement. La table de traduction est constituée en symétrisant les alignements selon l'heuristique *grow-diag-final-and* de Moses, et contient des segments dont la longueur va jusqu'à 7 mots. L'outil SAPA (chapitre 4) est utilisé pour le prétraitement de l'arabe.

5.5.1 Intégration de dictionnaires pour la traduction des EN

L'idée est de proposer des traductions des EN tout en utilisant le même système de traduction. Lors de la phrase de prétraitement, le texte en arabe est segmenté et les EN sont détectées. Selon le type d'EN détectée, le dictionnaire bilingue approprié est consulté (lieux, personnes et organisations) afin d'éviter les ambiguïtés : un nom de rue (LOC) peut être identique au nom d'une personne (PERS), comme par exemple *Charles de Gaulle* qui peut être un nom de rue, un nom d'aéroport ou un nom de personne.

[17] http://moses.statmt.org/
[18] http://geek.kyloo.net/software/doku.php/mgiza:overview

Des propositions de traductions de cette EN – extraites des dictionnaires – sont ajoutées dans le texte à traduire sous forme de balises. Par exemple, pour le nom propre باراك اوباما (*Barack Obama*), les traductions proposées au décodeur sont représentés sous le format suivant :
<n translation="Barack Hussein Obama||Barack Hussein Obama II||Barack Obama||Barak Obama"> bArAk AwbAmA < /n>

Ces informations sont proposées à Moses selon deux modes : inclusive et exclusive.

- le mode *exclusive* impose au décodeur de choisir pour le segment concerné une des suggestions proposées dans la balise. Seules les propositions de traductions figurant dans la balise sont utilisées dans le segment à l'entrée. Tous les segments de la table de traduction qui sont en concurrence avec ce segment sont ignorés.
- le mode *inclusive* permet de mettre en concurrence les traductions spécifiées dans la balise avec toutes les entrées de la table de traduction pour ce segment. Cette option permet au décodeur d'utiliser les traductions soit à partir de la table de traduction ou de la traduction proposée dans la balise. Pour chaque segment de la phrase, les scores pour les différentes hypothèses se trouvant dans la table de traduction, ainsi que pour les hypothèses de traduction, sont calculés. Toutes les hypothèses sont ensuite comparées afin de choisir celle qui a le meilleur score.

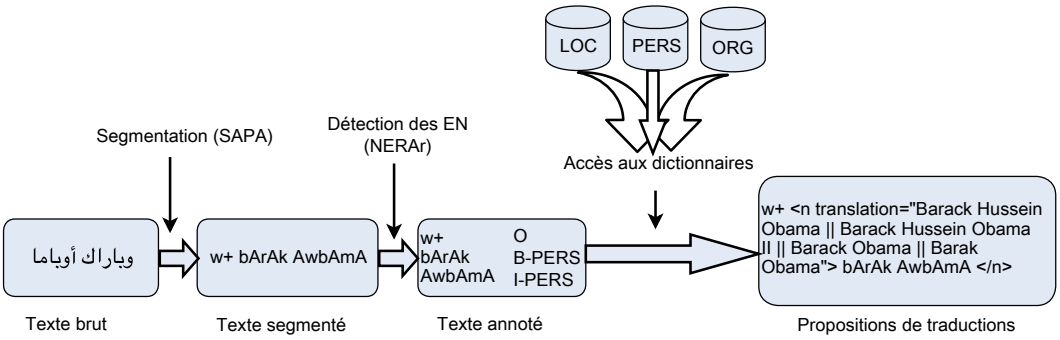

Figure 5.5: *Méthode de prétraitement du texte en arabe en combinant segmentation du texte, détection des EN et proposition de traductions.*

La figure 5.5 montre les étapes de la méthode de prétraitement des EN en arabe que nous proposons.

5.5.2 Expériences et résultats

Les expériences sont effectuées sur deux corpus différents : le corpus de test AFP (section 5.3.2), et le corpus Arcade II annoté en EN à la fois en arabe et en français. Le corpus Arcade permet de vérifier plus précisément l'impact du traitement spécifique des EN sur leur traduction automatique.

Ainsi, nous avons évalué comparativement trois situations : sans traitement a priori des EN (que nous avons appelé par la suite dans les expériences *default*), ou en proposant des traductions soit en mode *exclusive* ou en mode *inclusive*. Dans les trois cas, un deuxième poids est utilisé pour le modèle de langue permettant de pénaliser les mots hors vocabulaire.

Des évaluations en utilisant les métriques BLEU (Papineni et al., 2002) et METEOR (Banerjee et Lavie, 2005; Lavie et Agarwal, 2007) ont été réalisées. METEOR intègre des

équivalents sémantiques pour les mots et évalue les hypothèses de traduction automatique en les alignant à une ou plusieurs traductions de référence. Les alignements sont basés sur des exactitudes, des lemmes, des synonymes, et des paraphrases. Les scores sont calculés sur la base des alignements entre les paires hypothèse et référence.

Pour nos expériences, nous avons utilisé les modules d'exactitude et de paraphrases. Une table de paraphrases d'EN est constituée à partir des dictionnaires bilingues et contient toutes les paraphrases de toutes les EN en français. La table de paraphrases permet à METEOR de considérer un ensemble de paraphrases (un nom propre écrit avec différentes translittérations par exemple) correct lors de son évaluation par rapport à la référence. La table de paraphrases permet également d'inclure des synonymes ou aussi des expressions synonymes, puisque le module synonyme inclus dans METEOR est valable que pour l'anglais.

Des évaluations manuelles ainsi que des analyses détaillées des résultats ont été également effectuées.

5.5.2.1 Tests sur le corpus AFP

Le tableau 5.14 présente les résultats de traduction automatique du corpus AFP pour les trois cas *default*, *exclusive* et *inclusive* avec deux mesures de traduction BLEU et METEOR (qui intègre des équivalents sémantiques des EN). Le nombre de mots hors vocabulaire (#mots OOV) est également présenté.

	#mots OOV	BLEU	METEOR
default	288	34,62	52,64
exclusive	285	33,58	51,65
inclusive	285	34,21	52,39

Tableau 5.14: *Scores BLEU et METEOR en traduction arabe-français sur le corpus de test AFP de 1 000 phrases extraites de dépêches de décembre 2010.*

Une évaluation globale par les deux mesures BLEU et METEOR ne permet pas d'observer des améliorations en ajoutant des traductions d'EN (*inclusive* ou *exclusive*).

Les mesures BLEU et METEOR ne donnant pas de détails concernant la traduction des EN, nous avons vérifié manuellement les 100 premières lignes du corpus de test.

Évaluation manuelle :

Sur les 100 lignes analysées, 165 EN ont été détectées, parmi lesquelles 118 ont été trouvées dans les dictionnaires et ont pu être pré-traduites. Si l'on compare les systèmes *default* et *exclusive*, on observe des différences pour seulement 11 entités nommées : 8 pour lesquelles la traduction diffère de celle proposée dans la référence, tout en restant acceptable, voire corrigeant des traductions erronées de la référence. Par exemple, بيلاروسيا a été traduit par *Bélarus* (par le système *default*) qui est la traduction existante dans la référence, alors que l'option *exclusive* propose *Biélorusse*, qui est plus correcte en français. On note également qu'en utilisant cette option, 3 traductions d'EN sont améliorées (erronées dans le cas *default*, mais correctes avec notre approche).

La comparaison de la sortie de traduction *default* avec la sortie *inclusive* montre qu'en utilisant l'option *inclusive*, les différences sont encore plus réduites, les hypothèses du système *default* continuant à être préférées sauf dans trois cas : pour deux ces propositions sont correctes mais ne sont pas dans la référence et dans un cas la traduction a été améliorée.

Au final, nous observons peu de différences entre les différents systèmes; les propositions des dictionnaires sont dans l'ensemble correctes, mais trop rares pour influer positivement sur le score BLEU. À l'inverse, forcer la segmentation de la phrase source, voire imposer l'insertion de certains mots inconnus du modèle de langue dans la cible peut avoir un effet négatif sur la traduction du voisinage de ce mot. Ainsi, le système *default* propose la traduction *Séoul a annoncé que ces manoeuvres ...* alors que la meilleure hypothèse du système *exclusive* est *Séoul a indiqué que ces manoeuvres ...*, qui est induite par une autre segmentation. Le mot *indiqué* est dans cet exemple pénalisé car ne faisant pas partie de la référence.

Dictionnaires : La qualité de la traduction des EN est dépendante de leur détection. L'évolution du nombre d'EN détectées varie en fonction de la taille des dictionnaires. Ceci est reflété par les résultats représentés sur la figure 5.6, qui montre l'évolution du nombre d'EN détectées en fonction de la taille des dictionnaires. Nous avons initialement utilisé des

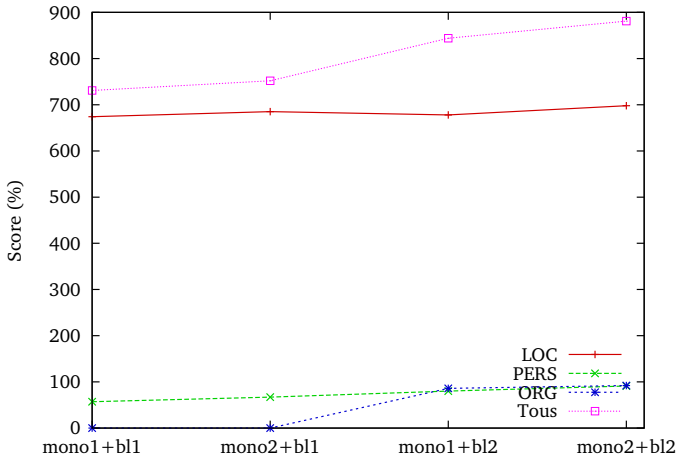

Figure 5.6: *Le nombre d'entités nommées trouvées dans les dictionnaires bilingues en fonction de l'évolution de la taille des dictionnaires.*

dictionnaires monolingues et bilingues de taille réduite, qui en particulier ne contiennent aucun nom d'organisation, et correspondent aux points situés à gauche sur la figure 5.6. Ces dictionnaires sont ensuite enrichis : les dictionnaires monolingues par 1143 lieux, 4466 personnes et 726 organisations supplémentaires; les dictionnaires bilingues par 1586 paires de lieux, 3129 paires de personnes et 726 paires d'organisations, correspondant aux deux points intermédiaires. Les points les plus à droite correspondent à l'ajout simultané des deux dictionnaires.

Les dictionnaires comme corpus : Nous avons enfin effectué une dernière expérience afin de savoir si l'intégration des dictionnaires en tant que modèles de traduction pouvait améliorer les scores de traduction. Nous avons effectué les tests, soit (i) en concaténant les données des dictionnaires avec les données du système AFP et en construisant un seul modèle (1 table) de traduction, soit (ii) en utilisant un deuxième modèle appris à partir des seules données contenues dans les dictionnaires (2 tables). Les scores BLEU obtenus sont donnés dans le tableau 5.15.

Cette dernière manière de procéder s'avère donc ici paradoxalement la meilleure et améliore de 0,3 points BLEU les performances du système de base. Intégrer les dictionnaires

	#mots OOV	BLEU	METEOR
1 table de traduction	272	34,97	52,99
2 tables de traduction	277	34,28	52,14

Tableau 5.15: *Scores BLEU et METEOR sur le corpus de test AFP en utilisant les dictionnaires sous forme de modèles de traduction.*

directement au sein des données parallèles permet au système d'optimiser leurs scores relatifs par rapport aux autres entrées du modèle de traduction, tout en assurant, comme précédemment, une meilleure couverture pour les EN.

5.5.2.2 Tests sur le corpus Arcade II

Afin d'évaluer au mieux l'impact de la détection des EN sur la traduction, nous avons utilisé le corpus Arcade II qui est annoté en EN en source et en cible. Environ 60 % des phrases du corpus Arcade II (dans les deux langues) contiennent des EN distribuées comme représenté sur la figure 5.7. L'annotation en source et en cible en EN du corpus Arcade II permet

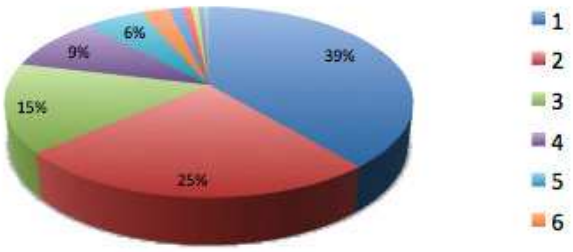

Figure 5.7: *Pourcentage des phrases selon le nombre d'entités nommées par phrase.*

d'évaluer automatiquement la qualité de la traduction des EN.

Le tableau 5.16 donne une comparaison du score BLEU des différentes variantes de nos systèmes de traduction. La détection des EN est effectuée avec le modèle adapté ANER+AFP. Une évaluation a été également réalisée sur le corpus de test en proposant les traductions correctes des EN afin de mesurer le maximum d'amélioration atteignable si l'on pouvait trouver toutes les EN de la référence dans les dictionnaires (Oracle). Un tel test nécessite d'avoir un corpus annoté en EN.

	#mots OOV	BLEU	METEOR
default	2634	20,87	40,17
exclusive	2604	20,52	39,81
inclusive	2604	20,83	40,22
Oracle			
exclusive	2493	20,60	39,24
inclusive	2493	21,17	40,13

Tableau 5.16: *Scores BLEU et METEOR en traduction arabe-français sur le corpus de test Arcade II avec et sans proposition de traductions, en comparaison avec l'Oracle.*

Comme précédemment, le tableau 5.16 ne montre pas d'amélioration du score BLEU lorsque l'on utilise les options *inclusive* et *exclusive*, bien que le pourcentage des mots hors vocabulaire (OOV) soit réduit de 1,13 %. L'option *exclusive* n'améliore pas les résultats de traduction (*default*), même dans la situation idéale où toutes les EN proposées sont correctes (configuration Oracle). Nous nous limitons dans les prochaines expériences à tester seulement l'option *inclusive*.

Pour ce corpus, l'approche qui consiste à utiliser les dictionnaires comme données parallèles s'avère moins positive pour le score de traduction. Les scores BLEU obtenus sont donnés dans le tableau 5.17.

	#mots OOV	BLEU	METEOR
1 table de traduction	2536	20,21	40,18
2 tables de traduction	2552	20,27	39,19

Tableau 5.17: *Scores BLEU et METEOR en traduction arabe-français sur le corpus de test Arcade II en utilisant les dictionnaires comme corpus d'entraînement des modèles de traduction.*

5.5.2.3 Analyse des résultats

Cette section est consacrée à l'analyse des résultats obtenus. Cette analyse s'effectue à trois niveaux (i) comparaison des deux types de corpus de test, (ii) couverture et pourcentage des EN existantes dans les dictionnaires et (iii) statistiques sur les EN traduites. Ces deux derniers niveaux d'analyse sont réalisés grâce au corpus Arcade II permettant d'analyser plus en détails les EN.

AFP vs Arcade II : Une observation troublante est la grande différence entre les scores BLEU obtenus sur le corpus AFP et le corpus Arcade (14 points de différence). Nous avons donc procédé à une étude comparative de ces deux corpus qui montre que le style des données Arcade - issues du journal 'le Monde Diplomatique' - est sensiblement différent de celui des dépêches de l'AFP. Ainsi, par exemple, le corpus de test AFP contient 43,2 % de phrases verbales (c-à-d qui débutent par un syntagme verbal), alors que le corpus de test Arcade contient seulement 9,6 % de telles phrases.

La figure 5.8 présente deux phrases en arabe issues respectivement du corpus Arcade II et du corpus AFP avec leurs annotations morphosyntaxiques – extraites automatiquement avec SAPA – et leurs traductions en français. Dans cet exemple, la phrase extraite du corpus Arcade est une phrase nominale alors que la phrase extraite du corpus AFP est une phrase verbale qui commence par une conjonction suivie d'un verbe. Ces deux derniers sont agglutinés dans la forme brute (avant segmentation).

Il existe également un décalage temporel entre les deux corpus, Arcade contient des données collectées entre 2001 et 2003 alors que les données AFP (de test) sont de 2010. Le système de traduction est constitué de données issues de dépêches datant de 2009 à 2012. Toutes ces variations entre les deux corpus expliquent la baisse du score de traduction (BLEU) sur le corpus de test Arcade.

Couverture des EN par les dictionnaires bilingues : Afin de savoir si les connaissances ajoutées au décodeur sont intéressantes, nous avons utilisé le corpus Arcade. Nous avons calculé le pourcentage des EN arabes pour lesquelles notre système propose des traductions : 16,4 % ont été retrouvées dans les dictionnaires bilingues (24,2 % des lieux, 20 % des noms de personnes et seulement 5 % des organisations).

Cette faible couverture s'explique par deux raisons : d'une part, le corpus Arcade contient des noms d'entités très spécifiques, comme par exemple *Barrio Alicia Pietri de Caldera* (un nom de quartier baptisé en hommage à la femme du précédent président du venezuela), ou

Arcade :	وبعد قليل من عودة السيد كيلي الي واشنطن اعلن مسوؤل اميركي امام الصحفيين ان اتفاق العام ١٩٩٤ الاساسي حول وقف العمل في مفاعل يونغ اصبح كانه لم يكن.
POS :	conj prep noun prep noun noun noun_prop prep noun_prop verb noun adj prep noun conj_sub noun adj noun_num adj prep noun noun prep noun noun_prop noun_prop verb prep conj_sub part_neg verb punc
Réf. :	peu de temps après le retour de M. Kelly à Washington , un responsable américain déclarait à des journalistes que l'accord-cadre de 1994 sur le gel du réacteur de Yongbyon était nul et non avenu .
AFP :	وقال برنار فاليرو المتحدث باسم الخارجية الفرنسية باشرنا عملية متواسلة لضمان الامن حفاضا علي سرية الوثائق.
POS :	conj verb noun_prop noun_prop adj prep noun noun adj punc verb noun adj prep noun noun noun prep noun noun punc punc
Réf. :	" afin de préserver la confidentialité des documents , nous sommes engagés dans un processus permanent de sécurisation " , déclare Bernard Valero , le porte-parole du ministère .

Figure 5.8: *Comparaison de la structure de deux phrases issues de deux corpus différents : Arcade II et AFP*

encore *Organisation Libérale* (en Tunisie) et *Tribunal pénal international pour l' ex-Yougoslavie de La Haye*.

D'autre part, un certain nombre de noms de lieux qui ne sont pas reconnus existent dans nos dictionnaires mais sous une autre forme: c'est le cas de *territoires palestiniens* qui existe en tant que *Palestine* ou encore de *zones kurdes* qui existe en tant que *Kurdistan*, voire qui ont une orthographe différente de celle des dictionnaires : c'est le cas par exemple pour *l'Angleterre*, انكلترا au lieu de انجلترا ou pour *les Philippines*, الفيليبين au lieu de الفلبين qui n'ont pas la même orthographe dans les dictionnaires et dans le corpus. Ces noms sont à l'origine des noms qui ont été translittérés vers l'arabe, donc on peut en trouver plusieurs translittérations différentes mais qui sont correctes.

Parmi les propositions de traductions, 84,2 % existent dans la référence en langue cible et se répartissent comme suit : 75,6 % de lieux – pour lesquelles il y a des propositions de traductions –, 81 % de noms de personnes et 96,1 % d'organisations. Donc au maximum 13,8 % (16,4 % x 84,2 %) des ENs peuvent être améliorées avec les dictionnaires existants.

Statistiques sur les EN traduites : Le taux d'EN traduites correctement a été calculé pour chaque type sur le corpus Arcade en comparant les EN traduites aux EN dans la référence

(français). Les résultats sont dans le tableau 5.18.

	LOC	PERS	ORG
default	70,4 %	51,2 %	44,6 %
inclusive	70,7 %	52,8 %	44,6 %
Oracle			
inclusive	70,9 %	53,0 %	45,2 %
Modèles de traduction			
1 table de traduction	72,0 %	57,3 %	44,4 %
2 tables de traduction	66,7 %	54,9 %	45,8 %

Tableau 5.18: *Pourcentage des EN traduites correctement en utilisant notre approche de proposition de traductions en comparaison avec l'Oracle et avec l'utilisation des dictionnaires sous forme de modèles de traduction.*

D'après le tableau 5.18, on note qu'en utilisant l'option *inclusive*, la traduction de tous les types d'EN a été améliorée (jusqu'à 1,52 %).

Ce résultat est à comparer du score qui serait atteint en traduisant correctement toutes les EN *détectées et qui existent dans les dictionnaires* (ligne Oracle) : les prétraductions proposées par le dictionnaire sont donc souvent justes. Les scores de l'Oracle pourraient être améliorés si les dictionnaires bilingues utilisés pour proposer des traductions étaient plus riches et contenaient plus de paires d'EN.

L'utilisation des dictionnaires sous forme de modèles de traduction est encore meilleure et réduit jusqu'à 2,5 % le taux d'EN inconnues lors de la traduction. Ces résultats quantitatifs montrent que l'utilisation de dictionnaires améliore effectivement la traduction des EN, même si cela ne se traduit pas toujours par une augmentation des métriques de traduction automatique.

5.6 Conclusion

Les approches statistiques sont plus robustes que les approches à base de règles et permettent l'utilisation de centaines de milliers de descripteurs. Notre approche pour la détection des EN, NERAr[19], était donc fondée sur les CRF. Les corpus annotés manuellement en EN sont des ressources très rares et surtout pour la langue arabe. Le corpus ANER construit par Benajiba et ses co-auteurs était notre seul corpus de référence permettant de se comparer avec d'autres travaux sur la détection des EN en arabe. NERAr ainsi que la chaîne complète de prétraitement et de pré-traduction est téléchargeable sur Internet[19].

Dans ces travaux nous nous sommes intéressés à la réduction du taux d'erreur des mots hors vocabulaire et à l'amélioration de la traduction des Entités Nommées (EN). Les EN représentent entre 25 à 40 % des mots inconnus. Nous nous sommes donc intéressés au traitement automatique des EN afin de réduire les erreurs de traduction de ces entités. Nous avons développé un outil de détection des EN en arabe (NERAr) construit par des méthodes d'apprentissage supervisé. Ce système, qui embarque des centaines de milliers de descripteurs, obtient des performances comparables aux meilleurs systèmes de l'état de l'art. Nous avons ensuite adapté ce système par auto-apprentissage et par hybridation conduisant à une légère amélioration des performances.

Notre deuxième contribution consiste à proposer une méthode de traduction des EN en arabe à l'aide de dictionnaires bilingues. Nous avons montré que la traduction préalable des EN améliore légèrement les résultats de la traduction automatique (BLEU et METEOR). Le

[19]NERAr est téléchargeable à partir de https://github.com/SouhirG/NERAr

nombre de mots hors vocabulaires a été réduit de 1,65 %. Une évaluation manuelle ainsi qu'une évaluation plus détaillée des EN obtenues en sortie de traduction montre que la traduction automatique des EN a été améliorée pour les trois types d'EN : personnes, lieux et organisations.

Une analyse des résultats à 3 niveaux a été réalisée : tout d'abord une comparaison des styles des deux corpus de test, ensuite des statistiques sur la couverture des EN par les dictionnaires bilingues (afin de voir si les données ajoutées au décodeur sont intéressantes), et finalement une étude sur l'effet de la traduction des EN. Les résultats obtenus sur le corpus Arcade II ainsi que l'analyse détaillée des résultats montrent que notre approche améliore la traduction des EN, mais peut également affecter la traduction.

Le corpus Arcade II a permis une évaluation plus fine et plus détaillée de la traduction des EN, c'est le seul moyen de vérifier leur traduction. Ce corpus nous a permis également de mesurer le maximum d'amélioration atteignable (Oracle) en utilisant les EN annotées. En utilisant des dictionnaires bilingues d'EN, nous avons pu réduire le taux de mots hors vocabulaire. Ces résultats ne sont pas visibles sur les scores de traduction.

Comme perspective, nous envisageons (i) d'améliorer notre méthode de traduction des EN en rajoutant des translittérations pour les noms propres détectés par NERAr mais qui n'existent pas dans les dictionnaires, et (ii) d'enrichir nos dictionnaires en explorant des corpus parallèles ou aussi des corpus comparables. La relative rareté de corpus parallèles (arabe-français) annotés en EN constitue un obstacle pour l'évaluation. Nous étions limités à effectuer nos expériences sur le corpus Arcade II dont le style est différent des styles de corpus que nous utilisons. D'un autre côté l'annotation manuelle est une approche très coûteuse en temps. À moyen terme nous envisageons d'exploiter des méthodes d'alignement au niveau des mots pour projeter des annotations depuis le français, langue dans laquelle les entités sont plus faciles à détecter, vers l'arabe, à l'instar du travail de Zhang (2012) pour le chinois: ceci nous permettrait d'avoir des corpus d'adaptation au moins partiellement annotés en EN.

Il serait également intéressant de détecter les EN sur un corpus comparable dans les deux langues chaque semaine (si on utilise le corpus comparable de l'AFP par exemple). Ensuite faire correspondre les EN en utilisant des techniques de mise en correspondance à l'aide de translittérations par exemple. Une autre perspective consiste à utiliser une liste des nbests afin d'améliorer la traduction. L'idée consiste à traduire le texte de l'arabe vers le français, ensuite détecter les EN en français et insérer les différentes valeurs des EN en français dans des listes nbest pour trouver les meilleures traductions.

