La construction automatique de ressources multilingues à partir des réseaux sociaux : application aux données dialectales du Maghreb

Karima Abidi

▶ **To cite this version:**

Karima Abidi. La construction automatique de ressources multilingues à partir des réseaux sociaux : application aux données dialectales du Maghreb. Informatique [cs]. Université de Lorraine, 2019. Français. NNT : 2019LORR0274 . tel-02510812

HAL Id: tel-02510812
https://hal.univ-lorraine.fr/tel-02510812

Submitted on 18 Mar 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

AVERTISSEMENT

Ce document est le fruit d'un long travail approuvé par le jury de soutenance et mis à disposition de l'ensemble de la communauté universitaire élargie.

Il est soumis à la propriété intellectuelle de l'auteur. Ceci implique une obligation de citation et de référencement lors de l'utilisation de ce document.

D'autre part, toute contrefaçon, plagiat, reproduction illicite encourt une poursuite pénale.

Contact : ddoc-theses-contact@univ-lorraine.fr

LIENS

Code de la Propriété Intellectuelle. articles L 122. 4
Code de la Propriété Intellectuelle. articles L 335.2- L 335.10
http://www.cfcopies.com/V2/leg/leg_droi.php
http://www.culture.gouv.fr/culture/infos-pratiques/droits/protection.htm

UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

La construction automatique de ressources multilingues à partir des réseaux sociaux : application aux données dialectales du Maghreb

THÈSE

présentée et soutenue publiquement le 20 Décembre 2019

pour l'obtention du

Doctorat de l'Université de Lorraine

(mention informatique)

par

ABIDI Karima

Composition du jury

Rapporteurs :	Pr Yannick Estève : Université Avignon, France Pr Karim Bouzoubaa : Université Mohammed-V, Maroc
Examinateurs :	HDR Anja Habacha : Université de Tunis, Tunsie Pr Denis Jouvet : Université de Lorraine, France Pr Kamel Smaïli : Université de Lorraine, France. MCF David Langlois : Université de Lorraine, France

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Remerciements

Je tiens à remercier mon directeur de thèse, Kamel Smaïli, et mon codirecteur, David Langlois, pour le sujet si passionnant qu'ils m'ont proposé. Un grand merci pour votre temps, votre confiance, votre patience, vos conseils, vos encouragements, votre disponibilité et tant d'autres... Grâce à vous j'ai appris beaucoup de choses notamment la rigueur dans le travail. Je les remercie également pour la patience et l'attention avec laquelle ils ont relu et corrigé mon manuscrit. Pour tout cela, je tiens à leurs témoigner toute ma gratitude.

J'exprime mes remerciements à Monsieur Denis JOUVET, Professeur à l'Université de Lorraine, qui a bien voulu être président de mon jury de thèse.

J'adresse tous mes remerciements à Monsieur Yannick Estève, Professeur à l'Université de d'avignon, ainsi qu'à Monsieur Karim Bouzoubaa, Professeur de l'université du Maroc, pour l'honneur qu'ils m'ont fait en acceptant d'évaluer mon travail et de participer au jury.

Je remercie également Madame Anja Habacha, Professeur à l'Université de Tunis, pour avoir accepté d'être un examinateur de ma thèse.

Enfin, mes remerciements les plus forts vont à mes parents, ma famille et mes proches pour leur amour inconditionnel et leurs encouragements. Cette thèse est dédiée à vous !

Enfin, je remercie tous mes amis et collègues du laboratoire LORIA. Un merci particulier à : Amine et sa femme Amina, Sara, Cherifa, Abir et Nouha.

À la mémoire de mon très chère père
J'espère que tu te reposes en paix là où tu es

Sommaire

Liste des tableaux **viii**

Introduction générale **2**

Chapitre 1 État de l'art

1.1 Introduction . . . 7
1.2 Les corpus . . . 7
1.3 Les corpus monolingues . . . 8
1.4 Les corpus multilingues . . . 8
1.4.1 Les corpus parallèles . . . 8
1.4.2 Les corpus comparables multilingues . . . 10
1.4.3 Évaluation du degré de comparabilité des corpus multilingues . . . 12
1.5 L'analyse de sentiments . . . 18
1.5.1 Méthodes d'attribution d'une orientation sémantique à une entrée lexicale 19
1.5.2 Méthodes et ressources pour une analyse à granularité fine . . . 22
1.6 La représentation des données . . . 23
1.6.1 Représentation discrète . . . 23
1.6.2 Représentations distributionnelles . . . 24
1.7 Conclusion . . . 27

Chapitre 2 Collecte et analyse de corpus pour les dialectes du Maghreb **28**

2.1 Introduction . . . 29
2.2 La langue arabe . . . 29
2.2.1 L'arabe standard moderne . . . 30
2.2.2 L'arabe dialectal . . . 31
2.3 Sélection et collecte des données dialectales . . . 36
2.4 Étude analytique sur les trois corpus maghrébins . . . 37
2.4.1 L'écriture du dialecte maghrébin dans les réseaux sociaux . . . 37
2.4.2 L'utilisation du code-switching dans les dialectes maghrébins . . . 38

2.5 Conclusion 39

Chapitre 3 Mesurer le code-switching **41**

3.1 Introduction 42
3.2 Les approches pour mesurer le degré de code-switching 43
3.2.1 Facteur de complexité de Ghosh 43
3.2.2 CESAR : nouvelle méthode pour mesurer le code-switching 44
3.3 Protocole d'évaluation 46
3.3.1 Étude comparative des deux métriques sur des exemples simples 46
3.3.2 Évaluation de CESAR et Ghosh 47
3.3.3 Expérimentation sur les dialectes maghrébins 49
3.4 Conclusion 50

Chapitre 4 La construction automatique des corpus comparables **52**

4.1 Introduction 53
4.2 L'alignement automatique des tweets multilingues : MSA-anglais 54
4.2.1 Acquisition des données à partir de Twitter 54
4.2.2 Pré-traitements des tweets 55
4.2.3 Alignement des tweets 59
4.2.4 Expérimentations 63
4.3 Méthode de création d'un corpus comparable pour les dialectes maghrébins . . 65
4.3.1 Approche basée sur le dictionnaire 65
4.3.2 Utilisation de la représentation phonétique pour l'appariement de textes 66
4.3.3 Une approche basée sur le *word embedding* 68
4.3.4 Une approche itérative basée sur le *multilingual word embedding* 70
4.3.5 Expérimentations 72
4.4 Conclusion 75

Chapitre 5 La construction automatique de ressources lexicales pour les dialectes maghrébins **77**

5.1 Introduction 78
5.2 Construction d'un lexique des différentes formes d'un même mot 78
5.2.1 La méthode d'extraction 79
5.2.2 Protocole d'évaluation 81
5.3 L'extraction de lexique de sentiments 85
5.3.1 La méthodologie proposée 85
5.3.2 Protocole d'évaluation 87
5.4 Conclusion 90

Chapitre 6 Contribution au projet AMIS **91**
6.1 Introduction 92
6.2 Corpus AMIS 92
6.3 Schéma d'analyse de sentiments 93
6.4 L'identification des vidéos comparables 93
6.4.1 Méthode fondée sur le dictionnaire bilingue 94
6.4.2 Méthode basée sur le *word embedding* 94
6.4.3 Expérimentations 94
6.5 Analyse de sentiments multilingues à granularité fine 95
6.5.1 La théorie *appraisal* 96
6.5.2 Travaux sur l'*appraisal* 97
6.5.3 La construction automatique d'un lexique d'*appraisal* 98
6.5.4 Modèle de prédiction de sentiments à granularité fine 100
6.5.5 Évaluation des vidéos du projet AMIS 102
6.6 Conclusion 104

Conclusion et perspectives **105**

Annexes **109**

Annexe A **109**

Annexe B **110**

Annexe C **112**

Annexe D Liste des publications **113**
D.1 Revue internationale 113
D.2 Conférences internationales 113

Bibliographie **115**

Table des figures

1.1 Exemple de deux textes comparables extraits de Wikipédia. 11
1.2 Les deux architectures de *Word2vec* proposée par [Mikolov et al., 2013b]. 26

2.1 La racine علم agglutinée avec les quatre affixes : l'antéfixe, le préfixe, le suffixe et le postfixe. 31
2.2 La répartition des commentaires en fonction du script dans les trois corpus. . . 37
2.3 Pourcentage de segments utilisés pendant le code-switching dans les trois corpus en fonction du nombre de mots. 39

3.1 La progression de Ghosh et CESAR en fonction de l'existence ou non code-switching. 48

4.1 Typologie de la morphologie des mots en arabe. 58
4.2 Exemple des lemmes obtenus pour le mot arabe mal écrit en utilisant l'approche basée sur *Light Stemming*. 59
4.3 L'encodage phonétique du nom propre arabe جَميلة et ses différentes formes en latin. 63
4.4 Un exemple de deux commentaires comparables. 66
4.5 Les mots qu'on peut faire correspondre avec le dictionnaire et la phonétique. . 67
4.6 Les mots corrélés avec le mot طبيب (médecin) obtenus avec CBOW. 69
4.7 Les mots corrélés avec le mot طبيب (*médecin*) écrits en caractères différents. . 70
4.8 Processus itératif de *mutlilingual word embedding*. 72
4.9 Les listes de mots corrélés avec l'entrée طبيب obtenues lors des quatre premières itérations. 73
4.10 Le taux de mots corrélés avec l'entrée طبيب durant le processus itératif du *Word2vec*. 74

5.1 l'évolution des rappels, de la précision et de la F-mesure en fonction des itérations *Word2vec*. 84
5.2 L'évolution du nombre d'entrées du lexique en fonction des itérations. 84
5.3 La progression du nombre de formes ajoutées entre la première et la dernière itération. 85

6.1 Une vue globale du schéma d'analyse de sentiments. 93
6.2 Taxonomie de l'*appraisal*. 96
6.3 Un exemple d'une fiche produite par la méthode de [Zhang and Ferrari, 2010] . 97

6.4 Les mots proches avec le mot *criminal* obtenus en utilisant les vecteurs *Word2vec* de *Google*. 99
6.5 Nouvelle taxonomie de l'*appraisal*. 101
6.6 Le canevas utilisé pour produire une évaluation qualitative. 102
6.7 Un exemple d'évaluation de sentiments pour deux vidéos comparables dans deux langues différentes. 103

Liste des tableaux

1.1 Exemple d'encodage en utilisant une table de correspondance. 23

2.1 Les quatre représentations graphiques de la lettre ع selon sa position dans le mot. 30
2.2 Quelques sens possibles du mot كتب en rajoutant les voyelles. 30
2.3 Les dérivés possibles de la racine كتب. 31
2.4 Exemples des mots empruntés dans les trois dialectes. 32
2.5 Quelques emprunts récents dans le dialecte algérien. 33
2.6 Des écritures possibles dans le dialecte marocain. 34
2.7 Comparaison entre l'arabe et les dialectes maghrébins. 35
2.8 Exemple de mots-clés utilisés pour la recherche des vidéos. 36
2.9 Quelques statistiques sur les corpus collectés. 36
2.10 La distribution du français, de l'anglais et de l'arabizi dans la partie SL des trois corpus collectés. 38
2.11 La distribution du MSA et du dialecte dans la partie écrite en script arabe des trois corpus. 38
2.12 Les cinq segments les plus utilisés en anglais et en français de longueur trois dans les trois corpus maghrébins. 39

3.1 Les scores de code-switching obtenus avec les facteurs de Ghosh et CESAR sur des exemples. 46
3.2 Mesurer le code-switching des trois corpus maghrébins avec CESAR et Ghosh. 49
3.3 Résultats du code-switching dans les trois corpus en utilisant le français comme langue de référence. 50
3.4 Résultats du code-switching en désactivant les mots MSA dans le dialecte. . . . 50

4.1 Nombre de tweets anglais collectés pour chaque *hashtag*. 55
4.2 Nombre de tweets arabes collectés pour chaque *hashtag*. 55
4.3 Exemples de dates homogénéisées. 57
4.4 Exemples des dates et des nombres avant et après les pré-traitements. 57
4.5 La classification des affixes proposée par [Kadri and Nie, 2006]. 58
4.6 Les particules utilisées dans les noms propores. 61
4.7 Table de groupes de lettres et les index correspondants de Soundex en anglais et en arabe. 62
4.8 Les rappels R@1, R@5 et R@10 obtenus avec les deux mesures LG et LGT sur les deux corpus parallèles C_{ANN} et $C_{t_{340}}$. 64
4.9 Des exemples de tweets comparables identifiés automatiquement. 64

4.10 Exemples de commentaires du corpus comparable. 65
4.11 Variantes d'écriture et mots corrélés du mot **kho** en dialecte algérien. 68
4.12 Résultats obtenus par les méthodes proposées en terme de rappel $R@1$, $R@5$, et $R@10$ sur un corpus de développement construit manuellement. 74

5.1 Quelques exemples montrant la variabilité lexicale existant dans les dialectes maghrébins. 78
5.2 Quelques exemples de mots mal orthographiés. 79
5.3 Des exemples des mots après la suppression des voyelles arabes. 80
5.4 Encodage des caractères utilisés en arabizi. 80
5.5 Des exemples des mots après le traitement. 80
5.6 Différentes façons d'écrire le mot يرحمك (*Dieu vous bénisse*) en script latin. . . 81
5.7 Exemples de quelques entrées du lexique algérien. 81
5.8 Exemples de quelques entrées du lexique marocain. 82
5.9 Exemples de quelques entrées du lexique tunisien. 82
5.10 La taille des trois lexiques. 82
5.11 Un exemple des candidats potentiels obtenus grâce à l'encodage de Soundex. . 83
5.12 Un exemple d'une entrée du lexique de *Word2vec* référence. 83
5.13 Quelques exemples de mots germes positifs et négatifs pour les trois dialectes. . 86
5.14 La taille des trois lexiques de sentiments. 87
5.15 Quelques exemples de mots positifs ou négatifs extraits des trois lexiques de sentiments construits automatiquement. 87
5.16 Résultats expérimentaux sur les corpus des dialectes du Maghreb. 88
5.17 Une analyse détaillée des résultats d'analyse de sentiment sur les trois corpus des dialectes du Maghreb. 89
5.18 Résultats expérimentaux sur des corpus de même taille de trois dialectes. . . . 90
5.19 Une analyse détaillée des résultats d'analyse de sentiment sur des corpus de même taille. 90

6.1 Le nombre de vidéos par langue. 92
6.2 La performance de différentes méthodes de comparabilité en termes Top1, Top5, et Top10 sur un corpus de référence d'Euronews. 95
6.3 Quelques exemples de la liste *MW363*. 98
6.4 Quelques exemple du lexique BingApp. 100
6.5 Les quatre classes de modifieurs de l'intensité. 101
6.6 Les résultats de l'analyse d'opinions en utilisant un lexique de polarité (*Bing*) et un lexique d'*appraisal* (*BingApp*). 101

B.1 Exemple de tweets parallèles extrait du corpus développé par [Ling et al., 2013]. 110
B.2 Table de translittération utilisée. 111

C.1 Exemples de commentaires comparables de CALYOU. 112

43

Introduction générale

Le domaine du Traitement Automatique des Langues (TAL) a réalisé, ces dernières années, des progrès considérables dans plusieurs domaines, notamment, le développement d'outils touchant à plusieurs axes du langage, la disponibilité de quantités considérables de données, l'existence de librairies d'apprentissage prêtes à l'emploi, le développement d'architectures d'apprentissage profond, etc. Tout cela a permis d'atteindre de très bonnes performances dans plusieurs applications. Citons à titre d'exemple, les systèmes de traduction automatique qui ont fait un bond en avant en termes de qualité de traduction et ce, pour plusieurs paires de langues. Cela est dû non seulement aux méthodes d'apprentissage profond adoptées, mais également à la disponibilité de grandes quantités de corpus parallèles. En effet, d'après les résultats de WMT18 [Ott et al., 2018], les auteurs ont entraîné un système de traduction automatique pour la paire de langues anglais-allemand en utilisant un corpus parallèle de WMT14 et un corpus issu de *ParaCrawl*[1] comportant 140M de paires de phrases identifiées automatiquement.

En dépit de la particularité de la langue arabe, actuellement cette langue, comme beaucoup de langues naturelles dispose de suffisamment de ressources pour envisager des applications en traitement automatique de la langue grandeur nature. La preuve en est que plusieurs projets de recherche ont porté sur le traitement automatique de l'arabe, citons à titre d'exemple les projets MEDAR (*Mediterranean Arabic Language and Speech Technology*), AQMAR (*American and Qatari Modeling of Arabic*) et AMIS (*Access to Multilingual Information and Opinions*). Néanmoins, tous ces travaux de recherche se sont focalisés sur l'arabe standard, mettant de côté l'arabe dialectal. En effet, dans le monde arabe, deux formes de langues coexistent : une connue sous le terme d'arabe standard, notée par la suite MSA (*Modern Standard Arabic*). Il s'agit de la langue des journaux, de l'école, de l'administration, de certaines émissions de télévision, etc. La deuxième forme de langue est connue sous le terme de dialecte qui est propre à chaque région du monde arabe. Il s'agit d'une forme informelle qui correspond au moyen de communication le plus naturel dans le monde arabe. Cette forme parlée de l'arabe est généralement fondée sur l'arabe standard. Cependant, plusieurs contraintes morpho-syntaxiques de la langue d'origine ont été écartées pour laisser place à une langue informelle plus facile d'usage. Jusqu'à un passé récent, les différentes formes de ces dialectes n'ont pas intéressé les chercheurs puisque elles étaient limités seulement à une utilisation orale. Avec l'arrivée des réseaux sociaux, les dialectes arabes sont désormais écrits. Cela est dû au fait que les usagers s'expriment et débattent plus facilement en utilisant leurs langues maternelles, en l'occurrence les dialectes arabes. Cela a conduit la communauté TAL à s'intéresser au traitement automatique des dialectes qui recouvre plusieurs défis scientifiques.

1. https://paracrawl.eu/releases.html

L'intérêt suscité par les formes vernaculaires du MSA est généralement motivé par la surveillance des réseaux sociaux par des administrations de sécurité, par des intérêts militaires ou politiques [Aransa, 2015], etc. Nous pouvons citer également l'intérêt porté au domaine de l'analyse de sentiments de textes dialectaux extraits des réseaux sociaux concernant les pays touchés par "le printemps arabe".

L'intérêt de traiter les formes vernaculaires de l'arabe est reconnu déjà depuis un certain temps. Cependant, la disponibilité de certaines ressources constitue la condition *sine qua non* pour le développement d'applications supportant les dialectes. Malgré, la disponibilité de textes d'une manière abondante dans les réseaux sociaux, les dialectes font partie, pour l'instant, de la classe des langues peu dotées. En effet, une langue est dotée en ressources, si elle dispose évidemment de corpus de textes de taille importante, mais aussi de dictionnaires, d'analyseurs morphologiques, etc. ce qui n'est malheureusement pas le cas des dialectes arabes. C'est pourquoi, dans cette thèse, nous menons un travail de recherche afin de développer des méthodes automatiques permettant la construction de certains types de ressources pour les dialectes du Maghreb telles que : les corpus comparables et les lexiques comprenant les différentes formes d'un même mot et leur polarité. Ces ressources ont été mises à la disposition de la communauté travaillant sur le traitement automatique du dialecte.

Par ailleurs, les ressources et les méthodes développées ont été utilisées dans le cadre du projet AMIS (*Access to Multilingual Information and Opinions*). Ce projet visait à développer un système d'aide à la compréhension de l'information multilingue sans aucune intervention humaine. Ce qu'on entend par compréhension dans ce contexte consiste à construire un résumé pertinent dans une langue cible à partir d'une vidéo dont la langue est étrangère. Par ailleurs, ce projet permet aux utilisateurs, non seulement, de visionner l'information dans sa propre langue, mais aussi de la comparer à une autre vidéo portant sur le même sujet et diffusée dans une langue étrangère. Cette comparaison est réalisée sur le plan des sentiments identifiés dans chaque vidéo. Notre travail dans ce projet consistait à construire une revue d'analyse de sentiments à partir de deux vidéos dans deux langues différentes et portant sur le même sujet. Pour ce faire, nous nous sommes basés sur la théorie de l'*appraisal*, pour proposer une analyse plus fine que celle donnée par l'approche fondée sur la polarité des mots.

Ce manuscrit comprend six chapitres. Le premier et le deuxième concernent l'état de l'art. Les quatre autres chapitres concernent nos contributions.

Ce manuscrit est structuré de la manière suivante :

— Le Chapitre 1, est consacré à la présentation des différents travaux liés aux problématiques soulevées dans cette thèse. Nous commençons le chapitre par la présentation des travaux qui concernent la construction des corpus multilingues et plus particulièrement les corpus comparables multilingues. Puis nous recensons les méthodes permettant la détermination automatique de l'orientation sémantique des mots afin de construire des lexiques de sentiments. Nous terminons le chapitre par la présentation des principaux types de représentations de données, en détaillant plus particulièrement les représentations distribuées, sur lesquelles s'appuient les méthodes que nous avons développées dans cette thèse.

— Dans le Chapitre 2, nous décrivons brièvement les particularités de l'arabe standard et de l'arabe dialectal en mettant l'accent sur les dialectes concernés par notre étude. Ensuite, nous comparons ces deux formes et nous mettons l'accent sur les spécificités de la forme dialectale de l'arabe. Pour ce faire, nous proposons une analyse détaillée effectuée sur un corpus de dialectes que nous avons collecté dans le cadre de ce travail. Dans cette étude, nous présentons quelques chiffres liés à la spécificité des dialectes maghrébins comme le style d'écriture et le phénomène du code-switching.

— Dans le Chapitre 3, nous proposons une métrique qui permet de quantifier le bruit dans un corpus dû au phénomène du code-switching. Cette métrique est nommée CESAR (CodE-Switching According to a Reference language), elle associe 0 à un document dans lequel le texte est écrit entièrement dans la langue de référence. Une valeur maximale égale à 1 est affectée à tout document étranger à la langue de référence. Cette mesure pourrait être utilisée pour extraire des corpus comportant un maximum de mots du dialecte étudié.

— Le Chapitre 4 décrit les méthodes que nous avons utilisées pour construire des corpus comparables à partir des données provenant des réseaux sociaux. Les corpus obtenus contiennent des documents alignés. Ces derniers peuvent être écrits dans une ou plusieurs langues en sachant qu'un même document peut utiliser plusieurs langues. Ce chapitre comporte deux parties. Dans la première, nous présentons nos propositions pour aligner des tweets en arabe standard et en anglais. Nous citons, au passage, les traitements liés à la particularité des tweets et plus spécialement à ceux de l'arabe standard. La deuxième partie de ce chapitre traite de la problématique d'alignement des données vernaculaires avec comme objectif de construire un corpus comparable. À notre connaissance, ce type de corpus n'existe pas. Nous avons ainsi proposé une méthode permettant d'aligner des textes qui peuvent être écrits en : dialecte, arabe standard, français et anglais. Contrairement à ce qui est fait pour l'alignement des tweets arabe-anglais, où nous disposons de dictionnaires, pour les dialectes, nous proposons une approche fondée sur le *multilingual word embedding* pour aligner automatiquement des données vernaculaires sans faire appel à des ressources externes.

— Dans le Chapitre 5, nous proposons des méthodes permettant de produire des ressources lexicales. La première contient la variabilité lexicale d'un même mot, une particularité des dialectes arabes. La variabilité lexicale est due au manque de standardisation d'écriture, à l'utilisation de l'arabizi et au manque de rigueur grammaticale dans l'écriture dans les réseaux sociaux. La méthode proposée, nous a permis de construire automatiquement des lexiques de ce type pour les trois dialectes traités dans cette thèse. Chaque entrée comporte les différentes formes d'écriture de celle-ci. Cette ressource peut être très utile dans de nombreuses applications en traitement automatique des langues. On peut l'utiliser par exemple en reconnaissance automatique de la parole ou en traduction automatique en phase de test. La deuxième ressource concerne un lexique de sentiments dans lequel un degré de polarité est associée à chaque entrée. L'orientation sémantique des mots dialectaux a été déterminée automatiquement en se basant sur des mots proches ayant une orientation prédéterminée. La méthode proposée a conduit à l'obtention de lexiques de sentiments pour les trois dialectes : l'algérien, le marocain et le tunisien.

— Dans le Chapitre 6, nous avons été confrontés à deux défis du projet AMIS. Le premier a consisté à extraire les paires de vidéos comparables parmi l'ensemble de vidéos collectées dans le cadre du projet AMIS. Pour ce faire, nous avons testé plusieurs approches, ensuite avec la méthode ayant donné la meilleure performance, nous avons aligné les transcriptions des vidéos obtenues grâce aux systèmes de reconnaissance automatique de la parole développés dans le cadre de ce projet. Quant au deuxième défi, il consiste à proposer une revue d'analyse en termes de sentiments pour les paires de vidéos portant sur le même sujet et que nous avons alignées automatiquement.

Cette thèse se termine par une conclusion qui récapitule nos réalisations ainsi que quelques perspectives.

1

État de l'art

Sommaire

1.1 Introduction 7
1.2 Les corpus 7
1.3 Les corpus monolingues 8
1.4 Les corpus multilingues 8
1.4.1 Les corpus parallèles 8
1.4.2 Les corpus comparables multilingues 10
1.4.3 Évaluation du degré de comparabilité des corpus multilingues 12
1.5 L'analyse de sentiments 18
1.5.1 Méthodes d'attribution d'une orientation sémantique à une entrée lexicale 19
1.5.2 Méthodes et ressources pour une analyse à granularité fine 22
1.6 La représentation des données 23
1.6.1 Représentation discrète 23
1.6.2 Représentations distributionnelles 24
1.7 Conclusion 27

Dans ce chapitre, nous mettons l'accent sur des approches de l'état de l'art destinées à la collecte et la construction des ressources types : corpus multilingues, et plus particulièrement les corpus comparables et les ressources lexicales comme les lexiques de sentiments. Nous consacrons également une partie de ce chapitre pour présenter les représentations vectorielles des mots.

