Statistical Machine Translation: Application to low resourced languages

Salima Harrat

République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

École Supérieure d'Informatique

Thèse

Présentée pour l'obtention du grade de DOCTEUR

de l'École Supérieure d'Informatique

Par

Salima Harrat

Traduction Automatique Fondée sur des Méthodes Statistiques : Application aux Langues peu Dotées en Ressources

---

Spécialité : Informatique

Soutenue le 29 Novembre 2018 devant un jury composé de :

Président du jury	**Prof. BALLA Amar**	(ESI)
Examinatrice	**Prof. SI-TAYEB Fatima**	(ESI)
Examinatrice	**Prof. KHELLAF-HANED Hadja-Faiza**	(USTHB)
Examinateur	**Prof. MALKI Mimoun**	(ESIBA)
Directeur de thèse	**Prof. HIDOUCI Khaled-Walid**	(ESI)
Co-Directeur de thèse	**Prof. SMAILI Kamel**	(LORIA, France)

A la mémoire de mon cher père

Merci...

Professeur Kamel Smaili pour la rigueur, la patience et surtout la modestie avec lesquelles vous avez dirigé ce travail. Les mots de remerciements ne seront jamais à la hauteur de ma gratitude.

Professeur Khaled-Walid Hidouci pour votre aide et collaboration.

Professeurs : Amar Balla, Si-Tayeb Fatima, Khellaf-Haned Hadja-Faiza et Malki Mimoun pour avoir accepté de juger mon travail.

Mustapha, Maman et Lila pour votre soutien inconditionnel.

A vous tous qui m'avez aidée et soutenue parents, amis et collègues.

Résumé

Le présent travail s'articule autour de la traduction automatique statistique dans le cadre des langues peu dotées en ressources. On s'intéresse aux dialectes arabes qui représentent le parlé quotidien de tous les peuples arabes. Ces dialectes différent d'un pays arabe à un autre et même dans un même pays on constate l'existence de plusieurs variantes de dialectes. Ces dialectes de par leur nature orale et non-standard représentent un défi pour le domaine de traitement automatique des langues. Dans le cadre précis de la traduction automatique statistique, ces dialectes sont difficiles à prendre en charge à cause de l'absence de ressources (de toutes natures) notamment les corpus monolingues et surtout parallèles nécessaires pour l'apprentissage des différents modèles statistiques. Dans cette thèse, on s'intéresse à cette problématique avec une attention particulière au dialecte algérien et plus précisément le dialecte algérois. Un corpus parallèle multi-dialecte PADIC (pour Parallel Arabic Dialect Corpus) a été créé, il s'agit d'une ressource textuelle importante qui comprend, jusqu'à présent, six dialectes arabes en plus de l'arabe standard. Ce corpus a fait l'objet d'une étude analytique pour mettre en relief la relation entre les dialectes (entre eux) et l'arabe standard. Au moyen du corpus PADIC, on s'est attaqué au problème de la traduction automatique statistique entre les différentes paires de dialectes et l'arabe standard. Plusieurs résultats ont été obtenus et vont tous dans le sens de la difficulté de la traduction des dialectes. Par ailleurs, plusieurs outils dédiés au dialecte algérois ont été réalisés dans le cadre de cette thèse. Le problème du code-switching a été aussi abordé au cours de ce travail où un outil d'identification a été mis en œuvre grâce aux techniques du « Machine Learning ».

Table des matières

Liste des tableaux

Table des figures

Liste des acronymes

PADIC Parallel Arabic Dialect Corpus

ALG Dialecte Algérois

ANB Dialecte Bonois

PAL Dialecte Palestinien

MAR Dialecte Marocain

SYR Dialecte Syrien

TUN Dialecte Tunisien

MSA Modern Standard Arabic (Arabe moderne standard)

HD Distance de Hellinger

WER Word Error Rate

DER Diacritization Error Rate

WB Technique de lissage de Witten-Bell

KN Technique de lissage de Kneser-Ney

BLEU Bilingual Evaluation Understudy

TER Translation Error Rate

METEOR Metric for Evaluation of Translation with Explicit Ordering

Chapitre 1

Introduction

La traduction automatique dans le cadre des langues peu dotées en ressources ; deux axes redoutables dans le domaine du traitement automatique de la langue. D'un coté, la traduction automatique requiert des ressources de toutes natures, de l'autre les langues peu dotées en ressources (comme leur nom l'indique si bien) manquent de ressources de toutes sortes. En effet, comme nous le verrons plus loin dans cette thèse, les approches de la traduction se classent dans l'une des deux grandes familles : l'approche empirique basée sur la disponibilité de corpus et l'approche experte basée sur des connaissances linguistiques profondes des langues sujet à la traduction. Or, pour la quasi-totalité des langues peu dotées en ressources, il est difficile d'envisager telle ou telle approche. En général, les ressources de type corpus monolingue et bilingue sont rares voir même inexistantes pour les langues peu dotées et les connaissances linguistiques profondes qui permettent la mise en place de systèmes de traduction basés sur une approche experte sont souvent indisponibles ou du moins pas assez profondes pour un tel besoin.

C'est au cœur de cette problématique que cette thèse est inscrite, la traduction automatique statistique pour les langues peu dotées, nous abordons les dialectes arabes et en particulier le dialecte algérien. Loin du débat linguistique sur la question est ce que les dialectes arabes sont des langues à part entière. Nous considérons ces dialectes dans le cadre précis de traitement automatique. Par abus de langage, on désignera ces dialectes dans cette thèse de langues au regard du domaine TAL.

Le traitement automatique des dialectes arabes est un domaine de recherche florissant car les premiers travaux sur le sujets sont récents. L'intérêt aux dialectes arabes est souvent motivé par des objectifs militaires comme ceux relatifs à la guerre du Golf où une multitude de travaux sur le dialecte irakien ont été réalisés[Condon et al., 2008, Condon et al., 2010] ; ou bien des objectifs politiques et économiques comme récemment, avec le printemps arabe. Suite à l'émergence des réseaux sociaux et leur influence sur l'opinion publique ainsi que leur rôle dans les révolutions dans le monde arabe, l'intérêt de la communauté scientifique envers ces dialectes est de plus en plus croissant[Shoukry and Rafea, 2012, Ameur and Jamoussi, 2013, Abdul-Mageed and Diab, 2014]. En effet, ces dialectes sont répandus dans les réseaux sociaux et les forums de discussion ainsi que dans les émissions débat de télévision et de radio, et représentent le moyen de communication par excellence de la majorité écrasante des populations arabes.

Plusieurs travaux de recherche ont été initiés au profit des dialectes arabes plus précisément pour certains d'entre eux comme l'égyptien[Zbib et al., 2012, Salloum and Habash, 2013, Sajjad et al., 2013, Aminian et al., 2014], le levantin [Zbib et al., 2012, Salloum and Habash, 2013]et le tunisien[Hamdi et al., 2013, Sadat et al., 2014]. Malheureusement, aucun travail d'envergure n'existait pour le dialecte algérien au début de cette thèse. Nous n'avons trouvé dans les travaux dédiés aux dialectes arabes aucun intérêt alloué à ce dialecte malgré qu'il représente le parler quotidien de dizaines de millions d'Algériens. On constate amèrement que ce dialecte est démuni de toutes sortes de ressources dédiées à son traitement automatique. Travailler sur ce dialecte devient une urgence au regard de la démocratisation d'Internet avec l'avènement des réseaux sociaux d'un coté et l'évolution vertigineuse de la téléphonie mobile de l'autre coté. Ceci a crée de nouveaux besoins dans le domaine de traitement automatique des langues qui devraient prendre en charge les dialectes de plus en plus utilisés. C'est pourquoi

nous mettons le focus dans cette thèse sur le dialecte algérien (en particulier le dialecte algérois).

Organisation de la thèse

La présente thèse est organisée en chapitres. Le second chapitre est consacré au concept des langues peu dotées en ressources, nous en avons expliqué la problématique et avons présenté une méthode de leur évaluation. La seconde partie de ce chapitre est dédiée aux dialectes arabes. Après en avoir donné un bref aperçu, nous avons tenté de situer ces dialectes par rapport aux langues peu dotées en ressources. Quant à la dernière partie de ce chapitre, elle est dédié à une étude linguistique du dialecte algérois.

Le troisième chapitre de cette thèse s'attelle à présenter la traduction automatique en général et celle des dialectes arabes en particulier. Nous y avons donné un bref historique retraçant son évolution au fil des années. Nous avons par la suite décrit les différentes approches de la traduction automatique à savoir : l'approche linguistique et l'approche empiriques ainsi que l'approche hybride. Nous avons expliqué les principes de base de chaque approche et avons donné des exemples des systèmes de traduction s'y afférant. Une partie de ce chapitre traite la traduction automatique statistique. Nous y avons abordé les fondements mathématiques sur lesquels elle est basée ainsi que les composants essentiels d'un système adoptant cette approche en détaillant chacun d'eux. La partie consacrée à la traduction automatique se termine par une présentation de l'évaluation des systèmes de traduction ainsi que les métriques les plus utilisées. Ce chapitre inclut aussi un état de l'art détaillé sur la traduction des dialectes arabes. Nous avons mis en relief les travaux existant dans ce domaine pour la traduction entre dialectes et arabe standard ainsi qu'entre dialectes et langues étrangères (notamment l'anglais). Tous les travaux ont été décrits en termes d'approche, de données et de résultats

obtenus.

Les chapitres qui suivent sont tous dédiés à nos contributions. Dans le chapitre quatre, nous avons présenté PADIC (Parallel Arabic Dialect Corpus), un corpus parallèle multi-dialectes que nous avons construit dans le cadre de cette thèse. Il s'agit d'une ressource textuelle importante incluant pour la première fois deux dialectes algériens (algérois et bonois), les dialectes tunisien, marocain, syrien et palestinien ainsi que l'arabe standard. Nous avons explicité la méthodologie de construction de ce corpus. La suite de ce chapitre est réservée à une étude analytique où le corpus PADIC a été décrit en détail : ses statistiques, le taux de recouvrement entre les différents vocabulaires au niveau corpus et au niveau phrastique entre les dialectes d'une part et entre les dialectes et l'arabe standard d'autre part. Dans cette même optique et dans le but d'étudier la distance entre les dialectes et l'arabe standard, nous avons utilisé la distance de Hellinger appliquée à un modèle de langage uni-gramme des dialectes utilisés, nous avons donc présenté la démarche adoptée et les résultats obtenus. Ce chapitre se termine par un module d'identification qui permettra de distinguer les dialectes sus-cités et l'arabe standard.

Le chapitre cinq de ce manuscrit est consacré à la traduction automatique statistique des dialectes arabes à la lumière du corpus multi-dialectes PADIC. Nous y avons décrit toutes les systèmes de traduction que nous avons construits ainsi que toutes expérimentations que nous avons conduites dans ce cadre.

Le dernier chapitre quant à lui couvre l'ensemble des ressources construites dans le cadre de ce travail pour le dialecte algérois. Ce chapitre comporte plusieurs parties, chacune dédiée à une ressource. Toutes les parties sont structurées de la même manière : un état de l'art sur la ressource en question, l'approche que nous avons adoptée, la description des données utilisées ainsi que les résultats obtenus. Nous avons d'abord présenté le système de vocalisation automatique que nous

avons développé. Une partie du corpus algérois a été vocalisée manuellement et a servi de données d'apprentissage et de test pour ce système. Le même système a été aussi entraîné sur des corpus arabe standard pour des besoins de comparaison, nous avons présenté tous les résultats obtenus ainsi que leur interprétation. La seconde partie de ce chapitre est dédiée au convertisseur graphème-phonème élaboré dans le cadre de cette thèse, alors que la partie suivante aborde la création de l'analyseur morphologique dédié aux textes en dialecte algérois.

Cette thèse se termine par une conclusion récapitulant l'essentiel de ce qui a été réalisé ainsi que quelques perspectives pour assurer la continuité de ce travail.

Chapitre 2

Les dialectes arabes des «langues» peu dotées en ressources

2.1 Introduction

Ce chapitre inclut trois parties, la première est consacrée à la notion de langues peu dotées en ressources, nous expliquons la problématique relative à ces langues ainsi que leur évaluation. La seconde partie de ce chapitre est dédiée à la langue arabe et ses dialectes, nous y avons donné une brève description en particulier des dialectes objets de notre travail. Dans cette même partie, nous avons tenté d'évaluer ces dialectes par rapport à cette notion de langue peu dotées en ressources. Enfin, ce chapitre se termine par une étude linguistique du dialecte algérois, le dialecte de la capitale Alger et sa périphérie. Nous avons présenté les traits les plus importants relatifs à ce dialecte.

2.2 Définition des langues peu dotées en ressources et leur classification

Les langues peu dotées en ressources [1] sont l'ensemble des langues qui ne disposent (ou disposent de peu) de ressources du point de vue du traitement automatique. Ce terme a été introduit par [Krauwer, 2003], dans le cadre d'un projet piloté par ELSNET [2], et dans lequel le concept BLARK [3](pour Basic Language Resource Kit) a été défini comme étant l'ensemble minimum de

1. En anglais ça correspond aux termes under-ressourced language, less-ressourced language.
2. European Network of Excellence in Human Language Technologies
3. http://www.blark.org/

ressources nécessaires pour le traitement automatique d'une langue donnée. Ce minimum de ressources requises comprend à titre d'exemple :

— Les corpus monolingues et multilingues

— Les corpus vocaux

— Les dictionnaires monolingues et bilingues

— Des outils d'analyse morphologique et syntaxique, d'étiquetage...

— Des outils d'exploitation et d'exploration de corpus

Cette notion a été introduite aussi dans [Berment, 2004] où l'on propose une classification des langues du point de vue de leur ressources TAL en trois classes :

— Les langues-π : les langues peu ou non dotées en ressources

— Les langues-μ : les langues moyennement dotées

— Les langues-τ : les langues très bien dotées

Cette classification est faite en appliquant une technique simple d'évaluation d'une langue du point de vue ressource. Il s'agit pour un groupe de locuteurs (natifs de la langue) d'affecter à chaque type de ressource une note et une criticité (un poids relatif à l'importance de la ressource). La moyenne pondérée de ces notes appelée indice-σ (une note sur une échelle de 20) permettra de classer la langue dans l'une des trois classes citées ci-dessus. Le tableau 2.1 décrit les différentes ressources à noter ainsi que la formule de calcul de l'indice-σ.

Services / ressources		Criticité C_k (0 à 10)	Note N_k (/20)	Note pondérée (C_k, N_k)
Traitement du texte	Saisie simple			
	Visualisation / impression			
	Recherche et remplacement			
	Sélection du texte			
	Tri lexicographique			
	Correction orthographique			
	Correction grammaticale			
	Correction stylistique			
Traitement de l'oral	Synthèse vocale			
	Reconnaissance de la parole			
Traduction	Traduction automatisée			
ROC	Reconnaissance optique de caractères			
Ressources	Dictionnaire bilingue			
	Dictionnaire d'usage			
Total		$\sum C_k$		$\sum C_k N_k$
Moyenne (/20)				$\sum C_k N_k / \sum C_k$

Tableau 2.1 : Évaluation des ressources TAL d'une langue[Berment, 2004]

À la lumière de cette évaluation, les langues peuvent être classées comme indiqué dans le tableau 2.2 :

indice-σ	Valeur	Classification
langue-π	$0 - 0.99$	Peu dotées
langue-μ	$10 - 13.99$	Moyennement dotée
langue-τ	$14 - 20$	Très bien dotée

Tableau 2.2 : Classification des langues en fonction de l'indice-σ

2.3 Problématique

On recense actuellement prés de 7097 langues à travers le monde dont 920 ont disparu ou en voie de disparition. [4]Les langues toujours vivantes différent du point de vue de leur prise en charge en termes de technologie du langage humain. En effet, seules une petite portion de ces langues possède des ressources pour leur

4. La source http://www.ethnologue.com/

traitement automatique et sont prises en charge par les logiciels de traitement de textes, les moteurs de recherche d'information, les traducteurs automatiques, les outils de traitement et de synthèse de la parole, etc. Si l'on prend comme exemple Google, seules quelques 139 langues sont disponibles pour le moteur de recherche, et 103 langues sont présentes dans Google translate.

Cependant, ces langues diffèrent du point de vue de leur prise en charge par les technologies de traitement automatique. Dans ce contexte, MetaNet [5] (Un réseau d'excellence de 60 centres de recherches européens) a publié les premiers résultats de ses travaux dédiés aux langues européennes (MetaNet White paper Series [6]) il en ressort que certaines langues européennes, principalement l'anglais, le français et l'espagnole sont mieux dotées en ressources de traitement automatiques que d'autres langues européennes telles que l'islandais, l'irlandais, le Gallois, le roumain et les langues slaves (l'ukrainien, le Belarussien, le tchèque, le Slovaque, etc). [7] Si telle est la situation pour les langues européennes, qu'en est-il pour les autres langues ? Il n'existe aucune cartographie classifiant les langues du monde en termes de ressources. Néanmoins, il est évident que la majorité des langues africaines et certaines langues de pays asiatiques sont des langues peu dotées en ressources vu leur présence rare dans les applications TAL. En effet, ces langues appartiennent à des pays non influents sur la scène politique et économique internationale.

L'enjeu économique et politique est un facteur qui peut « booster » la recherche scientifique pour la création de ressources TAL. Nous citons par exemple l'intérêt croissant de la communauté scientifique aux langues des pays économiquement influents tels que la chine. Avant l'émergence de cette puissance sur la scène

5. dédié à la promotion des fondements technologiques d'une société multilingue de l'information en Europe, financé par l'Union Européenne

6. http://www.meta-net.eu/whitepapers/overview

7. Plus de détails peuvent être trouvés à l'adresse : http://www.meta-net.eu/whitepapers/key-results-and-cross-language-comparison

internationale, peu de travaux s'intéressaient aux langues chinoises. Du point de vue politique, beaucoup de travaux TAL ont été dédiés à la langue Arabe après les événements du 11.09.2001. Dans cette même optique, Le dialecte Irakien a lui aussi été sujet de plusieurs travaux à cause de la guerre du Golf et ensuite à l'invasion de l'Irak. L'enjeu économique et politique est souvent derrière le financement de projets de recherche d'envergure internationale visant à la création de ressources TAL pour une langue ou un groupe de langues.

A l'ère du développement des nouvelles technologies de l'information, il existe toujours des langues qui manquent de ressources TAL. Par ailleurs, il convient de noter que cette notion de langue peu dotée en ressource est indépendante de la taille de la population. Certaines langues peu dotées sont même des langues officielles dans leurs pays et sont parlées par des populations importantes. En revanche, certaines langues de minorité sont bien dotées en termes de ressources TAL, comme par exemple le Catalan[8] qui est même disponible sur Google Translate[Besacier et al., 2014].

La problématique des langues peu dotées en ressources s'articule autour de la difficulté de la mise en place des systèmes TAL en raison de l'absence des composants basiques sur lesquels de tels systèmes sont implémentés. Le défi de ces langues réside dans le fait que toute application TAL requiert des outils basiques souvent indisponibles. Pour n'en citer que la création de corpus textuel, le scénario suivant reste envisageable même pour les autres types de ressources :

— Comment créer des corpus textuels pour une langue si on n'a pas de fond documentaire électronique à partir duquel on peut puiser l'information ?

— Saisir des quantités de textes à partir de manuscrits ?

— Comment faire si à l'origine cette langue n'est pas écrite ?

8. Langue parlée dans la Catalogne, une communauté autonome en Espagne, située dans le nord-est de la péninsule Ibérique dont la capitale est la ville de Barcelone.

— Dans ce cas, procéder à la collecte d'enregistrement sonore et transcrire le son en texte ?
— Et si cette langue n'a pas de règles orthographiques définies, comment transcrire ?
— Quelles règles établir pour la transcription ?
— Une fois ces règles établies comment les valider ?

La construction des ressources TAL est souvent confrontées à ce genre de questions. Le problème peut-être plus compliqué pour d'autres types de ressources telles que les corpus parallèles et les dictionnaires bilingues. Quant aux modules applicatifs comme les analyseurs morphologiques et syntaxiques ainsi que les générateurs, ils ont aussi leur degré de difficulté car ils requièrent des informations linguistiques encore plus poussées souvent non disponibles dans le cadre des langues pauvres en ressources.

2.4 Les dialectes arabes, des langues peu dotées ?

Avant de présenter la problématique relative aux ressources TAL des dialectes arabes, nous introduisons dans ce qui suit une brève description de la langue arabe et de ses dialectes.

2.4.1 Langue arabe VS dialecte arabe

L'arabe est une langue sémitique utilisée par environ 420 millions de personnes. C'est la langue officielle de 22 pays (répartis entre l'Afrique du nord et l'Asie). L'arabe est un terme générique couvrant trois variétés distinctes :

— L'arabe classique : est principalement défini comme la langue du Coran et celle de la littérature ancienne de la péninsule arabe, mais existe toujours dans certaines productions littéraires contemporaines.

— L'arabe standard moderne : Généralement dénommé MSA (Modern Standard Arabic, Fusha en arabe), une forme modernisée mais peu différenciée de l'arabe classique, retenue comme langue officielle dans tous les pays arabes. L'arabe standard n'est pas acquis en tant que langue maternelle, c'est une langue apprise à l'école, utilisée aussi dans les programmes de télévision et de radio (bulletins d'information), dans les pratiques religieuses et dans la presse écrite [Kirchhoff et al., 2003].

— Les dialectes Arabes : appelés également l'arabe dialectal ou vernaculaires sont des variétés parlées issues de l'arabe avec des emprunts différents selon les régions. Ces dialectes sont influencés à la fois par les anciennes langues locales et par les langues européennes comme le français, l'espagnol, l'anglais et l'italien.[9] La différence entre ces variétés de dialectes à travers le monde arabe peut être assez importantes, elles sont même difficilement compréhensibles sans apprentissage. En effet, au vu des différences entre ces dialectes, ces derniers peuvent être considérés comme des langues disparates dépendant de la région géographique où elles sont parlées. Ainsi, dans la littérature ces dialectes sont décrits du point de vue de la dichotomie Est-Ouest [Hetzron, 1997] :[10]

    — Les dialectes du moyen-orient Il s'agit de l'arabe parlé en péninsule arabe (Pays du Golf et le Yémen), le dialecte du Levant (Syrie, Liban, Palestine et la Jordanie), l'irakien, l'égyptien et le dialecte soudanais.

    — Les dialectes du Maghreb Essentiellement parlés en Algérie, en Tunisie, au Maroc, en Libye et en Mauritanie. Notons que le Maltais est une forme de dialecte arabe parlé à Malte.

9. L'influence des langues européennes est due au fait que la plupart des pays arabes étaient des colonies européennes au cours du thème siècle .

10. Une autre classification est donnée dans [Watson, 2007] où les dialectes arabes ruraux et bédouins se distinguent en raison de la diversité ethnique et sociale des locuteurs Arabes. L'auteur affirme que les dialectes bédouins montrent une tendance conservatrice et homogène, tandis que les dialectes urbains ont tendance à être plus évolutifs.

Après avoir défini les dialecte arabes et au regard de la définition des langues peu dotées évoquée plus haut, peut-on considérer les dialectes arabes comme des langues peu dotées en ressources ?

Les applications dans les domaines de la traduction automatique, de la reconnaissance de la parole, la synthèse de la parole, etc doivent prendre en compte ces dialectes, jusqu'à présents démunis du minimum de ressources informatiques à l'exception d'un nombre réduits (qu'on abordera plus loin dans ce manuscrit). De ce point de vue, nous pouvons considérer les dialectes arabes comme des langues peu dotées en ressources. Par ailleurs, comparés aux autres langues peu dotées en ressources et de par leur nature orale, ces dialectes présentent des défis supplémentaires dont nous citons :

— Ces dialectes sont des parlers locaux qui ne sont pas écrits, ils ne possèdent pas de règles d'écriture normalisant leur transcription en texte. Un mot peut être écrit avec différentes formes orthographiques toutes acceptables puisqu'il n'y a pas de référence.

— Ils sont caractérisés par une flexibilité au niveau lexical et grammatical malgré qu'ils soient des variantes de la langue arabe.

— En plus du fait que ces dialectes sont différents de l'arabe standard, ces dialectes sont différents les uns des autres. Les dialectes du Maghreb différent des dialectes du moyen orient, et même dans un même pays ces dialectes sont différents.

— Ces dialectes sont largement influencés par d'autres langues comme le français, l'espagnole, le turque, et le berbère.

— Le vocabulaire dialectal évolue de façon continue et spontanée, de nouveaux mots apparaissent et sont adoptés implicitement par les locuteurs.

— Le phénomène du code switching qui fait qu'une même phrase peut inclure des mots de deux langues (voir même trois), alterner entre le dialecte, l'arabe

standard, le Français ou l'anglais (et même d'autres langues locales comme le berbère dans les pays du Maghreb) est un phénomène courant dans les conversations quotidiennes .

2.4.2 Aperçu des dialectes utilisés

Nous avons travaillé sur un certain nombres de dialectes dans le contexte de cette thèse. En plus des deux dialectes Algériens (algérois et bonois), le choix des autres dialectes a été guidé par la disponibilité des personnes parlant ces dialectes à collaborer gracieusement dans le cadre de ce projet. Dans ce qui suit, nous donnons un aperçu des dialectes que nous avons utilisés.

Le dialecte dlgérien

Le dialecte algérien est le parler quotidien et informel en Algérie qui généralement n'est pas utilisé dans les discours officiels. Son vocabulaire est relativement le même à travers le pays. Cependant, dans l'est du pays, le dialecte est plus proche du dialecte tunisien alors que dans l'ouest, il se rapproche du dialecte marocain. La plupart des mots du dialecte arabe algérien proviennent de l'arabe standard, mais il y a une variation significative de la vocalisation dans la plupart des cas, et l'omission de quelques lettres dans d'autres cas. Dans le cadre de cette thèse, nous nous sommes intéressés à deux variantes du dialecte algérien : le dialecte algérois (parlé dans la capitale Alger et sa périphérie) et le dialecte bonois parlé à Annaba (une grande ville de l'est Algérien). Il convient de noter que le choix de ces deux variantes est justifié par la disponibilité des personnes qui les parlent. L'auteur de ce manuscrit entre-autre étant algéroise, un focus est mis sur le dialecte algérois plus loin dans ce chapitre.

Le dialecte tunisien

Le dialecte tunisien est un dialecte arabe parlé en Tunisie. Il est relativement proche des dialectes algériens, marocain et libyens puisque tous ces parlers appartiennent tous au consortium du dialecte du Maghreb. Le vocabulaire du dialecte tunisien est principalement arabe, avec des substrats berbères importants et des mots d'origine italienne, française, turque, espagnole et même maltaise. A l'instar de tous les dialectes arabes, le tunisien diffère de l'arabe standard au niveaux phonologique, morphologique, et syntaxique.

Le dialecte marocain

Le dialecte marocain recouvre le parler quotidien de la grande majorité de la population marocaine. Il est utilisé (comme le cas de tous les dialectes arabes) dans les communications non-officielles, les émissions de télévision, le cinéma, etc. Ce dialecte appartient au groupe des dialectes arabes maghrébins, il est influencé par le berbère, le français et l'espagnol.

Les dialectes syrien et palestinien

Les dialectes syriens et palestiniens font partie du Levant arabe, qui couvre également les dialectes du Liban et de Jordanie. Les dialectes Levantins arabes sont assez similaires du point de vue phonologique, syntaxique et lexical. Cependant, d'un point de vue géographique ces dialectes diffèrent. En effet les dialectes parlés dans les milieux urbains diffèrent de ceux parlés en milieux ruraux.

Le dialecte Arabe syrien est influencé par la langue syriaque, une langue sémitique du Moyen-Orient, appartenant au groupe de la langue araméenne. Son vocabulaire inclut une importante proportion de mots arabes et aussi des mots empruntés du turc et du français.

Le dialecte palestinien quant à lui est légèrement différent au niveau phonétique des

dialectes du nord du Levant. En plus des différences entre les dialectes palestiniens parlés en milieu rural et en milieu urbain, on note que le dialecte parlé au sud de la Palestine est différent de celui parlé au nord.

Pour une lecture aisée de ce manuscrit nous désignons les dialectes concernés par respectivement : ALG pour le dialecte algérois, ANB pour le dialecte bonois, TUN pour le tunisien, MAR pour le dialecte marocain, SYR pour le syrien et PAL pour le palestinien. Quant à l'arabe standard on gardera une notation anglo-saxonne MSA (pour Modern Standard Arabic).

2.4.3 Évaluation des dialectes en termes de l'indice-σ

Nous procédons dans cette section à l'évaluation des dialectes décrits ci-dessus à la lumière de la définition des langues peu dotées en utilisant l'indice-σ.

Pour ce faire, nous avons repris le Tableau de l'évaluation 2.1 que nous avons adapté comme indiqué ci-dessous dans le tableau 2.3.

Services / ressources		Criticité C_k (0 à 10)	Note N_k (/20)	Note pondérée (C_k, N_k)
Traitement du texte	Saisie simple et prise en charge des caractères non arabes			
	Visualisation / impression			
	Recherche et remplacement			
	Correction orthographique			
	Correction grammaticale			
	Correction stylistique			
Traitement de l'oral	Synthèse vocale			
	Reconnaissance de la parole			
Traduction automatique	Traduction du texte			
	Traduction de la parole			
ROC	Reconnaissance optique de caractères			
Ressources	Dictionnaire bilingue			
	Dictionnaire d'usage			
	Corpus monolingue			
	Corpus multilingue			
Total		$\sum C_k$		$\sum C_k N_k$
Moyenne (/20)				$\sum C_k N_k / \sum C_k$

Tableau 2.3 : Tableau d'évaluation des ressources TAL d'un dialecte arabe

Les adaptations que nous avons effectuées au niveau du tableau d'évaluation 2.1 se résument comme suit :

1. Traitement de texte :
    - Les critères relatifs à la sélection du texte et au tri lexicographique ont été retirés en raison du fait que ces deux critères s'appliquent aux langues non segmentées (dont les mots ne sont pas délimités par le caractère espace). Ce critère n'est pas applicable aux dialectes arabes.
    - Nous avons ajouté le critère de prise en charge des caractères non arabe utilisés dans les dialectes. Il s'agit des caractères correspondant aux phonèmes :/P/, /V/ et /G/
2. La traduction automatique : Nous avons segmenté la traduction automatique en traduction du texte et de la parole pour une évaluation plus fine de ce volet.
3. Ressources : Nous avons introduit les corpus monolingue et multilingue comme critère d'évaluation vu leur importance dans l'implémentation des outils TAL qui utilisent des algorithmes d'apprentissage.

Le tableau 2.4 présente les résultats de l'évaluation des dialectes concernés par la présente étude.

Dialectes			ALG		ANB		TUN		MAR		SYR		PAL	
Services / ressources		Criticité C_k	Note N_k	Note pondérée	Note N_k	Note pondérée	Note N_k	Note pondérée	Note N_k	Note pondérée	Note N_k	Note pondérée	Note N_k	Note pondérée
Traitement du texte	Saisie simple et Prise en charge des caractères non arabes	8	5	40	5	5	40	40	5	40	5	40	5	40
	Visualisation / impression	6	8	48	8	48	8	48	8	48	8	48	8	48
	Recherche et remplacement	6	8	48	8	48	8	48	8	48	8	48	8	48
	Correction orthographique	6	0	0	0	0	0	0	0	0	3	18	3	18
	Correction grammaticale	6	0	0	0	0	0	0	0	0	2	12	2	12
	Correction stylistique	5	0	0	0	0	0	0	0	0	0	0	0	0
Traitement de l'oral	Synthèse vocale	8	0	0	0	0	0	0	0	0	0	0	0	0
	Reconnaissance de la parole	8	0	0	0	0	0	0	0	0	0	0	0	0
Traduction automatique	Traduction du texte	8	0	0	0	0	0	0	2	16	2	16	2	16
	Traduction de la parole	8	0	0	0	0	0	0	0	0	0	0	0	0
ROC	Reconnaissance optique de caractères	5	5	25	5	25	5	25	5	25	5	25	5	25
Ressources	Dictionnaire bilingue	8	0	0	0	0	0	0	0	0	0	0	0	0
	Dictionnaire d'usage	8	0	0	0	0	0	0	0	0	0	0	0	0
	Corpus monolingue	10	0	0	0	0	3	30	0	0	3	30	0	0
	Corpus multilingue	8	0	0	0	0	0	0	2	16	3	24	0	0
Moyenne (/20)			1,49		1,49		2.06		1,49		2,42		1,92	

Tableau 2.4 : Évaluation des ressources TAL des dialectes arabes utilisés

Dans le processus d'évaluation les poids (criticité telle que appelée dans [Berment, 2004]) relatifs au traitement de la parole, à la traduction automatique ainsi qu'aux ressources sont les plus élevés en raison de l'importance de ces volets dans les applications récentes du TAL. Des poids plus faibles ont été affectés au volet traitement de texte, vu que les dialectes arabes sont pris en charge par les outils de traitement de textes pour l'arabe standard.

Au vue des résultats, il est apparaît clairement que les dialectes arabes que nous avons évalués sont peu dotés en ressources. En effet, tous ces dialectes sont de classe indice-π avec une moyenne maximale de 2.42 (l'indice relatif au dialecte tunisien). On notera que les autres dialectes arabes peuvent être classés dans cette même catégorie car le manque de ressources est observé pour la majorité d'entre eux à l'exception du dialecte égyptien pour lequel un certain nombre de ressources existent, mais on souligne que ces ressources ne sont pas de taille à le classer comme étant moyennement doté (indice-μ) ou très doté en ressources (indice-τ). [11]

2.5 Étude linguistique du dialecte algérois

Comme mentionné plus haut, dans cette section, nous mettons en relief le dialecte algérois en abordant ses différentes caractéristiques. Nous avons décrit l'alphabet sur lequel il se base, ensuite nous nous sommes intéressés à sa morphologie en identifiant les classes de son lexique : les verbes, les noms, les pronoms et les mots outils. Ces différentes classes ont été décrites en mettant en évidence leurs sous-classes et leurs propriétés. Nous avons ensuite entamé la notion de flexion aussi bien pour les noms et les verbes ainsi que les mots outils.

11. Nous attirons l'attention du lecteur à un travail de synthèse intéressant dédié au traitement automatique des dialectes arabes[Shoufan and Al-Ameri, 2015]. Malheureusement, aucun travail pour le dialecte algérien n'a été cité, sauf ceux réalisés dans le cadre de la présente thèse. Le dialecte algérien est en net décalage par rapport aux autres dialectes arabes en termes d'outils TAL. Nous pensons que les causes de ce retard ne sont pas dues uniquement à des raisons scientifiques et techniques, mais des facteurs politiques et économiques peuvent aussi être derrière cela. Nous ne pouvons dans le contexte de cette thèse les discuter.

Par la suite une partie de cette étude a été dédiée à l'aspect syntaxique du dialecte algérois. Nous y avons explicité les types de phrases dans ce dialecte, ainsi que l'ordre des mots dans les phrases dialectales.

Le dialecte algérois est le dialecte arabe parlé à Alger et sa périphérie. Ce dialecte est différent des dialectes parlés dans les autres régions du pays. Ce dialecte n'est pas utilisé dans les établissements scolaires, ni dans les discours officiels ou dans la presse. Le dialecte Arabe Algérien est le parlé de tous les jours d'une vaste majorité d'Algériens [Boucherit, 2002].