CHAPITRE 6

Adaptation thématique des systèmes de traduction

Dans le langage quotidien, on s'exprime de façon différente si on s'adresse à un familier ou à un inconnu, à un enfant ou à un supérieur hiérarchique mais aussi selon l'âge de la personne à qui on s'adresse, son milieu social, son niveau culturel. L'être humain adapte spontanément son langage naturel selon la situation dans laquelle il est et selon la (ou les) personne(s) avec qui il interagit.

Dans le traitement automatique des langues, de nombreux travaux s'intéressent à l'adaptation des modèles de traitement automatique des langues. Il existe plusieurs types d'adaptation, comme l'adaptation des modèles acoustiques au genre des locuteurs (homme/femme) qui suscite des recherches dans le domaine de la reconnaissance de la parole comme par exemple ceux de Ferràs et al. (2007). D'autres travaux s'intéressent à l'adaptation de registre de langue – c'est-à-dire à l'utilisation sélective et cohérente des procédés d'une langue afin d'adapter l'expression à un auditoire particulier – comme ceux de Banerjee (2012) qui adapte des modèles de traduction au domaine des forums[1] de discussion. La langue manipulée dans les forums n'est pas conforme aux normes de langue écrite, elle est "déviante" par rapport à cette dernière. De plus, il est très difficile de trouver un corpus parallèle extrait de forums puisque ce sont généralement des données qui existent dans une langue et ne sont pas en général traduits. Dans la plupart des cas, des systèmes de traduction construits à partir de données *hors-domaine* sont utilisées pour traduire. Par exemple pour traduire un texte issu d'un fichier audio, généralement des systèmes de traduction construits à partir de corpus écrits sont utilisés puisque ces derniers sont plus fréquents que les corpus audio. De même des corpus issus de forums sont traduites avec des systèmes de traduction construits à partir de corpus écrits et donc non adaptés. Ces systèmes de traduction non adaptés introduisent des ambiguïtés lors de la traduction surtout pour les mots ayant plusieurs sens.

Dans le domaine de la traduction automatique, ce problème de polysémie est très connu et c'est un problème de recherche toujours d'actualité. Si le système de traduction utilisé n'est pas approprié au contexte du texte à traduire, alors il est très probable qu'un mot polysémique ne sera pas traduit correctement. Dans une même langue, un terme peut être ambigu selon les contextes dans lesquels ce terme apparaît. Dans la phrase *Il a un grand bureau* par exemple, le mot bureau signifie le meuble ou la pièce contenant ce meuble.

De même le mot *bourse* peut avoir plusieurs sens : d'après le *TLF*[2], le mot bourse peut avoir

[1] Les forums du site web de Symantec.
[2] Le Trésor de la Langue Française : `http://atilf.atilf.fr`

principalement cinq sens. Le contexte dans lequel le mot apparaît joue un rôle primordial pour aider à la compréhension de la phrase. Par exemple le terme *bourse* sera interprété de façons différentes s'il apparaît dans un contexte de finances (marché financier) ou d'éducation et de recherche (allocation accordée à un étudiant).

Nous avons montré dans le chapitre 3 que l'utilisation d'un modèle de traduction adapté améliore les performances de traduction même si la taille de ce modèle est très inférieure au modèle *hors-domaine*. En utilisant le modèle *hors-domaine* dans le chapitre 3, le mot *discussions* a été traduit par *pourparlers*. Ceci est dû simplement au fait que le système de traduction *hors-domaine* a été entraîné principalement sur des débats politiques où le mot *pourparlers* est plus fréquent que le mot *discussions*.

Les dépêches AFP sont des fichiers de format NewsML (voir annexes 7.6 et 7.6) qui est un format standard basé sur XML créé par l'organisme international de codification IPTC[3] (*International Press Telecommunications Council*). Ces dépêches contiennent des mots clés thématiques décrivant les éléments importants de l'information, ainsi que les éléments de la nomenclature proposée par l'IPTC. Les dépêches de l'AFP couvrent différentes thématiques notamment la politique, l'économie, le sport, etc. Les dépêches AFP sont classifiées en 17 catégories principales selon les 17 catégories de l'IPTC présentées dans le tableau 6.1.

Code de la catégorie	Code IPTC	Nom de la catégorie
ACE	01000000	arts, culture and entertainment
CLJ	02000000	crime, law and justice
DIS	03000000	disaster and accident
FIN	04000000	economy, business and finance
EDU	05000000	education
EVN	06000000	environmental issue
HTH	07000000	health
HUM	08000000	human interest
LAB	09000000	labour
LIF	10000000	lifestyle and leisure
POL	11000000	politics
REL	12000000	religion and belief
SCI	13000000	science and technology
SOI	14000000	social issue
SPO	15000000	sport
WAR	16000000	unrest, conflicts and war
WEA	17000000	weather

Tableau 6.1: *Les 17 catégories de l'IPTC*

Ces différentes thématiques et catégories peuvent être exploitées pour adapter et améliorer les systèmes de traduction automatique. Nous allons présenter dans ce chapitre nos différentes propositions d'adaptation de modèles de traduction et de modèles de langues. La particularité de nos données est qu'elles sont extraites d'une application réelle utilisée par l'AFP et dans laquelle nous avons plusieurs thématiques.

La section 6.1 présente un état de l'art sur des travaux effectués sur l'adaptation. Les motivations à l'origine de nos travaux sont décrites dans la section 6.2. La section 6.4 présente les trois scénarios proposés pour traduire un corpus de test multicatégorique. Cette section présente également les deux approches de classification des données de l'AFP. Les expériences que nous avons effectuées ainsi que les résultats sont présentés dans la section 6.5.

[3] www.iptc.org

6.1 État de l'art

En apprentissage automatique, l'adaptation consiste à développer un système de traitement pour un domaine cible à partir de données et/ou d'un système de traitement développé pour un domaine source.

On peut situer les premiers travaux sur l'adaptation de modèle de langue vers la fin des années quatre-vingt-dix, notamment dans le cadre d'applications de reconnaissance de la parole (De Mori et Federico, 1999). Parmi les premiers travaux dans le domaine de la traduction automatique ceux de Langlais (2002) qui implémente une stratégie pour intégrer des lexiques adaptés dans le modèle de traduction.

Snover, Dorr et Schwartz (2008) utilisent des documents comparables pour adapter le système de traduction automatique (modèle de langue et modèle de traduction) afin d'augmenter les probabilités de générer des textes qui ressemblent aux documents comparables.

Schwenk et Senellart (2009) et Lambert et al. (2011) traduisent des données monolingues *du domaine* en utilisant un système de traduction *hors-domaine* pour construire automatiquement un corpus parallèle artificiel du domaine et adapter des modèles de traduction. Zhao, Eck et Vogel (2004) explorent des techniques d'adaptation non-supervisée du modèle de langue. Des modèles de langue sont construits à partir de données monolingues extraites selon des scores de similarité et des caractéristiques sémantiques. Ces modèles sont ensuite interpolés avec le modèle de langue général.

Hildebrand et al. (2005) sélectionnent les phrases similaires à l'ensemble de test en utilisant des techniques de recherche d'information (IR - *Information Retrieval*) pour construire un corpus d'entraînement adapté.

Différentes méthodes d'interpolation de modèles du domaine et hors domaine sont réalisées par Koehn et Schroeder (2007). Yamamoto et Sumita (2007); Yamamoto et Sumita (2008) proposent une approche utilisant des modèles spécifiques pour la traduction automatique : pour chaque phrase à traduire, le domaine est détecté et les modèles spécifiques à ce domaine sont utilisés pour traduire cette phrase.

Des méthodes d'interpolation log-linéaire sont proposées par Nakov (2008) qui effectue une adaptation en combinant trois modèles de traduction de différents domaines et en ajoutant un paramètre pour chaque modèle. Ce paramètre indique pour chaque segment s'il existe ou pas dans le modèle de traduction concerné. Les modèles de réordonnancement sont combinés de la même façon.

Haque et al. (2009) étudient différentes méthodes d'adaptation sur des systèmes de traduction de l'anglais vers l'Hindi : adaptation de modèles de langue, adaptation de modèles de traduction soit en utilisant plusieurs modèles de traduction ou en utilisant un marqueur d'appartenance au thème comme dans Nakov (2008). L'adaptation par classification des phrases à traduire par domaine, comme dans Yamamoto et Sumita (2008), a été aussi effectuée. Niehues et Waibel (2010) utilisent un identifiant pour chaque corpus pour distinguer entre les données du domaine et les données *hors domaine* dans un modèle de traduction factorisé. Trois paramètres supplémentaires sont ajoutés pour modéliser la probabilité du corpus auquel appartient le segment.

Foster, Goutte et Kuhn (2010) décrivent une approche d'adaptation dans laquelle les paires de phrases hors domaine ont des poids selon leur pertinence par rapport au domaine cible, et ceci en déterminant leur degré de similarité et leur appartenance ou non au modèle général.

Différentes combinaisons d'adaptation de modèles de langue et de modèles de traduction sont effectuées par Ceausu et al. (2011) qui expérimentent différents types d'adaptation sur des données réparties en 8 domaines différents. Les meilleurs résultats sont obtenus en utilisant un modèle de traduction *du domaine* et un modèle de langue général.

Bisazza et al. (2011) proposent d'améliorer un système de traduction entraîné sur des données *du domaine* (corpus TED : parole transcrite) en ajoutant des données supplémentaires (appelées *fill-up*) *hors-domaine* (données ONU et Europarl). Les résultats montrent que cette

technique donne les mêmes résultats que l'utilisation de deux modèles de traduction, et de meilleurs résultats par rapport à la concaténation de deux modèles de traduction (*du domaine* et *hors-domaine*).

Des méthodes d'extraction de données du même domaine que le corpus de test en utilisant la différence de l'entropie croisée (*cross entropy difference*) et des combinaisons de modèles sont proposées par Axelrod, He et Gao (2011). Ces méthodes sont moins performantes que l'utilisation de plusieurs tables de traduction.

Différentes méthodes d'adaptation en effectuant une interpolation de modèles de traduction ont été réalisés par Mansour et Ney (2012). Des modèles de mélange (*mixture modeling*) ont été également effectués par Foster et Kuhn (2007) et par Sennrich (2012a) ou encore par Lavergne et al. (2011).

Récemment, Sennrich (2012b) enrichit les travaux proposés par Foster, Goutte et Kuhn (2010) en minimisant la perplexité des scores du modèle de traduction. Il propose également d'adaptater les modèles de traduction en optimisant des scores pondérés. Il montre que l'ajout de données *hors-domaine* dégrade les performances de traduction en introduisant des ambiguïtés lexicales. Dans d'autres travaux, Sennrich (2012a) montre qu'il est possible d'appliquer des méthodes de classification non supervisée pour construire des modèles adaptés. Chaque phrase dans le corpus de test est traduite par le modèle approprié.

Dans ce qui suit, nous proposons une approche de classification par phrase afin de traduire chaque phrase par le modèle approprié. Notre approche est assez proche de celles de Sennrich (2012a) et de Yamamoto et Sumita (2008). Nous comparons trois scénarios de traduction : traduction "normale" sans aucune classification, traduction avec une classification a priori des phrases et traduction avec une classification automatique des phrases.

6.2 Motivations

Dans la plupart des travaux existants sur l'adaptation, les données adaptées et spécifiques à des domaines différents sont issues de sources bien séparées. Dans les travaux de Banerjee (2012) par exemple, des données de deux domaines chez *Symantec* ont été utilisés : la validité – *availability* – (sauvegarde, récupération des données) et la sécurité (protection contre les attaques, vulnérabilité de logiciels malveillants). De même, parmi les données utilisées par Sennrich (2012b), il y a par exemple des données extraites du journal *Alpine Club* (dédié à l'alpinisme et aux alpinistes) et des données *Europarl* (débats politiques). Toutes ces données proviennent de sources bien distinctes.

La difficulté de notre approche est que nous nous intéressons à un cas où les données proviennent toutes de la même source et dont les frontières thématiques sont floues. Les catégories que nous manipulons sont donc très proches les unes des autres et certaines phrases peuvent être étiquetées par plusieurs catégories, ce qui n'est pas le cas des autres travaux. Dans notre cas, nous traitons un cas réel d'une application qui est celle de l'AFP, qui contient des dépêches qui se répartissent en 17 catégories différentes, et qui peuvent en même temps être multicatégoriques.

Notre objectif est d'étudier l'impact de l'utilisation de modèles spécifiques (de traduction et de langue) construits à partir d'une classification selon les catégories IPTC. Il est possible de savoir pour chaque phrase du corpus de test, sa catégorie IPTC (extraite de la dépêche à partir de laquelle la phrase est extraite). Il est donc possible de traduire chaque phrase par le système approprié à la catégorie concernée.

Certaines catégories IPTC étant très proches comme par exemple POL et WAR. Nous proposons cette méthode de classification afin de tenter d'optimiser la traduction et d'éviter les ambiguïtés pour les mots polysémiques. Prenons l'exemple du verbe سجل (enregistre) en arabe qui peut apparaître dans des dépêches extraites de différentes catégories. La phrase سجل فرحان شكور هدف العراق (*Farhan Chakour* ***a marqué*** *le but de l'Irak ...*) par exemple est

extraite d'une dépêche de type SPO (sport), alors que la phrase سجل سنودن ومساعدته اسميهما علي رحلة ايرفلوت (*Snowden et son assistante **ont enregistré** leurs noms sur le vol Airfloat ...*) est extraite d'une dépêche de type POL. Le mot سجل est traduit soit par *marquer* ou *enregistrer*. Ceci dépend du contexte de la dépêche et du texte.

6.3 Les données AFP

Le corpus parallèle utilisé dans ce chapitre est extrait automatiquement avec la méthode d'extraction de corpus parallèle présentée dans le chapitre 3. Les données utilisées pour l'entraînement proviennent de dépêches collectées entre décembre 2009 et juillet 2012 et constituent un corpus de 265K paires de phrases. Les corpus de développement et de test sont extraits du mois de novembre 2011 et sont constitués respectivement de 1 178 et 1 000 paires de phrases.

Les données de base que nous allons utiliser pour les expériences sont classifiées par l'AFP. Le tableau 6.2 montre le pourcentage des phrases AFP appartenant à chaque catégorie IPTC pour les corpus d'entraînement, de test et de développement. Le calcul des pourcentages est effectué en se basant sur le nombre de phrases pour chaque type de corpus (entraînement 265K phrases, développement 1 178 phrases et test 1 000 phrases). Une phrase peut appartenir à une ou plusieurs catégories, raison pour laquelle la somme de tous les pourcentages n'est pas égale à 100 %.

On note que le pourcentage n'est pas le même pour l'arabe et le français. Ceci est dû au fait de l'inexactitude et l'infidélité de la traduction manuelle des dépêches AFP (voir figure 3.3, chapitre 3). La répartition des dépêches est donc basée sur la classification des phrases en arabe. On remarque également que certaines dépêches ne sont pas catégorisées. Ceci peut être dû à des oublis de la part des journalistes.

	Entraînement		test		dev	
Catégorie	%AR	%FR	%AR	%FR	%AR	%FR
ACE	1,1	2,7	0,6	1,6	0,7	1,9
CLJ	9,2	12,1	9,4	11,6	8,1	10,0
DIS	5,3	6,8	1,5	2,6	2,1	2,9
FIN	**14,8**	**15,3**	**20,7**	**23,3**	**19,7**	**21,2**
EDU	0,2	0,2	0	0	0,2	0,2
EVN	0,8	1,3	0,1	0,5	0,1	0,7
HTH	0,9	1,0	0	0,3	0,2	0,5
HUM	0,4	0,7	0,1	1,2	1,3	1,0
LAB	0,8	2,0	0,7	2	0,3	1,0
LIF	0,03	0,03	0	0	0	0
POL	**58,9**	**62,3**	**66,6**	**74**	**69,9**	**77,2**
REL	3,4	3,9	2,9	3,3	2,8	3,5
SCI	1,0	1,3	0,7	0,1	0,4	0,4
SOI	2,0	2,6	0,8	1,7	0,4	1,4
SPO	1,3	1,3	0,8	0,8	0,7	0,7
WAR	**38,3**	**42,9**	**35**	**39,3**	**34,8**	**39,3**
WEA	0,9	0,1	0,2	0	0,17	0
-	0,1	2,3	-	-	-	-

Tableau 6.2: *Distribution des catégories IPTC en arabe et en français pour le corpus d'entraînement de 265K phrases et les corpus de développement et de test de novembre 2011.*

Le tableau montre que la répartition des données est très inégale et que la catégorie dominante (sur les trois types de corpus) est la catégorie POL avec environ 60 % de phrases appartenant à cette catégorie. Les catégories que nous allons utiliser pour nos expériences sont celles qui apparaissent le plus dans les dépêches et sont donc *politique* (POL), *guerres* (WAR), et *finances* (FIN).

Comme déjà mentionné, une phrase peut appartenir à une ou plusieurs catégories. Le tableau 6.3 montre le pourcentage de phrases en commun pour les catégories POL et WAR, POL et FIN ainsi que pour WAR et FIN dans le corpus d'entraînement, de test et de développement. On note que dans le corpus d'entraînement, il existe un peu plus de 16 % de phrases

		# phrases en commun AR (%)		
Catégorie 1	Catégorie 2	Entraînement	Dev	Test
POL	WAR	16,35	19,86	18,3
POL	FIN	5,97	8,65	8,7
WAR	FIN	1,36	1,86	1,6

Tableau 6.3: *Pourcentage de phrases en commun entre certaines paires de catégories dans les corpus d'entraînement, de développement et de test.*

en commun entre les catégories POL et WAR.

			Pourcentage des phrases communes dans chaque cat.	
Catégorie 1	Catégorie 2	phrases en commun (%)		
POL	WAR	16,35	27,7 % POL	42,7 % WAR
POL	FIN	5,97	12,5 % POL	40,3 % FIN
WAR	FIN	1,36	3,6 % WAR	9,2 % FIN

Tableau 6.4: *Pourcentage des phrases appartenant à la deuxième catégorie.*

Le tableau 6.4 montre le pourcentage de phrases appartenant à chaque catégorie et faisant partie des phrases communes pour chacune des paires de catégories POL-WAR, POL-FIN et WAR-FIN.

Différents scénarios d'adaptation ont été proposés. Pour chacune des catégories considérées un corpus de test et un corpus de développement spécifiques ont été créés de tailles respectives 1 000 et 1 178 phrases. Ces corpus sont créés indépendamment des corpus de test et de développement généraux (ce ne sont pas des sous corpus).

Les données en arabe sont prétraitées avec SAPA (chapitre 4) et pour la partie français, la tokenisation consiste à séparer les mots des ponctuations et à remplacer les majuscules par des minuscules, sauf pour les noms propres. Les données parallèles sont ensuite alignées avec MGiza++[4] (Gao et Vogel, 2008). Les tables de traduction sont constituées en symétrisant les alignements selon l'heuristique *grow-diag-final-and* de Moses. Les modèles de langue sont construits à l'aide de l'outil SRILM[5] (Stolcke, 2002). Les poids des paramètres sont fixés après optimisation avec MERT (Och, 2003).

6.4 Scénarios de traduction sans et avec classification

Dans ce chapitre, nous proposons trois scénarios pour traduire un corpus de test général constitué de phrases appartenant à différentes catégories. Dans le premier scénario, la

[4] http://geek.kyloo.net/software/doku.php/mgiza:overview
[5] http://www.speech.sri.com/projects/srilm/

traduction est réalisée de manière ordinaire, sans aucune classification préalable des phrases. Le deuxième et troisième scénarios sont basés sur la classification : le deuxième scénario (section 6.4.2) consiste à affecter *a priori* à chaque phrase la (les) catégorie(s) de la dépêche d'où elle est extraite, alors que dans le troisième scénario (section 6.4.3), chaque phrase d'une même dépêche peut être étiquetée par une catégorie qui lui est propre, indépendamment de la dépêche d'où elle est issue. Pour chaque catégorie, un modèle de langue et un modèle de traduction spécifiques sont créés. La figure 6.1 illustre le processus de traduction en utilisant les systèmes de traduction spécifiques pour chaque phrase du corpus de test.

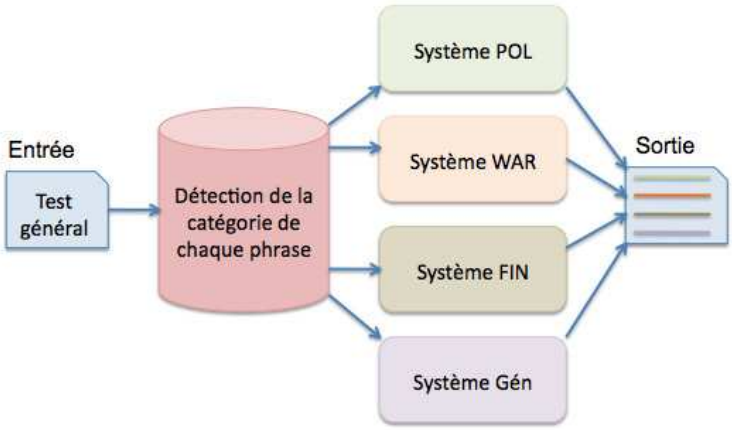

Figure 6.1: *Illustration de la traduction phrase par phrase d'un corpus de test selon la catégorie.*

6.4.1 Scénario sans classification

Dans ce scénario, la traduction automatique est réalisée sans avoir à classifier les phrases dans des catégories spécifiques. Dans ce cas, le modèle de traduction général et le modèle de langue général sont utilisés.

6.4.2 Classification de documents a priori

Dans cette approche, chaque phrase est catégorisée selon la catégorie affectée *a priori* à la dépêche d'où elle est extraite. Dans le cas où la dépêche est monocatégorique, la phrase est classée dans cette catégorie. Si une phrase appartient à plusieurs catégories, alors elle est considérée comme appartenant à toutes les catégories dont elle fait partie. Puisque le test général contient 60 % des phrases multicatégoriques (voir tableau 6.5), une vérification manuelle a été effectuée afin d'assigner une seule catégorie pour chaque phrase du corpus de test général et de pouvoir effectuer la combinaison des différents systèmes pour la traduction (voir tableau 6.6). Le pourcentage des phrases monocatégoriques et multicatégoriques – pour les corpus d'entraînement, de développement et de test – est donné dans le tableau 6.5.

On observe que la plupart des phrases sont issues de dépêches multicatégoriques. D'après le tableau 6.5, le corpus de test contient 40 % de phrases monocatégoriques dont 49,8 % sont de la catégorie POL, 26 % appartiennent à WAR et 14,8 % sont de la catégorie FIN. 43,2 % des phrases du corpus de test appartiennent à deux catégories dont les paires les plus dominantes appartiennent à la fois aux catégories POL et WAR (28,2 %), aux catégories CLJ et POL (9,3 %) ainsi qu'aux catégories FIN et POL (7,6 %).

Corpus	monocatégoriques (%)	multicatégoriques (%)			
	1 cat.	2 cat.	3 cat.	4 cat.	5 cat.
Entraînement	45,8	40,3	11,8	1,8	0,1
Développement	38,7	43,4	14,9	2,5	0,6
Test	40,0	43,2	14,1	2,2	0,5

Tableau 6.5: *Pourcentage des phrases monocatégoriques et multicatégoriques pour le corpus d'entraînement, de développement et de test*

Le tableau 6.6 montre le pourcentage des phrases pour chacune des catégories POL, WAR et FIN ainsi que pour les autres catégories après vérification manuelle du corpus de test général.

Catégorie	# phrases (%)
POL	56
WAR	21
FIN	16
Autres	13

Tableau 6.6: *Pourcentage des phrases pour chaque catégorie du test général après la vérification manuelle.*

6.4.3 Développement d'un classifieur automatique

Dans cette approche, nous partons du principe qu'il peut être opportun d'assigner à une phrase une catégorie qui soit différente de celles de la dépêche dont elle est issue et qui serait visiblement mieux dans une autre catégorie. Également, certaines phrases sont de façon évidente à classer en dehors de toute catégorie IPTC (comme par exemple des phrases situant la date de l'événement relaté par la dépêche). Nous introduisons ainsi une catégorie *générale* supplémentaire, permettant de ne pas assigner à ce type de phrase une catégorie spécifique.

La figure 6.2 montre un extrait d'une dépêche AFP assignée à la catégorie WAR.

1 PARIS, 20 mai 2011 (AFP) -
2 Voici l'agenda pour la journée du samedi 21 mai.
3 ...
4 Les violences ont fait près de 3.000 morts selon les autorités.
5 ...
6 ALGERIE
7 ALGER - Début des consultations du président du Sénat avec les partis et personnalités sur les réformes à mettre en oeuvre en Algérie, sans la présence de l'opposition.
8 ...

Figure 6.2: *Extrait d'une dépêche AFP catégorisée par l'AFP dans la catégorie WAR.*

On observe que parmi les phrases de la dépêche, certaines phrases comme la phrase numéro 7 serait visiblement mieux dans la catégorie POL. Les phrases numéro 1, 2 et 6 contiennent seulement des noms de lieux et/ou des dates, donc peuvent appartenir à n'importe quelle catégorie et pas forcément la catégorie WAR. Il serait donc plus judicieux d'affecter chaque phrase à une seule catégorie qui lui correspond.