1.1 Introduction

Le traitement automatique des langues nécessite l'utilisation de ressources telles que les corpus (monolingues et multilingues), les corpus vocaux, les lexiques et les dictionnaires, etc. Certaines ressources sont le résultat d'un effort humain considérable. En effet, des linguistes ont consacré beaucoup de temps à la création de lexiques ou de dictionnaires bilingues qui constituent une référence solide sur laquelle se sont basés plusieurs travaux dans le domaine. Des traducteurs professionnels ont traduit une quantité gigantesque de documents provenant d'institutions officielles, ce qui a permis de mettre à la disposition de la communauté des corpus parallèles indispensables au développement de systèmes de traduction automatique.

Malgré la qualité de ces ressources, ce travail fastidieux est à refaire à chaque fois que l'on doit changer de langue. C'est pourquoi, plusieurs recherches se sont focalisées sur le développement des ressources obtenues d'une manière semi-automatique ou automatique afin de réduire au maximum le coût lié à l'intervention humaine.

Le présent chapitre est consacré à l'état de l'art lié à la construction semi-automatique ou automatique de certaines ressources nécessaires au traitement de la langue en général et au traitement des dialectes arabes en particulier. Ces ressources concernent : les corpus multilingues et les lexiques de sentiments. De ce fait, dans ce chapitre, nous décrirons dans les premières sections, les méthodes proposées dans la littérature pour construire des corpus multilingues en mettant en exergue ceux concernant les corpus comparables. Nous décrirons également les différentes méthodes existant dans la littérature pour construire des lexiques de sentiments, nous nous inspirerons de ces méthodes pour proposer des ressources similaires pour les trois dialectes du Maghreb. Nous terminerons ce chapitre par la description des méthodes utilisées pour la représentation des données et plus particulièrement les représentations distribuées que nous utiliserons dans nos contributions.

Pour ne pas alourdir ce chapitre, nous présenterons certains travaux liés à nos contributions directement dans les chapitres concernés.

1.2 Les corpus

Dans cette thèse, les corpus constituent le matériel de base nécessaire à tous les traitements que nous réaliserons dans ce mémoire. C'est pourquoi, il est nécessaire de définir, de catégoriser et de recenser les corpus. Les corpus constituent une ressource importante dans le traitement des langues. Plusieurs définitions ont été proposées par des linguistes, notamment celle de Sinclair qui convient à une étude linguistique ou à un traitement informatique [Harris, 1954], traduit par Prochasson [Prochasson, 2009] :

Definition 1. *Un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques et extra-linguistiques explicites pour servir d'échantillon d'emplois déterminés d'une langue*

Un corpus est une collection de données homogène construit ou collecté à des fins de traitement linguistique ou informatique. Ces données peuvent être de différentes natures : texte, audio, image, vidéo, etc. Il existe plusieurs critères de classement des corpus, si l'on

choisit le critère de la langue, nous pouvons les classer en deux catégories : les corpus monolingues et les corpus multilingues.

1.3 Les corpus monolingues

Les corpus monolingues représentent une collection de données dont la langue est unique. Ces corpus sont répandus et volumineux, puisque dès l'apparition d'Internet, les chercheurs en traitement automatique de langue se sont tournés vers ce média pour collecter le matériel nécessaire à leurs travaux. Ces corpus ont notamment été utilisés en reconnaissance automatique de la parole et en traduction automatique pour le développement des modèles de langage. Ces corpus sont généralement volumineux, c'est le cas de *Gigaword* [Parker et al., 2011] qui comporte un milliard de mots. Les réseaux sociaux constituent une source de données très intéressante pour la collecte des données qui nécessite des traitement spécifiques et qui représentent un défi scientifique auquel la communauté s'intéresse de plus en plus.

1.4 Les corpus multilingues

Le Web par nature est multilingue, autrement dit, on trouve sur Internet des textes et des données en plusieurs langues. Pour toucher un maximum de monde, certains médias et sites diffusent leurs contenus en plusieurs langues. Par exemple, le site *Euronews* publie des articles en 12 langues. Citons un autre exemple : Wikipédia qui diffuse des contenus concernant le même sujet en plusieurs langues. Ces plateformes et sites constituent désormais des sources d'information extrêmement utiles pour la communauté travaillant sur des données multilingues. À travers les liens proposés par ces plateformes, il est donc possible de construire des corpus multilingues dans lesquels on regroupe des données concernant généralement le même sujet.

Dans cette thèse, nous définissons un corpus multilingue comme une collection de données en plusieurs langues dans lequel, il existe des textes portant sur des sujets similaires.

Bien que les données multilingues sont aujourd'hui disponibles, elles sont néanmoins relativement peu nombreuses par rapport à la pléthore de corpus monolingues existants. En outre, pour les langues peu dotées, ce type de corpus est encore plus rare.

Les corpus multilingues sont utilisés dans plusieurs domaines comme : la traduction automatique, l'extraction des lexiques multilingues, la recherche d'information multilingue, l'analyse de sentiments dans des corpus multilingues, etc. Il est possible de classer ces corpus multilingues en deux catégories : les corpus parallèles et les corpus comparables. Dans ce qui suit, nous présenterons ces deux catégories de corpus, mais nous consacrerons plus de temps aux corpus comparables, objet de notre étude.

1.4.1 Les corpus parallèles

Un corpus est dit parallèle s'il est composé de textes en langue source et de leurs traductions exactes en une ou plusieurs langues cibles. Ces corpus jouent un rôle primordial pour le développement de systèmes multilingues, en général et des systèmes de traduction

automatique, en particulier. Malheureusement, les corpus parallèles ne sont pas toujours librement accessibles et il y a peu de paires de langues pour lesquelles ces corpus sont disponibles en quantité permettant de développer un système de traduction automatique. Pour certaines langues comme l'anglais, le français, l'espagnol, l'allemand, l'arabe, le chinois, etc. les corpus sont généralement disponibles, en revanche pour les langues peu dotées il est difficile de trouver des corpus parallèles. Dans ce qui suit nous donnons quelques exemples des corpus utilisés dans la littérature. Ces corpus ont été créés particulièrement par les institutions internationales où le multilinguisme est de rigueur. Ou bien par des institutions travaillant sur la diffusion d'informations multilingues.

— **Europarl** [Koehn, 2005] : un corpus largement utilisé en traduction automatique. Ce corpus regroupe des textes parallèles extraits des débats du Parlement européen dans 21 langues. Le nombre de phrases parallèles dépend de la paire de langues, ainsi pour la paire roumain-anglais il y a 0,4 millions de phrases parallèles alors que pour la paire français-anglais il y a 2 millions de phrases parallèles.

— **MultiUN** [Eisele and Chen, 2010a] : ce corpus est extrait des documents des nations unies traduit dans les six langues officielles. La paire de langues ayant le plus de phrases parallèles est la paire français-anglais et le corpus parallèle le moins volumineux comporte 8 millions de phrases anglaises-chinoises.

— **TED** : TED [2] fournit des corpus parallèles de sous-titres de discussions consacrées à la diffusion de présentations et de conférences couvrant des sujets très divers et comportant 14 paires de langues. Cependant la taille de ces corpus n'est pas volumineuse.

— **Hansard** : ce corpus rassemble des textes parallèles issus du Sénat canadien. Ce corpus contient 2,8 millions de phrases parallèles français-anglais [3].

— **Hong-Kong Hansard** [Ma, 1999] : ce corpus a été rassemblé et nettoyé par LDC à partir des comptes rendus de réunions hebdomadaires du parlement de Hong Kong. Il comporte 238 721 phrases parallèles pour la paire de langues anglais-chinois.

Tous les corpus cités ci-dessus concernent des langues pour lesquels les ressources langagières sont disponibles. Malheureusement, on trouve très peu de corpus parallèles destinés aux langues dites peu dotées. En général, ce terme fait référence à des langues qui ne disposent pas de ressources langagières (corpus monolingues et multilingues, corpus vocaux, dictionnaires monolingues et bilingues, etc). Ces dernières années, la communauté TAL a commencé à travailler sur la création de telles ressources. Les langues vernaculaires sont de bons exemples de langues peu dotées en ressources. Dans [Honnet et al., 2018], les auteurs ont collecté un corpus parallèle de 60k mots afin de développer un système de traduction pour le dialecte suisse-allemand [4]. Pour la création de corpus parallèles pour les dialectes arabes, nous pouvons citer les travaux suivants :

— **Le corpus de Bouamor** : les auteurs de [Bouamor et al., 2014] ont demandé à 4 traducteurs (locuteurs natifs de Palestine, Syrie, Jordanie et Tunisie) de traduire 2 000 phrases écrites en dialecte égyptien dans leurs dialectes. L'égyptien a été choisi comme langue de départ parce que c'est le dialecte le mieux compris dans le monde arabe.

2. http ://nlp.ffzg.hr/resources/corpora/ted-talks/
3. https ://catalog.ldc.upenn.edu/LDC95T20
4. Le suisse allemand désigne l'ensemble des dialectes largement parlés en Suisse mais qui sont rarement écrits.

— **PADIC** [Meftouh et al., 2015, Meftouh et al., 2018] : ce corpus rassemble des phrases en arabe standard et leurs traductions en français et en six dialectes arabes : le dialecte d'Alger et de ses environs, le dialecte d'Annaba (est de l'Algérie), un dialecte tunisien, un dialecte marocain, un dialecte syrien et un dialecte de Gaza (Palestine). PADIC est composé de 6 400 phrases parallèles. Ce corpus a été créé en traduisant manuellement des conversations en dialecte algérois vers l'arabe standard et ensuite vers les autres dialectes et vers le français. Ce corpus a été utilisé pour développer le premier système de traduction pour certains dialectes du Maghreb.

— **MADAR** [Bouamor et al., 2018] : MADAR est une collection de phrases parallèles couvrant les dialectes de 25 villes du monde arabe dont plusieurs du même pays, en plus de l'anglais, du français et de l'arabe standard. Ce corpus a été créé en traduisant 12 000 phrases du corpus de voyage BTEC *Basic Traveling Expression Corpus*.

Malgré l'existence de PADIC et de MADAR, il est important de fournir plus d'efforts pour créer ou pour construire automatiquement des corpus plus conséquents afin de pouvoir développer des systèmes de traduction automatique performants pour les dialectes arabes. En effet, si on compte le nombre total de phrases parallèles des corpus précédemment cités, on ne dispose que de 20 400 phrases parallèles pour des dizaines de dialectes dans le monde arabe.

Comme nous l'avons mentionné précédemment, ces corpus parallèles sont le résultat du travail de traducteurs humains qui traduisent d'une langue vers une autre. Dans certains cas, ce travail est plutôt semi-automatique, on traduit automatiquement avec un système et ensuite on corrige manuellement les erreurs de traduction.

Une alternative aux corpus parallèles, qui sont coûteux en développement, est l'exploitation des corpus comparables. En effet, comme le Web est par définition multilingue, il est possible d'aligner des documents écrits dans des langues différentes et ensuite d'en extraire des segments parallèles pour construire finalement un corpus parallèle. Dans la section suivante nous détaillons ces corpus comparables.

1.4.2 Les corpus comparables multilingues

Selon Teubert [Teubert, 1996] les corpus comparables sont des corpus comportant des données en deux ou plusieurs langues ayant une composition ou une structure similaire (ou quasi-similaire). Les auteurs de [Fung and Cheung, 2004a, Fung and Cheung, 2004b] ont défini les corpus comparables comme étant des corpus composés de phrases qui ne sont pas nécessairement des traductions les unes des autres comme dans les corpus parallèles. De ce fait, l'alignement dans ces corpus se fait sur le domaine, le sujet, le thème, etc. Une autre définition, permettant de quantifier la comparabilité a été proposée par Déjean et Gaussier. dans [Déjean and Gaussier, 2007] :

Definition 2. *Deux corpus de deux langues l_1 et l_2 sont dits comparables s'il existe une sous-partie non-négligeable du vocabulaire du corpus de langue l_1, respectivement l_2, dont la traduction se trouve dans le corpus de langue l_2, respectivement l_1*

Dans la figure 1.1 nous donnons un exemple de deux documents comparables extraits de Wikipédia. Cette figure montre également que les documents comparables peuvent partager des segments parallèles.

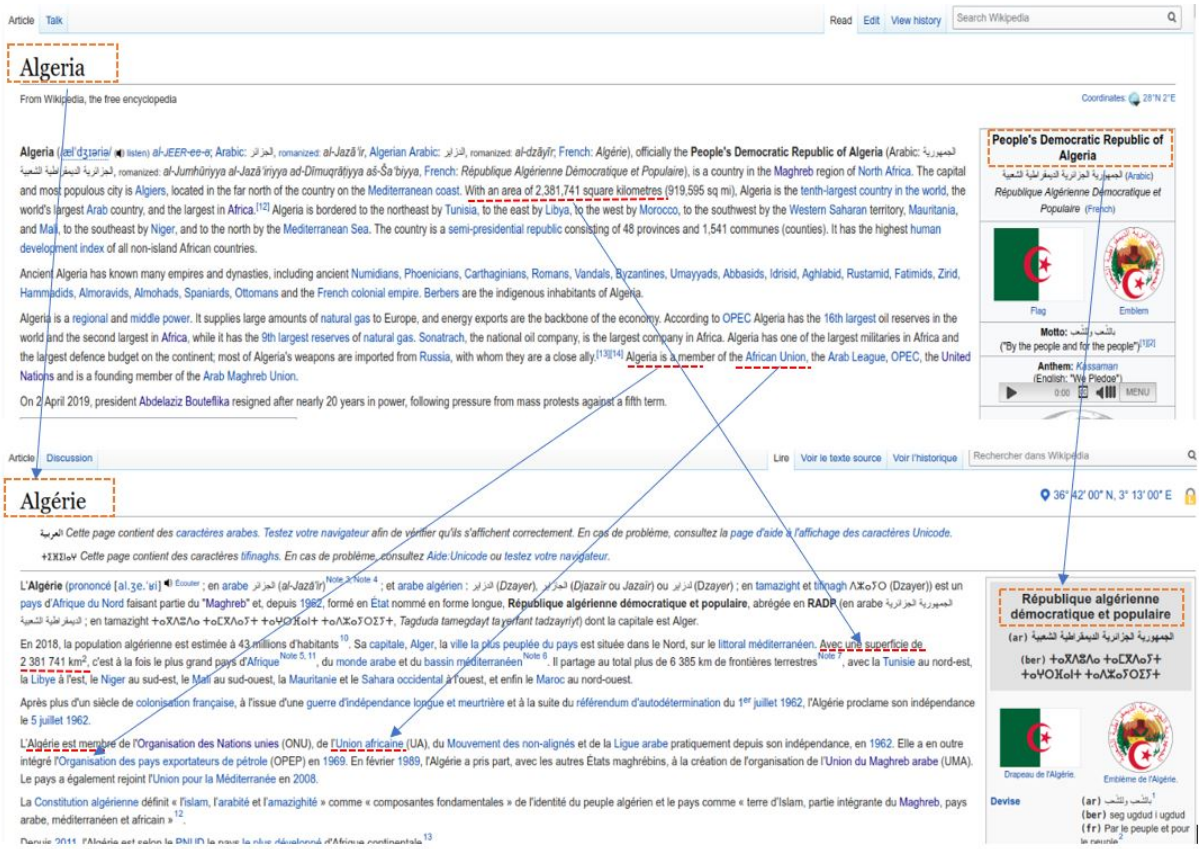


FIGURE 1.1 – Exemple de deux textes comparables extraits de Wikipédia.

Les corpus comparables ont été exploités dans plusieurs domaines du traitement automatique des langues, citons, entre autres, l'extraction des lexiques bilingues, la traduction automatique, la fouille de données multilingues, etc.

L'idée d'utiliser des corpus comparables afin d'extraire des lexiques bilingues est ancienne, elle remonte aux travaux de [Rapp, 1995]. Plusieurs autres travaux ont suivi qui avaient comme objectif de repérer des termes dans des textes sources et leurs traductions dans des textes cibles [Fung, 1997], [Déjean and Gaussier, 2007], [Morin et al., 2008], [Ismail and Manandhar, 2010], [Bo et al., 2011], [Tamura et al., 2012], [Hazem and Morin, 2018] et [Liu et al., 2018]. La plupart de ces travaux se basent sur l'hypothèse de la co-occurrence des mots pour identifier les paires multilingues dont une partie est la traduction de l'autre. Dans [Lavecchia et al., 2007a], les auteurs ont construit à partir d'un corpus comparable de sous-titres de films un dictionnaire bilingue français-anglais. Dans les travaux de [Li and Gaussier, 2010, Bo et al., 2011, Li and Gaussier, 2013], les auteurs ont d'abord mis l'accent sur la qualité des corpus comparables en développant des méthodes assurant une meilleure qualité des documents comparables avant de les exploiter pour en extraire des lexiques bilingues.

Des travaux similaires ont été proposés pour étendre les travaux précédents à l'identification de segments plus longs dans deux langues différentes. En effet, l'idée étant d'associer à une suite de mots dans une langue source la traduction dans une langue cible. Les auteurs

de [Munteanu et al., 2004] ont pu améliorer les performances d'un système de traduction anglais-arabe en utilisant des phrases parallèles extraites automatiquement à partir d'un corpus comparable journalistique. Dans [Lavecchia et al., 2007b] les auteurs ont construit un corpus parallèle français-anglais, à partir d'une base de données de sous-titres de 40 films. Le résultat de ce travail est un corpus de 37625 paires alignées obtenu grâce à une méthode de type programmation dynamique dont les performances étaient de 92,3% de précision.

Dans [AbduI-Rauf and Schwenk, 2009], les auteurs ont pu améliorer considérablement les performances de leur système de traduction en utilisant des corpus comparables. Les auteurs ont traduit automatiquement la partie source du corpus comparable et ensuite chaque texte traduit a été considéré comme une requête au sens de la recherche d'information. Le système développé permet d'utiliser en plus des informations supplémentaires comme la date de publication, cette requête pour identifier dans le corpus cible les textes les plus proches. Cette approche s'apparente à ce qui se fait dans le domaine de la recherche d'information inter-langues (*Cross-Language Information Retrieval*) (CLIR), un domaine qui vise à extraire des documents pertinents dans une langue différente de celle de la requête formulée par l'utilisateur. Généralement, la CLIR repose sur des ressources linguistiques telles que les systèmes de traduction automatique, les dictionnaires bilingues ou les corpus parallèles. Dans [Shakery and Zhai, 2013], les auteurs ont montré qu'il est possible d'effectuer une recherche d'information multilingue en utilisant des corpus comparables sans avoir recours à des ressources externes. Pour ce faire, les auteurs ont utilisé une méthode d'association de mots multilingues (*cross-lingual word association*) basée sur la similitude de distribution de fréquence de paires de mots. Puis, pour une requête d'utilisateur en langue L_1, les auteurs estiment les mots les plus corrélés en langue cible. À l'aide de ces mots, ils construisent une requête en langue L_2. Cette dernière est utilisée pour faire correspondre les documents cibles à la requête de l'utilisateur.

Citons une approche intéressante proposée par [Ramesh and Sankaranarayanan, 2018], elle est fondée sur un réseau de neurones récurrent bidirectionnel de type *end-to-end* siamois permettant de générer des phrases parallèles à partir d'articles multilingues comparables extraits de Wikipédia. Les auteurs se sont intéressés aux couples de langues anglais–tamoul et anglais-hindi. Le modèle développé apprend à estimer la probabilité de traduction d'un couple de phrases donné en utilisant, en plus des corpus parallèles, un corpus d'échantillons négatifs dans lequel on présente des phrases sources et des phrases cibles qui ne sont pas des traductions les unes des autres. Avec 10000 articles comparables, les auteurs ont identifié 78000 phrases parallèles anglais-hindi et 76000 phrases parallèles anglais-tamoul.

1.4.3 Évaluation du degré de comparabilité des corpus multilingues

Comme mentionné dans la section précédente, il existe plusieurs sites Web à partir desquels des corpus comparables sont collectés. Toutefois, l'alignement de ces textes est une tâche difficile qui nécessite l'utilisation ou le développement de méthodes de mise en correspondance des textes similaires. Ces méthodes utilisent généralement des mesures permettant d'évaluer le degré de comparabilité de deux textes.

Dans [Rose et al., 1998], les auteurs mentionnent, que mesurer la comparabilité de deux corpus est une tâche complexe et comporte de multiples facettes et que différentes mesures

peuvent être nécessaires pour atteindre l'objectif de rendre les corpus similaires. Il est à noter que dans ce travail il s'agissait de corpus monolingues. Rendre des corpus multilingues comparables est une tâche encore plus complexe.

Dans [Saralegi et al., 2008], les auteurs ont proposé une méthode basée sur le contexte des mots pour mesurer la similarité entre les documents de deux corpus. Cette approche est fondée sur l'hypothèse suivante : plus les documents sont similaires, plus le contexte des mots est proche. Soit C_1 un corpus en langue L_1 constitué de m documents $d^i_{l_1}$ avec $i \in [1...m]$ et un corpus C_2 en langue L_2 composé de n documents $d^j_{l_2}$ avec $j \in [1...n]$. Chaque document de ces corpus est représenté par un vecteur contenant ses termes les plus représentatifs. Entre deux documents les auteurs utilisent une mesure de similarité pour calculer leur degré de rapprochement [Saralegi and Alegria, 2007]. Ensuite, ces scores sont reportés dans une matrice DM. Chaque élément S_{ij} de cette matrice correspond à la similarité entre les documents $d^i_{l_1}$ et $d^j_{l_2}$. Afin d'estimer le score global entre C_1 et C_2, les auteurs ont utilisé la distance connue sous le terme EMD (*Earth Movers Distance*) dont la matrice DM est un paramètre.

$$DM = \begin{array}{c} \\ d^1_{l2} \\ d^2_{l2} \\ \vdots \\ d^n_{l2} \end{array} \begin{array}{c} \begin{array}{ccccc} d^1_{l1} & d^2_{l1} & d^3_{l1} & \cdots & d^m_{l1} \end{array} \\ \begin{pmatrix} S_{11} & S_{12} & S_{13} & \dots & S_{1m} \\ S_{21} & S_{22} & S_{23} & \dots & S_{2m} \\ \vdots & \vdots & \ddots & \vdots & \\ S_{n1} & S_{n2} & S_{n3} & \dots & S_{nm} \end{pmatrix}_{(n \times m)} \end{array}$$

D'après [Goeuriot, 2009] cette méthode est efficace sur un nombre limité de documents, car le calcul de la similarité de toutes les paires de documents multilingues d'un corpus de taille importante rend cette mesure peu pratique et inefficace.

Les auteurs de [Otero and Lopez, 2011] construisent un vocabulaire de termes V_s correspondant aux liens se trouvant dans une page Wikipédia. De la même manière, ils exploitent les liens inter-langues de chaque page pour construire un dictionnaire bilingue tel que $\forall x \in V_s \ \exists y = trans(x)$. $trans$ étant une fonction renvoyant la traduction de son paramètre identifié à partir des liens inter-langues de Wikipédia. y pouvant renvoyer la valeur Null s'il n'existe pas de page Wikipédia correspondant à x dans la langue cible. Pour deux documents C_s et C_t, les auteurs se proposent de calculer le degré de comparabilité de C_s et C_t. Pour ce faire, à partir du vocabulaire de termes de C_s et du dictionnaire bilingue construit grâce aux liens de traductions, les auteurs mettent en correspondance ces traductions avec les termes du vocabulaire de C_t. Le degré de comparabilité est calculé selon la formule suivante :

$$Dice_{bin}(C_s, C_t) = \frac{2 \sum_{t_s \in V_s} Trans(t_s, V_t)}{|V_s| + |V_t|} \tag{1.1}$$

Avec $trans(w, V)$ qui renvoie 1 si la traduction de w est dans V, 0 sinon.

L'inconvénient de cette méthode est le fait qu'elle soit destinée aux documents extraits de Wikipédia, par conséquent, on ne peut pas l'exploiter pour évaluer la comparabilité entre des corpus provenant d'une autre source que Wikipédia.

Par ailleurs, les auteurs de [Su and Babych, 2012] ont proposé trois approches différentes pour mesurer la comparabilité de documents : une basée sur l'alignement lexical, une basée sur les mots-clés et une autre basée sur la traduction automatique.