2.5.1 L'alphabet

Le dialecte algérois comme tous les dialectes arabes simplifie les règles morphologiques et syntaxiques de l'arabe standard. On notera que ce phénomène n'est pas propre à la langue arabe et ses dialectes mais il est général est constaté pour chaque paire langue standard et dialecte. Dans [Ferguson, 1959],l'un des premiers travaux sur les dialectes arabes du point de vue linguistique, l'auteur décrit l'importance des différences entre l'arabe écrit et le dialecte dans différents niveaux du langage : « les différences phonologiques entre l'arabe classique et l'arabe parlé sont modérées (comparées à d'autres paires langue-dialecte), alors que les différences grammaticales sont les plus frappantes. Au niveau lexical, les différences sont marquées par la variation des formes avec des usages et des sens distincts ».

En effet, au niveau phonologique, le dialecte algérois est caractérisé par la majorité des traits relatifs à l'arabe. En plus des 28 phonèmes consonantiques de l'arabe[12](voir le Tableau 2.5), le système consonantique du dialecte inclut des phonèmes non-arabes : /*g*/ comme dans le mot قَاع (tout), et les phonèmes /*p*/

12. incluant trois voyelles longues (ا *ā*, و *w* et ي *y*).

et /V/ utilisés essentiellement dans les mots empruntés au Français comme le cas du mot پُومپة (adapté du mot français « pompe ») et le mot فَليزة (adapté du mot Français « valise »). Il convient aussi de noter que l'utilisation des phonèmes (ظ) et (ذ) est très rare, la plupart du temps ظ est prononcé /d‘/(ض) et ذ est prononcé /d/(د). Il en est de même pour le /T/ (ث) qui est prononcé /t/(ت). Ces deux dernières substitutions sont aussi observées pour le dialecte Jordanien [Amer et al., 2011]. (voir les détails dans le chapitre 6).

Lettre	Phonème	Lettre	Phonème	Lettre	Phonème
أ	/?/	ز	/z/	ق	/q/
ب	/b/	س	/s/	ک	/k/
ت	/t/	ش	/S/	ل	/l/
ث	/T/	ص	/s‘/	م	/m/
ج	/Z/	ض	/d‘/	ن	/n/
ح	/x/	ط	/t‘/	ه	/h/
خ	/X/	ظ	/D‘/	و	/w/
د	/d/	ع	/?‘/	ج	/j/
ذ	/D/	غ	/G/		
ر	/r/	ف	/f/		
ـَ	/a/	ـِ	/i/	ـُ	/u/
ا	/a :/	ى	/i :/	و	/u :/

Tableau 2.5 : Les phonèmes Arabe avec SAMPA [13]

2.5.2 Le vocabulaire

Le vocabulaire du dialecte algérois est inspiré de l'arabe dont les mots ont été phonologiquement altérés, avec des substrats berbères, et de nombreux mots empruntés au français, au turc et à l'espagnol. Même si ce vocabulaire est principalement arabe, il inclut des variations significatives au niveau de la

13. Nous utilisons SAMPA (Speech Assessment Methods Phonetic Alphabet) pour la représentation des phonèmes, http://www.phon.ucl.ac.uk/home/sampa/index.html.

voyellation dans la plupart des cas, ainsi que l'omission ou la modification de certaines lettres dans d'autres cas (principalement la Hamza) [14]. Le vocabulaire du dialecte algérois comporte quatre classes à savoir les verbes, les noms, les pronoms et les particules. Dans la section qui suit, nous donnons une description de chaque classe et éventuellement les sous-classes qui lui sont adjacentes.

— Les verbes

Certains verbes du dialecte algérois adoptent le même schème que les verbes en arabe standard en respectant la même vocalisation, comme dans le cas du verbe سَمَّى (nommer) ou سلّم (saluer). D'autres verbes sont prononcés différemment des verbes arabes correspondants en adoptant différentes marques diacritiques comme le cas des verbes شْرَبْ (boire) vs شَرِبَ en arabe standard . Une autre catégorie de verbes est obtenue par l'omission ou la modification de certaines lettres. Dans le tableau 2.6 nous donnons quelques exemples de chaque cas cité.

Verbe dialectal	Verbe correspondant en Arabe	Sens	Case
سَلَّم	سَلَّم	saluer	même schème
قَابَل	قَابَل	confronter	même signes diacritiques
شْرَبْ	شَرِب	boire	même schème
كْتَبْ	كَتَبَ	écrire	Différents signes diacritiques
جَا	جَاء	venir	Omission ou modification de lettres
بْقَى	بَقِيَ	rester	
كْلَا	أَكَلَ	manger	
كَمَّلْ	أَكْمَلَ	finir	

Tableau 2.6 : Exemples de différence de schèmes verbaux entre le dialecte algérois et l'arabe standard

Une autre catégorie de verbes du dialecte algérois sont ceux issus de langues étrangères spécialement le Français comme les cas des verbes شَارجَا correspondant au verbe « charger » et قَارَا altération du verbe « garer ».

14. La Hamza est une lettre de l'alphabet arabe, représentant la consonne occlusive glottale.

— Les noms

Les noms du dialecte algérois peuvent être primitifs (c'est à dire non dérivés d'aucune racine verbale) ou dérivés de verbe comme les noms verbaux et les participes (actifs et passifs), voir le Tableau 2.7 où des exemples illustratifs sont donnés.

Verbe	Nom Verbal	Participe actif	participe Passif
باع	بيع	بايع	مبيوع
vendre	vente	vendeur	vendu

Tableau 2.7 : Exemples de mots dialectaux dérivés de racine verbale

Il convient de noter que les noms du dialecte algérois comportent une importante proportion de mots Français. La majorité d'entre eux est le résultat d'altération phonologique comme موتور (moteur), لاطونسيون (la tension) et پوليس (policier).

Les noms dialectaux incluent aussi les nombres représentant les unités, les dizaines, les centaines, etc. De un à dix, les chiffres en algérois sont les mêmes que les chiffres Arabes (moyennant une légère modification des signes diacritiques), à l'exception des chiffres zéro et deux : le premier est prononcé comme en Français /zero/, et le second est زوج alors qu'en arabe c'est إثنين. Pour les nombres de onze à dix-neuf, la prononciation diffère de celle de l'arabe standard avec altération des lettres et de signes diacritiques. Cependant, le nombre peut être facilement perçu par un locuteur arabe. Les nombres supérieures à vingt sont proches de l'arabe standard avec aussi une légère modification des voyelles.

— Les pronoms

La liste des pronoms est une liste fermée ; elle contient les pronoms démonstratifs et personnels. Pour les pronoms relatifs, il n'y a qu'un seul en dialecte algérois أَلّى (qui) ; Ce pronom est utilisé pour le féminin, le

masculin, le singulier et le pluriel. Nous donnons dans les tableaux 2.8 and 2.9 tous les pronoms du dialecte algérois

	Singulier		Pluriel
	Féminin	Masculin	Féminin & Masculin
1ère Personne	(je) أنا		(Nous) حنا
2ème Personne	(Tu) انتِ	(Tu) انتَ	(Vous) انتوما
3ème Personne	(Elle) هى	(Il)هو	(Elles ou Ils) هوما

Tableau 2.8 : Les pronoms Personnels du dialecte algérois.

Il est important de signaler que le duel n'existe pas dans le dialecte algérois, il n'y a pas d'équivalent dans ce dialecte des pronoms arabes أنتما (seconde personne, duel) et هما (troisième personne, duel). Identiquement, les pronoms personnels correspondant au féminin pluriel أنتنّ et هنّ relatifs respectivement à la seconde et troisième personne n'existent pas en dialectes algérois.

Singulier		Pluriel
Féminin	Masculin	Féminin & Masculin
(cette/celle)هادى	(ce/cet/celui) هادا	(ces/celles/ceux) هادو
(celle-ci/ celle-là)هاديک	(celui-ci/celui-là) هاداک	(ceux-ci/ceux-là) هادوک

Tableau 2.9 : Les pronoms démonstratifs du dialecte algérois.

— Les particules

Les particules sont utilisées pour situer des faits et des objets relativement au lieu et au temps. Elles sont classées en différentes catégories :

- Les prépositions (فى dans, على sur, بـ avec)
- Les conjonctions de coordination (و et, أومبعد après)
- Les quantificateurs (كل, كلش, قاع tout/tous, شوية, peu)

2.5.3 La flexion

On s'intéresse dans cette section à l'aspect flexionnel du dialecte algérois, ce trait caractéristique aussi de la langue arabe. En effet, la flexion des mots

dans le dialecte algérois intervient pour exprimer un ensemble d'informations grammaticales. La flexion des verbes plus communément connue par la conjugaison est utilisée pour déterminer le temps, la voix (passive ou active), la personne, le nombre et le genre. La déclinaison : la flexion des noms, des adjectives et des pronoms exprime aussi la personne, le nombre et le genre.

La conjugaison des verbes

La conjugaison des verbes en dialecte algérois est affectée (comme pour l'arabe standard) par la personne (première, deuxième ou troisième personne), le nombre (singulier ou pluriel), le sexe (féminin ou masculin), le temps (passé, présent ou futur), et la voix (active ou passive). Ce dialecte utilise des formes similaires à l'arabe standard, à savoir :

— Le passé : Ses formes sont obtenues par suffixation (ajout de suffixes relatifs au nombre et au genre à la racine verbale) ainsi que par la modification des signes diacritiques (voir le Tableau 2.10).

Pronoms		ALG	MSA	Sens
1ère Personne	أنا	كْتَبْتْ	كَتَبْتُ	j'ai écrit
	حنا	كْتَبْنَا	كَتَبْنَا	nous avons écrit
2ème Personne	أنتِ	كْتَبْتِي	كَتَبْتِ	tu as écrit
	أنتَ	كْتَبْتْ	كَتَبْتَ	tu as écrit
	أنتوما	كْتَبْتُو	كَتَبْتُمْ	vous avez écrit
3ème Personne	هى	كَتْبَتْ	كَتَبَتْ	elle a écrit
	هو	كْتَبْ	كَتَبَ	il a écrit
	هوما	كَتْبُو	كَتَبُوا	ils/elles ont écrit

Tableau 2.10 : Conjugaison du verbe كتب au passé

— Le présent et le futur : La forme du présent en Arabe dialectal Algérien est obtenue par affixation : les préfixes يـ, نـ et تـ et les suffixes ى and و (Tableau

2.11). Le verbe peut être précédé par la particule راه (avec ses formes fléchies [15]) pour exprimer une action qui est en cours. La forme du futur est obtenue de la même façon que le présent (les mêmes préfixes et affixes) mais avec l'antéposition d'une particule ou d'une expression indiquant le futur comme أومبعد (après) or غدوا (demain), le mois prochain, ...etc

Pronoms		ALG	MSA	Sens
1ère Personne	أنا	نَلْعَبْ	أَلْعَبُ	je joue
	حنا	نَلَعْبُو	نَلْعَبُ	Nous jouons
2ème Personne	أنتِ	تَلَعْبِى	تَلْعَبِينَ	tu joues
	أنتَ	تَلْعَبْ	تَلْعَبُ	tu joues
	أنتوما	تَلَعْبُو	تَلْعَبُونَ	vous jouez
3ème Personne	هى	تَلْعَبْ	تَلْعَبُ	elle joue
	هو	يَلْعَبْ	يَلْعَبُ	il joue
	هوما	يَلَعْبُو	يَلْعَبُونَ	ils/elles jouent

Tableau 2.11 : Conjugaison du verbe لعب au présent.

— L'impératif : il exprime l'ordre et il est utilisé (comme l'arabe standard) avec la seconde personne. Il est obtenu généralement par l'ajout du préfixe أ et des suffixes ى and و au verbe.

Pronoms	ALG	MSA	Sens
أنتِ	أَخْرْجِى	أُخْرُجِى	Sors (tu, singulier, féminin)
أنتَ	أُخْرَجْ	أُخْرُجْ	Sors (tu, singulier, masculin)
أنتوما	أَخْرْجُو	أُخْرُجُوا	Sortez (vous, pluriel, féminin & masculin)

Tableau 2.12 : Conjugaison du verbe خرج à l'impératif

15. voir la section 2.5.3

La déclinaison du nom

En arabe standard, la déclinaison du nom possède trois cas : le nominatif (الرفع), le génitif (الجر), et l'accusatif (النصب). Ces trois cas du nom sont utilisés pour indiquer sa fonction grammaticale dans la phrase. La plupart des mots singuliers sont suffixés respectivement par les voyelles courtes ـُ, ـِ et ـَ afin de se mettre aux trois cas cités plus haut. [16] Dans le dialecte algérois, ces marques du cas sont supprimées à l'instar des autres dialectes arabes ; la disparition des voyelles courtes finales et du /h/ (ة) dans certaines conditions dans de nombreux dialectes arabes sont des changements très significatifs [Ferguson, 1957] (par rapport à l'arabe standard). Le même auteur dans [Ferguson, 1959] stipule que l'arabe classique possède pour le nom trois cas, les dialectes n'en ont pas. Donc une des caractéristiques du dialecte arabe algérois est qu'il n'accepte pas la déclinaison à trois cas pour les noms singuliers et les adjectifs.

Pour la déclinaison des noms au pluriel, le dialecte algérois possède les mêmes classes du pluriel que l'arabe standard :

1. Le pluriel masculin régulier : il est toujours obtenu par la post-fixation du mot au singulier par ين, contrairement à l'arabe standard où le mot (en fonction de sa catégorie grammaticale) est post-fixé par le suffixe ون (pour le nominatif), et ين (pour l'accusatif et le génitif). Par exemple, le pluriel régulier du mot arabe معلم (enseignant) pourrait être معلمون (nominatif) ou معلمين (accusatif ou génitif) suivant sa fonction dans la phrase. Alors que, par exemple le mot dialectal صابر (patient) prend toujours صابرين (patients) pour le pluriel régulier quelle que soit sa catégorie grammaticale.

2. Le pluriel féminin régulier : Il est obtenu par l'ajout du suffixe ات au mot singulier sans altérer sa structure comme pour l'arabe standard avec une seule

16. Les voyelles doubles (ـً, ـٌ, ـٍ) expriment aussi les mêmes marques du même du cas ainsi que l'indéfinition nominale.

différence dans la marque du cas. En effet, en arabe standard le pluriel féminin régulier possède les marques اتُ ou اتٌ pour le nominatif et اتِ ou اتٍ pour l'accusatif et le génitif. Le dialecte algérois possède une seule marque pour tous les cas اتْ. Par exemple en arabe standard, le pluriel du mot جميلة (belle) est جميلاتُ ou جميلاتِ (belles)[17] dépendant de la fonction du mot dans la phrase, tandis que le pluriel du mot dialectal شابة (belle) est toujours شاباتْ (belles) quelle que soit sa catégorie grammaticale .

3. Le pluriel brisé ou interne : il s'agit d'une forme du pluriel qui modifie la structure du mot singulier pour avoir son pluriel. Comme pour l'arabe standard, ce pluriel obéit à plusieurs règles dépendant du schème du mot singulier.

Dans le tableau 2.13, on donne un exemple pour chaque type de pluriel évoqué plus haut ainsi que son équivalent en arabe standard.

Pluriel	Dialecte algérois		Arabe standard		Sens
	Singulier	Pluriel	Singulier	Pluriel	
Masculin régulier	فلّاح	فلّاحين	فلّاح	فلّاحين/فلّاحون	Fermier/Fermiers
Marque du cas	pas de voyelles	ين	ـَ, ـُ, ـِ , ـً, ـٌ, ـٍ	ين/ون	
Féminin régulier	طبيبة	طبيبات	طبيبة	طبيبَاتِ/طبيبَاتُ	Docteur/Docteurs
Marque du cas	pas de voyelles	ات	ـَ, ـُ, ـِ , ـً, ـٌ, ـٍ	اتٍ/اتٌ اتِ/اتُ	
Irrégulier	طير	طيور	طير	طيور	Oiseau/oiseaux
	يوم	ايّامات/ايّام	يوم	أيّام	Jour/Jours
Marque du cas	pas de voyelles	pas de voyelles	ـَ, ـُ, ـِ , ـً, ـٌ, ـٍ	ـَ, ـُ, ـِ , ـً, ـٌ, ـٍ	

Tableau 2.13 : Exemples des formes du pluriel dans le dialecte algérois

Une différence importante dans la déclinaison du nom en dialecte algérois par rapport à l'arabe standard est l'absence du duel (une sorte de pluriel qui désigne deux items). En effet, en arabe standard par exemple le duel du mot وَلَد (un garçon) est وَلَدان (deux garçons), le mot est post-fixé par ان or ين en fonction du cas.[18]. En Arabe dialectal algérois, le duel est généralement obtenu avec le mot

17. جميلاتٌ et جميلاتٍ aussi.
18. ان pour le nominatif et ين pour l'accusatif et le génitif.

زوج (deux) suivi du pluriel (Féminin ou masculin) du nom ou de l'adjective. [19] Par exemple, le duel du mot وْلَد (un garçon) est زوج وْلاد (deux garçons).

2.5.4 Le niveau syntaxique

La forme déclarative

L'ordre des mots d'une phrase déclarative dans le dialecte algérois est relativement flexible. En effet, dans l'usage courant les phrases de ce dialecte peuvent commencer par le verbe, le sujet ou même l'objet. Cet ordre est basé sur l'importance donnée par le locuteur à chacun de ces constituants. Dans le tableau 2.14, nous donnons un exemple de différents ordres de mots pour une même phrase.

Ordre	Phrase dialectale	Sens
SVO	الولد راح لـلمسيد	Le garçon va à l'école
VSO	راح الولد لـلمسيد	
OVS	لـلمسيد الولد راح	
OSV	الولد لـلمسيد راح	

Tableau 2.14 : Exemple de l'ordre des mots dans une phrase en dialecte algérois

Il convient de noter que l'ordre le plus courant dans les conversations quotidiennes est (SVO)[Souag, 2006] même si les autres sont aussi permis. [20] que les deux premières formes (SVO, VSO) sont les plus utilisées dans les conversations courantes.

La forme Interrogative

En arabe dialectal algérois, n'importe quelle phrase peut être transformée en question, dans l'une des manières suivantes :

19. On note une exception pour les mots tels que عينين (deux yeux), ودنين (deux oreilles), يدين (deux mains), etc.

20. L'ordre SVO est aussi le plus courant pour les dialectes marocain[Ennaji, 2005] et tunisien[Mahfoudhi, 2002].

1. En utilisant un ton de voix interrogatif comme dans : راح تقرا؟ ou رايح تقرا؟(tu vas réviser ?).
2. Par l'introduction d'un pronom ou d'une particule d'interrogation comme dans : وين راح تقرا؟ (où est-ce-que tu vas réviser ?).

Nous donnons dans le tableau 2.15 les particules et pronoms interrogatifs les plus utilisés en dialecte algérois. Il est important de signaler l'existence de la particule ياك utilisée dans les questions supportant une réponse de oui ou non.

Dialecte algérois	Arabe standard	Sens
شكون	من	qui
أما	أى	quel/quelle
وين	أين	où
منين	من أين	d'où
واش / واشن	ماذا	quoi
باش	بماذا	avec quoi
فاش	فى ماذا	en quoi
وقتاش	متى	quand
وعلاش	لماذا	pourquoi
كفاش	كيف	comment
شحال	كم	combien

Tableau 2.15 : Pronoms et particules interrogatifs en dialectes algérois et leurs équivalents en arabe standard.

La forme négative

Les particules ما et ماشى sont généralement utilisées pour exprimer la négation. ما est utilisée en arabe standard et en dialecte, mais les formes de la négation différent entre les deux variantes, alors que ماشى est spécifique au dialecte algérois. En utilisant ces particules, la forme négative est obtenue de différentes manières.

— La négation avec la particule ما

1. En ajoutant les affixes ما et ش au verbe conjugué (ما comme préfixe ش comme suffixe).

2. En utilisant la particule راه équivalente à l'auxiliaire être au présent de l'indicatif, la négation est obtenue en ajoutant les affixes ما et ش à la particule راه éventuellement combinée avec un pronom personnel. On notera que cette particule ne peut être considérée comme un verbe car elle ne peut être conjuguée à aucun temps.

— La négation avec la particule ماشى

3. La particule ماشى est ajoutée au début d'une phrase verbale déclarative sans aucune modification de la phrase.

4. La particule ماشى peut être ajoutée au début d'une phrase verbale déclarative en introduisant le pronom relatif أَلّى.

5. Dans le cas d'une phrase nominale, ماشى peut être introduite au début en inversant l'ordre de ses constituants.

6. ماشى peut aussi être introduite au milieu d'une phrase nominale sans aucune modification de celle-ci.

Dans le tableau 2.16 des exemples sont donnés pour chaque cas cité.

Cas	Dialecte algérois	Arabe standard	Sens
1	لعبت	لعبت	elle a joué
	ما لعبتش	لم تلعب / ما لعبت	elle n'a pas joué
2	راهى مريضة	إنها مريضة	elle est malade
	ما راهيش مريضة	ليست مريضة	elle n'est pas malade
3	هوما كتبو	هم كتبوا	ils écrivaient
	ماشى هوما كتبو	ليسوا هم من كتبوا	Ce ne sont pas eux qui ont écrit
4	هوما كتبو	هم كتبوا	ils écrivaient
	ماشى هوما ألّى كتبو	ليسوا هم الذين كتبوا	Ce ne sont pas eux qui ont écrit
5	الولد مريض	الولد مريض	le garçon est malade
	ماشى مريض الولد	ليس الولد بمريض	le garçon n'est pas malade
6	الولد مريض	الولد مريض	le garçon est malade
	الولد ماشى مريض	الولد ليس مريضا	le garçon n'est pas malade

Tableau 2.16 : Exemples de phrases déclaratives avec leur négation en dialecte algérois

2.6 Conclusion

Nous avons abordé dans ce chapitre la notion de langues peu dotées en ressources ainsi que leur évaluation en termes de l'indice-σ. Nous avons abordé la difficulté de création de ressources TAL pour une langue donnée. Par la suite, nous avons présenté les dialectes arabes en commençant par une brève définition de la langue arabe. Une attention particulière a été accordée aux dialectes concernés par notre travail. A la lumière de la notion de langues peu dotées en ressources, nous avons procédé à l'évaluation de ces dialectes. En effet, nous avons montré en termes de l'indice-σ que ces dialectes sont pauvres en ressources. Enfin, La dernière partie de ce chapitre a été dédié à l'étude linguistique du dialecte algérois. Nous y avons abordé les différents niveaux du langage relatifs à ce dialecte. Nous avons mis en relief les caractéristiques les plus importantes et les plus distinctives de ce dialecte.

Chapitre 3

La traduction automatique, cas des dialectes arabes

3.1 Introduction

La traduction automatique est l'une des problématiques les plus redoutables et les plus passionnantes dans le traitement automatique des langues. Bien que les premiers travaux sur ce thème datent du milieu du 20^{me} siècle, les résultats de la traduction automatiques sont encore loin de la qualité de la traduction humaine. Nous abordons dans une première partie de ce chapitre la traduction automatique. Nous consacrons une petite partie à son histoire, ensuite nous abordons ses différentes approches en citant les travaux les plus connus dans la littérature pour chacune d'entre elles. Nous mettons l'accent ensuite sur la traduction automatique statistique, puisque tous les systèmes présentés dans le cadre de cette thèse se basent sur cette approche. Nous expliquons ses fondements mathématiques ainsi que les composants essentiels d'un système de traduction statistique. Enfin, cette partie est close par l'évaluation des systèmes de traduction. On y présente les métriques les plus utilisées.

La seconde partie de ce chapitre est consacrée aux travaux dédiés à la traduction automatique des dialectes arabes. Nous avons choisi de tracer un état de l'art détaillé de tous les efforts fournis dans ce domaine, car jusqu'à présent aucun travail de synthèse n'a été réalisé sur ce thème important. En effet, dans la littérature, il n'existe que des travaux de synthèse relatifs à la traduction de la langue arabe, principalement entre les paires arabe-anglais et arabe-français. Nous verrons tout au long de cette partie que les recherches dans le domaine de la

traduction des dialectes exploitent d'une manière ou d'une autre la proximité entre ces dialectes et l'arabe standard et tentent d'utiliser ou d'adapter les ressources disponibles pour cette langue au profit des dialectes.

3.2 La traduction automatique

3.2.1 La traduction automatique, un peu d'histoire

La traduction automatique est le processus de traduction d'un texte ou d'une conversation audio d'une langue (source) vers une autre langue (cible) sans intervention humaine. Elle n'est pas à confondre avec la traduction assistée par ordinateur où l'intervention humaine est nécessaire. Les outils de traduction automatique existant actuellement (Google translate, Microsoft Bing, Systran,...) sur la toile et laissent penser que l'histoire de la traduction automatique est récente, alors que les premiers essais de traduction automatiques ont commencé vers 1933 : le savant Russe P. P. Smirnov-Trojanskij et l'ingénieur Français G. Artsruni ont déposé (chacun de son côté) un brevet pour leurs travaux sur la traduction automatique[Buchmann, 1984].

Durant les deux décennies suivantes, les tentatives de traduction automatiques continuent. Le fait marquant de cette période fut le célèbre mémorandum du scientifique Américain Warren Weaver[Warren, 1955] intitulé simplement « Translation » qui a suscité l'intérêt de la communauté scientifique à la traduction automatique et a donné lieu à plusieurs projets de recherches notamment aux Etats-Unis et en Grande Bretagne. En 1952, la première conférence sur la traduction automatique a eu lieu au MIT (Massachusetts Institute of Technology). En 1954, IBM et Geogetown University présentent une démonstration de traduction automatique de l'Anglais vers le Russe. il s'agissait de traduire une soixantaine de phrases utilisant un vocabulaire de quelques 250 mots et 6 règles syntaxiques. La limite de cette expérience n'est pas à démontrer (à l'heure

actuelle), mais à cette époque-là ce travail a eu beaucoup d'impact sur la presse. Il faut noter que cet intérêt est motivé en premier lieu par des motivations politiques et militaires (la guerre froide) et dans une même optique l'intérêt américain aux projet spatiaux soviétiques notamment après l'avènement du spoutnik en 1957. L'utilité de traduire les documents scientifiques Russes devenait plus que nécessaire. Parallèlement, en ex-URSS, l'Académie des sciences initie des travaux de traduction automatiques dirigés par D.Y. Panov. En Grande Bretagne les essais de A.D. Booth ont été largement médiatisés.

En 1959, lors de la conférence sur le traitement numérique de l'information à Paris, le système de traduction japonais Yamato a été présenté, il traduisait de l'anglais vers le japonnais.

Les systèmes de cette période sont qualifiés de « première génération »[Vauquois, 1979]. Il s'agissait d'une traduction mot-à-mot en utilisant un dictionnaire constitué sur la base d'un corpus. L'analyse syntaxique ne dépasse pas la reconnaissance automatique des parties du discours des mots ainsi que des fonctions syntaxiques des syntagmes[Villard, 1989].

Un autre fait marquant dans le domaine de la recherche en traduction automatique était le rapport ALPAC (Automatic Language Processing Advisory Committee). Il s'agit d'une commission d'évaluation des résultats obtenus par les groupes de recherches en traduction automatique financés par le gouvernement Américain. Le rapport ALPAC (publié en 1966) était accablant et jugeait que les résultats obtenus par les différents groupes de recherches n'étaient pas à la hauteur de leurs financements. À la lumière de ces résultats, beaucoup de projets ont vu leurs financements diminués voir même supprimés. Il convient de noter que même si ce rapport concernait la situation de la traduction automatique aux Etats-Unis, ces recommandations ont eu un impact négatif sur le financement des projets de recherches en traduction automatique partout dans le monde.

Durant la décennie post-ALPAC, quelques projets continuent comme même mais sans l'engouement et l'enthousiasme de l'époque d'avant ALPAC. Au Canada, dans le cadre du projet TAUM (Traduction Automatique de l'Université de Montréal) le système Météo a été développé pour traduire des textes de prévisions météorologiques de l'anglais vers français[Chandioux, 1976].

En France, à l'université de Grenoble un système de traduction de textes mathématiques et de physique russes vers le Français[Vauquois, 1975, Vauquois, 1979] a été mis au point. Le système était basé sur la notion d'interlingue : l'introduction d'un langage pivot. Un modèle similaire METAL a été développé à l'université de Texas durant les années 70 et qui traduisait initialement de l'allemand vers l'anglais.

Un autre exemple de réussite est le système Systran[Toma, 1976b, Toma, 1976a, Toma, 1978], sa première version a été installée en 1971 pour l'armée américaine pour la traduction entre le russe et l'anglais (dans les deux sens). En 1976, la communauté européenne achète la version anglais-français. Systran est le premier système commercialisé en traduction automatique. Jusqu'à la fin des années 80, l'approche à base de règles était adoptée par la quasi-totalité des systèmes de traduction. La dominance de cette approche a été interrompue par l'apparition de nouvelles méthodes basées sur l'apprentissage à partir de corpus de données. Les premiers modèles proposés étaient des systèmes à base d'exemples[Nagao, 1984] qui traduisaient à partir d'exemples de traduction appris sur des corpus parallèles.

La fin des années 80 a connu un développement notable qui a donné un autre souffle à la traduction automatique, IBM[Brown et al., 1988, Brown et al., 1990] présentait une approche purement statistique qui se basait elle aussi sur l'apprentissage à partir de corpus de données. Le succès de cette approche n'est pas à prouver, c'est toujours un domaine de recherche florissant de la traduction automatique. Les firmes internationales telles que Google, Microsoft,

IBM(toujours), et même Facebook financent des projets de traduction automatique de grande envergure basés sur des modèles statistiques. ces firmes tentent de couvrir le maximum de langues et se tournent même vers les langues de minorité et vers les dialectes. Le succès de la traduction automatique statistique est dû en grande partie à la simplicité et la rapidité de la mise en œuvre de ces systèmes et la disponibilité des outils open source pour leur implémentation. Plus loin dans ce chapitre, la traduction automatique statistique sera présentée en détails.

3.2.2 Approches de la traduction automatique

La traduction automatique est le processus de traduction à partir d'une langue source vers une langue cible sans aucune intervention humaine. Pour ce faire, la plupart des travaux dans le domaine se basent sur l'une des deux approches : approche linguistique ou approche empirique. L'approche linguistique est fondée sur des connaissances linguistiques des langues source et cible. Elle suppose la maîtrise parfaite des deux langues pour l'implémentation du système de traduction. L'approche empirique quant à elle, ne nécessite aucune connaissance linguistique préalable. Le processus de traduction se résume en l'apprentissage à partir des corpus parallèles.

Approche linguistique

1. Approche à base de règles Les systèmes à base de règles (RBMT, Rule Based MT) requièrent des connaissances linguistiques approfondies des deux langues (source et cible) pour pouvoir établir des règles de transfert pour la traduction. Les premiers systèmes à base de règles adoptaient une traduction directe qui se limitait à une analyse morphologique simple de la phrase source. La traduction se faisait mot par mot à l'aide d'un dictionnaire bilingue. Les mots traduits sont alors réordonnés en utilisant des règles simples pour obtenir une

phrase dans la langue cible (il s'agit du niveau basique du triangle de Vauquois présenté dans la figure 3.1). L'approche directe ayant vite montré ses limites a

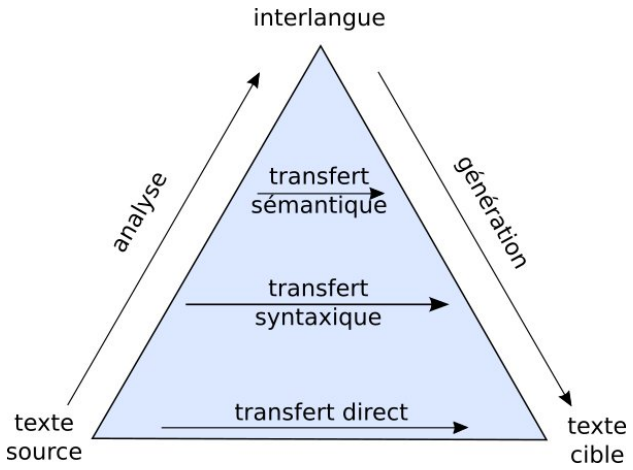

FIGURE 3.1 : Le triangle de Vauquois[Vauquois, 1968]

donné lieu à une autre génération de systèmes à base de règles fondés sur des niveaux d'analyse et génération plus fin des phrases. La traduction se fait en trois étapes :

1. Une phase d'analyse syntaxique et/ou sémantique de la phrase source qui générera une représentation structurale de celle-ci.
2. Une phase de transfert qui convertit la représentation structurale de la phrase cible en une représentation dans la langue cible en utilisant des règles de transfert.
3. Une étape de génération qui génère la phrase cible à partir sa représentation (résultat de la phase 2).

Cette approche bien qu'elle paraissait intuitive présente un certains nombre d'inconvénients : elle requiert en plus des connaissances profondes des langues source et cible des analyseurs syntaxiques et sémantiques performants qui doivent tenir compte de tous les phénomènes langagiers ; cette couverture est quasi-impossible vu l'évolution continue des langues. Aussi, un système développé en utilisant cette approche n'est applicable que pour la paire de langue pour

laquelle il a été conçu. Certains systèmes bien connus dans la littérature adoptent l'approche à base de règles à l'instar de Systran, Météo (voir section 3.2) et Eurotra[King, 1981].

2. Approche inter-langue Cette approche est fidèle à l'objectif de la traduction en général qui est l'élimination des barrières entre les langues. En effet, elle suppose l'existence d'un langage formel qui jouerait le rôle de pivot pour passer d'une langue à une autre. Le principe de l'interlangue est de représenter le sens de la phrase source dans un langage formel indépendamment de la langue utilisée et de générer par la suite la phrase dans la langue cible à partir de cette même présentation. L'avantage majeur de cette approche est la facilité d'intégrer de nouvelles paires de langues dans un tel système de traduction, il suffit de développer les modules nécessaires pour la représentation interlangue des deux langues. Cependant, son inconvénient est de taille : il est presque impossible de trouver un langage formel dont les concepts sont capables de représenter le sens d'une phrase dans n'importe quelle langue d'une part et de représenter tous les phénomènes langagiers relatifs à une même langue d'autre part. Parmi les systèmes de traduction basés sur l'interlangue on cite : DLT (pour Distributed Language Translation)[Witkam, 1988] système de traduction multilingue comprenant douze langues européennes, le projet Rosetta [Landsbergen, 1989, Rosetta, 1994] initialement pour l'anglais et le néerlandais ensuite pour l'espagnole et le néerlandais et KANT (Knowledge-based Accurate Natural-language Translation)[Mitamura et al., 1991] développé par une équipe de l'université de Carnegie-Mellon.

Approche empirique L'évolution rapide de la puissance des calculateurs d'une part et la vulgarisation de l'utilisation de l'informatique d'une autre part a donné lieu à la création de quantités importantes de données textuelles exploitables pour

diverses fins logicielles principalement dans le domaine du TAL. La traduction automatique basée sur une approche data a vu le jour vers le début des années 90. L'idée centrale sur laquelle cette approche repose est l'exploitation de corpus de textes parallèles pour apprendre les modèles de traduction. Cette approche ne requière a priori aucune connaissance linguistique contrairement à l'approche à base de règles de transfert et l'approche inter-langue. Cependant, elle suppose la disponibilité de corpus de taille importante pour une couverture maximale. L'approche empirique de la traduction automatique inclut : la traduction à base d'exemples et la traduction statistique.