C'est essentiellement pour cette raison que nous avons proposé une méthode de classification des dépêches par phrase.

Le classifieur de dépêches AFP (en arabe) que nous avons construit est entraîné sur les données classifiées par l'AFP, selon une classification naïve bayésienne (Lewis, 1998) avec modèles multinomiaux (Yvon, 2011). Nous avons construit quatre modèles pour les catégories les plus répandues dans notre corpus d'apprentissage POL, WAR, FIN (voir tableau 6.2) et un modèle qui englobe toutes les données (Gén). Pour la classification, on procède comme suit : la probabilité a postériori de chaque phrase par rapport aux quatre modèles est calculée. La catégorie attribuée à la phrase est celle du modèle qui a obtenu le meilleur score de probabilité a postériori.

La figure 6.3 montre le processus de classification automatique que nous avons mis en place.

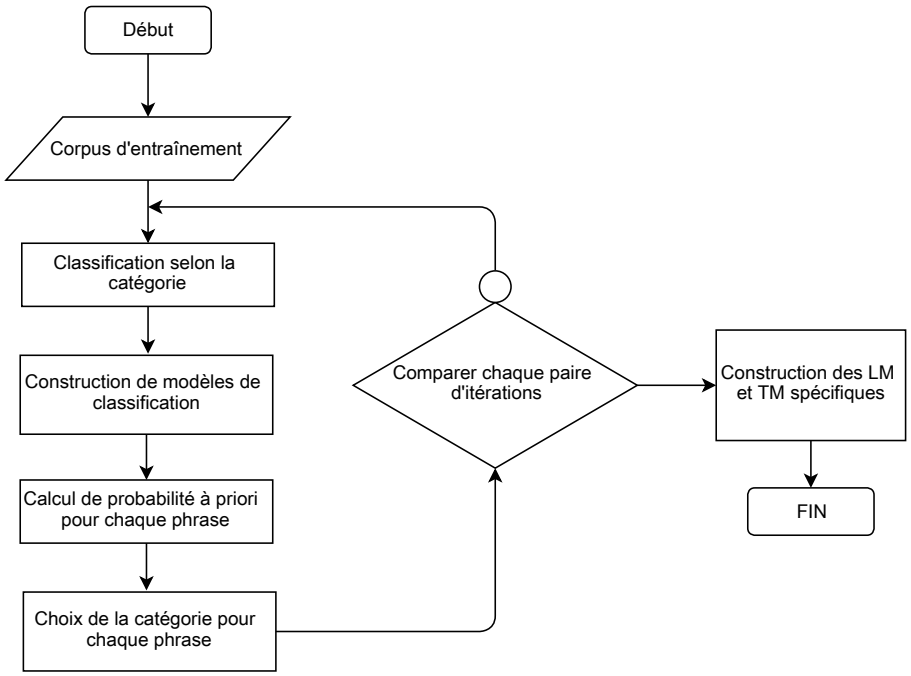

Figure 6.3: *Processus itératif de classification automatique des phrases.*

La classification automatique du corpus d'entraînement (265 milles phrases) ainsi obtenue est différente de celle effectuée par l'AFP avec une différence de 41,01 %. Les distributions sont initialisées en faisant l'hypothèse que toutes les phrases du document sont dans la même classe. Les classes/catégories assignées par notre classifieur sont utilisées pour reclassifier les phrases du corpus jusqu'à la convergence en utilisant les k-moyennes (*k-means*). Les résultats que nous avons obtenus convergent à la sixième itération avec une différence de 1,32 % entre la sixième et la cinquième itération. La courbe de la figure 6.4 montre la variation de classification entre deux itérations.

Afin d'optimiser nos résultats, nous avons utilisé quelques contraintes pour le classifieur. La première contrainte consiste à placer les phrases qui ont une longueur inférieure à dix tokens dans la catégorie générale. Cette contrainte est valable surtout pour les phrases qui contiennent le nom de ville et l'abréviation AFP, comme par exemple la phrase *rAm Allh (A f b)* - (*Ram Allah (A F P)* -) constituée de huit tokens ainsi que les phrases qui contiennent

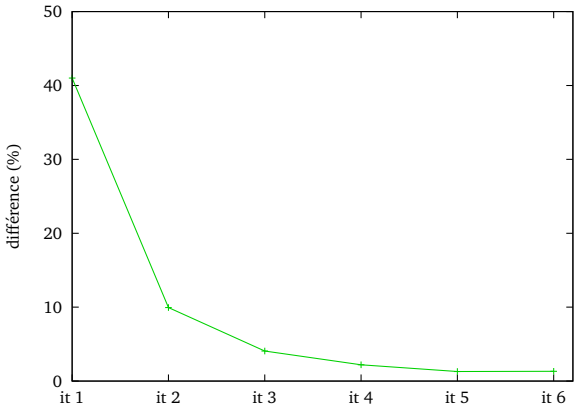

Figure 6.4: *Pourcentage de différence des catégories IPTC entre deux paires d'itérations successives.*

également le token *(A F P)*[6] -. Les phrases les plus courtes sont souvent de ce format. La longueur moyenne d'une phrase en arabe dans les dépêches AFP est de 25 mots. Un score de confiance est également utilisé pour filtrer les phrases qui ont un score normalisé (probabilité a priori/nombre de mots) très élevé. Ces phrases sont classifiées dans la catégorie Gén.

Nous avons donc utilisé la classification obtenue après 6 itérations pour construire les modèles (de traduction et de langue) spécifiques.

Une comparaison des catégories IPTC attribuées par les journalistes selon les dépêches, avec les catégories attribuées par notre classifieur a été effectuée sur le test général. Une différence de 58 % a été constatée et elle est répartie comme le montre le tableau 6.7. On

Pourcentage (%)	Cat. par l'AFP (ref)	Cat. automatique (hyp)
16,60	POL	Gén
5,40	POL	WAR
4,70	POL	FIN
4,40	WAR	POL
0,90	WAR	FIN
7,30	WAR	Gén
2,10	FIN	POL
1,20	FIN	WAR
4,90	FIN	FIN
4,80	Autres	POL
3,60	Autres	WAR
2,10	Autres	FIN

Tableau 6.7: *Comparaison entre les catégories IPTC assignées par l'AFP et les catégories IPTC assignées par notre classifieur automatique. Il est à noter que la catégorie Gén assignée par l'AFP concerne toutes les catégories qui ne sont pas POL, WAR ou FIN.*

constate qu'environ 73 % des phrases étiquetées par l'AFP par la catégorie POL sont étiquetées

[6]Comme mentionné dans le chapitre 2, les lettres en arabe s'écrivent différemment si elles sont isolés, c'est le cas du mot AFP (ا ف ب). De plus il n'y a pas de notion de majuscules et minuscules en arabe. C'est pour cette raison qu'en arabe le mot AFP est constitué de trois tokens.

de la même façon par notre classifieur automatique. De même pour la catégorie WAR (environ 87 %) et par la catégorie FIN (environ 91,8 %). Finalement, environ 90 % des phrases catégorisées par l'AFP dans d'autres catégories que POL, WAR et FIN, ont été classifiées dans la catégorie Gén par le classifieur automatique.

6.5 Expériences et résultats

Dans cette section, nous allons décrire les expériences effectuées pour l'adaptation thématique selon les catégories IPTC. Les expériences de traduction du premier scénario (sans aucune classification) sont présentés dans la section 6.5.1. La section 6.5.2 décrit les expériences effectuées en utilisant la catégorisation a priori de l'AFP. La section 6.5.3 présente les expériences effectuées en utilisant la classification naïve bayésienne itérative que nous avons développée.

6.5.1 Traduction automatique sans classification

Dans cette section, on présente les résultats de traduction du test général sans considérer les codes ni les catégories IPTC. Le tableau 6.8 présente les résultats de traduction automatique "normale" sans aucune classification des phrases à traduire. La traduction est évaluée sur deux corpus de test : le corpus de test 2010 et le corpus de test 2011. Nous rappelons le score du test 2010 utilisé pour les expériences dans les chapitres précédents. Le test 2011 est plus récent que le test 2010.

Test	2010	2011
BLEU	33,30	33,47

Tableau 6.8: *Traduction automatique "normale" des tests généraux 2010 et 2011 sans aucune classification.*

On note que les résultats des deux corpus de test sont comparables et les performances sont les mêmes pour les deux tests. Le test 2011 sera donc utilisé pour la suite des expériences de ce chapitre puisqu'il est extrait de dépêches plus récentes que celles du test 2010.

6.5.2 Adaptation en utilisant la classification a priori

À partir de la classification effectuée par l'AFP, nous avons construit des modèles de traduction (TM) et des modèles de langue (LM) spécifiques. La taille des modèles spécifiques et généraux classifiés selon la catégorisation de l'AFP est présentée dans le tableau 6.9. La taille des modèles de traduction est exprimée en nombre de phrases et de segments et la taille des modèles de langue est exprimée en nombre de mots constituant le corpus monolingue de base. Les modèles Gén sont développés en utilisant l'ensemble des données (POL, WAR, FIN

Domaine	# phrases	# segments TM	# mots LM
Gén	265K	19,9M	7,8M
POL	156K	12,2M	4,7M
WAR	101K	7,6M	3M
FIN	39K	3,3M	1,1M

Tableau 6.9: *Taille des modèles de traduction (TM), de langue (LM) pour chaque type de catégorie ainsi que pour les modèles généraux (Gén) contenant toutes les catégories.*

ainsi que les autres catégories).

6.5.2.1 Systèmes de base

Nous avons constitué quatre systèmes de base pour chacune des catégories POL, WAR et FIN. Quatre combinaisons de modèles de langue et de traduction sont possibles. La première combinaison (Système général) consiste à utiliser les modèles de traduction et de langue généraux. La deuxième combinaison (Système LM-spécifique) consiste à utiliser un modèle de langue spécifique à la catégorie considérée et un modèle de traduction général. La troisième combinaison (Système TM-spécifique) consiste à utiliser un modèle de traduction spécifique et un modèle de langue général. La quatrième combinaison (Système spécifique) consiste à utiliser un modèle de langue et un modèle de traduction spécifiques à la catégorie considérée.

Le tableau 6.10 présente les résultats de traduction des tests POL, WAR, FIN et la combinaison des trois tests obtenus en traduisant avec les systèmes de traduction de base. Pour chacune des catégories considérées, deux optimisations ont été réalisées : une optimisation avec le corpus de développement général (dev Gén) et une optimisation avec le corpus de développement spécifique à la catégorie concernée. La colonne *Comb.* contient les scores de traduction des trois tests spécifiques combinés (un test de 3 milles phrases) où chaque test est traduit avec le système de traduction approprié à la catégorie lui correspondant.

	BLEU (dev Gén)				BLEU (dev Spé)			
Test / Système	POL	WAR	FIN	Comb.	POL	WAR	FIN	Comb.
Général	**32,08**	**34,87**	**31,71**	**33,10**	31,44	**34,37**	**31,50**	**32,68**
LM-spécifique	31,67	34,38	30,05	32,14	**31,59**	34,20	30,42	32,30
TM-spécifique	31,26	33,66	29,09	31,48	31,23	33,81	29,03	31,48
Spécifique	30,84	33,25	28,05	30,89	30,94	32,96	28,00	30,77

Tableau 6.10: *Traduction avec les systèmes de traduction de base, et évaluation sur les corpus de tests spécifiques POL, WAR et FIN et la combinaison des trois tests spécifiques (Comb.).*

Pour les systèmes optimisés avec le corpus de développement général, on note que le meilleur résultat pour chaque test est donné par le système général. Ceci est dû au fait que le système de traduction général contient plus de données. On note qu'il y a une légère amélioration du score BLEU pour le système LM-spécifique – optimisé sur le corpus de développement spécifique – par rapport au système général (0,15 points BLEU). On observe que l'adaptation de modèle de langue (LM-spécifique) est meilleure que l'adaptation de modèles de traduction (TM-spécifique).

On note qu'en utilisant un corpus de développement spécifique POL, le système spécifique POL améliore le score de 0,1 points BLEU. Le système TM-spécifique pour la catégorie WAR optimisé sur le corpus de développement WAR améliore le score BLEU de 0,15 points par rapport au même système optimisé sur le corpus de développement général. Une amélioration de 0,16 points BLEU pour le test (Comb.) traduit avec le LM-spécifique optimisé sur le corpus de développement spécifique par rapport au même système optimisé sur le corpus de développement général. Pour la catégorie FIN, le système LM-spécifique a un meilleur score qu'en utilisant un corpus de développement général (+0,4). Pour les autres scores, il n'y a pas une variation importante entre l'optimisation avec un corpus de développement général ou spécifique FIN. Les modèles spécifiques du domaine FIN sont quatre fois plus petits que les modèles spécifiques du domaine POL. C'est sans doute pour cette raison que la variation du score BLEU entre le système spécifique pour la catégorie FIN et le système général FIN est importante (3,5 points BLEU). Pour la suite des expériences, les systèmes de

traduction adaptés seront optimisés sur les corpus de développement spécifiques à chacune des catégories concernées (c'est plus logique d'optimiser un système de traduction sur un corpus de développement du même type que le corpus de test). Des scores de traduction supplémentaires du test général traduit avec les systèmes spécifiques sont donnés dans l'annexe 7.6.

En utilisant des modèles spécifiques, on n'utilise pas l'ensemble des données disponibles. Ces modèles sont donc limités ce qui influe sur les performances des systèmes de traduction.

Un croisement des trois modèles spécifiques (POL, WAR et FIN) sur les trois tests spécifiques a été réalisé. Le tableau 6.11 donne les résultats de traduction des trois systèmes spécifiques testés sur chacun des trois tests spécifiques POL, WAR et FIN. On note que le

	BLEU		
Système \ Test	POL	WAR	FIN
POL	**30,94**	32,82	**29,89**
WAR	28,10	**32,96**	25,61
FIN	25,25	26,67	**28,00**

Tableau 6.11: *Évaluation de tous les systèmes spécifiques optimisés sur le corpus de développement spécifique sur les trois tests spécifiques POL, WAR et FIN.*

système de traduction adapté au domaine POL donne les meilleures performances sur le test POL. De même le système de traduction adapté au domaine WAR donne les meilleures performances sur le corpus de test WAR. Le système POL donne les meilleurs résultats sur le corpus de test FIN. Ceci est dû essentiellement au fait que 40 % du corpus utilisé pour l'entraînement du système FIN a été aussi utilisé pour l'entraînement du système POL (voir tableau 6.4). De plus, les modèles de langue et de traduction pour le système POL sont plus gros que les systèmes WAR et FIN. Si on compare seulement les scores de traduction donnés par les systèmes adaptés WAR et FIN qui ont seulement 1,36 % de données communes (voir tableau 6.3), on voit que la différence entre les résultats est plus claire. Le système spécifique WAR donne un score de 32,96 sur le test WAR contre 26,67 par le système spécifique FIN. Ce dernier donne un score de 28,00 sur le corpus de test FIN (contre 25,61 par le système WAR). On voit bien donc que les résultats de traduction sont meilleurs si le système de traduction utilisé est adapté aux données à traduire – et surtout s'il ne contient pas beaucoup de données communes avec d'autres catégories.

6.5.2.2 Systèmes adaptés

Dans cette section, on propose et on compare les quatre approches d'adaptation suivantes:

(a) Fusion des modèles de traduction : tous les modèles de traduction spécifiques (POL, WAR et FIN) ainsi que le modèle général sont fusionnés dans un seul modèle. Pour chaque segment donné, s'il existe dans un ou plusieurs modèles de traduction, les scores sont recopiés dans le modèle global, sinon des scores de probabilité faibles sont assignés aux paramètres correspondants. Le modèle contient donc dix-sept paramètres : les quatre paramètres classiques pour chaque modèle ($P(s|t)$, $P(t|s)$, $lex(s|t)$ et $lex(t|s)$) et un score qui représente la probabilité de distorsion (qui est toujours constante – 2,718 – et donc elle a été assignée une seule fois pour tous les modèles).

(b) Interpolation log-linéaire de modèles de traduction : utilisation de deux modèles de traduction (spécifique et général) en privilégiant le premier modèle de traduction spécifique par rapport au modèle de traduction général (mode *either* dans Moses avec l'option

decoding-graph-backoff). Une interpolation est log-linéaire si deux ou plusieurs modèles (de langue ou de traduction) sont introduits séparément dans le système. C'est-à-dire qu'ils sont introduits en parallèle dans le fichier de configuration de Moses. Leurs poids sont optimisés séparément avec MERT (Och, 2003) afin d'optimiser les performances de traduction;

(c) Interpolation log-linéaire de modèles de langue : utilisation de deux modèles de langues (spécifique et général);

(d) Interpolation linéaire de modèles de langue : les coefficients d'interpolation sont estimés en optimisant la perplexité sur un corpus de développement spécifique. L'interpolation linéaire consiste à estimer les coefficients pour chacun des modèles de langue (spécifique et hors-domaine) à interpoler. En premier lieu, chacun des modèles de langues est créé séparément, ensuite les deux modèles de langue sont interpolés linéairement à l'aide de l'outil SRILM en donnant les poids estimés pour chacun des modèles.

Le tableau 6.12 présente (i) les résultats de traduction des corpus de tests spécifiques (POL, WAR et FIN) en utilisant les différentes approches d'adaptation et (ii) les résultats de traduction du test Comb. constitué de 3 milles phrases et qui contient la combinaison des 3 tests spécifiques traduits chacun avec son système approprié.

			BLEU			
Approche	TM	LM	POL	WAR	FIN	Comb.
(a.1)	Global	Gén	31,68	34,18	31,35	32,61
(a.2)	Global	Spé	31,36	34,01	29,32	31,75
(b.1)	Spé+Gén	Gén	31,68	34,36	31,54	32,74
(b.2)	Spé+Gén	Spé	31,55	33,32	30,32	31,85
(c.1)	Gén	Spé+Gén (log-lin)	31,36	34,16	31,40	32,57
(c.2)	Spé	Spé+Gén (log-lin)	31,26	34,04	29,19	31,68
(b)+(c)	Spé+Gén	Spé+Gén (log-lin)	**32,04**	34,26	31,63	32,91
(d.1)	Gén	Spé+Gén (lin)	31,84	**34,42**	**31,78**	**32,98**
(d.2)	Spé	Spé+Gén (lin)	31,18	33,68	29,00	31,50
(b)+(d)	Spé+Gén	Spé+Gén (lin)	32,03	34,19	31,57	32,85

Tableau 6.12: *Traduction en utilisant les différentes méthodes d'adaptation, et évaluation sur les test spécifiques pour les domaines POL, WAR et FIN ainsi que les trois tests combinés (Comb.). Chaque système de traduction est optimisé sur le corpus de développement spécifique à la catégorie concernée.*

Pour chacune des approches (a) et (b), deux tests sont réalisés, un test en utilisant le modèle de langue général et un test en utilisant le modèle de langue spécifique. De même pour les approches (c) et (d), un test est effectué en utilisant le modèle de traduction général et un deuxième test en utilisant le modèle de traduction spécifique. Les systèmes de traduction sont optimisés sur le corpus de développement spécifique à la catégorie concernée. On observe que l'utilisation des modèles généraux améliore les scores jusqu'à 2 points BLEU pour la catégorie FIN pour laquelle les modèles spécifiques de langue et de traduction sont les plus petits. On note que la combinaison des approches (b) et (c) donne les meilleurs scores BLEU pour la catégorie POL. En ce qui concerne les catégories WAR et FIN, et la combinaison des 3 tests, le meilleur score BLEU est donné par l'approche (d.1).

D'après les résultats précédents, on note que généralement les meilleurs résultats sont obtenus en utilisant un modèle de langue adapté par interpolation linéaire. Il est clair que l'utilisation d'un modèle de langue interpolé est meilleure que d'utiliser un modèle de langue

spécifique seul. L'interpolation de deux modèles de traduction est meilleure que l'utilisation d'un modèle de traduction seul. Les coefficients d'interpolation étaient respectivement 0,5, 0,5 et 0,6 pour les catégories POL, WAR et FIN. On observe également qu'un modèle de traduction de taille réduite a un impact sur le score de traduction même si le modèle de langue est de grande taille.

En comparant avec le tableau 6.10, on note, qu'en optimisant avec le corpus de développement spécifique, le score de traduction global (Comb.) est amélioré de 32,68 à 32,98. On note également que les scores donnés par les systèmes adaptés sont en général améliorés par rapport à ceux donnés par les systèmes de base pour chacune des trois catégories indépendamment (jusqu'à 0,5 points BLEU). Nous allons nous limiter donc dans la section 6.5.3 aux expériences sur les systèmes adaptés.

6.5.3 Adaptation en utilisant la classification automatique

De nouveaux modèles spécifiques ont été entraînés sur la base de la classification automatique décrite dans la section 6.4.3. La taille des modèles est précisée dans le tableau 6.13. On observe

Domaine	# phrases	# segments TM	# mots LM
Gén	265K	19,9M	7,8M
POL	60K	5,8M	1,8M
WAR	38K	3,5M	1,0M
FIN	29K	3,1M	868K

Tableau 6.13: *Taille des modèles de traduction (TM) et de langue (LM) classifiés automatiquement.*

que ces nouveaux modèles sont plus petits que ceux entraînés à partir de la classification a priori de l'AFP (tableau 6.9). Ceci est dû au fait que chaque phrase est classifiée dans une et une seule catégorie.

Le tableau 6.14 montre les résultats obtenus pour la traduction des tests spécifiques POL, WAR et FIN ainsi que la combinaison des trois tests (colonne Comb.). Les différentes versions d'adaptation du système de traduction présentées en section 6.5.2.2 ont été proposées avec les nouveaux modèles construits à partir de la classification automatique. Les coefficients d'interpolation étaient les mêmes pour les trois catégories : 0,8 pour le modèle de langue général et 0,2 pour le modèle de langue spécifique. En comparant les résultats du tableau 6.14 avec le tableau 6.12, on observe une légère amélioration (+0,16) des scores de traduction du test Comb. et une amélioration d'environ 0,5 points BLEU par rapport au score initial du tableau 6.10. On note également que les scores de traduction baissent pour les systèmes (b), (b)+(c) et (b)+(d) constitués de deux modèles de traduction interpolés log-linéairement. Ceci est dû probablement au fait que nous avons privilégié le modèle de traduction (trop petit) par rapport au modèle de traduction général qui englobe la totalité des données. Pour les systèmes (c.1) et (d.1) les scores de traduction pour les 4 tests s'améliorent par rapport aux systèmes avec classification a priori.

6.5.4 Comparaison des trois scénarios

Afin d'avoir une vue plus claire sur les résultats obtenus et de pouvoir les comparer d'une manière plus explicite, nous résumons les résultats obtenus dans la figure 6.5 qui montre une comparaison des scores BLEU obtenus pour les trois scénarios étudiés dans ce chapitre sur le corpus de test Comb.

Approche	TM	LM	BLEU POL	WAR	FIN	Comb.
(a.1)	Global	Gén	30,60	34,68	31,15	32,41
(a.2)	Global	Spé	29,63	31,72	29,43	30,45
(b.1)	Spé+Gén	Gén	31,06	34,26	31,08	32,41
(b.2)	Spé+Gén	Spé	27,06	31,43	29,15	29,50
(c.1)	Gén	Spé+Gén (log-lin)	31,50	34,62	31,62	32,85
(c.2)	Spé	Spé+Gén (log-lin)	28,77	29,46	27,89	28,97
(b)+(c)	Spé+Gén	Spé+Gén (log-lin)	31,52	33,96	30,43	32,26
(d.1)	Gén	Spé+Gén (lin)	**31,94**	**34,84**	**31,83**	**33,14**
(d.2)	Spé	Spé+Gén (lin)	29,60	29,79	27,97	29,31
(b)+(d)	Spé+Gén	Spé+Gén (lin)	31,37	34,31	31,54	32,66

Tableau 6.14: *Évaluation de la traduction des corpus de test spécifiques POL, WAR, FIN ainsi que les trois tests combinés (Comb.) avec les approches d'adaptation.*

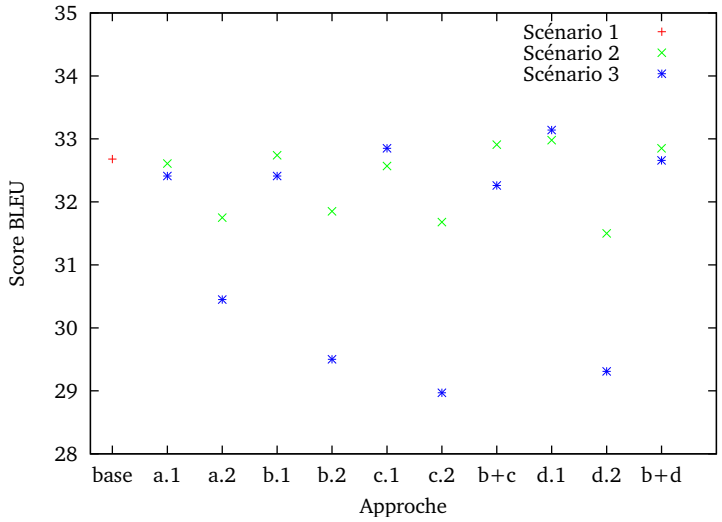

Figure 6.5: *Score BLEU évalué sur le corpus de test Comb. pour les 3 scénarios : sans classification (Scénario 1), avec classification a priori (Scénario 2) et avec classification automatique (Scénario 3).*

On observe que les systèmes adaptés améliorent les résultats par rapport aux systèmes sans classification. Le meilleur score obtenu est donné par le scénario avec classification automatique. Le score donné pour le scénario 1 est celui obtenu par la traduction avec le système général (modèles de langue et de traduction généraux) et optimisé sur le corpus de développement spécifique.