1. **Mesure fondée sur l'alignement lexical**
   Afin de palier au manque des dictionnaires bilingues pour certaines langues, les auteurs ont proposé de les construire à partir de corpus parallèles. Cela est fait, par l'utilisation de l'outil d'alignement des mots *GIZA++*. Les dictionnaires obtenus sont utilisés par la suite pour aligner les paires de documents mot-à-mot. Enfin, une mesure de similarité de type cosinus est utilisée par les auteurs afin d'estimer le degré de comparabilité entre les documents.
2. **Métrique basée sur les mots clés**
   Cette méthode est similaire à la précédente exceptée le fait que pour comparer deux documents on ne prend en compte que les mots les plus représentatifs de chacun d'eux. Pour ce faire, les auteurs gardent pour chaque document les n mots les plus pertinents en terme de TF-IDF.
3. **Métrique fondée sur la traduction automatique**
   Dans ce cas les auteurs utilisent un système de traduction au lieu de se contenter d'un outil d'alignement mot-à-mot. Pour ce faire, les auteurs utilisent une API de traduction de Microsoft et exploite d'autres caractéristiques pour le calcul de la mesure de comparabilité.
   - Caractéristique lexicale : les documents sont représentés par des sacs de lemmes ne contenant pas les mots outils de la langue (*stop words*). Ensuite, une similarité lexicale SL de chaque paire de documents est calculée en utilisant le cosinus.
   - Caractéristique de structure : la structure d'un document est représentée par le nombre de mots porteurs de sens W et par le nombre de phrases S. L'idée étant que si deux documents D_1 et D_2 sont proches, alors ils sont composés par approximativement le même nombre de phrases et le même nombre de mots porteurs de sens. La similarité de structure ST de deux documents D_1 et D_2 est définie comme suit : $ST = 0.5 * (W_{D_1}/W_{D2}) + 0.5 * (S_{D_1}/S_{D_2})$
   - Caractéristique de mots-clés : les auteurs ont gardé les 20 premiers mots en fonction de leur valeur TF-IDF. Puis la similarité de cette représentation en mots-clés SK de deux documents est mesurée par le cosinus.
   - Caractéristique des entités nommées : cette mesure est fondée sur l'utilisation des entités nommées de chaque document. Si plusieurs entités nommées co-occurrent dans deux documents, alors elles sont susceptibles de faire référence au même événement ou au même sujet et par conséquent les documents sont probablement comparables. Pour identifier les entités nommées, les auteurs ont utilisé l'outil de *Stanford* (Finkel et al., 2005). Comme pour les autres caractéristiques la mesure cosinus a été utilisée pour calculer la similarité *entités nommées* SN.

   Enfin, ils ont combiné les quatre valeurs de similarité pour définir un score global de comparabilité SC :

$$SC = \alpha * SL + \beta * ST + \gamma * SK + \delta * SN$$

   Avec $\alpha + \beta + \gamma + \delta = 1$.

Ensuite, les auteurs ont testé leurs trois méthodes sur 6 corpus comparables collectés et annotés manuellement selon leur degré de comparabilité. Les auteurs ont montré que la métrique fondée sur la traduction automatique (dont la mesure est nommée SC) donne les meilleurs résultats. Cette dernière méthode nous semble intéressante, cependant il est difficile de l'utiliser sur nos corpus de dialectes arabes pour lesquels nous n'avons pas de système de traduction automatique ni de système d'identification d'entités nommées.

Une autre mesure de comparabilité fondée sur l'analyse sémantique latente (LSA) a été utilisée par les auteurs de [Saad et al., 2014] qui se sont inspirés de la méthode de [Dumais et al., 1998]. Cette méthode a été employée auparavant pour certaines paires de langues, mais les auteurs de [Saad et al., 2014] l'ont utilisée pour la première fois pour l'arabe. Cette méthode peut être employée pour aligner et comparer des documents multilingues. L'avantage de cette méthode est qu'elle n'a pas besoin d'un dictionnaire bilingue, ni d'analyseurs morphologiques ni de systèmes de traduction automatique. Contrairement à la méthode de [Dumais et al., 1998], les auteurs ont appliqué la méthode LSA non seulement sur des données parallèles, mais également sur des paires de documents comparables extraits de Wikipédia. Comme nous l'avons mentionné, leur approche est basée sur l'analyse sémantique latente, une technique largement utilisée pour la recherche de documents similaires, pour la recherche de relations entre termes, pour l'identification de concepts, etc. Cette technique se base sur l'hypothèse suivante : deux mots sont sémantiquement proches s'ils apparaissent dans des contextes similaires et deux contextes sont similaires s'ils utilisent des mots proches.

Avant de considérer l'utilisation de la LSA dans un contexte multilingue, nous donnons ci-dessous la formalisation de la méthode dans un contexte monolingue. Soit une matrice M de termes-documents, où les lignes correspondent aux termes utilisés dans les différents documents et les colonnes représentent les documents.

$$M = \begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \\ \vdots \\ t_n \end{array} \begin{array}{c} \begin{array}{ccccc} d_1 & d_2 & d_3 & \cdots & d_m \end{array} \\ \begin{pmatrix} w_{11} & w_{12} & w_{13} & \dots & w_{1m} \\ w_{21} & w_{22} & w_{23} & \dots & w_{2m} \\ w_{31} & w_{32} & w_{33} & \dots & w_{3m} \\ \vdots & \vdots & \ddots & \vdots & \\ w_{n1} & w_{n2} & w_{n3} & \dots & w_{nm} \end{pmatrix}_{(n\times m)} \end{array}$$

L'analyse sémantique latente consiste à calculer la décomposition en valeurs singulières de M :

$$\begin{array}{c} \\ \text{Termes} \end{array} \begin{array}{c} \overset{Documents}{} \\ \begin{pmatrix} \\ M \\ \\ \end{pmatrix}_{(n\times m)} \end{array} = \begin{pmatrix} \\ U \\ \\ \end{pmatrix}_{(n\times k)} \times \begin{pmatrix} \\ S \\ \\ \end{pmatrix}_{(k\times k)} \times \begin{pmatrix} \\ V^t \\ \\ \end{pmatrix}_{(k\times m)} \tag{1.2}$$

Où U et V sont des matrices orthonormales et S est une matrice diagonale contenant les valeurs singulières de M. Quelques propriétés intéressantes concernant cette méthode sont données ci-dessous :

— Le produit de deux vecteurs termes $t_i \times t_j^t$ donne la corrélation entre les deux mots i et j.

— Le produit matriciel $M^t M$ donne pour chaque cellule, de la matrice résultat du produit, la valeur de la corrélation entre les deux termes correspondant à cette cellule.

— Le produit matriciel MM^t donne pour chaque cellule, de la matrice résultat du produit, la valeur de la corrélation entre les deux documents correspondant à cette cellule.

— Lorsqu'on sélectionne les k plus grandes valeurs singulières et les vecteurs singuliers correspondants dans U et V on obtient une approximation de la matrice M d'occurrences termes-documents. Cette réduction de rang permet de projeter les données de départ dans un ensemble de k concepts avec $M_k = U_k \times S_k \times V_k^t$

— Le vecteur ligne t_i possède alors k composantes. Chaque valeur de ce vecteur donne le lien du terme i et chaque concept latent.

— Le vecteur colonne d_j possède alors k composantes. Chaque valeur de ce vecteur donne l'importance du document j dans chacun des différents concepts latents.

— Pour comparer deux documents d_i, d_j, on calcul le cosinus de $S_k \times d_i$ et $S_k \times d_j$.

— Même principe pour les termes t_i et t_j.

— En phase de test lorsqu'on souhaite mesurer le degré de similarité d'un nouveau document, il suffit de le considérer comme une requête q qu'on transforme en vecteur. On calcule $q_k = {S_k}^{-1} {U_k}^t q$. Ce vecteur est ensuite comparé au reste des documents de la matrice.

Pour l'utilisation de la LSA dans un contexte multilingue, les auteurs de [Saad et al., 2014] ont construit une matrice M_{mul} composée de la concaténation des textes sources et des textes cibles alignés. Chaque d_i est la concaténation du texte source d_i^s et de sa cible d_i^t. Par conséquent, la matrice M_{mul} représente un corpus composé de n documents et $l + m$ termes dont l est le nombre des termes dans la langue source et m est le nombre des termes dans la langue cible.

$$M_{mul} = \begin{array}{c} \\ t_1^s \\ t_2^s \\ t_3^s \\ \vdots \\ t_l^s \\ \\ t_1^t \\ t_2^t \\ t_3^t \\ \vdots \\ t_m^t \end{array} \begin{array}{c} \begin{array}{ccccc} d_1 & d_2 & d_3 & \cdots & d_n \end{array} \\ \begin{pmatrix} w_{11}^s & w_{12}^s & w_{13}^s & \cdots & w_{1n}^s \\ w_{21}^s & w_{22}^s & w_{23}^s & \cdots & w_{2n}^s \\ w_{31}^s & w_{32}^s & w_{33}^s & \cdots & w_{3n}^s \\ \vdots & \vdots & \ddots & \vdots & \\ w_{l1}^s & w_{l2}^s & w_{l3}^s & \cdots & w_{ln}^s \\ & & & & \\ w_{11}^t & w_{12}^t & w_{13}^t & \cdots & w_{1n}^t \\ w_{21}^t & w_{22}^t & w_{23}^t & \cdots & w_{2n}^t \\ w_{31}^t & w_{32}^t & w_{33}^t & \cdots & w_{3n}^t \\ \vdots & \vdots & \ddots & \vdots & \\ w_{m1}^t & w_{m2}^t & w_{m3} & \cdots & w_{mn}^t \end{pmatrix}_{((l+m)\times n)} \end{array}$$

Où w_{ij}^x sont les poids TF-IDF des termes.

De la même manière que pour l'utilisation de la LSA sur des corpus monolingues, les auteurs ont utilisé ce principe en l'appliquant à la matrice M_{mul} sur des corpus multilingues.

Grâce à cette méthode, les auteurs ont aligné de nouveaux corpus multilingues. Les performances ont été meilleures que celles obtenues avec la méthode basée sur les dictionnaires. En ce qui nous concerne, nous ne pourrons pas utiliser cette méthode pour l'alignement de données provenant des réseaux sociaux puisque la construction de la matrice M_{mul} nécessite l'utilisation de documents comparables ou parallèles pour les dialectes arabes qui sont très peu répandus.

Les auteurs de [Li and Gaussier, 2010] ont proposé une méthode fondée sur l'utilisation d'un dictionnaire bilingue. L'idée sous-jacente est que plus les documents sont comparables plus il y a de chances de trouver des mots dans un document source pour lesquels il existe des traductions dans le document cible. Pour un corpus comparable C constitué d'une partie C_{l_s} en langue l_s et d'une partie C_{l_t} en langue l_t, si on considère dans un premier temps la comparabilité dans le sens $l_s \Rightarrow l_t$, la mesure de comparabilité $M_{l_s \Rightarrow l_t}$ peut être définie comme le nombre de mots de C_{l_s} pour lesquels il existe une traduction dans C_{l_t}. Plus formellement, on peut exprimer cette contrainte de la manière suivante :

$$\sigma(w_s, V_t) = \begin{cases} 1 \ si \ T(w_s) \cap V_t \neq \emptyset \\ 0 \ sinon \end{cases} \tag{1.3}$$

La fonction σ renvoie 1 si une des traductions de w_s notées $T(w_s)$ appartient au vocabulaire V_t composé des mots du corpus C_{l_t}. Pour établir ce type de correspondance, les auteurs font appel à un dictionnaire bilingue $Dic_{l_s \Rightarrow l_t}$.

La recherche des mots correspondant dans le document cible C_{l_t} ne concerne que les mots du document source C_{l_s} qui existent dans la partie gauche du dictionnaire $Dic_{l_s \Rightarrow l_t}$. Par conséquent, la mesure de comparabilité $M_{l_s \Rightarrow l_t}$ est définie comme suit :

$$M_{l_s \Rightarrow l_t} = \sum_{w_s \ \in \ \{C_{l_s} \cap Dic_{l_s \Rightarrow l_t}\}} \sigma(w_s, V_t) \tag{1.4}$$

En fait, les auteurs de cette méthode ont proposé une mesure symétrique, autrement dit, ils cherchent des mots dans le document cible correspondant aux mots du document source et inversement. Par conséquent, la formule finale de la mesure est donnée par :

$$M_{l_s \Leftrightarrow l_t} = \frac{\sum_{w_s \ \in \ \{C_{l_s} \cap Dic_{l_s \Rightarrow l_t}\}} \sigma(w_s, V_t) + \sum_{w_t \ \in \ \{C_{l_t} \cap Dic_{l_t \Rightarrow l_s}\}} \sigma(w_t, V_s)}{\mid C_{l_s} \cap Dic_{l_s \Rightarrow l_t} \mid + \mid C_{l_t} \cap Dic_{l_t \Rightarrow l_s} \mid} \tag{1.5}$$

Cette méthode de comparabilité a été testée d'abord sur un corpus parallèle où les mesures de comparabilité sont sensées donner les plus grandes valeurs puisque la comparabilité dans ce cas là est maximale. Ensuite, pour évaluer la méthode sur des corpus comparables, ils les ont construits à partir du corpus parallèle. En effet, le corpus parallèle a été subdivisé en 10 sous-corpus $(C_1, C_2, C_3, ..., C_{10})$. Puis, on a remplacé certaines parties des corpus C_i par des textes non-parallèles. L'objectif étant de dégrader la qualité des corpus parallèles C_i et de les transformer dans une certaine mesure en des corpus comparables. Les auteurs ont montré que la mesure baissait au fur à mesure de la dégradation de la qualité des corpus parallèles.

1.5 L'analyse de sentiments

Savoir ce que les personnes pensent à propos d'un produit, d'un sujet, d'un film, etc. offre des perspectives très intéressantes dans de nombreux domaines. Par exemple, en politique, on peut connaître le positionnement des citoyens par rapport à un sujet délicat (par exemple le mariage pour tous) à travers les discussions dans les réseaux sociaux. On peut connaître également la tendance d'intention de vote pour une élection présidentielle, ce qui a été le cas pour l'élection de *Trump* que tous les sondages classiques avaient donné perdant alors que l'analyse des forums de discussions a montré le contraire. Dans le domaine du Marketing, connaître les avis positifs ou négatifs des clients envers un produit peut aider les entreprises à améliorer la qualité de leurs marchandises et mieux cibler la clientèle. Ces applications intéressent plus particulièrement la communauté TAL qui travaille sur l'analyse de sentiments (*sentiment analysis*) ou la fouille d'opinions (*opinion mining*). Ce domaine permet d'étudier les opinions, les sentiments, les appréciations, les attitudes, etc. exprimés par les personnes envers une entité.

Ces études visent à déterminer la position des gens envers un sujet. Cette position est connue sous le nom d'*orientation sémantique* qui est généralement matérialisée par la polarité, l'émotion ou l'objectivité versus la subjectivité. La polarité est le concept le plus étudié dans le domaine de l'analyse de sentiments. La polarité regroupe principalement deux classes positive et négative. La première classe indique si le jugement porté est positif par exemple *une machine puissante*, tandis que la deuxième concerne un jugement négatif par exemple, *une machine lente*. Nous présenterons plus loin, dans cette section, d'autres concepts plus raffinés associés à l'orientation sémantique.

L'attribution d'une orientation sémantique à un sujet est traitée dans la littérature par deux approches. La première est une approche non-supervisée, elle s'appuie sur l'utilisation de ressources comportant des informations liées aux sentiments comme les lexiques, les ontologies, etc. La deuxième consiste en une approche supervisée qui se base sur des corpus annotés et des méthodes d'apprentissage automatique. Dans cette section de l'état de l'art nous nous focalisons seulement sur la première approche. Cela se justifie par le fait que dans nos travaux nous nous intéressons à la création de ressources pour l'analyse de sentiments et aussi parce que la littérature concernant les méthodes supervisées est abondante.

L'approche non-supervisée s'appuie sur l'utilisation de ressources externes comme les lexiques de sentiments étiquetés au préalable. Ce type de ressources est généralement composé de termes avec leurs orientations sémantiques. Ensuite, pour un document donné, l'orientation sémantique de celui-ci est estimé en fonction de la polarité de chaque terme de ce document. Les lexiques de sentiments disponibles et les plus couramment utilisés dans la littérature sont :

— **General Inquirer** : ce lexique a été construit manuellement par les auteurs de [Stone et al., 1962], il est disponible à l'adresse [5]. Ce lexique a été développé pour l'analyse des discussions visant à la découverte des thèmes psychologiques sous-jacents. Il n'est pas dédié spécifiquement à l'analyse de sentiments, mais à une étude beaucoup plus complète. Il comporte 182 catégories correspondant à différents thèmes, tels que la polarité, la faiblesse, la force, la douleur, la sensation, la famille, etc. En ce qui concerne

5. http://www.wjh.harvard.edu/~inquirer/homecat.htm

la partie destinée à l'analyse de sentiments, il existe une liste de 1915 mots positifs et 2291 mots négatifs en anglais.

— **Bing Liu's Opinon Lexicon** : le lexique *Bing* a été construit par les auteurs de [Hu and Liu, 2004], il est disponible à l'adresse [6]. Il est composé de 4 913 mots négatifs et 2 718 mots positifs.

— **MPQA** : ce lexique a été créé par les auteurs de [Wiebe et al., 2005], il est disponible à l'adresse [7]. Il fournit une liste de 8 222 entrées composées chacune du mot, de l'intensité de sa subjectivité, de sa catégorie grammaticale et de sa polarité.

— **SentiWordNet** : ce lexique a été construit par les auteurs de [Baccianella et al., 2010] à partir de *WordNet*. Il est composé d'un ensemble de groupes de mots partageant une relation de synonymie. Chacun de ces groupes correspond à une entrée appelée *synset* ou *synonym set*. Chaque *synset* dispose de trois scores numériques représentant la positivité (Pos), la négativité (Neg) et l'objectivité (Obj). La dernière version de ce lexique contient 117 659 entrées [8].

Dans beaucoup de travaux l'estimation du score d'un document se fait par un simple comptage des mots positifs et négatifs sans tenir compte de la variabilité de la polarité de chaque mot et de son impact sur le sentiment global du document. Certaines autres études ont suggéré d'affecter des poids différents aux termes en fonction de leur degré de polarité. À titre d'exemple, dans le lexique *SentiWordNet* à chaque *synset* (s) est associé trois scores numériques entre 0 et 1 : Obj(s), Pos(s) et Neg(s). Ces scores décrivent à quelle point les termes contenus dans le *synset* sont objectifs, positifs et négatifs. Dans [Tian et al., 2018] les auteurs attribuent à chaque terme un score entre -3 et +3 pour prendre en compte le degré de variabilité des sentiments : fortement négatif (-3), négatif (-2), faiblement négatif (-1), neutre (0), faiblement positif (+1), positif (+2), fortement positif (+3).

1.5.1 Méthodes d'attribution d'une orientation sémantique à une entrée lexicale

Il existe plusieurs approches conduisant à l'attribution d'une orientation sémantique à un mot. Dans les sections suivantes on décrira les grandes lignes de ces méthodes.

Méthode manuelle

Cette approche est la plus simple, mais aussi la plus fastidieuse. En effet, dans cette méthode, chaque mot est analysé par un expert qui lui attribue la polarité correspondante. Lorsque ce travail est effectué par un linguiste, le lexique obtenu devient une ressource pertinente et permet, par conséquent, d'effectuer des analyses conduisant à de très bonnes performances. Plusieurs travaux ont adopté cette approche, citons notamment le cas de [Stone et al., 1962], [Asher et al., 2008], [El-Halees, 2011]. Une fois que cette ressource est construite, on procède à l'estimation de l'orientation sémantique d'un document qui se calcule d'une manière simple par une fonction combinant la polarité de chacun des mots de celui-ci.

6. https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
7. http://mpqa.cs.pitt.edu/
8. http://ontotext.fbk.eu/sentiwn.html

Plus récemment, la communauté TAL s'est intéressée à ce type de travail afin de développer des ressources pour les dialectes arabes ou pour les langues peu dotées [Amiri et al., 2015], [Duwairi et al., 2015], [Mataoui et al., 2016].

Méthodes semi-automatiques

Dans ces méthodes, on utilise une ressource externe comportant une liste de termes étiquetés préalablement avec leurs polarités. Ensuite, un algorithme est développé, qui a comme objectif de compléter la ressource externe en allant identifier des mots dont la polarité est similaire à celle des mots de la ressource de départ.

Les auteurs de [Hatzivassiloglou and McKeown, 1997] partent du principe que les adjectifs liés par des conjonctions peuvent partager une orientation similaire ou différente selon la conjonction utilisée. Par exemple, dans la phrase *elle est belle et intelligente* la conjonction *et* relie deux adjectifs de même orientation. Alors que dans l'exemple *elle est belle mais stupide*, la conjonction *mais* oppose les deux adjectifs et donc ils se verront affectées des polarités différentes. Pour pouvoir appliquer cette heuristique, il est nécessaire de disposer d'un dictionnaire des adjectifs préalablement annotés en terme de polarité. Pour ce faire, les auteurs ont identifié, à partir d'un corpus extrait du *Wall Street journal* annoté avec des *part-of-speech* , tous les adjectifs apparaissant plus de 20 fois. Ensuite, ils ont traité manuellement la liste obtenue en supprimant tous les adjectifs neutres et en supprimant les adjectifs dont la polarité est ambiguë. Cette liste comporte 657 adjectifs positifs et 679 adjectifs négatifs. Ensuite, en appliquant l'heuristique citée ci-dessus, ils ont construit une liste plus large.

Dans [Hu and Liu, 2004] les auteurs ont proposé une méthode simple qui est basée sur le constat que les mots et leurs synonymes partagent la même orientation sémantique et une orientation opposée avec leurs antonymes. Partant de cette idée, les auteurs ont construit manuellement une liste de 30 mots positifs et négatifs, puis ils ont développé un algorithme itératif qui cherche dans le dictionnaire *WordNet* les mots synonymes et leur affecte la même orientation et attribue une orientation opposée aux mots antonymes. De même, une approche similaire a été proposée par les auteurs de [Mohammad et al., 2009] afin de créer une ressource lexicale de polarité. Pour ce faire, ils se sont servis du thésaurus *Macquarie Thesaurus* pour identifier les synonymes des mots positifs et négatifs d'une liste construite manuellement.

Dans [Turney, 2002], les auteurs ont proposé une approche permettant de calculer l'orientation sémantique d'un mot donné sur la base de sa proximité sémantique avec d'autres mots appelés par les auteurs "mots germes" (*seed words*) dont l'orientation sémantique est connue. Les auteurs ont proposé deux listes de mots germes regroupant les mots ayant une polarité non ambiguë, une liste de mots germes positifs comportant les mots suivants : *{good, nice, excellent, positive, fortunate, correct, et superior}* ; une liste de mots germes négatifs comportant les mots suivants : *{bad, nasty, poor, negative, unfortunate, wrong, et inferior}*. L'idée de cette approche est en quelque sorte similaire à ce que nous avons présenté dans cette section, c'est-à-dire qu'un mot est considéré comme étant positif s'il est plus proche des mots germes positifs et plus éloigné des mots germes négatifs. De la même manière, un mot est considéré comme étant négatif s'il est plus proche des mots germes négatifs et plus éloigné des mots germes positifs. Les auteurs proposent d'estimer l'orientation sémantique d'un mot selon la

formule suivante :

$$OS(w) = \sum_{W_p \in POS} A(w, W_p) - \sum_{W_n \in NEG} A(w, W_n) \tag{1.6}$$

Avec POS et NEG qui représentent respectivement la liste des mots germes ayant une orientation sémantique positive et négative. $A(w_1, w_2)$ est une mesure d'association entre les mots w_1 et w_2. Pour mesurer l'association entre w_1 et w_2, les auteurs ont testé la *Pointwise Mutual Information* (PMI) et l'analyse sémantique latente.

Une approche similaire à celle de Turney et al. a été proposée par [Bestgen, 2002]. Elle s'appuie sur une liste des mots germes beaucoup plus large que celle de Turney et al. En effet, cette liste comporte 3000 mots germes dont la valence de l'orientation sémantique correspond à la moyenne des scores attribués par une trentaine d'experts. Ensuite, pour un mot donné w dont la valence est inconnue, les auteurs lui affecte la valeur moyenne de la valence de ses 30 plus proches voisins. Ces derniers sont identifiés en calculant la similarité cosinus entre les vecteurs représentatifs du mot w et les mots germes obtenus grâce à l'analyse sémantique latente. Dans un autre travail de recherche [Bestgen, 2006], Bestgen a comparé sa méthode avec celle de [Turney, 2002] sur des corpus provenant de plusieurs domaines. Il montre que l'utilisation d'un dictionnaire de germes plus riche et l'utilisation de corpus diversifiés pour calculer l'orientation sémantique des mots ont un impact considérable sur les performances.

Un autre travail adoptant la même méthodologie a été proposé dans [Htait et al., 2017]. Pour construire les listes des mots germes, les auteurs ont collecté les 100 mots les plus fréquents à partir d'un très grand corpus de tweets [Go, 2009]. Après un filtrage de ces listes, les auteurs ont conservé 38 mots pertinents pour chaque polarité. Puis, ils ont ajouté à ces listes les mots germes de Turney et al. Le calcul de la similarité entre un nouveau mot dont la polarité n'est pas connue et les mots germes se fait comme pour la méthode de Turney et al. (voir formule 1.6) sur des vecteurs obtenus grâce au modèle *Word2vec* appris sur un corpus comportant 300 millions de tweets.

Pour résumer, les méthodes précédentes sont globalement similaires, elles se basent toutes sur une ressource externe comportant un nombre de mots orientés sémantiquement plus au moins grand et une approche généralement simple permettant de calculer le degré de proximité entre un mot et des mots pour lesquels l'orientation est connue. Plusieurs autres travaux fondés sur le même principe continuent à être proposés, c'est notamment le cas de [Mahyoub et al., 2014], [Rouvier and Favre, 2016], [Rouces et al., 2018].

Méthode fondée sur la traduction des ressources

Étant donné la difficulté de disposer de ressources orientées sémantiquement, certains auteurs ont profité de l'existence de ces données pour l'anglais pour les traduire vers les langues sur lesquelles ils travaillent [Mihalcea et al., 2007, Meng et al., 2012, Mohammad et al., 2016]. Ces approches traduisent :

— soit le texte de la langue cible vers une langue riche en ressources telle que l'anglais, et utilisent ensuite un système d'analyse de sentiments de l'anglais sur le texte obtenu.

— soit les ressources telles que les lexiques de sentiments ou les corpus étiquetés de l'anglais vers la langue cible, et les utilisent comme une ressource principale ou supplémentaire dans le système d'analyse de sentiments de la langue étudiée.