1. Approche à base d'exemples On l'appelle aussi la traduction par analogie ou par recombinaison de fragments déjà traduits (EBMT, exemple Based MT). Cette approche modélise le processus de traduction en simulant le raisonnement humain. En effet, pour traduire une phrase ou un texte l'être humain se réfère toujours à des exemples qu'il connait déjà en essayant de repérer des phrases, des segments ou même des mots qu'il a déjà rencontrés. Le processus de traduction à base d'exemples se déroule en trois étapes : une première étape de correspondance qui consiste à rechercher dans le corpus parallèle d'apprentissage les séquences sources qui se rapprochent le plus aux séquences à traduire. La seconde étape, celle de l'extraction des séquences cibles associés aux séquences sources précédemment identifiées. La dernière étape est celle de la génération de la phrase cible en arrangeant les séquences obtenues dans la deuxième phase de façon à avoir une phrase cohérente. Le premier système à base d'exemple a été présenté dans [Nagao, 1984] pour la traduction entre l'Anglais et le Japonais. D'autres systèmes adoptent la même approche nous en citons [Brown, 1996] et [Way and Gough, 2003].

2. Approche basée sur des modèles statistiques L'approche statistique de la traduction automatique se base sur des modèles probabilistes appris sur des corpus monolingue et bilingue. Au début des années 90 et suite au succès des modèles statistiques dans le domaine de la reconnaissance de la parole, une équipe d'IBM présentait une approche statistique pour la traduction automatique. L'approche s'appuie sur deux modèles probabilistes : un modèle de traduction appris à partir d'un corpus bilingue de textes en langues source et cible, un modèle de langage appris sur un corpus monolingue de textes en langue cible. Le modèle de traduction permet de trouver la traduction la plus probable pour une phrase source. Le modèle de langage en revanche garantit la qualité de la phrase cible. La dernière section de ce chapitre est consacré à la traduction automatique statistique.

Approche hybride Nous avons présenté plus haut les deux principales approches adoptées par la quasi-totalité des systèmes de traduction existants. Les deux approches s'opposent radicalement, l'une s'appuyant sur des règles fondées sur la connaissance linguistique parfaite des deux langues source et cible (à tous les niveaux du langage). L'autre ne requiert à priori aucune connaissance linguistique mais s'appuie essentiellement sur des quantités de textes parallèles pour l'estimation des paramètres de ces modèles. L'approche hybride se place à la médiane de ces deux familles de systèmes de traductions. Elle tente conjuguer les avantages des deux écoles dans l'objectif de produire des traductions de meilleure qualité.

Dans la littérature plusieurs systèmes de ce type existent. Dans le cadre du projet Euromatrix, dans [Eisele et al., 2008] les auteurs adoptent une approche à base de règles et une approche statistique. De même, les auteurs de [Shirai et al., 1997] proposent un système à base base d'exemples qui relaie sur des modules à base

de règles linguistiques pour la traduction du Japonais vers l'anglais. Les auteurs dans [España Bonet et al., 2011] associent aussi une approche statistique à une approche par règles pour la paire de langue espagnole-basque. Les auteurs stipulent que les résultats obtenus dépassent de loin ceux du système statistique. Dans [Groves and Way, 2005] un système fondé sur la traduction à base d'exemples et la traduction statistique est proposé pour la paire de langues Français-Anglais. Ce système hybride affiche une performance nettement meilleure par rapport aux systèmes statistique et à base d'exemples considéré chacun à part.

Pour les systèmes commerciaux, Systran présente un système de traduction hybride dans lequel son moteur à base de règles est associé à un système statistique utilisé pour la post-édition. Dans la campagne d'évaluation WMT'09 (Workshop on Statistical Machine Translation 2009), le système [Schwenk et al., 2009] a remporté la première place pour la paire de langue français-Anglais. Pour la traduction entre le Chinois et l'Anglais, Systran augmenté d'un module de post-édition statistique[Yang et al.,] a remporté aussi la première place dans la campagne d'évaluation CWMT'2011 (China Workshop on Machine Translation 2011).

Les travaux dans ce domaine peuvent être consultés dans [Costa-Jussa and Fonollosa, 2015] où un état de l'art détaillé pour l'hybridation des systèmes de traduction automatique a été présenté.

3.2.3 Traduction automatique statistique

Principe et fondement mathématique

Un texte en langue source est traduit vers une langue cible en fonction de la loi de probabilité conditionnelle qui calcule la probabilité que la phrase en langue cible f soit une traduction de la phrase e en langue source. Cela signifie que le système de traduction traduit e en f avec une probabilité p. Le problème

de la traduction automatique consiste à trouver la phrase f^* qui maximise $p(f|e)$, formellement :

$$f^* = \underset{f}{\operatorname{argmax}}\, p(f|e) \tag{3.1}$$

1. Relation de Bayes La relation de Bayes intervient pour faire introduire un modèle de langage de la langue cible et un modèle de traduction dans la fonction 3.1. D'après la relation de Bayes :

$$p(f|e) = \frac{p(e|f).p(f)}{p(e)} \tag{3.2}$$

On aboutit alors à l'équation 3.3

$$f^* = \underset{f}{\operatorname{argmax}} \frac{p(e|f).p(f)}{p(e)} \tag{3.3}$$

Étant donné que le terme $p(e)$ est indépendant de f, il n'a donc aucune incidence sur la fonction de maximisation, d'où on peut aboutir à l'équation fondamentale relative à la traduction automatique [1] qui suit :

$$f^* = \underset{f}{\operatorname{argmax}}\, p(e|f).p(f) \tag{3.4}$$

où le premier facteur $p(f)$ est le modèle de langage de la langue cible dont les paramètres sont extraits d'un corpus monolingue de la langue cible, et le second facteur $p(e|f)$ est le modèle de traduction dont les paramètres sont extraits à partir d'un corpus parallèle.

1. On parle dans ce cas du modèle de canal bruité « Noisy chanel model » terme émanant du domaine de la reconnaissance de la parole.

2. Le modèle log-linéaire Pratiquement, il est bénéfique d'affecter des poids aux sources d'information (modèle de langage et modèle de traduction) impliquées dans l'expression 3.4. Ceci donnera lieu à l'expression suivante :

$$f^* = \underset{f}{\operatorname{argmax}}\, p(e|f)^{\alpha}.p(f)^{\beta} \tag{3.5}$$

En plus de la pondération des modèles, il est possible d'introduire n'importe quelle autre caractéristique qui pourrait améliorer le résultat de la traduction de la phrase source e en f. L'expression à maximiser sera alors de la forme :

$$\underset{f}{\operatorname{argmax}}\, p(f|e) = \underset{f}{\operatorname{argmax}} \prod_i h_i(f,e)^{\lambda_i} \tag{3.6}$$

Par la suite, le modèle log-linéaire a été introduit pour aboutir à l'expression qui suit :

$$\underset{f}{\operatorname{argmax}}\, p(f|e) = \underset{f}{\operatorname{argmax}} \exp\left(\sum_i \lambda_i \log h_i(f,e)\right) \tag{3.7}$$

Où h_i est une fonction caractéristique et λ_i son poids affecté dans la combinaison. En somme, le problème de la traduction automatique peut être décomposé comme suit

— Calcul des paramètres du modèle de langage de la langue cible $p(f)$

— Calcul des paramètres du modèle de traduction $p(e|f)$

— Opération de maximisation de l'équation 3.4 en un temps acceptable (cette opération est appelée le décodage).

Un système de traduction automatique statistique peut être schématisé comme illustré dans la Figure 3.3.

Le modèle de langage

Ce modèle est responsable de la prise en considération des contraintes imposées par la langue cible, il est estimé sur un corpus monolingue. Son rôle est

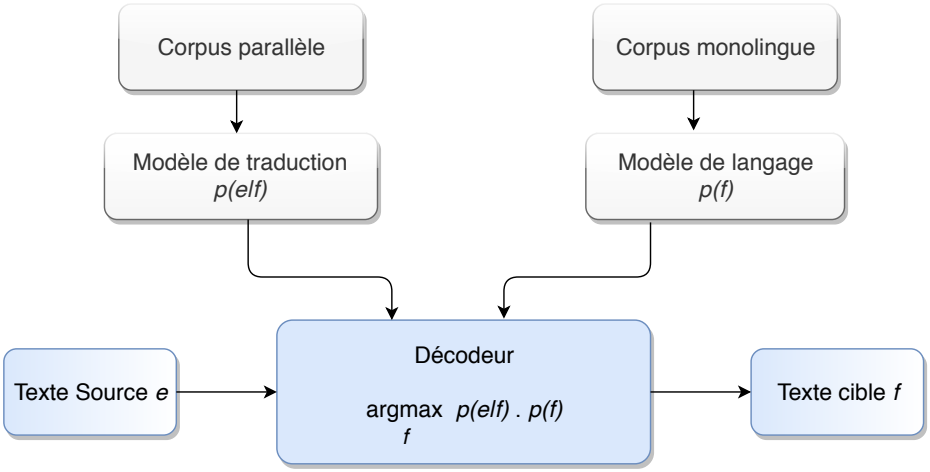

FIGURE 3.2 : Modélisation d'un système de traduction automatique statistique

d'estimer la probabilité d'une séquence de mots (phrase) : plus cette probabilité est élevée plus cette phrase ou séquence de mots est conforme au modèle de langage de la langue cible. Il spécifie une distribution $p(f)$ sur les phrases f_i de la langue cible telle que :

$$\sum_i p(f_i) = 1 \tag{3.8}$$

Si l'on considère une séquence de mots f (une phrase de I mots) tel que : $f = w_1 ... w_I$, alors :

$$p(w_1 w_2 ... w_I) = \prod_{i==1}^{I} p(w_i | w_1 ... w_{i-1}) \tag{3.9}$$

où $h = w_1 ... w_{i-1}$ est appelé l'historique du mot w_i Il s'agit de calculer pour chaque mot w_i la probabilité d'apparition étant donné son historique $h = w_1 ... w_{i-1}$.

$$p(w_i|h) = \frac{Count(hw_i)}{Count(h)} \tag{3.10}$$

La fonction $Count(x)$ renvoie le nombre d'occurrences de x dans le corpus d'apprentissage. Ce modèle de langage est appelé un modèle n-gramme. A partir d'un corpus d'apprentissage, on estime une distribution de probabilité pour le prochain mot avec un historique de taille n. Il s'agit au fait d'un modèle de Markov

d'ordre n, où l'on utilise les n dernières observations pour la prédiction du mot suivant. La probabilité d'un mot sachant son historique (de taille n) est donnée alors :

$$p(w_i|w_1w_2...{}_{i-1}) \approx p(w_i|w_{i-n+1}...w_{i-1}) \tag{3.11}$$

Cette approximation est due au fait qu'aucun corpus d'apprentissage ne peut contenir tous les historiques possibles de tous les mots, d'où l'impossibilité d'estimer toutes les probabilités $p(w_i|h)$. Un problème de probabilité nulle peut apparaître dans ces modèles ; en effet un mot n-gramme peut avoir une probabilité nulle s'il n'existe pas dans le corpus d'apprentissage ; même si ce mot est une réalisation possible dans la langue cible. Pour palier à ce problème les techniques de lissage ont été introduites pour éviter l'attribution d'estimation nulle causée par manque de données dans le corpus d'apprentissage.

Il existe plusieurs techniques de lissage qui tentent d'affecter des probabilités non-nulles à des n-grammes non observés dans les corpus d'apprentissage. Le lissage par repli[Katz, 1987] par exemple (appelé backoff en Anglais) qui consiste à se replier sur le n-gramme d'ordre $k-1$ lorsque le n-gramme d'ordre k n'est pas observé dans le corpus d'apprentissage. Dans ce cas, La formule 3.11 sera alors de la forme :

$$p(w_i|w_1...w_{i-k+1}...w_{i-1}) = \begin{cases} p(w_i|w_{i-k+1}...w_{i-1}) & \text{si } w_ih \quad \text{existe} \\ \lambda(w_{i-k+1}...w_{i-1}).p(w_i|w_{i-k+2}...w_{i-1}) & \text{sinon.} \end{cases} \tag{3.12}$$

Où $h = w_{i-k+1}...w_{i-1}$ est l'historique du mot w_i et $\lambda(w_{i-k+1}...w_{i-1})$ est un poids de repli.

Une autre technique de lissage assez répandue est le lissage par interpolation où tous les niveaux du modèle n-gramme sont combinés pour le calcul de la probabilité d'une séquence de mots. Lorsqu'un n-gramme n'est pas rencontré dans le corpus

d'apprentissage, son score n'est pas nul car il correspond à une somme pondérée des niveaux inférieurs du modèle n-gramme. La formule de calcul donnée en 3.11 devient alors :

$$p_{intpl}(w_n|w_1w_2...{}_{n-1}) = \lambda_1 p_1(w_n) + \lambda_2 p_2(w_n|w_{n-1})... + \lambda_n p_n(w_n|w_1...w_{n-1}) \quad (3.13)$$

telle que $0 \leq \lambda_i \leq 1$ et $\sum_i \lambda_i = 1$.
D'autres techniques existent, telles que le lissage de Witten-Bell[Witten and Bell, 1991] basé sur l'interpolation et le lissage Kneser-Ney modifié[Chen and Goodman, 1999] initialement proposé dans [Kneser and Ney, 1995] qui est une variante du lissage par repli.

1. Évaluation des modèles de langage La perplexité est la mesure la plus plus utilisée pour l'évaluation des modèles de langage afin de déterminer leur qualité. Le calcul de la perplexité d'un modèle de langage se base sur l'entropie croisée définie comme suit :

$$H(P_{LM}) = -\frac{1}{n}\sum_{i=1}^{n} \log P_{LM}(w_i|w_1...w_{i-1})) \quad (3.14)$$

Par une simple transformation, la perplexité sera de la forme :

$$PPL = 2^{H(P_{LM})} \quad (3.15)$$

On notera que plus cette valeur est petite, moins est le modèle perplexe sur le choix d'un n-gramme donné. Enfin, cette métrique n'est utilisable pour la comparaison de deux modèles de langage différents que si ces deux derniers ont le même vocabulaire.

2. Interpolation des modèles de langage Pour la modélisation du langage, une méthode assez répandue pour augmenter la taille des corpus d'apprentissage est la combinaison de plusieurs modèles en un seul modèle de langage. On parle dans ce cas de l'interpolation des modèles de langage. Pratiquement, si on dispose de M Modèles de langage correspondant à m distribution P_m où $m = 1, ..., M$ le modèle interpolé suivra alors la distribution suivante :

$$p_{interpol}(w_i|w_{i-n+1}...w_{i-1}) = \sum_{m=1}^{M} \lambda_m P_m(w_i|w_{i-n+1}...w_{i-1})) \tag{3.16}$$

avec $\sum_{m=1}^{M} \lambda_m = 1$
Les λ_m représentent les poids de chaque modèle de langage estimés le plus souvent avec le maximum de vraisemblance qui minimise la perplexité du modèle sur un corpus de données représentatives pour lesquelles le modèle est estimé.

Le modèle de traduction

Appris sur des corpus parallèles alignés, les modèles de traduction se chargent d'estimer la probabilité $p(e|f)$ (formule 3.4) en se basant sur une table de traduction et un modèle d'alignement.

1. L'alignement La notion d'alignement intervient dans la majorité des systèmes de traduction statistique. Il s'agit en effet de mettre en correspondance les mots d'une phrase de la langue source et les mots de la phrase de la langue cible qui est sa traduction possible. La figure qui suit illustre un exemple simple d'alignement entre une phrase en anglais et sa traduction en Français :

Il faut noter que l'alignement entre groupes de mots est autorisé et qu'il existe formellement un mot Null dans les deux langues. Ce mot Null est utilisé lorsqu'un ou plusieurs mots d'une phrase n'ont pas de correspondance dans l'autre

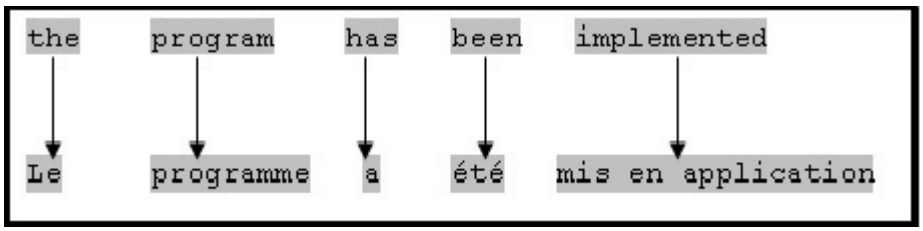

FIGURE 3.3 : Exemple d'alignement entre une phrase et sa traduction

phrase. En introduisant cette notion d'alignement (la variable A), la probabilité de traduction $p(e|f)$ sera de la forme :

$$p(e|f) = \sum_{A} p(e, A|f) \approx \max_{A} p(e, A|f) \tag{3.17}$$

En fonction de l'unité de traduction qui apparaît dans les lois de probabilités, il existe deux approches des modèles de traduction : Modèles à base de mots et modèles à base de séquence de mots.

2. Modèles à base de mots

Les modèles IBM Les auteurs de [Brown et al., 1993] ont défini cinq modèles statistiques de traduction dont la complexité et la performance croient d'un modèle à un autre. Ces modèles visent tous à calculer la probabilité $p(f|e)$ en utilisant différents paramètres estimés sur un corpus parallèle. Notons dans ce qui suit : $e = e_1...e_i$ la phrase source et $f = f_1...f_j$ la phrase cible.

Le modèle IBM1 Il considère que tous les mots source peuvent être alignés à tous les mots cible avec la même probabilité (tous les alignements sont équiprobables). Le modèle IBM1 repose sur une seule loi de probabilité, une loi lexicale notée $T(f|e)$; une table de traduction de mots où seront stockées les probabilités de traduction entre tous les mots sources et cibles.

Le modèle IBM2 Contrairement au 1^{er} modèle, le modèle IBM2 exclut l'équiprobabilité des alignements entre mots. A chaque alignement possible entre les mots de deux phrases est affectée une probabilité. Cette probabilité dépend des mots alignés et de la taille des phrases. Cet alignement est de la forme $A = a_1...a_J$, où, pour tout j de l'intervalle $[1, J]$, a_j est inclus dans l'intervalle $[0, I]$ tel que :

$$\begin{cases} a_j = i & \text{Signifie que le mot cible } f_j \text{ est aligné à } e_i \\ a_j = 0 & \text{Signifie que le mot cible } f_j \text{ est aligné au mot null} \end{cases} \tag{3.18}$$

Ainsi, un alignement de cette forme autorise l'alignement de plusieurs f_j à un seul e_i, par contre un mot cible f_j est aligné au plus à un mot source e_i. C'est pour cette propriété que les modèles IBM2 à 5 sont asymétriques. Enfin il est acceptable qu'un mot e_i (de la phrase source) ne soit aligné à aucun mot f_j (de la phrase cible). Le modèle IBM2 repose donc sur une loi de traduction lexicale $T(f|e)$ et une loi d'alignement ou de distorsion $p(a_j|J)$.

Le modèle IBM3 Ce modèle de traduction intègre en plus des lois de traduction lexicale et de distorsion une loi de fertilité, de la forme $N(\phi|e)$. Pour chaque position source i de l'intervalle $[1, I]$, ϕ_i est le nombre de mots cible alignés à f_i, c'est à dire $\phi_i = Card\{j|a_j = i\}$. Le modèle IBM3 considère que chaque mot de e produit un ou plusieurs mots dans la phrase f (c'est la notion de fertilité), les mots générés sont par la suite identifiés et puis liés à une position dans la phrase e (distorsion). Le mot $Null$ est considéré comme cas spécial : aucune probabilité de fertilité ne lui est affectée. Le modèle définit ainsi une probabilité $p_1 = 1–p_0$ de génération spontanée d'un mot cible aligné à $Null$ après toute génération de mot cible aligné à un mot source. Ces mots « libres » appelés aussi « spontanés » ne portent pas de sens ; en pratique, ils sont insérés de façon à respecter la grammaire de la langue cible.

Modèles IBM4 et IBM5 Pour les modèles 2 et 3, la probabilité de connexion $P(f|e)$ dépend des positions des mots et de la longueur des deux phrases source et cible. En revanche, pour le modèle 4, la probabilité de distorsion dépend, en plus de la position du mot considéré dans e et du mot dans f, des mots eux-mêmes mais aussi de la position des autres mots dans f également liés au mot considéré dans e. Malgré cette précision, les modèles 3 et 4 ont des limites, le modèle 5 reste toujours le plus utilisé, il est certes identique au modèle 4 mais il corrige certaines insuffisances de ce dernier. Les modèles 1-4 peuvent être des moyens d'initialisation pour le modèle 5.

2. Modèles à base de séquence de mots l'unité de traduction des modèles à base de séquence de mots (comme leur nom l'indique) est le mot. Généralement, les modèles à base de séquences découpent le processus de traductions en trois étapes :

— La phrase source e est découpée en I séquences notées $\tilde{e}_1 ... \tilde{e}_I$

— Chaque séquences $\tilde{e}_i$ est ensuite traduite dans une langue cible par le biais d'une table de traduction de séquences de la forme $T(f|e)$ qui attribue une probabilité de traduction, à tous les couples $\tilde{e}_l, \tilde{f}_j$, avec $\tilde{e}_l$ et $\tilde{f}_j$ et respectivement une séquence de mots sources et une séquence de mots cibles.

— Enfin, les séquences cibles sont éventuellement réordonnées à l'aide d'un modèle de distorsion pour produire la phrase cible finale $f = \tilde{f}_1, \tilde{f}_2, ... \tilde{f}_J$.

Le processus de traduction par séquence de mots repose donc en grande partie sur une table de traduction qui va établir les correspondances lexicales entre les séquences de mots de la langue source et celles de la langue cible associées à des probabilités de traduction. L'utilisation des groupes de mots comme unité de traduction permet d'aligner n mots source à m mots cible et d'éviter ainsi les alignements parfois inadéquats que les modèles à base de mots produisent. Il

est important de signaler que ces modèles produisent des traductions directes qui évitent donc le ré-ordonnancement des mots (du fait que cet ordre est déjà produit dans la séquence cible), c'est là une propriété importante de ces modèles.

Apprentissage des paramètres des modèles

L'apprentissage des paramètres des modèles à base de mots se fait de façon itérative en appliquant l'algorithme « Expectation-Maximization »[Dempster et al., 1977]. Un programme (disponible publiquement) appelé Giza++[Och, 2000] réalise cet apprentissage. Giza++ est une extension du programme Giza, qui fait partie de la boîte à outils pour la traduction automatique statistique Egypt[Al-Onaizan et al., 1999].

L'apprentissage (appelé aussi l'entraînement dans la littérature) est réalisé à partir d'un corpus parallèle, aligné par phrases. Aucune information a priori n'est nécessaire. Giza++ entraîne successivement des modèles de complexité croissante. Une fois déterminés, les paramètres d'un modèle servent de paramètres de départ du modèle qui suit.

Quant aux modèles à base de séquence de mots, l'apprentissage fait toujours recours aux modèles IBM, avec utilisation du modèle d'alignement qui permettra de calculer automatiquement les alignements de mots entre les paires de traduction (à travers l'outil Giza++) pour en extraire la tables de traduction des séquences.

Le décodage Le processus de traduction qui consiste à transformer une phrase source en phrase cible est appelé décodage dans le domaine de la traduction automatique probabiliste. Le décodeur est un composant clé dans un système de traduction automatique statistique ; il produit pour une phrase source les n meilleures traductions possibles en se basant sur les paramètres déjà appris (à partir d'un corpus bilingue aligné) des modèles sur lesquels se base le

système de traduction. Le décodeur implémente l'algorithme de la fonction Argmax de la formule 3.4. Plusieurs décodeurs disponible en open-source existent dont les plus connus sont Jane[Markus et al., 2014], Cdec[Chris et al., 2010] et Moses[Koehn et al., 2007].

Cas du décodeur Moses Moses [2] est système de traduction automatique statistique dont le décodeur est implémenté à base d'un algorithme de recherche en faisceau (beam search decoder). Il emploie un modèle de langage tri-gramme à repli et une table de traduction pour générer une liste des n meilleures traductions. Celles-ci sont ensuite réévaluées à l'aide d'un modèle de langage neuronal quadri-gramme afin de sélectionner la traduction cible. Dans sa version standard, Moses utilise huit fonctions caractéristiques modélisant le processus de traduction. Pour aboutir à la phrase cible, Ces fonctions implémentent les contraintes : les probabilités de traduction des séquences de mots dans les deux sens, les probabilités de traduction des mots dans les deux sens, une mesure de distorsion, deux pénalités d'insertion de mots et de séquences de mots, et la probabilité calculée par le modèle de langage de la langue cible. Moses est distribué sous licence libre et est activement développé et dispose de nombreuses caractéristiques intéressantes, comme la possibilité d'exploiter des modèles de traduction factorisés ou des modèles de distorsion lexicalisés etc.

3.2.4 Évaluation automatique des systèmes de traduction

L'évaluation de la traduction automatique est une phase importante dans la mise en œuvre d'un système de traduction quelque soit l'approche qu'il adopte. l'évaluation de la qualité de traduction peut se faire selon plusieurs critères tels que des critères de correction grammaticale et de fidélité au sens du texte. L'évaluation la plus intuitive et la plus crédible qui tiendrait compte de ces critères

2. http ://www.statmt.org/moses

demeure l'évaluation humaine. Néanmoins, cette évaluation est subjective d'une part et coûteuse en terme d'intervention humaine. C'est pourquoi des évaluations automatiques ont été introduites. Les évaluations automatiques (évaluation objective) nécessitent une ou plusieurs traductions qui seront considérées comme des références pour la traduction d'une phrase source. L'évaluation consiste soit à mesurer la distance entre la phrase candidate et les traductions de référence, comme le cas des scores (TER, PER, WER) ; ou bien à mesurer la ressemblance comme c'est le cas du score (BLEU). La qualité de la traduction de référence est donc très importante.

1. BLEU Le score BLEU (Bilingual Evaluation Understudy) a été proposé par[Papineni and al., 2001]. Il varie de 0 à 1 et il est d'autant meilleur qu'il est grand. BLEU a gagné le statut de mesure automatique de référence au sein de la communauté de traduction automatique. L'idée principale est la comparaison de la sortie du traducteur avec une/des traductions de référence. Les statistiques de co-occurrence basées sur les ensembles de n-grammes pour les segments de traduction et de référence, sont calculées pour chacun de ces segments et sommées sur tous les segments, le but étant de trouver combien de n-grammes sont retrouvés dans la traduction de référence. Cette moyenne est multipliée par une pénalité de brièveté, afin de pénaliser les systèmes qui essaieraient d'augmenter artificiellement leurs scores en produisant des phrases délibérément courtes.

2. NIST NIST[Doddington, 2002] (National Institute of Standards and Technology) est une variante du BLEU, alors que ce dernier affecte des poids égaux aux n-grammes corrects, NIST favorise les n-grammes corrects rares. Il diffère aussi du BLEU dans le calcul de la pénalité de brièveté dans ce sens que les petites variations de longueur des phrases traduites n'influent pas sur le score global pour autant.

3. WER WER[Nießen et al., 2000] (Word Error Rate) est une métrique d'évaluation à l'origine pour mesurer la performance des systèmes de reconnaissance de la parole. WER calcule la distance minimale d'édition entre hypothèse et référence. WER est basé sur la distance de Levenshtein[Gilleland, 2009], utilisée pour comparer deux chaînes de caractères en calculant le nombre minimum d'opérations requises pour transformer une chaîne à l'autre. Ces opérations sont la suppression, l'insertion et la substitution d'un seul caractère. WER opère au niveau du mot et calcule la somme des insertions, substitutions et suppression normalisée par la longueur de la phrase de référence.

4. PER Une métrique similaire au WER est le PER[Leusch et al., 2003] (Position-independent word error rate), qui ne tient pas compte de l'ordre des mots dans la comparaison des traduction et référence. PER considère les traduction et référence comme des sacs à mots pour lesquels il calcule la différence normalisée avec la longueur de la référence.

5. METEOR METEOR[Banerjee and Lavie, 2005] (Metric for Evaluation of Translation with Explicit Ordering) est une métrique d'évaluation basée sur la précision et rappel. Elle repose sur l'alignement mot à mot entre hypothèse et référence. L'alignement se fait soit par la correspondance exacte des mots ou en introduisant des informations linguistiques telles que la dé-suffixation (stemming) pour aligner des mots ayant le même stem ou la synonymie en utilisant Wordnet. METEOR inclut aussi une pénalité pour l'ordre des segments.

6. TER Le TER[Snover et al., 2006] (Translation Edit Rate), cette métrique mesure le nombre d'opérations d'édition nécessaires pour transformer la traduction hypothèse en l'une des traductions de référence, en normalisant avec la longueur

des références. Les opérations d'éditions sont l'insertion, substitution et suppression d'un seul mot ou le déplacement d'une séquence de mots.

3.3 État de l'art de la traduction automatique des dialectes arabes

En dépit de leur utilisation massive dans le monde arabe, les dialectes arabes comptent un nombre réduit de travaux dédiés à la traduction de et vers ces dialectes et les résultats de ces travaux sont toujours à un stade précoce. Ces dialectes ne sont pas pris en charge par les systèmes de traduction pour la langue arabe, pourtant ces dialectes sont des variantes de cette langue. De plus ils sont largement plus utilisés dans le quotidien des populations arabes, comparés à l'arabe standard qui n'est utilisé que dans les discours et correspondances officiels et dans les discours religieux. De part la négligence de ces dialectes et leur exclusion des outils TAL dédiés à l'arabe standard, leur prise en charge dans ce domaine (le TAL) est problématique, surtout pour la traduction automatique. En effet, les systèmes de traduction à base de règles sont difficiles à mettre en œuvre pour ces dialectes en absence de règles bien définies mises en place par des linguistes. De plus la majorité d'entre eux (ces dialectes) ne disposent pas d'outils TAL même basiques. Les systèmes fondés sur des corpus sont aussi difficiles à envisager du fait du manque de ressources textuelles (corpus parallèles et monolingue) pour la majorité de ces dialectes.

Il convient de noter, que la problématique relative au manque de ressources TAL n'est pas propre aux dialecte arabes, différentes paires de langue standard-dialecte sont aussi concernées. Plusieurs travaux dans ce contexte existent. L'idée principale de ces contributions est l'exploitation de la proximité entre les langues standards (dotées en ressources) et les dialectes ou les langues non dotées de ressources les plus proches. Dans le cadre précis de la traduction automatique, le lecteur peut se référer aux travaux suivants : [Zhang, 1998] pour le cantonais et le mandarin,

[Hajič et al., 2000] pour le tchèque et le slovaque, [Altintas and Cicekli, 2002] pour le turque et le tatar de Crimée, ainsi que [Scannell, 2006] pour l'irlandais et le gaélique écossais, [Nakov and Ng, 2012] pour la paire indonésien-anglais en pivotant par le malais et [Haddow et al., 2013] pour l'allemand d'Autriche et le dialecte viennois. On notera aussi l'avancée de ces travaux sur les dialectes arabes et la prise de conscience précoce de l'importance de la problématique liée aux dialectes dans les autres pays (en dehors du monde arabe).

3.3.1 La traduction entre l'arabe standard et les dialectes

Les auteurs de [Bakr et al., 2008] présentent une approche générique pour convertir une phrase en arabe dialectal égyptien en une phrase vocalisée en arabe standard. Les auteurs ont combiné une approche statistique pour la segmentation et l'étiquetage des phrases en arabe standard et dialectal ainsi qu'une approche par règles qui repose sur un dictionnaire bilingue pour la génération des phrases diacritisées de l'arabe standard. Notons que la segmentation et l'étiquetage du texte dialectal a été faite manuellement. Ce travail a été évalué sur un corpus de 1K de phrases dialectales réparties en 800 phrases pour l'apprentissage et 200 pour l'évaluation. Pour la conversion des mots dialectaux en mots arabes standards un taux de rappel de 88% a été reporté, cependant le taux de rappel relatif au bon ordonnancement de ces mots est de 70%.

Elissa [Salloum and Habash, 2012] est un système de traduction à base règles qui traduit depuis l'arabe dialectal vers l'arabe standard, les dialectes concernés sont l'égyptien, le levantin, l'irakien et à un degré moindre l'arabe du Gulf. Après une phase d'identification (réalisée par l'utilisateur), les mots dialectaux sont convertis par Elissa en paraphrase en arabe standard à l'aide de l'analyseur morphologique dialectal ADAM [Salloum and Habash, 2011], des règles de transfert morphologique et de dictionnaires dialecte-arabe standard. Les

paraphrases générées forment un treillis à partir duquel les meilleures hypothèses seront sélectionnées à travers un modèle de langage 5-grammes appris sur un corpus de 200M mots arabes.

Dans [Mohamed et al., 2012], les auteurs ont présenté un système de traduction à base de règles qui traduit de l'arabe standard vers l'arabe dialectal égyptien. Le système a été évalué au regard de l'étiquetage morpho-syntaxique et en termes de couverture des mots hors vocabulaire. En termes d'étiquetage le taux de rappel s'est vu améliorer de 73.24% à 86.84%, alors que le taux des mots hors vocabulaire a chuté de 28.98% à 16.66%.

Pour le dialecte sana´ani du Yemen, dans [Al-Gaphari and Al-Yadoumi, 2012] les auteurs ont aussi adopté une approche à base de règles qui traduit vers l'arabe standard. Le système a atteint un taux de rappel de 77.32% sur un corpus de test de 9386 mots.

Pour la traduction du marocain vers l'arabe standard, une approche à base de règles combinée avec un modèle de langage a été retenue dans [Tachicart and Bouzoubaa, 2014]. Le système est basé sur une analyse morphologique avec l'analyseur morphologique Alkhalil[Boudlal et al., 2010] adapté et enrichi avec les affixes du dialecte marocain, un dictionnaire bilingue(construit à partir de scénarios de télévision et de pages WEB). Après une étape d'identification qui sépare le contenu dialectal de celui de l'arabe standard, le texte est analysé et segmenté en unités dialectales annotées. Ces unités sont alors alignées à une ou plusieurs unités arabes standard correspondantes en utilisant le dictionnaire bilingue. Dans l'étape de génération, les phrases en arabe standard sont générées et passées au modèle de langage pour en produire les phrases les plus correctes.

Dans [Hamdi et al., 2013], les auteurs proposent un système de traduction entre les formes verbales du dialecte tunisien et l'arabe standard. Ce travail est basé sur une analyse morphologique profonde à base de racine et de schème. Cette

Référence	Source	Destination
[Bakr et al., 2008]	égyptien	arabe standard
[Salloum and Habash, 2012]	levantin, égyptien, irakien, dialecte du golfe	arabe standard
[Mohamed et al., 2012]	arabe standard	égyptien
[Al-Gaphari and Al-Yadoumi, 2012]	sana´ani (Yéménite)	arabe standard
[Tachicart and Bouzoubaa, 2014]	marocain	arabe standard
[Sadat et al., 2014]	tunisien	arabe standard
[Hamdi et al., 2013]	tunisien	arabe standard
	MSA	tunisien

Tableau 3.1 : Travaux de la traduction automatique entre les dialectes arabes et l'arabe standard :Source/destination

approche est similaire à celle utilisée dans [Mohamed et al., 2012], [Sawaf, 2010] et [Salloum and Habash, 2013] mais elle se distingue par l'analyse morphologique profonde basée sur MAGEAD [Habash and Rambow, 2006a] (un analyseur morphologique et générateur pour l'arabe dialectal). Le système traduit dans les deux sens (de l'arabe standard vers le dialecte tunisien et vice et versa). Son taux de rappel est de 84% pour la traduction du dialecte vers l'arabe standard et 80% dans le sens opposé.