D'après la figure, on voit clairement pour les approches où des modèles spécifiques (de langue ou de traduction) sont utilisés, la classification a priori (scénario 2) est meilleure que la classification automatique (scénario 3) : la taille des modèles spécifiques utilisés pour les expériences du scénario 3 sont plus petits que ceux utilisés pour le scénario 2. Pour les approches où les modèles généraux sont utilisés, les résultats sont parfois meilleurs pour le scénario 3. Ce dernier améliore légèrement les résultats par rapport au scénario 2 lorsque les modèles de langue et de traduction sont interpolés.

Afin de comparer les 3 scénarios pour le corpus de test général, nous présentons les résultats de traduction de ce test où **chaque phrase du test général est traduite avec le système approprié à la catégorie détectée** comme l'approche de Yamamoto et Sumita (2008) et de Sennrich (2012a). Les résultats sont présentés par la figure 6.6.

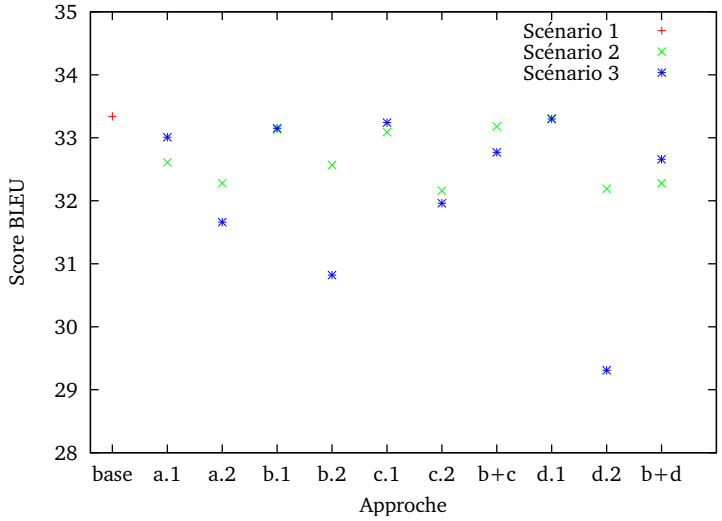

Figure 6.6: *Score BLEU évalué sur le corpus de test général – où chaque phrase est traduite avec le système approprié à la catégorie détectée – pour les 3 scénarios : sans classification (Scénario 1), avec classification a priori (Scénario 2) et avec classification automatique (Scénario 3).*

On observe que le meilleur score obtenu pour les scénarios avec classification est le même que celui obtenu par le premier scénario. Le score donné pour le scénario 1 est celui obtenu par la traduction avec le système général (modèles de langue et de traduction généraux) et optimisé sur le corpus de développement spécifique. On observe comme dans la figure 6.5 pour les approches où des modèles spécifiques (de langue ou de traduction) sont utilisés qu'il y a une baisse du score BLEU.

La figure 6.7 montre une phrase extraite du test traduit par le système de traduction général et une phrase extraite du corpus de test traduit en classification par phrase (ClassPhrase) avec l'approche (d.1). Cette phrase appartient à la catégorie POL.

Gén :	*le ministre israélien des Affaires étrangères Avigdor Lieberman* , cité par le quotidien Maariv mardi à de nouvelles sanctions internationales " qui paralyse " de l' Iran .
ClassPhrase :	*le chef de la diplomatie israélienne Avigdor Lieberman* , cité par le quotidien Maariv mardi à de nouvelles sanctions internationales " paralysée " de l' Iran .
Réf. :	l' Iran doit être frappé par des sanctions internationales " paralysantes " , a affirmé *le chef de la diplomatie israélienne Avigdor Lieberman* selon ses propos publiés mardi par le journal Maariv .

Figure 6.7: *Comparaison d'une phrase traduite par le système de traduction.*

On note que même si le score BLEU ne montre pas d'amélioration des performances, la méthode de classification par phrase améliore la traduction de la phrase spécifique à la

catégorie POL par rapport à celle traduite avec le système général.

6.6 Conclusion

Dans ce chapitre, nous avons présenté un ensemble d'expériences sur l'adaptation thématique. Un corpus pré-classifié par l'AFP selon les 17 catégories de l'IPTC est utilisé pour réaliser cette étude. Les catégories utilisées pour ces travaux sont celles qui sont les plus présentes dans les dépêches de l'AFP : politique (POL), guerres (WAR) et finances (FIN). Une catégorie générale qui englobe toutes les données (Gén) est également considérée.

Trois scénarios de traduction ont été proposés : une traduction "normale" sans classification, une traduction avec classification a priori des phrases et une traduction avec classification automatique.

Différentes méthodes d'adaptation de modèles de langue et de traduction ont été expérimentées. Les meilleurs résultats de la traduction avec classification sont donnés par l'approche d'adaptation (d.1) qui propose une interpolation linéaire du modèle de langue spécifique et du modèle de langue général.

Nous avons constaté qu'en utilisant les modèles construits par classification automatique pour traduire le corpus de test (Comb.), les approches (c.1) et (b)+(d) améliorent les résultats avec respectivement 0,3 et 0,16 points BLEU. Les scores en utilisant seulement les modèles spécifiques n'améliorent pas les résultats puisque chaque phrase est classifiée dans une et une seule catégorie afin d'éviter l'aspect de redondance des données dans plusieurs modèles. La conséquence de ce choix est que les modèles spécifiques sont devenus encore plus petits que les premiers modèles construits à partir de la classification de l'AFP et donc moins performants. Finalement, les systèmes adaptés améliorent les scores de traduction sur le test Comb. jusqu'à 0,5 points BLEU par rapport au test initial optimisé le corpus de développement spécifique.

La méthode de classification par phrase du corpus de test général afin de traduire chaque phrase avec le système approprié à la catégorie assignée ne montre pas de gains importants. Ceci est dû principalement à deux raisons : la taille très limitée et très petite des modèles spécifiques et les frontières floues entre les catégories IPTC.

Récemment, Mansour et Ney (2013) montrent qu'en utilisant leur méthode d'adaptation, les modèles adaptés n'améliorent pas la traduction puisque leur taille est inférieure de 10 % par rapport aux modèles complets.

Dans les travaux présentés précédemment sur l'adaptation comme ceux de Banerjee (2012) ou de Sennrich (2012b), les données thématiques sont créées à partir de données extraites de corpus différents, ce qui n'est pas le cas pour notre étude. La présence des phrases multicatégoriques complique la tâche de classification et de séparation des données dans des catégories différentes; par contre il s'agit d'une application réelle dans laquelle les données – issues d'une même source – sont multicatégoriques. Certaines phrases pouvant apparaître dans n'importe quelle catégorie sont considérées appartenant à une catégorie bien spécifique. Nous avons tenté dans ce chapitre différentes méthodes d'adaptation de données journalistiques extraites d'une application réelle, celle de l'AFP.

Un des obstacles que nous avons rencontré lors de cette étude est l'absence d'un corpus de référence classifié selon la catégorisation IPTC. Nous avons utilisé la classification de l'AFP comme référence, mais cette classification contient beaucoup d'erreurs, et classifie une dépêche dans plusieurs catégories, ce qui complique la tâche de catégorisation.

Il serait intéressant d'ajouter des données parallèles et monolingues afin d'améliorer le classifieur automatique.

Plusieurs perspectives sont envisageables pour ces travaux. Une perspective intéressante serait de pondérer les phrases à traduire au moment de la traduction afin de contourner le problème de la taille réduite des catégories et surtout les modèles spécifiques. Ceci permettrait de combiner plusieurs systèmes lors de la traduction et de garder toutes les

catégories attribuées à chaque phrase pour le corpus de test à traduire. Une optimisation permettrait d'avoir pour chaque système un poids qui sera utilisé au moment du décodage lorsque plusieurs systèmes (ou modèles) sont utilisés pour la traduction. Une autre perspective intéressante consiste à construire des modèles sur l'ensemble des catégories de la phrase à traduire, ce qui permettrait d'avoir des modèles plus enrichis et de pallier au problème de manque de données.

Ces expériences peuvent être enrichies également en adaptant les modèles de réordonnancement comme les travaux de Chen, Foster et Kuhn (2013) par exemple.

CHAPITRE 7

Vers une application réelle : intégration de la traduction dans une plateforme d'analyse multimédia

L'objectif du projet SAMAR, comme présenté dans l'introduction de ce manuscrit, est de développer une plateforme de traitement multimédia en langue arabe. Les données journalistiques (dépêches AFP) que nous avons utilisées pour nos recherches et pour construire des systèmes de traduction étaient fournies dans le cadre de ce projet.

La plateforme SAMAR contient plusieurs fonctionnalités. Elle permet de transcrire automatiquement de la parole (débats diffusés sur les radios et télévisions, et principalement des nouvelles journalistiques en langue arabe). Une analyse sémantique est également effectuée : les entités nommées (noms de personnes, de lieux, d'organisations ou de marques) sont extraites et constituent les sujets des informations analysées. Les textes en arabe sont catégorisés suivant la taxonomie standardisée par l'IPTC (International Press Telecommunication Council) en vigueur dans les agences de presse, et permet une recherche de concepts par thème. Une gestion cross-média et cross-lingue est effectuée pour les documents multimédias avec métadonnées. Cette gestion permet des recherches dans les textes et les vidéos en arabe, français et anglais.

L'objectif final du projet est essentiellement d'avoir un moteur de recherche qui indexe les dépêches dans les trois langues (arabe, français et anglais) quelle que soit la langue dans laquelle la recherche est effectuée. L'intégration de la traduction s'effectue donc dans le cadre du moteur de recherche indexé multilingue.

Dans ce chapitre, on présente essentiellement comment la traduction automatique est utilisée dans le cadre d'une application réelle. Nous présentons dans la section 7.1 le projet SAMAR, son architecture et le rôle que joue la traduction automatique dans le projet. La section 7.2 présente les différentes versions de systèmes de traduction arabe-français et arabe-anglais qui ont été livrées pour le projet SAMAR. La section 7.3 présente l'intégration des systèmes dans la plateforme. On présente dans une dernière section (7.5), nos participations à des campagnes d'évaluation pour la traduction automatique ainsi que les résultats que nous avons obtenu.

7.1 Le projet SAMAR

Le projet SAMAR est un projet financé par Cap Digital. De nombreux partenaires industriels et laboratoires académiques ont participé à la réalisation de ce projet, qui a été piloté par la société Temis. Ce projet regroupe onze partenaires. L'*AFP* est le fournisseur des flux multimédia radio et télévisuels. Ces flux sont la source essentielle des corpus arabe, français et anglais. Du côté du traitement de la parole : la société *Vecsys*, spécialiste de la reconnaissance vocale et du traitement automatique de la parole est le fournisseur des transcriptions manuelles du texte à partir de la parole et l'entreprise *Vocapia* est le partenaire qui travaille sur le traitement automatique de la parole. Pour l'aspect linguistique on cite le *Laboratoire du Langage, Langues et Cultures d'Afrique Noire*, (*LLACAN*) pour l'analyse de l'arabe littéraire et dialectal ainsi que l'*Institut National des Langues et Civilisations Orientales* (l'*Inalco*) pour la validation de modèles. Pour la traduction automatique : le *Laboratoire d'Informatique pour la Mecanique et les Sciences de l'Ingénieur* (*LIMSI*) pour la modélisation de traduction et le *Groupe de REcherche en Informatique, Image, Automatique et Instrumentation de Caen*, le *Greyc* pour les alignements de textes. En ce qui concerne la fouille et la classification du texte : la société *Temis* pour l'extraction de connaissances à partir de textes, la société *Antidot* pour la recherche cross-lingue et la société *Mondeca* pour la gestion des ontologies. Finalement la société *Nuxeo* pour la gestion de contenu multimédia et l'intégration. La figure 7.1 montre le rôle de chaque partenaire pour la plateforme SAMAR.

Figure 7.1: *Rôle de chacun des partenaires du projet dans la plateforme SAMAR*[1].

L'objectif est de réaliser une station d'analyse multimédia en langue arabe. C'est une plateforme modulaire permettant (i) l'intégration et les traitements homogènes de flux écrits audio et vidéo (transcription, analyse linguistique, indexation, etc) ainsi que (ii) la traduction automatique.

7.1.1 Architecture

La figure 7.2 montre l'architecture de la plateforme SAMAR. Cette architecture a été illustrée par Nuxeo (le partenaire qui s'occupe de la partie intégration). Cette figure présente les interactions entre les partenaires du projet pour la construction de la plateforme SAMAR.

[1]Schéma extrait de la plaquette de présentation du projet SAMAR (www.samar.fr/wp-content/uploads/2012/11/Samar_Ok2_Web.pdf)

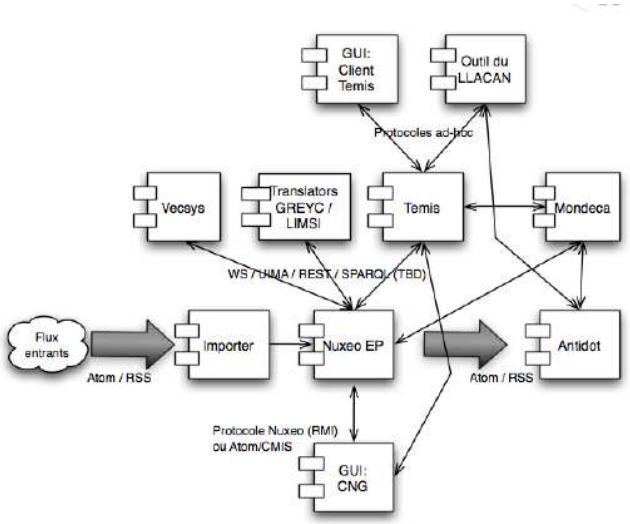

Figure 7.2: *Architecture de la plateforme SAMAR telle que illustrée par Nuxeo*[2].

Neuf partenaires apparaîssent sur la figure. L'AFP n'a pas d'interaction directe sur la plateforme SAMAR puisque son rôle principal est de fournir les corpus. Vecsys regroupe Vecsys et Vocapia qui travaillent en collaboration sur la transcription de l'audio. L'INALCO travaille en collaboration avec le LLACAN et Temis puisqu'il vérifie et valide les modèles.

On note que Vecsys-Vocapia intéragissent directement avec la plateforme de Nuxeo. Le LIMSI en collaboration avec le GREYC interagissent directement avec la plateforme. Le LLACAN, Mondeca et Antidot interagissent avec Temis qui, lui, interagit directement avec la plateforme. On note que la plupart des partenaires interagissent directement avec Nuxeo (le partenaire responsable d'intégration). L'INALCO et le LLACAN interagissent seulement avec la société Temis pour la validation des modèles. Mondeca gère les ontologies et interagit donc avec Antidot et Temis.

7.1.2 Le module traduction du LIMSI

La plateforme SAMAR est constituée de plusieurs composantes comme le montre son architecture. La composante LIMSI est liée directement à la plateforme d'intégration Nuxeo. Les données à traduire sont données au système de traduction sous leur forme d'origine : des dépêches AFP sous format NewsML. Afin de traduire une dépêche, il est donc nécessaire d'effectuer un certain nombre de traitements.

Cette chaîne de prétraitement est constituée de plusieurs étapes. Les dépêches à traiter sont reçues au format NewsML utilisé par l'AFP. La première étape de prétraitement consiste à extraire le texte du format NewsML, ensuite vient l'étape du prétraitement de l'arabe, la segmentation. Une fois prétraité, le texte est donné au décodeur qui traduit le texte à l'aide du décodeur Moses. La sortie de traduction est par la suite détokenisée, c'est-à-dire que la première lettre de chaque phrase est mise en majuscule et que les espaces avant les ponctuations sont supprimés. Finalement, le texte est remis en format NewsML pour qu'il puisse être lu par les outils de l'AFP.

La figure 7.3 montre la chaîne de traitement réalisée pour traduire une dépêche AFP en format NewsML et ensuite la remettre dans le format original NewsML.

[2]Schéma extrait de la présentation du premier bilan d'étape du projet SAMAR (http://www.samar.fr/?p=91)

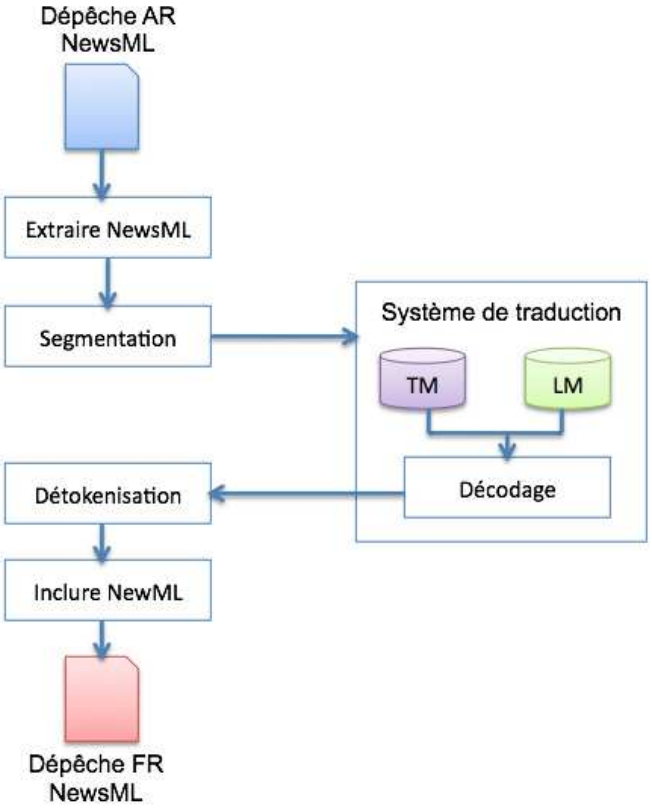

Figure 7.3: *Chaîne de traitement d'une dépêche AFP pour la traduction de l'arabe vers le français.*

À la sortie du système de traduction, le texte traduit est affiché directement sous forme de dépêche AFP.

7.2 Expériences et résultats

Le système que nous utilisons dans le cadre du projet SAMAR est le système *open source* Moses[3]. Il intègre un certain nombre d'autres composants *open source*, et repose en particulier sur l'aligneur sous-phrastique MGIZA++[4] Gao et Vogel (2008) pour la phase d'entraînement. Les tables de traduction sont constituées en symétrisant les alignements selon l'heuristique *grow-diag-final-and* de Moses, et contient des segments dont la longueur va jusqu'à 7 mots. Un modèle de distorsion (réordonnancement) lexical a été entraîné sur les parties arabe et française. Un modèle de langue en langue cible (français) a été appris à l'aide de l'outil SRILM[5] (Stolcke, 2002).

Le modèle de langue utilisé est binarisé afin d'avoir un temps d'exécution plus rapide. Le modèle de traduction n'est pas binarisé puisque sa taille n'est pas très volumineuse.

[3] http://moses.statmt.org/
[4] http://geek.kyloo.net/software/doku.php/mgiza:overview
[5] http://www.speech.sri.com/projects/srilm/

7.2.1 Système de traduction arabe-français

Le corpus arabe-français fourni par l'AFP est un corpus comparable qui ne peut pas être utilisé directement pour construire un système de traduction. Comme mentionné dans le chapitre 3, nous avons développé une méthode d'extraction de corpus parallèle à partir du corpus comparable. Le corpus obtenu a été utilisé pour construire ce système de traduction arabe-français SAMAR-1. Ce dernier est constitué donc de 145K phrases extraites pour la période décembre-2009 jusqu'à novembre-2010, et il a été prétraité avec MADA. Dans le chapitre 4, une nouvelle approche de prétraitement de l'arabe a été proposée avec SAPA. Nous avons donc construit un nouveau système arabe-français amélioré, SAMAR-2, qui a été prétraité avec notre outil de prétraitement, SAPA. Le système SAMAR-2 contient exactement le même nombre de phrases que SAMAR-1.

Une troisième version du système de traduction, SAMAR-3, a été construite en incluant des données supplémentaires qui ont été extraites avec notre méthode d'extraction de corpus parallèle (période de janvier 2011 jusqu'a juillet 2012). Le nombre total des phrases utilisées pour SAMAR-3 est de 265K phrases. Le prétraitement a été effectué avec SAPA. La dernière version du système de traduction, SAMAR-4, a été construite avec les mêmes données que le système SAMAR-3. Par contre, le modèle de langue utilisé pour le système SAMAR-4 est adapté aux données de traduction et c'est un modèle de langue constitué à partir de dépêches AFP provenant du Moyen-Orient. Les résultats sont dans le tableau 7.1.

Système ar-fr	Taille corpus	Date données	Prétraitement	BLEU	Vitesse (mots/sec)
SAMAR-1	145K	Dec2009-Nov2010	MADA	32,8	90
SAMAR-2	145K	Dec2009-Nov2010	SAPA	33,3	2 960
SAMAR-3	265K	Dec2009-Nov2010 +Janv2011-Juil2012	SAPA	34,4	2 960
SAMAR-4	265K	Dec2009-Nov2010 +Janv2011-Juil2012	SAPA	35,8	2 960
Google	-	Novembre 2012[6]	-	25,3	-

Tableau 7.1: *Scores BLEU en traduction arabe-français sur un test de 1 000 phrases extraites de dépêches de décembre 2010.*

D'après le tableau 7.1, on note que le système SAMAR-2 améliore les résultats de SAMAR-1 grâce à l'outil de prétraitement. Le rajout de données (SAMAR-3) permet de gagner 1 point BLEU. L'utilisation d'un modèle de langue adapté, SAMAR-4, améliore les résultats de 1 point BLEU. Ce modèle adapté aux dépêches AFP a une taille 7 fois inférieure au modèle de langue général, mais donne de meilleures performances pour la traduction. On constate que lorsque les données utilisées pour la construction du modèle de langue proviennent de sources bien séparées (contrairement aux données utilisées dans le chapitre 6), l'adaptation améliore les scores de traduction. Nous avons donc amélioré les performances du système de traduction de base SAMAR-1 de 3 points BLEU.

Afin d'avoir une idée des performances de notre système de traduction, nous avons traduit notre test avec Google de l'arabe vers le français. Ce dernier donne un score BLEU largement inférieur au score BLEU donné par notre meilleur système (25,3 contre 35,8 pour notre système). Ceci est dû au fait que le système de traduction de Google est très général et il a été entraîné sur beaucoup de données très hétérogènes (il peut contenir beaucoup de données provenant de l'ONU par exemple). Par contre, les systèmes SAMAR-1, SAMAR-2,

[6]La traduction avec le système de traduction de Google a été effectuée en mois de novembre 2012.

SAMAR-3 et SAMAR-4 ont été entraînés sur des données du même type que le corpus de test.

On note également que les données extraites pour la deuxième période, plus longue que la première période, constituent un corpus de 120K phrases seulement. Le nombre de phrases de l'ensemble des fichiers similaires pour la partie arabe est d'environ 396 mille phrases pour la première période et d'environ 480 mille phrases pour la deuxième période. Par contre les dépêches pour la deuxième période sont constitués de phrases courtes (4% des dépêches sont constitués de moins de 5 lignes contre 0% pour la première période). Nous avons particulièrement remarqué pour les dépêches courtes de la deuxième période, beaucoup de phrases sont regroupés sur une même ligne. La longueur moyenne d'une phrase pour les phrases de la première période est de 13,34 mots alors que la longueur moyenne d'une phrase pour les phrases de la deuxième période est de 17,78 mots.

Le tableau 7.1 présente les résultats en termes de scores BLEU sur le jeu de test datant de décembre 2010. La traduction a été effectuée sur le corpus de test tokenisé. Dans ce tableau, on compare le système de base arabe-français aux systèmes de base améliorés.

On note que les résultats donnés pour le système SAMAR-1 sont différents de ceux donnés dans le chapitre 3. Ceci est dû principalement au fait que le modèle de langue utilisé est différent et plus riche que celui utilisé dans les premières expériences.

7.2.2 Traduction de transcriptions audio

Une parmi les tâches que doit effectuer la plateforme SAMAR, est la traduction de l'audio. Malheureusement l'AFP ne dispose pas d'un corpus parallèle de type audio pour la paire de langue arabe-français et arabe-anglais. Or on sait que, d'après le chapitre 2 que pour la langue arabe, le dialecte est très différent de l'écrit, et diffère aussi d'une région à une autre, ce qui rend la tâche de sa traduction plus complexe.