Les auteurs des références citées ci-dessus [Mihalcea et al., 2007, Meng et al., 2012, Mohammad et al., 2016] ont appliqué cette approche pour traiter les langues suivantes : arabe, chinois et roumain.

1.5.2 Méthodes et ressources pour une analyse à granularité fine

Comme nous l'avons mentionné auparavant, les premiers travaux sur l'analyse de sentiments se contentaient d'analyser le texte selon sa polarité (positive ou négative) ou selon son aspect global exprimé en terme d'objectivité ou de subjectivité. Cependant, lorsque l'on souhaite analyser le contenu d'une discussion ou d'un document pour découvrir des thèmes psychologiques sous-jacents, ou de faire une analyse plus profonde, il faudrait se baser sur des concepts décrivant d'autres états émotionnels. Avec l'apparition des forums de discussions, on a constaté la profusion de débats et de discussions animés et chargés de différents types d'émotions. Les chercheurs disposaient désormais de matériaux permettant de faire une analyse plus profonde de ces débats et discussions. Ces émotions se matérialisent généralement par des concepts plus fins représentant l'état émotionnel d'un individu ou la description d'une entité autre. Citons à titre d'exemple le défi *DEFT* concernant l'analyse des émotions dans les tweets [Hamon et al., 2015]. Les auteurs de la campagne ont proposé plusieurs tâches autour de l'analyse des opinions exprimées dans les messages postés sur Twitter. Certaines tâches proposaient une analyse à granularité fine allant du niveau d'analyse le plus global (polarité, classes génériques) au plus fin (accord, amour, apaisement, colère, déplaisir, ennui, etc).

Les travaux portant sur l'identification de classes sémantiques plus fines reposent souvent sur des lexiques affectifs. Il s'agit de ressources dans lesquelles on trouve des classes appelées *affects* et les mots correspondants. Parmi les ressources les plus utilisées dans la littérature, on peut citer :

— **WordNet Affect** : comme *SentiWordNet*, ce lexique a été construit par les auteurs de [Strapparava and Valitutti, 2004], à partir de *WordNet*, il est disponible gratuitement à l'adresse [9]. Cette ressource est composée de *synsets* annotés avec des étiquettes affectives (*emotion, cognitive state, trait, behavior, attitude,* et*feeling*).

— **NRC-EmoLex** : ce lexique a été construit manuellement en utilisant le service *Amazon Mechanical Turk* [10] par les auteurs de [Mohammad and Turney, 2013], il est disponible à l'adresse [11]. Il est composé de plus de 14000 mots en anglais étiquetés selon la polarité et l'émotion (joie, confiance, anticipation, tristesse, surprise, dégoût, peur et colère). Chaque mot de ce lexique a un score entre 0 et 1 qui représente l'intensité de l'affect.

— **FEEL** : cette ressource est destinée à l'analyse de sentiments pour le français. Elle a été construite par les auteurs de [Abdaoui et al., 2017], elle est disponible à l'adresse [12]. Cette ressource est composée de 5704 mots positifs et 8423 mots négatifs classés selon les émotions suivantes : joie, confiance, anticipation, tristesse, surprise, dégoût, peur et colère.

9. http://hltdistributor.fbk.eu/index.php
10. https://www.mturk.com/
11. https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm
12. http://www.lirmm.fr/~abdaoui/FEEL.html

1.6 La représentation des données

Réussir à développer des méthodes d'apprentissage efficaces dépend généralement de la représentation des données, car de celle-ci, des relations cachées peuvent être exploitées pour mieux interpréter les données ou pour fournir une bonne représentation pour les méthodes d'apprentissage automatique. Cependant, il est important d'identifier les unités linguistiques nécessaire à une représentation pertinente des corpus. La représentation la plus couramment utilisée est celle connue sous le terme sac-de-mots (*bag-of-words*). À partir d'un corpus, on collecte les mots les plus représentatifs généralement ceux qui sont les plus fréquents. Dans certains cas, les représentations les plus pertinentes ne sont pas constituées de mots mais de structures infra-mots : lemmes, syllabes, racines, etc. Dans d'autres cas, il est utile de représenter un même corpus par des mots, des infra-mots et des supra-mots. Dans [Meftouh et al., 2019], les auteurs ont participé à une campagne d'évaluation dont la tâche consistait à identifier le dialecte d'une phrase parmi 26 variétés de dialectes du monde arabe, les auteurs ont montré que la meilleure représentation pour cette tâche était celle qui combine des bigramm, 3-grammes, 4-grammes, 5-grammes de caractères, 1-gramme et bigrammes de mots. Cette représentation leur a permis d'être classés premier sur 19 participants.

Dans ce qui suit, nous présentons les principales représentations, en détaillant plus particulièrement les représentations distribuées de mots, sur lesquelles s'appuient les méthodes que nous avons développées dans cette thèse.

1.6.1 Représentation discrète

Dans ce type de représentation, il existe plusieurs manières de représenter le sac-de-mots.

Encodage par des valeurs numériques dans une table de correspondance

Dans cette représentation une fois que le texte est pré-traité et que l'unité linguistique est choisie, on associe à chaque item un index. On construit ainsi une table de correspondance dans laquelle chaque terme (caractère, lemme, mot ou autres) est associé à une valeur numérique. Cette représentation a un avantage certain du point de vue informatique, mais ne permet pas d'avoir des liens sémantiques entre les mots étant donné que les index sont attribués au fur et à mesure de la rencontre des mots dans le corpus. Par exemple dans la table 1.1, malgré le fait que les mots maison et maisonnette sont proches leurs index sont très différents.

Mot	Index
Maison	3622
Maisonnette	5825
Voiture	1

Table 1.1 – Exemple d'encodage en utilisant une table de correspondance.

Encodage One-hot

Dans cette représentation, l'idée est de créer un vecteur de la taille du vocabulaire dans lequel toutes les dimensions sont à 0 sauf le mot que l'on veut représenter pour lequel sa

dimension est initialisée à 1. En fait, chaque vecteur est encodé avec $n+1$ éléments où n est la taille de vocabulaire et la dimension supplémentaire est consacrée au mot inconnu comme dans l'exemple *Tototo*.

$$\begin{array}{ll} \text{Maison :} & (0,\ 0,\ \ldots,\ 1,\ 0,\ \ldots,\ 0) \\ \text{Maisonnette :} & (0,\ 0,\ \ldots,\ 0,\ \ldots,\ 1,\ 0) \\ \text{Voiture :} & (1,\ 0,\ \ldots,\ 0,\ \ldots,\ 0,\ 0) \\ \text{Tototo :} & (0,\ 0,\ \ldots,\ 0,\ \ldots,\ 0,\ 1) \end{array}$$

Encodage fréquentiel

Plusieurs encodages sont fondés sur l'existence ou non de mots d'un vocabulaire donné.

— Un encodage binaire où chaque document est représenté par un vecteur de n dimensions. Si le mot existe dans le document sa dimension est positionnée à 1 sinon à 0.

— Un encodage par fréquence où chaque document est représenté par un vecteur de n dimensions. Si le mot existe dans le document alors la dimension correspondante est initialisée par sa fréquence sinon on lui attribue la valeur 0.

— Un encodage par pondération où l'importance et la pertinence d'un terme sont prises en compte dans la représentation. Une des plus utilisées en fouille de textes est la TF-IDF (*term frequency-inverse document frequency*). Cette dernière attribue un coefficient à chaque terme qui est proportionnel à son occurrence dans un document et sa rareté dans l'ensemble des documents.

1.6.2 Représentations distributionnelles

Dans les représentations précédentes, les mots sont représentés d'une manière discrète. Or, dans les applications en traitement automatique de la langue, il est de coutume de traiter des données de dimentionnalités élevées. Dans la représentation discrète, un mot comme *garçon* peut être associé arbitrairement à l'index *325* et le mot *fille* à l'index *245* sans aucune relation entre ces deux mots, ce qui est regrettable puisque ces deux mots peuvent se retrouver dans plusieurs contextes identiques. Dans ces méthodes fondées sur ces représentations, il n'est pas possible d'utiliser les connaissances apprises sur le mot *fille* lors du traitement du mot *garçon* ou inversement. En effet, la représentation discrète des mots conduit à l'utilisation de vecteurs de très grandes dimensions, ce qui nécessite plus de données pour un entraînement efficace des méthodes d'apprentissage statistiques ou neuronales. L'utilisation de représentations distributionnelles permet de repousser cette limite.

La représentation distributionnelle se base sur l'hypothèse formulée par Harris dans [Harris, 1954] qui consiste à affirmer que les mots ayant des significations proches se retrouvent généralement dans des contextes similaires, c'est pourquoi ces mots ont des distributions identiques ou similaires. Chaque mot est encodé par un vecteur qui a une représentation proche de tout mot semblable. Ces représentations sont réalisées à l'aide de vecteurs dits vecteurs distributionnels. Ils représentent les mots en incluant des informations liées au contexte dans lesquels ils apparaissent. L'hypothèse sous-jacente est qu'il est possible de projeter des connaissances d'un mot donné sur un autre si leurs représentations distributionnelles sont similaires.

Représentations distribuées de type *Word embedding*

Le *word embedding* est une des représentations les plus populaires des mots d'un corpus. Cette méthode est intéressante parce qu'elle est capable de capturer des liens lexicaux, syntaxiques et sémantiques entre les mots d'un même document. Le *word embedding* est une représentation distribuée d'un document dans un espace de dimensions n. Le principe du *word embedding* prédit un mot à partir de son contexte à l'aide de vecteurs denses ayant des dimensions inférieures à celle du vocabulaire de départ.

Dans ce qui suit nous présentons les modèles les plus utilisés.

Word2vec

Le *Word2vec* [Mikolov et al., 2013c] est un des modèles les plus performants du principe du *word embedding*. Il apprend à partir d'un texte brut des représentations distributionnelles des mots constituant ce texte. Autrement dit, à partir d'un corpus le *Word2vec* produit un ensemble de vecteurs en sortie. À partir d'un corpus d'apprentissage, il construit un vocabulaire et apprend une représentation de type *word embedding* des mots. Ces vecteurs peuvent être utilisés dans plusieurs applications en traitement automatique des langues. L'apprentissage de ces vecteurs est basé sur l'utilisation d'un réseau de neurones artificiels permettant de capturer des régularités sémantiques et syntaxiques [Mikolov et al., 2013b] et [Mikolov et al., 2013a]. La capture de ces régularités est rendue possible grâce à une fenêtre d'exploration fixée par un paramètre du modèle qui permet non seulement de rapprocher des mots qui se trouvent dans un même contexte, mais aussi des mots qui ne se sont pas contigus. Ces représentations possèdent également des propriétés algébriques surprenantes. Si l'apprentissage est effectué sur un corpus suffisamment grand, on peut constater par exemple que le résultat de Vecteur(*Alger*) - Vecteur(*Algérie*) + Vecteur(*Maroc*) produit un vecteur qui sera proche du vecteur du mot *Rabat*.

Deux architectures ont été proposées, CBOW et Skip-gram, pour l'apprentissage de la représentation distributionnelle *word embedding*. Elles sont illustrées dans la figure 1.2.

CBOW

CBOW (*Continuous Bag of Words*) a pour objectif de prédire un mot w_t à partir de son contexte gauche $(w_{t-k} \dots w_{t-1})$ et droit $(w_{t+1} \dots w_{t+k})$. L'objectif de l'apprentissage dans ce cas est de maximiser la probabilité d'observer le mot w_t sachant son contexte. CBOW est présentée dans la partie gauche de la figure 1.2. La couche d'entrée de ce réseau de neurones représente d'une manière binaire la présence ou l'absence des mots du contexte de w_t. Deux matrices de poids sont utilisées, la première W de taille $V \times N$ où V est la taille du vocabulaire et N le nombre de neurones de la couche cachée. Chaque ligne de W correspond à la projection d'un vecteur d'entrée dans un espace de plus faible dimension et donne par conséquent, la représentation de ce vecteur d'entrée dans cet espace. La seconde matrice $W\prime$ de poids et de dimension $N \times V$ permet de calculer un score pour chaque mot du vocabulaire. Ensuite, on utilise un classifieur *softmax* pour calculer la distribution a *posteriori* des mots.

Dans CBOW l'objectif est de maximiser la probabilité du mot w_t sachant le contexte $w_{t-k}, \dots w_{t-1}, w_{t+1}, \dots, w_{t+k}$:

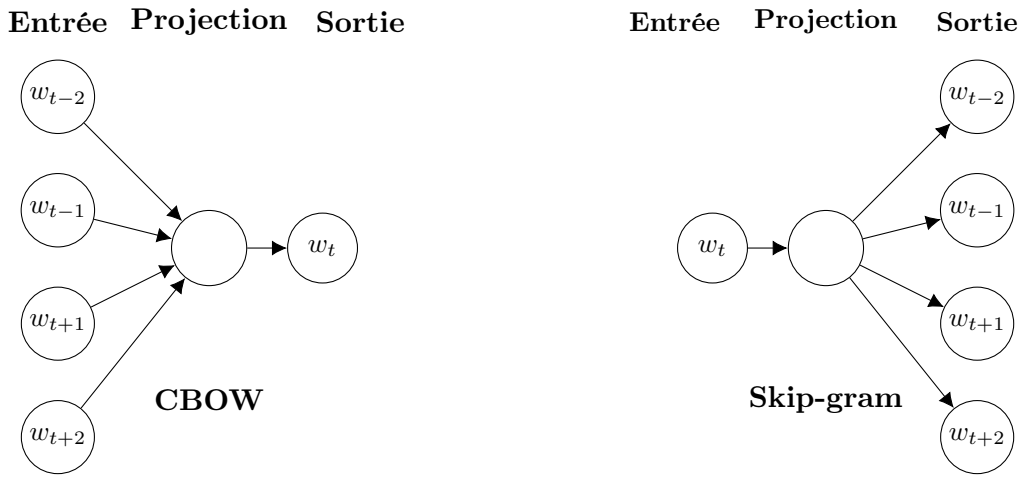


FIGURE 1.2 – Les deux architectures de *Word2vec* proposée par [Mikolov et al., 2013b].

$$S = \frac{1}{n}\sum_{t=1}^{n} log\ p(w_t|w_{t-k},\ldots,w_{t-1},w_{t+1},\ldots,w_{t+k}) \tag{1.7}$$

Où n est la taille du corpus d'apprentissage et k est la taille du contexte.

Skip-gram

Le *Skip-gram* est une reformulation du modèle précédent, l'idée est de prédire le contexte du mot w_t. L'architecture correspondant à ce modèle est présentée dans la partie droite de la figure 1.2. Le modèle *Skip-gram* est fondé sur un réseau de neurones artificiels où la couche d'entrée du réseau correspond au vecteur *one-hot* du mot w_t. Ce mot est projeté par la suite dans la matrice de poids, puis transmis à la couche de sortie qui va prédire son contexte. L'architecture *Skip-gram* maximise l'équation suivante :

$$S = \frac{1}{n}\sum_{t=1}^{n}\sum_{j=t-k,\ j\neq t}^{t+k} log\ p(w_j|w_t) \tag{1.8}$$

D'après [Mikolov et al., 2013b], les deux modèles précédents sont équivalents, leur complexité algorithmique est faible, ils peuvent être utilisés en apprentissage sur de gros corpus. Il a été établi que le modèle *Skip-gram* fonctionne mieux sur de petits corpus.

Glove *Global Vectors for Word Representation*

Glove est un autre algorithme non supervisé pour apprendre une représentation vectorielle des mots. Ce modèle a été conçu par l'équipe *Natural Language Processing* de l'université de Stanford [Pennington et al., 2014] [13]. Les auteurs défendent l'idée d'une nouvelle approche différente de celle du *Word2vec* par le fait que cette dernière est sous-optimale, puisqu'elle n'utilise pas les informations statistiques concernant les co-occurrences de mots. L'idée de

13. https://nlp.stanford.edu/projects/glove/

cette approche est de prendre en compte dans le calcul la co-occurrence des mots qui auraient un potentiel d'encodage sémantique plus puissant. La méthode consiste à construire une matrice $V \times V$ où V est la taille du vocabulaire extrait du corpus d'apprentissage. Chaque cellule (i, j) de cette matrice contient le nombre de fois où un mot w_i a été rencontré dans le contexte du mot w_j. Comme cette matrice peut être gigantesque, elle est factorisée d'une manière récursive afin d'obtenir une matrice de plus faible dimension $V \times N$ où N est la taille de la représentation distribuée d'un mot.

Les auteurs de cette approche montrent que leur méthode produit une représentation distribuée meilleure et plus rapide que celle du *Word2vec*. Néanmoins, plusieurs travaux ont montré que les deux méthodes sont globalement similaires en terme de performance. D'autre travaux ont montré que le *Word2vec* et notamment le *Skip-gram* donne de meilleurs résultats [Berardi et al., 2015]. Dans [Wang et al., 2019], les auteurs ont effectué plusieurs expériences en utilisant entre autres le *Word2vec* et Glove et ils montrent que les performances dépendent du domaine d'application, mais que globalement les résultats sont meilleurs avec le *Word2vec*.

1.7 Conclusion

Dans ce chapitre, nous avons présenté l'état de l'art lié aux travaux correspondant à nos contributions qui seront présentées dans les chapitres suivants. Il s'agit notamment de la construction de ressources : les corpus multilingues et plus particulièrement les corpus comparables et les lexiques de sentiments. Parce que nos données textuelles nécessitent une représentation, nous avons décrit les principales représentations vectorielles de données existant dans la littérature (discrète, distributionnelle). D'autres travaux liés directement à certaines de nos contributions seront présentés dans les chapitres concernés.

2

Collecte et analyse de corpus pour les dialectes du Maghreb

Sommaire

2.1 Introduction 29
2.2 La langue arabe 29
2.2.1 L'arabe standard moderne 30
2.2.2 L'arabe dialectal 31
2.3 Sélection et collecte des données dialectales 36
2.4 Étude analytique sur les trois corpus maghrébins 37
2.4.1 L'écriture du dialecte maghrébin dans les réseaux sociaux 37
2.4.2 L'utilisation du code-switching dans les dialectes maghrébins 38
2.5 Conclusion 39

Dans les pays arabes, généralement, deux formes de langue coexistent : une langue formelle, connue sous le terme d'arabe standard, que l'on notera par la suite MSA (Modern Standard Arabic) et une langue informelle qui correspond au moyen de communication dans la vie courante.

Cette langue informelle est généralement fondée sur l'arabe standard, cependant plusieurs contraintes morpho-syntaxiques de la langue d'origine sont écartées pour constituer une langue informelle plus facile d'usage.

Le présent chapitre est dédié à la présentation des particularités de l'arabe standard et de la langue informelle que l'on appellera dans ce qui suit : arabe dialectal. Nous présenterons également, les différents corpus des dialectes maghrébins que nous avons collectés et sur lesquels nous mènerons nos études et expérimentations.

2.1 Introduction

Les réseaux sociaux sont devenus aujourd'hui incontournables, les usagers de beaucoup de pays ont pris possession de ces réseaux et y postent des contenus variés : informations textuelles, images, vidéos, etc. L'utilisation des réseaux sociaux dans le monde arabe a connu aussi une progression importante, plus particulièrement après le printemps arabe.

Comme dans le monde arabe, deux langues coexistent, les internautes de ces pays ont la possibilité de poster leurs messages soit en arabe standard soit dans leurs propres dialectes (langues vernaculaires). Un des intérêts, de cette profusion de messages, pour la communauté scientifique travaillant sur la langue arabe est de pouvoir collecter facilement des masses de données pour les exploiter dans les différents traitements de type TAL (Traitement Automatique de la Langue).

Ce chapitre est consacré à l'étude de l'arabe standard et des particularités de trois dialectes du Maghreb, à savoir ceux de l'Algérie, du Maroc et de la Tunisie. Cette étude sur les dialectes sera de type analytique (comparaison des usages des trois dialectes dans les réseaux sociaux) et elle sera menée à travers les corpus dialectaux collectés à partir des réseaux sociaux. La section 2.2 de ce chapitre est consacrée à la définition et la présentation de la langue arabe et de l'arabe vernaculaire spécifique au pays du Maghreb. Nous terminons cette section par un tableau récapitulatif qui résume la différence entre ces deux formes de l'arabe sur plusieurs aspects. Dans la section 2.3 nous décrivons la méthode utilisée pour collecter des données dialectales convenables aux trois pays du Maghreb à partir de YouTube. Enfin, dans la section 2.4 nous présentons une analyse comparative à la lumière des trois corpus collectés. Dans cette étude nous abordons quelques spécificités liées au dialecte maghrébin comme le style d'écriture et le phénomène du code-switching.

2.2 La langue arabe

L'arabe est une langue sémitique millénaire, elle est utilisée comme langue officielle dans 22 pays arabes dont la majorité se trouve au Moyen-Orient. Elle est également la langue utilisée par de nombreux musulmans dans le monde. Elle occupe la 5ème position au classement des langues avec un nombre de locuteurs d'environ 422 millions [14]. L'arabe est également une des six langues des Nations Unies.

D'après [Farghaly and Shaalan, 2009], on distingue deux catégories principales : l'arabe littéraire et l'arabe dialectal. La première catégorie est la forme formelle et standard employée dans tous les pays arabes, tandis que, la deuxième fait référence aux variétés dialectales utilisées localement dans chaque pays, dans les conversations quotidiennes.

L'arabe littéraire regroupe deux formes : l'arabe classique et l'arabe standard moderne dénommé MSA (*Modern Standard Arabic*). L'arabe classique est la forme la plus ancienne de l'arabe qui existe seulement dans le Coran, les anciens ouvrages et les textes religieux. Cependant, l'arabe standard est la forme simplifiée et standardisée de l'arabe classique avec quelques modifications grammaticales. Elle est employée dans les communications officielles

14. Selon les statistiques publiées dans : http://www.axl.cefan.ulaval.ca/Langues/1div_inegalite.htm

orales ou écrites, dans l'administration et dans le monde éducatif.

Dans les paragraphes qui suivent, nous dressons quelques particularités de l'arabe standard et de l'arabe dialectal tout en mettant en exergue les principales difficultés relatives à leurs traitements dans le cadre de cette thèse.

2.2.1 L'arabe standard moderne

Cette forme officielle de l'arabe s'inspire fortement des concepts linguistiques de l'arabe classique. Dans ce qui suit, nous présentons quelques caractéristiques du MSA ayant un lien direct avec le travail que nous décrivons dans le cadre de cette thèse.

L'arabe est morphologiquement riche du fait des variations orthographiques et du phénomène d'agglutination. Elle s'écrit de droite à gauche, son système graphique est composé de 28 lettres (25 consonnes et 3 voyelles longues و, ي et ا). Chaque lettre peut avoir quatre écritures différentes selon sa position dans le mot (début, milieu, fin, isolé). Un exemple illustrant ce phénomène est donné dans le tableau 2.1 :

Position	**début**	**milieu**	**fin**	**isolé**
Les différentes formes de la lettre ع	عـ	ـعـ	ـع	ع
Exemple d'utilisation	علم	معلم	مصنع	يزرع
Traduction	drapeau	enseignant	usine	Il cultive

TABLE 2.1 – Les quatre représentations graphiques de la lettre ع selon sa position dans le mot.

Dans la première ligne de ce tableau, nous présentons un exemple d'écriture de la lettre arabe (ع) qui peut avoir plusieurs formes graphiques en fonction de sa position dans le mot : au début, au milieu, à la fin, et isolée. Ensuite, nous donnons des exemples de mots arabes contenant cette lettre et leurs traduction respectivement dans la 2ème et la 3ème ligne.

L'écriture arabe comporte aussi des voyelles courtes qui sont facultatives dans l'arabe moderne standard, parce que les arabophones restaurent intuitivement ces signes diacritiques. C'est pourquoi, en arabe standard, il est difficile de lire un texte sans en comprendre le contexte. Toutefois, l'absence de ces voyelles augmente le degré d'ambiguïté dans cette langue, car les formes non vocalisées de l'arabe peuvent avoir plusieurs significations. Par exemple le mot كتب dans le tableau 2.2 peut avoir des sens différents selon les voyelles utilisées.

En ajoutant les voyelles	كَتَبَ	كُتُبُ	كَتَّبَ
Traduction	écrire	livres	il a été écrit

TABLE 2.2 – Quelques sens possibles du mot كتب en rajoutant les voyelles.

En outre, le mot arabe peut être décomposé en cinq éléments : racine, antéfixe, préfixe, suffixe et postfixe. La racine constitue l'élément de base du mot arabe, elle est généralement composée de seulement trois consonnes, rarement quatre ou cinq. La concaténation de la racine avec les autres éléments (antéfixe, préfixe, suffixe et postfixe) permet de générer plusieurs autres formes dérivées. Le tableau 2.3 montre quelques dérivés que l'on peut produire en concaténant la racine كتب avec des antéfixe, préfixe, suffixe ou postfixe.

Les dérivés	كَتَبَ	كِتَاب	مَكتُوبِ	كِتَابه	كُتُبه
Traduction	écrire	livre	écrit	son livre	(ces livres)

TABLE 2.3 – Les dérivés possibles de la racine كتب.

L'association de l'ensemble des éléments précédents à une racine produit la forme la plus agglutinée en arabe. La figure qui suit donne un exemple illustratif de ce phénomène, le mot arabe ليعلموهم qui correspond à une phrase en langue française (*Pour leur apprendre*) est obtenue à partir de la combinaison de la racine علم, de l'antéfixe ل, du préfixe ي, du suffixe و et du postfixe هم comme le montre la figure 2.1.