3.3.2 La traduction entre les dialectes arabes et l'anglais

Dans [Sawaf, 2010] un système de traduction hybride est proposé, combinant une approche statistique et à base de règles. Ce système traduit des textes en arabe dialectal (contenu WEB et transcriptions d'émissions diffusées) vers l'anglais en utilisant l'arabe standard comme langue pivot. Les textes dialectaux sont normalisés (traduits vers l'arabe standard) en utilisant des règles de transfert au niveau caractères. Le texte est ensuite analysé à l'aide d'analyseurs morphologiques pour l'arabe dialectal et standard. Un processus de décodage basée sur la recherche en faisceau est utilisé ainsi qu'un modèle de langage pour générer le texte normalisé en arabe standard. Le travail concerne une grande variété de dialectes à savoir : les dialectes du Levant (libanais, syrien, palestinien et jordanien), les dialectes du golf arabe (Irak,Arabie-Saoudite, et le sud de la

Péninsule arabe), la région du Nil (Égypte et Soudan) et les dialectes du Maghreb arabe (libyen, marocain et Tunisien). Les résultats de ce travail montrent que l'approche hybride est plus performante que les approches statistique et à base de règles utilisée chacune à part, et que la normalisation des textes dialectaux (données d'apprentissage et de test) améliore en absolu le score BLEU de 2% pour le texte extrait du WEB et 1% pour le texte d'émission diffusées et informations.

Dans [Salloum and Habash, 2013], Elissa (cité plus haut) a été manuellement évalué, 93% des phrases arabes produites par Elissa étaient correctes. Par ailleurs, Elissa a été utilisé comme pivot dans un système de traduction automatique de l'arabe dialectal vers l'Anglais. Ce pivotage a amélioré le score BLEU du système de traduction entre 0.6% et 1.4%.

Dans [Sajjad et al., 2013], les auteurs présentent un système de traduction automatique statistique de l'arabe égyptien vers l'anglais. Des règles de transfert morphologiques et phonologiques sont utilisées pour adapter le dialecte égyptien à l'arabe standard. Ces transformations ont réduit le taux des mots hors-vocabulaire de 5.2% à 2.6% et amélioré le score BLEU du système de traduction automatique statistique de 1.87 points.

Dans [Jeblee et al., 2014], les auteurs présentent un système qui traduit (contrairement aux autres travaux) de l'anglais vers le dialecte égyptien en utilisant l'arabe standard comme langue pivot. Le traducteur est basé sur un système de traduction statistique anglais-arabe appris sur un corpus de 5M de paires de phrase qu'on convient d'appeler système noyau. Les sorties de ce système sont alors traduites en dialecte égyptien en utilisant les techniques d'adaptation au domaine. Pour le besoin de l'adaptation, un corpus tri-parallèle (Anglais, arabe standard et dialecte égyptien) de 100K phrases a été crée en utilisant des règles de transfert. Par facilité de lecture, on se référera à chaque partie de ce corpus par Ang-100K, MSA-100K et Egy-100K. Deux variantes de système d'adaptation ont été présentées. La

première variante traduit avec le système noyau la partie Anglaise (Ang-100K) du corpus tri-parallèle vers l'arabe standard (nous appelons le résultat de cette traduction MSA-100K-Trad). MSA-100K-Trad et la partie égyptienne (Egy-100K) du corpus tri-parallèle sont utilisés comme corpus d'apprentissage pour traduire de l'arabe standard vers le dialecte égyptien. La seconde variante inclut deux étapes d'adaptation. La première étape consiste en l'adaptation de la sortie arabe standard du système noyau au domaine de la partie arabe standard du corpus tri-parallèle, il s'agit là de traduire en utilisant un système entraîné sur le corpus (MSA-100K-Trad,MSA-100K). La seconde étape d'adaptation consiste à traduire la sortie arabe sortie du premier système d'adaptation vers le dialecte égyptien en utilisant le corpus parallèle (MSA-100k, Egy-100k). Ce travail a démontré que l'adaptation entre le dialecte égyptien et l'arabe standard peut être vue comme l'adaptation entre différents domaines de la même langue et que cette adaptation permet d'améliorer les performances du système de traduction. Par ailleurs, l'utilisation de l'arabe standard comme langue pivot ensuite l'adapter au dialecte peut aussi influer positivement sur la qualité du système de traduction.

Les auteurs de [Al-Mannai et al., 2014] ont utilisé la segmentation morphologique non-supervisée de l'arabe dialectal pour améliorer les résultats du système de traduction du dialecte qatari vers l'anglais. Dans ce travail, il s'est avéré que la segmentation en utilisant Morfessor[Siivola et al., 2007] (un outil pour la segmentation morphologique non-supervisée) améliore les performance du système de traduction comparé à un système de traduction sans segmentation du tout ou avec une segmentation Arabic Treebank (ATB). De plus un modèle de segmentation multi-dialecte a été appris sur un corpus multi-dialecte incluant Le dialecte qatari, l'égyptien, le Levantin et l'arabe standard. Un système de traduction qatari-anglais a été entraîné sur ce corpus parallèle segmenté (multi-dialecte/arabe standard-Anglais). Le score bleu de ce système a été amélioré de

1.5 points comparé au même système sans segmentation. Dans l'autre sens de la traduction (de l'anglais vers le dialecte qatari), les auteurs ont présenté un système de traduction préliminaire en utilisant le même corpus parallèle sans segmentation avec un modèle de langage incluant d'autres corpus dialectaux. Ce système affiche un gain en score BLEU de 0.22 comparé au système avec un modèle de langage entraîné seulement sur le corpus qatari.

Dans [Durrani et al., 2014], la qualité du système de traduction de l'égyptien vers l'anglais a été amélioré en réduisant le taux des mots hors vocabulaire. L'approche adoptée consiste en la conversion de l'égyptien vers l'arabe standard en utilisant un large modèle de langage pour la sélection des meilleures hypothèses arabe standard candidates pour les mots dialectaux hors vocabulaire (via un algorithme de recherche en faisceaux). Les hypothèses de l'arabe standard sont obtenues à travers des corrections orthographiques des mots dialectaux ou par la suggestion de synonymes. Après sélection des meilleures hypothèses, les résultats sont traduits vers l'anglais à l'aide d'un système de traduction automatique statistique. Il s'est avéré que la correction orthographique améliore le score BLEU de 1.7 points par rapport à un système basique qui traduit directement le dialecte égyptien sans édition vers l'anglais.

Le projet Bolt(2011-2014) DARPA[3] (l'agence Américaine responsable des projets en recherche avancée pour la Défense) a lancé le programme BOLT[4] (Broad Operational Language Translation). L'objectif du projet est la création de nouvelles techniques pour la traduction automatique et l'analyse linguistique qui peuvent être appliquées pour des genres informels de textes et de parole dans la communication en ligne ou personnelle. Il est dédié notamment à la traduction entre l'anglais et le chinois d'une part et l'anglais et le dialecte arabe égyptien

3. Defense Advanced Research Projects Agency
4. http://www.darpa.mil/program/broad-operational-language-translation

d'une autre part. BOLT inclut trois domaines techniques : le développement d'algorithmes et de systèmes pour la traduction automatique, la collection de données et la création de corpus ainsi que l'évaluation.

Dans le cadre de ce programme, dans [Zbib et al., 2012] deux corpus parallèles levantin-anglais (1.1M mots) et égyptien-anglais (380K mots) ont été construits en traduisant vers l'anglais des parties extraites d'un large corpus WEB de textes arabes. La classification par dialecte et la traduction ont été faites en utilisant le crowdsourcing. Plusieurs expérimentations ont été réalisées avec des systèmes de traduction statistiques entrainés sur ces corpus en plus d'un corpus parallèle arabe standard-Anglais (150M mots pour la partie arabe). Il a été constaté que la segmentation utilisant l'analyseur morphologique MADA[Habash and Rambow, 2005] améliore la qualité de la traduction. Il a été aussi montré que le système de traduction appris sur des corpus combinant des données dialectales et arabe standard tendent à renvoyer de meilleures résultats, ce résultat est naturel car les textes dialectaux dans leur majorité sont une mixture de l'arabe standard et dialectal. Au regard de l'utilisation de l'arabe standard comme langue pivot pour traduire du dialecte vers l'anglais (cette expérimentation concerne le dialecte du Levant seulement), les auteurs ont constaté que le score BLEU a été amélioré de 2.3 points (pour les premières expérimentations d'augmentation de la taille du corpus dialectal) mais après l'ajout de plus données dialectales au corpus d'apprentissage (400K mots), la traduction directe retourne de meilleurs résultats malgré le taux relativement faible des mots dialectaux hors vocabulaires lorsque le système pivote par l'arabe standard.

Pour traiter le problème des mots dialectaux hors vocabulaire dans le contexte de la traduction automatique du dialecte arabe vers l'anglais, les auteurs de [Aminian et al., 2014] ont adopté une approche qui normalise les

mots dialectaux en mots arabe standard en utilisant AIDA [5][Elfardy et al., 2014] et MADAMIRA [6][Pasha et al., 2014], pour l'identification et le remplacement respectivement. Cette approche a amélioré la qualité du système de traduction automatique en score BLEU absolu de 0.4% et 0.3% pour AIDA et MADAMIRA dans cet ordre.

Dans le cadre de ce même programme, les contributions apportées dans [Aransa, 2015] concernent aussi la traduction des dialectes arabes vers l'anglais, avec un focus sur l'égyptien en particulier. Différentes techniques ont été implémentées comme l'adaptation au domaine puisque les corpus d'apprentissage dans le cadre du projet BOLT incluent l'arabe standard et différents dialectes (égyptien, levantin et irakien). Les performances des systèmes de traduction ont été améliorées en considérant les différents dialectes comme des domaines différents. Les techniques d'adaptation ont été utilisées pour les modèles de langage et de traduction. Dans le cadre de ce travail, différents schèmes de segmentation morphologique ont été évalués pour l'amélioration de la qualité de la traduction automatique. Par ailleurs, pour traiter le problème des mots hors vocabulaires, l'auteur a proposé une technique qui consiste à la translittération les mots rares tels que les noms propres.

Le projet MuDMAT(2014-2017) MuDMAT[Sadat, 2015] (pour Multi-Dialect Machine Translation) est un autre projet de traduction automatique des dialectes arabes financé par NSERC [7]. L'objectif de ce projet est la traduction automatique entre les dialectes arabes du Maghreb (algérien, marocain et tunisien), l'arabe standard et le français. D'après l'auteur, un premier système de traduction

5. Un outil d'identification de dialecte arabe qui opère au niveau du mot et de la phrase.
6. Un système d'analyse morphologique et de désambiguisation pour l'arabe standard et le dialecte égyptien.
7. National Science and Engineering Research Council of Canada

Référence	Source	Pivotage par l'arabe standard	destination
[Sawaf, 2010]	levantin (Liban, Nord de la Syrie, Damas, Palestine, Jordanie) Dialecte arabe du Golfe (Nord de l'Irak, Baghdad, Sud de l'Irak, Arabie-Saoudite, Sud de la Péninsule arabe), Région du nil (Égypte et Soudan) Maghreb arabe(Libye, Maroc, Tunisie)	Oui	anglais
[Zbib et al., 2012]	levantin, égyptien	Non	anglais
[Salloum and Habash, 2013]	levantin, égyptien, irakien, arabe du Golfe	Oui	anglais
[Sajjad et al., 2013]	égyptien	Oui	anglais
[Jeblee et al., 2014]	anglais	Oui	égyptien
[Al-Mannai et al., 2014]	qatari	Non	anglais
[Durrani et al., 2014]	égyptien	Oui	anglais
[Aminian et al., 2014]	égyptien	Oui	anglais
[Aransa, 2015]	égyptien	Non	anglais

Tableau 3.2 : Travaux de la traduction automatique entre les dialectes arabes et l'anglais : Source/destination et pivotage par l'arabe standard

à base de règles a été réalisé pour traduire le dialecte tunisien vers l'arabe standard et le français.

Tous les travaux cité précédemment sont relatifs à la traduction automatique textuelle. Pour la traduction de la parole, seuls des travaux financés par le DARPA existent à notre connaissance. Le projet TRANSTAC (The Spoken Language Communication and Translation System for Tactical), est le prédécesseur de BOLT, un projet de traduction de la parole pour le dialecte arabe Irakien et l'anglais. Son objectif était le développement rapide de système de traduction automatique bi-directionnels qui permettent à des personnes parlant différentes langues de communiquer. Plusieurs prototypes ont été développés pour les domaines militaire et médical permettant les conversations avec les populations locales parlant l'arabe irakien, le mandarin, le persan, le pachto et le thaï. Les travaux de traduction de l'irakien dans le cadre de ce projet ont été évalués dans [Condon et al., 2010, Condon et al., 2008].

Dialecte	Références
égyptien	[Bakr et al., 2008], [Salloum and Habash, 2012],[Mohamed et al., 2012]
levantin	[Salloum and Habash, 2012], [Meftouh et al., 2015],
irakien	[Salloum and Habash, 2012]
arabe du Golfe	[Salloum and Habash, 2012]
sana´ani (Yéménite)	[Al-Gaphari and Al-Yadoumi, 2012]
tunisien	[Hamdi et al., 2013], [Sadat et al., 2014]
marocain	[Tachicart and Bouzoubaa, 2014]

Tableau 3.3 : Les dialectes arabes concernés par les travaux de la traduction automatique (dialectes/arabe standard)

3.3.3 Discussion

Nous résumons dans ce qui suit les constations les plus importantes des différentes contributions cités plus haut :

— Tous les travaux cités sont dédiés à la traduction entre l'arabe standard, les dialectes arabes et l'anglais, la majorité d'entre eux utilise les dialectes comme langue source.

— Le nombre limité des dialectes couverts dans le cadre de ces travaux montre que la traduction des dialectes est à un stade précoce.

— Au regard des dialectes utilisés (voir tableaux 3.3 et 3.4), il est clair que les dialectes du moyen-orient spécialement l'égyptien sont les plus étudiés en termes de traduction automatique. Par ailleurs les dialectes du Maghreb sont moins présents, alors que certains dialectes comme le koweïtien, le bahrini, le omanni et le mauritanien n'ont jamais fait l'objet d'une telle étude.

— En termes de méthodologie, on notera que l'approche hybride combinant des méthodes statistiques et à base de règles est la plus utilisée. La voix dominante est l'utilisation de règles de transfert pour la production de paraphrases arabe standard ou pour traiter les mots dialectaux hors vocabulaire.

— Un autre aspect important de ces dialectes est l'utilisation de l'arabe standard comme langue pivot pour la traduction vers l'anglais. Toutes les contributions s'accordent sur le fait que ce pivotage améliore la qualité des systèmes de

Dialecte	Références
égyptien	[Sawaf, 2010], [Zbib et al., 2012],[Salloum and Habash, 2013] ,[Sajjad et al., 2013], [Jeblee et al., 2014], [Durrani et al., 2014], [Aminian et al., 2014], [Aransa, 2015]
levantin	[Sawaf, 2010], [Zbib et al., 2012],[Salloum and Habash, 2013]
irakien	[Sawaf, 2010], [Salloum and Habash, 2013]
arabe du Golfe	[Sawaf, 2010], [Salloum and Habash, 2013]
soudanais	[Sawaf, 2010]
Libyen	[Sawaf, 2010]
marocain	[Sawaf, 2010]
tunisien	[Sawaf, 2010]
qatari	[Al-Mannai et al., 2014]

Tableau 3.4 : Les dialectes arabes concernés par les travaux de la traduction automatique (dialectes/Anglais)

traduction automatique, sauf un seul travail où les auteurs montrent que l'ajout de données d'apprentissage dialectales améliore mieux les scores de la traduction. Cependant ce même travail souligne que ceci est au détriment du taux des mots hors vocabulaire qui est plus faible pour les systèmes pivotant par l'arabe standard

— Du point de vue corpus, nous notons un manque significatif des ressources dialectales. Tous les travaux sont confrontés à ce problème. En effet, nous constatons que dans la quasi-totalité de ces contributions, une importante portion du travail est allouée à la création de ressources : production de corpus dialectaux artificiellement, pré-traitement des données WEB, ou pour les travaux mieux financés l'utilisation du crowdsourcing.

Références	Approche	Description des corpus	Résultats
[Bakr et al., 2008]	Segmentation et étiquetage statistique+ + Règles de transfert	800 phrases pour l'apprentissage 200 phrases pour l'évaluation	Précision : 88%
[Salloum and Habash, 2012]	Approche à base de règles + Modèle de langage	-	-
[Mohamed et al., 2012]	Approche à base de règles	100 commentaires d'utilisateur	Évaluation avec Etiq. morpho. syn. rappel : 73.24%
[Al-Gaphari and Al-Yadoumi, 2012]	Approche à base de règles	9386 mots	Précision : 77.32%
[Tachicart and Bouzoubaa, 2014]	Approche à base de règles +Lexique bilingue+Modèle de langage	-	-
[Sadat et al., 2014]	Approche à base de règles +Lexique bilingue+Modèle de langage	50 phrases	Score BLEU : 14.32
[Hamdi et al., 2013]	Approche à base de règles (analyse morphologique profonde de racines et de schèmes)	Corpus Parallèle tunisien/arabe standard de 1500 paires de phrase Dev set 750 paires de phrases , test set 750 paires phrases	Précision : tunisien-arabe standard 84% arabe standard-tunisien 80%

Tableau 3.5 : Travaux de la traduction automatique entre l'arabe standard et ses dialectes : Approche, description des corpus et résultats

Références	Approche	Description des corpus	Résultats
[Sawaf, 2010]	Traduction automatique statistique +Approche à base de règles	Apprentissage/test(Dialecte/anglais) : Informations diffusées 14.3M/ 12.4K phrases Contenu web 38.5K/ 547 phrases	Scores BLEU : Informations diffusées 36.4 Contenu web 42.1
[Zbib et al., 2012]	Traduction automatique statistique + Segmentation Morphologique	Apprentissage(Dialecte/anglais) : 180k paires de phrases (1.1M Levantin ,380k égyptien, anglais 2.3M mots) Apprentissage(arabe standard/anglais) : 8M paires de phrases arabe standard-anglais	Score BLEU : Pour l'égyptien 20.66 Pour le levantin 19.29
[Salloum and Habash, 2013]	Traduction automatique statistique +Approche à base de règles	Apprentissage(arabe standard/anglais) :64M de mots (partie arabe) Dev10 test 1568 phrases(données audio dev DARPA GALE program) Test Levantin : 2728 phrases [Zbib et al., 2012], Test égyptien : test 1553 phrases (BOLT program)	Scores BLEU : Dev10 set : 39.13 aevantin : 10.54 égyptien : 19.59
[Sajjad et al., 2013]	Règles de transfert morphologique et phonologique (niveau caractère) + Traduction automatique statistique + Adaptation au domaine	Corpus parallèle (Dialecte/anglais) de 38k phrases[Zbib et al., 2012], Apprentissage : 32k phrases, Test : 4k phrases Apprentissage(arabe standard/anglais)200k phrases du : (LDC2004T17,LDC2004E72,corpus parallèle du programme GALE)	Score BLEU : 16.96
[Jeblee et al., 2014]	Traduction automatique statistique +Adaptation de domaine +Adaptation de dialecte	Corpus arabe standard/anglais :5M paires de phrases(NIST 2012) Test :1313 phrases (NIST MT09) Corpus tri-parallèle artificiel de 100k (égyptien-arabe standard-anglais)	Score BLEU : 42.9
[Al-Mannai et al., 2014]	Traduction automatique statistique +Segmentation+Adaptation (arabe standard et dialectes)	Corpus parallèle arabe-qatari/anglais[Elmahdy et al.,] Apprentissage : 12k phrases Test : 1k phrases	Score BLEU : 15.2
[Durrani et al., 2014]	Décodeur Égyptien-arabe standard + Décodeur arabe standard-anglais	Gale-dev10 et Bolt égyptien (tahyyes dev)	Score BLEU : 23.72
[Aminian et al., 2014]	Traduction automatique statistique + Identification et remplacement + des mots dialectaux	Apprentissage : 29M mots arabe standard, 5M de mot dialectaux test (BOLT-arz-test)1065 sentences(LDC2012E30) 16177 mots test arabe standard(MT09-test) 1445 phrases (LDC2010T23), 40858 mots	Score BLEU : AIDA 25.9 MADAMIRA : 28.8
[Aransa, 2015]	Traduction automatique statistique + Adaptation du modèle de langage et du modèle de traduction +Différents schème de segmentation + translittération des noms propres	Différents corpus forum de discussion SMS/système de Chat conversation téléphoniques	Plusieurs scores : (Ter - Bleu)/2 [Servan and Schwenk, 2011] et BLEU

Tableau 3.6 : Travaux de la traduction automatique entre les dialectes arabes et l'anglais : Approche, description des corpus et résultats

3.4 Conclusion

Nous avons vu dans ce chapitre un bref historique sur la traduction automatique, à partir duquel on constate que la traduction automatique est presque aussi vieille que les premiers ordinateurs. Nous avons abordé les bases théoriques sur lesquelles les systèmes de traduction sont construits. Une partie de ce chapitre a été consacrée à la traduction automatique statistique. Nous y avons vu qu'un système fondé sur ce principe utilise un corpus bilingue aligné qui permet de construire une table de traduction qui donne la probabilité de traduction d'un mot ou un groupe de mots de la langue source en un mot ou groupe de mots de la langue cible. Un corpus monolingue pour calculer un modèle de langage qui permet d'estimer la probabilité d'une séquence de mots ou d'une phrase. Et un décodeur qui se charge de trouver la meilleure traduction d'une phrase étant donnés les paramètres de ces modèles. Tous ces composants ont été décrits en détail. La section consacrée à la traduction automatique se termine par une revue des métriques les plus utilisées dans l'état de l'art pour l'évaluation des systèmes de traduction. Nous avons par la suite présenté un état de l'art sur la traduction automatique des dialectes arabes. Nous avons présenté en détail tous les travaux réalisés jusqu'à aujourd'hui dans ce domaine. Tous les travaux recensés ont été décrits en termes d'approche adoptée, de données utilisées (corpus), et de résultats obtenus. Nous avons constaté à travers ces différentes contributions que la traduction des dialectes arabes (en particulier le dialecte algérien) est encore à un stade embryonnaire, les résultats obtenus sont encore loin de ceux de la traduction des langues standards.

Chapitre 4

PADIC un corpus arabe multi-dialecte

4.1 Introduction

Ce chapitre est dédié à la création de ressources dans le contexte des langues peu dotées. Nous nous intéressons particulièrement aux dialectes arabes. Nous évoquons la méthodologie de création du corpus multi-dialectes PADIC (pour Parallel Arabic Dialect Corpus). Nous abordons la démarche que nous avons adoptée pour passer d'un dialecte à un autre en pivotant par l'arabe standard. Aussi, nous consacrons une partie de ce chapitre à une étude analytique de ce corpus, à savoir les recouvrements entre vocabulaires ainsi que la proximité inter-dialectes au moyen de la distance de Hellinger. Enfin, dans la dernière section de ce chapitre, nous abordons le problème de l'identification des dialectes.

4.2 Méthodologie de construction du corpus parallèle

Le premier corpus crée dans le cadre de cette thèse concerne le dialecte algérois. Nous avons transcrit les scénarios de films et de sitcoms en texte, en éliminant les génériques, les passages musicaux ainsi que tout autre passage vide de parole. Cette transcription a permis la collecte de 2500 phrases que nous avons traduites vers l'arabe standard.

Un corpus du dialecte bonois aussi a été crée de la même manière mais en transcrivant des enregistrement sonores de la vie quotidienne de certaines personnes de Annaba (université, salle d'attente de médecins, ménages). La transcription a permis de collecter 3900 phrases qui ont été traduites aussi vers l'arabe standard.

Afin d'augmenter la taille des deux corpus, ils ont été traduits l'un vers le dialecte de l'autre. Ainsi à l'issue de cette opération, un premier corpus trilingue de 6400 phrases a été crée pour les dialectes algérois et bonois ainsi que pour l'arabe standard.

4.2.1 Pivotage par l'arabe standard et passage à d'autres dialectes

Afin d'introduire les dialectes tunisien, marocain, syrien et palestinien, nous avons utilisé l'arabe comme une langue pivot. La partie arabe standard du corpus trilingue construit comme expliqué ci-dessus a été traduite vers les dialectes tunisien, marocain, syrien et palestinien.

Le corpus tunisien a été produit par 20 locuteurs natifs. Chacun d'entre eux était responsable de la traduction de près de 320 phrases de l'arabe standard vers TUN. Ces locuteurs sont tous du sud de la Tunisie, où les gens ont tendance à utiliser des mots d'origine arabe plutôt que Français contrairement aux gens du nord. En effet, le dialecte utilisé dans le Sud est plus proche de l'arabe standard que celui utilisé dans le nord de la Tunisie (ce phénomène est observé même dans le sud Algérien).

Les corpus marocain, syriens et palestiniens ont été créés de la même manière comme le tunisien, sauf que le marocain a été pris en charge par une seule personne native du Maroc (vivant dans la ville de Rabat), et le syrien ainsi que le palestinien ont été réalisés par deux traducteurs locuteurs natifs de ces dialectes. Le dialecte palestinien concerné par cette étude est principalement le dialecte de la bande de Gaza. Le dialecte syrien quant à lui est celui de la ville de Damas (voir la figure 4.1 et le tableau 4.1).

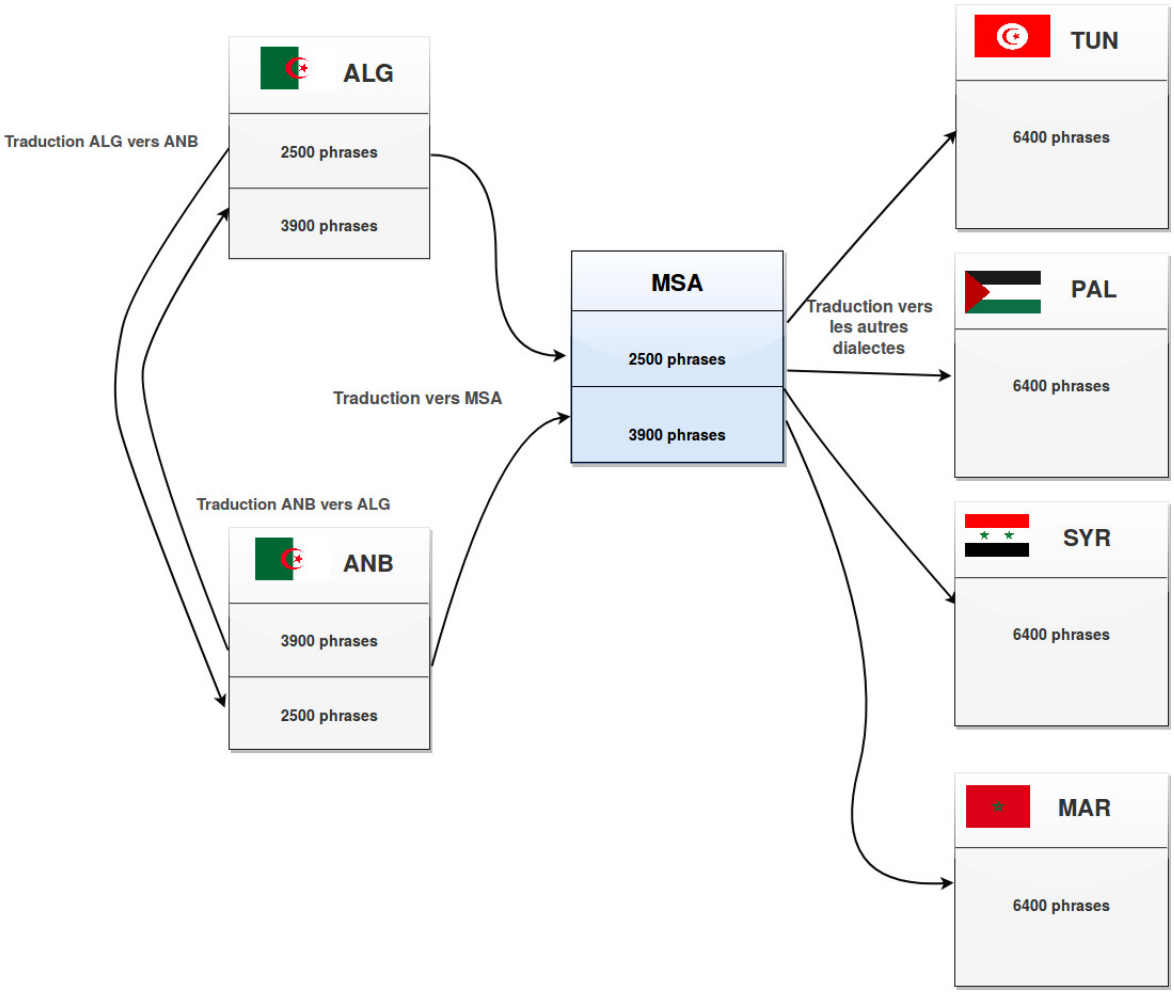

FIGURE 4.1 : Construction du corpus multi-dialecte

Dialecte	Ville du locuteur natif	Nombre d'intervenants	Tache effectuée
ALG	Alger	1	Transcription et traduction
ANB	Annaba	20	Enregistrement de conservation, transcription et traduction
TUN	Sfax	20	Traduction
MAR	Rabat	1	Traduction
SYR	Damas	2	Traduction
PAL	Gaza	2	Traduction

Tableau 4.1 : Ville et nombre de personnes ayant participé à la création du corpus dialectal parallèle

4.2.2 Nécessité d'adopter des règles d'écriture standard

La transcription de la parole en texte ainsi que la traduction vers les dialectes ont nécessité l'adoption d'un certain nombres de règles d'écriture. Cela est dû au fait qu'à l'origine ces dialectes ne sont pas écrits et ne possèdent aucun standard d'écriture. Nous avons donc convenu d'utiliser les règles orthographiques suivantes :

— Chaque mot dialectal a été transcrit en adoptant la notation arabe : si un mot dialectal existe dans l'arabe standard, on adopte sa forme arabe standard, sans aucun changement (cas par exemple des mots ساعة (une montre) et وردة (une rose)), sinon le mot est écrit comme il est prononcé (par exemple مسيد (une école)).

— Adopter l'article de définition ال de l'arabe standard pour les mots dialectaux (par exemple الشكارة (le sac)) et même pour les mots d'origine française (par exemple الكار (le bus)) .

— Utiliser ة pour les mots féminins dialectaux (exemple الزرودية (la carotte)) et même pour les mots d'origine française (par exemple صالة (une salle)).

— Les pronoms liés (à un verbe ou à un nom) s'écrivent comme en arabe standard, comme dans les exemples كتبها (il l'a écrite), ou كتابهم (son livre).

— A l'exception du Tanween (la voyelle double) qui n'existe pas dans le dialecte Arabe, tous les signes diacritiques sont autorisés même la gémination.

— Écrire les mots Français tels qu'ils sont prononcés avec des lettres arabes et utiliser lorsque nécessaire les lettres Françaises telles que le V et le P.

— Lorsqu'il s'agit d'une expression en Français article+nom (éventuellement avec apostrophe ou liaison) écrire l'article et le nom en un seul mot comme par exemple لسيتى (la cité).

— Transcrire le "R" Français tel qu'il est prononcé en ر ou غ.

4.3 Comparaison Analytique

Dans ce qui suit, nous allons comparer les dialectes entre eux et avec MSA. L'idée est de comprendre quels sont les dialectes les plus proches, les plus divergents, etc. Nous espérons que cela nous aidera dans un travail futur à adapter les outils TAL dédiés au MSA pour développer des outils pour le traitement des dialectes.

4.3.1 Statistiques du Corpus Multi-dialectes PADIC

Le corpus parallèle obtenu est constitué de 6400 phrases parallèles.[1]La partie MSA contient 40906 mots, dont 9,131 différents. Les six dialectes, ALG, ANB, TUN, MAR, SYR et PAL comprennent en moyenne de 37500 mots avec un vocabulaire qui ne dépasse pas 10215 mots(voir le tableau 4.2). Le nombre moyen de mots dans une phrase dialectale est de 6 alors qu'il est de 7 pour MSA.

Corpus	#Mots distincts	#Mots
ALG	8966	38707
ANB	9060	38428
TUN	10215	36648
MAR	10897	39178
SYR	9825	37259
PAL	9196	39286
MSA	9131	40906

Tableau 4.2 : Description du corpus parallèle

4.3.2 Unités lexicales communes entre dialectes et l'arabe standard

Il est évident que la langue arabe est la même dans tout le monde arabe, alors que les dialectes varient selon l'emplacement géographique. Nous nous sommes intéressés à la mesure de la proximité entre les vocabulaire dialectaux et

1. PADIC est disponible en téléchargement libre à l'adresse http://smart.loria.fr/pmwiki/pmwiki.php/PmWiki/

MSA en utilisant le corpus parallèle précité. Les résultats que nous avons obtenus, montrent que les dialectes emploient de nombreux mots MSA, même si l'énoncé de ces mots dépend fortement de chaque dialecte. Prenons comme exemple le mot arabe باسل qui signifie courageux, en dialecte algérois ce mot décrit une personne ennuyeuse. Au regard des résultats du Tableau 4.3, nous constatons que PAL est Particulièrement plus proche de MSA que les autres dialectes.

Dialecte	ALG	ANB	TUN	MAR	SYR	PAL
%	21.18	21.07	37.60	27.57	37.36	51.68

Tableau 4.3 : Pourcentage des mots communs entre les dialectes et MSA

Ces résultats ne sont pas surprenants. En effet, les dialectes arabes au Moyen-Orient ont tendance à être plus proche de l'arabe standard que ceux du Maghreb. En outre, il est à noter que les dialectes arabes parlés dans le sud des pays du Maghreb sont plus proches de l'arabe standard que les dialectes parlés au nord. Cela explique les différents taux en termes de mots communs pour les dialectes algériens, marocain et le dialecte tunisien (21.18, 21.07 et 27.57 vs 37.60). En effet TUN est parlé dans le sud de la Tunisie alors que ALG, ANB et MAR sont des dialectes du nord de l'Algérie et du Maroc respectivement.
Dans cette même optique, nous avons aussi identifié les mots les plus fréquents dans le vocabulaire de chaque dialecte appartenant aussi au vocabulaire MSA. Nous donnons dans le tableau 4.4 quelques exemples de ces mots.

De la même manière, nous avons calculé le pourcentage des mots communs entre toutes les paires de dialectes (voir le Tableau 4.5). Les résultats obtenus distinguent les paires ALG et ANB ainsi que PAL et SYR. En effet, ALG et ANB se partagent le plus grand nombre de mots suivis de PAL et SYR. Ces valeurs ne sont pas surprenantes car ALG et ANB sont des dialectes appartenant au continuum

Dialecte	Mots les plus fréquents			
ALG	واحد *un*	صح *vrai*	راح *il va*	كامل *tout*
ANB	واحد *un*	صح *vrai*	راح *il va*	عندک *tu as ou attention*
TUN	كان *il était*	وقت *temps*	واحد *un*	الكل *tout*
MAR	كان *il était*	نهار *le jour*	واحد *un*	عندى *j'ai*
SYR	اليوم *aujourd'hui*	مرة *une fois*	واحد *un*	عندى *j'ai*
PAL	اليوم *aujourd'hui*	واحد *un*	طيب *bien*	راح *il va*

Tableau 4.4 : Les mots communs les plus fréquents entre chaque dialecte et MSA relativement au corpus parallèle

des dialectes arabe algériens. Les dialectes PAL et SYR quant à eux sont utilisés dans la même zone géographique séparées par seulement 175 miles.