Un corpus de test issu des données recueillies et transcrites a été sélectionné par le partenaire Vecsys comme représentatif de l'ensemble des données du projet. Ce corpus est constitué de débats politiques provenant de différentes chaines télévisées et de vidéos de presse : Al Nile TV, Al Jazeera TV, Al Arabia, France 24, BBC et AFP. Ces données sont donc constituées de textes en arabe moderne standard la plupart du temps (90 % des cas) qui contiennent des mots, des segments et des phrases dialectales (10 %). Le projet ne prévoyant pas de budget pour des traductions manuelles, ce corpus a été traduit manuellement de l'arabe vers le français par les partenaires arabophones du projet pour pouvoir constituer un corpus de test. Une transcription automatique a été également donnée par Vocapia. Le corpus obtenu est constitué de 650 phrases en français. Un travail de nettoyage a été également réalisé, puisque les traductions manuelles ont été réalisées par plusieurs personnes qui ne respectent pas toujours la même annotation et ponctuation. Les traductions manuelles sont par la suite alignées avec les transcriptions automatiques et manuelles. Le corpus parallèle final est constitué de 400 phrases (arabe transcrit manuellement, arabe transcrit automatiquement et français traduit manuellement).

La traduction de l'audio s'effectue sur un corpus transcrit automatiquement. Les systèmes de traduction ont été adaptés sur les données issues de l'audio afin de s'en rapprocher le plus de ces données. L'approche consiste donc à transformer les données d'apprentissage textuelles pour les rendre les plus proches possibles de données recueillies à la sortie du système de reconnaissance de la parole. Principalement, l'adaptation a porté sur quatre points : normalisation, traitement des ponctuations, optimisation des paramètres sur des données de développement et utilisation du lexique de transcription automatique.

Le système de traduction arabe-français SAMAR-2 adapté à l'audio est utilisé donc pour traduire les transcriptions. Le tableau 7.2 montre les résultats obtenus pour la traduction des transcriptions sur le corpus de test construit manuellement. On note une grande baisse des scores de traduction sur la traduction de l'audio par rapport aux scores de traduction

	BLEU
Transcription automatique	16,25
Transcription manuelle	19,55

Tableau 7.2: *Évaluation de la traduction automatique des transcriptions automatiques et manuelles*

obtenus sur l'écrit (19,55 contre 33,3). Globalement, l'ensemble des méthodes utilisées pour l'adaptation de la traduction à des données issues de l'audio, a permis une amélioration des performances de traduction des transcriptions automatiques d'environ 1 point BLEU. Les résultats restent très modestes et ceci est dû à plusieurs raisons : l'arabe parlé est très différent de l'arabe écrit, la transcription automatique a un taux d'erreur de 23,6% et la ponctuation y est incomplète. Compte tenu de ces résultats modestes, il a été décidé de ne pas intégrer ces propositions dans la plateforme.

7.2.3 Système de traduction arabe-anglais

Pour la paire de langues arabe-anglais, il n'y a pas de dépêches AFP en arabe traduites de ou vers l'anglais. Nous avons donc construit notre corpus parallèle artificiel en appliquant notre méthode de construction d'un corpus parallèle artificiel décrite dans le chapitre 3. Le LIMSI dispose d'un corpus NIST extrait des données LDC pour la paire de langues arabe-anglais. Comme déjà mentionné dans le chapitre 2, les données LDC proviennent de plusieurs sources incluant les flux de journaux et du web (blogs, newsgroups, email) dans plusieurs domaines. Ce corpus a été utilisé pour construire le système de référence arabe-anglais, Référence-AR:EN. Ce dernier a été utilisé pour traduire les dépêches AFP du projet de l'arabe vers l'anglais. Les prétraitements ont été effectués avec SAPA. Les dépêches traduites ont été par la suite filtrées pour ne garder que 50% des données incluant les phrases qui ont le meilleur score de traduction. La figure 7.4 illustre notre méthode de construction du corpus parallèle arabe-anglais à partir d'un corpus monolingue.

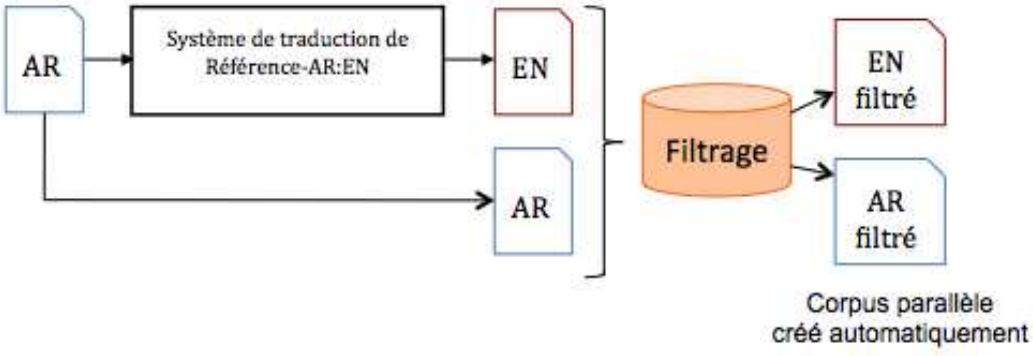

Figure 7.4: *Construction de corpus parallèle arabe-anglais à partir d'un corpus monolingue.*

Le corpus parallèle construit automatiquement a été utilisé par la suite pour entraîner le système de traduction adapté arabe-anglais, SAMAR-AR:EN.

Pour la paire de langues arabe-anglais, on ne dispose pas de corpus de test issu de dépêches AFP provenant du Moyen-Orient. Le seul corpus de test disponible est le corpus

de test NIST constitué de données journalistiques. Nous avons donc utilisé ce corpus pour l'évaluation.

Le tableau 7.3 montre les scores de traduction du système Référence-AR:EN et le système SAMAR-AR:EN sur un corpus de test (NIST) constitué de 813 phrases extraites de données journalistiques. Les scores de traduction ont été comparés au système de traduction de Google.

Système ar-en	Taille corpus	Prétraitement	BLEU
Référence-AR:EN	5M (dont 1,2M news)	SAPA	44,7
SAMAR-AR:EN	400K	SAPA	42,6
Google	-	-	50,2

Tableau 7.3: *Scores BLEU en traduction arabe-anglais évalués sur un corpus de test NIST constitué de 813 phrases – extraites de données journalistiques – sur 4 références.*

Le tableau 7.3 montre que le système Référence-AR:EN donne de meilleures performances que le système SAMAR-AR:EN. Ceci est dû au fait que les données de test sont adaptées au système de référence : elles proviennent de la même source et sont extraites à partir de données LDC.

On voit que bien que le système SAMAR-AR:EN est construit à partir de corpus parallèle artificiel, le score de traduction obtenu est un score respectable, qui n'est pas très inférieur à celui obtenu par le système Référence-AR:EN. Le système de traduction de Google a le meilleur score puisqu'il a été entraîné fort probablement sur beaucoup plus de données y compris des données journalistiques. La figure 7.5 montre une phrase en arabe et sa traduction en anglais par les systèmes Référence-AR:EN et SAMAR-AR:EN.

Arabe	مهما كان الكفاح صعبا في العراق ينبغي علينا الانتصار.
Référence	however difficult the fight is in Iraq , we must win it .
Référence-AR:EN	no matter how difficult was the struggle in Iraq should be us victory .
SAMAR-AR:EN	whatever the difficult struggle in Iraq should be us victory .

Figure 7.5: *Comparaison de traduction d'une phrase avec les systèmes de traduction Référence-AR:EN et SAMAR-AR:EN.*

Rappelons que cette méthode a été évaluée sur la paire de langues arabe-français dans le chapitre 3 puisqu'on dispose d'un corpus de test arabe-français issu de dépêches AFP provenant du Moyen-Orient. Les résultats ont montré que le système SAMAR-AR:FR (qui donne un score BLEU de 24) améliore les résultats par rapport au système Référence-AR:FR (qui donne un score BLEU de 26,7) de 2,7 points BLEU. La traduction automatique donnée par le système de traduction de Google donne un score BLEU de 23,6.

On peut donc raisonnablement en déduire que les performances du système SAMAR-AR:EN sur des dépêches AFP se situe au dessus de celles obtenues sur le corpus de test NIST.

7.3 Intégration dans la plateforme

Comme nous l'avons déjà mentionné, la plateforme SAMAR contient plusieurs fonctionnalités. Le système SAMAR comporte outre la traduction automatique développée par le LIMSI, la transcription de la parole en arabe (Vocapia, Vecsys), la détection des entités nommées (noms propres de lieux, organisations, personnes) développée par Temis et le laboratoire LLACAN

ainsi que l'INALCO, et la catégorisation automatique IPTC des documents. La traduction automatique arrive en bout de chaîne du traitement de l'information.

En tapant un mot clé à rechercher sur la plateforme, toutes les dépêches audio et texte contenant ce mot en arabe, français ou anglais apparaîssent. Ceci est rendu possible grâce à l'indexation multilingue.

Figure 7.6: *Traduction automatique de la transcription de la vidéo.*

La figure 7.6 montre une capture d'écran des différentes fonctionnalités la plateforme SAMAR. À gauche on voit la vidéo consacrée à *la victoire historique de Barack Obama pour son deuxième mandat*. À droite, la transcription automatique en arabe de la bande son. Les entités nommées en arabe sont extraites du texte et sont affichées en dessous du texte transcrit en arabe dans des boîtes bleues. La traduction automatique vers le français de la transcription automatique est affichée sous la bande son.

7.4 Réalisations

Dans cette section on présente un bilan de toutes les réalisations qui ont été effectuées pour le projet SAMAR.

- Rapport livrable 7.1 sur les ressources en arabe qui énumère un ensemble des ressources numériques sur l'arabe. Ce rapport a été rédigé par le LIMSI et le GREYC.
- Rapport livrable 7.2 qui décrit le système de base arabe-français. Ce rapport a été rédigé et livré par le LIMSI
- Système de traduction arabe-français SAMAR-4.
- Système de traduction arabe-anglais SAMAR-AR:EN.
- Corpus parallèle arabe-français constitué automatiquement à partir du corpus comparabe AFP
- Corpus parallèle arabe-anglais constitué automatiquement à partir du corpus monolingue AFP
- Outil de prétraitement de l'arabe SAPA

- Chaîne complète de traitement pour traduire une dépêche AFP en format NewsML.

7.5 Autres évaluations

Nous avons participé régulièrement à des évaluations annuelles de traduction de l'arabe vers le français dans le cadre du projet Quaero[7]. Le programme Quaero est fondé sur un consortium public-privé de 32 partenaires franco-allemands, composé de groupes industriels, de PME (Petite et Moyenne Entreprises) et de laboratoires et d'institutions publiques. Il vise à développer des technologies de traitement automatique des contenus multimédia et multilingues permettant d'offrir de nouveaux produits et services au grand public comme aux professionnels. Dans le cadre du projet Quaero, le LIMSI participe à des évaluations annuelles pour la tâche de traduction du texte (WP 4.1).

Nous avons participé aux évaluations Quaero pour la tâche de traduction de texte arabe-français pour les années 2010, 2011 et 2012. Le tableau 7.4 montre les résultats obtenus par le LIMSI pour les systèmes P2, P3, P4 et P5 construits respectivement en 2009, 2010, 2011 et 2012. Les scores ont été évalués avec les métriques BLEU et TER sur le même corpus de test 2012 (livré par Quaero).

Système ar-fr	BLEU	TER
LIMSI_P2	45,3	54,5
LIMSI_P3	46,1	52,9
LIMSI_P4	47,4	52,1
LIMSI_P5	49,3	52,4

Tableau 7.4: *Résultats de traduction obtenus lors de la campagne d'évaluation Quaero pour la tâche de traduction de l'arabe vers le français sur le test 2012.*

Le système LIMSI_P2 est entraîné sur un corpus de débats politiques constitué de 7,6M de phrases, avec les prétraitements MADA. Ce système est le premier système de base arabe-français du LIMSI construit avant l'intégration des travaux sur l'arabe présentés dans cette thèse. Le système LIMSI_P3 est le premier système que nous avons construit. Il est entraîné sur les mêmes données mais en utilisant les pré-traitements de MADA avec la normalisation. Pour le système LIMSI_P4, nous avons reçu dans le cadre du programme Quaero beaucoup de données provenant de l'ONU, donc un corpus parallèle constitué de 14 millions de phrases. Ce corpus a été filtré pour obtenir un système de traduction constitué de trois modèles de traduction : données journalistiques (Projet Syndicate, 37K), ONU (7M) et traduit automatiquement (1M). SOUL (Le et al., 2011; Le et al., 2013) qui représente un nouveau modèle de langue basé sur les réseaux neuronaux a été également appliqué sur ce système. Finalement, le système LIMSI_P5 a été entraîné sur des données de type news constitué de deux modèles de traduction contenant des données journalistiques : Project Syndicate (59K) et AFP (234K), prétraité avec SAPA, et pour lequel nous avons appliqué SOUL.

Finalement, on note que les performances du système de traduction de l'arabe vers le français du LIMSI ont été régulièrement améliorées et au final de 4 points BLEU par rapport à la première version.

7.6 Conclusion

Depuis le début du projet SAMAR, nous recevons quotidiennement des dépêches AFP en arabe, français et anglais. Ces dépêches sont stockées sur nos disques, et sont traduites

[7] http://www.quaero.org/

systématiquement de l'arabe vers le français. 176 463 dépêches (= 2,6M phrases) ont été donc traduites depuis le début du projet.

Il est à noter que nous étions menés à rédiger des rapports livrables pour le projet, et à fournir des corpus parallèles, des systèmes de traduction, ainsi que la chaîne complète de traitement et traduction d'une dépêche. Le système de traduction arabe-français *SAMAR-4* a été livré au projet en septembre 2012 pour son intégration sur la plateforme de démonstration. Le système de traduction arabe-anglais *SAMAR-AR:EN* a été également livré au projet SAMAR en novembre 2012.

Afin de pouvoir construire les systèmes de traduction de l'arabe vers le français et de l'arabe vers l'anglais, nous avons développé deux méthodes différentes pour construire les corpus parallèles nécessaires pour cela.

Au début du projet SAMAR, il était prévu d'utiliser un outil de prétraitement de l'arabe fourni par l'un des partenaires du projet. Cet outil n'a pas pu être utilisé pour plusieurs raisons dont les principales est que l'outil fonctionne exclusivement sur Windows et il est trop lent. Nous avons donc développé notre propre outil de prétraitement de l'arabe SAPA. Le développement de SAPA n'était pas prévu au début du projet, mais s'est avéré nécessaire pour pouvoir améliorer notre chaîne de prétraitement.

Nous avons également développé notre propre outil de détection des entités nommées, qui n'était pas non plus parmi les tâches que le LIMSI devrait réaliser. Ceci est principalement dû au fait que l'outil de Temis tourne exclusivement sur Windows.

La plateforme SAMAR est en premier lieu destinée aux journalistes travaillant en langue arabe pour la gestion de contenus multimédia à destination de la presse et des médias du Proche-Orient et du Maghreb. Elle permet de produire des contenus ou d'enrichir des contenus existants en langue arabe et de les diffuser vers des sites de présentation aux utilisateurs finaux de l'information. Le système SAMAR permet à des non arabophones d'avoir une idée, parfois approximative, du contenu des documents traduits en français et en anglais.

Conclusion générale

Ce dernier chapitre résume brièvement nos contributions et discute des résultats obtenus à partir de ces contributions. Nous décrivons également plusieurs perspectives envisageables pour poursuivre ces travaux.

Contributions

Tout au long de cette thèse, nous avons travaillé sur la langue arabe. Le point de départ de mes recherches a consisté à collectionner tout d'abord les ressources disponibles sur l'arabe au LIMSI, dans le cadre du projet SAMAR et en libre service sur Internet. L'objectif en premier lieu était de construire un corpus parallèle arabe-français pour pouvoir construire un système de traduction de base permettant de traduire les dépêches de l'AFP. Pour cela nous avons développé une méthode pour extraire un corpus parallèle à partir du corpus comparable arabe-français issu des dépêches journalistiques de l'AFP. Ensuite, nous avons proposé différentes approches d'adaptation du modèle de traduction (i) en utilisant le corpus parallèle extrait automatiquement ou (ii) en utilisant un corpus parallèle construit automatiquement – en traduisant un corpus monolingue avec un système de traduction *hors-domaine*. Notre première contribution a donc consisté à explorer un corpus comparable afin d'adapter un modèle de traduction (chapitre 3). Nous avons montré que l'adaptation améliore les performances de traduction avec une hausse du score BLEU jusqu'à 6 points.

Lors de nos premiers travaux sur la construction et l'adaptation de corpus parallèle, nous avons observé que la construction d'un système de traduction, et plus particulièrement le prétraitement de l'arabe, est lent et nécessite l'utilisation de beaucoup de ressources (si la taille de données dépasse quelque milliers de phrases). Nous avons donc développé notre propre outil de prétraitement de l'arabe, SAPA[8] dans le but de réduire ces temps de prétraitements et d'alléger nos traitements. Ce dernier a la particularité d'être (i) beaucoup plus rapide que les outils de prétraitement de l'arabe les plus utilisés et (ii) indépendant de toute autre ressource d'analyse morphologique et de désambiguïsation (chapitre 4). SAPA prédit simultanément l'étiquette morphosyntaxique et les proclitiques d'un mot s'ils existent. Nous avons remarqué que la prédiction simultanée (de l'étiquette morphosyntaxique et des proclitiques) fait baisser le taux d'erreur de la prédiction des étiquettes morphosyntaxiques qui passe de 5,3 % à 4,2 %.

Ces résultats encourageants nous ont motivé pour compléter la chaîne de prétraitement de l'arabe (segmentation) avec la reconnaissance des EN, afin de ne pas segmenter les entités nommées et essayer d'améliorer leur traduction. D'après nos évaluations et celles de Habash (2008), entre 25 % et 40 % des mots hors vocabulaires sont des entités nommées. Nous avons donc développé un outil de détection des entités nommées en arabe, NERAr[9] (chapitre 5). Nous avons montré qu'en utilisant NERAr on réduit faiblement le taux de mots hors vocabulaire. Bien que l'amélioration soit faible, cette étude reste intéressante puisqu'elle présente une étude complète sur les entités nommées de l'arabe : depuis la détection des

[8]Téléchargeable sur Internet : `https://github.com/SouhirG/SAPA`
[9]Téléchargeable sur Internet : `https://github.com/SouhirG/NERAr`

entités nommées jusqu'à leur traduction. L'impact de la traduction des entités nommées a été étudié sur deux corpus différents.

Une étude sur l'adaptation thématique des modèles de traduction et de langue a été réalisée dans le chapitre 6. Deux approches de classification des phrases ont été proposées pour traduire chaque phrase avec le système approprié à la catégorie détectée : une approche de classification a priori et une approche de classification automatique à base de classification naïve bayésienne itérative. La particularité de notre approche d'adaptation est qu'elle est extraite d'une application réelle contenant un corpus constitué d'un ensemble de phrases multicatégoriques.

Finalement, des systèmes de traduction arabe-français et arabe-anglais ont été intégrés dans une plateforme d'analyse multimédia (chapitre 7). Nous avons présenté toute la chaîne de traitement (pré-traitement et post-traitement) que nous avons développée pour traduire une dépêche en format NewsML.

Le projet SAMAR m'a permis d'avoir une idée du rôle de la traduction automatique dans une application réelle et d'intéragir avec plusieurs partenaires dont chacun a produit ses modules technologiques qui ont été intégrés dans l'application.

Le projet SAMAR était une bonne opportunité pour voir de plus près le déroulement d'un projet avec toutes ses phases. Ce fut une expérience enrichissante scientifiquement et industriellement qui m'a permis de réaliser une thèse dans un laboratoire de recherche tout en ayant une vue sur le monde extérieur des entreprises.

Perspectives

Plusieurs possibilités s'offrent pour poursuivre nos travaux.

Perspectives à court-terme

Dans le chapitre 5, nous avons effectué une étude complète des entités nommées depuis la construction d'un outil de détection des EN jusqu'à leur traduction. Afin de compléter cette chaîne de prétraitement des entités nommées, il serait intéressant de créer notre propre outil de translittération de l'arabe vers le français et/ou vers l'anglais. Cette tâche nécessitera tout d'abord un corpus d'entraînement, ainsi qu'une réflexion sur les bases de translittération des noms propres de l'arabe. Certains travaux ont été réalisés dans ce cadre comme ceux de Saadane et al. (2012) par exemple. Nous avons également constaté lors de la phase d'alignement que l'ordre des mots entre l'arabe et le français varie.

Carpuat, Marton et Habash (2010) montrent qu'en arabe les sujets qui apparaîssent après les verbes (VS) sont difficiles à traduire vers l'anglais. Ils proposent donc de réordonner les phrases verbales VS en phrases nominales SV seulement pour la phase d'alignement de mots. Ils constatent que 98 % des SV sont traduits dans un ordre monotone, 64,7 % des VS sont traduits dans un ordre inversé, tandis que 27,3 % sont traduits dans un ordre monotone. Cette stratégie améliore les performances de traduction (+0,3 points BLEU) et serait intéressante à explorer surtout dans le cas où les corpus à traduire ont un style particulier différent du corpus d'entraînement (comme le cas du corpus Arcade II dans le chapitre 5).

Les résultats du chapitre 6 pourraient être améliorés avec un corpus d'entraînement plus volumineux. Plusieurs perspectives sont envisageables pour ces travaux, en particulier sur l'amélioration du classifieur automatique avec des modèles de classification plus volumineux. Une perpective intéressante serait de pondérer les phrases à traduire au moment de la traduction et de combiner plusieurs systèmes lors de la traduction. Ceci permettrait de garder toutes les catégories attribuées à chaque phrase pour le corpus de test à traduire.

Perspectives à plus long-terme

Notre étude principale a en effet porté sur la traduction automatique à partir de l'arabe. Il serait intéressant d'étudier cet aspect dans le sens inverse, c'est-à-dire étudier la traduction automatique vers l'arabe.

Nos tentatives d'amélioration des prétraitements de l'arabe ont révélé qu'il est possible d'avoir de bons résultats pour la prédiction des étiquettes morphosyntaxiques, la segmentation ainsi que le prétraitement pour la tâche de traduction. Il serait donc intéressant de compléter SAPA afin de l'adapter pour les post-traitements pour la traduction, non pas depuis mais vers l'arabe.

Un autre sujet intéressant à aborder avec ce type de données est l'adaptation temporelle. Il serait intéressant d'étudier l'aspect d'ajout progressif des données pour construire des systèmes de traduction adaptés à des données récentes. Comme le montre la figure 5.1 dans le chapitre 5, toutes les semaines voire tous les jours de nouvelles entités nommées apparaissent. En dehors des EN, le vocabulaire s'enrichit constamment par l'ajout de nouveaux termes comme c'est le cas avec les mots récemment parus dans les nouvelles technologies comme par exemple le mot *wifi* ou *facebook*.

Quelques travaux sur l'adaptation temporelle de modèles de langue ont été effectués dans le domaine de la reconnaissance de la parole, comme ceux de (Whittaker, 2001) qui adaptent les modèles de langue en utilisant des données datant de la même période de temps que les données à traiter.

Allauzen (2003) propose une adapatation *ad-hoc* basée sur la mise à jour régulière du vocabulaire utilisé pour créer des modèles de langue temporellement proches des documents à transcrire. Levenberg et Osborne (2009) montrent que pour une traduction à flux courant (*stream-based translation*), l'utilisation de données récentes améliore les performances. Ils proposent un modèle de langue aléatoire (*randomised language model*) basé sur le hashage dynamique. Ce nouveau modèle de langue est plus rapide et plus efficace que les modèles de langue classiques et a la capacité de s'adapter aux données et textes publiés sur le web quotidiennement.

Quelques résultats préliminaires sur l'adaptation temporelle et l'impact de l'ajout de données en fonction du temps sont présentés par la figure 1 qui montre l'évolution du nombre de mots hors vocabulaire et du score BLEU de traduction au fur et à mesure de l'ajout du vocabulaire pour chaque semaine.

Neuf systèmes de traduction avec des modèles de traduction différents ont été construits en utilisant le corpus présenté dans le chapitre 3. Le premier système au point j1 est constitué avec des données à partir de décembre 2009 jusqu'au dernier jour de l'avant dernière semaine du mois de novembre 2010. Les systèmes j2 jusqu'à j9 ont été construits en ajoutant à chaque fois les données parallèles extraites pour le jour suivant.

Le premier système au point j1 est constitué de 141 757 phrases. À l'état initial (j1), le nombre de mots uniques dans le vocabulaire est 74 234. Une semaine après (j9), le nombre de mots uniques passe à 75 049, par un corpus d'entraînement de taille 145 192.

On observe que le nombre de mots hors-vocabulaire est réduit de 50 mots au bout d'une semaine avec parallèlement un score BLEU qui augmente pour les modèles de plus en plus à jour. L'étude de l'adaptation de modèles de traduction et de modèles de langues adaptés temporellement serait intéressante. Allauzen (2003) par exemple créé des modèles de langue adaptés temporellement et qui contiennent toujours la même quantité de vocabulaire : le plus ancien vocabulaire est remplacé par le récent et ainsi de suite (puisque les modèles de langue pour la reconnaissance de la parole ont une taille limitée).