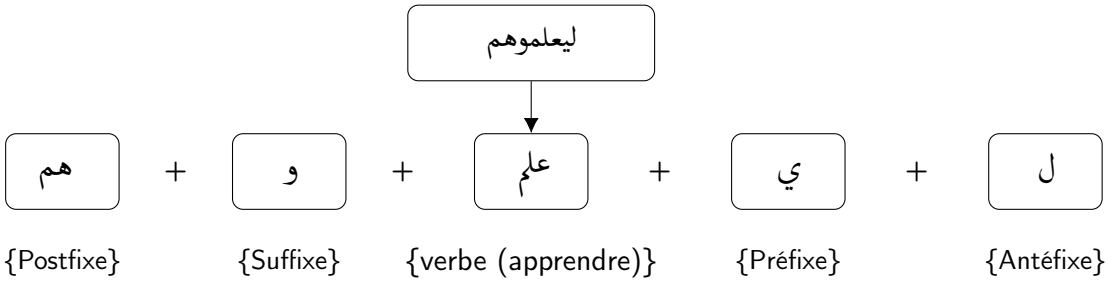


FIGURE 2.1 – La racine علم agglutinée avec les quatre affixes : l'antéfixe, le préfixe, le suffixe et le postfixe.

Par ailleurs, le mot arabe peut avoir des représentations graphiques différentes selon le genre (masculin et féminin). À titre d'exemple, le mot معلمون (*enseignants*) correspond au pluriel masculin tandis que معلمَات (*enseignantes*) correspond au féminin pluriel. En plus des formes singulière et plurielle, les mots arabes peuvent avoir une autre forme qui se situe entre les deux. Cette forme est appelée le duel, qui fait référence à un cas particulier du pluriel, lorsque le nombre est deux, par exemple معلمتَان (*deux enseignantes*), منزلَان (*deux maisons*).

On notera aussi que, la langue arabe possède également d'autres spécificités au niveau syntaxique et phonétique, qu'on n'aborde pas ici, car ce n'est pas l'intérêt de cette thèse.

2.2.2 L'arabe dialectal

Bien que l'arabe moderne standard soit la langue officielle dans le monde arabe, elle ne constitue la langue maternelle d'aucun peuple de ce monde. Chaque pays a développé

un dialecte qu'il utilise dans la communication informelle. Par conséquent, cette forme vernaculaire de l'arabe est restée limitée géographiquement à chaque région pour un pays donné.

Dans ce travail, nous nous sommes intéressés aux dialectes de trois pays du Maghreb : l'Algérie, le Maroc et la Tunisie. Les dialectes parlés dans ces pays sont à peu près similaires avec quelques variations liées à l'histoire et à la position géographique de chacun.

Les dialectes maghrébins se différencient de l'arabe standard du point de vue : phonologique, morphologique, lexical et syntaxique. Dans ce qui suit, nous expliquons brièvement les particularités de ces dialectes et les principales difficultés à leurs traitements. À la fin de la section nous proposons un tableau comparatif qui illustre la différence entre l'arabe et l'arabe dialectal sur plusieurs plans : l'utilisation, la phonologie, la morphologie, le lexique et la syntaxe.

Les dialectes maghrébins ont une morphologie moins complexe que l'arabe standard, car certains modes de flexion ont disparu dans ces dialectes. Par exemple, l'arabe standard a une forme duale pour désigner un ensemble de deux choses ou de deux personnes (voir la section précédente 2.2.1) et des formes singulières et plurielles, alors que dans les dialectes, le duel (masculin et féminin) et le pluriel féminin (pour la conjugaison des verbes) ont disparu. Par conséquent, les maghrébins dans leurs dialectes, ajoutent les mots (زوز,زوج ou جوج) qui signifient "deux" pour exprimer le duel. En outre, pour certains dialectes du Maghreb, les gens ne font pas la distinction de genre du pluriel ou du singulier.

Pour des raisons historiques, le dialecte maghrébin a été influencé par les anciennes langues locales comme l'amazigh et certaines langues étrangères notamment le français, l'espagnol, l'anglais, l'italien et le turc. Cela a conduit à un nombre considérable d'emprunts lexicaux dans ces dialectes. Dans le tableau 2.4, nous présentons quelques exemples des mots empruntés dans les trois dialectes.

	Mot	**Dialecte**	**Écriture Originale**	**Traduction**	**Origine d'emprunt**
1	بلَاك (balak)	algérien	bālāk	peut-être	turc
2	بوقَادو (bogado)	algérien	abogado	avocat	espagnol
3	مكينة (mekina)	tunisien	mekinâ	machine	italien
4	طوبيس (toubis)	marocain	autobus	autobus	français
5	كوميرة (comira)	marocain	comer	manger	espagnol

TABLE 2.4 – Exemples des mots empruntés dans les trois dialectes.

L'emprunt dans les dialectes maghrébins s'articule généralement autour de plusieurs types. Un emprunt qu'on peut qualifier de "direct", qui consiste à utiliser le sens et la forme du mot étranger avec une légère adaptation phonétique ou orthographique comme le montrent les trois premiers exemples du tableau 2.4. Un emprunt sémantique qui garde le sens du mot et change la forme comme le montre le $4^{\text{ème}}$ exemple. Un autre type emprunt existe ainsi dans ces dialectes qui consiste à garder la forme et changer le sens comme dans le $5^{\text{ème}}$ exemple.

Par ailleurs, les dialectes maghrébins, comme toute autre langue, se développent et s'adaptent à chaque époque. Néanmoins, les dialectes empruntent fréquemment des mots des autres langues. Il arrive aussi souvent que l'on crée de nouveaux mots. Dans le tableau 2.5, nous présentons quelques nouveaux mots ajoutés récemment au dialecte algérien après l'apparition des réseaux sociaux avec la signification de chacun.

Mot	**Traduction**
جمجمولي *ğmğmwly*	Mettez moi des "j'aime"
ابونيلي *abwnyly*	Abonnez-vous
ديسلايكات *dyslāykāt*	Des aversions

TABLE 2.5 – Quelques emprunts récents dans le dialecte algérien.

Jusqu'à un passé récent, les dialectes maghrébins ne s'écrivaient pas, mais l'apparition des réseaux sociaux a fait passer cette forme parlée à l'écriture. De ce fait, les gens commencent à écrire les mots sans aucune contrainte linguistique, ce qui conduit à l'apparition de plusieurs graphies juste pour un seul mot. Citons à titre d'exemple le mot du dialecte algérien مانسوطيش *mānswṭyš* (*Je ne saute pas*), ce mot peut avoir plusieurs graphies dépendant de la prononciation des personnes : (مانسوتيش, منسوطيش, منصوطيش, منسوتيش, etc).

En outre, les mots du dialecte sont écrits en utilisant les caractères arabes et latins. Pour illustrer ce phénomène, revenons à l'exemple précédent مانسوطيش. Il peut être écrit également en caractères latins :

mansotich, mansotiiche, mansautich, mnsotich, mansoutich, mansotiche, mansotiwch, manesotich, manssotiche, mansoutiche, manssotich, etc.

Ce phénomène d'écriture du dialecte arabe avec le script latin est connu sous le nom d'arabizi par la communauté travaillant sur le traitement automatique de la langue arabe [Darwish, 2013]. L'utilisation de l'arabizi est due à plusieurs raisons, citons entre autres le peu de claviers arabes et particulièrement dans les smartphones. Par ailleurs, l'écriture en utilisant le script latin facilite l'utilisation de mots étrangers dans un message.

Évidement, adapter un système d'une langue latine, qui est graphiquement et phonétiquement différent de celui de l'arabe pour écrire dans cette langue peut poser quelques difficultés. Par conséquent, l'inexistence de lettres latines représentant certains sons arabes (غ, خ , ع et ش) a engendré la création de plusieurs notations non standardisées pour remédier à ce problème. Pour certains sons manquants, une combinaison de lettres latines dont le son résultant s'approche plus ou moins du son arabe est proposée. Par exemple, les trois lettres arabes غ, خ et ش peuvent être représentées respectivement comme suit (*gh*), (*kh*) et (*ch* ou *sh*). Dans d'autres cas, le son manquant est remplacé par un chiffre dont la graphie ressemble généralement à la lettre d'origine en arabe. Citons à titre d'exemple ع, qui est remplacé par 3 ; ح par 7 et ق par 9, etc.

Pour illustrer nos propos, dans le tableau 2.6 nous donnons un exemple de commentaire en dialecte marocain extrait de YouTube avec son équivalent en arabizi. Nous donnons

également le script arabizi utilisant des chiffres. Il est important de noter que, cette écriture n'est pas standard, pour le même commentaire arabe, on peut avoir plusieurs écritures différentes.

En script arabe	عفاك عندي مشكل فترتيب حقايب دياولي عفاك شي فكره
En arabizi	aafak aandi mochkil fitartib lhakayab dyawli aafak chifekra
En arabizi avec chiffres	3afak 3andi mochkil fitartib l7a9ayab dyawli 3afak chifekra
Traduction	S'il te plaît, j'ai un problème de rangement de mes valises, as-tu une idée stp.

TABLE 2.6 – Des écritures possibles dans le dialecte marocain.

Cette écriture présente un problème difficile pour le traitement automatique du dialecte, car la majorité des outils développés pour le dialecte sont dédiés aux données écrites en lettres arabes. Par conséquent, l'arabizi ne peut pas être traité avec ces outils.

Comme mentionné plus haut, dans les pays du Maghreb coexistent plusieurs langues, l'arabe dialectal, l'arabe moderne, le français et récemment l'anglais. Cela a conduit à l'apparition d'un phénomène connu sous le terme de code-switching appelé aussi alternance ou mélange codique. Ce phénomène consiste en l'utilisation simultanée de deux ou plusieurs langues dans une même conversation ou dans le même segment textuel. Le code-switching dans la conversation maghrébine peut concerner un mot isolé ou plusieurs mots consécutifs comme le montrent les deux commentaires extraits de YouTube pour les dialectes algérien et marocain :

ALG : Verry simple for you *on veut toujours les ingrédients* **okhti** *merci.*
Traduction : "*Cela paraît simple pour vous, ma soeur, on aimerait bien avoir les ingrédients, merci".*

MAR : : تبارك الله عليك *merci pour la vidéo* عجبني *les boucles d'oreilles* منين خديتهم **thanks.**
Traduction : "*Que Dieu vous bénisse, merci pour cette vidéo, j'ai bien aimé les boucles d'oreilles vous les avez achetées où. Merci.*"

Afin de faciliter la lecture de ces exemples, nous changeons de style d'écriture, en passant du gras à l'italique et vice-versa, à chaque fois qu'une nouvelle langue est utilisée. Dans le commentaire algérien (ALG), entre le segment écrit en anglais et celui écrit en français, nous remarquons que l'usager a utilisé le mot arabe (**okhti**) écrit en caractères latins. Dans l'exemple marocain (MAR), l'utilisateur est passé plusieurs fois du dialecte écrit en caractères arabes vers le français avant de terminer le commentaire par un mot en anglais.

D'après ces exemples, nous constatons que malgré la brièveté de ces commentaires, les maghrébins alternent entre les langues à plusieurs endroits dans un même commentaire. Dans ce qui suit, nous appellerons ces emplacements : les points de rupture de la langue.

Le système phonologique des dialectes maghrébins est plus large que l'arabe standard, car ce système inclut en plus des phonèmes de l'arabe standard, des phonèmes empruntés de

langues latines comme : /p/, /g/ et /v/.

Le tableau 2.7 présente les différences entre l'arabe standard et les dialectes maghrébins selon plusieurs aspects : phonologique, syntaxique, lexical, etc.

	Dialecte maghrébin	**L'arabe moderne standard**
Statut	— La forme vernaculaire de l'arabe utilisée dans le quotidien (conversation orale au sein de la communauté).	— La forme formelle de l'arabe qui est employée dans les communications officielles orales et écrites.
Phonologique	— Possède les mêmes phonèmes que ceux du MSA en plus d'autres phonèmes des langues latines : /g/,/p/ et /v/. — Prononciation variable en fonction des usagers.	— Possède une trentaine de phonèmes en plus de quelques voyelles. — Prononciation uniforme.
Lexical	— Une partie importante du vocabulaire du dialecte est inspirée de celui du MSA. Certains mots lorsqu'ils sont transférés vers le dialecte, changent de sens. — Pas de règles grammaticales et un manque de dictionnaires. — Vocabulaire limité à une région donnée.	— Un vocabulaire standard et riche qui obéit à des règles strictes. — Un vocabulaire commun employé par tous les arabophones. — L'existence de dictionnaires.
Morphologique	— La disparition de plusieurs modes comme le duel et le féminin pluriel. — Un système d'affixation compliqué. — La négation est affixée, elle est souvent exprimée avec les affixes suivants : ش+mot+ما (dans les trois dialectes).	— Une morphologie plus riche que celle des dialectes. — Un système de cliticisation moins complexe que dans le dialecte. — La négation est construite avec des particules non agglutinées au mot.
Syntaxique	— Les phrases sont simples et courtes. — L'adoption en général de la forme active. — Complexité syntaxique due au phénomène du code-switching.	— Structures syntaxiques riches comme pour les autres langues naturelles. — Utilisation de la forme active et passive. — Le code-switching est beaucoup moins fréquent que dans le dialecte.

TABLE 2.7 – Comparaison entre l'arabe et les dialectes maghrébins.

2.3 Sélection et collecte des données dialectales

Développer des systèmes efficaces permettant le traitement des dialectes maghrébins nécessite de disposer de corpus en grande quantité. Pour ce faire, nous avons décidé de les collecter à partir des réseaux sociaux. En effet, jusqu'à un passé récent le dialecte arabe ne s'écrivait pas, mais les réseaux sociaux ont produit une masse importante de données dialectales.

YouTube est parmi les plate-formes les plus fréquentés dans le monde entier et en particulier au Maghreb. Nous avons donc utilisé cette plate-forme pour collecter des données qui répondent à notre problématique. Pour ce faire, nous avons porté une attention particulière aux critères suivants :

— Extraire de YouTube des commentaires du dialecte approprié.

— Collecter des corpus volumineux afin d'entraîner nos modèles fondés sur les réseaux de neurones que nous proposons plus loin dans ce manuscrit.

Le problème est qu'il n'existe pas de méthode standard pour savoir comment sélectionner les commentaires correspondant au dialecte de chaque pays. C'est pourquoi, nous avons opté pour une méthode qui consiste à dresser au préalable une liste de mots-clés relatifs à chaque pays. Le fait de choisir des mots-clés spécifiques correspondant à des évènements ou à des personnalités connus principalement par des habitants d'un pays peut nous assurer de collecter une grande quantité de données correspondant au dialecte recherché. Nous n'avons sélectionné que les personnages ou les célébrités connus localement pour ne pas tomber sur d'autres dialectes. Dans le tableau 2.8, nous donnons quelques exemples de mots-clés utilisés liés à chaque dialecte (pour plus de détails, voir annexe A).

Algérien	شمس الدين, Zanga Crazy, زروطة, Dzjoker
Marocain	سكينة درابيل, Laila Hadioui, حسن الفد
Tunisien	نجلاء تونسية, Alaa Chabi, Ghannouchi

Table 2.8 – Exemple de mots-clés utilisés pour la recherche des vidéos.

En utilisant l'API [15] de Google nous avons collecté trois corpus correspondant à trois dialectes du Maghreb. Ensuite, un processus de filtrage et de nettoyage est lancé pour supprimer les commentaires courts, ceux contenant uniquement des liens URLs, etc. Le tableau 2.9 donne quelques statistiques sur les corpus collectés, où $|C|$ est le nombre de commentaires, $|W|$ est le nombre de mots et $|V|$ est le nombre de mots distincts.

	Algérien	**Marocain**	**Tunisien**
$\|C\|$	1,1M	1,6M	1,2M
$\|W\|$	18,3M	22,1M	16M
$\|V\|$	1M	1,3M	1M

Table 2.9 – Quelques statistiques sur les corpus collectés.

15. Disponible sur : https ://developers.google.com/YouTube

2.4 Étude analytique sur les trois corpus maghrébins

Dans ce qui suit, nous présentons une étude analytique détaillée concernant les trois corpus collectés précédemment. L'objectif principal de cette étude est d'avoir une idée sur l'utilisation des dialectes maghrébins dans les réseaux sociaux en particulier dans YouTube. En effet, dans cette étude nous avons accordé une attention au style d'écriture, langue dominante dans ces corpus, ainsi qu'au phénomène du code-switching.

2.4.1 L'écriture du dialecte maghrébin dans les réseaux sociaux

Comme nous l'avons évoqué auparavant, les maghrébins font partie des communautés qui mélangent plusieurs langues dans leurs discours : l'arabe standard, le dialecte, le français et l'anglais. De ce fait, les commentaires peuvent être écrits dans toutes ces langues ou bien ces langues peuvent être mélangées dans une même phrase.

Comme nous l'avons évoqué précédemment, dans les réseaux sociaux il est possible d'écrire l'arabe en script latin comme illustré par les diagrammes en secteurs de la figure 2.2. Nous présentons ainsi le pourcentage de commentaires écrits en Script Latin (SL), ceux écrits en Script Arabe (SA) et enfin ceux écrits en Script Mixte (SM). Ces derniers correspondent aux commentaires dans lesquels on trouve les deux scripts SL et SA en même temps. Il est important de noter que les textes écrits en script latin peuvent correspondre à des commentaires en français ou en anglais comme ils peuvent correspondre à des textes écrits en arabizi (arabe écrit en caractères latins).

Ces chiffres montrent clairement que la façon d'écrire dans les réseaux sociaux dans les trois pays du Maghreb est similaire. Cette analyse montre que dans les réseaux sociaux, on préfère l'utilisation du script latin pour poster ses commentaires. Ainsi dans le dialecte tunisien 53% des commentaires sont postés en SL. Alors que pour les dialectes algérien et marocain les taux sont respectivement de 46% et 45%. Nous constatons que, les textes dans lesquels il y a le plus de mélange de script (SM) se retrouvent plutôt dans le dialecte marocain.

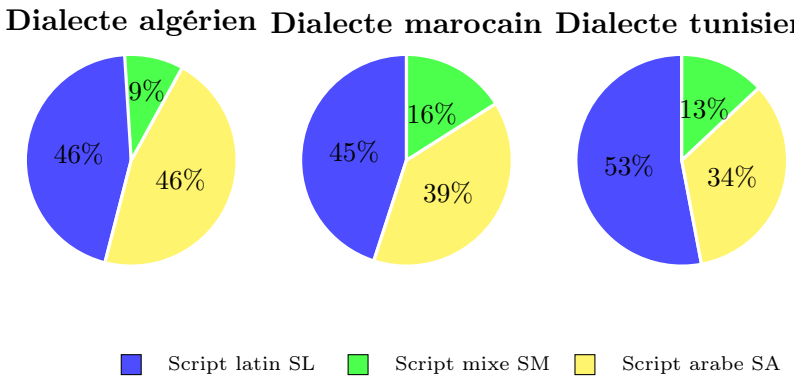


FIGURE 2.2 – La répartition des commentaires en fonction du script dans les trois corpus.

Dans le tableaux 2.10, nous donnons les pourcentages de commentaires en français en anglais et en arabizi.

Corpus (SL)	arabizi	français	anglais
algérien	83%	**14%**	3%
marocain	**93%**	5%	2%
tunisien	87%	7%	**6%**

TABLE 2.10 – La distribution du français, de l'anglais et de l'arabizi dans la partie SL des trois corpus collectés.

Pour établir ces statistiques, il faut réussir à identifier un commentaire en français ou en anglais du commentaire en arabizi. Pour ce faire, un texte est considéré comme un texte en français si chacun de ses mots est écrit en français. Idem pour l'anglais. Un texte est considéré comme un commentaire en arabizi s'il existe au moins un mot qui n'est pas en français ni en anglais. Malheureusement, cette règle ne permet pas de récupérer des commentaires exclusivement en dialecte. En effet, l'application de la règle précédente peut renvoyer des phrases en français ou en anglais dans lesquelles certains mots sont mal orthographiés. Pour savoir si un mot est un mot français ou anglais, on fait appel à des dictionnaires extraits d'une table de traduction français-anglais. Cette table comporte 6 millions d'entrées [Ameur et al., 2016a].

Le tableau 2.10 montre que le pourcentage d'arabizi est très élevé dans les trois corpus. Le corpus marocain est celui dans lequel l'arabizi est le plus largement utilisé. Alors que, le corpus algérien est celui dans lequel la langue française est la plus utilisée.

Par ailleurs, dans le tableau 2.11 nous donnons le pourcentage de commentaires écrits en arabe standard (MSA) et en dialecte.

Corpus(SA)	Dialecte	MSA
algérien	**63%**	37%
marocain	30%	**70%**
tunisien	**64%**	36%

TABLE 2.11 – La distribution du MSA et du dialecte dans la partie écrite en script arabe des trois corpus.

Un commentaire est considéré comme un commentaire MSA si chacun de ses mots existe dans un dictionnaire arabe. Le dictionnaire que nous avons utilisé contient 9 millions d'entrées [Menacer et al., 2017a]. Le tableau 2.11 montre que le corpus marocain contient un nombre important de commentaires en MSA par rapport aux deux autres corpus. Nous pouvons également mentionner que les distributions dialecte-MSA pour les dialectes tunisien et algérien sont similaires.

2.4.2 L'utilisation du code-switching dans les dialectes maghrébins

Le code-switching consiste à glisser des segments de mots en langue étrangère dans une conversation qui est généralement en langue maternelle. Ces segments peuvent concerner un ou plusieurs mots qui peuvent être placés à des endroits différents dans une même phrase. Afin d'avoir une idée précise sur le nombre de mots insérés dans ces corpus, nous présentons dans la figure 2.3, le taux de segments en langue étrangère, dans chaque corpus, concernés

par le phénomène du code-switching, en fonction du nombre de mots par segment. Pour ce faire, nous avons identifié les segments anglais et français de chaque corpus et nous les avons comptés en fonction de leurs longueur. Ne sont comptabilisés que les segments pour lesquels il y a une rupture langagière par rapport au dialecte ou au MSA. On peut remarquer que le nombre de ces segments est plus élevé dans le corpus algérien comparativement aux corpus marocain et tunisien, et ce, quelle que soit la taille du segment. Cette analyse, nous montre également que les usagers utilisent peu de segments longs servant au code-switching.

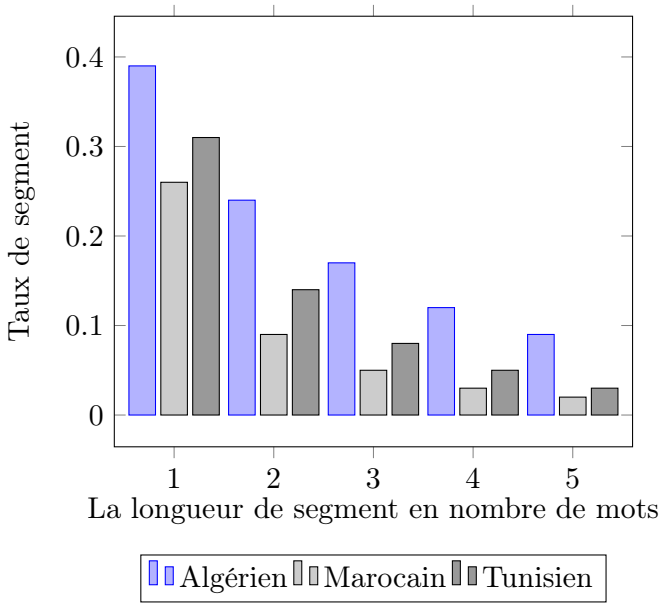


FIGURE 2.3 – Pourcentage de segments utilisés pendant le code-switching dans les trois corpus en fonction du nombre de mots.

Dans le tableau 2.12, à titre d'exemple, nous présentons les cinq segments les plus utilisés comportant trois mots et concernés par le code-switching.

Algérien		**Marocain**		**Tunisien**	
Segment	$Nb-Occ$	Segment	$Nb-Occ$	Segment	$Nb-Occ$
très belle chanson	342	merci ma belle	621	vive la tunisie	541
top top top	238	top top top	382	participation du jour	422
moi je suis	227	merci pour la	376	je participe et	328
et bonne continuation	209	vive le maroc	338	merci ma belle	264
est ce que	204	merci ma cherie	322	the good work	227

TABLE 2.12 – Les cinq segments les plus utilisés en anglais et en français de longueur trois dans les trois corpus maghrébins.

2.5 Conclusion

Dans ce chapitre nous avons présenté l'arabe standard (MSA) et l'arabe dialectal et nous nous sommes concentrés plus particulièrement sur les dialectes utilisés au Maghreb. Nous

avons abordé les particularités de ces deux formes tout en mettant en exergue les problèmes relatifs à leur traitement dans le cadre de notre travail. Nous avons présenté les différences entre l'arabe standard et l'arabe dialectal au niveau phonologique, lexical, morphologique, etc. Nous avons proposé également une procédure pour collecter à partir de YouTube des corpus pour chacun des trois dialectes étudiés du Maghreb : algérien, marocain et tunisien. Ensuite, nous avons effectué une étude analytique des trois corpus pour comprendre la façon dont on écrit dans les réseaux sociaux. Cette étude montre qu'au pays du Maghreb, les utilisateurs d'Internet préfèrent utiliser leurs dialectes pour poster leurs commentaires et ils choisissent souvent le script latin.

Nous avons abordé aussi une des caractéristiques importantes du dialecte arabe, il s'agit du code-swhitching ou l'alternance codique. En effet, ce phénomène est très répandu dans les réseaux sociaux. Nous avons pu constater, par exemple, que pour le dialecte algérien, le nombre de segments utilisés en français et en anglais de longueur supérieure à 5 est de respectivement 952k, 361k et 394k pour les corpus algérien, marocain et tunisien. Par conséquent, les corpus que nous avons collectés ne sont pas entièrement en dialecte. Ce résultat oblige à poser une question importante : existe-t-il un moyen permettant la sélection de corpus composés que de mots du dialecte, à partir des réseaux sociaux ?