Aussi, Les résultats obtenus mettent en relief d'une part les dialectes Maghrébins (ALG, ANB, TUN et MAR), d'autre part les dialectes du Moyen-orient (PAL et PAL). Les taux des mots communs entre les paires des dialectes du Maghreb sont plus élevés que les taux calculés entre les paires de dialectes du Maghreb et le Moyen-orient. Par exemple, ALG se partage plus de mots avec MAR et TUN qu'avec SYR et PAL. Notons, par ailleurs que TUN se distingue des dialectes du Maghreb par rapport au recouvrement avec PAL. Cela peut s'expliquer par la proximité de ces deux dialectes par rapport à l'arabe standard (TUN est un dialecte du sud de la Tunisie proche de l'arabe standard (comme mentionné plus haut).

	Pourcentage des mots communs					
Ref.	ALG	ANB	TUN	MAR	SYR	PAL
ALG	-	73.62	35.43	33.28	24.16	25.43
ANB	72.86	-	34.25	31.39	23.59	25.00
TUN	31.10	30.38	-	25.57	29.79	33.49
MAR	29.91	28.50	26.18	-	26.50	27.67
SYR	21.01	20.73	29.52	25.65	-	44.00
PAL	24.79	24.63	37.20	30.02	49.33	-

Tableau 4.5 : Pourcentage des mots communs inter-dialectes

Nous avons aussi calculé la moyenne des pourcentages des mots communs au niveau de la phrase entre chaque paire de dialectes et entre chaque dialecte et MSA. Pour chaque paire de phrases $k-$alignés $S^k_{L_i}$ et $S^k_{L_j}$ du bitexte (L_i, L_j). Le pourcentage des mots communs est calculé par la formule 4.1 qui correspond au rapport du nombre des mots communs avec le nombre total de mots dans les deux phrases. Ensuite la moyenne de ce rapport est calculé pour toutes les paires de phrases.

$$Ovp(S^k_{L_i}, S^k_{L_j}) = \frac{|S^k_{L_i} \cap S^k_{L_j}|}{|S^k_{L_i} \cup S^k_{L_j}|} \tag{4.1}$$

Dans le Tableau 4.6 nous présentons le recouvrement entre les dialectes arabes et MSA au niveau de la phrase. Les résultats obtenus confirment ceux des deux dernières expériences. PAL est le dialecte le plus proche de MSA suivi par TUN, SYR et MAR, tandis que ALG, ANB sont les plus loin.

	ALG	ANB	TUN	SYR	PAL	MAR
MSA	0.12	0.10	0.16	0.14	0.21	0.14
MAR	0.13	0.11	0.13	0.10	0.13	
PAL	0.13	0.11	0.17	0.21		
SYR	0.09	0.09	0.13			
TUN	0.16	0.13				
ANB	0.32					

Tableau 4.6 : Recouvrement au niveau phrase entre les vocabulaires dialectes-MSA

Cette expérience met également en lumière la proximité entre les dialectes algériens (ALG et ANB) et les dialectes du Levant (PAL et SYR). Elle montre également que TUN est plus proche de PAL et de SYR que ALG, ANB et MAR.

4.3.3 Mesure de la distance de Hellinger entre dialectes et l'arabe standard

Les expérimentations présentées plus haut ont été réalisées sur les corpus ainsi que leurs vocabulaires respectifs. Dans cette section, nous utilisons des modèles de langage uni-grammes pour mesurer la divergence entre les dialectes et MSA. Nous voulons à travers ces expérimentations connaître les paires de dialectes les plus proches et la distance de chaque dialecte par rapport à l'arabe standard. Pour ce faire, nous avons choisi d'utiliser une métrique qui permet de calculer la distance/divergence entre toutes les paires dialecte-dialecte et dialecte-arabe standard. Nous avons trouvé dans la littérature plusieurs métriques qui pourraient être applicables à notre cas, à l'instar de la divergence de Kullback-libleir la mesure du $\chi 2$ ainsi que la distance de Hellinger. Nous avons opté pour cette dernière car contrairement aux deux premières mesures qui sont asymétriques, cette mesure est symétrique et remplie les trois conditions relatives à la distance. Dans la section suivante nous définissons la distance de Hellinger et présentons ses principales propriétés.

La distance de Hellinger

La distance de Hellinger est une métrique qui quantifie la similarité entre deux distributions de probabilité. Elle est appelée aussi la distance de Bhattacharyya car elle a été introduite à l'origine dans [Bhattacharyya, 1943]. Cette distance est utilisée dans différents domaines notamment, pour la détection des défaillances dans la classification[Cieslak and Chawla, 2009] et pour l'estimation des classes dans le domaine de machine learning[González-Castro et al., 2013]. Elle est aussi utilisée pour mesurer la perte d'information en matière de protection des données[Torra and Carlson, 2013].

La distance de Hellinger pour toute mesure de probabilité absolument continue P et Q sur un ensemble fini X est définie par :

$$HD(P,Q) = \sqrt{\frac{1}{2}\int (\sqrt{P(x)} - \sqrt{Q(x)})^2 \mathrm{d}x} \tag{4.2}$$

Pour le cas de deux distributions discrètes $P = (p_1, ..., p_k)$ et $Q = (q_1, ..., q_k)$, la distance de Hellinger est définie par :

$$HD(P,Q) = \sqrt{\frac{1}{2}\sum_{i=1}^{k}(\sqrt{p_i} - \sqrt{q_i})^2} \tag{4.3}$$

Les principales propriétés de la distance de Hellinger se résument dans ce qui suit :

— Elle est symétrique ie $HD(P,Q) = HD(QP)$

— Elle est Positive $HD(P,Q) \geq 0$

— Elle est bornée[2] entre 0 et 1

2. Dans la littérature certains auteurs éliminent le facteur $\frac{1}{2}$ la distance est alors bornée entre 0 et $\sqrt{2}$

— Elle obéit à la règle d'inégalité triangulaire : $HD(P,Q) \leq HD(P,R) + HD(Q,R)$

La distance de Hellinger appliquée aux modèles de langage

Afin de mesurer l'écart entre deux langages avec HD, considérons le bi-texte (L_i, L_j) avec les vocabulaires V_i et V_j respectivement. Nous constituons V, un vocabulaire de 10K mots, constitués des mots communs entre V_i et V_j, et des mots restants des deux vocabulaires nous incluons les plus fréquents pour compléter V. Pour chaque partie du bi-texte, une distribution de probabilités uni-gramme $P(w|L_i)$ est calculée sur V (voir la figure 4.2). Pour éviter les probabilités nulles (dues aux mots n'appartenant pas à la langue considérée), nous avons utilisé une technique de lissage des probabilités. La comparaison des deux distributions est alors calculée comme suit :

$$HD(L_i, L_j) = \sqrt{\frac{1}{2}\sum_{w \in V}(\sqrt{P(w|L_i)} - \sqrt{P(w|L_j)})^2} \tag{4.4}$$

Le Tableau 4.7 retrace les valeurs HD calculées entre tous les dialectes et MSA. Ces valeurs montrent que PAL est le dialecte le plus proche de MSA suivi par TUN puis SYR et MAR, alors que ALG et ANB sont les plus divergents. Les dialectes les plus proches selon HD sont ALG et ANB et également PAL et SYR. La distance de Hellinger calculée entre les différentes paires de dialectes d'une part et entre chaque dialecte et l'arabe standard appuie les résultats obtenus dans les expérimentations précédentes. Pour les dialectes, nous voyons deux groupes se distinguer, il s'agit des dialectes du Maghreb ALG, ANB, TUN et MAR d'une part, PAL et SYR d'autre part. Toutes les expériences confirment la proximité entre les dialectes du Maghreb. Nous constatons aussi, que TUN par rapport aux autres dialectes maghrébins) semble être le plus proche des dialectes PAL et SYR alors que ANB en est le plus loin.

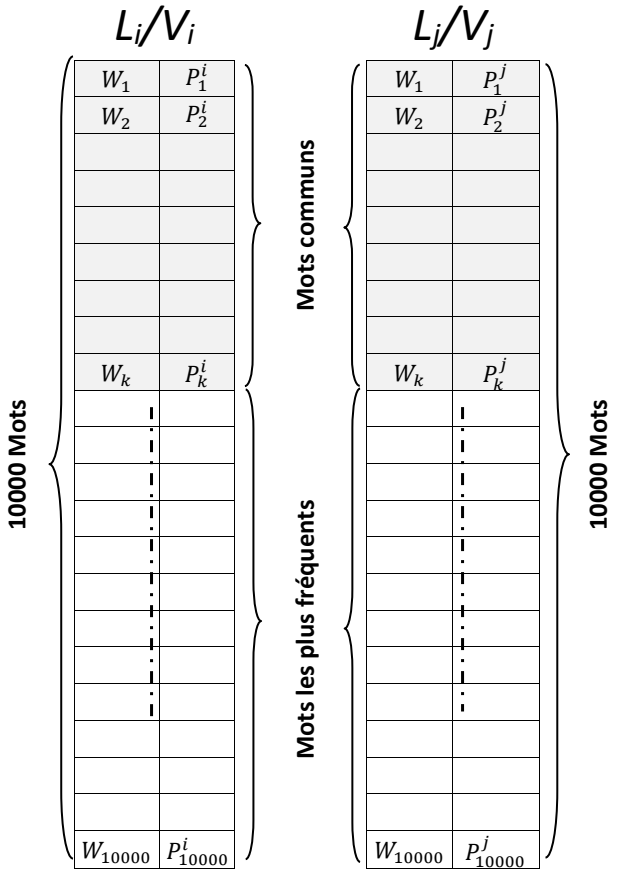

FIGURE 4.2 : Constitution des vecteurs de calcul de la distance HD

4.4 Identification des dialectes

Dans cette partie, nous avons abordé le phénomène du code-switching qui est connu par l'utilisation de mots appartenant à plusieurs langues dans un même discours. Ce phénomène est fréquent dans le monde Arabe et en particulier au Maghreb. Par exemple, en Algérie, les gens dans leurs conversations quotidiennes peuvent passer du dialecte, à l'arabe standard et au Français voir même à la langue berbère. Ce phénomène est caractéristique des dialectes arabes et a crée

	ALG	ANB	TUN	SYR	PAL	MAR
MSA	0.72	0.72	0.60	0.62	0.55	0.64
MAR	0.77	0.79	0.78	0.83	0.84	
PAL	0.85	0.86	0.81	0.76		
SYR	0.84	0.86	0.81			
TUN	0.79	0.80				
ANB	0.73					

Tableau 4.7 : Les valeurs de la distance de Hellinger entre les différentes paires de langages

de nouveaux besoins en matière de modules applicatifs qui les prennent en charge. En effet, nombreuses applications requièrent des outils capables d'identifier ces dialectes et de séparer les contenus dialectaux (en recensant les différents dialectes présents) des autres contenus non-dialectaux notamment arabe standard, Français et même Anglais (pour les dialectes du moyen-orient).

4.4.1 État de l'art

Les auteurs de [Habash et al., 2008] ont proposé un ensemble de règles d'annotation pour identifier le basculement entre l'arabe standard et au moins un dialecte arabe. Ces directives peuvent être utilisées pour annoter de larges corpus de données à contenu hautement dialectal destinés aux applications TAL arabe.

Dans [Zaidan and Callison-Burch, 2012], les auteurs ont présenté un large corpus de données collectées à partir des commentaires d'un journal en ligne. Ce corpus a été annoté en utilisant le crowdsourcing. [3] où les participants à l'opération d'annotation devaient identifier la portion dialectale de chaque phrase ainsi que le dialecte dans lequel la phrase est écrite. Le corpus annoté ainsi obtenu a été utilisé pour l'apprentissage et l'évaluation de classificateurs utilisant des modèles de langage pour l'identification des dialectes.

3. appelée approvisionnement par la foule, il s'agit du transfert de processus de travail(au profit d'une personne ou d'une organisation) vers une main d'œuvre constituée d'un grand nombre d'internautes à titre gracieux ou payant.

Dans [Elfardy and Diab, 2013], les auteurs ont présenté une approche supervisée pour distinguer l'arabe standard du dialecte arabe égyptien. L'approche se base sur l'identification du dialecte au niveau du mot d'abord. Elle utilise un modèle de langaue n-gramme, une analyse morphologique et des techniques de normalisation orthographiques pour annoter chaque mot de la phrase. Plus tard, dans [Elfardy et al., 2014] les mêmes auteurs introduisent un outil d'analyse morphologique muni d'un module de désambiguisation pour réduire au maximum les possibilités d'analyse de chaque mot.

Dans [Cotterell and Callison-Burch, 2014], les auteurs ont présenté un corpus multi-dialecte annoté manuellement avec un module d'identification des dialectes capable d'identifier cinq dialectes arabes. Pour ce faire, ils ont adopté les classificateurs Naïve bayes et les SVM.

Dans [Darwish et al., 2014], l'identification se base sur les traits lexicaux, morphologiques et phonologiques des dialectes considérés. Une amélioration du taux de l'identification de 10% a été rapportée dans ce travail en comptant sur ce type d'information.

4.4.2 Approche adoptée pour l'identification

Nous considérons l'identification des dialectes comme un problème de classification qui sera traité par un classificateur Naïve Bayes (NBC). C'est l'une des modèles d'apprentissage les plus pratiques et les plus utilisés dans les problématiques liées à la classification et catégorisation en raison des performances démontrées dans ce domaine. Ces classifiers présentent beaucoup d'avantages, parmi les plus cités dans la littérature :

— Ils sont fondés sur une théorie mathématique probabiliste précise.

— Simplicité et facilité d'implémentation.

— Efficacité des résultats même avec de petits corpus d'apprentissage.

— L'incrémentation des données d'apprentissage est relativement facile.

Les classificateurs Naïve Bayes

NBC est un algorithme d'apprentissage probabiliste, il est utilisé pour de nombreuses applications TAL. Un classificateur naïf Bayes suppose que toutes les traits liés à un problème donné sont conditionnellement indépendants compte tenu des valeurs des variables de classification. Les classificateurs NB sont utilisés dans différentes applications telles que le filtrage de courrier électronique [Kim et al., 2006], la classification de documents [Pop, 2006, Dumais et al., 1998, Sahami et al., 1998] et désambiguïsation du sens des mots [Pedersen, 2000, Gale et al., 1992, Farag and Nurnberger, 2008]. Pour notre cas, Pour N classes correspondant à n langues, le but est d'assigner la classe la plus adaptée C_i conformément à un ensemble de caractéristiques (features) $F = f_1, ..., f_n\}$ qui maximise la probabilité conditionnelle :

$$p(C_i \mid f_1, ..., f_n) = p(C_i) \prod_{j=1}^{n} p(f_j|C_i) \tag{4.5}$$

où $p(C_i)$ est la probabilité de la classe C_i et $p(f_j|C_i)$est la probabilité conditionnelle de la caractéristique f_j observée avec la classe C_i. Ces probabilités sont calculées par estimation du maximum de vraisemblance :

$$p(C_i) = \frac{Count(C_i)}{\sum_{i=1}^{k} Count(C_i)} \tag{4.6}$$

où $Count(C_i)$ est le nombre d'occurrence de la classe C_i dans le données d'apprentissage, $\sum_{i=1}^{k} Count(C_i)$ est le nombre total d'occurrences de toutes les classes C_i dans ces mêmes données.

Et

$$p(f_j|C_i) = \frac{Count(f_j, C_i)}{\sum_{i=1}^{k} Count(f_j, C_i)} \tag{4.7}$$

où $Count(f_j, C_i)$ est le nombre d'occurrences de la caractéristique f_j et la classe C_i observées simultanément dans les données d'entrainement, et $\sum_{i=1}^{k} Count(f_j, C_i)$ est le total des occurrences de toutes les classes C_i avec les contraintes f_j. Pour éviter les probabilités nulles nous avons introduit le lissage de Laplace qui consiste à ajouter 1 à toutes les occurrences l'équation 4.7 sera alors :

$$p(f_j|C_i) = \frac{Count(f_j, C_i) + 1}{\sum_{i=1}^{k} Count(f_j, C_i) + k} \tag{4.8}$$

Pour l'identification, nous utilisons donc un classifier Naïve Bayes entrainé sur des modèles de langage n-gramme (d'ordre 3)$w_1, ..., w_n$ annotés avec leurs classes qui correspondent au différents dialectes et à l'arabe standard $\{L_i, i = 1..6\}$

.

$$p(L_i \mid w_1, ..., w_n) = p(L_i) \prod_{j=1}^{n} p(w_j|L_i) \tag{4.9}$$

4.4.3 Résultats de l'identification sur PADIC

Pour l'expérimentation, nous avons créé un corpus en fusionnant MSA, ALG, ANB, TUN, MAR, SYR et PAL où chaque phrase est annotée par sa classe de langue correspondante. En choisissant au hasard, 80 % des données, nous avons créé le corpus d'apprentissage et le reste 20% a été consacré au test.

Les résultats de la classification dans le tableau 4.8 montrent que le rappel pour MSA est le plus élevé (75 %) ; cela pourrait être expliqué par le fait que l'écriture MSA obéit à des règles strictes contrairement aux dialectes pour lesquels des règles d'écriture formelles n'existent pas : un mot dialectal pourrait être écrit sous différentes formes qui sont toutes acceptables. Par conséquent, ce phénomène génère une plus grande distribution des probabilité pour les dialectes que pour MSA. Par ailleurs on constate que MAR se distingue des autres dialectes avec le taux de rappel le plus élevé (6%), cela est du à la nature de ce dialecte qui possède

des caractéristiques particulières distinctives par rapport aux autres dialectes. Nous citons comme exemple l'emploi de suffixes verbaux tels que ك et غ et l'utilisation du ك à la place du ق dans les mots courants tels que قالت, قال et قلت.

Langue/Dialecte	Précision	Rappel	F
ALG	0,46	0,47	0,47
ANB	0,48	0,48	0,48
TUN	0,63	0,48	0,55
MAR	0,70	0,68	0,69
SYR	0,61	0,55	0,57
PAL	0,52	0,55	0,53
MSA	0,61	0,76	0,68

Tableau 4.8 : Résultats de l'identification des dialectes

Le tableau 4.9 représente la matrice de confusion du classificateur. Pour les dialectes, il est clairement démontré que les taux de confusion les plus élevés sont ceux entre ALG et ANB et entre PAL et SYR, cette confusion est justifiée par la proximité entre ces paires de dialectes ; ALG et ANB par exemple partagent un vocabulaire important en dépit de leur différence. Du même tableau nous pouvons

	Classe estimée						
Classe réelle	ALG	ANB	TUN	MAR	SYR	PAL	MSA
ALG	**47**	32	5	7	2	3	4
ANB	34	**48**	4	5	2	4	3
TUN	9	7	**48**	8	5	8	14
MAR	8	5	6	**68**	4	3	6
SYR	3	2	4	5	**55**	21	10
PAL	3	3	3	3	15	**55**	18
MSA	2	2	4	3	4	10	**76**

Tableau 4.9 : Matrice de confusion (en %) de l'identification des dialectes

constater que les taux les plus élevés de confusion liés à MSA sont ceux avec PAL, tandis que pour ALG et ANB, les taux de confusion relatifs à MSA ne dépassent pas 4% pour les deux dialectes.

4.4.4 Résultats sur des corpus hors PADIC

Nous avons choisi de tester ce classifier sur un autre corpus de données dialectales et arabe standard. Contrairement à PADIC, ces corpus ne sont pas parallèles. Néanmoins, en termes de taille ces corpus sont plus volumineux (voir le tableau 4.10). Du point de vue couverture, ces corpus les seuls qu'on a pu avoir) sont relatifs à l'algérois, au tunisien et à l'arabe standard puisque nous n'avons trouvé aucun corpus disponible pour le bonois, le palestinien, le syrien ainsi que le marocain. Le corpus algérois a été téléchargé à partir de forum de discussions Algériens, le corpus tunisien téléchargé de pages Facebook [Ameur and Jamoussi, 2013] et le corpus Arabe standard a été aussi téléchargé du web. La répartition apprentissage et test suit la même répartition que plus haut, 80% et 20% respectivement.

Corpus	#Phrases	#mots	#mots différents
ALG	3186	38260	11702
TUN	16289	132090	32883
MSA	21120	268284	45297

Tableau 4.10 : Description des corpus non-parallèles

Le tableau 4.11 retrace les résultats du classifier pour des expérimentations effectuées sur les corpus non-parallèles décrits plus haut. il est clair que ces résultats sont meilleurs que ceux du corpus parallèle PADIC, cela est du principalement à la taille de ces corpus. Notons que même pour cette expérience le taux de rappel de l'arabe standard reste le plus élevé (95%). Pour les dialectes le rappel est de 0.65% pour le tunisien vs 0.56% pour l'algérois. Ces deux taux de rappel devraient être interprétés au regard de la taille de chaque corpus.

On présente dans le Tableau 4.12 la matrice de confusion du classifier entrainé sur les corpus non-parallèles. Il en ressort que le taux de confusion entre le tunisien et l'arabe standard est le plus élevé (31%) comparé à celui de

Langue/Dialecte	Précision	Rappel	F
ALG	0.72	0.56	0.63
TUN	0.79	0.65	0.72
MSA	0.85	0.95	0.90

Tableau 4.11 : Résultats de l'identification en utilisant des corpus non-parallèles

la confusion entre l'algérois et l'arabe (17%). Ces résultats appuient ceux des expériences décrites précédemment. Aussi, l'arabe standard est bien identifié, les taux de confusion relatifs sont tous les deux inférieur à 5%.

	Classe estimée		
Classe réelle	**ALG**	**TUN**	**MSA**
ALG	**56**	27	17
TUN	4	**65**	31
MSA	1	4	**95**

Tableau 4.12 : Matrice de confusion de l'identification en utilisant des corpus non-parallèles

4.5 Conclusion

Dans ce chapitre, nous avons expliqué la méthodologie de construction du corpus multi-dialectes PADIC en pivotant par l'arabe standard et en adoptant des conventions d'écriture orthographique normalisées. Nous avons ensuite procédé à la description de PADIC en fournissant les statistiques relatives à ce corpus. La suite du chapitre a été allouée à une étude comparative des dialectes au regard du corpus PADIC. Nous avons calculé les recouvrements entre les vocabulaires dialecte/dialecte et dialecte/arabe standard (au niveau corpus et au niveau phrastique).

Par la suite, nous avons calculé la distance entre les différentes paires de dialectes et entre chaque dialecte et l'arabe standard au moyen de la distance de Hellinger. Toutes ces expérimentations, vont dans le sens de la proximité entre les

deux dialectes Algériens : l'algérois et le bonois ainsi qu'entre les deux dialectes du moyen-orient palestinien et syrien. Nous avons aussi constaté que la distance de Hellinger entre les différents dialectes montrent que le tunisien et le marocain sont proches des deux dialectes algériens qu'aux dialectes syrien et palestinien. Par ailleurs, les dialectes du Moyen-orient sont plus proches de l'arabe standard suivis du tunisien, du marocain et deux dialectes algériens . Ces résultats nous semblent naturels et conformes à nos attentes.

En termes d'identification, les résultats obtenus montrent que l'arabe standard se distingue facilement des dialectes avec un taux de rappel important. Pour les dialectes, on a pu constater que le marocain se distingue avec le taux de rappel le plus élevé. Cette distinction est due à notre sens à la particularité de ce dialecte par rapport aux autres. Les autres résultats vont tous dans la proximité des deux dialectes algériens et des dialectes palestinien et syrien.

Chapitre 5

La traduction automatique des dialectes arabes au moyen du corpus PADIC

5.1 Introduction

Nous avons pu constater dans le chapitre 2 que les dialectes arabes du corpus PADIC sont des dialectes dépourvus de ressources, en particulier les dialectes maghrébins. Les efforts en termes de traduction automatique sont peu nombreux et demeurent à un stade précoce. Dans ce chapitre, nous présentons les différents systèmes de traduction que nous avons montés au moyen du corpus PADIC. Nous abordons pour la première fois à notre connaissance la traduction inter-dialectes jusqu'à présent non abordée dans les différents travaux que nous avons pu voir.

5.2 Les défis de la traduction automatique des dialectes arabes

L'état de l'art sur les traduction automatique des dialectes tracé dans le second chapitre de cette thèse donne une idée sur l'intérêt accordé aux traitement des dialectes arabes dans le cadre de la TA. Ces dialectes émergent comme de véritables langues pour lesquelles beaucoup de ressources restent à créer et à développer.

La traduction automatique de ces dialectes reste à un état de recherche encore embryonnaire. Plusieurs défis sont à surmonter. Les dialectes comme ils sont présentés dans sont classifiés par pays, par région : Le levantin, l'égyptien, l'algérien, le tunisien,...etc. Cette classification simplifie considérablement la situation linguistique dans le monde arabe. En réalité, dans chaque pays arabe il existe une variante de dialectes avec des spécificités propres à chacun d'entre eux.

Les systèmes de traduction dédiés à ces dialectes doivent tenir compte des toutes ces caractéristiques.

Un autre défi relatif aux dialectes arabes et qui doit être pris en considération est l'alternance de code linguistique plus connu par le code-switching. Les populations arabes alternent dans leur conversation entre le dialecte arabe, l'arabe standard et même d'autres langues. Ce phénomène est observé surtout dans les pays du Maghreb arabe où les gens tendent à utiliser le dialecte, l'arabe, le Français et dans certaines régions même le Berbère. Ce phénomène ajoute un degrés de complexité à la mise en place de systèmes de traduction fiables puisque ces derniers doivent en tenir compte. Ce code-switching peut être une source importante de mots hors vocabulaire spécialement pour les dialectes du Maghreb pour lesquels la normalisation des mots dialectaux en arabe standard et le pivotage via l'arabe peuvent être insuffisants, contrairement aux dialectes du moyen orient. Dans cette même optique, l'évolution rapide des dialectes est un fait que les systèmes de traduction doivent surmonter. Au quotidien de nouveaux mots dialectaux apparaissent et sont adoptés par les communautés spontanément sans validation aucune. Ces mots sont une source évidente de mots hors vocabulaire.

5.3 La traduction automatique statistique, apprentissage avec PADIC

Comme le montre l'étude comparative du chapitre 4, section 4.3, les dialectes, même s'ils sont fortement inspirés de l'arabe, des différences significatives peuvent exister et rendre la communication entre les gens du monde arabe peu confortable. En effet, on observe dans notre quotidien la difficulté que l'on peut rencontrer lorsqu'il s'agit de parler à une personne arabe surtout du moyen-orient. Même si parfois on a tendance à penser que cette communication est simple. On cite à titre d'exemple une conversation entre une personne Algérienne et une personne égyptienne. A première vue, on pensera que la communication se

passera de façon très fluide du fait que le dialecte égyptien (à travers le cinéma et la télévision) est connu dans tout le monde arabe. Beaucoup d'expériences réelles ont montré que cette communication n'est pas toujours évidente et n'est pas aussi facile que l'on imaginait. On a souvent recours à l'arabe standard ou même à l'utilisation du français ou l'anglais pour passer une idée qu'on arrive pas communiquer. Dans cette optique, nous proposons la traduction automatique entre les dialectes arabes et l'arabe standard en utilisant notre corpus multi-dialectes PADIC.

5.3.1 Description de l'environnement des expérimentations

Nous avons développé un système de traduction statistique pour chaque paire de langues (dialectes) de PADIC. Tous ces systèmes sont entraînés sur des bi-textes de 5900 paires de phrases, et 500 paires de phrases ont été consacrées à l'évaluation. Nous avons utilisé le modèle standard d'un système de traduction automatique à base de segments en l'occurrence Moses [Koehn et al., 2007] pour l'apprentissage et le décodage, Giza++[Och and Ney, 2003]pour l'alignement ainsi que SRILM[Stolcke, 2002] pour le calcul des modèles de langage. On notera que nous avons migré vers KenLM [Heafield, 2011] pour les modèles de langage car cet outil est beaucoup plus rapide que SRILM et gère plus efficacement l'espace mémoire. Vu que la taille du corpus parallèle est faible, nous avons expérimenté les techniques de lissage Kneser-Ney et Witten-Bell dans l'espoir d'identifier celui qui convient le mieux.

5.3.2 Résultats obtenus

Les résultats réalisés sur l'ensemble de test sont présentés en termes de BLEU dans le tableau 5.1, de TER dans le tableau 5.2 et METEOR dans 5.3.

Au vue des scores BLEU obtenus(tableau 5.1), nous pouvons aboutir à des conclusions très intéressantes. Tout d'abord, pour les petits corpus parallèles, il semble que la technique de lissage a un impact sur les résultats de la traduction. Une différence de près de 2 points en termes de scores BLEU a été observée pour la traduction entre ANB et ALG. Mais, nous ne pouvons pas généraliser en affirmant que l'une technique de lissage est certainement mieux que l'autre.

	Cible													
	ALG		**ANB**		**TUN**		**MAR**		**SYR**		**PAL**		**MSA**	
Source	KN	WB	KN	WB	KN	WB	KN	WB	KN	WB	KN	WB	KN	WB
ALG	-	-	61.06	60.81	9.67	9.36	10.22	9.55	7.29	7.95	10.61	10.14	15.1	14.64
ANB	67.31	65.55	-	-	9.08	8.64	10.00	9.41	7.52	7.95	10.12	9.84	14.44	13.95
TUN	9.89	9.48	9.34	9.01	-	-	14.37	14.26	13.05	12.93	22.55	22.21	25.99	25.21
MAR	10.13	9.71	10.16	9.52	14.68	14.20	-	-	9.68	9.50	18.91	18.75	24.93	24.44
SYR	7.57	7.50	7.50	7.64	13.67	13.23	9.93	9.74	-	-	26.60	25.74	24.14	22.96
PAL	11.28	10.67	9.53	9.15	17.93	16.64	16.08	15.83	23.29	23.07	-	-	40.48	39.76
MSA	13.55	13.05	12.54	11.72	20.03	20.44	20.02	20.17	21.38	20.32	42.46	41.37	-	-

Tableau 5.1 : Le score BLEU pour la traduction automatique entre dialectes et MSA en utilisant deux techniques de lissage

Un score élevé de la traduction a été réalisé entre ANB et ALG dans les deux côtés. Ce résultat est naturel puisque ces deux dialectes sont utilisés dans le même pays et partagent jusqu'à 60 % des mots. La même observation est faite pour la paire SYR et PAL puisque ces deux dialectes appartiennent à la même famille de langues (le levant).

	Cible													
	ALG		**ANB**		**TUN**		**MAR**		**SYR**		**PAL**		**MSA**	
Source	KN	WB	KN	WB	KN	WB	KN	WB	KN	WB	KN	WB	KN	WB
ALG	-	-	21.41	21.75	75.17	76.37	75.06	76.10	79.54	79.51	69.63	70.75	65.63	66.85
ANB	17.12	17.81	-	-	74.83	75.62	75.10	76.90	79.13	79.13	69.10	70.26	67.40	68.47
TUN	71.10	71.76	73.13	73.71	-	-	64.65	64.25	66.03	66.55	51.20	51.57	50.85	51.30
MAR	74.42	75.13	74.83	75.7	66.47	67.20	-	-	74.92	73.80	56.86	56.93	52.97	53.12
SYR	76.89	77.67	76.89	77.67	66.54	67.91	71.18	72.20	-	-	32.28	33.24	52.81	53.59
PAL	71.43	72.51	72.25	73.47	58.51	59.65	61.17	61.80	32.44	33.86	-	-	36.74	36.87
MSA	67.02	67.91	68.94	70.16	57.18	57.28	56.96	57.03	56.60	57.08	34.00	34.66	-	-

Tableau 5.2 : Le score TER (en %) pour la traduction automatique entre dialectes et MSA en utilisant deux techniques de lissage

Un autre résultat intéressant et attendu est le score BLEU obtenu par la traduction entre MSA et les dialectes. En fait, le score le plus élevé est celui relatif à la paire PAL et MSA puisque ce dialecte est le plus proche à MSA comme il a été constaté dans les expériences des sections 4.3.2 et 4.3.3.

	Cible													
	ALG		**ANB**		**TUN**		**MAR**		**SYR**		**PAL**		**MSA**	
Source	KN	WB	KN	WB	KN	WB	KN	WB	KN	WB	KN	WB	KN	WB
ALG	-	-	0.452	0.450	0.181	0.178	0.186	0.183	0.161	0.164	0.202	0.199	0.222	0.218
ANB	0.472	0.464	-	-	0.172	0.172	0.179	0.179	0.156	0.159	0.196	0.194	0.200	0.200
TUN	0.186	0.183	0.182	0.182	-	-	0.209	0.207	0.203	0.203	0.261	0.260	0.268	0.266
MAR	0.179	0.181	0.179	0.184	0.202	0.205	-	-	0.175	0.176	0.240	241	0.263	263
SYR	0.155	0.154	0.159	0.157	0.195	0.190	0.174	0.173	-	-	0.359	0.356	0.259	0.256
PAL	0.189	0.185	0.187	0.183	0.229	0.225	0.224	0.224	0.365	0.360	-	-	0.341	0.339
MSA	0.205	0.203	0.201	0.199	0.242	0.245	0.246	0.245	0.247	0.247	0.359	0.356	-	-

Tableau 5.3 : Le score METEOR pour la traduction automatique entre dialectes et MSA en utilisant deux techniques de lissage

Les résultats les plus surprenants sont ceux relatifs à SYR et TUN. Il semble qu'il est plus facile de traduire TUN à MSA que SYR à MSA. En outre, la traduction de MSA à TUN donne de meilleurs résultats que de MSA aux dialectes algériens. Dans la partie symétrique de traduction nous obtenons la même échelle de résultats. Cela montre certainement la proximité de TUN à MSA en comparaison avec les dialectes algériens ou-bien qu'il existe un biais dans la traduction du corpus d'apprentissage du MSA vers le tunisien. Il est clair qu'au vue des résultats en termes de score TER et METEOR, les constations énumérées plus haut restent valides. Il apparaît que les techniques de lissage n'ont pas d'impact sur les scores, les différences sont négligeables.

Dans les travaux cités dans l'état de l'art, on remarque que la traduction entre les dialectes arabes et l'arabe standard fait appel à de petits corpus et que les résultats de la traduction sont donnés en termes de précision et rappel. Aussi pour la traduction entre dialectes nous n'avons trouvé aucun travail qui permet de comparer les résultats obtenus. Nous avons alors jugé intéressant d'évaluer des systèmes de traduction automatique entraînés sur des corpus du même ordre de

grandeur que PADIC. Pour ce faire, nous avons mis en place plusieurs systèmes de traduction entre l'arabe standard et l'Anglais avec des corpus parallèles de tailles différentes téléchargés du WMT. Les corpus utilisés étaient de $20K$, $50K$, $120K$ et $150K$ phrases parallèles qu'on notera C_{20K}, C_{50K}, C_{120K} et C_{150K} respectivement. L'environnement du système de traduction est comme celui décrit plus haut. Nous avons constitué 20 corpus de test de 500 phrases chacun (pour rester dans les mêmes conditions que les expériences sur PADIC). Nous avons alors calculé les différentes valeurs du score BLEU en utilisant aussi les deux techniques de lissage (KN et WB). Nous avons retenu juste les valeurs maximales et minimales du BLEU pour chaque corpus d'apprentissage. Le Tableau 5.4 montre les résultats obtenus. En comparant les scores BLEU obtenus, on constate que les valeurs relatives aux

	KN		**WB**	
Corpus	Min	Max	Min	Max
C_{20k}	4.25	9.56	4.1	8.97
C_{50k}	5.15	11.75	5.18	11.32
C_{120k}	5.94	14.38	5.95	14.13
C_{150k}	6.13	14.39	6.19	14.27

Tableau 5.4 : Valeurs maximales et minimales du score BLEU pour des systèmes de traduction entraînés sur de petits corpus avec deux techniques de lissage

expériences de PADIC sont meilleures. A titre d'exemple, les scores obtenus sur le corpus C_{20K}, le plus élevé étant de 9.56 et le plus faible de 4.25, alors que pour PADIC le plus faible score BLEU enregistré est 7.57 pour la traduction du syrien vers l'algérois. Sachant que ce corpus (C_{20K}) est de taille 3 fois plus grande que PADIC. La proximité entre les dialectes d'une part et la proximité entre l'arabe standard et ses dialectes d'autre part justifient à notre avis les résultats du BLEU relativement acceptables comparés aux résultats des expérimentations sur de petits corpus pour la traduction entre l'arabe standard et l'Anglais.