Nous n'avons pas eu le temps pour finir nos expériences et de développer complètement les idées que nous avions l'intention d'explorer. Ces premiers résultats montrent que l'aspect temporel est important et influe sur les performances de traduction. Bien que ces premiers résultats soient positifs, les performances pourraient être améliorées sur différents aspects

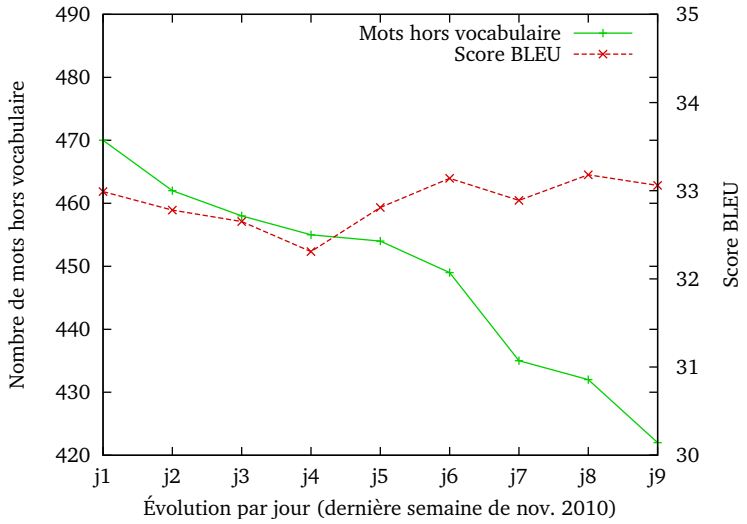

Figure 1: *Évolution du nombre de mots hors vocabulaire et du score BLEU au cours du temps pour chaque jour (dernière semaine de novembre 2010). Les scores présentés sont évalués sur le corpus de test 2010 (décembre) utilisé tout au long des travaux de ce manuscrit.*

comme par exemple l'exploration de différentes adaptations de modèles de langue et de traduction (comme dans le chapitre 6) au niveau temporel.

L'adaptation dynamique de modèles de langue et de traduction est une perspective intéressante pour améliorer et adapter les systèmes de traduction pour traduire systématiquement des dépêches journalistiques de l'AFP (chose que nous faisons depuis le début du projet SAMAR mais sans adaptation dynamique).

Liste des abréviations

- AFP: Agence France-Presse
- ATB: Arabic Tree Bank
- BAMA: Buckwalter Arabic Morphological Analyzer
- BLEU: Bilingual Evaluation Understudy
- ELDA: Evaluations and Language resources Distribution Agency
- ELRA: European Language Resources Association
- EN: Entités Nommées
- GALE: Global Autonomous Language Exploitation
- IPTC: International Press Telecommunications Council
- LDC: Linguistic Data Consortium
- LM: Language Model
- MERT: Mininum Error Rate Training
- MIRA: Margin Infused Relaxed Algorithm
- MT: Machine Translation
- NERAr: Named Entity Recognition for Arabic
- NewsML: News Markup Language
- NIST: National Institute of Standards and Technology
- NLP: Natural Language Processing
- POS: Part-Of-Speech Tagging
- SAMA: Standard Arabic Morphological Analyzer
- SAMAR: Station d'Analyse Multimédia pour l'Arabe
- SAPA: Segmentor and Part-of-Speech Tagger for Arabic
- SMT: Stastical Machine Translation
- TAL: Traitement Automatique des Langues
- TALN: Traitement Automatique du Langage Naturel
- TM: Translation Model
- WER: Word Error Rate
- XML: Extensible Markup Language

Dépêche AFP en arabe en format NewsML

La figure 2 montre un exemple d'une dépêche AFP. Le format NewsML inclut beaucoup d'informations concernant chaque dépêche : date et heure d'apparition, la langue de la dépêche, la ou les catégories IPTC auxquelles appartient la dépêche, le lieu où la dépêche a été écrite, les mots clés, l'identifiant de la dépêche, le nombre de mots, la priorité de la dépêche.

La plupart de ces informations sont utilisées en interne par l'AFP. Toutes ces informations sont concervées après la traduction automatique par nos systèmes de traduction automatique arabe-français et arabe-anglais.

La dépêche ci-dessous a été publiée le 10 mai 2012 à 16h12 et 10 secondes (20120510T161210Z). Elle est multicatégorique et appartient aux catégories politique (11011000) et guerres (16003000).

Figure 2: *Exemple d'une dépêche AFP en arabe en format NewsML.*

Dépêche AFP en français en format NewsML

La dépêche ci-dessous a été publiée le 10 mai 2012 à 06h31 et 49 secondes (20120510T063149Z). Elle est monocatégorique et appartient à la catégorie finances (04000000, 04008000, 04008004). Les mots clés qui ont été choisis par l'auteur (uh) sont *Economie, indicateur, croissance*.

```
<?xml version="1.0" encoding="utf-8" ?>
- <NewsML Version="1.2">
    <!-- AFP NewsML text-photo profile evolution2  -->
    <!-- Processed by Xafp1-4ToNewsML1-2 rev18  -->
    <Catalog Href="http://www.afp.com/dtd/AFPCatalog.xml" />
  - <NewsEnvelope>
      <TransmissionId>0678</TransmissionId>
      <DateAndTime>20120510T063152Z</DateAndTime>
      <NewsService FormalName="DGTE" />
      <NewsProduct FormalName="FRS" />
      <NewsProduct FormalName="FIL" />
      <NewsProduct FormalName="FRA" />
      <NewsProduct FormalName="DVBA" />
      <NewsProduct FormalName="DAGI" />
      <NewsProduct FormalName="DPSE" />
      <NewsProduct FormalName="DILI" />
      <NewsProduct FormalName="DVBP" />
      <NewsProduct FormalName="DGTE" />
      <NewsProduct FormalName="DGTI" />
      <Priority FormalName="3" />
    </NewsEnvelope>
  - <NewsItem xml:lang="fr">
    - <Identification>
      - <NewsIdentifier>
          <ProviderId>afp.com</ProviderId>
          <DateId>20120510T063149Z</DateId>
          <NewsItemId>TX-PAR-TKS09</NewsItemId>
          <RevisionId PreviousRevision="0" Update="N">1</RevisionId>
          <PublicIdentifier>urn:newsml:afp.com:20120510T063149Z:TX-PAR-
            TKS09:1</PublicIdentifier>
        </NewsIdentifier>
        <NameLabel>Economie-indicateur-croissance</NameLabel>
      </Identification>
    - <NewsManagement>
        <NewsItemType FormalName="News" />
        <FirstCreated>20120510T063149Z</FirstCreated>
        <ThisRevisionCreated>20120510T063149Z</ThisRevisionCreated>
        <Status FormalName="Usable" />
        <Urgency FormalName="3" />
        <DerivedFrom NewsItem="urn:newsml:afp.com:20120510T063010Z:TX-PAR-TKR91" />
      </NewsManagement>
    - <NewsComponent>
      - <NewsLines>
          <DateLine xml:lang="fr">PARIS, 10 mai 2012 (AFP) -</DateLine>
          <HeadLine xml:lang="fr">La Banque de France prévoit une croissance nulle au 2e
            trimestre 2012</HeadLine>
        </NewsLines>
      - <AdministrativeMetadata>
        - <Provider>
            <Party FormalName="AFP" />
          </Provider>
        </AdministrativeMetadata>
      - <DescriptiveMetadata>
          <Language FormalName="fr" />
        - <SubjectCode>
            <Subject FormalName="04000000" />
          </SubjectCode>
        - <SubjectCode>
            <SubjectMatter FormalName="04008000" />
          </SubjectCode>
        - <SubjectCode>
            <SubjectDetail FormalName="04008004" />
          </SubjectCode>
          <OfInterestTo FormalName="FRF-TFG-1=FAF" />
          <DateLineDate>20120510T082137+0200</DateLineDate>
        - <Location HowPresent="Origin">
            <Property FormalName="Country" Value="FRA" />
            <Property FormalName="City" Value="PARIS" />
          </Location>
          <Property FormalName="Keyword" Value="Economie" />
          <Property FormalName="Keyword" Value="indicateur" />
          <Property FormalName="Keyword" Value="croissance" />
          <Property FormalName="GeneratorSoftware" Value="Cafp32" />
        </DescriptiveMetadata>
      - <ContentItem>
          <MediaType FormalName="Text" />
          <Format FormalName="NITF3.1" />
        - <Characteristics>
            <SizeInBytes>347</SizeInBytes>
            <Property FormalName="Words" Value="57" />
          </Characteristics>
        - <DataContent>
          - <nitf>
            - <body>
              - <body.content>
                  <p>La Banque de France prévoit une croissance nulle de l'économie
                    française au deuxième trimestre 2012, selon sa première
                    estimation annoncée jeudi dans un communiqué.</p>
                  <p>Selon la Banque de France, le Produit intérieur brut (PIB) de la
                    France devrait rester "stable" au deuxième trimestre par rapport
                    aux trois premiers mois de l'année.</p>
                  <p>uh/fz/bfa</p>
                </body.content>
              </body>
            </nitf>
          </DataContent>
        </ContentItem>
      </NewsComponent>
    </NewsItem>
  </NewsML>
</NewsML>
```

Figure 3: *Exemple d'une dépêche AFP en français en format NewsML.*

Scores de traduction des modèles spécifiques sur le test général

Le tableau 1 montre les scores BLEU du test général traduit en utilisant les différentes combinaisons des modèles spécifiques et généraux pour chaque catégorie.

	BLEU (dev Gén)			BLEU (dev Spé)		
Catégorie / Système	POL	WAR	FIN	POL	WAR	FIN
Général		**33,47**		33,02	33,22	32,95
LM-spécifique	32,88	31,93	30,88	32,95	32,18	30,91
TM-spécifique	32,26	31,26	19,07	32,13	30,93	28,56
Spécifique	31,54	29,64	26,34	31,91	29,76	26,30

Tableau 1: *Traduction automatique du test général en utilisant les systèmes de traduction de base appropriés aux catégories POL, WAR et FIN.*

Lorsque le corpus de développement général est utilisé, on note que le meilleur score est obtenu en utilisant le système de traduction général. Dans les scores suivants, le deuxième et troisième meilleurs scores sont obtenus en utilisant les modèles spécifiques de la catégorie POL puisque cette dernière représente 60% de l'ensemble des données. Le plus mauvais score est obtenu avec le système TM-spécifique pour la catégorie FIN (19,07) puisque les modèles FIN sont les plus petits. De plus, seulement 20% des phrases du corpus de test général appartiennent à la catégorie FIN.

En utilisant des modèles spécifiques, on utilise de manière sous optimale les données disponibles pour les volumes de données considérées.

Publications de l'auteure

2012

- Souhir Gahbiche-Braham, Hélène Bonneau-Maynard, Thomas Lavergne, François Yvon. Repérage des entités nommées pour l'arabe : adaptation non-supervisée et combinaison de systèmes. *Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN.* Grenoble, France, 04-07 Juin, 2012.
- Souhir Gahbiche-Braham, Hélène Bonneau-Maynard, Thomas Lavergne, François Yvon. Joint Segmentation and POS Tagging for Arabic Using a CRF-based Classifier. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12).* Istanbul, Turkey, May 23-25, 2012.

2011

- Souhir Gahbiche-Braham, Hélène Bonneau-Maynard, François Yvon. Two Ways to Use a Noisy Parallel News Corpus for Improving Statistical Machine Translation. *Proceedings of the 4th ACL Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web* Portland, Oregon. June, 2011.

Autres publications

Outils

- SAPA, *Segmentor and Part-of-speech tagger for Arabic*, https://github.com/SouhirG/SAPA
- NERAr, *Named Entity Recognizer for Arabic*, https://github.com/SouhirG/NERAr

Rapports livrables pour le projet SAMAR

- Wigdan Mekki, Souhir Gahbiche-Braham, Yves Lepage, Franccois Yvon, Hélène Maynard. Livrable 7.1, Ressources nécessaires à l'amélioration des modèles de traduction automatique.
- Adrien Lardilleux, Souhir Gahbiche-Braham, Hélène Bonneau-Maynard, François Yvon. Livrable 7.2, Description du système de traduction *baseline*.
- Souhir Gahbiche-Braham, Hélène Bonneau-Maynard, François Yvon. Livrable 7.4, Description des systèmes de traduction améliorés.

Présentations pour le projet SAMAR

- Bilan d'étape numéro 1, 15 octobre 2010, Paris.
- Bilan d'étape numéro 2, 21 novembre 2011, Paris.
- Bilan d'étape numéro 3, 20 novembre 2012, Paris.

Apparition dans la presse

- Pierre Vandeginste. *Ils créent des programmes sur mesure*. Les dossiers de la recherche, Edition N4, Juin-Juillet 2013, p92.

Bibliographie

Abdul Hamid, Ahmed and Kareem Darwish (2010). "Simplified Feature Set for Arabic Named Entity Recognition". In: *Proc. of the 2010 Named Entities Workshop*. Uppsala, pp. 110–115. URL: http://www.aclweb.org/anthology/W10-2417 (*Cité pages* 30, 65).

Abdul Rauf, Sadaf (2012). "Efficient corpus selection for Statistical Machine Translation". PhD thesis. Le Mans, France (*Cité page* 67).

Abdul-Rauf, Sadaf and Holger Schwenk (2009). "On the use of comparable corpora to improve SMT performance". In: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. EACL '09. Athens, Greece: Association for Computational Linguistics, pp. 16–23. URL: http://portal.acm.org/citation.cfm?id=1609067.1609068 (*Cité page* 37).

Ahmed, Farag and Andreas Nürnberger (2008). "Arabic/English Word Translation Disambiguation using Parallel Corpora and Matching Schemes". In: *Proceedings of EAMT'08*. EAMT '08. Hamburg, Germany (*Cité page* 28).

Aker, Ahmet, Yang Feng, and Robert J. Gaizauskas (2012). "Automatic Bilingual Phrase Extraction from Comparable Corpora". In: *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference*. Ed. by Martin Kay and Christian Boitet. Mumbai, India, pp. 23–32 (*Cité page* 47).

Al-Haj, Hassan and Alon Lavie (2012). "The impact of Arabic morphological segmentation on broad-coverage English-to-Arabic statistical machine translation". In: *Machine Translation* 26.1-2, pp. 3–24 (*Cité page* 30).

Al-Jumaily, Harith T., Paloma Martínez, José Luis Martínez-Fernández, and Erik Van der Goot (2012). "A real time Named Entity Recognition system for Arabic text mining". In: *Language Resources and Evaluation* 46.4, pp. 543–563 (*Cité page* 65).

Al-Onaizan, Yaser and Kevin Knight (2002a). "Machine Transliteration of Names in Arabic Text". In: *In ACL Workshop on Comp. Approaches to Semitic Languages*, pp. 34–46 (*Cité pages* 27, 30, 66).

Al-Onaizan, Yaser and Kevin Knight (2002b). "Translating named entities using monolingual and bilingual resources". In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. ACL '02. Philadelphia, Pennsylvania: Association for Computational Linguistics, pp. 400–408. DOI: 10.3115/1073083.1073150. URL: http://dx.doi.org/10.3115/1073083.1073150 (*Cité page* 66).

Al-Sughaiyer, Imad A. and Ibrahim A. Al-Kharashi (2004). "Arabic morphological analysis techniques: A comprehensive survey". In: *JASIST* 55.3, pp. 189–213 (*Cité page* 50).

Allauzen, Alexandre (2003). "Modélisation linguistique pour l'indexation automatique de documents audiovisuels". PhD thesis. Université Paris Sud, Orsay (*Cité pages* 66, 121).

Allauzen, Alexandre and François Yvon (2012). "Textual Information Access". In: *Statistical Methods for Machine Translation*. Ed. by Eric Gaussier and François Yvon. ISTE/Wiley, Paris. Chap. 7, pp. 223–304 (*Cité pages* 4, 9).

Alotaiby, Fahad, Ibrahim Alkharashi, and Salah Foda (2009). "Processing Large Arabic Text Corpora: Preliminary Analysis and Results". In: *Proceedings of the Second International Conference on Arabic Language Resources and Tools*. Ed. by Khalid Choukri and Bente Maegaard. Cairo, Egypt: The MEDAR Consortium. ISBN: 2-9517408-5-9 (*Cité page* 30).

Attia, Mohammed A. (2008). "Handling Arabic morphological and syntactic ambiguities within the LFG framework with a view to machine translation". PhD thesis. University of Manchester (*Cité pages* xvi, 20, 22, 30, 50).

Attia, Mohammed (2012). *Ambiguity In Arabic Computational Morphology And Syntax: A Study within the Lexical Functional Grammar Framework*. LAP Lambert Academic Publishing. ISBN: 978-3-8484-4967-5 (*Cité page 20*).

Axelrod, Amittai, Xiaodong He, and Jianfeng Gao (2011). "Domain adaptation via pseudo in-domain data selection". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '11. Edinburgh, United Kingdom: Association for Computational Linguistics, pp. 355–362. ISBN: 978-1-937284-11-4. URL: http://dl.acm.org/citation.cfm?id=2145432.2145474 (*Cité page 90*).

Babych, Bogdan and Anthony Hartley (2003). "Improving machine translation quality with automatic named entity recognition". In: *Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools: Resources and Tools for Building MT*. EAMT '03. Budapest, Hungary: Association for Computational Linguistics, pp. 1–8 (*Cité page 30*).

Baloul, Sofiane and Philippe Boula de Mareüil (2002). "Un modèle syntactico-prosodique pour la synthèse de la parole à partir du texte en arabe standard voyellé". In: *7ème Conference Maghrebine sur les sciences informatiques*. Annaba (Algérie) (*Cité page 28*).

Banerjee, Pratyush (2012). "Domain Adaptation for Statistical Machine Translation of Corporate and User-Generated Content". Way, Andy and Van Genabith, Josef and Roturier, Johann. PhD thesis. Dublin City University (*Cité pages 87, 90, 104*).

Banerjee, Satanjeev and Alon Lavie (2005). "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments". In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation*. Ann Arbor, Michigan, pp. 65–72 (*Cité pages 12, 57, 78*).

Bar-haim, Roy and Yoad Winter (2005). "Choosing an optimal architecture for segmentation and POS-tagging of Modern Hebrew". In: *In Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, pp. 39–46 (*Cité page 51*).

Béchet, Frédéric, Benoît Sagot, and Rosa Stern (2011). "Coopération de méthodes statistiques et symboliques pour l'adaptation non-supervisée d'un système d'étiquetage en entités nommées". In: *actes de la conférence TALN*. Montpellier, France. URL: http://hal.inria.fr/inria-00617068 (*Cité page 66*).

Bellagarda, Jérôme R. (2001). "An overview of statistical language model adaptation". In: *Proceedings of the ISCA Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition*. Sophia Antipolis, France, pp. 165–174 (*Cité pages 37, 66*).

Benajiba, Yassine, Mona Diab, and Paolo Rosso (2008). "Arabic named entity recognition using optimized feature sets". In: *Proc. of EMNLP*. EMNLP. Honolulu, Hawaii, pp. 284–293. URL: http://dl.acm.org/citation.cfm?id=1613715.1613755 (*Cité pages 30, 65*).

Benajiba, Yassine and Paolo Rosso (2007). "Conquering the NER task for the Arabic language by combining the Maximum Entropy with POS-tag information". In: *Proceedings of Workshop on Natural Language-Independant Engineering*. IJCAI (*Cité pages 29, 65*).

Benajiba, Yassine and Paolo Rosso (2008). "Arabic named entity recognition using Conditional Random Fields". In: *Proceedings of the Conference on Language Resources and Evaluation*. Marrakech, Marroco (*Cité pages 30, 65, 74*).

Benajiba, Yassine, Paolo Rosso, and José-Miguel Benedí (2007). "ANERsys: An Arabic Named Entity Recognition System Based on Maximum Entropy". In: *CICLing*, pp. 143–153 (*Cité pages 25, 29, 68*).

Bies, Ann, Denise DiPersio, and Mohamed Maamouri (2012). "Challenges for Arabic Machine Translation". In: ed. by Abdelhadi Soudi, Ali Farghaly, Günter Neumann, and Rabih Zbib. Natural language processing. Amsterdam ; Philadelphia: John Benjamins Pub. Co. Chap. Linguistic resources for Arabic machine translation, pp. 15–22 (*Cité page 23*).

Bisazza, Arianna, Nick Ruiz, Marcello Federico, and FBK-Fondazione Bruno Kessler (2011). "Fill-up versus Interpolation Methods for Phrase-based SMT Adaptation". In: *International Workshop on Spoken Language Translation (IWSLT)* (*Cité page 89*).

Blitzer, John (2008). "Domain adaptation of natural language processing systems". Adviser-Pereira, Fernando. PhD thesis. Philadelphia, PA, USA. ISBN: 978-0-549-57740-9 (*Cité page 66*).

Bouamor, Dhouha, Nasredine Semmar, and Pierre Zweigenbaum (2013). "Context Vector Disambiguation for Bilingual Lexicon Extraction from Comparable Corpora". In: *ACL*. Sofia, Bulgaria (*Cité page 35*).

Broder, Andrei Z. (2000). "Identifying and Filtering Near-Duplicate Documents". In: *Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching*. COM '00. London, UK: Springer-Verlag, pp. 1–10 (*Cité page* 37).

Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin (06/1990). "A statistical approach to machine translation". In: *Comput. Linguist.* 16.2, pp. 79–85. ISSN: 0891-2017. URL: http://dl.acm.org/citation.cfm?id=92858.92860 (*Cité pages* 4, 5).

Brown, Peter F., Vincent J.Della Pietra, Stephen A. Della Pietra, and Robert. L. Mercer (1993). "The Mathematics of Statistical Machine Translation: Parameter Estimation". In: *Computational Linguistics* 19, pp. 263–311 (*Cité pages* 5, 6).

Buckwalter, Tim (2002). *Buckwalter Arabic Morphological Analyzer Version 1.0*. Catalog No. LDC2002L49. University of Pennsylvania: Linguistic Data Consortium. ISBN: 1-58563-257-0 (*Cité pages* 23, 43).

Buckwalter, Tim (2004). *Buckwalter Arabic Morphological Analyzer Version 2.0*. Catalog No. LDC2004L02. University of Pennsylvania: Linguistic Data Consortium. ISBN: 1-58563-324-0 (*Cité pages* 23, 50, 51).

Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Josh Schroeder (2009). "Findings of the 2009 workshop on statistical machine translation". In: *Proceedings of the Fourth Workshop on Statistical Machine Translation*. StatMT '09. Athens, Greece: Association for Computational Linguistics, pp. 1–28 (*Cité page* 11).

Carpuat, Marine, Yuval Marton, and Nizar Habash (2010). "Improving Arabic-to-English statistical machine translation by reordering post-verbal subjects for alignment". In: *Proceedings of the ACL 2010 Conference Short Papers*. ACLShort '10. Uppsala, Sweden: Association for Computational Linguistics, pp. 178–183 (*Cité page* 120).

Carpuat, Marine, Yuval Marton, and Nizar Habash (03/2012). "Improved Arabic-to-English statistical machine translation by reordering post-verbal subjects for word alignment". In: *Machine Translation* 26.1-2, pp. 105–120. ISSN: 0922-6567. URL: http://dx.doi.org/10.1007/s10590-011-9112-y (*Cité page* 31).

Ceausu, Alexandru, John Tinsley, Jian Zhang, and Andy Way (2011). "Experiments on Domain Adaptation for Patent Machine Translation in the PLuTO project". In: *Proceedings of the 15th Annual Conference of the European Association for Machine Translation (EAMT 2011)*. EAMT '11. Leuven, Belgium, pp. 21–28 (*Cité page* 89).

Cettolo, Mauro, Marcello Federico, and Nicola Bertoldi (2010). "Mining Parallel Fragments from Comparable Texts". In: *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*. Ed. by Marcello Federico, Ian Lane, Michael Paul, and François Yvon. Paris, France, pp. 227–234 (*Cité pages* 37, 46).

Che, Wanxiang, Mengqiu Wang, Christopher D. Manning, and Ting Liu (2013). "Named Entity Recognition with Bilingual Constraints". In: *NAACL-HLT* (*Cité page* 65).

Chen, Boxing, George Foster, and Roland Kuhn (2013). "Adaptation of Reordering Models for Statistical Machine Translation". In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Georgia: Association for Computational Linguistics, pp. 938–946. URL: http://www.aclweb.org/anthology/N13-1114 (*Cité page* 105).

Chen, Stanley F. and Joshua Goodman (1996). "An empirical study of smoothing techniques for language modeling". In: *Proceedings of the 34th annual meeting on Association for Computational Linguistics*. ACL '96. Santa Cruz, California: Association for Computational Linguistics, pp. 310–318 (*Cité page* 8).

Chiang, David (2005). "A hierarchical phrase-based model for statistical machine translation". In: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. ACL '05. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 263–270 (*Cité page* 13).

Chiang, David, Kevin Knight, and Wei Wang (2009). "11,001 new features for statistical machine translation". In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. NAACL '09. Boulder, Colorado: Association for Computational Linguistics, pp. 218–226. ISBN: 978-1-932432-41-1 (*Cité page* 10).

Chiao, Yun-chuang, Olivier Kraif, Dominique Laurent, Thi Minh, Huyen Nguyen, Nasredine Semmar, François Stuck, Jean Véronis, and Wajd i Zaghouani (2006). "Evaluation of multilingual text alignment systems: the ARCADE

II project". In: *In Proceedings of LREC2006* (*Cité pages* 24, 68, 69).

Chomsky, Noam (1981). *Lectures on Government and Binding*. Dordrecht, Holland: Foris Publications (*Cité page* 22).