Répondre cette question est le sujet du chapitre suivant.

3

Mesurer le code-switching

Sommaire

3.1 Introduction 42
3.2 Les approches pour mesurer le degré de code-switching 43
3.2.1 Facteur de complexité de Ghosh 43
3.2.2 CESAR : nouvelle méthode pour mesurer le code-switching 44
3.3 Protocole d'évaluation 46
3.3.1 Étude comparative des deux métriques sur des exemples simples . . . 46
3.3.2 Évaluation de CESAR et Ghosh 47
3.3.3 Expérimentation sur les dialectes maghrébins 49
3.4 Conclusion 50

Ce chapitre a pour objectif de répondre à la question soulevée précédemment qui concerne la possibilité de développer une mesure qui nous permette de vérifier si les corpus que nous avons collectés sont composés majoritairement par des commentaires en dialecte ou non. De ce fait, dans ce chapitre, nous proposons une métrique pour quantifier le bruit dû au code-switching dans un corpus relativement à une langue donnée. L'idée est de récolter un corpus qui soit le plus pur possible en termes de contenu relativement à une langue de référence. La mesure est nommée CESAR (CodE-Switching According to a Reference language).

3.1 Introduction

Comme nous l'avons déjà évoqué dans le chapitre précédent le processus de code-switching consiste à glisser des segments de mots en langue étrangère dans une conversation (orale ou écrite) qui est à la base dans une langue différente. Ce phénomène a attiré l'attention des psycholinguistes pendant de nombreuses années [Joshi, 1982], [Auer, 1999], [Gafaranga and Torras, 2002], [Tawwab and Eldin, 2014], et [Redouan, 2005]. Ils ont étudié, entre autres, les raisons qui poussent les individus à vouloir migrer d'une langue à une autre dans une même conversation. Ils montrent que cela peut être dû à un manque de vocabulaire technique ou non, ou à une maîtrise partielle d'une langue, etc.

Récemment, ce problème a intéressé également la communauté travaillant sur le traitement automatique de la langue (TAL). Par conséquent, plusieurs travaux de recherche tentent de constituer des corpus code-switchés, de mesurer le degré de code-switching dans un document, d'identifier des points de rupture de la langue dans un texte et de proposer des solutions et des modèles à d'autres phénomènes liés au code-switching.

Le code-switching concerne principalement les communautés parlant au moins deux langues. Ce phénomène est plus fréquent dans la communication informelle (par exemple dans les réseaux sociaux) que dans la communication formelle. Au Maghreb, le code-switching est un phénomène très répandu. En effet, on utilise l'arabe standard comme une langue officielle, l'arabe dialectal pour communiquer dans la vie quotidienne, le français pour des raisons historiques et l'anglais à cause de l'émigration récente vers les pays anglophones.

Dans le chapitre précédent (voir section 2.4.2) nous avons montré à travers une étude analytique sur nos corpus, la façon dont les langues sont combinées dans les réseaux sociaux. L'étude menée a montré que dans le code-switching on utilise non seulement de simples mots issus d'autres langues, mais aussi des structures plus longues.

Dans ce chapitre, nous proposons une nouvelle métrique appelée CESAR (*CodE-Switching According to a Reference language*) qui consiste à mesurer le code-switching dans un corpus par rapport à une langue de référence donnée. Autrement dit, cette mesure évalue la quantité de bruit dans un corpus par rapport à une langue, que nous appelons ici langue de référence. Dans notre cas, nous avons choisi le dialecte comme une langue de référence. CESAR est une métrique bornée, qui donne une valeur égale à 0 si le corpus est entièrement écrit dans la langue de référence (le dialecte dans notre cas) et une valeur égale à 1 si le corpus est entièrement écrit dans une langue différente de la langue de référence. Cette mesure a l'avantage de permettre l'extraction de corpus contenant un nombre réduit de mots étrangers à la langue de référence selon l'instanciation de la valeur de CESAR. Par exemple, un seuil de CESAR fixé à 0 ne récupérera que des phrases dans la langue de référence ce qui permettra pour des langues peu dotées de construire un lexique des mots typiques de cette langue.

Dans ce chapitre, nous détaillons une des mesures d'évaluation du niveau de code-switching dans un texte proposée récemment par [Ghosh et al., 2017] et nous présentons ensuite notre proposition de mesure appelée CESAR. Nous proposons par la suite une étude comparative de ces deux métriques pour montrer que les mesures existantes notamment celle de Ghosh n'est pas adaptée à notre problématique. Puis, nous réalisons un ensemble d'expérimentations sur les trois corpus que nous avons collectés à partir de YouTube.

3.2 Les approches pour mesurer le degré de code-switching

Comme nous l'avons mentionné dans le chapitre 1, de nombreux travaux de recherche ont tenté de mesurer le code-switching que ce soit au niveau du document ou du corpus. Dans ce travail, notre objectif est différent de ces travaux, car nous tentons à travers la mesure que nous avons proposée de mesurer le bruit produit par le phénomène de code-switching en prenant en considération la langue de référence dans lequel le document est rédigé. Afin de montrer la particularité de notre mesure, nous la comparons avec une mesure proposée récemment par Ghosh [Ghosh et al., 2017]. Nous réalisons plusieurs expérimentations dans ce chapitre sur ces deux métriques pour montrer l'inadéquation des mesures existantes par rapport à l'objectif que l'on s'est fixé.

3.2.1 Facteur de complexité de Ghosh

Dans [Ghosh et al., 2017], les auteurs définissent un facteur de complexité permettant de mesurer la complexité multilingue d'un corpus. Ce que nous avons appelé une mesure de code-switching. Le facteur de complexité proposé par Ghosh est parmi les tentatives récentes pour quantifier le degré de combinaison des langues existant dans un document ou dans un corpus. Ce facteur est fondé principalement sur la mesure proposée par [Das and Gambäck, 2014]. Il prend en considération plusieurs paramètres lors de l'analyse d'un texte : *Languages Factor* (LF), *Switching Factor* (SF) et *Mix Factor* (MF). La combinaison de ces trois facteurs est utilisée afin d'estimer le degré de code-switching selon l'équation suivante :

$$\mathbf{CF} = \frac{a * MF + b * SF}{LF} \tag{3.1}$$

Où a et b sont des poids pour les deux facteurs MF et SF, qui sont fixés à 50.

Dans les paragraphes qui suivent, nous détaillons chacun de ces facteurs séparément.

Languages Factor (LF)

Ce facteur dépend du nombre de langues utilisées dans une phrase. LF est obtenu en divisant le nombre de langues distinctes dans la phrase par le nombre total de mots de la phrase.

$$LF = \frac{W}{N} \tag{3.2}$$

W le nombre de mots de la phrase et N est est le nombre de langues distinctes utilisées dans la phrase.

Switching Factor (SF)

Ce facteur correspond au nombre de fois où l'on change de langue dans une phrase. L'endroit de changement de langue est nommé : point de rupture de la langue. Par conséquent, changer plusieurs fois de langue augmente la complexité de traitement de la phrase. Pour une phrase, ce facteur est obtenu par la division du nombre de points de rupture sur le nombre

de mots.

$$SF = \frac{S}{W-1}, \quad \text{if} \quad W > 1$$
$$SF = 0, \quad \text{if} \quad W = 1 \tag{3.3}$$

S est le nombre de points de rupture dans la phrase et W est le nombre de mots de la phrase.

Mix Factor (MF)

C'est le ratio entre le nombre de mots qui ne sont pas écrits dans la langue dominante de la phrase sur le nombre de mots des langues distinctes utilisées dans la phrase.

$$MF = \frac{W' - max\{w\}}{W'}, \quad \text{if} \quad W' > 0$$
$$MF = 0, \quad \text{if} \quad W' = 0 \tag{3.4}$$

W' est le nombre de mots des langues distinctes et $max\{w\}$ est le nombre de mots appartenant à la langue la plus fréquente dans la phrase.

À noter aussi que les trois facteurs sont applicables au niveau de la phrase, du paragraphe ou du document.

Les auteurs de [Ghosh et al., 2017] ont proposé deux autres variantes du facteur CF en l'occurrence $CF2$ et $CF3$. La différence se situe au niveau du calcul de LF qui dans un cas utilise une fonction linéaire et dans l'autre une fonction géométrique comme suit :

$$\mathbf{CF2} = \frac{a * MF + b * SF}{\frac{0.25}{W-1}(LF-1)+1} \tag{3.5}$$

$$\mathbf{CF3} = \frac{a * MF + b * SF}{\frac{arctan(LF)}{\pi} + 0.75} \tag{3.6}$$

D'après [Ghosh et al., 2017] $CF2$ et $CF3$ sont plus pertinents et mesurent mieux la complexité multilingue dans un corpus.

3.2.2 CESAR : nouvelle méthode pour mesurer le code-switching

Le facteur de complexité de Ghosh présenté dans la section précédente ne correspond pas à nos objectifs, car il mesure le code-switching dans un corpus indépendamment de toute langue. Partant de ce fait, dans cette section nous proposons une métrique afin de mesurer le code-switching dans les trois corpus des dialectes maghrébins que nous avons collectés dans le chapitre 2. Cette métrique est nommée CESAR (*CodE-Switching According to a Reference language*), elle prend en considération, pendant le calcul, la langue de référence. Cela signifie que si un commentaire est rédigé complètement en dialecte (langue de référence), CESAR lui affecte la valeur 0 et plus le nombre de langues utilisées dans le commentaire augmente, plus le score s'accroît et se rapproche de 1. Si le commentaire est complètement en langue étrangère (différente de la langue de référence) CESAR affecte la valeur 1 au commentaire. L'évaluation du niveau de code-switching dans un commentaire est fondée sur

deux paramètres Pr et Br.

$P_r(C)$ désigne la précision du commentaire relativement à la langue de référence r. Autrement dit, ce facteur mesure à quel point le commentaire est mixé de mots étrangers à la langue de référence.

$$P_r(C) = \frac{1}{n}\sum_{i=1}^{n}\delta(d_i)LF(d_i) \tag{3.7}$$

Où $\delta(d_i)$ est défini comme suit :

$$\delta(d_i) = \begin{cases} 1 \; si \; \exists w \in d_i \;\; \text{tel que} \;\;\; L(w) \neq L_r \\ 0 \; sinon \end{cases} \tag{3.8}$$

Où $\delta(d_i)$ affecte la valeur 1 à d_i si au moins un point de rupture de langue existe dans le document d_i.

— $L(x)$ est la langue du mot x.

— d_i est un document du corpus C.

— n est le nombre de documents de C.

— L_r est la langue de référence.

$LF(d_i)$ mesure la proportion des langues utilisées dans d_i.

Pour résumer, $P_r(C)$ vaut 0 si le corpus C est entièrement écrit dans la langue de référence et vaut 1 si le corpus ne comporte aucun mot de la langue de référence. Entre les deux valeurs, $P_r(C)$ indique le nombre moyen de langues utilisées dans le corpus C.

Quant à $B_r(C)$, il permet de quantifier le bruit introduit par les mots différents de ceux de la langue de référence.

$$B_r(C) = \frac{1}{n}\sum_{i=1}^{n}\frac{N(w_{d_i})}{N_{d_i}}LF(d_i) \;\; \text{où } L(w_{d_i}) \neq L_r \tag{3.9}$$

— $N(w_{d_i})$ est le nombre de mots du document d_i dont la langue est différente de la langue de référence.

— N_{d_i} est le nombre total de mots du document d_i.

Enfin, la mesure CESAR permet d'estimer le changement de code dans un corpus C conformément à la langue de référence r, à travers la formule suivante :

$$CESAR(C) = \alpha P_r(C) + \beta B_r(C) \tag{3.10}$$

Avec α et β les poids assignés à P_r et B_r, ils sont déterminés empiriquement en respectant la contrainte $\alpha + \beta = 1$.

3.3 Protocole d'évaluation

Afin de montrer la spécificité de CESAR par rapport aux autres mesures existantes et notamment les facteurs de complexité proposés par Ghosh [Ghosh et al., 2017], nous avons réalisé un ensemble d'expériences sur plusieurs corpus. Ces corpus sont : ceux que nous avons collectés à partir de YouTube et ceux que nous avons confectionnés manuellement. Dans les sections qui suivent, nous proposons un ensemble d'expérimentations pour évaluer et comparer les deux métriques : CESAR et les variantes de la mesure de complexité multilingue proposées par Ghosh.

3.3.1 Étude comparative des deux métriques sur des exemples simples

Nous présentons à travers trois exemples simples S_1, S_2, S_3 et S_4 les valeurs de CESAR et des mesures proposées par Ghosh. L'idée étant de montrer l'inadéquation des mesures de complexité multilingues de Ghosh par rapport à notre objectif.

Nombre	phrase	CF	CF2	CF3	CESAR
S_1	$x_1^a x_2^a x_3^a x_4^a x_5^a$	0	0	0	1
S_2	$x_1^r x_2^r x_3^r x_4^r x_5^r$	0	0	0	0
S_3	$x_1^r x_2^e x_3^a x_4^r x_5^r$	34,5	55,19	53,34	0,29
S_4	$x_1^r x_2^e x_3^a x_4^r x_5^r x_1^r x_2^e x_3^a x_4^r x_5^r$	15,99	50,08	46,08	0,43

TABLE 3.1 – Les scores de code-switching obtenus avec les facteurs de Ghosh et CESAR sur des exemples.

La deuxième colonne du tableau 3.1 présente les exemples sur lesquels nous calculerons les scores du code-switching. Chaque mot x_i^l est écrit dans une langue l où $l \in$ {a : arabe, r : référence et e : anglais}.

CF, CF_2, CF_3 et CESAR correspondent aux valeurs de code-switching obtenues par les facteurs de complexité de Ghosh, et de celle de CESAR.

Pour la phrase S_1, on remarquera qu'il n'y a aucun mot de la langue de référence, c'est pourquoi CESAR lui affecte la valeur maximale de code-switching, tandis que les trois facteurs de Ghosh lui attribuent la valeur minimale qui correspond à une absence de code-swotching.

Pour S_2, la phrase est entièrement écrite dans la langue de référence. CESAR lui attribue la valeur minimum, alors que Ghosh lui attribue exactement les mêmes valeurs que pour la phrase S_1. En effet, pour CESAR cette phrase est intéressante puisqu'elle est composée que de mots de la langue de référence. Alors que pour les mesures de Ghosh, elles ne font aucune différence entre S_2 et S_1 puisqu'elles ne sont pas fondées sur cette notion de langue de référence.

Dans l'exemple S_3, nous avons une phrase comportant trois points de rupture de langue. CESAR lui attribue une valeur faible parce que la langue dominante dans cette phrase est la langue de référence. Tandis que Ghosh lui donne des valeurs difficilement compréhensibles

dues à une mesure non bornée.

L'exemple S_4 correspond à la phrase S_3 dupliquée. Cette phrase, par conséquent, est plus complexe que la précédente. Cependant, les trois variantes de Ghosh affectent à cette phrase des valeurs qui sont inférieures à celles données pour l'exemple S_3. Cela signifie que, la mesure de Ghosh considère que cette phrase est moins complexe que S_3. Au contraire, CESAR considère que cette phrase est plus complexe et lui attribue une valeur plus élevée.

3.3.2 Évaluation de CESAR et Ghosh

Le corpus nommé C sur lequel porte nos expériences contient 1535 commentaires entièrement écrits dans la langue de référence (le dialecte), extraits du corpus algérien que nous avons collecté (voir le chapitre 2). Signalons que pour des raisons de rigueur de comparaison, tous les commentaires de C ont la même taille.

Nous avons effectué trois expériences distinctes E_1, E_2 et E_3 sur le corpus C. Pour chaque expérience E_i, nous injectons de nouveaux mots (un par un) dans les commentaires de C afin de mesurer la sensibilité de CESAR et des trois facteurs CF, $CF2$, et $CF3$ de Ghosh aux nouveaux mots introduits. Les mots sont injectés dans des positions choisies aléatoirement. Si ces mots proviennent de la langue de référence, cela signifie qu'il n'y a pas de code-switching dans le commentaire, par conséquent, les mesures initiales ne devraient pas changer. Dans le cas où les mots ajoutés proviennent d'autres langues, les valeurs des mesures devraient évoluer de manière croissante, car les commentaires sont désormais code-switchés.

Ces expériences sont reportées dans la figure 3.1.

— Dans l'expérience E_1, les mots ajoutés proviennent de la même langue que celle du corpus C (langue de référence).

— Dans l'expérience E_2, les mots ajoutés sont en français.

— Dans l'expérience E_3, les mots ajoutés sont en français, en MSA et en anglais.

L'expérience E_1 (courbes en vert) a comme objectif d'analyser le comportement des mesures lorsque les commentaires ne contiennent aucun mot en langue étrangère. On constate que CESAR et les valeurs de Ghosh sont constantes ce qui prouvent que les mesures évaluent correctement le niveau de code-switching dans le corpus.

Dans l'expérience E_2 (courbes bleues), nous avons ajouté des mots français aux commentaires composés uniquement de mots en dialecte (la langue de référence). Cela devrait conduire à augmenter les valeurs de Ghosh et de CESAR, car ces commentaires sont désormais code-switchés. Pour les trois facteurs de Ghosh, nous constatons qu'elles ont le même comportement. Pour les trois courbes CF, CF_2 et CF_3, les valeurs augmentent dès l'introduction du premier mot en français, jusqu'à un certain pallier, ce qui constitue un comportement normal et ensuite les valeurs de Ghosh dégringolent, ce que nous considérons comme un comportement surprenant puisqu'on continue à introduire des mots d'une autre langue. Alors que CESAR évolue d'une manière logarithmique avec le nombre de mots français ajoutés. Cette évolution montre que la mesure CESAR continue à pénaliser un texte dans lequel le nombre de mots étrangers devient de plus en plus important.

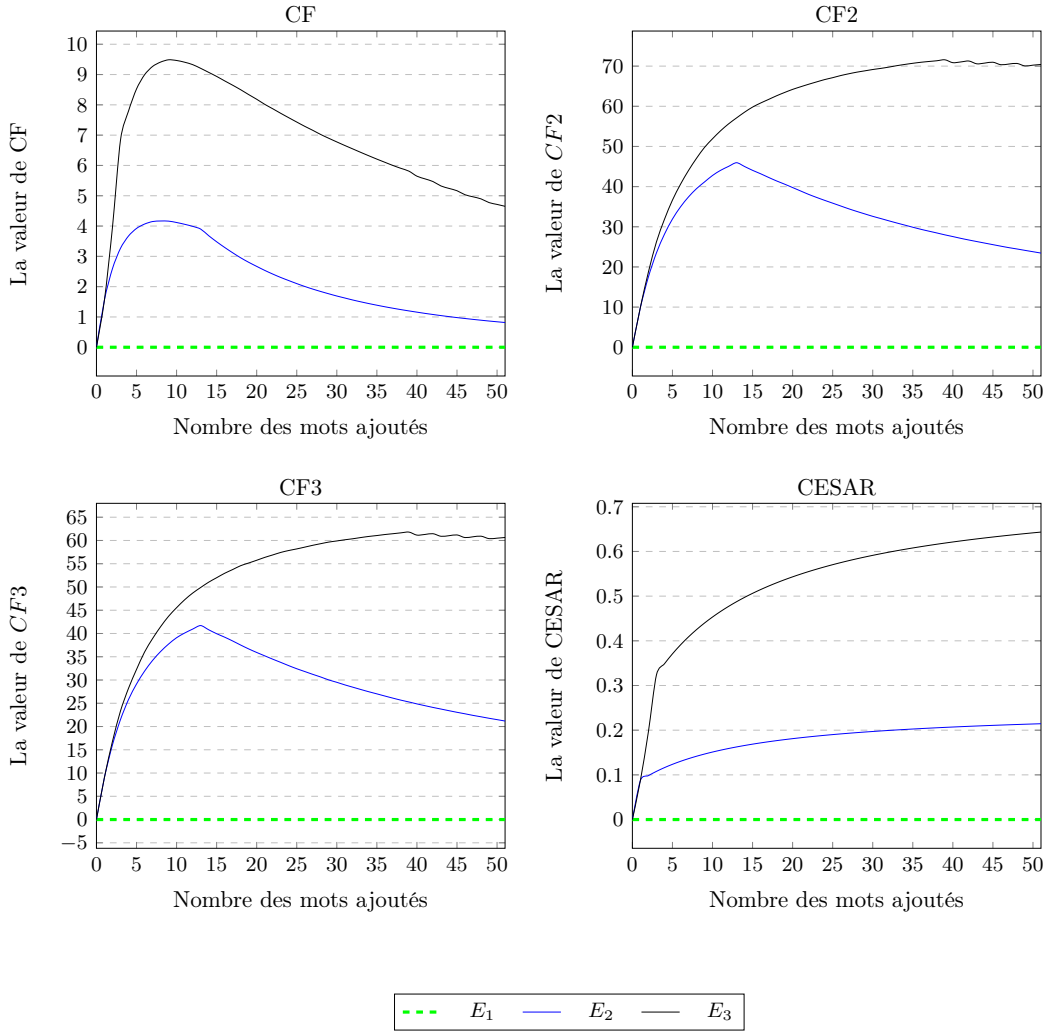


FIGURE 3.1 – La progression de Ghosh et CESAR en fonction de l'existence ou non code-switching.

Dans l'expérience E_3 (courbes en noir), nous injectons des mots, mais dans plusieurs langues. On constate que pour CF la mesure croît jusqu'à un certain niveau et décroît ensuite alors que le niveau de code-switching augmente. Pour CF_2 et CF_3 le comportement est meilleur que celui de CF puisque leurs valeurs augmentent mais finissent par se stabiliser à la fin même si le niveau de code-switching augmente. Signalons que l'introduction de plus de langues dans l'expérience E_3 n'a pas donnée lieu a un écart important dans les valeurs par rapport à l'expérience E_2 où on a introduit qu'une seule langue. Pour CESAR la mesure augmente avec le nombre de mots injectés et la mesure prend bien en compte le nombre de langues étrangères puisque les valeurs obtenues dans la majorité des cas sont supérieures à celles de l'expérience E_2.

3.3.3 Expérimentation sur les dialectes maghrébins

Après avoir évalué les deux métriques sur quelques exemples simples du corpus C. Dans cette section, nous mesurons le code-switching sur les trois corpus maghrébins que nous avons collectés à partir de YouTube.

Estimation de la valeur de CESAR et Ghosh sur les trois corpus du Maghreb

Le tableau 3.2 récapitule les valeurs obtenues avec CESAR et Ghosh sur ces corpus. Nous constatons que CESAR donne des valeurs similaires pour l'algérien et le tunisien et une valeur plus faible pour le marocain. Selon notre mesure le corpus marocain comporte moins de code-switching que les deux autres. En ce qui concerne les résultats obtenus par les mesures de Ghosh, l'interprétation est différente selon le facteur utilisé. Pour CF_2 et CF_3, les valeurs sont similaires pour les trois dialectes. Autrement dit, le taux de code-swtching pour les trois corpus du Maghreb est quasi le même. Alors que, pour le facteur CF, c'est le corpus tunisien qui semble être le plus code-switché.

	Algérien	**Marocain**	**Tunisien**
CESAR(C)	0,27	0,23	0,26
CF	15,01	16,24	17,52
CF_2	50,29	50,75	50,86
CF_3	47,92	48,35	48,80

TABLE 3.2 – Mesurer le code-switching des trois corpus maghrébins avec CESAR et Ghosh.

Mesure du code-switching en utilisant le français comme référence

Dans cette expérience, nous nous intéressons à mesurer le code-switching relativement au français. L'objectif étant d'identifier le dialecte qui utilise le plus la langue française. Parce que CESAR a la possibilité de mesurer le code-switching par rapport à une langue de référence qui est le français dans cette expérience et que les mesures de Ghosh ne le permettent pas, nous ne donnerons dans le tableau 3.3 que les résultats obtenus avec la mesure CESAR. Nous rappelons que si toutes les phrases sont écrites en français, CESAR donnera un score de zéro.

Dans le tableau 3.3, nous présentons les résultats sur tout le corpus (C) et ainsi que sur la partie écrite en script latin (SL). Les résultats montrent que le dialecte algérien est celui qui est le plus proche de la langue française, par rapport aux deux autres dialectes lorsque le calcul est effectué sur la partie SL du corpus C. On remarque que le corpus marocain, lorsque le calcul est effectué sur tout le corpus C, est celui qui est le plus code-switché par rapport au français. Autrement dit, c'est le corpus qui comporte le moins de segments en français.

Mieux identifier les mots du dialecte

Dans toutes les expériences précédentes pour calculer les mesures de code-switching (CESAR et Ghosh), nous avons dû affecter à chaque mot une étiquette correspondant à sa langue. Pour ce faire, nous avons utilisé des dictionnaires de MSA, français et anglais.

	Algérien	**Marocain**	**Tunisien**
CESAR(SL)	0,35	0,51	0,49
CESAR(C)	0,69	0,74	0,70

TABLE 3.3 – Résultats du code-switching dans les trois corpus en utilisant le français comme langue de référence.

Or, il n'y a pas de dictionnaires existants pour les dialectes traités dans ce chapitre. C'est pourquoi, nous avons procédé par élimination. Nous avons considéré que tous les mots qui ne se retrouvent pas dans les dictionnaires de MSA, français ou anglais sont des mots du dialecte.