5.3.3 Étude de l'impact des techniques de lissage sur les scores de la traduction

Nous voulons dans cette partie de chapitre confirmer ou infirmer l'impact des techniques de lissage sur le score de la traduction automatique pour de petits corpus de données tels que PADIC. En effet, nous avions dans la section précédente constaté les différences entre les scores BLEU pour les mêmes paires de langue en utilisant les deux techniques de lissage (KN et WB). Malheureusement, la taille du corpus PADIC ne nous permet pas de le diviser en sous corpus de tailles acceptables afin de procéder aux tests de significativité statistique. Nous avons alors effectué ces tests sur les systèmes de traduction entre l'arabe standard et l'Anglais qu'on a décrits dans la section 5.3.2. Le tableau 5.5 trace les données des tests de signification sur les distributions du score BLEU relativement aux deux techniques de lissage où *Min*, *Max*, *E[X]*, σ^2 représentent respectivement le *minimum*, le *maximum*, la *moyenne* et la *variance*. σ_{XY} et *valeur-p* correspondent respectivement à la *covariance* et à la *valeur-p* des deux distributions. Nous avons utilisé le test statistique T-test après s'être assurés que les deux distributions suivent une loi Gaussienne. L'hypothèse émise H_0 est que les deux distributions sont similaires en termes de moyenne.

	KN				**WB**					
Corpus	Min	Max	$E[X]$	σ^2	Min	Max	$E[X]$	σ^2	σ_{XY}	*Valeur-p*
C_{20k}	4.25	9.56	6.64	2.47	4.1	8.97	6.43	2.23	2.33	0.33
c_{50k}	5.15	11.75	8.15	3.26	5.18	11.32	7.99	2.92	3.16	0.35
C_{120k}	5.94	14.38	9.58	4.62	5.95	14.13	9.35	4.32	4.45	0.36
C_{150k}	6.13	14.39	9.91	4.85	6.19	14.27	9.74	4.72	4.75	0.39

Tableau 5.5 : Valeurs du test de significativité statistique

En interprétant les valeurs du tableau 5.5, il apparaît que les résultats obtenus pour les deux techniques de lissage ne sont pas distinguables. En effet, pour chaque corpus d'apprentissage et pour les 20 tests différents les résultats sont

équivalents en termes de *minimum*, *maximum*, *moyenne* et *variance* des scores BLEU.

Par ailleurs, la covariance est positive pour toutes les expérimentations, ce qui signifie que les deux distributions sont linéairement dépendantes. Nous constatons aussi que la *valeur-p* est supérieure au risque α fixé à 0.05 quelque soit le corpus d'apprentissage. Cela implique qu'il y a au moins un risque de 33% pour accepter l'hypothèse alternative H_1. Nous pouvons donc conclure que même s'il y a des différences dans les scores BLEU calculés pour les deux techniques de lissage, ces différences ne sont pas significatives.

5.3.4 Impact de l'interpolation des modèles de langage sur les scores de la traduction

On s'intéresse dans cette section à l'incidence de l'interpolation des modèles de langage sur les scores de la traduction. Nous essayons d'exploiter la proximité entre l'arabe standard et ses dialectes en interpolant les modèles de langage basés sur PADIC avec des modèles de langage arabe standard, dans l'espoir de voir les scores BLEU s'améliorer. Nous avons gardé le même contexte d'expérimentation que plus haut, mais nous avons calculé des modèles de langages interpolés pour chaque système de traduction. Tous les modèles de langage ont été interpolés un à un avec le modèle de langage du corpus LDC Arabic Treebank (Part3,V1.0) après avoir déterminé les poids d'interpolation optimaux pour chaque modèle. On notera aussi que nous avons gardé la techniques de lissage (KN) pour ces tests vu que cela n'a pas une grande influence sur les résultats d'après les expérimentations conduites plus haut. Nous donnons dans le tableau 5.6 les scores BLEU obtenus pour les différents systèmes de traductions avec des modèles de langage interpolés.

Dans le tableau 5.7 nous calculons les variations du BLEU (avant et après interpolation) exprimé en pourcentage pour les différents systèmes de traduction.

	Cible						
Source	ALG	ANB	TUN	MOR	SYR	PAL	MSA
ALG	-	61,02	9,71	6,90	6,99	10,01	15,11
ANB	67,86	-	9,15	6,45	6,64	9,44	14,45
TUN	9,86	9,32	-	9,77	12,84	21,15	26,26
MAR	6,99	7,41	10,94	-	6,83	13,10	19,89
SYR	7,68	7,50	13,66	6,03	-	24,27	24,24
PAL	11,30	9,54	18,25	10,46	21,17	-	40,10
MSA	13,61	12,52	19,88	13,83	20,13	39,47	-

Tableau 5.6 : Score BLEU des systèmes de traduction avec interpolation des modèles de langage

	Cible						
Source	ALG	ANB	TUN	MOR	SYR	PAL	MSA
ALG	-	-0,065	0,413	-32,48	-4,11	-5,65	0,066
ANB	0,81	-	0,77	-35,5	-11,7	-6,71	0,069
TUN	-0,303	-0,214	-	-32,01	-1,609	-6,221	1,038
MAR	-30,99	-27,06	-25,47	-	-29,44	-30,72	-20,21
SYR	1,453	0,00	-0,073	-39,27	-	-8,75	0,414
PAL	0,177	0,104	1,765	-34,95	-9,102	-	-0,938
MSA	0,442	-0,159	-0,748	-30,91	-5,825	-7,041	-

Tableau 5.7 : Variations en (%) du score BLEU des systèmes de traduction avec interpolation des modèles de langage

Les scores enregistrés avec l'interpolation n'affichent aucun gain par rapport aux résultats de la section 5.3.2. Au contraire dans certains cas ils sont même pénalisant. Pour la traduction de l'arabe standard vers les dialectes, nous avons une chute de (6.19) points pour pour la paire MSA-MAR et (2.99) pour la paire MSA-PAL. Dans l'autre sens de la traduction (des dialectes vers l'arabe standard), on ne constate aucun effet signifiant de cette interpolation, les scores varient avec un maximum de ($\pm$0.38) en termes de BLEU, une exception est à noter pour la paire MAR-MSA ou la différence est de ($-$5.04). Pour la traduction inter-dialectes, la plus large différence entre les scores BLEU est enregistrée pour la paire de dialectes MAR-PAL (5.81).

5.4 Conclusion

Dans ce chapitre, nous avons commencé par présenter les défis relatifs à la traduction automatique des dialectes arabes. Par la suite, nous avons décrit les différents systèmes de traductions statistiques (inter-dialectes et dialecte-arabe standard) que nous avons entraînés avec le corpus PADIC. Nous avons donné les résultats de la traduction entre chaque paire de langue/dialectes en termes de BLEU, TER et METEOR. Les résultats de la traduction appuient ceux de la comparaison analytique du chapitre 4. En effet, les scores de la traduction des dialectes algériens entre eux et des dialectes du Moyen-Orient (entre eux aussi) sont les meilleurs, ce qui est expliqué par la proximité entre ces paires de dialectes. Aussi, le tunisien est plus facile à traduire vers l'arabe standard que les deux dialectes algériens et le dialecte marocain. Parmi tous les dialectes, le palestinien affiche le meilleur score de traduction de et vers l'arabe standard. Outre ces résultats, nous avons aussi monté des systèmes de traduction entre l'arabe standard et l'anglais. Ces systèmes ont été appris sur des corpus de tailles de même ordre que PADIC. Les scores de traduction obtenus pour ces systèmes sont moins performants que ceux de PADIC en dépit de sa taille. Nous avons aussi étudié l'impact des techniques de lissage sur les scores de la traduction. En effectuant le test de significativité statistique, nous avons conclu que la technique de lissage n'a pas d'incidence sur le score de la traduction automatique des dialectes arabes. De même, nous avons interpolé des modèles de langage afin de voir leur répercussion sur les résultats de la traduction, cette interpolation n'avait pas d'effet notable, elle a même influé négativement sur certains scores.

Chapitre 6

Outils TAL pour le dialecte algérois

6.1 Introduction

La langue Anglaise par exemple, la langue du web par excellence est la langue la plus dotée en termes de ressources. On recense pour cette langue des milliers d'outils de traitement automatique couvrant tous les niveaux du langage. Viennent ensuite des langues comme le français l'espagnole, le japonais et la langue chinoise qui a émergé cette dernière décennie. La langue arabe quant à elle est dotée d'un certains nombre d'outils qui restent toujours limités du point de vue de la prise en charge de tous les phénomènes langagiers vu sa richesse et sa robustesse. Les dialectes arabes moins dotés que l'arabe standard ont été longtemps ignorés par les applications relatives au TAL, les premiers travaux s'y afférant sont récents comparés à la langue arabe et remontent au années 2000. En outre, cet intérêt pour les dialectes arabes ne concerne que certains dialectes du moyen orient, peu d'attention est accordée aux dialectes du Maghreb ; encore moins au dialecte algérien. C'est pourquoi nous avons dédié une partie importante de nos efforts à la création des ressources basiques pour le dialecte algérois dans l'optique de généraliser ces ressources pour tous les dialectes arabes Algériens. Ce chapitre décrit toutes les ressources que nous avons pu créer.

6.2 La voyellation automatique pour les textes en dialectes algérois

La voyellation, la diacritisation ou encore la restauration des signes diacritiques représente l'un des défis majeurs pour le traitement automatique de la langue arabe. En effet, l'absence de voyelles dans les textes arabes (standards et

dialectaux) génère un nombre d'ambiguïtés considérables au niveau morphologique, syntaxique et même sémantique et pragmatique. Différentes applications liées au TAL de la langue Arabe requièrent des textes voyellés et la majorité d'entre elles font appel à des modules de restauration des signes diacritiques. La quasi-totalité des textes arabes contemporains (journaux, livres, bulletin d'information, …) sont non voyellés, cette absence de voyelles ne pose en général pas de problèmes à la compréhension humaine des textes ; même si dans certains cas la voyellation n'est pas intuitive même pour un être-humain. La voyellation automatique quant à elle demeure un problème non encore résolu pour de nombreuses langues dont la langue arabe. Les dialectes arabes sont aussi concernés par ce problème, l'absence des marques diacritiques dans les textes dialectaux peut être un défi pour certaines applications telles que la conversion Graphème-Phonème.

6.2.1 La voyellation pour la langue arabe et le dialecte algérien

L'alphabet arabe se compose de 28 lettres [1] dénotant des consonnes et trois voyelles longues (ا, و et ى). La graphie arabe comprend les voyelles courtes (ـَ *a*, ـُ *u*, ـِ *i*) et d'autres symboles phonétiques qui sont représentés par les signes diacritiques (placés au-dessus ou au-dessous des consonnes et des voyelles longues). Les voyelles courtes peuvent apparaître partout dans le mot, le *Tanween* (la voyelle double) inclut les trois cas (ـً *an*, ـٌ *un*, ـٍ *in*), il apparaît seulement à la fin des mots et exprime l'indétermination nominale. Les signes diacritiques arabes comprennent aussi la marque de gémination *Shadda* qui désigne une consonne double (qui peut être combiné avec les voyelles courtes et doubles), et le *Sukun* qui dénote l'absence de voyelle. Selon sa fonction dans la phrase, un mot arabe peut se mettre dans l'un des trois cas : l'accusatif, le nominatif et le génitif dénoté respectivement par les voyelles courtes et doubles (ـَ *a*,ـً *an*, ـُ *u*,ـٌ *un*, ـِ *i*,

1. comprenant 14 consonnes solaires qui assimilent le ل d'un article défini précédant ال et 14 consonnes lunaires qui ne l'assimilent pas.

ٍ *in*) en combinaison éventuellement avec la gémination.[2] Il reste à noter que l'absence des signes diacritiques en Arabe est une source d'ambiguïté. Un mot sans signes diacritiques peut avoir de nombreuses vocalisations valides au niveau lexical et même au niveau syntaxique en fonction de sa catégorie grammaticale (voir le Tableau 6.1).

Forme diacritisée	Sens	Catégorie Grammaticale
فَسَّرَ	il a expliqué	verbe (voix active)
فُسِّرَ	il a été expliqué	verbe (voix passive)
فَسِّرْ	explique	verbe (impératif)
فَسِرْ	alors marche	conjonction+ verbe
فَسِرٌّ	et un secret	conjonction+noun
فَسْرٌ	Déclaration/notification	Noun

Tableau 6.1 : Vocalisations possibles du mot arabe فسر

Le dialecte algérois écrit en script arabe utilise le système vocalique de l'arabe standard à l'exception du *Tanween* (la voyelle double), il élimine aussi la marque du cas en remplaçant les signes diacritiques à la fin des mots par le *Sukun* (absence de voyelles). Ceci simplifie considérablement le processus de vocalisation mais génère une ambiguïté syntaxique. A titre d'exemple, il est parfois difficile dans des phrases simples de distinguer l'agent de l'objet sans recourir à des informations sémantiques (voir un exemple détaillé dans le Tableau 6.2). Au niveau lexical, l'absence de voyelles en dialecte algérois peut aussi être source d'ambiguïté (voir le Tableau 6.3).

2. Nous avons évoqué juste la flexion des noms aux trois cas pour montrer l'importance des voyelles, nous attirons l'attention du lecteur que la flexion des noms (lorsqu'il se mettent au pluriel), des particules ainsi que la conjugaison des verbes obéissent à d'autres règles.

Phrase Dialectale	Phrase en Arabe standard	Sens	Agent	Objet
شَمَعْ وَلِيدْ عُمَرْ	سَمِعَ وَلِيدٌ عُمَرَ	Walid a entendu Omar	Walid	Omar
	سَمِعَ وَلِيداً عُمَرُ	Omar a entendu Walid	Omar	Walid

Tableau 6.2 : Exemple de l'ambiguïté syntaxique causée par la suppression de la marque du cas dans le dialecte algérois

Mot Dialectal	Vocalisation Valide	sens
جوز	جَوَّزْ	il a passé/ passe (verbe transitif)
	جُوزْ	passe (verbe intransitif)
	جُوزْ	noix
قريت	قْرِيتْ	j'ai lu/étudié ou tu as lu/étudié
	قَرِّيتْ	j'ai enseigné/tu as enseigné

Tableau 6.3 : Exemple de l'ambiguïté lexicale causée par l'absence des voyelles

6.2.2 Travaux de la voyellation automatique pour l'arabe standard

La plupart des travaux sur la diacritization sont dédiés à l'arabe standard. Plusieurs approches ont été adoptées. Dans [Emam and Fischer, 2004], les auteurs présentent un système basé sur une recherche hiérarchique aux niveaux de la phrase, du syntagme, du mot et du caractère. A partir du niveau de la phrase, le système tente de récupérer des exemples diacritisés à partir des données d'apprentissage, si aucun exemple approprié n'est trouvé pour la phrase donnée, le système segmente la phrases en syntagmes et recherche des exemples diacritisés. Si la recherche échoue au niveau du syntagme, celui-ci est à son tour segmenté en mots, le système recherche des exemples de mots diacritisés. Une recherche au niveau du caractère est lancée si la recherche au niveau du mot échoue . Le système utilise des n-grammes pour la définition des exemples.

Dans [El-Sadany and Hashish, 1989] la diacritisation a été considérée comme un système de traduction automatique à base de règles entre textes voyellés et non-voyellés. Les insuffisances de ce système sont ceux des systèmes

de traduction à base de règles : l'ajout de nouveaux cas nécessite la définition de nouvelles règles qui rend la maintenance du système coûteuse.

Dans [Nelken and Shieber, 2005], les auteurs ont combiné les transducteurs à états finis probabilistes appris sur le corpus LDC arabic Treebank. Le système intègre un modèle de langage tri-gramme de niveau mot et un modèle de langage basé caractère ainsi qu'un module d'analyse morphologique très simple basé sur un ensemble réduits de clitiques (préfixes et suffixes).

Dans [Schlippe et al., 2008] deux approches sont utilisées, la première combine un système de traduction statistique automatique avec un voyelleur à base de règles, la seconde considère la diacritisation comme un problème d'étiquetage de séquences. La première approche opère à différents niveaux (la phrase entière, le mot, le caractère, et une combinaison du mot et du caractère). Le système de traduction statistique est utilisé pour la post-édition, le corpus parallèle d'apprentissage comprend la sortie du système de voyellation à base de règles alignée au texte correctement voyellé de cette sortie. Le second système utilise les champs aléatoires conditionnels (Conditional Random Fields CRF) afin de prédire la séquence correcte des signes diacritiques pour une séquence de consonnes non-diacritisées.

Dans [Gal, 2002] et [Elshafei et al., 2006] les modèles de Markov (HMM) sont utilisés pour la diacritisation, les mots non-diacritisés sont considérés comme des observations et leurs diacritisations possibles sont les états cachés qui ont produit ces observations. L'algorithme de Viterbi est utilisé pour définir les états cachés les plus adéquats. Il convient de noter que ces deux travaux utilisent le Coran comme corpus d'apprentissage et de test.

Dans [Zitouni et al., 2006] la voyellation est considérée comme un problème de classification de séquences : étant donné une séquence de caractères X, chaque caractère est marqué par son signe diacritique (une séquence d'étiquettes Y est

alors obtenue). L'objectif du système est d'attribuer la séquence d'étiquettes Y pour X séquence de consonnes, les auteurs proposent une approche statistique fondée sur le principe de maximum d'entropie en utilisant des traits lexicaux et les étiquettes morpho-syntaxiques.

6.2.3 Le système de voyellation automatique réalisé

Nous avons construit un système de traduction automatique statistique basé sur un corpus parallèle de textes voyellés et non voyellés. Le système est à base de segments avec les paramètres par défaut : un modèle de distorsion (à sept contraintes), une pénalité au niveau du mot et la phrase ainsi qu'un modèle de langage. Le système utilise un alignement au niveau des mots à l'aide de l'outil GIZA++, une table de traduction avec des entrées voyellées et non voyellées, un modèle de langage tri-gramme entrainés sur des textes voyellés.

Corpus utilisés

L'utilisation de la traduction automatique statistique suppose la disponibilité de corpus parallèles pour les langues source et cible. Pour construire une telle ressource dans notre cas, il suffit de procéder à la suppression des signes diacritiques d'un corpus voyellé, ou de diacritiser un corpus non voyellé. Les deux voies ont été exploitées dans notre travail, la première pour l'arabe standard et la seconde pour dialecte algérois.

1. **Corpus arabe standard**

   Tashkeela Au début de ce travail, le seul corpus disponible pour nous était Tashkeela un corpus voyellé libre sous licence GPL. Ce corpus est une collection de livres arabes classiques téléchargés depuis une bibliothèque en ligne . Il se compose de plus de 6M mots. Nous avons commencé par supprimer les symboles

spéciaux, des chiffres et certains caractères non Arabes du corpus vocalisé, nous avons par la suite segmenté les longues phrases composant le corpus en phrases plus courtes en utilisant les signes de ponctuation.Une fois le corpus nettoyé, les signes diacritiques ont été supprimés et donc nous avons pu obtenir un corpus non voyellé. Les données issues de ce corpus se composent d'environ 1200K paires de phrases. Leur répartition était comme suit : Données d'apprentissage (80%), développement (10%) et test (10%), cette répartition a été faite en attribuant au hasard une série de livres pour chaque tâche.

LDC Arabic Treebank (Part3,V1.0) Pour pouvoir situer nos résultats par rapport aux travaux réalisés dans la domaine de la restauration des signes diacritiques dédiés à la langue Arabe, nous avons été contraints d'utiliser le corpus LDC Arabic Treebank (Part3,V1.0) [3], qui est un corpus largement utilisé dans plusieurs applications TAL en particulier celles dédiées à la restauration de signes diacritiques [Nelken and Shieber, 2005, Schlippe et al., 2008, Zitouni et al., 2006]. Ce corpus est un ensemble de 600 documents collectés auprès du Journal Annahar. Il comprend 340K mots. Pour exploiter cette ressource, nous avons construit un corpus vocalisé en explorant toutes les étiquettes annotant ce corpus (l'annotation a été réalisée manuellement par l'équipe de LDC), nous avons extrait pour chaque mot sa diacritisation correspondante. La répartition du corpus en données d'apprentissage, développement et test a suivi celle du 1_{er} corpus à savoir (80%,10% et 10%) en allouant au hasard un ensemble de documents à chaque tâche.

2. **Le corpus algérois**

Le grand défi de ce travail est la disponibilité du corpus algérois. Nous avons

3. Voir en annexe A les détails sur la licence gratuite à ce travail dans le cadre du programme LDC Scholarship

exploité une partie du corpus cité plus haut (crée par nos soins). Nous avons effectué une diacritisation manuelle de cette partie du corpus pour l'utiliser comme données d'apprentissage. Cette tâche a été coûteuse en temps et en effort humain. Le corpus du dialecte algérois voyellé se compose de 4K paires de phrases avec 23K mots.

6.2.4 Évaluation du système

Pour l'évaluation du système, nous avons utilisé WER (Word Error Rate) : le pourcentage des mots dont la diacritisation est erronée. Au niveau du caractère, nous utilisons DER (Diacritization Error Rate). On calcule ces taux d'erreur en utilisant Sclite [4] qui trouve les alignements entre une référence et une hypothèse aux niveaux des mots et des caractères. Un mot est considéré comme incorrectement voyellé si au moins l'un de ses caractères a au moins un signe diacritique incorrect. Nous avons évalué aussi notre système par le calcul des taux de précision et de rappel. Ces mesures sont largement utilisées pour évaluer de nombreux systèmes de TAL.

Résultats pour l'arabe Standard

Le tableau 6.4 qui suit récapitule les différentes valeurs relatives à la diacritisation pour les corpus arabe standard.

Corpus	Taskeela		LDC ATB	
Distribution	WER	DER	WER	DER
Substitution	16.2	1.8	23.1	0.5
Suppression	0.0	1.9	0.0	5.1
Insertion	0.0	0.3	0.0	0.1
Total(WER/DER)	16.2	4.1	23.1	5.7

Tableau 6.4 : Les résultats WER/DER (pour les corpus d'arabe standard)

4. Sclite : Partie du NIST SCTK Scoring Toolkit http://www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm

Pour les valeurs de rappel et de précision, le tableau suivant (6.5) en donne les résultats.

Niveau	Mot		Caractère	
Corpus	Rappel	Précision	Rappel	Précision
Taskeela	83.8%	85.2%	95.9%	93.1%
ATB	76.9%	89.2%	94.3%	96%

Tableau 6.5 : Résultats Précision/Rappel(pour les corpus d'arabe standard)

Nous avons constaté en observant les erreurs de diacritisation que la majorité de ces erreurs sont relatives aux marques du cas. Nous avons effectué une évaluation en ignorant les marques du cas, le WER et DER ont régressé de près de 50

Résultats pour le dialecte algérois

Du tableau 6.6, nous constatons que les WER et DER sont élevés par rapport aux valeurs obtenues pour l'arabe standard, ces résultats doivent être analysés en prenant en compte la taille des différents corpus utilisés. En effet, au regard de la taille minime du corpus algérois par rapport à Tashkeela et ATB, ces résultats sont très satisfaisants.

Distribution	WER	DER
Substitution	25.8	0.1
Suppression	0.0	12.6
Insertion	0.0	0.1
Total(WER/DER)	25.8	12.8

Tableau 6.6 : Les résultats WER/DER pour le corpus du dialecte algérois

En termes de rappel et précision le tableau 6.7 résume les résultats obtenus.

Level	Mot		Caractère	
Corpus	Rappel	Précision	Rappel	Précision
Dialecte algérois	74.2%	96.3%	87.2%	98%

Tableau 6.7 : Les résultats Précision/Rappel pour le corpus du dialecte algérois

Il est important de noter que nous avons effectué plusieurs tests en augmentant à chaque fois la taille du corpus d'apprentissage, nous avons observé que lorsque la taille du corpus augmente d'un petit taux, le WER et DER régressent. A titre de comparaison, nous avons également fait de nombreux tests pour de petits Corpus d'arabe standard extraits de Tashkeela et ATB et dont les tailles sont du même ordre que la taille du corpus algérois, Nous avons enregistré un WER et un DER respectivement de plus de 52,5 et 20,5 à chaque fois (pour les deux corpus). Ces résultats sont expliqués par le fait qu'en Arabe standard un mot non voyellé a plus de possibilités de vocalisation qu'un mot non voyellé en dialecte (absence de la marque du cas pour le dialecte et des voyelles doubles diminuent le nombre de possibilités).

6.2.5 Vocalisation du corpus algérois

Compte tenu du taux de précision élevé (au regard de la taille du corpus d'apprentissage), nous avons utilisé un processus itératif pour vocaliser le reste du corpus dialectal. La partie du corpus non voyellée a été segmentée en petites parties de 200 phrases. Nous avons vocalisé les premières 200 phrases en utilisant le système ci-dessus. Par la suite, nous avons procédé à la correction des erreurs (manuellement), cette tâche n'a pas été coûteuse en temps (cela est dû au taux de précision comme indiqué plus haut). Le corpus de 200 phrases ainsi voyellé est ajouté au corpus d'apprentissage du système de voyellation. Nous avons itéré cette opération jusqu'à vocalisation du corpus entier (Voir la Figure 6.1).

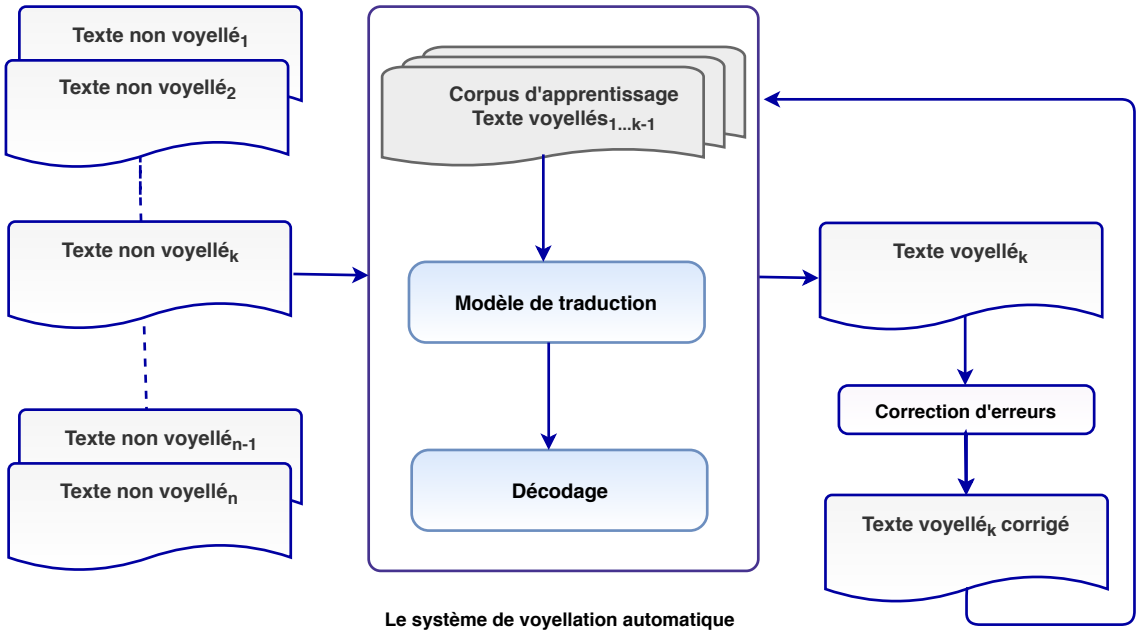

FIGURE 6.1 : Le processus de diacritisation itérative

6.3 La conversion graphème-phonème

La conversion graphème-phonème (CGP appelée aussi phonétisation ou encore la transcription orthographique-phonétique) consiste à transformer un texte orthographique à prononcer en une suite de phonèmes. La conversion graphème phonème n'est pas toujours une opération simple à réaliser surtout pour les langues non transparentes telle que l'anglais où la relation entre graphèmes et phonèmes n'est pas toujours biunivoque : un phonème peut être représenté par un graphème ou un groupe de graphèmes et vice-versa. Contrairement à la langue anglaise par exemple, l'arabe est considéré comme une langue transparente ; en fait, la relation entre graphèmes et phonèmes est presque biunivoque, mais notons que cette particularité est conditionnée par la présence de signes diacritiques. L'absence de la voyellation génère des ambiguïtés au niveau phonétique et par conséquent aux niveaux lexical, syntaxique et sémantique. Le mot كتب / KTB / par exemple, peut avoir différentes transcriptions phonétiques comme /kataba/,

/Kutiba/, /kutubun/, /Kutubi/, /katbin/ suivant ses marques diacritiques. Le dialecte algérois obéit au même principe : sans signes diacritiques, la conversion GP est un problème difficile à résoudre. Pour palier à cette difficulté, nous avons développé un système de restauration des signes diacritiques pour les textes en dialecte algérois.

6.3.1 Approches de conversion graphème phonème

Il existe deux grandes familles d'approches pour la CGP

Approche par dictionnaire

La conversion graphème-phonème se fait grâce à un lexique qui associe à une entrée sa forme phonétisée, la conversion se voit réduite à la recherche de la représentation phonétique d'un mot dans ce dictionnaire.

Approche par règles

La conversion se fait par application de règles de phonétisation, ces règles peuvent être déduites d'étude phonologique et phonétique de la lange considérée, ou bien apprises sur un corpus phonétisé (il s'agit là d'une approche statistique basée sur des quantités de donnés considérables). La plupart des systèmes de conversion actuels sont considérés comme des systèmes hybrides, il combinent à la fois l'utilisation des règles et de lexiques (surtout des lexiques d'exception pour le traitement des cas particuliers).

Étant donné que le dialecte algérois est non doté de ressources, nous ne pouvions envisager d'utiliser une approche basée sur dictionnaire vu qu'aucun dictionnaire phonétisé n'est disponible. Nous avons alors opté pour deux solutions l'une à bases de règles et l'autre statistique afin de prendre en considération toutes les spécificités phonétiques du dialecte algérois.

6.3.2 Problèmes liés à la CGP pour le dialecte algérois

La CGP pour le dialecte algérois obéit aux mêmes règles que celle relatives à l'arabe standard. En effet, lorsque le texte dialectal est voyellé l'alignement entre graphèmes et phonèmes est de type un à un. Ceci nous conduit à adopter une approche à base de règles pour construire un convertisseur GP. Par ailleurs, le dialecte algérois comporte des mots empruntés de langues étrangères (surtout du Français). Son vocabulaire contient de nombreux mots français utilisés dans les conversations quotidiennes. Ces mots Français peuvent être classés en deux catégories : la première comprend des mots français phonologiquement altérés comme le mot فَاملية (famille) et la seconde comprend des mots qui sont prononcés comme en français comme le mot (سور (sûr). Cette dernière catégorie constitue un problème pour la conversion GP puisque ces mots ne respectent pas les règles de prononciation arabe.

Mot dialectal	transcription phonétique	Mot français
كوزينة	/ku :sina/	Cuisine
طَابلة	/t'a :bla/	Table
كونكسيون	/kɔnnɛksjɔ̃/	Connexion
دوفيز	/dəviz/	Devise

Tableau 6.8 : Exemples de mots Français utilisés dans le dialecte algérois

Dans les exemples du Tableau 6.8, bien que les deux premiers mots soient des mots Français, ils sont phonétisés comme mots arabes. Le mot français «table» est phonologiquement modifié et écrit en dialecte algérois (en script Arabe) /t'a :bla/. Par contre, les deux derniers mots sont phonétisés comme mots Français, car ils sont prononcés comme en français par les locuteurs du dialecte algérois. Afin de tenir compte de cette catégorie de mots, les phonèmes français comme /ɛ/,/ɔ̃/ and /ə/ doivent être inclus dans la liste des phonèmes du dialecte. Notons que, selon l'interlocuteur, on peut trouver un même mot dans les deux

catégories. Un exemple de ceci, le mot فَامليّة (famille), qui est un mot français (famille) phonologiquement modifié est largement utilisé dans dialecte algérois, mais même le mot فَامى (famille) avec la prononciation française est utilisé.

6.3.3 Approche à base de règle

Après l'étape de diacritisation décrite en 6.2.5, le texte est converti en phonèmes en appliquant un ensemble de règles décrites dans ce qui suit. On soulignera que la plupart de ces règles sont celles adoptées pour l'arabe standard et ne sont applicables que pour les mots arabes et les mots étrangers phonologiquement altérés dans notre corpus. Soit :

— BS la marque de début d'une phrase,

— ES la marque de fin d'une phrase,

— BL le caractère espace,

— C une consonne,

— V une voyelle,

— LC une consonne lunaire,

— SC une consonne solaire,

— LV une voyelle longue.

Une règle de conversion peut être écrite comme suit :

$$LFT + GR + RGT \Longrightarrow /PH/$$

$LFTetRGT \in \{BS, ES, BL, C, V, LC, SV, LV\}$ La règle peut être lue comme suit : un graphème GR ayant respectivement comme contexte gauche et droit LFT et RGT est converti en phonème PH. Les contextes gauche et droit peuvent être un graphème, un séparateur de mot, le début ou la fin d'une phrase ou vide. Nous donnons dans ce qui suit quelques règles de conversion (la totalité des règles peuvent être trouvées en annexe dans l'article relatif à la CGP).

1. Les règles relatives aux ذ , ظ et ث

   Dans le dialecte algérois les graphèmes ذ , ظ et ث ne sont très utilisé, dans la plupart des cas ils sont réalisés comme les graphèmes د, ض and ت, respectivement.

2. Règles relatives aux phonèmes non Arabes : L'alphabet du dialecte algérois comporte trois lettres non Arabes G, V et P.

3. L'article de définition ال

   — L'article de définition est prononcé lorsqu'il est suivi d'une consonne lunaire (qui n'assimile pas le ا). Exemple : القمَر (la lune) $\Longrightarrow$ /laqmar/

   Cette règle s'applique aussi l'arabe standard avec la différence que le ا est prononcé si l'article de définition est en début de phrase.

   — Lorsque l'article de définition ال est suivi d'une consonne solaire le ل n'est pas prononcé et la consonne suivante est géminée (doublée). Exemple : السقَف (le plafond)$\Longrightarrow$ / ?assqaf/

   — Lorsqu'il est précédé par la voyelle longue ى et suivi par une consonne solaire, l'article de définition ainsi que la voyelle longue sont omis et la consonne solaire est géminée (doublée).