Chung, Wingyan (05/2008). "Web searching in a multilingual world". In: *Commun. ACM* 51.5, pp. 32–40. ISSN: 0001-0782. DOI: 10.1145/1342327.1342335. URL: http://doi.acm.org/10.1145/1342327.1342335 (*Cité page* 15).

Cohen, D. (1970). *Études de linguistique sémitique et arabe*. Janua linguarum: Series practica. Walter de Gruyter GmbH & Co. KG (*Cité page* 26).

Cori, Marcel and Jacqueline Léon (2002). "La constitution du TAL, Étude historique des dénominations et des concepts". In: *TAL* 43.3, pp. 21–56 (*Cité page* xiii).

Crammer, Koby and Yoram Singer (03/2003). "Ultraconservative online algorithms for multiclass problems". In: *J. Mach. Learn. Res.* 3, pp. 951–991. ISSN: 1532-4435 (*Cité page* 10).

Crego, Josep Maria and José B. Mariño (09/2006). "Improving statistical MT by coupling reordering and decoding". In: *Machine Translation* 20.3, pp. 199–215. ISSN: 0922-6567 (*Cité page* 8).

Darwish, Kareem (2002). "Building a Shallow Arabic Morphological Analyser in One Day". In: *Proceedings of the ACL-02 Workshop on Computational Approaches to Semitic Languages*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics. URL: http://www.aclweb.org/anthology/W02-0506 (*Cité page* 27).

Daume III, Hal (2007). "Frustratingly Easy Domain Adaptation". In: *Proc. of the 45th Annual Meeting of the ACL*. Prague, Czech Republic (*Cité page* 66).

Daumé III, Hal and Jagadeesh Jagarlamudi (2011). "Domain Adaptation for Machine Translation by Mining Unseen Words". In: *ACL*. Portland, OR (*Cité page* 63).

Daume III, Hal and Daniel Marcu (2006). "Domain Adaptation for Statistical Classifiers". In: *Journal of Artificial Intelligence Research* 26, pp. 101–126 (*Cité page* 66).

Daume III, Hal, Tejaswini Deoskar, David McClosky, Barbara Plank, and Jörg Tiedemann, eds. (2010). *Proc. of the 2010 Workshop on Domain Adaptation for NLP*. Uppsala, Sweden. URL: http://www.aclweb.org/anthology/W10-26 (*Cité page* 66).

De Mori, Renato and Marcello Federico (1999). "Language Model Adaptation". In: *In K. Ponting, editor, Computational models of speech pattern processing volume 169, NATO ASI*. Ed. by Springer Verlag. Prague, Czech Republic, pp. 280–303 (*Cité page* 89).

Debili, F., H. Achour, and E. Souissi (2002). *De l'étiquetage grammatical à la voyellation automatique de l'arabe*. Tech. rep. vol 71. Correspondances de l'Institut de Recherche sur le Maghreb Contemporain 17 (*Cité page* 28).

Debili, F., Z. Ben Tahar, and E. Souissi (2008). "Automatic versus interactive analysis for the massive vowelization, tagging and lemmatization of Arabic". In: Marrakech, Maroc: LREC (*Cité page* 19).

Debili, Fathi, Zied Ben Tahar, and Emna Souissi (2007). "Analyse automatique vs analyse interactive : un cercle vertueux pour la voyellation, l'étiquetage et la lemmatisation de l'arabe". In: *Traitement Automatique des Langues Naturelles*. Toulouse, France, pp. 347–356 (*Cité page* 19).

Debili, Fathi and Emna Souissi (1998). "Etiquetage grammatical de l'arabe voyellé ou non". In: *Proc. of the Workshop on Computational Approaches to Semitic Languages*. Semitic '98. Montreal, Quebec, Canada, pp. 16–25. URL: http://dl.acm.org/citation.cfm?id=1621753.1621757 (*Cité page* 29).

Dewaele, Jean-Marc (2004). *L'acquisition simultanée de trois langues maternelles : exploration d'un « miracle » linguistique*. Birkbeck College, Université de Londres (*Cité page* xiii).

Diab, Mona T. and Nizar Habash (2012). "Arabic Dialect Processing Tutorial". In: *HLT-NAACL* (*Cité page* 18).

Diab, Mona (2009). "Second Generation Tools (AMIRA 2.0): Fast and Robust Tokenization, POS tagging, and Base Phrase Chunking". In: *Proc. of the Second International Conference on Arabic Language Resources and Tools*. Ed. by Khalid Choukri and Bente Maegaard. Cairo, Egypt: The MEDAR Consortium. ISBN: 2-9517408-5-9 (*Cité pages* 27, 50, 74).

Dice, Lee Raymond (1945). "Measures of the Amount of Ecologic Association Between Species". In: *Ecology* 26.3, pp. 297–302. URL: http://www.jstor.org/pss/1932409 (*Cité page* 40).

Dinarelli, Marco and Sophie Rosset (2011). "Models Cascade for Tree-Structured Named Entity Detection". In: *Proceedings of 5th International Joint Conference on Natural Language Processing*.

Chiang Mai, Thailand: Asian Federation of Natural Language Processing, pp. 1269–1278. URL: http://www.aclweb.org/anthology/I11-1142 (*Cité page 76*).

Djili, Abdelaziz (2011). *L'arabe une langue des défis*. URL: http://www.youtube.com/watch?v=LbMQ4HVOyo4 (*Cité page 15*).

Dorr, Bonnie J, P.W. Jordan, and J.W. Benoit (1998). "A Survey of Current Paradigms in Machine Translation". In: (*Cité page 4*).

Dukes, Kais, Eric Atwell, and Nizar Habash (2013). "Supervised collaboration for syntactic annotation of Quranic Arabic". In: *Language Resources and Evaluation* 47.1, pp. 33–62 (*Cité page 26*).

Durgar El-Kahlout, Ilknur and François Yvon (2010). "The pay-offs of preprocessing for German-English Statistical Machine Translation". In: *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*. Ed. by Marcello Federico, Ian Lane, Michael Paul, and François Yvon. Paris, France, pp. 251–258 (*Cité page 30*).

El Isbihani, Anas, Shahram Khadivi, Oliver Bender, and Hermann Ney (2006). "Morpho-syntactic Arabic Preprocessing for Arabic to English Statistical Machine Translation". In: *Human Language Technology Conf. / North American Chapter of the ACL Annual Meeting (HLT-NAACL), Workshop on SMT*. New York: Association for Computational Linguistics, pp. 15–22. URL: http://www.aclweb.org/anthology/W/W06/W06-3103 (*Cité pages 28, 50, 51, 53*).

El Kassas, Dina and Sylvain Kahane (2004). "Modélisation de l'ordre des mots en arabe standard". In: *JEP-TALN 2004*. Fès (*Cité page 28*).

El Kholy, Ahmed and Nizar Habash (2012). "Orthographic and morphological processing for English-Arabic statistical machine translation". In: *Machine Translation* 26.1-2, pp. 25–45 (*Cité pages 31, 61*).

Farghaly, Ali (2010). "Introduction in Arabic computational linguistics". In: (*Cité page 16*).

Farghaly, Ali and Khaled Shaalan (12/2009). "Arabic Natural Language Processing: Challenges and Solutions". In: 8.4, 14:1–14:22. ISSN: 1530-0226 (*Cité pages xvi, 15, 30, 50, 56*).

Federico, Marcello, Fondazione Bruno, Kessler Irst, and Mauro Cettolo (2007). "Efficient Handling of N-gram Language Models for Statistical Machine Translation". In: *Proceedings of the ACL Workshop on Statistical Machine Translation*. Prague, Czech Republic, pp. 88–95 (*Cité page 8*).

Ferràs, Marc, Cheung Chi Leung, Claude Barras, and Jean-Luc Gauvain (2007). "Constrained MLLR for Speaker Recognition". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Honolulu, Hawaii, USA, pp. 53–56 (*Cité page 87*).

Foster, George, Cyril Goutte, and Roland Kuhn (2010). "Discriminative instance weighting for domain adaptation in statistical machine translation". In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. EMNLP '10. Cambridge, Massachusetts: Association for Computational Linguistics, pp. 451–459. URL: http://dl.acm.org/citation.cfm?id=1870658.1870702 (*Cité pages 89, 90*).

Foster, George and Roland Kuhn (2007). "Mixture-model adaptation for SMT". In: *Proceedings of the Second Workshop on Statistical Machine Translation*. StatMT '07. Prague, Czech Republic: Association for Computational Linguistics, pp. 128–135. URL: http://dl.acm.org/citation.cfm?id=1626355.1626372 (*Cité page 90*).

Foster, George and Roland Kuhn (2009). "Stabilizing minimum error rate training". In: *Proceedings of the Fourth Workshop on Statistical Machine Translation*. StatMT '09. Athens, Greece: Association for Computational Linguistics, pp. 242–249 (*Cité page 10*).

Fournier, Bas (2008). "Preprocessing on bilingual data for Statistical Machine Translation". PhD thesis. University of Twente, The Netherlands (*Cité page 43*).

Fung, Pascale and Percy Cheung (2004). "Multilevel Bootstrapping for Extracting Parallel Sentences from a Quasi Parallel Corpus". In: *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP 04)*, pp. 1051–1057 (*Cité page 35*).

Fung, Pascale and Lo Yuen Yee (1998). "An IR approach for translating new words from nonparallel, comparable texts". In: *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics, pp. 414–420 (*Cité page 35*).

Galliano, Sylvain, Guillaume Gravier, and Laura Chaubard (2009). "The ester 2 evaluation campaign for the rich transcription of French radio

broadcasts". In: *10th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Brighton, United Kingdom: ISCA, pp. 2583–2586 (*Cité page 77*).

Gao, Qin and Stephan Vogel (2008). "Parallel implementations of word alignment tool". In: *In Proc. of the ACL 2008 Software Engineering, Testing, and Quality Assurance Workshop*. SETQA-NLP '08. Columbus, Ohio, pp. 49–57 (*Cité pages 6, 11, 43, 61, 77, 92, 110*).

Germann, Ulrich, Michael Jahr, Kevin Knight, Daniel Marcu, and Kenji Yamada (2001). "Fast decoding and optimal decoding for machine translation". In: *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*. ACL '01. Toulouse, France: Association for Computational Linguistics, pp. 228–235. DOI: 10.3115/1073012.1073042. URL: http://dx.doi.org/10.3115/1073012.1073042 (*Cité pages 8, 10*).

Giménez, Jesús and Lluís Màrquez (2004). "SVMTool: A general POS tagger generator based on support vector machines". In: *In Proc. of the 4th International Conference on Language Resources and Evaluation*, pp. 43–46 (*Cité page 51*).

Goldwater, Sharon and David Mc Closky (2005). "Improving statistical MT through morphological analysis". In: *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. HLT '05. Vancouver, British Columbia, Canada: Association for Computational Linguistics, pp. 676–683 (*Cité page 30*).

Goutte, Cyril, Marine Carpuat, and George Foster (2012). "The Impact of Sentence Alignment Errors on Phrase-Based Machine Translation Performance". In: *Conference of the Association for Machine Translation in the Americas (AMTA)*. San Diego, CA (*Cité page 44*).

Graff, David (2007). *Arabic Gigaword Third Edition*. LDC2007T40. Linguistic Data Consortium. ISBN: 1-58563-460-3 (*Cité page 30*).

Graff, David, Junbo Kong, Ke Chen, and Kazuaki Maeda (2007). *English Gigaword Third Edition*. LDC2007T07. Linguistic Data Consortium. ISBN: 1-58563-416-6 (*Cité page 30*).

Guillemin-Lanne, Sylvie, Fathi Debili, Zied Ben Tahar, and Chafik Gaci (2007). "Reconnaissance des entités nommées en arabe". In: *Colloque VSST, Veille Stratégique Scientifique et Technologique*. Marrakech, Marroco (*Cité page 72*).

Habash, Nizar (2008). "Four techniques for online handling of out-of-vocabulary words in Arabic-English statistical machine translation". In: *Proc. of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*, pp. 57–60 (*Cité pages 63, 119*).

Habash, Nizar (2010). *Introduction to Arabic Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers (*Cité pages 15–17, 21, 27, 29, 56*).

Habash, Nizar and Owen Rambow (2005). "Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop". In: *Proc. of the 43rd Annual Meeting on ACL*. ACL '05. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 573–580. DOI: http://dx.doi.org/10.3115/1219840.1219911. URL: http://dx.doi.org/10.3115/1219840.1219911 (*Cité pages 28, 50*).

Habash, Nizar and Owen Rambow (2006). "MAGEAD: A Morphological Analyzer and Generator for the Arabic Dialects". In: *In Proceedings of COLING-ACL*. Sydney, Australia (*Cité pages 28, 50*).

Habash, Nizar, Owen Rambow, and Ryan Roth (2009). "MADA+TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization". In: *Proc. of the 2nd International Conference on Arabic Language Resources and Tools*. Ed. by Khalid Choukri and Bente Maegaard. Cairo, Egypt: The MEDAR Consortium. ISBN: 2-9517408-5-9 (*Cité pages 28, 43, 50, 51*).

Habash, Nizar and Fatiha Sadat (2006). "Arabic preprocessing schemes for statistical machine translation". In: *Proc. of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*. NAACL-Short '06. New York, New York: ACL, pp. 49–52. URL: http://dl.acm.org/citation.cfm?id=1614049.1614062 (*Cité pages 43, 54, 59, 61*).

Habash, Nizar and Fatiha Sadat (2012). "Challenges for Arabic Machine Translation". In: ed. by Abdelhadi Soudi, Ali Farghaly, Günter Neumann, and Rabih Zbib. Natural language processing. Amsterdam ; Philadelphia: John Benjamins Pub. Co. Chap. Arabic preprocessing for statistical machine translation, pp. 73–94 (*Cité pages 30, 53*).

Hajic, Jan, Otakar Smrz, Petr Zemanek, Petr Pajas, Jan Snaidauf, Emanuel Beska, Jakub Krac-

mar, and Kamila Hassanova (2004). *Prague Arabic Dependency Treebank 1.0*. Catalog No. LDC2004T23. University of Pennsylvania: Linguistic Data Consortium. ISBN: ISBN 1-58563-319-4 (*Cité page* 24).

Halek, O., R. Rosa, and O Tamchyna A.and Bojar (2011). "Named entities from wikipedia for machine translation". In: *In Proceedings of the Conference on Theory and Practice of Information Technologies*. Vratna dolina, Slovak Republic (*Cité page* 30).

Hálek, Ondřej, Rudolf Rosa, Aleš Tamchyna, and Ondřej Bojar (2011). "Named Entities from Wikipedia for Machine Translation". In: *ITAT 2011 Information Technologies – Applications and Theory*. Ed. by Markéta Lopatková. Vol. 788, pp. 23–30. ISBN: 978-80-89557-02-8 (*Cité pages* 64, 67).

Haque, Rejwanul, Sudip Kumar Naskar, Josef van Genabith, and Andy Way (2009). "Experiments on Domain Adaptation for English–Hindi SMT". In: *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, PACLIC*. Hong Kong, China: City University of Hong Kong Press, pp. 670–677 (*Cité page* 89).

Hassan, Ahmed, Haytham Fahmy, and Hany Hassan (2007). "Improving Named Entity Translation by Exploiting Comparable and Parallel Corpora". In: *Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP '07)*. Borovets, Bulgaria (*Cité page* 66).

Hazem, Amir and Emmanuel Morin (2012). "QAlign: A New Method for Bilingual Lexicon Extraction from Comparable Corpora". In: *Computational Linguistics and Intelligent Text Processing - 13th International Conference, CICLing 2012*. Ed. by Alexander F. Gelbukh. Vol. 7182. Lecture Notes in Computer Science. New Delhi, India: Springer, pp. 83–96. ISBN: 978-3-642-28600-1 (*Cité page* 36).

Hazem, Amir, Emmanuel Morin, and Sebastian Peña Saldarriaga (2011). "Bilingual lexicon extraction from comparable corpora as metasearch". In: *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*. BUCC '11. Portland, Oregon: Association for Computational Linguistics, pp. 35–43. ISBN: 978-1-937284-015. URL: http://dl.acm.org/citation.cfm?id=2024236.2024244 (*Cité page* 35).

Heafield, Kenneth (2011). "KenLM: Faster and Smaller Language Model Queries". In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Edinburgh, UK: Association for Computational Linguistics (*Cité page* 8).

Heintz, Ilana (2008). "Arabic Language Modeling with Finite State Transducers". In: *Proc. of the ACL-08: HLT Student Research Workshop*. Columbus, Ohio: ACL, pp. 37–42. URL: http://www.aclweb.org/anthology/P/P08/P08-3007 (*Cité page* 63).

Hermjakob, Ulf, Kevin Knight, and Hal Daumé III (2008). "Name Translation in Statistical Machine Translation - Learning When to Transliterate". In: *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA*. Ed. by Kathleen McKeown, Johanna D. Moore, Simone Teufel, Allan James, and Sadaoki Furui. The Association for Computer Linguistics, pp. 389–397 (*Cité pages* 30, 63).

Hildebrand, Almut Silja, Matthias Eck, Stephan Vogel, and Alex Waibel (2005). "Adaptation of the Translation Model for Statistical Machine Translation based on Information Retrieval". In: *Proceedings of the 10th Annual Conference on European Association for Machine Translation*. EAMT '05. Budapest, Hungary (*Cité page* 89).

Hutchins, John (01/1998). "From First Conception to First Demonstration: the Nascent Years of Machine Translation, 1947–1954.A Chronology". In: *Machine Translation* 12.3, pp. 195–252. ISSN: 0922-6567 (*Cité page* 3).

Hutchins, W. John (2001). "Machine Translation over Fifty Years". In: *HISTOIRE, EPISTEMOLOGIE, LANGAGE, TOME XXII, FASC. 1 (2001)* 23, pp. 7–31 (*Cité page* 4).

Hutchins, W. John (2010). "Machine translation: A concise history". In: *Journal of Translation Studies on Special issue: The teaching of computer-aided translation* 13.1–2, pp. 29–70 (*Cité page* 4).

Jiang, Jing and ChengXiang Zhai (2007). "Instance Weighting for Domain Adaptation in NLP". In: *Proc. of the 45th Annual Meeting of the ACL*. Prague, Czech Republic, pp. 264–271 (*Cité page* 66).

Jiang, Long, Ming Zhou, Lee feng Chien, and Cheng Niu (2007). "Named Entity Translation with Web Mining and Transliteration". In: *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pp. 1629–1634 (*Cité page* 66).

Jing, Hongyan, Radu Florian, Xiaoqiang Luo, Tong Zhang, and Abraham Ittycheriah (2003). "How to get a Chinese Name (Entity): Segmentation

and Combination Issues". In: *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*. Ed. by Michael Collins and Mark Steedman, pp. 200–207. URL: http://www.aclweb.org/anthology/W03-1026.pdf (*Cité page* 30).

Kahn, Juliette (2011). "Parole de locuteur : performance et confiance en identification biométrique vocale". PhD thesis. Université d'Avignon et des Pays de Vaucluse (*Cité page* xiii).

Kashani, Mehdi M., Fred Popowich, and Anoop Sarkar (2006). "Automatic Transliteration of Proper Nouns from Arabic to English. The Challenge of Arabic For NLP/MT". In: *Proceedings of the Second Workshop on Computational Approaches to Arabic Script-based Languages* (*Cité pages* 27, 30).

Kashani, Mehdi M., Eric Joanis, Roland Kuhn, George Foster, and Fred Popowich (2007). "Integration of an Arabic transliteration module into a statistical machine translation system". In: *Proceedings of the Second Workshop on Statistical Machine Translation*. StatMT '07. Prague, Czech Republic: Association for Computational Linguistics, pp. 17–24 (*Cité page* 66).

Kneser, Reinhard and Hermann Ney (1995). "Improved backing-off for M-gram language modeling". In: *International Conference* 1, 181—184 (*Cité page* 8).

Knight, Kevin (12/1999). "Decoding complexity in word-replacement translation models". In: *Comput. Linguist.* 25.4, pp. 607–615. ISSN: 0891-2017. URL: http://dl.acm.org/citation.cfm?id=973226.973232 (*Cité page* 8).

Knight, Kevin and Jonathan Graehl (12/1998). "Machine transliteration". In: *Comput. Linguist.* 24.4, pp. 599–612. ISSN: 0891-2017 (*Cité page* 26).

Knight, Kevin and Philipp Koehn (2004). *What's New in Statistical Machine Translation*. Boston (*Cité page* 7).

Knight, Kevin and Daniel Marcu (2005). "Machine translation in the year 2004". In: *In Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP*. IEEE Computer Society, pp. 965–968 (*Cité page* 4).

Koehn, Philipp (2004). "Pharaoh: A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models". In: *Machine Translation: From Real Users to Research, 6th Conference of the Association for Machine Translation in the Americas, AMTA 2004, Washington, DC, USA, September 28-October 2, 2004, Proceedings*. Ed. by Robert E. Frederking and Kathryn Taylor. Vol. 3265. Lecture Notes in Computer Science. Springer, pp. 115–124. ISBN: 3-540-23300-8 (*Cité page* 9).

Koehn, Philipp (2010). *Statistical Machine Translation*. 1st. New York, NY, USA: Cambridge University Press. ISBN: 0521874157, 9780521874151 (*Cité pages* 4, 9).

Koehn, Philipp, Franz Josef Och, and Daniel Marcu (2003). "Statistical phrase-based translation". In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*. NAACL '03. Edmonton, Canada: Association for Computational Linguistics, pp. 48–54. DOI: 10.3115/1073445.1073462. URL: http://dx.doi.org/10.3115/1073445.1073462 (*Cité page* 6).

Koehn, Philipp and Josh Schroeder (2007). "Experiments in domain adaptation for statistical machine translation". In: *Proceedings of the Second Workshop on Statistical Machine Translation*. StatMT '07. Prague, Czech Republic: Association for Computational Linguistics, pp. 224–227. URL: http://dl.acm.org/citation.cfm?id=1626355.1626388 (*Cité page* 89).

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst (2007). "Moses: open source toolkit for statistical machine translation". In: *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. ACL '07. Prague, Czech Republic: Association for Computational Linguistics, pp. 177–180 (*Cité pages* 10, 43, 61, 77).

Kouloughli, Djamel E. (1994). *Grammaire de l'arabe d'aujourd'hui*. Les Langues pour Tous. Pocket. ISBN: 9782266039123. URL: http://books.google.fr/books?id=4Lv5PQAACAAJ (*Cité pages* xvi, 20).

Kouloughli, E. Djamel (2007). *Moyen arabe et questions connexes*. URL: http://cle.ens-lyon.fr/arabe/langue-et-langues-11865.kjsp?RH=CDL&RF=CDL_ARA120000 (*Cité page* 15).

Kulick, Seth (2010). "Simultaneous tokenization and part-of-speech tagging for Arabic without a morphological analyzer". In: *Proc. of the ACL 2010 Conference Short Papers*. ACLShort '10. Uppsala, Sweden: Association for Computational Linguistics, pp. 342–347. URL: http:

//dl.acm.org/citation.cfm?id=1858842.1858905 (*Cité pages* 27, 50).

Kulick, Seth (2011). "Exploiting Separation of Closed-Class Categories for Arabic Tokenization and Part-of-Speech Tagging". In: *ACM Transactions on Asian Language Information Processing (TALIP)* 10 (1), p. 4. ISSN: 1530-0226 (*Cité page* 27).

Kumano, Tadashi and Hideki Tanaka Takenobu Tokunaga (2007). "Extracting Phrasal Alignments from Comparable Corpora by Using Joint Probability SMT Model". In: *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'07)*. Ed. by Andy Way and Barbara Gawronska. Skövde, Sweden (*Cité page* 37).

Kumar, Shankar and William J. Byrne (2003). "A Weighted Finite State Transducer Implementation of the Alignment Template Model for Statistical Machine Translation". In: *HLT-NAACL* (*Cité page* 8).

Kumar, Shankar, Yonggang Deng, and William Byrne (2006). "A weighted finite state transducer translation template model for statistical machine translation". In: *Natural Language Engineering* 12.1, pp. 35–75 (*Cité page* 9).

Lafferty, John, Andrew McCallum, and Fernando Pereira (2001). "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". In: *Proc. ICML*. San Francisco, pp. 282–289 (*Cité pages* 40, 54).

Lambert, Patrik, Holger Schwenk, Christophe Servan, and Sadaf Abdul-Rauf (2011). "Investigations on Translation Model Adaptation Using Monolingual Data". In: *Empirical Methods in Natural Language Processing / Workshop on statistical Machine Translation (EMNLP/WMT)*. Edinburgh (UK) (*Cité page* 89).

Lamel, Lori, Abdel Messaoudi, and Jean-Luc Gauvain (2007). "Improved Acoustic Modeling for Transcribing Arabic Broadcast Data". In: *Interspeech*. Antwerp, Belgium (*Cité page* 31).

Langé, Jean-Marc (1995). "Modèles statistiques pour l'extraction de lexiques bilingues". In: *Traitement Automatique des Langues* 36.1-2, pp. 133–155 (*Cité page* 35).