Comme nous l'avons présenté dans les chapitres précédents, beaucoup de mots des dialectes du Maghreb proviennent de l'arabe standard. La conséquence est que lors de la phase d'identification, un mot qui appartient en même temps à l'arabe standard et au dialecte ne pourra pas être recensé comme un mot dialectal puisqu'il sera reconnu comme un mot appartenant au dictionnaire de MSA. Cela a comme conséquence de créer un biais dans le calcul de CESAR.

Ce que nous proposons est de considérer les mots communs au dialecte et à MSA comme des mots dialectaux. Pour ce faire, nous utilisons PADIC (*Parallel Arabic DIalectal Corpus*), développé il y a plusieurs années [Meftouh et al., 2015] et récemment étendu au dialecte marocain [Meftouh et al., 2018]. Ce corpus parallèle contient six dialectes arabes parmi lesquels il y a les trois dialectes concernant le sujet de cette thèse. Chaque mot appartenant aux dictionnaires des dialectes algérien, marocain et tunisien construits à partir de PADIC et au dictionnaire MSA est considéré comme un mot du dialecte correspondant.

Dans le tableau 3.4, nous présentons les nouveaux résultats de calcul de code-switching.

	Algérien	**Marocain**	**Tunisien**
CESAR	0,20	0,16	0,20
CF	15,75	16,22	17,80
CF_2	50,69	50,84	51,29
CF_3	48,22	48,43	49,16

TABLE 3.4 – Résultats du code-switching en désactivant les mots MSA dans le dialecte.

Les résultats obtenus montrent que pour CESAR on a gardé les mêmes proportions pour les trois dialectes, mais les valeurs ont baissé pour les trois puisque dans ce calcul il y a plus de mots qui sont considérés comme des mots des dialectes. Concernant les valeurs de Ghosh, globalement les valeurs n'ont pas changé.

3.4 Conclusion

Dans ce chapitre nous avons présenté CESAR, une nouvelle méthode pour estimer le degré de code-switching par rapport à une langue de référence. L'idée de base est de permettre d'évaluer le bruit existant dans les corpus collectés à partir des réseaux sociaux par rapport

à une langue de référence.

CESAR est une métrique bornée qui associe 0 à un document dans lequel le texte est entièrement dans la langue de référence (dialecte dans notre cas). Une valeur maximale égale à 1 est attribuée à tout document ne contenant aucun mot de la langue de référence. Lorsque le texte comporte plusieurs langues, CESAR lui attribue une valeur comprise entre 0 et 1 en fonction du degré du code-switching. Cette mesure a été comparée avec les facteurs proposés par Ghosh dans plusieurs expériences.

Enfin, cette mesure pourrait être utilisée pour extraire des corpus contenant un nombre maximum de mots de la langue de référence en minimisant la valeur de CESAR.

4

La construction automatique des corpus comparables à partir de données des réseaux sociaux

Sommaire

4.1 Introduction 53
4.2 L'alignement automatique des tweets multilingues : MSA-anglais 54
4.2.1 Acquisition des données à partir de Twitter 54
4.2.2 Pré-traitements des tweets 55
4.2.3 Alignement des tweets 59
4.2.4 Expérimentations 63
4.3 Méthode de création d'un corpus comparable pour les dialectes maghrébins 65
4.3.1 Approche basée sur le dictionnaire 65
4.3.2 Utilisation de la représentation phonétique pour l'appariement de textes 66
4.3.3 Une approche basée sur le *word embedding* 68
4.3.4 Une approche itérative basée sur le *multilingual word embedding* . . . 70
4.3.5 Expérimentations 72
4.4 Conclusion 75

Les corpus multilingues sont des ressources cruciales pour le développement et l'amélioration des outils multilingues. Parmi ces ressources, nous pouvons mentionner : les corpus parallèles et les corpus comparables. Les corpus parallèles regroupent des textes dans une langue source et leurs traductions dans une langue cible, tandis que les corpus comparables sont constitués de textes multilingues qui partagent la même thématique.

Dans ce chapitre, nous décrirons la méthode utilisée pour construire ces corpus comparables en exploitant la masse de données accessibles actuellement sur les réseaux sociaux. Dans les deux parties que nous présenterons dans ce chapitre, nous nous focaliserons sur la construction de corpus comparables pour les langues formelles et pour les dialectes.

4.1 Introduction

Le développement des modèles pour le domaine du TAL liés au multilingue (comme la traduction automatique) dépend essentiellement de la disponibilité de données en quantité considérable. Ces données sont regroupées généralement dans des corpus que nous nommerons corpus multilingues. Ils sont composés de textes dans deux ou plusieurs langues. La communauté travaillant dans ce domaine en recense deux types : les corpus parallèles et les corpus comparables. Les corpus parallèles sont constitués de textes et de leurs traductions. Quant aux corpus comparables, ils rassemblent des documents multilingues traitant du même sujet.

La traduction automatique est l'une des problématiques difficiles en traitement automatique des langues. La mise en œuvre de ces systèmes dépend de la disponibilité de collections de données parallèles en grande quantité. Pour la plupart des langues naturelles, il est souvent possible de trouver ce genre de corpus. Cependant, pour certaines langues peu dotées, ces données n'existent pas ou sont disponibles en petite quantité. Pour remédier à ce problème, une attention particulière a été portée sur l'exploitation des segments parallèles existant dans les corpus comparables. D'où l'intérêt de notre travail sur la création de corpus comparables. Plusieurs travaux se sont intéressés à l'exploitation des segments parallèles à partir de corpus comparables, parmi ces travaux citons, entre autres, [Smith et al., 2010], [Afli et al., 2012], et [Andoni Azpeitia and Garcia, 2018].

Actuellement, il existe plusieurs sources Web desquelles on peut extraire des corpus comparables. Citons à titre d'exemple, les articles de presse ayant des versions dans plusieurs langues [Aker et al., 2012] et Wikipédia qui dispose des textes dans plusieurs langues qui peuvent être associés via les liens inter-langues [Saad et al., 2013].

Cependant, en dehors des cas cités précédemment, il est nécessaire de disposer de métriques permettant d'identifier le degré de similitude entre les documents qui traitent du même sujet, mais qui sont écrits dans des langues différentes.

Un intérêt grandissant, pour l'exploitation des données des réseaux sociaux pour la construction des corpus multilingues, a été observé ces dernières années. Cela se fait, par l'alignement des textes multilingues publiés par les usagers sur un sujet particulier. Cependant, dans la plupart des cas, ces données ne sont pas directement exploitables, elles nécessitent des traitements particuliers notamment pour certaines langues comme l'arabe.

Dans ce chapitre, nous nous intéressons principalement à la création automatique des corpus comparables à partir des réseaux sociaux. Notre intérêt porte non seulement sur l'alignement de données en langue formelle, mais également en langue vernaculaire (dialecte). Pour ce faire, nous avons été contraints de répondre à plusieurs questions au cœur de cette problématique :

- Comment traiter efficacement le bruit qui existe dans les données provenant des réseaux sociaux ?
- Comment traiter la comparabilité alors que souvent dans les réseaux sociaux les messages sont très brefs ?

– Comment aligner des données vernaculaires alors que nous manquons de ressources linguistiques pour ces "langues" ?
– Comment gérer automatiquement la variété des mots, leurs écritures multiples, et l'évolution de ces formes d'écriture dans le temps ?

Ce chapitre est constitué de deux parties : la première partie est consacrée à la création d'un corpus comparable multilingue (MSA-anglais). Nous traitons de la problématique d'alignement des tweets multilingues. Nous commençons par la présentation du corpus collecté à partir de Twitter pour les deux langues formelles : arabe standard et anglais. Nous montrons aussi que la mesure classique [Li and Gaussier, 2010] de calcul de comparabilité basée sur un dictionnaire ne peut pas être efficace si elle est utilisée directement sur des données aussi brèves que celles des tweets. Pour pouvoir utiliser une mesure à base de dictionnaire, nous ferons appel à des traitements particuliers liés à la langue arabe et nous prendrons en compte également les phénomènes spécifiques à l'écriture de textes dans Twitter.

La deuxième partie est consacrée à la création d'un corpus comparable multilingue (Dialecte/MSA-Autres). Dans ce corpus on aligne des phrases écrites en script arabe avec des phrases écrites en script latin. À notre connaissance, ce type de corpus n'existe pas. Comme pour les dialectes, nous n'avons pas de dictionnaires, nous proposons une approche basée sur le *multilingual word embedding* pour aligner automatiquement des données vernaculaires sans faire appel à des ressources extérieures.

4.2 L'alignement automatique des tweets multilingues : MSA-anglais

Dans cette section, nous décrivons nos propositions pour aligner automatiquement des données provenant de Twitter. Nos expériences se focalisent sur les deux langues formelles : l'anglais et l'arabe standard.

4.2.1 Acquisition des données à partir de Twitter

Nous créons un corpus de tweets à partir d'un sujet d'actualité abondant : "*la guerre en Syrie*". Pour ce faire, nous avons collecté les tweets publiés entre septembre et décembre 2016 et correspondant aux Top-7 *hashtags* anglais et arabes employés pendant cette période.

En raison de la liberté d'écriture des mots arabes dans les réseaux sociaux, nous avons lancé la recherche des tweets en utilisant plusieurs possibilités d'écriture d'un même *hashtag*. Nous donnons à titre d'exemple les trois *hashtags* #سوريا, #سورية, et #سوريه qui correspondent au mot *Syrie*. Le premier se termine par *alif* (ا), le deuxième par *Ta* (ة), tandis que le troisième se termine par Ha (ه).

Dans les tableaux 4.1 et 4.2, nous donnons le nombre de tweets collectés pour chaque *hashtags* en utilisant l'API de Twitter.

Nous remarquons que le nombre de tweets en anglais concernant le *hashtag réfugiés syriens* est deux fois plus important que le nombre de tweets en arabe. Ceci peut être justifié par le fait que ce sujet était très populaire dans le monde occidental, car ces pays étaient directement concernés par le problème des réfugiés. Alors que, le nombre de tweets concernant le thème de

Les *hashtags* Anglais	**Nombre de tweets**
#SyrianRefugees	10,89k
#refugeescrisis	2,85k
#Syrianarmy	3,21k
#freesyrianarmy	3,12k
#SyriaCrisis	6,26k
#syria	17k
#syrian	17k
Total	**60,33k**

TABLE 4.1 – Nombre de tweets anglais collectés pour chaque *hashtag*.

Les *hashtags* en arabe	**Traduction**	**Nombre de tweets**
#اطفال_سوريا	Enfants de Syrie	1,59k
#الثورة_السورية	Révolution syrienne	10,09k
#اللاجيئين_السورين	Réfugiés syriens	4k
#سوريا	Syrie	17k
#سورية	Syrie	17k
#الجيش_العربي_السوري	L'armée arabe syrienne	0,91k
#الجيش_السوري_الحر	Armée syrienne libre	4,32k
Total		**54,91k**

TABLE 4.2 – Nombre de tweets arabes collectés pour chaque *hashtag*.

la Syrie est beaucoup plus important en arabe qu'en anglais parce que le monde arabe était directement concerné par la crise syrienne.

4.2.2 Pré-traitements des tweets

L'une des principales difficultés de l'utilisation de données issues de Twitter est le fait que l'information est condensée à cause de la taille limitée des tweets [16]. Cela a eu pour conséquence que les utilisateurs exploitent le peu de caractères autorisés pour faire passer un maximum d'information. Cela rend les traitements de type TAL plus difficiles. De ce fait, une étape de pré-traitement additionnel par rapport aux données standards est toujours recommandée afin d'homogénéiser l'écriture. Plus ces traitements d'homogénéisation dans les deux langues sont efficaces et plus le processus d'identification de tweets comparables est pertinent.

Avant de procéder aux pré-traitements séparés pour chaque langue, nous avons effectué sur les deux corpus quelques pré-traitements communs :

- La suppression de tweets vides et les re-tweets.
- La suppression des liens https.
- La suppression des caractères spéciaux et des émoticônes.

16. La taille de tweets a été limitée à 140 caractères jusqu'au mois de septembre 2018.

– La correction de mots étirés comme *woahhh*, عاااجل en supprimant les caractères répétés séquentiellement si leur nombre dépasse 2.

Pré-traitement de tweets anglais

À cause du nombre restreint des caractères de tweets, les usagers ont tendance à utiliser les abréviations et les acronymes pour écrire dans Twitter, dans le but de gagner de l'espace et du temps. Donnons à titre d'exemple, l'abréviation *ppl* qui est utilisée pour écrire le mot *people*. En arabe standard, en général il y a moins d'abréviations. Pour traiter ce problème dans le corpus des tweets anglais, nous avons remplacé ces abréviations par leur forme littéraire en utilisant un dictionnaire de SMS [17].

Les *hashtags* nécessitent également un traitement. Rappelons qu'il s'agit de mots-clés précédés par un (#) qui jouent un rôle important dans l'indexation des tweets postés par les usagers et leur facilitent la recherche des publications qui correspondent à leurs intérêts. Nous avons profité de ces *hashtags* afin de les utiliser comme une autre alternative pour gagner de l'espace. Le problème est que certains *hashtags* peuvent être composés de plusieurs mots, par exemple *#SyrianArabArmy, #SyrianCivilians, etc.* Il est donc nécessaire de séparer les mots du *hashtag* composé afin d'accéder à chaque constituant qui peut être utile lors de la comparabilité. Pour ce faire, nous avons effectué deux traitements sur ces *hashtags* :

– Dans le cas où le *hashtag* est composé d'une suite de mots concaténés où chaque mot commence par une majuscule, la séparation est simple.
– Dans les cas où les mots sont séparés par un caractère spécial, le traitement de séparation est également évident.
– Dans le cas où le *hashtag* est entièrement écrit en minuscule, le problème est plus difficile. À partir du corpus des tweets, nous avons construit un dictionnaire de tous les mots qui ne sont pas dans le *hashtags*, et nous l'avons trié par ordre alphabétique et par taille. Ensuite, grâce à cette liste une recherche de motifs dans le *hashtag* est effectuée pour séparer les mots.

Pour rendre l'alignement de tweets plus performant, nous avons décidé d'utiliser toutes les informations qu'on peut trouver dans ces messages, notamment les dates, les nombres et les noms propres, etc. Le problème est qu'il n'existe pas dans Twitter une manière standard pour écrire ces informations. C'est pourquoi, nous avons décidé d'homogénéiser leurs écritures. Nous avons réécrit les nombres dans un format standard, par exemple le nombre *13,5* est transformé en *13.5*. Pour les dates nous avons choisi le format : *JJ / MM / AAAA*. Dans le tableau 4.3, nous donnons quelques exemples avant et après le processus de réécriture.

Pré-traitement de tweets arabes

Comme nous l'avons évoqué dans le chapitre 2, la langue arabe est très différente des langues indo-européennes. Elle est caractérisée par un ensemble de particularités qu'il faut prendre en considération. C'est pourquoi des traitements supplémentaires doivent être appliqués sur les tweets arabes.

17. Disponible dans : http://www.illumasolutions.com/omg-plz-lol-idk-idc-btw-brb-jk.htm

Avant le pré-traitement	**Après le pré-traitement**
April 13, 2016	13/4/2016
February 1, 2016	1/2/2016
22 feb 2016	22/2/2016
2.2.2016	2/2/2016

TABLE 4.3 – Exemples de dates homogénéisées.

Parce que dans Twitter, les utilisateurs s'octroient le droit d'écrire librement, ils remplacent parfois la lettre ة par ه uniquement à la fin des mots. Pour un non arabophone, cela pourrait être surprenant puisque ces deux lettres ont des rôles linguistiques différents, cependant elles sont similaires au niveau de la forme, sauf que la première a deux points au-dessus. Pour les mêmes raisons, nous avons remplacé toutes les formes du symbole *Alif* avec *Hamza*, comme dans : آ, أ, إ par un simple *Alif* ا. Concernant les signes diacritiques, nous les avons supprimés, car généralement on arrive à lire un texte en arabe sans voyelles.

En arabe, deux systèmes de numération sont utilisés pour écrire les nombres, le premier type est celui utilisé dans le monde entier qui utilise des chiffres arabes et le second type est celui qui utilise des chiffres indiens [18]. Par conséquent, nous avons décidé de conserver la notation utilisant les chiffres arabes afin de faciliter la mise en correspondance entre les chiffres dans les deux langues. En ce qui concerne les nombres décimaux en arabe, nous les avons réécrits en utilisant les chiffres arabes et nous avons opté pour la même homogénéisation que pour l'anglais.

En ce qui concerne les dates, dans le monde arabe, trois types de calendriers pourraient être utilisés : Assyrien, Hijri et Grégorien. Le premier et le second sont plus utilisés au Moyen-Orient qu'à l'ouest du monde arabe. Par exemple : le mois de *Janvier* pourrait être écrit en arabe : كانون التاني, جانفي ou يناير en fonction de la région du monde arabe. C'est pour cette raison, que chaque date, quelle que soit sa forme, est réécrite dans notre traitement selon le format utilisé précédemment *JJ / MM / AAAA* (voir les exemples dans le tableau 4.4).

Avant	**Après**
١٣ من نيسان ٢٠١٦	13/4/2016
الاول من شباط ٢٠١٦	1/2/2016
22 فبراير 2016	22/2/2016
٢٠٢٠٢٠١٦	2/2/2016
13٫5	13.5

TABLE 4.4 – Exemples des dates et des nombres avant et après les pré-traitements.

18. Ils sont utilisés au Moyen-orient : ٩ .. ١٢٣

Lemmatisation

Comme mentionné plus haut, l'arabe est une langue riche morphologiquement en raison de règles fondamentales utilisées pour la construction des mots. En fait, une racine est souvent considérée comme le générateur de mots, car sa concaténation avec un ensemble d'affixes produit de nouveaux mots. Dans [Kadri and Nie, 2006] ces affixes sont classés en quatre catégories : des antéfixes, des préfixes, des suffixes et des postfixes comme le montre le tableau 4.5. Ces affixes sont attachés avec la racine selon un ordre spécifique, que nous appelons ici niveau interne et externe (voir la figure 4.1).

	Arabic affixes
Antéfixes	ل, ب, و, ف, ك, وس, فل, فب, فس, ول, و ب, ال, ولّل, كال, فال, وال, وبال
Préfixes	ا, ن, ي, ت
Suffixes	ت, ن, ا, ي, و, ين, وا, تا, تم, تن, نا, تما, يون, تين, تان, ات, ان, ون
Postfixes	ي, ه, ك, كم, هم, نا, ها, تي, هن, كن, هما, كما

TABLE 4.5 – La classification des affixes proposée par [Kadri and Nie, 2006].

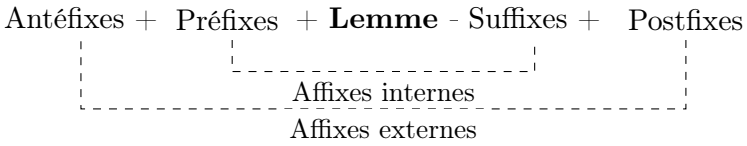


FIGURE 4.1 – Typologie de la morphologie des mots en arabe.

La plupart des travaux dédiés à la langue arabe concluent que la lemmatisation est une étape primordiale pour améliorer la précision des systèmes traitant cette langue. La lemmatisation consiste à décomposer un terme en sa forme canonique.

Afin d'améliorer l'identification de tweets comparables, nous devons résoudre le problème de la variabilité de mots arabes due à la richesse morphologique de la langue. Pour ce faire, nous avons lemmatisé les tweets en arabe. De surcroît, les données provenant des réseaux sociaux contiennent souvent des mots mal orthographiés. Par exemple dans notre corpus, le mot لايكلمونهم, qui signifie "*Ils ne leur parlent pas*", est mal orthographié, car en arabe standard, on doit mettre un espace entre لا et يكلمونهم. Par conséquent, les analyseurs morphologiques conçus pour l'arabe standard sont incapables de proposer une décomposition morphologique à ces mots. Pour remédier à ce problème, dans ce travail nous proposons une méthode fondée sur une combinaison utilisant l'analyseur morphologique de Buckwalter [Tim, 2004] et un analyseur de type *Light Stemming* que nous avons développé. Ce dernier est fondé sur la troncation progressive de l'ensemble des affixes concaténés à ce mot. Ce processus de troncation se déroule de la manière suivante :

— **Étape 1 (La suppression des affixes du niveau externe)** : cette étape consiste à supprimer du mot l'ensemble des affixes du niveau externe (postfixes et/ou antéfixes).

Trois possibilités s'offrent à nous et cela produit trois lemmes : la première consiste à tronquer du mot l'antéfixe, le postfixe pour la seconde et les deux pour la troisième.

— **Étape 2 (La suppression des affixes du niveau interne)** : on procède à une suppression, à partir de chaque lemme candidat produit dans l'étape 1, des affixes du niveau interne (préfixes et suffixes) de la même manière que dans l'étape précédente. Cela produira trois nouveaux lemmes pour chacun des lemmes de l'étape précédente. Pour la suite, nous gardons les 12 lemmes produits par ces deux étapes.

— **Étape 3 (Filtrage de la liste des lemmes obtenues)** : cette suppression peut produire des lemmes qui n'appartiennent pas à la langue arabe. Pour pallier ce problème, nous vérifions chaque lemme engendré grâce à un dictionnaire de 9 millions de mots arabes utilisé dans [Menacer et al., 2017a]. Tous les lemmes produits n'appartenant pas à cette liste sont écartés.

Pour illustrer le processus décrit précédemment, revenons à l'exemple : لايكلمونهم (*Ils ne leur parlent pas*). Cette séquence devrait être écrite en séparant le mot لا et le mot يكلمونهم. La figure 4.2 montre comment on peut produire des lemmes pour cette séquence mal orthographiée avec la méthode que nous avons proposée. Dans cette figure nous utilisons la notation suivante : Ant : préfixe, Pos : postfixe, Pre : préfixe, Suf : suffixe}.

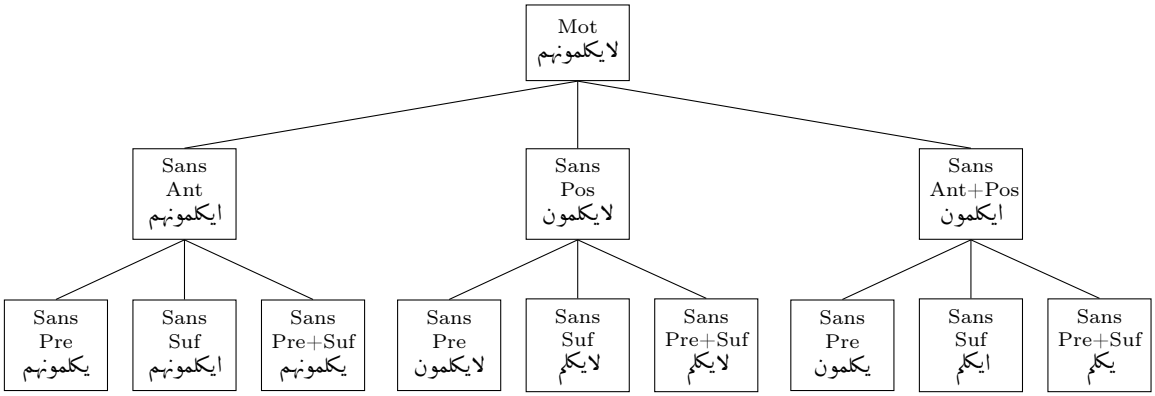


FIGURE 4.2 – Exemple des lemmes obtenus pour le mot arabe mal écrit en utilisant l'approche basée sur *Light Stemming*.

L'application de l'étape 1 et l'étape 2 de la méthode précédente produit les 12 lemmes de la figure 4.2. L'application de l'étape 3 produit 4 lemmes que nous considérons être corrects : { يكلم, يكلمون, ايكلم, يكلمونهم, ايكلمونهم }.

Pour lemmatiser les mots arabes de notre corpus, nous appliquons d'abord l'analyseur morphologique de Buckwalter et en cas d'échec et notamment pour les mots mal orthographiés, nous appliquons la procédure citée ci-dessus. Pour l'anglais, nous avons utilisé l'analyseur morphologique TreeTagger. [19]

4.2.3 Alignement des tweets

La création automatique de corpus comparables multilingues est réalisée à l'aide de l'utilisation de mesures de similarité permettant d'aligner deux tweets exprimés dans deux langues

19. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

différentes. Dans notre travail, nous avons décidé d'utiliser une mesure fondée sur le vocabulaire commun, la mesure proposée par Li et Gaussier dans [Li and Gaussier, 2010]. Cette mesure est basée sur la quantité d'équivalents de traduction partagée entre deux documents, obtenue par l'utilisation d'un dictionnaire bilingue. Dans la suite, nous calculons la similitude entre deux tweets T_a (tweet en arabe) et T_e (tweet en anglais). La mesure de comparabilité est définie comme le score maximum entre un tweet arabe T_a et tous les tweets T_i pour $1 \leq i \leq N_{T_e}$, où N_{T_e} est le nombre total de tweets en anglais.
Pour chaque couple de tweets arabe et anglais, le score est calculé comme suit :

$$Score(T_e, T_a) = \frac{\sum_{w \in \{T_e \cap D_e\}} \sigma(w, T_a) + \sum_{w \in \{T_a \cap D_a\}} \sigma(w, T_e)}{\mid T_e \cap D_e \mid + \mid T_a \cap D_a \mid} \tag{4.1}$$

Où D_e et D_a sont respectivement les parties anglaise et arabe du dictionnaire bilingue. σ est une fonction qui prend la valeur 1 si une des traductions de w données par le dictionnaire bilingue ($T(w)$) existe dans la liste des mots du tweet de la langue cible désigné ci-dessous par V_x.

$$\sigma(w, V_x) = \begin{cases} 1 \ si \ T(w) \cap V_x \neq \emptyset \\ 0 \ sinon \end{cases} \tag{4.2}$$

La mesure LG donne le score du meilleur tweet à aligner avec T_e :

$$LG(T_e) = \max_{1 \leq i \leq N_{T_a}} Score(T_e, T_a^i) \tag{4.3}$$

Afin de faire la correspondance entre les mots source et cible nous avons utilisé le dictionnaire bilingue OMWN (*Open Multilingual WordNet*) [20]. Ce dictionnaire est disponible en de nombreuses langues, parmi ces langues les langues arabe et anglaise. OMWN contient 148k lemmes anglais, extraits de WordNet anglais et 17k lemmes arabes extraits de WordNet arabe.