   Exemple : فى الدَار (à la maison) $\Longrightarrow$ فدَّار $\Longrightarrow$ /fddAr/

4. La marque du cas

   La marque du cas en dialecte algérois est le Sukun (absence de voyelles), la dernière consonne du mot dans ce dialecte ne prend aucun signe diacritique.

   Exemple : قبل (avant) $\Longrightarrow$ /qbal/

5. Règles relatives aux voyelles longues

   Lorsque ا , و and ى apparaissent dans le mot précédées par les voyelles courtes ـَ , ـُ and ـِ , respectivement, leur voyelles longues correspondantes sont générées.

   Exemples :

كَاس (un verre) $\Longrightarrow$ /ka :s/

فُول (beans) $\Longrightarrow$ /fu :l/

كبِير (a well) $\Longrightarrow$ /kbi :r/

6. Règle relative à la Hamza

   Lorsqu'un mot en dialecte algérois commence par la Hamza, sa représentation phonétique débute avec la consonne occlusive glottale. A la fin d'un mot lorsque la hamza est précédée par ا elle n'est pas prononcée.

   Exemple : أسكت (tais-toi) $\Longrightarrow$ / ?askut/ and سماء (le ciel) $\Longrightarrow$ /sm ?/

   lorsqu'elle se trouve en milieu du mot, La hamza est remplacée par les voyelles longues ا or ى. Par exemple les mots بئر (un puits) et فأس (poleax) correspondent à /bi :r/ et /fa :s/, respectivement.

7. Règle relative au Alif Maqsura ى

   Alif Maqsura ى (est toujours précédée par la fathaا) et se trouve toujours en fin du mot, elle est réalisée comme la voyelle courte /a/.

   Exemple : رمى (il a jetté) $\Longrightarrow$ /rmaa/

8. La règle relative au Alif madda آ

   Alif madda آ est réalisée comme alef / ?/ avec la voyelle longue /a :/. Exemple : آمن (il a cru)$\Longrightarrow$ / ?a :man/

9. les mots se terminant avec ة Le ة n'est pas prononcé en Dialecte algérois contrairement à l'arabe standard où il est réalisé avec les deux phonèmes /t/ et /h/ (en fonction de la position du mot dans la phrase).

$$\{BL, ES\}+ة+\{C, V\} \Longrightarrow /Null/$$

   Exemple : طفلة (une fille)$\Longrightarrow$ /t'afla/

10. Mots se terminant par ه Le ه n'est pas prononcé en Dialecte algérois lorsqu'il est précédé par ـُ . Exemple : كتَابُه (son livre) $\Longrightarrow$ /kta :bu/

11. Mots contenant la séquence ب, ن Lorsque ن est suivi par ب , le ن est prononcé comme م

$$\{\text{ ب}\}+\text{ن}+\{C, V\} \Longrightarrow /m/$$

Exemple : مَنبَر (estrade) $\Longrightarrow$ /mambar/

12. Règle de la gémination

lorsque la shadda (la voyelle dénotant la gémination) apparait sur une consonne, cette consonne est doublée.

Exemple : سُكُّر (sucre) $\Longrightarrow$ /sukkur/

Il convient de noter que la plupart de ces règles pourraient être appliquées à d'autres dialectes algériens et aussi aux dialectes arabes des pays voisins tels que le tunisien et le marocain.

Résultats de la CGP avec l'approche à base de règles

Nous avons utilisé le corpus voyellé du dialecte ALgérois décrit plus haut. Il se compose de plus de 6K phrases dont 10.7k mots différents. Ce corpus est composé de trois catégories de mots :

1. Mots arabes.
2. Mots français phonologiquement altérés dont la prononciation est réalisée avec des phonèmes arabes.
3. Mots français dont la prononciation est réalisée avec des phonèmes français.

En considérant seulement les mots Arabes et les mots Français phonologiquement altérés (catégorie 1 et 2) les résultats de la conversion sont de 100%. Ce résultat régresse à 92%. en incluant les mots Français réalisés avec des phonèmes Français. En effet nous ne pouvons introduire de règles pour les mots français écrits en caractères arabes, puisque la relation entre graphèmes arabes et phonèmes français n'est pas de type un à un. Par exemple, le graphème و dans un mot français écrit

en caractères arabes pourraient correspondre aux phonèmes français /y/, /u/, /ɔ/ ou /O/ (voir quelques exemples dans le tableau 6.9).

Mot dialectal	Transcription phonétique française	Mot Français
سور	syʀ	Sûre
پور	pɔʀ	Port
سودور	sudœʀ	Soudeur

Tableau 6.9 : Exemples de mappings entre le graphème arabe ـو et les phonèmes Français

6.3.4 Approche statistique

L'approche à base de règles que nous avons adoptée plus haut ne tient pas compte des mots français utilisés dans le dialecte algérois dont la réalisation est en phonèmes Français. L'approche statistique nous parait une piste pour résoudre ce genre de problème. Nous avons considéré la CGP comme un problème de traduction automatique dans lequel la langue source est le texte (un ensemble de graphèmes) et la langue cible est sa représentation phonétique (un ensemble de phonèmes). La principale motivation d'utiliser une approche statistique est que nous pouvons inclure les phonèmes français dans les données d'apprentissage.

Pour construire ce système, le premier composant requis est un corpus parallèle comprenant un texte et sa représentation phonétique. Cette ressource n'étant pas disponible, nous l'avons créée à partir du corpus voyellé du dialecte algérois (voir Figure 6.2). Nous avons utilisé le convertisseur à base de règles décrit ci-dessus pour convertir les mots arabes et les mots français phonologiquement altérés (catégorie 1 et 2) du corpus algérois en phonèmes arabes. Alors que les mots français (de catégorie 3) ont été identifiés puis translittérés en caractères latins (manuellement) ensuite convertis en phonèmes français en utilisant un convertisseur français libre. Par exemple, le mot كونكسيون est translittéré en "connexion", puis converti à /kɔnnɛksjɔ̃/.

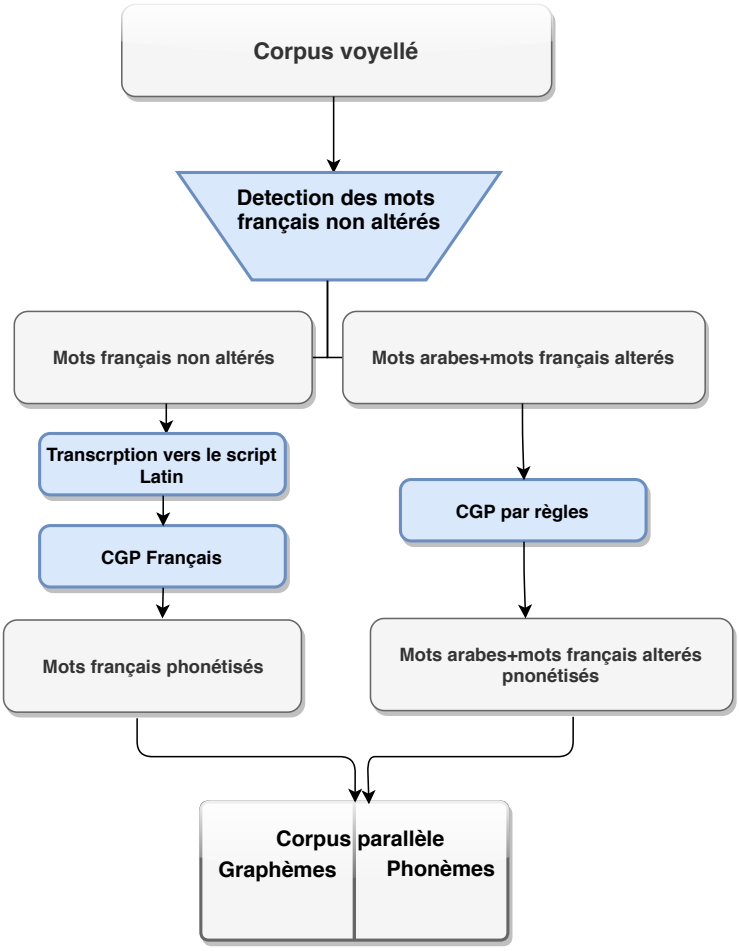

FIGURE 6.2 : Création de corpus parallèle graphème-phonème

Ce système opère au niveau du graphème et phonème, nous avons divisé les corpus parallèles en graphèmes et phonèmes individuels, en incluant un caractère spécial comme séparateur de mots pour rétablir le mot après le processus de conversion (voir exemple ci-dessous dans le tableau 6.10).

◌	ت	ـُـ	ك	◌	سـ	ـَـ	ن
Null	/t/	/u/	/k/	Null	/s/	/a/	/n/

Tableau 6.10 : Exemples de graphèmes et phonèmes alignés

Résultats de la CGP avec l'approche statistique

Pour les besoins d'évaluation, nous avons divisé (au hasard) le corpus parallèle graphème-phonème construit plus haut en trois ensembles de données : données d'apprentissage (80%), les données de tuning (10%) et les données de test (10%). Nous avons d'abord testé l'approche statistique sur un ensemble de test ne contenant que des mots arabes et des mots français phonologiquement altérés (catégorie 1 et 2). Nous avons obtenu une précision de 93%. Puis nous avons procédé à un test sur un corpus contenant les trois catégories de mots, la précision du système à régressé à 85%. Ce résultat est dû à l'augmentation du nombre d'hypothèses de chaque graphème en raison de l'introduction de phonèmes français dans les données d'apprentissage. Le graphème و par exemple dans certains mots arabes (catégorie 1) est phonétisé en phonèmes français /y/, /u/, /ɔ/ ou /O/ au lieu de la voyelle longue / u : /, le phonème /ɔ̃/ au lieu de /u : n /, alors que, quelques mots dans la catégorie 3 sont phonétisés avec des phonèmes arabes en remplaçant par exemple les phonèmes / y /, / u /, / ɔ / ou / O / par / u : / et / ɛ / en / a : /. Dans la tableau 6.11 nous résumons les résultats obtenus.

Approche	Approche Statistique		Approche à base de règles	
Mots	Mots de cat. 1 et 2	Mots de cat.1, 2 et 3	Mots de cat. 1 et 2	Mots de cat.1, 2 et 3
Précision	93%	85%	100%	92%

Tableau 6.11 : Résultats obtenus

À première vue, nous pouvons déduire que l'approche à base de règles est plus efficace que l'approche statistique, mais réellement cette approche ne tient pas compte des mots français de la catégorie 3 ; les résultats sont efficaces seulement pour les mots arabes et français phonologiquement altérés (catégories 1 et 2). Les résultats de l'approche statistique doivent être analysés en tenant compte de la faible taille du corpus d'apprentissage. Le taux de 85% peut être facilement amélioré en augmentant la taille des données d'apprentissage.

6.4 Un analyseur morphologique pour dialecte algérois

Le dialecte algérois simplifie considérablement les règles de l'arabe classique d'une part, comporte une base lexicale assez riche principalement des mots d'origine Arabe, Turque, Berbère et Française. Cependant, le dialecte Algérien en général est classé comme une langue peu ou non dotée en termes de ressources en matière de traitement automatique des langues. Nous dédions une partie de ce travail au développement de ressources pour ce langage. La composante de base de tout système TAL est bien évidemment l'analyseur morphologique. Dans cette optique, nous essayons de réaliser un analyseur de façon rapide en tirant profit des outils disponibles pour l'arabe standard et en prenant en considération les spécificités du dialecte algérois.

6.4.1 Les analyseurs morphologiques dialectaux

Comparé à l'arabe standard, peu d'analyseurs sont dédiés aux dialectes ou du moins prennent en charge les dialectes arabes en plus de l'arabe standard. Les travaux dans ce domaine peuvent être divisés en deux catégories :

— La première catégorie inclut des analyseurs entièrement développés à partir de zéro, comme celui décrit dans [Habash and Rambow, 2006b, Altantawy et al., 2011] baptisé MAGEAD.[5] Il s'agit d'un analyseur (et générateur aussi) basé sur des règles linguistiques de racines et de schèmes ne nécessitant aucun lexique. MAGEAD a été évalué pour l'arabe standard et le Levantin.

— La seconde catégorie inclut des travaux qui tentent de tirer profit des analyseurs existants pour l'arabe standard pour les adapter à l'analyse des dialectes. Il

5. Dans [Altantawy et al., 2011] il s'agit MAGEAD-Express, une version plus rapide de MAGEAD.

convient de noter que cette approche est la plus utilisée du fait de son cout réduit en temps et en efforts.

Différents travaux adoptant cette approche existent. Dans [Salloum and Habash, 2011], les auteurs ont utilisé l'analyseur morphologique BAMA[Buckwalter, 2002](pour Buckwalter Arabic Morphological Analyser), un analyseur bien connu par la communauté scientifique. Ils ont étendu la table des affixes de BAMA avec les affixes des dialectes Levantin et égyptien.

De façon identique, les auteurs de [Afify et al., 2006] ont conçu un analyseur pour le dialecte Irakien analyseur dont l'objectif était l'amélioration d'un système de reconnaissance la parole en dialecte Irakien.

Les auteurs de [Almeman and Lee, 2012] ont adapté par contre l'analyseur morphologique Al-Khalil [Boudlal et al., 2011] en enrichissant son dictionnaire d'affixes avec une liste d'affixes de quatre dialectes arabes.

Les auteurs de [Habash et al., 2012] ont converti un lexique Egyptien (ECAL, Egyptian Colloquial Arabic Lexicon) en une représentation similaire au dictionnaire de SAMA[Graff et al., 2009] (Standard Modern Arabic Analyser).

6.4.2 Approche

Pour le développement de notre analyseur dédié au dialecte algérois, nous avons choisi d'adapter un analyseur existant pour l'arabe standard. Ce choix est justifié par l'économie de temps et d'effort. Nous avons opté pour l'adaptation de l'analyseur morphologique de Buckwalter, baptisé AraMorph, bien connu dans la littérature. Il est employé au sein de l'étiqueteur grammatical de l'arabe au LDC, Penn Treebank de l'arabe, et la Dependancy Treebank de l'arabe à Prague. La simplicité de mise en œuvre de cet analyseur nous a poussés à l'adapter pour le dialecte algérois. En effet, ce dialecte obéit aux mêmes règles que l'arabe standard même si ces règles sont simplifiées d'une part. D'autre part, la majorité

des décisions de l'algorithme sont codées à l'intérieur du lexique : il nous suffit donc de modifier le lexique arabe standard par celui du dialecte et de procéder à des modifications minimes au niveau du code pour la prise en charge du dialecte algérois. Cette approche nous a parue la plus optimale en termes d'efforts et de temps.

Aramorph

Le système adopte une approche procédurale pour la représentation des différentes ressources linguistiques. Certaines règles orthographiques nécessaires à l'analyse sont construites directement dans le lexique lui-même au lieu d'être spécifiées en tant que règles générales qui interagissent pour produire les sorties de l'analyse. Les données lexicales sont représentées dans trois tables : la table des préfixes, des suffixes et des lexèmes. Chaque entrée du lexique prend un lexème comme forme de base et fournit des informations qui concernent la racine, la catégorie grammaticale et la traduction anglaise, excepté les préfixes et suffixes.

Ces trois tables sont en interaction avec trois tables de compatibilité morphologique afin de pouvoir traiter les concaténations. Ces tables sont utilisées pour contrôler les combinaisons entre préfixes et lexèmes, entre lexèmes et suffixes, et entre préfixes et suffixes. Elles sont alors utilisées pour préciser les possibilités de co-apparition des catégories morphologiques. Par exemple, la catégorie morphologique de la conjonction de coordination و (et), Pref-Wa, est compatible avec tous les radicaux nominaux et radicaux verbaux à l'accompli. Toutefois, Pref-Wa n'est pas compatible avec les radicaux verbaux à l'inaccompli puisque ces derniers nécessitent un préfixe sujet.

Ce système d'analyse à base de lexèmes utilise un algorithme d'analyse assez simple puisque toutes les décisions d'analyse sont codées dans le lexique et les matrices de compatibilité. Cependant, lorsqu'il s'agit de l'analyse d'une forme

agglutinée, les segmentations ne seraient valables que si les différents composants existaient dans le lexique et sont triplement compatibles (préfixe - lexème, lexème - suffixe et préfixe - suffixe).

L'adaptation

Nous avons commencé par la modification des dictionnaires. Nous avons remplacé les dictionnaires par ceux du dialecte algérois : les principaux changements ont eu lieu au niveau de l'élimination des marques du cas, du duel et du féminin pluriel.

1. **Modification du dictionnaire des préfixes**

   La taille de ce dictionnaire a été réduite à une soixantaine d'entrées pour le dialecte. Cette réduction est du au fait que certains préfixes de l'arabe standard n'existent pas pour le dialecte algérois tels que :

   - les prépositions س (dénotant le futur) et ف (conjonction de coordination
   - Les préfixes relatifs au duel féminin et masculin
   - Les préfixes relatifs au féminin pluriel

   Ces préfixes ont été supprimés de la tables des préfixes dialectaux ainsi que tous les préfixes composés où ils apparaissent(voir le tableau 6.12).

Préf.	**Description**
ف	Préf. de conjonction
س	Préf. du verbe au futur
فبال	Préf. de conjonction+ Préf. de Préposition Pre.+Préf. art. Définition

Tableau 6.12 : Exemples des préfixes Arabe standard supprimés de la table des préfixes du dialecte algérois

Nous avons aussi gardé certains préfixes inchangés comme les préfixes يـ et تـ (relatifs à la troisième personne du singulier masculin et féminin

respectivement) qui précédent les verbes au futur, des exemples sont fournis dans le tableau 6.13.

Préfixe retenu	**Description**
ي , ت	Préfixe du verbe au futur(sing.,3eme personne,masc.,fém.)
ال	Préfixe nominale (article de définition)
ل , ب	Préfixe de préposition

Tableau 6.13 : Exemples des préfixes Arabe standard retenus dans la table des préfixes du dialecte algérois

Certains préfixes propres au dialecte algérois ont été aussi rajoutés à la table des préfixes pour tenir compte de toutes les spécificités morphologiques de ce dialecte (voir le tableau 6.14).

Préfixes	**Description**
ف	Préfixe de préposition
فال	Préf. de préposition+ Préf. article de définition
ين	Préf. du verbe au passé (voix passive, (sing., masc.) et (plu, masc/fém.))
تن	Préf. du verbe au passé (voix passive, (sing. fem.)

Tableau 6.14 : Exemples des préfixes dialectaux ajoutés à la table des préfixes du dialecte algérois

2. **Modification du dictionnaire des suffixes**

   La taille de ce dictionnaire a été réduite à plus de 240 entrées pour le dialecte. Pour ce dictionnaire aussi nous avons supprimé tous les suffixes qui ne s'appliquent pas au dialecte algérois (voir Tableau 6.15). On en recense principalement :

   - Les suffixes relatifs au duel féminin et Masculin
   - Les suffixes relatifs au féminin pluriel
   - Les suffixes des marques du cas le nominatif, l'accusatif et le génitif (qui s'appliquent au noms singuliers)
   - Le suffixe ون relatif au masculin pluriel (dans le cas du nominatif)

— Les suffixes des verbes relatifs au nominatif, l'apocopé et l'accusatif

Il faut noter qu'en plus des suffixes simples supprimés, nous avons aussi supprimé tous les suffixes composés où ils apparaissent.

Suffixe supprimé	**Description**
ن	Suf. du verbe au passé et futur(sujet, plu., fém.)
تما	Suf. du verbe au passé et futur (sujet, duel, fém/masc., 2eme personne)
هما	Suf. du verbe au passé et futur (objet direct, duel., fém/masc., 3eme personne)
ون	Suf. nominatif (masc.,plu.)
ان	Suf. nominatif (masc.,duel)
هن	Suf. du verbe au passé et futur (objet direct, plu., fem.)
تهن	Suf. du verbe au passé (sujet sing.,2eme personne, masc., objet direct, plu., 3eme personne, fém.)

Tableau 6.15 : Exemples des suffixes supprimés de la tables des suffixes

Nous avons par ailleurs introduit les suffixes purement dialectaux :

— Le suffixe ش relatifs à la négation

— Les suffixes composés qui combinent le suffixe ش avec tous les suffixes relatifs aux pronoms personnels comme تش et همش

Pour tenir compte des variations orthographiques de la transcription des mots dialectaux, nous avons introduit certains suffixes qui en tiennent compte comme par exemple :

— Le suffixe verbal و qui exprime le pluriel(féminin et masculin) lorsqu'il est rajouté à la fin du verbe [6]

— Le suffixe nominal و qui exprime un pronom nominal lorsqu'il se trouve en fin d'un nom à l'instar du suffixe arabe standard ه

Dans le tableau 6.16 sont donnés des exemples.

Cependant certains suffixes arabe standard appartiennent aussi au dialectes algérois, donc on les a préservés dans la table des suffixes (voir des exemples dans le Tableau 6.17).

6. L'équivalent des suffixes arabe standard وا et ون.

Suff.	Description
ش	Suf. de la négation du verbe au passé et futur
همش	Suf. de la négation du verbe au passé et futur (objet direct,plu., 3eme personne, masc./fém)
كمش	Suf. de la négation du verbe au passé et futur (objet direct, plu., 2eme personne, masc./fém.)
و	Suf. du verbe au passé et futur (objet direct, plu.,masc.,fém.)

Tableau 6.16 : Exemple des suffixes du dialecte algérois ajoutés dans la table des suffixes

Suff.	Description
ين	Suf. du nom à l'accusatif/génitif(masc.,plu.)
ات	Suf. du nom(fém.,plu.)
ت	suf. du verbe au passé (fém.,sing)

Tableau 6.17 : *Exemples des suffixes arabe standard retenus dans la table des suffixes du dialecte algérois*

3. **Modification du dictionnaire des lexèmes**

   Il s'agit du dictionnaire le plus volumineux, il a été modifié en entier. Seule la partie relative aux noms propres a été maintenue sans majeure modification. Ce dictionnaire a été construit en entier, pour ce faire nous avons exploité les lexèmes de l'arabe standard contenus dans le dictionnaire de BAMA ainsi que le vocabulaire voyellé du corpus algérois cité plus haut (85%, correspondant à 9170 mots distincts, les 15% restant et correspondant à 1619 mots distincts ont été retenus pour le test).

   Lexèmes à partir du vocabulaire du corpus algérois Nous avons commencé par extraire une liste des noms facilement identifiables par les affixes ة et l'article de définition ٱلـ (compatible uniquement avec les noms). Nous avons supprimé ces deux affixes de tous les mots extraits. A partir de la liste de mots obtenue nous avons créé des lexèmes selon le modèle BAMA. Le reste du corpus a été analysé et classé en trois catégories : les mots de

fonction, les verbes et les noms (qui ne comprennent pas les suffixes ة et اَل) puis converti en lexèmes BAMA.

Nous avons aussi introduit de nouvelles catégories de lexèmes pour tenir compte de quelques caractéristiques propres au dialecte algérois. pour illustrer cette idée, nous donnons l'exemple des lexèmes arabe standard relatifs à la catégorie des verbes conjugués au passé et qui ont le schème فَعَل couvrent les trois personnes, les deux genres, le singulier, le dual et le pluriel. Le lexème reste inchangé il suffit de lui rajouter juste les affixes adéquats pour avoir ses différentes formes fléchies. Pour le dialecte algérois, chaque catégorie de lexèmes de ce type est éclatée en deux catégories distinctes :فعَل et فَعل pour la prise en considération de toutes les formes fléchies possibles. Nous donnons dans la table 6.18 un exemple du verbe سمع (entendre).

Pro. Fr	pro. Dia	verb Dia.	lexème Dia.	pro. MSA	verbe MSA	lexème MSA
Elle	هى	سَمْعَت	سَمْع	هى	سَمِعَت	سَمِع
Elles/Ils	هومَا	سَمْعو		هم	سَمِعوا	
Il	هو	سْمَع	سْمَع	هو	سَمِعَ	
Nous	هنَا	سْمَعنَا		نحن	سَمِعنَا	

Tableau 6.18 : Exemple de l'éclatement d'un lexème arabe standard en deux lexèmes dialectaux .

Exploitation du dictionnaire des lexèmes de BAMA

(a) Les Verbes

L'idée principale derrière la création des lexèmes verbaux à partir de ceux de l'arabe standard est l'utilisation de la notion de schème. Par exemple les verbes dialectaux ayant le schème فعَل sont pour la plupart des cas des verbes arabes ayant les schèmes فَعَلَ, فَعُلَ ou فَعِلَ. D'autres verbes dialectaux sont conformes aux même schème qu'en arabe standard comme les verbes ayant le schème فَعَّل.

Nous avons donc extrait de la table des lexèmes tous les verbes ayant les schèmes فَعَل, فَعُل, فَعِل et فَعَّل (relatifs au passé). Ensuite, les verbes ayant les trois premiers schèmes ont été convertis vers le dialecte en modifiant leur signe diacritiques de telle sorte qu'ils soient conformes au schème dialectale فعَل.

Cependant, les verbes ayant comme schème فَعَّل ont été retenus tels qu'ils sont puisque ce schème est conforme au dialecte algérois.

A ce stade, nous avons crée une liste de lexèmes verbaux arabe standard conformes à des schèmes dialectaux. Ces lexèmes ont été analysés, seuls ceux appartenant au dialecte algérois ont été retenus dans la table des lexèmes (des exemples illustratifs sont données dans la table 6.19.

lexèmes	ALG	MSA	Sens
ضرب	ضْرَب	ضَرَب	frapper
شرب	شْرَب	شرِب	boire
بدل	بَدَّل	بَدَّل	changer
كبر	كْبَر	كَبُر	grandir

Tableau 6.19 : Exemples de lexèmes convertis de l'arabe standard vers le dialecte algérois

Nous avons par la suite procédé de la même façon pour les autres schèmes comme تَفَعَّل, تَفَاعَل, فَاعَل,استَفعَل. Une fois que la liste des lexèmes relatifs aux passé a été établie, nous l'avons utilisée pour la construction des lexèmes verbaux relatifs au futur et à l'impératif.

(b) Les noms

Les noms propres contenus dans la table des lexèmes ont été retenus dans la table des lexèmes du dialectes algérois, car les lexèmes de

cette classe couvrent un grand nombre de types de noms propres tels que les pays, les monnaies, les noms personnels,...etc

(c) Les mots outils

Les mots outils de l'arabe standard qui n'existent pas dans le dialecte algérois ont été supprimé aussi de la table des lexèmes nous en citons à titre d'exemple :

— Les pronoms relatifs et personnels du dual

— Les pronoms relatifs et personnels du féminin pluriel

— Les prépositions telles لم, لن

Nous avons aussi introduit dans la tables des lexèmes du dialecte algérois, des lexèmes avec les lettres non arabes ڤ *G*, ف *V*, and پ *P*. Nous avons aussi modifié le code de BAMA pour tenir compte de ces lettres

Par ailleurs, on souligne que à chaque lexème dans la table des suffixes de BAMA correspond une traduction en anglais. Nous avons gardé cette traduction en anglais, ainsi que le lexème en arabe pour chaque lexème dialectal.

Après avoir construit les trois dictionnaires de BAMA (suffixes, affixes, et lexèmes) propres au dialecte Algérien, nous avons mis à jour les trois tables de compatibilité en fonction des relations qu'entretiennent les lexèmes et affixes dialectaux.

6.4.3 Expérimentation

Nous avons expérimenté notre analyseur sur un corpus de test de 600 phrases extraites au hasard de notre corpus du dialecte algérois. Nous avons fait deux sortes d'expérimentation. Au début nous avons utilisé seulement les dictionnaires construit à partir du corpus algérois, ensuite nous avons intégré les

entrées des dictionnaires que nous avons construites du dictionnaire de l'arabe standard, les résultats sont donnés dans la table 6.20.

Résultats	Lexèmes à partir du corpus algérois	Lexèmes issus du corpus algérois et des lexèmes MSA
# mots analysés	703	1115
Pourcentage	43%	69%
# mots non analysés	915	503
Pourcentage	57%	31%

Tableau 6.20 : Résultats de l'analyse morphologique

Nous avons examiné les mots pour lesquels l'analyseur morphologique n'a donné aucune sortie, et nous avons constaté les cas les plus courants qui suivent (des exemples sont données dans le tableau 6.21) :

— Des mots Français qui n'existent pas dans la table des lexèmes tels que تريسيتى (électricité)

— Des mots Français aussi qui existent dans la table des lexèmes mais avec un orthographe différent tels que أنجنيور (ingénieur) et ألنيمرو (numéro) qui sont inclus dans la table des lexèmes mais avec l'orthographe suivant أنجينيور et ألنيميرو respectivement.

— Dans la même optique certains mots ne sont pas analysés à cause de l'utilisation de la voyelle longue ا à la place de ة, comme dans le cas du mot پلاسا, qui existe dans la tables des lexèmes mais écrit avec l'orthographe پلاسا (place).

— Certains mots sont sont écrits avec des lettres omises tel que le mot ألنسا (les femmes) qui apparaît dans la table des stems avec l'orthographe ألنساء. Le même cas est observé pour les mots tels que قالى qui existe dans la table des stems avec l'orthographe قالّى or قال لى (il m'a dit) ou قتلو (je lui ai dit) à la place de قلتلو ou قلت لو.

— Les noms propres apparaissant dans le test n'ont pas été analysés.

Mots non analysés	Lexème correspondant	Sens
أنترنَات	أنترنت	Internet
أمبعد	أومبعد	Après
تريمَاستر	تريمستر	Trimestre
تيليفون	تليفون	Téléphone

Tableau 6.21 : Exemples de mots non analysés

6.5 Conclusion

Nous avons tout au long de ce chapitre présenté les approches que nous avons adoptées pour la création d'outils pour le traitement du dialecte algérois. Le lecteur remarquera sans peine que nous avons tout fait à partir de zéro. Nous avons été contraints d'intervenir manuellement dans plusieurs phases de ce travail, car nous ne trouvions aucun travail déjà initié pour le dialecte Algérien en général. Nous avons tenté d'exploiter au maximum les ressources que nous avons créées au cours de ce travail de thèse. Le corpus algérois que nous avons créé nous a servi pour l'élaboration des outils décrits plus haut. Nous avons utilisé une partie de ce corpus pour développer un système de vocalisation automatique qui, par la suite, nous a servi à l'aide d'un processus itératif semi-automatique de vocaliser tout le corpus algérois. Nous avons abordé le problème de la conversion graphème-phonème qui représente un des défis les plus épineux pour le traitement du dialecte algérois et du dialecte algérien en général. En effet, le phénomène du code-switching implique la prise en considération de l'existence de plusieurs variantes de prononciations incluant des phonèmes arabes et des phonèmes français. Ce phénomène tel que nous l'avons décrit plus haut influe négativement sur les performances de la conversion graphème-phonème. Nous avons utilisé le corpus algérois vocalisé pour entrainer un convertisseur graphème-phonème qui prend en charge les mots français. La dernière partie de ce chapitre a été dédiée à la création d'un analyseur morphologique pour

le dialecte algérois. Nous avons adopté une approche adaptative qui nous a permis de tirer profit de l'analyseur BAMA de l'arabe standard. Nous avons aussi exploité le vocabulaire du corpus algérois pour enrichir les dictionnaires de notre analyseur morphologique.

Chapitre 7

Conclusions et perspectives

Les dialectes arabes s'imposent à travers tout le monde arabe comme la « langue » de communication par excellence non plus pour les conversations quotidiennes comme il a été question depuis longtemps, mais même dans les débats politiques, les émissions de télévision et surtout sur les réseaux sociaux (facebook, youtube, etc.) et dans le domaine de la téléphonie mobile. Ce phénomène concerne aussi l'Algérie, le dialecte gagne du terrain sur les langues standards, l'arabe et le français entre-autre. Les gens s'expriment en dialecte à travers des textos envoyés par SMSs ou au moyen de commentaires postés sur les réseaux sociaux. Ces nouvelles pratiques ont généré de nouveaux besoins en traitement automatique des langues. Il s'agit de prendre en charge ces dialectes dans les applications TAL qui auparavant étaient dédiées à l'arabe standard. C'est dans cette optique que s'est inscrit ce travail. Cette thèse a démarré à partir de zéro, aucune ressource n'était disponible pour le dialecte algérien lors du démarrage de ce travail. Nous avons abordé la problématique relative à la traduction automatique des dialectes dans le cadre des langues peu dotées en ressources. Nous avons été heurtés dés le début au manque criant de corpus, surtout dans le cadre du dialecte algérien. Nous avons consacré un effort à l'élaboration de corpus pour ce dialecte.

Deux corpus dialectaux (algérois et bonois) ont été manuellement créés et traduits vers l'arabe standard. La dimension arabe standard du corpus parallèle tri-lingue nous a permis d'atteindre d'autres dialectes. En effet, ce corpus baptisé PADIC (comme déjà mentionné) compte à l'heure actuelle six dialectes en plus de l'arabe standard. En termes de nombre de phrases parallèles, c'est le corpus le plus important dans le domaine de la traduction des dialectes arabes. PADIC a

été utilisé pour étudier la relation entre les dialectes qui le composent et l'arabe standard. Il a fait l'objet de plusieurs expérimentations intéressantes qui ont mis en relief la distance entre ces dialectes entre eux et la distance avec l'arabe standard. L'étude analytique réalisée a confirmé la proximité entre les dialectes du Maghreb (algérien, tunisien et marocain), ainsi que la proximité entre les deux dialectes palestinien et syrien. Par rapport à l'arabe standard, les dialectes du Moyen-orient sont plus proches que ceux du Maghreb.

Le corpus PADIC a été aussi utilisé pour l'apprentissage d'un classifier dédié à l'identification. Les résultats expérimentaux ont montré que l'arabe standard est mieux identifié que les dialectes. Ce résultat est naturel au regard du fait qu'il s'agit d'une langue standard possédant un solide système d'écriture contrairement aux dialectes qui s'écrivent de façon non-standardisée, ce qui rend leur identification plus difficile. Dans le continuum de ces dialectes, le marocain se distingue et s'identifie mieux par rapport à tous les dialectes. Cela est dû aux traits particuliers de ce dialecte par rapport aux autres dialectes de PADIC.

La suite de nos travaux a porté sur la traduction automatique statistique en exploitant PADIC la seule ressource qui nous est disponible dans le cadre de cette thèse. Nous avons conduit une multitude d'expérimentations qui vont dans le sens de la traduction des dialectes entre eux ainsi que la traduction de ces dialectes de et vers l'arabe standard. Nous avons entraîné des systèmes de traductions statistiques entre toutes les paires de « langues » . Nous avons montré à travers le test de significativité statistique que les techniques de lissage n'ont pas d'impact important sur la performance des systèmes de traductions statistiques des dialectes arabes. Nous avons aussi investigué l'incidence de l'interpolation des modèles de langage sur les scores de la traduction. Les résultats obtenus n'affichent aucune amélioration significative sur les performances des systèmes de traduction, certains scores ont même baissé suite à cette interpolation.

Tout au long de ce travail, un focus particulier a été mis sur le dialecte algérois. Nous avons commencé par développer des ressources pour ce dialecte dans une perspective de les généraliser sur les autres dialectes arabes algériens. Notre but était de construire des ressources rapidement en essayant d'adapter celles dédiées à l'arabe standard d'une part, et d'autre part d'adopter une approche machine learning pour exploiter les ressources que nous avons créées manuellement. Pour ce faire, nous avons développé un système de vocalisation automatique de textes en dialecte algérois. Les résultats obtenus sont encourageants au regard des ressources disponibles. Le système de vocalisation nous a permis de procéder à la vocalisation du corpus alégrois.