Langlais, Philippe (2002). "Improving a general-purpose Statistical Translation Engine by terminological lexicons". In: *COLING-02 on COMPUTERM 2002: second international workshop on computational terminology - Volume 14*. COMPUTERM '02. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 1–7. DOI: 10.3115/1118771.1118776. URL: http://dx.doi.org/10.3115/1118771.1118776 (*Cité page* 89).

Larkey, Leah S., Lisa Ballesteros, and Margaret E. Connell (2002). "Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis". In: *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. SIGIR '02. Tampere, Finland: ACM, pp. 275–282. ISBN: 1-58113-561-0 (*Cité page* 30).

Lavecchia, Caroline (2010). "Les Triggers Interlangues pour la Traduction Automatique Statistique". PhD thesis. Université Nancy (*Cité page* 13).

Lavergne, Thomas, Olivier Cappé, and François Yvon (2010). "Practical Very Large Scale CRFs". In: *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 504–513 (*Cité pages* 40, 54).

Lavergne, Thomas, Hai-Son Le, Alexandre Allauzen, and François Yvon (2011). "LIMSI's experiments in domain adaptation for IWSLT11". In: *Proceedings of the eigth International Workshop on Spoken Language Translation (IWSLT)*. Ed. by Mei-Yuh Hwang and Sebastian Stüker. San Francisco, CA, pp. 62–67 (*Cité page* 90).

Lavie, Alon and Abhaya Agarwal (2007). "METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments". In: *Proceedings of the ACL Workshop on Statistical Machine Translation*. Prague, Czech Republic, pp. 338–231 (*Cité pages* 12, 78).

Le, Hai Son, Ilya Oparin, Abdelkhalek Messaoudi, Alexandre Allauzen, Jean-Luc Gauvain, and François Yvon (2011). "Large Vocabulary SOUL Neural Network Language Models". In: *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association*. Florence, Italy: ISCA, pp. 1469–1472 (*Cité page* 116).

Le, Hai Son, Ilya Oparin, Alexandre Allauzen, Jean-Luc Gauvain, and François Yvon (2013). "Structured Output Layer Neural Network Language Models for Speech Recognition". In: *IEEE Transactions on Audio, Speech & Language Processing* 21.1, pp. 195–204 (*Cité page* 116).

Lebert, Marie (2010). *L'Ethnologue recense les 6.909 langues vivantes de la planète*. URL: http://archive.wikiwix.com/cache/?url=http://www.actualitte.com/actualite/17394-

Ethnologue - recenser - langues - vivantes - monde.htm&title=L%27 (*Cité page* xiii).

Lee, Young-suk (2004). "Morphological analysis for statistical machine translation". In: *In Proceedings of the Human Language Technology, Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL*, pp. 57–60 (*Cité page* 27).

Levenberg, Abby and Miles Osborne (2009). "Stream-based randomised language models for SMT". In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*. EMNLP '09. Singapore: Association for Computational Linguistics, pp. 756–764 (*Cité page* 121).

Lewis, David D. (1998). "Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval". In: *Machine Learning: ECML-98, 10th European Conference on Machine Learning*. Ed. by Claire Nedellec and Céline Rouveirol. Vol. 1398. Lecture Notes in Computer Science. Chemnitz, Germany: Springer, pp. 4–15. ISBN: 3-540-64417-2 (*Cité page* 95).

Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig (2013). *Ethnologue: Languages of the World*. Dallas, Texas: SIL International. URL: http://www.ethnologue.com/statistics/size (*Cité page* 15).

Li, Bo and Eric Gaussier (2010). "Improving corpus comparability for bilingual lexicon extraction from comparable corpora". In: *Proceedings of the 23rd International Conference on Computational Linguistics*. COLING '10. Beijing, China: Association for Computational Linguistics, pp. 644–652. URL: http://dl.acm.org/citation.cfm?id=1873781.1873854 (*Cité page* 35).

Lin, Dekang (1998). "An Information-Theoretic Definition of Similarity". In: *In Proceedings of the 15th International Conference on Machine Learning*. Morgan Kaufmann, pp. 296–304 (*Cité page* 40).

Ling, Wang, Pável Calado, Bruno Martins, Isabel Trancoso, Alan Black, and Luísa Coheu (2011). "Named Entity Translation using Anchor Texts". In: *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*. San Francisco, USA (*Cité page* 66).

Lopez, Adam (2008). "Statistical machine translation". In: *ACM Comput. Surv.* 40.3 (*Cité page* 4).

Maamouri, Mohamed, Ann Bies, Tim Buckwalter, Hubert Jin, and Wigdan Mekki (2005a). *Arabic Treebank: Part 3 (full corpus) v 2.0 (MPG + Syntactic Analysis)*. LDC2005T20. Linguistic Data Consortium. ISBN: 1-58563-341-0 (*Cité pages* 24, 55, 71).

Maamouri, Mohamed, Ann Bies, Tim Buckwalter, Hubert Jin, and Wigdan Mekki (2005b). *Arabic Treebank: Part 4 v 1.0 (MPG Annotation)*. LDC2005T30. Linguistic Data Consortium. ISBN: 1-58563-343-7 (*Cité pages* 24, 55, 71).

Maamouri, Mohamed, Dave Graff, Basma Bouziri, Sondos Krouna, Ann Bies, and Seth Kulick (2010a). *LDC Standard Arabic Morphological Analyzer (SAMA) Version 3.1*. LDC2010L01. ISBN: 1-58563-555-3 (*Cité page* 23).

Maamouri, Mohamed, Dave Graff, Basma Bouziri, Sondos Krouna, Ann Bies, and Seth Kulick (2010b). *LDC Standard Arabic Morphological Analyzer (SAMA) Version 3.1*. LDC2010L01. Linguistic Data Consortium. ISBN: 1-58563-555-3 (*Cité page* 27).

Macherey, Wolfgang, Franz Josef Och, Ignacio Thayer, and Jakob Uszkoreit (2008). "Lattice-based minimum error rate training for statistical machine translation". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '08. Honolulu, Hawaii: Association for Computational Linguistics, pp. 725–734 (*Cité page* 10).

Maloney, John and Michael Niv (1998). "TAGARAB: a fast, accurate Arabic name recognizer using high-precision morphological analysis". In: *Proc. of the Workshop on Computational Approaches to Semitic Languages*. Semitic '98. Montreal, Quebec, Canada, pp. 8–15. URL: http://dl.acm.org/citation.cfm?id=1621753.1621756 (*Cité pages* 29, 65).

Mansour, Saab (12/2010). "MorphTagger: HMM-Based Arabic Segmentation for Statistical Machine Translation". In: *International Workshop on Spoken Language Translation*. Paris, France, pp. 321–327 (*Cité pages* 28, 50, 53, 54).

Mansour, Saab and Hermann Ney (12/2012). "A Simple and Effective Weighted Phrase Extraction for Machine Translation Adaptation". In: *International Workshop on Spoken Language Translation*. Hong Kong, pp. 193–200 (*Cité page* 90).

Mansour, Saab and Hermann Ney (2013). "Phrase Training Based Adaptation for Statistical Machine Translation". In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Georgia: Association for Computational Linguistics,

pp. 649–654. URL: http://www.aclweb.org/anthology/N13-1074 (*Cité page 104*).

Mansour, Saab, Khalil Sima'an, and Yoad Winter (2007). "Smoothing a lexicon-based POS tagger for Arabic and Hebrew". In: *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*. Semitic '07. Prague, Czech Republic: Association for Computational Linguistics, pp. 97–103 (*Cité pages 28, 51, 53*).

Marsi, Erwin, Antal van den Bosch, and Abdelhadi Soudi (2005). "Memory-based morphological analysis generation and part-of-speech tagging of Arabic". In: *Proc. of the ACL 2005 workshop on computational approaches to Semitic languages*. Ann Arbor, USA, pp. 1–8 (*Cité pages 27, 50*).

Meftouh, Karima, Najette Bouchemal, and Kamel Smaïli (05/2012). "A Study of a Non-Resourced Language: The Case of one of the Algerian Dialects". In: *The Third International Workshop on Spoken Languages Technologies for Under-resourced Language, SLTU'12*. Cape-Town, Afrique Du Sud, pp. – (*Cité page 17*).

Messaoudi, Abdel, Jean-Luc Gauvain, and Lori Lamel (2006). "Arabic broadcast news transcription using a one million word vocalized vocabulary". In: *In Proceedings of ICASSP*. volume 1, pp. 1093–1096 (*Cité page 31*).

Mihalcea, Rada (2004). "Co-training and Self-training for Word Sense Disambiguation". In: *HLT-NAACL Workshop: CoNLL-2004*. Ed. by Hwee Tou Ng and Ellen Riloff. Boston, pp. 33–40 (*Cité page 66*).

Mohamed, Emad and Sandra Kübler (2010). "Arabic Part of Speech Tagging". In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Ed. by Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias. Valletta, Malta: European Language Resources Association (ELRA). ISBN: 2-9517408-6-7 (*Cité page 50*).

Mohamed, Emad, Behrang Mohit, and Kemal Oflazer (2012). "Annotating and Learning Morphological Segmentation of Egyptian Colloquial Arabic". In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis. Istanbul, Turkey: European Language Resources Association (ELRA), pp. 873–877. ISBN: 978-2-9517408-7-7 (*Cité pages 28, 50*).

Moore, Robert C. (2003). "Learning Translations of Named-Entity Phrases from Parallel Corpora". In: *IN PROC. OF EACL*, pp. 259–266 (*Cité page 66*).

Moore, Robert C. and Chris Quirk (2007). "Faster beam-search decoding for phrasal statistical machine translation". In: *In Proceedings of MT Summit XI* (*Cité page 9*).

Morin, Emmanuel and Béatrice Daille (2004). "Extraction de terminologies bilingues à partir de corpus comparables d'un domaine spécialisé". In: *Traitement Automatique des Langues (TAL)*. Vol. 45. 3. Lavoisier, 103–122 (*Cité page 35*).

Morin, Emmanuel and Emmanuel Prochasson (2011). "Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora". In: *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*. BUCC '11. Portland, Oregon: Association for Computational Linguistics, pp. 27–34. ISBN: 978-1-937284-015. URL: http://dl.acm.org/citation.cfm?id=2024236.2024243 (*Cité page 35*).

Morin, Emmanuel, Béatrice Daille, Koichi Takeuchi, and Kyo Kageura (2010). "Brains, not brawn: The use of "smart" comparable corpora in bilingual terminology mining". In: *TSLP* 7.1 (*Cité page 35*).

Munteanu, Dragos Stefan and Daniel Marcu (2005). "Improving Machine Translation Performance by Exploiting Non-Parallel Corpora". In: *Computational Linguistics* 31.4, pp. 477–504 (*Cité pages 36, 37, 40*).

Munteanu, Dragos Stefan and Daniel Marcu (2006). "Extracting parallel sub-sentential fragments from non-parallel corpora". In: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. ACL-44. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 81–88 (*Cité page 37*).

Nagao, Makoto (1984). "A framework of mechanical translation beteween Japanese and English". In: *Artificial and Human Intelligence*. Ed. by Alick Elithorn and Rana Banerji. North-Holland, pp. 173–180 (*Cité page 3*).

Nakov, Preslav (2008). "Improving English-Spanish statistical machine translation: experiments in domain adaptation, sentence paraphrasing, tokenization, and recasing". In:

Proceedings of the Third Workshop on Statistical Machine Translation. StatMT '08. Columbus, Ohio: Association for Computational Linguistics, pp. 147–150. ISBN: 978-1-932432-09-1 (*Cité page 89*).

Neyreneuf, Michel and Ghalib Al-Hakkak (2001). *Grammaire active de l'arabe littéral*. Le Livre de poche. Méthode 90. Librairie Générale Française. ISBN: 9782253085614 (*Cité pages 21, 22*).

Nguyen, ThuyLinh and Stephan Vogel (2008). "Context-based Arabic morphological analysis for machine translation". In: *Proc. of the 12th Conference on Computational Natural Language Learning*. CoNLL '08. Manchester, United Kingdom: ACL, pp. 135–142. ISBN: 978-1-905593-48-4. URL: http://dl.acm.org/citation.cfm?id=1596324.1596348 (*Cité page 22*).

Niehues, Jan and Alex Waibel (2010). "Domain adaptation in statistical machine translation using factored translation models". In: *Proceedings of EAMT* (*Cité page 89*).

Nießen, Sonja and Hermann Ney (2000). "Improving SMT quality with morpho-syntactic analysis". In: *18th Int. Conf. on Computational Linguistics*, pp. 1081–1085 (*Cité page 30*).

Och, Franz Josef (2003). "Minimum error rate training in statistical machine translation". In: *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*. ACL '03. Sapporo, Japan: Association for Computational Linguistics, pp. 160–167. DOI: 10.3115/1075096.1075117. URL: http://dx.doi.org/10.3115/1075096.1075117 (*Cité pages 10, 41, 43, 61, 92, 100*).

Och, Franz Josef and Hermann Ney (03/2003). "A systematic comparison of various statistical alignment models". In: *Comput. Linguist.* 29.1, pp. 19–51. ISSN: 0891-2017 (*Cité page 6*).

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). "BLEU: a method for automatic evaluation of machine translation". In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 311–318 (*Cité pages 10, 11, 45, 57, 78*).

Popovic, Maja and Hermann Ney (2004). "Towards the Use of Word Stems and Suffixes for Statistical Machine Translation". In: *LREC* (*Cité page 30*).

Rapp, Reinhard (1995). "Identifying word translations in non-parallel texts". In: *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*. ACL '95. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 320–322 (*Cité page 35*).

Rapp, Reinhard, Marko Tadic, Serge Sharoff, and Pierre Zweigenbaum, eds. (2012). *Proceedings of the 5th Workshop on Building and Using Comparable Corpora, LREC 2012*. Istanbul, Turkey (*Cité pages xiii, 36*).

Resnik, Philip and Noah A. Smith (2003). "The Web as a Parallel Corpus". In: *Computational Linguistics* 29, pp. 349–380 (*Cité page 37*).

Rosset, Sophie, Cyril Grouin, Karën Fort, Olivier Galibert, Juliette Kahn, and Pierre Zweigenbaum (07/2012). "Structured Named Entities in two distinct press corpora: Contemporary Broadcast News and Old Newspapers". In: *Proc. of the 6th ACL Linguistic Annotation Workshop*. Jeju, Corée, pp. 40–48 (*Cité page 64*).

Saadane, Houda, Aurélie Rossi, Christian Fluhr, and Mathieu Guidère (2012). "Transcription of Arabic names into Latin". In: pp. 857–866 (*Cité page 120*).

Salloum, Wael and Nizar Habash (2012). "Elissa: A Dialectal to Standard Arabic Machine Translation System". In: *In Proceedings of the International Conference on Computational Linguistics (COLING 2012)*. Mumbai, India (*Cité page 31*).

Samy, Doaa, Antonio Moreno, and José M. Guirao (2005). "A Proposal For An Arabic Named Entity Tagger Leveraging a Parallel Corpus". In: RANLP '05 (*Cité pages 29, 65*).

Santanu, Pal, Sudip Kumar Naskar, Pavel Pecina, Sivaji Bandyopadhyay, and Andy Way (2010). "Handling Named Entities and Compound Verbs in Phrase-Based Statistical Machine Translation". In: *Proceedings of the COLING 2010 Workshop on Multiword Expressions: from Theory to Applications (MWE 2010)*. Beijing, China, pp. 46–54 (*Cité page 66*).

Schwenk, Holger (2008). "Investigations on large-scale lightly-supervised training for statistical machine translation". In: *Proceedings of the International Workshop on Spoken Language Translation*. Hawaii, USA, pp. 182–189 (*Cité page 36*).

Schwenk, Holger and Jean Senellart (2009). "Translation Model Adaptation for an Arabic/French News Translation System by Lightly-Supervised Training". In: *MT Summit* (*Cité page 89*).

Sennrich, Rico (2012a). "Mixture-Modeling with Unsupervised Clusters for Domain Adaptation

in Statistical Machine Translation". In: *Proceedings of the 16th EAMT Conference*. Trento, Italy, pp. 185–192 (*Cité pages* 90, 103).

Sennrich, Rico (2012b). "Perplexity Minimization for Translation Model Domain Adaptation in Statistical Machine Translation". In: *EACL 2012, 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, April 23-27, 2012*. Ed. by Walter Daelemans, Mirella Lapata, and Lluís Màrquez. The Association for Computer Linguistics, pp. 539–549 (*Cité pages* 90, 104).

Shaalan, Khaled and Hafsa Raza (2009). "NERA: Named Entity Recognition for Arabic". In: *Journal of the American Society for Information Science and Technology* 60.9, pp. 1652–1663 (*Cité pages* 29, 65).

Shah, Rushin, Paramveer S. Dhillon, Mark Liberman, Dean Foster, Mohamed Maamouri, and Lyle Ungar (2010). "A new approach to lexical disambiguation of Arabic text". In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. EMNLP '10. Cambridge, Massachusetts: Association for Computational Linguistics, pp. 725–735 (*Cité page* 28).

Sherif, Tarek and Kondrak Grzegorz (2007). "Bootstrapping a Stochastic Transducer for Arabic-English Transliteration Extraction". In: *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*. Ed. by John A. Carroll, Antal van den Bosch, and Annie Zaenen. The Association for Computational Linguistics (*Cité page* 27).

Smadja, Frank, Kathleen R. McKeown, and Vasileios Hatzivassiloglou (1996). "Translating collocations for bilingual lexicons: a statistical approach". In: *Computational Linguistics* 22, pp. 1–38 (*Cité page* 35).

Smith, Jason R., Chris Quirk, and Kristina Toutanova (2010). "Extracting parallel sentences from comparable corpora using document level alignment". In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. HLT '10. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 403–411 (*Cité page* 37).

Snover, Matthew, Bonnie Dorr, and Richard Schwartz (2008). "Language and translation model adaptation using comparable corpora". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '08. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 857–866 (*Cité pages* 37, 89).

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul (2006a). "A Study of Translation Edit Rate with Targeted Human Annotation". In: *Proceedings of Association for Machine Translation in the Americas (AMTA)* (*Cité page* 12).

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul (2006b). "A Study of Translation Edit Rate with Targeted Human Annotation". In: *Proceedings of the conference of the Association for Machine Translation in the America (AMTA)*, pp. 223–231 (*Cité page* 57).

Sokolovska, Nataliya, Olivier Cappé, and François Yvon (2009). "Sélection de caractéristiques pour les champs aléatoires conditionnels par pénalisation L_1". In: *TAL* 50.3, pp. 139–171 (*Cité page* 70).

Soltau, Hagen, George Saon, Brian Kingsbury, Hong-Kwang Jeff Kuo, Lidia Mangu, Daniel Povey, and Ahmad Emami (2009). "Advances in Arabic Speech Transcription at IBM Under the DARPA GALE Program". In: *IEEE Transactions on Audio, Speech & Language Processing* 17.5, pp. 884–894 (*Cité page* 31).

Stolcke, Andreas (2002). "SRILM - An Extensible Language Modeling Toolkit". In: *Proceedings International Conference on Spoken Language Processing*, pp. 901–904 (*Cité pages* 8, 51, 92, 110).

Tillmann, Christoph and Jian-ming Xu (2009). "A simple sentence-level extraction algorithm for comparable data". In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*. NAACL-Short '09. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 93–96 (*Cité page* 36).

Ueffing, Nicola, Gholamreza Haffari, and Anoop Sarkar (06/2008). "Semi-supervised model adaptation for statistical machine translation". In: *Machine Translation* 21.2, pp. 77–94. ISSN: 0922-6567 (*Cité page* 36).

Uszkoreit, Jakob, Jay M. Ponte, Ashok C. Popat, and Moshe Dubiner (2010). "Large scale parallel document mining for machine translation". In: *Proceedings of the 23rd International Conference on Computational Linguistics*. COLING '10. Stroudsburg, PA, USA: Association for Com-

putational Linguistics, pp. 1101–1109 (*Cité page 37*).

Varga, Dániel, László Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy (2005). "Parallel corpora for medium density languages". In: *Proceedings of the RANLP*, pp. 590–596 (*Cité page 40*).

Wang, Ye yi and Alex Waibel (1997). "Decoding Algorithm in Statistical Machine Translation". In: pp. 366–372 (*Cité page 8*).

Watanabe, Taro, Jun Suzuki, Hajime Tsukada, and Hideki Isozaki (2007). "Online large-margin training for statistical machine translation". In: *In Proc. of EMNLP* (*Cité page 10*).

Watson, Janet C. E. (2002). *The Phonology and Morphology of Arabic*. The phonology of the world's languages. Oxford: Oxford Univ. Press. ISBN: 9780199257591 (*Cité page 20*).

Whittaker, E.W.D (2001). "Temporal Adaptation of Language Models". In: *LM Adaptation for information retrieval of spoken news/radio programs (i.e. Speech-Bot)*. Sophia-Antipolis, France: Adaptation Methods for Speech Recognition, ISCA Tutorial and Research Workshop (ITRW) (*Cité page 121*).

Wright, William (1988). *A Grammar of the Arabic Language*. 3ème édition révisée. Cambridge (*Cité page 28*).

Yamada, Kenji and Kevin Knight (2001). "A syntax-based statistical translation model". In: *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*. ACL '01. Toulouse, France: Association for Computational Linguistics, pp. 523–530 (*Cité page 13*).

Yamamoto, Hirofumi and Eiichiro Sumita (2007). "Bilingual Cluster Based Models for Statistical Machine Translation". In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learnin (EMNLP-CoNLL)*. Prague, Czech Republic, pp. 514–523 (*Cité page 89*).

Yamamoto, Hirofumi and Eiichiro Sumita (2008). "Bilingual Cluster Based Models for Statistical Machine Translation". In: *IEICE Transactions* 91-D.3, pp. 588–597 (*Cité pages 89, 90, 103*).

Yvon, François (2011). "Introduction aux modèles probabilistes pour la fouille de textes". In: *Modèles statistiques pour l'accès à l'information textuelle*. Ed. by Eric Gaussier and François Yvon. Hermès, Paris. Chap. 7, pp. 423–477 (*Cité page 95*).

Zaghouani, Wajdi, Bruno Pouliquen, Mohamed Ebrahim, and Ralf Steinberger (2010). "Adapting a resource-light highly multilingual Named Entity Recognition system to Arabic". In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), pp. 563–567 (*Cité pages 29, 65*).

Zbib, Rabih and Ibrahim Badr (2012). "Challenges for Arabic Machine Translation". In: ed. by Abdelhadi Soudi, Ali Farghaly, Günter Neumann, and Rabih Zbib. Natural language processing. Amsterdam ; Philadelphia: John Benjamins Pub. Co. Chap. Preprocessing for English-to-Arabic Statistical Machine Translation, pp. 97–107 (*Cité page 31*).

Zbib, Rabih and Abdelhadi Soudi (2012). "Challenges for Arabic Machine Translation". In: ed. by Abdelhadi Soudi, Ali Farghaly, Günter Neumann, and Rabih Zbib. Natural language processing. Amsterdam ; Philadelphia: John Benjamins Pub. Co. Chap. Introduction: Challenges for Arabic Machine Translation, pp. 1–13 (*Cité pages xvi, 30*).

Zbib, Rabih, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris Callison-Burch (2012). "Machine translation of Arabic dialects". In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NAACL HLT '12. Montreal, Canada: Association for Computational Linguistics, pp. 49–59. ISBN: 978-1-937284-20-6. URL: http://dl.acm.org/citation.cfm?id=2382029.2382037 (*Cité page 31*).

Zens, Richard, Franz Josef Och, and Hermann Ney (2002). "Phrase-Based Statistical Machine Translation". In: Springer Verlag, pp. 18–32 (*Cité page 6*).

Zhang, Min, Haizhou Li, A Kumaran, and Ming Liu (2011). "Report of NEWS2011 Machin Transliteration Shared Task". In: *Proceedings of the 2011 Named Entities Workshop*. Chang Mai, Thailand (*Cité page 63*).

Zhang, Yuqi (2012). "The Application of Source Language Information in Statistical Machine Translation". PhD thesis. RWTH Aachen University (*Cité pages 63, 85*).

Zhao, Bing, Matthias Eck, and Stephan Vogel (2004). "Language model adaptation for statistical machine translation with structured query models". In: *Proceedings of the 20th international conference on Computational Linguistics*. COL-

ING '04. Geneva, Switzerland: Association for Computational Linguistics (*Cité pages* 37, 89).

Zhao, Bing and Stephan Vogel (2002). "Adaptive parallel sentence mining from Web bilingual news collection". In: *Proceedings of the International Conference on Data Mining*. IEEE Computer Society, pp. 745–748 (*Cité page* 37).

Zitouni, Imed, Jeffrey Sorensen, Xiaoqiang Luo, and Radu Florian (2005). "The Impact of Morphological Stemming on Arabic Mention Detection and Coreference Resolution". In: *Proc. of Workshop on Computational Approaches to Semitic Languages*. Ann Arbor, Michigan, pp. 63–70. URL: http://www.aclweb.org/anthology/W/W05/W05-0709 (*Cité pages* 29, 65).

Zollmann, Andreas, Ashish Venugopal, and Stephan Vogel (2006). "Bridging the inflection morphology gap for Arabic statistical machine translation". In: *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*. NAACL-Short '06. New York, New York: Association for Computational Linguistics, pp. 201–204 (*Cité page* 30).

Table des figures

Liste des tableaux