Grâce au corpus diacritisé ci-dessus, nous avons réalisé un convertisseur graphème-phonème dédié au dialecte algérois. L'approche adoptée a permis de prendre en charge les variantes de prononciation qui caractérisent le dialecte algérois à savoir l'amalgame entre phonèmes arabes et français.

Enfin, nous avons réalisé un analyseur morphologique pour le dialecte algérois grâce à l'adaptation de l'analyseur BAMA de l'arabe standard. Les dictionnaires de BAMA ont été redéfinis pour prendre en compte les phénomènes morphologiques dialectaux. On notera aussi qu'une partie de ces dictionnaires a été construite à partir du corpus algérois de PADIC.

La continuité de nos travaux portera sur l'enrichissement du corpus PADIC de façon à introduire d'autres dialectes et d'augmenter automatiquement sa taille en explorant les pistes de l'adaptation des modèles de langage et de traduction ainsi que le paraphrasage. Nous comptons aussi y introduire d'autres langues standards telles que le français et l'anglais. Par ailleurs, nous pensons aussi que les corpus comparables peuvent constituer une approche intéressante pour l'enrichissement de PADIC. Une attention particulière sera accordée aux outils que nous avons développés pour le dialecte algérois. Nous comptons étendre leur couverture pour

prendre en charge les traits du dialecte algérien de toutes les régions de l'Algérie. Un axe important auquel nous nous intéressons dans un futur proche est le passage au traitement de la parole relatif au dialecte algérien dans le but de la traduction automatique.

Bibliographie

[Abdul-Mageed and Diab, 2014] Abdul-Mageed, M. and Diab, M. T. (2014). Sana : A large scale multi-genre, multi-dialect lexicon for arabic subjectivity and sentiment analysis. In *LREC*, pages 1162–1169.

[Afify et al., 2006] Afify, M., Sarikaya, R., Kuo, H.-K. J., Besacier, L., and Gao, Y. (2006). On the use of morphological analysis for dialectal arabic speech recognition. In *INTERSPEECH*. Citeseer.

[Al-Gaphari and Al-Yadoumi, 2012] Al-Gaphari, G. and Al-Yadoumi, M. (2012). A method to convert sana'ani accent to modern standard arabic. *International Journal of Information Science and Management (IJISM)*, 8(1) :39–49.

[Al-Mannai et al., 2014] Al-Mannai, K., Sajjad, H., Khader, A., Al Obaidli, F., Nakov, P., and Vogel, S. (2014). Unsupervised word segmentation improves dialectal arabic to english machine translation. *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Langauge Processing ANLP 2014*, pages 207–216.

[Al-Onaizan et al., 1999] Al-Onaizan, Y., Curin, J., Jahr, M., Knight, K., Lafferty, J., Melamed, I., Och, F., Purdy, D., Smith, N., and Yarowsky, D. (1999). Statistical machine translation. In *Final Report, JHU Workshop*.

[Almeman and Lee, 2012] Almeman, K. and Lee, M. (2012). Towards developing a multi-dialect morphological analyser for arabic. In *4th International Conference on Arabic Language Processing*, pages 19–25.

[Altantawy et al., 2011] Altantawy, M., Habash, N., and Rambow, O. (2011). Fast yet rich morphological analysis. In *Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing*, pages 116–124. Association for Computational Linguistics.

[Altintas and Cicekli, 2002] Altintas, K. and Cicekli, I. (2002). A machine translation system between a pair of closely related languages. In *Proceedings of the 17th International Symposium on Computer and Information Sciences (ISCIS 2002)*, pages 192–196.

[Amer et al., 2011] Amer, F. H., Adaileh, B. A., and Rakhieh, B. A. (2011). Arabic diglossia : A phonological study. *Argumentum, Debreceni Egyetemi Kiado*, 7 :19–36.

[Ameur and Jamoussi, 2013] Ameur, H. and Jamoussi, S. (2013). Dynamic Construction of Dictionaries for Sentiment Classification. In *SENTIRE Workshop, IEEE International Conference on Data Mining ICDM'2013*, pages 896–903.

[Aminian et al., 2014] Aminian, M., Ghoneim, M., and Diab, M. (2014). Handling oov words in dialectal arabic to english machine translation. *LT4CloseLang*, pages 99–108.

[Aransa, 2015] Aransa, W. (2015). *Statistical Machine Translation of the Arabic Dialect*. PhD thesis, University of Maine, doctoral school STIM.

[Bakr et al., 2008] Bakr, H. A., Shaalan, K., and Ziedan, I. (2008). A hybrid approach for converting written egyptian colloquial dialect into diacritized arabic. In *The 6th International Conference on Informatics and Systems, INFOS2008. Cairo University*.

[Banerjee and Lavie, 2005] Banerjee, S. and Lavie, A. (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 and/or summarization*, volume 29, pages 65–72.

[Berment, 2004] Berment, V. (2004). *Méthodes pour informatiser les langues et les groupes de langues «peu dotées»*. PhD thesis, Université Joseph-Fourier-GrenobleI.

[Besacier et al., 2014] Besacier, L., Barnard, E., Karpov, A., and Schultz, T. (2014). Automatic speech recognition for under-resourced languages : A survey. *Speech Communication*, 56 :85–100.

[Bhattacharyya, 1943] Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.*, 35 :99–109.

[Boucherit, 2002] Boucherit, A. (2002). *L'Arabe parlé à Alger*. ANEP Edition.

[Boudlal et al., 2010] Boudlal, A., Lakhouaja, A., Mazroui, A., Meziane, A., Bebah, M. O. A. o., and Shoul, M. (2010). Alkhalil morpho sys1 : A morphosyntactic analysis system for arabic texts. In *International Arab Conference on Information Technology, ACIT 2010*.

[Boudlal et al., 2011] Boudlal, A., Lakhouaja, A., Mazroui, A., Meziane, A., Bebah, M. O. A. O., and Shoul, M. (2011). Alkhalil morpho sys : A morphosyntactic analysis system for arabic texts. In *Proceedings of 7th International Computing Conference in Arab ACIT*.

[Brown et al., 1990] Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Lafferty, J., Mercer, R., and Roossin, P. (1990). A statistical approach to machine translation. *Computational Linguistics*, 16 :79–85.

[Brown et al., 1988] Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Mercer, R., and Roossin, P. (1988). A statistical approach to language translation. *Proceedings of the 12th Conference on Computational Linguistics - Volume 1, COLING '88*, page 71–76.

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

[Brown, 1996] Brown, R. D. (1996). Example-based machine translation in the pangloss system. In *Proceedings of the 16th conference on Computational linguistics-Volume 1*, pages 169–174. Association for Computational Linguistics.

[Buchmann, 1984] Buchmann, B. (1984). Early history of machine translation. In *ISSC Tutorial on machine translation (MT)*.

[Buckwalter, 2002] Buckwalter, T. (2002). Buckwalter arabic morphological analyzer version 1.0. *Linguistic Data Consortium LDC2002L49*.

[Chandioux, 1976] Chandioux, J. (1976). meteo : A system translation of public weather forecasts. In *FBIS Seminar of Machine Translation, vol.1*.

[Chen and Goodman, 1999] Chen, S. F. and Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, 13 :359–394.

[Chris et al., 2010] Chris, D., Jonathan, W., Hendra, S., Lopez, A., Ferhan, T., Vladimir, E., Juri, G., Phil, B., and Philip, R. (2010). cdec : A decoder, alignment, and learning framework for finite-state and context-free translation models. In *In Proceedings of the ACL 2010 System Demonstrations*, page 7–12.

[Cieslak and Chawla, 2009] Cieslak, D. A. and Chawla, N. V. (2009). A Framework for Monitoring Classifiers' Performance : When and Why Failure Occurs ? *Knowledge and Information Systems*, pages 83–109.

[Condon et al., 2010] Condon, S., Parvaz, D., Aberdeen, J., Doran, C., Freeman, A., and Awad, M. (2010). Evaluation of machine translation errors in english and iraqi arabic. Technical report, DTIC Document.

[Condon et al., 2008] Condon, S. L., Phillips, J., Doran, C., Aberdeen, J. S., Parvaz, D., Oshika, B. T., Sanders, G. A., and Schlenoff, C. (2008). Applying automated metrics to speech translation dialogs. In *LREC*.

[Costa-Jussa and Fonollosa, 2015] Costa-Jussa, M. R. and Fonollosa, J. A. (2015). Latest trends in hybrid machine translation and its applications. *Computer Speech & Language*, 32(1) :3–10.

[Cotterell and Callison-Burch, 2014] Cotterell, R. and Callison-Burch, C. (2014). A multi-dialect, multi-genre corpus of informal written arabic. In *LREC*, pages 241–245.

[Darwish et al., 2014] Darwish, K., Sajjad, H., and Mubarak, H. (2014). Verifiably effective arabic dialect identification. In *EMNLP*, pages 1465–1468.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). likelihood from in-complete data via the em algorithm. *Journal of the Royal Statistical Society : Series B*, 39 :1–38.

[Doddington, 2002] Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Dumais et al., 1998] Dumais, S., Platt, J., Heckerman, D., and Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. In *Proceedings of the seventh international conference on Information and knowledge management*, pages 148–155.

[Durrani et al., 2014] Durrani, N., Al-Onaizan, Y., and Ittycheriah, A. (2014). Improving egyptian-to-english smt by mapping egyptian into msa. In *Computational Linguistics and Intelligent Text Processing*, pages 271–282. Springer.

[Eisele et al., 2008] Eisele, A., Federmann, C., Uszkoreit, H., Saint-Amand, H., Kay, M., Jellinghaus, M., Hunsicker, S., Herrmann, T., and Chen, Y. (2008). Hybrid machine translation architectures within and beyond the euromatrix project. In *Proceedings of the 12th annual conference of the European Association for Machine Translation (EAMT 2008)*, pages 27–34.

[El-Sadany and Hashish, 1989] El-Sadany, T. and Hashish, M. (1989). An arabic morphological system. *IBM Systems Journal*, 28(4) :600–612.

[Elfardy et al., 2014] Elfardy, H., Al-Badrashiny, M., and Diab, M. (2014). Aida : Identifying code switching in informal arabic text. *EMNLP*, page 94.

[Elfardy and Diab, 2013] Elfardy, H. and Diab, M. T. (2013). Sentence level dialect identification in arabic. In *ACL (2)*, pages 456–461.

[Elmahdy et al.,] Elmahdy, M., Hasegawa-Johnson, M., and Mustafawi, E. Development of a tv broadcasts speech recognition system for qatari arabic. In *9th edition of the Language Resources and Evaluation Conference, Reyk- javik, Iceland, May.*

[Elshafei et al., 2006] Elshafei, M., Al-Muhtaseb, H., and Alghamdi, M. (2006). Statistical methods for automatic diacritization of arabic text. In *Proceedings of the Saudi 18th National Computer Conference (NCC18)*, pages 301–306.

[Emam and Fischer, 2004] Emam, O. and Fischer, V. (2004). Hierarchical approach for the statistical vowelization of arabic text. Technical report. *IBM patent filed, DE9-2004-0006, US patent application US2005/0192809 A1.*

[Ennaji, 2005] Ennaji, M. (2005). *Multilingualism, cultural identity, and education in Morocco.* Springer Science & Business Media.

[España Bonet et al., 2011] España Bonet, C., Màrquez Villodre, L., Labaka, G., Díaz de Ilarraza Sánchez, A., and Sarasola Gabiola, K. (2011). Hybrid machine translation guided by a rule-based system. In *Machine translation summit XIII : proceedings of the 13th machine translation summit, September 19-23, 2011, Xiamen, China*, pages 554–561.

[Farag and Nurnberger, 2008] Farag, A. and Nurnberger, A. (2008). Arabic/English Word Translation Disambiguation Using Parallel Corpora and Matching Schemes. In *12th EAMT conference*, pages 6–11.

[Ferguson, 1957] Ferguson, C. A. (1957). Two problems in arabic phonology. *Word*, 13 :460–478.

[Ferguson, 1959] Ferguson, C. A. (1959). Diglossia. *Word*, 15 :325–340.

[Gal, 2002] Gal, Y. (2002). An hmm approach to vowel restoration in arabic and hebrew. In *Proceedings of the ACL-02 workshop on Computational approaches to semitic languages*, pages 1–7. Association for Computational Linguistics.

[Gale et al., 1992] Gale, W. A., Church, K. W., and Yarowsky, D. (1992). Using bilingual materials to develop word sense disambiguation methods. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 101–112.

[Gilleland, 2009] Gilleland, M. (2009). Levenshtein distance, in three flavors.

[González-Castro et al., 2013] González-Castro, V., Alaiz-Rodríguez, R., and Alegre, E. (2013). Class Distribution Estimation Based on the Hellinger Distance. *Information Sciences*, pages 146—164.

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

[Groves and Way, 2005] Groves, D. and Way, A. (2005). Hybrid data-driven models of machine translation. *Machine Translation*, 19(3-4) :301–323.

[Habash et al., 2012] Habash, N., Eskander, R., and Hawwari, A. (2012). Morphological analyzer for egyptian arabic. In *Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology SIGMORPHON*, pages 1–9. Association for Computational Linguistics.

[Habash and Rambow, 2005] Habash, N. and Rambow, O. (2005). Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 573–580. Association for Computational Linguistics.

[Habash and Rambow, 2006a] Habash, N. and Rambow, O. (2006a). Magead : a morphological analyzer and generator for the arabic dialects. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 681–688. Association for Computational Linguistics.

[Habash and Rambow, 2006b] Habash, N. and Rambow, O. (2006b). Magead : A morphological analyzer and generator for the arabic dialects. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 681–688. Association for Computational Linguistics.

[Habash et al., 2008] Habash, N., Rambow, O., Diab, M., and Kanjawi-Faraj, R. (2008). Guidelines for annotation of arabic dialectness. In *Proceedings of the LREC Workshop on HLT & NLP within the Arabic world*, pages 49–53.

[Haddow et al., 2013] Haddow, B., Huerta, A. H., Neubarth, F., and Trost, H. (2013). Corpus development for machine translation between standard and dialectal varieties. *Adaptation of Language Resources and Tools for Closely Related Languages and Language Variants*, page 7.

[Hajič et al., 2000] Hajič, J., Hric, J., and Kuboň, V. (2000). Machine translation of very close languages. In *Proceedings of the sixth conference on Applied natural language processing*, pages 7–12. Association for Computational Linguistics.

[Hamdi et al., 2013] Hamdi, A., Boujelbane, R., Habash, N., and Nasr, A. (2013). The effects of factorizing root and pattern mapping in bidirectional tunisian-standard arabic machine translation. In *MT Summit 2013*.

[Heafield, 2011] Heafield, K. (2011). Kenlm : Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197. Association for Computational Linguistics.

[Hetzron, 1997] Hetzron, R. (1997). *The Semitic Languages*. Routledge language family descriptions. Routledge.

[Jeblee et al., 2014] Jeblee, S., Feely, W., Bouamor, H., Lavie, A., Habash, N., and Oflazer, K. (2014). Domain and dialect adaptation for machine translation into egyptian arabic. *ANLP 2014*, pages 196–207.

[Katz, 1987] Katz, S. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE transactions on acoustics, speech, and signal processing*, 35(3) :400–401.

[Kim et al., 2006] Kim, S.-B., Han, K.-S., Rim, H.-C., and Myaeng, S. H. (2006). Some effective techniques for naive bayes text classification. *IEEE transactions on knowledge and data engineering*, 18(11) :1457–1466.

[King, 1981] King, M. (1981). Eurotra–a european system for machine translation. *Lebende Sprachen*, 26(1) :12–14.

[Kirchhoff et al., 2003] Kirchhoff, K., Bilmes, J., Das, S., Duta, N., Egan, M., Ji, G., He, F., Henderson, J., Liu, D., Noamany, M., Schone, P., Schwartz, R., and Vergyri, D. (2003). Novel approaches to arabic speech recognition : Report from the 2002 johns-hopkins workshop. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Hong Kong.

[Kneser and Ney, 1995] Kneser, R. and Ney, H. (1995). Improved backing-off for m-gram language modeling. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 181–184. IEEE.

[Koehn et al., 2007] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses : Open Source Toolkit for Statistical Machine Translation. *Proceedings of the Annual Meeting of the Association for Computational Linguistics, demonstation session*, pages 177–180.

[Krauwer, 2003] Krauwer, S. (2003). The basic language resource kit (blark) as the first milestone for the language resources roadmap. *Proceedings of SPECOM 2003*, pages 8–15.

[Landsbergen, 1989] Landsbergen, J. (1989). The rosetta project'. *Second MT Summit, Munich*, pages 82–87.

[Leusch et al., 2003] Leusch, G., Ueffing, N., Ney, H., et al. (2003). A novel string-to-string distance measure with applications to machine translation evaluation. In *Proceedings of MT Summit IX*, pages 240–247.

[Mahfoudhi, 2002] Mahfoudhi, A. (2002). Agreement lost, agreement regained : A minimalist account of word order and agreement variation in arabic. *California Linguistic Notes*, 27(2) :1–28.

[Markus et al., 2014] Markus, F., Matthias, H., and Hermann, N. J. (2014). Jane : Open source machine translation system combination. In *In Proc. of the Conf. of the European Chap- ter of the Assoc. for Computational Linguistics (EACL)*, page 29–32.

[Meftouh et al., 2015] Meftouh, K., Harrat, S., Jamoussi, S., Abbas, M., and Smaili, K. (2015). Machine translation experiments on padic :
a parallel arabic dialect corpus. In *Proceedings PaCLiC 29th Asia Conference on Language, Information and Computation*, pages 26–34.

[Mitamura et al., 1991] Mitamura, T., Nyberg, E. H., and Carbonell, J. G. (1991). An efficient interlingua translation system for multi-lingual document production. In *Machine Translation Summit III*, page 2–4.

[Mohamed et al., 2012] Mohamed, E., Mohit, B., and Oflazer, K. (2012). Transforming standard arabic to colloquial arabic. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics : Short Papers - Volume 2*, ACL '12, pages 176–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Nagao, 1984] Nagao, M. (1984). A framework of a mechanical translation between japanese and english by analogy principle. *Artificial and Human Intelligence, A. Elithorn and R. Banerji (eds). North-Holland*.

[Nakov and Ng, 2012] Nakov, P. and Ng, H. T. (2012). Improving statistical machine translation for a resource-poor language using related resource-rich languages. *Journal of Artificial Intelligence Research (4)*, pages 179–222.

[Nelken and Shieber, 2005] Nelken, R. and Shieber, S. M. (2005). Arabic diacritization using weighted finite-state transducers. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, pages 79–86. Association for Computational Linguistics.

[Nießen et al., 2000] Nießen, S., Och, F. J., Leusch, G., Ney, H., et al. (2000). An evaluation tool for machine translation : Fast evaluation for mt research. In *LREC*.

[Och, 2000] Och, F. (2000). Giza++ tools for training statistical translation models.

[Och and Ney, 2003] Och, F. J. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics, volume 29, number 1*, pages 19–51.

[Papineni and al., 2001] Papineni, K. and al. (2001). Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual of the Association for Computational linguistics*, pages 311–318, Philadelphia, USA.

[Pasha et al., 2014] Pasha, A., Al-Badrashiny, M., Diab, M., El Kholy, A., Eskander, R., Habash, N., Pooleery, M., Rambow, O., and Roth, R. M. (2014). Madamira : A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland*.

[Pedersen, 2000] Pedersen, T. (2000). A Simple Approach to Building Ensembles of Naive Bayesian Classifiers for Word Sense Disambiguation. In *Proceedings of 1st Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 63–69.

[Pop, 2006] Pop, I. (2006). An Approach of the Naive Bayes Classifier for the Dcument Classification. *General Mathematics, Vol14, No 4*, pages 135–138.

[Rosetta, 1994] Rosetta, M. (1994). Rosetta : Compositional translation.

[Sadat, 2015] Sadat, F. (2015). Multi-dialect machine translation (mudmat). In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, page 226, Antalya, Turkey.

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

[Sahami et al., 1998] Sahami, M., Dumais, S., Heckerman, D., and Horvitz, E. (1998). A bayesian approach to filtering junk e-mail. In *Learning for Text Categorization : Papers from the 1998 workshop*, volume 62, pages 98–105.

[Sajjad et al., 2013] Sajjad, H., Darwish, K., and Belinkov, Y. (2013). Translating dialectal arabic to english. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics ACL (2),Sofia, Bulgaria*, pages 1–6.

[Salloum and Habash, 2011] Salloum, W. and Habash, N. (2011). Dialectal to standard arabic paraphrasing to improve arabic-english statistical machine translation. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, pages 10–21. Association for Computational Linguistics.

[Salloum and Habash, 2012] Salloum, W. and Habash, N. (2012). Elissa : A dialectal to standard arabic machine translation system. In *COLING (Demos)*, pages 385–392.

[Salloum and Habash, 2013] Salloum, W. and Habash, N. (2013). Dialectal arabic to english machine translation : Pivoting through modern standard arabic. In *HLT-NAACL*, pages 348–358.

[Sawaf, 2010] Sawaf, H. (2010). Arabic dialect handling in hybrid machine translation. In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA), Denver, Colorado.*

[Scannell, 2006] Scannell, K. P. (2006). Machine translation for closely related language pairs. In *Proceedings of the Workshop Strategies for developing machine translation for minority languages*, pages 103–109. Citeseer.

[Schlippe et al., 2008] Schlippe, T., Nguyen, T., and Vogel, S. (2008). Diacritization as a machine translation problem and as a sequence labeling problem. In *Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 21–25.

[Schwenk et al., 2009] Schwenk, H., Abdul-Rauf, S., Barrault, L., and Senellart, J. (2009). Smt and spe machine translation systems for wmt'09. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 130–134. Association for Computational Linguistics.

[Servan and Schwenk, 2011] Servan, C. and Schwenk, H. (2011). Optimising multiple metrics with mert. *The Prague Bulletin of Mathematical Linguistics*, 96 :109–117.

[Shirai et al., 1997] Shirai, S., Bond, F., and Takahashi, Y. (1997). A hybrid rule and example-based method for machine translation. In *Proceedings of NLPRS*, volume 97, pages 49–54.

[Shoufan and Al-Ameri, 2015] Shoufan, A. and Al-Ameri, S. (2015). Natural language processing for dialectical arabic : A survey. In *ANLP Workshop 2015*, page 36.

[Shoukry and Rafea, 2012] Shoukry, A. and Rafea, A. (2012). Preprocessing egyptian dialect tweets for sentiment mining. In *The Fourth Workshop on Computational Approaches to Arabic Script-based Languages*, page 47.

[Siivola et al., 2007] Siivola, V., Creutz, M., and Kurimo, M. (2007). Morfessor and varikn machine learning tools for speech and language technology. In *INTERSPEECH*, pages 1549–1552.

[Snover et al., 2006] Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, pages 223–231.

[Souag, 2006] Souag, M. L. (2006). *Explorations in the Syntactic Cartography of Algerian Arabic*. PhD thesis, School of Oriental and African Studies (University of London.

[Stolcke, 2002] Stolcke, A. (2002). Srilm – an Extensible Language Modeling Toolkit. In *ICSLP*, pages 901–904, Denver, USA.

[Tachicart and Bouzoubaa, 2014] Tachicart, R. and Bouzoubaa, K. (2014). A hybrid approach to translate moroccan arabic dialect. In *Intelligent Systems : Theories and Applications (SITA-14), 2014 9th International Conference on*, pages 1–5. IEEE.

[Toma, 1976a] Toma, P. (1976a). An operational machine translation system. In *Berislin, R.W.(ed)*, pages 247–260.

[Toma, 1976b] Toma, P. (1976b). The systran system. In *FBIS Seminar of Machine Translation, vol. 2*.

[Toma, 1978] Toma, P. (1978). systran as a multilingual machine translation system. In *Commission of the European Communities (1978), vol. 1*, pages 569–581.

[Torra and Carlson, 2013] Torra, V. and Carlson, M. (2013). On the Hellinger Distance for Measuring Information Loss in Microdata. In *Joint UNECE/Eurostat work session on statistical data confidentiality*.

[Vauquois, 1968] Vauquois, B. (1968). A survey of formal grammars and algorithms for recognition and transformation in mechanical translation. In *IFIP Congress (2)*, page 1114–1122.

[Vauquois, 1975] Vauquois, B. (1975). La traduction automatique à grenoble. In *Paris : Dunod*.

[Vauquois, 1979] Vauquois, B. (1979). Aspects of mechanical translation in 1979. In *Conference for Japan IBM Scientific Program (GETA Report)*.

[Villard, 1989] Villard, M. (1989). Traduction automatique et recherche cognitive. *Histoire Épistémologie Langage*, 11(1) :55–84.

[Warren, 1955] Warren, W. (1955). Translation. In *Machine translation of languages, N. William & A. D. Booth (Eds.), Reprinted from Mimeographed , 1949, 12 pp.*, pages 15–23.

[Watson, 2007] Watson, J. C. (2007). *Phonology and Morphology of Arabic*. Phonology of the World's Languages. Oxford University Press, New York.

[Way and Gough, 2003] Way, A. and Gough, N. (2003). webmt : developing and validating an example-based machine translation system using the world wide web. *Computational Linguistics*, 29(3) :421–457.

[Witkam, 1988] Witkam, T. (1988). Dlt : an industrial r & d project for multilingual mt. In *Proceedings of the 12th conference on Computational linguistics-Volume 2*, pages 756–759. Association for Computational Linguistics.

[Witten and Bell, 1991] Witten, I. H. and Bell, T. C. (1991). The zero frequency problem : estimating the probabilities of novel events in adaptive text compression. *IEEE Trans. Inform. Theory*, 37 :1085–1094.

[Yang et al.,] Yang, J., Enoue, S., and Senellart, J. Systran chinese-english and english-chinese hybrid machine translation systems for cwmt2011. In *Proceedings of the 7th China Workshop on Machine Translation*.

[Zaidan and Callison-Burch, 2012] Zaidan, O. F. and Callison-Burch, C. (2012). Arabic dialect identification. *Computational Linguistics*, 1(1).

[Zbib et al., 2012] Zbib, R., Malchiodi, E., Devlin, J., Stallard, D., Matsoukas, S., Schwartz, R., Makhoul, J., Zaidan, O. F., and Callison-Burch, C. (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*, pages 49–59. Association for Computational Linguistics.

[Zhang, 1998] Zhang, X. (1998). Dialect mt : a case study between cantonese and mandarin. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pages 1460–1464. Association for Computational Linguistics.

[Zitouni et al., 2006] Zitouni, I., Sorensen, J. S., and Sarikaya, R. (2006). Maximum entropy based restoration of arabic diacritics. In *Proceedings of*

the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, pages 577–584. Association for Computational Linguistics.

Annexe A

Les règles de conversion graphème phonème pour le dialecte Algérois

Tableau A.1 : Les règles de conversion graphème phonème pour le dialecte Algérois

#	Titre de la règle	Règle
1	Règles relatives au ذ, ظ et ث	$\{C,V\}$+ذ +$\{C,V\} \Longrightarrow /d/$
		$\{C,V\}$+ظ +$\{C,V\} \Longrightarrow /d'/$
		$\{C,V\}$+ث +$\{C,V\} \Longrightarrow /T/$
2	Règles des lettres non arabe	$\{C,V\}$+ڤ +$\{C,V\} \Longrightarrow /g/$
		$\{C,V\}$+ف +$\{C,V\} \Longrightarrow /v/$
		$\{C,V\}$+پ +$\{C,V\} \Longrightarrow /p/$
3	Règle de l'article de définition اَل	$\{LC\}$+اَل+$\{BL,BS\} \Longrightarrow /l/ + /LC/$
		$\{SC\}$+اَل+$\{BL-BS\} \Longrightarrow$/ ?a/+/$SC$/ + /$SC$/
4	Règle de la marque du cas	$\{BL,ES\} + C + \{C,V\} \Longrightarrow /C/$
5	Règles des voyelles longues	{C+ ا}+ ـَ+$\{C\} \Longrightarrow /a:/$
		{C+و}+ـُ+$\{C\} \Longrightarrow /u:/$
		{C+ى}+ـِ+$\{C\} \Longrightarrow /i:/$
6	Règle de la glotte	$\{C,V\}$+أَ+$\{BS,BL\} \Longrightarrow /?/$
		$\{BL,ES\}$+ ء +{ اَ } $\Longrightarrow /Null/$
7	Règle du Alif Maqsura ى	$\{BL,ES\}$+ى+{ـَ+$C\} \Longrightarrow /a/$
8	Règle du Alif Madda آ	$\{C\}$+آ+$\{C\} \Longrightarrow /?a:/$
9	Règle des mots se terminant par ة	$\{BL,ES\}$+ة+$\{C,V\} \Longrightarrow /Null/$
10	Règle des mots se terminant par ه	$\{BL,ES\}$+ه+{ـُ} $\Longrightarrow /Null/$
11	Règle des mots contenant la sequence ب,ن	{ ب}+ن+$\{C,V\} \Longrightarrow /m/$
12	Règle de la gémination	{V}+ ّ +$\{C\} \Longrightarrow /CC/$

Annexe B

Obtention de la licence LDC ARB TreeBank

Pour des besoins de comparaison nous avons eu besoin d'un corpus arabe standard diacritisé à l'instar de Arabic TreenBank. Ce corpus est une ressource payante qui requiert soit un abonnement auprès du LDC (pour institutions membres), ou bien des frais d'exploitation pour les institutions non-membres du LDC (3500 $). Cependant, Le LDC possède un programme Nommé LDC Scholarship [1] dans lequel il attribue des copies libres de ses corpus pour les travaux de recherches solides n'ayant pas de moyens de financement. Ce programme est ouvert aux chercheurs de tous les pays en fournissant un dossier de candidature justifiant l'utilisation de la ressource LDC et l'intérêt que procurera le travail en question pour le LDC. Parmi les soumissions de la session Spring program 2013, seules trois d'entre elles ont été acceptées après étude minutieuse des dossiers (il existe deux sessions par an). Notre travail s'est vu attribué la licence d'utilisation sans frais du corpus Arabic treebank [2] (LDC2004T11, Arabic Treebank : Part 3 v 1.0). [3]

1. https://www.ldc.upenn.edu/language-resources/data/data-scholarships
2. https://www.ldc.upenn.edu/communications/newsletter/february-2013-newsletter
3. https://catalog.ldc.upenn.edu/LDC2004T11

LINGUISTIC DATA CONSORTIUM
3600 Market St., Suite 810, University of Pennsylvania, Philadelphia, PA 19104-2653 USA • Tel: (215) 898-0464 • Fax: (215) 573-2175 • email: ldc@ldc.upenn.edu

LDC User Agreement for Non-Members.

This User Agreement is provided by the Linguistic Data Consortium as a condition of accepting the databases named or described herein.

In the remainder of this document the term User refers to Salima HARRAT (Individual name) of Ecole Supérieure d'Informatique ESI (Affiliation),

and the term User's research group refers to: Ecole Supérieure d'Informatique ESI (Specific department or area within company, if appropriate, or University, Institute or Company name).

This Agreement describes the terms between User/User's Research Group and Linguistic Data Consortium (LDC), in which User will receive material, as specified below, from LDC. The terms of this Agreement supercede any previous Membership Agreement for the Corpora received in Exhibit A.

Under this agreement User will receive one or more CD-ROM discs, DVDs, electronic files or other media as appropriate, named below under "CORPORA RECEIVED" (Exhibit A), containing speech, video and/or text data. User agrees to use the material received under this agreement only for non-commercial linguistic education and research purposes. In the event that User's use of LDC Corpora results in the development of a commercial product, User must join the LDC as a Commercial Member and pay all applicable fees prior to release of said commercial product. Unless explicitly permitted herein, User shall have no right to copy, redistribute, transmit, publish or otherwise use the LDC Databases for any other purpose and User further agrees not to disclose, copy, or re-distribute the material to others outside of User's research group.

USER ACKNOWLEDGES AND AGREES THAT "CORPORA RECEIVED" ARE PROVIDED ON AN "AS-IS" BASIS AND THAT LDC, ITS HOST INSTITUTION THE UNIVERSITY OF PENNSYLVANIA, AND ITS DATA PROVIDERS AND CORPUS AUTHORS MAKE NO REPRESENTATIONS OR WARRANTIES OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, OR CONFORMITY WITH WHATEVER DOCUMENTATION IS PROVIDED. IN NO EVENT SHALL LDC, ITS HOST INSTITUTION, DATA PROVIDORS OR CORPUS AUTHORS BE LIABLE FOR SPECIAL, DIRECT, INDIRECT, CONSEQUENTIAL, PUNITIVE, INCID AGES, LOSSES, COSTS, CHARGES, CLAIMS, DEMANDS, FEES NATURE OR KIND ARISING IN ANY WAY FROM THE FURNISHING OF CORPORA RECEIVED.

For ESI (User Affiliation)

Signature  Date 03 Mar 2013

Name Salima HARRAT Title PHD STUDENT

For LDC

Chris Cieri, Executive Director

EXHIBIT A

CORPORA RECEIVED (Include Title and Catalog Number)

1 LDC2004T11 / Arabic Treebank: Part 3 v 1.0

2 _____

3 _____

4 _____

5 _____

Annexe C

Liste des publications de l'auteur

1. **Interspeech 2013**

   Diacritics restoration for Arabic dialect texts

   14th Annual Conference of the International Speech communication Association Interspeech 2013, Lyon - France, du 25 au 29 Aout 2013.

2. **SLTU 2014**

   Grapheme To Phoneme Conversion An Arabic Dialect Case

   4th International Workshop on Spoken Language Technologies for Under-resourced Languages SLTU 14, Saint Petersburg, Russie du 14 au 16 Mai 2014

3. **Interspeech 2014**

   Building resources for Algerian Arabic dialects

   15th Annual Conference of the International Speech communication Association Interspeech 2014, Singapour, du 14 au 18 Septembre 2014.

4. **Cicling 2015**

   Cross-Dialectal Arabic Processing

   Gelbukh, A., editions : Computational Linguistics and Intelligent Text Processing. Volume 9041 of Lecture Notes in Computer Science. Springer International Publishing (2015) pp 620–632, doi="10.1007/978-3-319-18111-0_47", http://dx.doi.org/10.1007/978-3-319-18111-0_47

5. **PACLIC 29**

   Machine Translation Experiments on PADIC : A Parallel Arabic DIalect Corpus The 29th Pacific Asia Conference on Language, Information and Computation Shanghai, du 30 Octobre au 1er Novembre 2015

6. **IJACSA 2016** An Algerian dialect : Study and Resources,

   International Journal of Advanced Computer Science and Applications(ijacsa), 7(3), 2016. http://dx.doi.org/10.14569/IJACSA.2016.070353

7. **Cicling 2017**

   Creating Arabic dialect corpora : pitfalls to avoid 18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING)

8. **IPM 2018/Elsevier** Machine translation for Arabic dialects,

   Information processing and management. http://doi.org/10.1016/j.ipm.2017.08.003 Article mis en ligne le 31.08.2017

9. **ISGA 2018**

   Maghrebi Arabic dialect processing : an overview

   Journal of International Science and General Applications, 2018, vol. 1, selected papers from International Conference on Natural Language, Signal and Speech Processing, (ICNLSSP Déc, 2017, Casablanca Maroc).

10. **ICAT 2018**

    PADIC : extension and new experiments

    7th International Conference on Advanced Technologies ICAT 2018, Antalya,Turkey.