Parce que, par définition, les tweets sont des messages courts, il est très difficile d'établir la correspondance entre deux tweets et par conséquent les aligner juste en se basant sur le peu de mots qui les composent. C'est pourquoi, en plus de la correspondance des mots, nous utilisons d'autres méthodes pour aligner les tweets.

Utilisation des dates de publication des tweets

Afin de réduire le nombre de paires de tweets candidats à l'alignement, nous proposons de comparer seulement les tweets dont les dates de publication sont proches. Cette procédure permet de comparer seulement des tweets concernant des événements semblables, voire identiques. Pour ce faire, à partir de deux corpus S et T, dans deux langues différentes et pour un tweet s_i^d publié à la date d, nous recherchons tous les tweets $t_j^{d'}$ publiés à la date d' en respectant la contrainte $d - 1 \leq d' \leq d + 1$.

20. http://compling.hss.ntu.edu.sg/omw/

Utilisation de marqueurs pertinents dans les tweets

Utilisation des marqueurs pertinents de type date et nombres Comme l'objectif est de cibler les tweets à aligner, nous avons décidé d'utiliser les marqueurs pertinents comme : les dates, les nombres et les noms propres. En effet, si ces mêmes marqueurs se retrouvent dans les tweets sources et cibles, il y a de fortes chances que les tweets en question soient de bons candidats à l'alignement. La mise en correspondance des dates et des nombres est facile à établir puisque nous avons homogénéisé leurs écritures dans la section 4.2.2.

Utilisation du marqueur pertinent de type nom propre Concernant la détection des noms propres, le problème est plus critique pour le traitement automatique en langue arabe. En effet, pour l'arabe, la notion de majuscule n'existe pas. De plus, les noms des personnes font référence, en général, à des noms communs en arabe. Par exemple, le prénom كريمة signifie *généreuse* et par conséquent il peut qualifier un substantif dans une phrase et dans ce cas, il ne doit pas être considéré comme un nom propre. Un autre problème des noms propres est qu'il arrive qu'ils soient agglutinés avec des affixes ce qui nécessite une étape de lemmatisation avant leur identification. Par exemple : بسوريا (*en Syrie*) doit être lemmatisé et séparé en deux termes : بـ (*en*) et سوريا (*Syrie*).

Les noms propres peuvent être simples tels que علي (*Ali*) ou composés tels que : ابن عبد الرحمان (*ibn Abdul Rahman*) ou علاء الدين (*Alaa Aldine*). Les noms propres composés sont généralement constitués d'un seul nom propre précédé ou suivi de particules. Des exemples de ces particules sont présentés dans le tableau 4.6. Par conséquent, pour faciliter leur traitement, nous avons décidé de concaténer le mot à ses particules. Par exemple, ابن عبد الرحمان (*ibn Abdul Rahman*) est réécrit en ابن_عبد_الرحمان *ibn_Abdul_Rahman.*

Particules	**arabe (translittération anglais)**
Préfixes	(ابن، بن) (ibn, bin, ben), عبد (3bd, abd), ابو (abu, abo, abou), بنت (bint, bent), ام (oum)
Suffixes	الدين (eldin, aldin, uldin, eldin)

TABLE 4.6 – Les particules utilisées dans les noms propores.

Pour trouver une solution à ce problème de mise en correspondance entre un nom propre en arabe et un nom propre en anglais, nous proposons de procéder à la translittération de certains mots potentiellement considérés comme nom propre. La translittération est l'action de représenter les signes d'un alphabet de la langue source par les signes de l'alphabet de la langue cible, par exemple le nom du président algérien بوتفليقة est translitéré en anglais par *Bouteflika*. Le problème de la translittération est qu'elle n'est pas unique surtout lorsque les données traitées proviennent des réseaux sociaux. Tout dépend de la façon dont le nom propre est prononcé dans la langue cible. À titre d'exemple, pour le prénom : سليمان, les translittérations suivantes sont possibles : *Sulayman, Seleiman, Sliman et Selayman.*

Les noms propres composés sont faciles à identifier grâce aux particules qui les précèdent. Par conséquent dès qu'une particule est concaténée à une suite de termes, le résultat est considéré comme un nom propre potentiel. Quant à l'identification des noms propres simples, il s'agit d'une tâche plus ardue. Pour résoudre le problème, nous avons décidé de les coder phonétiquement. En effet, certains mots dans la source et dans la cible seront codés phoné-

tiquement en utilisant Soundex [Aqeel et al., 2006]. L'idée est que les noms ayant la même prononciation disposeront de la même codification et par conséquent on saura les apparier. Dans ce qui suit, nous expliquons la procédure de codage phonétique des mots qu'on "soupçonne" d'être des noms propres. Nous avons utilisé deux versions de Soundex, une en anglais et une en arabe [Aqeel et al., 2006]. L'algorithme consiste à :

1. Conserver la première lettre du mot telle qu'elle est.
2. Remplacer toutes les autres lettres par leur index de groupe de caractères comme indiqué dans le tableau 4.7. Ces groupes sont constitués de lettres correspondant globalement à la même classe de sons.
3. Tous les caractères du groupe 0 sont ignorés, autrement dit ils ne sont pas codés. Sauf s'ils apparaissent en première position du nom et auquel cas il faut procéder comme dans l'étape 1.
4. Chaque mot sera codé avec une lettre suivie des trois premiers chiffres. Les autres seront ignorés.

Groupe de lettres en anglais	**Index**	**Groupe de lettres en arabes**
A E H I O U W Y	0	ى و ه ع ح ا
B P F	1	ب ف
C S K G J Q X Z	2	ك ق غ ص ش س ز ج ح
D T	3	ظ ط ض ذ د ث ت
L	4	ل
M N	5	م ن
R	6	ر

TABLE 4.7 – Table de groupes de lettres et les index correspondants de Soundex en anglais et en arabe.

Comme Soundex préserve la première lettre du mot, il ne sera donc pas possible de mettre en correspondance les deux codes associés à un même nom propre, l'un avec une lettre en arabe et le second avec une lettre en anglais. Par exemple, le nom propre arabe جميلة sera codé ج540 alors qu'en anglais il peut être codé J540. Il faut donc faire un post-traitement du code obtenu en arabe. Pour ce faire, nous avons utilisé un tableau de translittération des lettres arabes vers les lettres latines (dans le tableau de l'annexe B). Dans la figure 4.3, nous présentons l'exemple complet de codification du mot جميلة. On constate que l'application de Soundex en arabe et ensuite l'application du traitement de translittération de la première lettre donne les codes suivants : DJ540, J540 et G540. En effet, la lettre ج a trois translittérations possibles en latin {DJ, J et G}. Par conséquent, avec ces trois codes, on pourra faire correspondre جميلة avec un des trois noms propres *Djamila*, *Jamila* ou *Gamila* qui sont en effet toutes les écritures possibles de ce prénom en latin (voir Figure 4.3).

Pour la mise en correspondance entres les mots des tweets des deux langues, nous commençons d'abord par rechercher la traduction d'un mot du tweet source (anglais) dans le tweet cible (arabe) grâce au dictionnaire OMWN. En cas d'échec, autrement dit, si le mot en anglais n'existe pas dans le dictionnaire, il y a des chances qu'il s'agisse d'un nom propre, auquel cas on le code en utilisant la procédure Soundex, et dans ce cas, le tweet en

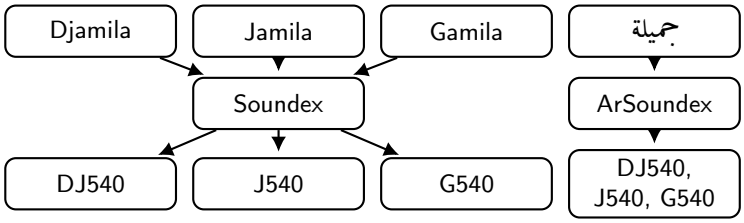


FIGURE 4.3 – L'encodage phonétique du nom propre arabe جميلة et ses différentes formes en latin.

arabe est codé entièrement avec Soundex. Le mot recherché initialement dans OMWN est désormais recherché dans la liste des codes du tweet arabe. En cas d'égalité de codes, la mise en correspondance est établie.

Grâce aux procédures précédentes, il est possible de traiter les noms propres, les dates, et les nombres et comme ces derniers ne se retrouvent pas dans les dictionnaires, nous avons été amenés à modifier légèrement le score de Li et Gaussier comme suit :

$$LGT(T_e) = \max_{1 \leq i \leq N_{T_a^i}} \frac{\sum_{w \in T_e} \sigma(w, T_a^i) + \sum_{w \in T_a^i} \sigma(w, T_e)}{|\ T_e| + |\ T_a^i|} \tag{4.4}$$

4.2.4 Expérimentations

Dans cette section, nous comparons les approches, celle de base proposée par [Li and Gaussier, 2010] (LG), et celle prenant en compte les traitements supplémentaires que nous avons proposés (LGT). Les corpus parallèles, pour le calcul de la comparabilité peuvent être considérés comme des corpus de référence puisqu'on connaît pour chaque texte celui avec lequel il faut l'apparier. C'est pourquoi nos tests seront effectués sur deux corpus parallèles.

Le premier corpus est composé de tweets alignés automatiquement par les auteurs de [Ling et al., 2013] [21], que l'on notera par la suite C_t. Ce corpus ne contient que 2006 tweets parallèles. À notre connaissance, c'est le seul corpus de tweets parallèles disponible pour la paire de langues arabe-anglais.

Parce que ce corpus est trop petit et que nous avons besoin de montrer la faisabilité de notre modèle, nous avons décidé de le tester sur un corpus plus large, mais qui malheureusement n'est pas constitué de tweets, mais de dépêches journalistiques. Il est composé de 11942 phrases parallèles extraites du journal ANN [22], nommé ci-après : C_{ANN}.

En analysant C_t, nous avons constaté que beaucoup de tweets sont alignés par erreur par la méthode automatique de [Ling et al., 2013] (des exemples de ce corpus sont présentés en annexe B). C'est pourquoi, nous avons décidé de n'utiliser qu'une partie de celui-ci. Nous

21. http://www.cs.cmu.edu/~lingwang/microtopia/
22. www.annahar.com

avons donc sélectionné manuellement 340 tweets que nous estimons être parallèles, nous noterons ce sous-ensemble $C_{t_{340}}$.

Nous avons mesuré la comparabilité avec les deux mesures décrites précédemment sur les deux corpus $C_{t_{340}}$ et C_{ANN}. Pour chaque texte source, nous avons généré la liste des textes cibles ordonnés du plus comparable au moins comparable, selon la mesure utilisée. Puis, nous avons calculé le rappel classique qui consiste à vérifier si le texte cible correspondant est en première position dans la liste (R@1), dans les 5 premiers textes (R@5) et dans les 10 premiers textes (R@10). Les résultats obtenus sont présentés dans le tableau 4.8.

Corpus	**Méthode**	**R@1**	**R@5**	**R@10**
$C_{t_{340}}$	LG	53	73	78
	LGT	**56**	77	**83**
C_{ANN}	LG	73	85	87
	LGT	**79,7**	89,4	**92**

TABLE 4.8 – Les rappels R@1, R@5 et R@10 obtenus avec les deux mesures LG et LGT sur les deux corpus parallèles C_{ANN} et $C_{t_{340}}$.

Ce tableau montre que la mesure que nous avons proposée (LGT) donne de meilleurs résultats par rapport à la mesure de la base proposée par [Li and Gaussier, 2010] (LG) et ce pour les deux corpus. Cela se justifie par le fait que nous prenons en compte des informations supplémentaires pour l'appariement.

Bien qu'il soit difficile d'apparier des tweets écrits par des personnes différentes et dans deux langues, la mesure LGT a permis, en rang 1, d'obtenir un rappel de 56%. Ce taux est d'une manière absolue faible, mais pour des tweets, il n'est pas négligeable. Dès que l'on recherche des correspondances dans les cinq premiers rangs, la mesure devient performante avec un rappel de 77%. Ces résultats sont confirmés sur le corpus de journaux C_{ANN} où l'on obtient un rappel de 79,7% en position 1 et 89,4% en position 5. Après ces résultats que l'on peut considérer comme encourageants, nous avons utilisé la méthode décrite précédemment pour apparier la totalité de nos corpus de tweets en arabe et en anglais. Cela a conduit à un corpus multilingue comparable de 11,5k tweets dont un échantillon est donné dans le tableau 4.9.

minister kerry : the diplomatic path is the only path that can isolate terrorist groups like daash and front victory 1 2 syria
الوزير كيري المسار الديبلوماسي هو المسار الوحيد الذي يمكن ان يعزل الجمعات الارهابيه مثل داعش
news : obama called putin on syria ceasefire : white house
البيت الابيض اوباما و بوتين يبحثان وقف اطلاق النار في سوريا
Syria president bashar al assad issues decree no 63 which sets wednesday 13/4/2016
سوريه اصدر الرئس بشار الاسد المرسوم رقم ٦٣ لّعام ٢٠١٦ القاضي بتحدد يوم الاربعاء ١٣/٤/٢٠١٦ موعدا

TABLE 4.9 – Des exemples de tweets comparables identifiés automatiquement.

4.3 Méthode de création d'un corpus comparable pour les dialectes maghrébins

Cette section est consacrée à la méthode que nous avons proposée pour la création d'un corpus comparable comprenant des textes en dialectes, en MSA, en français et en anglais. À notre connaissance, il n'existe pas de corpus comparables pour les dialectes arabes en général et pour les dialectes maghrébins en particulier.

Dans la suite, nous présenterons seulement les expériences menées sur le corpus algérien (collecté dans le chapitre 2) parce que l'évaluation nécessite des corpus de référence construits manuellement. Ce travail est coûteux, c'est pour cette raison que les tests sont effectués uniquement sur le corpus algérien. Néanmoins, la méthode proposée est valable pour tous les dialectes. La particularité du corpus que l'on propose de créer réside dans le fait qu'il comporte non seulement des paires de commentaires en dialecte, mais également des paires en langue formelle ou les deux en même temps.

Dans le tableau 4.10, on donne trois exemples de phrases comparables que l'on peut trouver dans le corpus final. *D-D* signifie que la phrase source est en dialecte et la cible également. *F-D* signifie que la phrase source est en français et la cible en dialecte et enfin pour la troisième *A-D*, la source est en anglais et la cible en dialecte.

Type	Source	Cible
D-D	vive tahar misoum rabi yahafdak "*Vive Tahar Missoum, que dieu te protège*"	تحيا طاهر ميسوم " *Vive Tahar Missoum*"
F-D	Merci chemsou je t'adore	نحبك شمسو بزاف "*Je t'aime beaucoup Chemsou*"
A-D	the best youtube video ever good luck Dz djoker	فيديو اكتر من رايع تشوفو 100 خطرة ما تكرهش "*Une superbe vidéo, tu peux la visionner* 1000 fois sans en avoir marre"

Table 4.10 – Exemples de commentaires du corpus comparable.

Dans la suite, nous présentons les différentes méthodes que nous avons testées pour la création d'un tel corpus.

4.3.1 Approche basée sur le dictionnaire

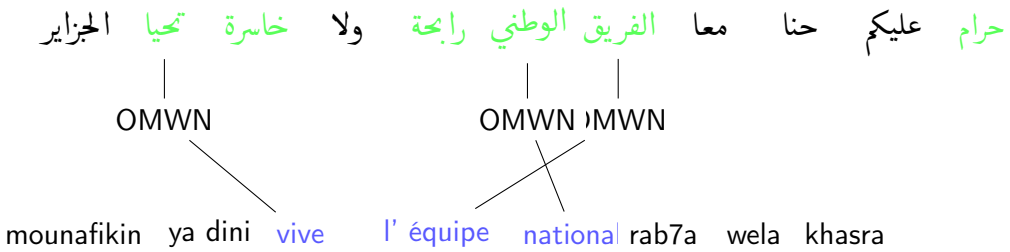
Comme présenté dans la section 4.2.3, la mesure de Li et Gaussier permettant de calculer la comparabilité entre deux textes [Li and Gaussier, 2010] est fondée sur un dictionnaire bilingue. Malheureusement, il n'existe pas de dictionnaire bilingue incluant le dialecte algérien. Par conséquent, pour utiliser la mesure précédente il faudrait trouver une solution. Ce que nous proposons est d'utiliser les dictionnaires de *Open Multilingual WordNet* (OMWN) [23]. En effet, comme le dialecte partage un certain nombre de mots avec l'arabe standard, nous espérons pouvoir aligner nos textes en se basant malheureusement seulement sur les mots communs entre le dialecte et l'arabe standard. De la même manière, les dictionnaires français et anglais de la série OMWN permettront d'identifier certains mots dans les textes puisque

23. http://compling.hss.ntu.edu.sg/omw/

dans ces corpus il y a un certain nombre de mots qui sont écrits en français ou en anglais.

L'exemple de la figure 4.4 illustre la manière dont les mots sont appariés en utilisant les dictionnaires de OMWN. Dans cet exemple la phrase source à apparier est écrite en dialecte en utilisant le script arabe. Alors que la phrase cible est écrite également en dialecte, mais en utilisant le script latin. On remarquera que dans la phrase cible, certains mots sont en français (mots en bleu). Également dans la phrase source certains mots appartiennent au vocabulaire de l'arabe standard (mots en vert). Dans cet exemple, l'alignement se fera grâce au dictionnaire de l'arabe standard et du français entre les mots qui sont reliés dans la figure 4.4. Nous constatons que très peu de mots sont associés pour l'alignement et donc le résultat en utilisant cette façon de faire ne peut donner de bons résultats sur l'ensemble du corpus.

Traduction : *Honte à vous, nous sommes avec l'équipe nationale qu'elle gagne ou qu'elle perde, vive l'Algérie*



Traduction : *Espèces d'hypocrites, vive l'équipe nationale qu'elle gagne ou qu'elle perde*

FIGURE 4.4 – Un exemple de deux commentaires comparables.

4.3.2 Utilisation de la représentation phonétique pour l'appariement de textes

Comme la méthode précédente est fondée sur un dictionnaire, elle ne permet malheureusement pas d'aligner les données en dialecte puisque nous n'avons pas de dictionnaire du dialecte. Pour pallier ce problème, nous avons décidé d'utiliser la méthode précédente, mais en codant phonétiquement comme nous l'avons présenté dans la section 4.2.3 tout mot du dialecte qui n'est pas dans le dictionnaire arabe. Pour établir la correspondance, la phrase ou le texte cible à aligner est codé entièrement phonétiquement. Comme certains mots du dialecte sont en fait des mots appartenant à d'autres langues mais écrits en arabe ou en arabizi, le fait de les coder phonétiquement permettra de les associer lorsqu'ils correspondent au même mot. L'avantage de cette méthode est que l'on pourra associer à un mot supposé être un mot du dialecte écrit en script latin à un mot écrit en script arabe, même si ces derniers sont mal orthographiés, mais néanmoins proches. Ces mots partagent en effet la même prononciation. Par exemple, le mot كومنتار n'est pas un mot arabe, mais il est utilisé dans le dialecte et il correspond au mot français *commentaire*. Imaginons la phrase suivante écrite en arabe ....كومنتار.... et la phrase écrite en script latin .... *comentaire*.... Notre méthode pourra associer le mot كومنتار et le mot français mal orthographié *comentaire*. Dans l'exemple vu précédemment on n'avait que 3 mots qui étaient alignés, cette fois-ci, grâce à l'application du codage phonétique des mots, nous avons pu associer 6 mots comme le montre la figure 4.5.

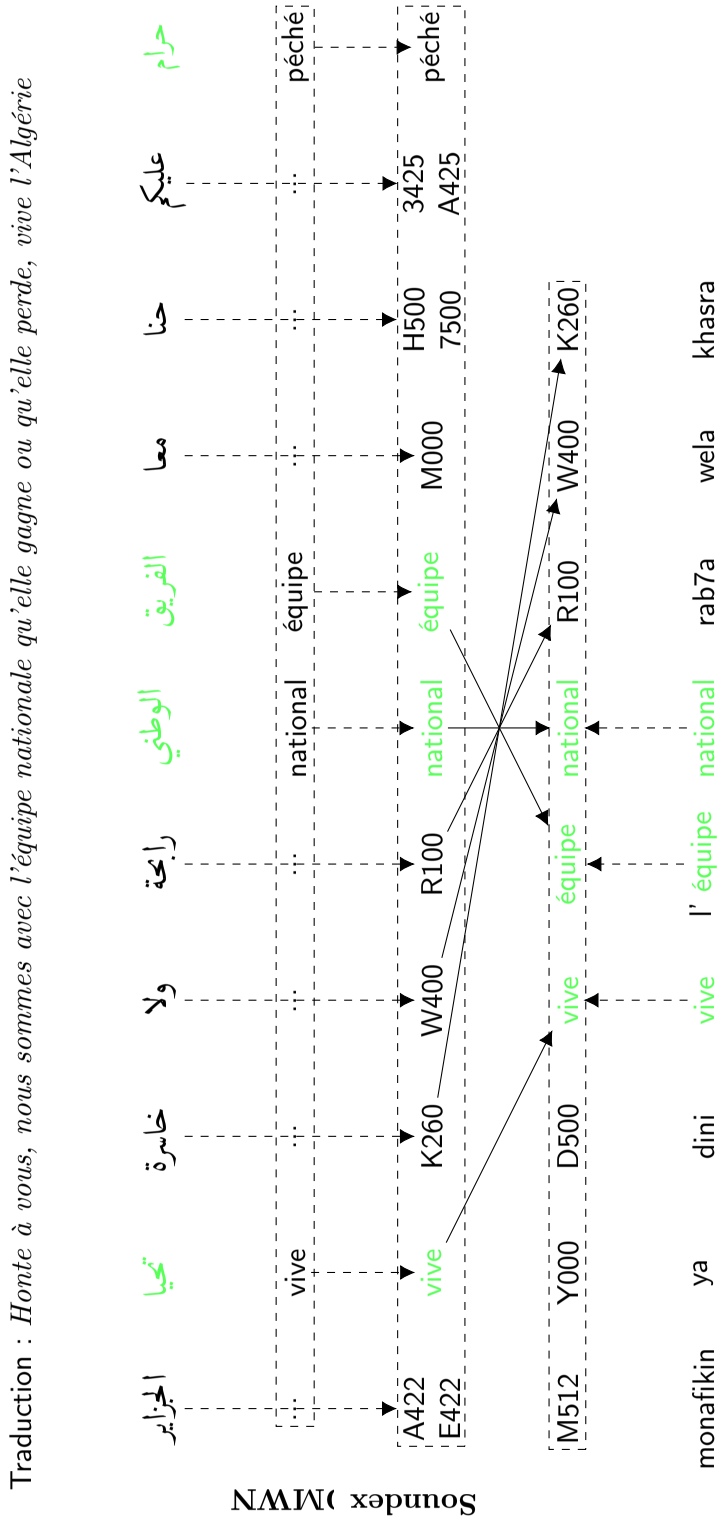


FIGURE 4.5 – Les mots qu'on peut faire correspondre avec le dictionnaire et la phonétique.

4.3.3 Une approche basée sur le *word embedding*

L'absence de règles linguistiques pour l'écriture du dialecte engendre au moins un problème, celui de la variabilité d'écriture d'un même mot. Cela entraîne évidemment un problème dans le traitement automatique du dialecte et notamment dans la création de vocabulaires. À titre d'exemple, le mot *kho*, en dialecte algérien, signifiant en français *frère*, possède plusieurs formes d'écriture. Il peut être écrit *khoo, khouya, kho, khyo* en script latin et خو, خويا en script arabe.

Il est donc indispensable de lier tous les mots qui correspondent à une même entrée lexicale. Par ailleurs, comme nous sommes confrontés au problème d'alignement de commentaires, il serait intéressant d'associer à cette même entrée tous les mots sémantiquement proches. Ainsi, le mot *hbibna*, dans le dialecte algérien peut être considéré, dans une certaine mesure, comme un mot proche, l'un pouvant se substituer à l'autre. Toujours dans cette optique d'identifier tous les mots proches, nous avons été confrontés, en traitant le corpus, à des écritures pour le moins surprenantes. Ainsi, en est-il de la chaîne de caractères *frr* correspondant au mot *frère* de laquelle les voyelles ont été supprimées comme on le fait en langue arabe. Ce problème est intéressant et il faut, par conséquent, le prendre en compte.

khoya	khoo	khouya	khou	khyo	kho	hbibna	frr	خويا	خو	حبيبنا

TABLE 4.11 – Variantes d'écriture et mots corrélés du mot **kho** en dialecte algérien.

Afin de construire un dictionnaire qui regroupe toutes les possibilités d'écriture d'un même mot ainsi que les mots qui lui sont proches sémantiquement, nous avons décidé d'utiliser l'approche *Word2vec* [Mikolov et al., 2013b] basée sur la représentation distribuée des mots qui permet de projeter les mots similaires dans le même espace. Ce modèle a montré des performances intéressantes dans de nombreux travaux en traitement automatique des langues. Ce modèle est fondé sur une projection des mots du vocabulaire dans un espace de faible dimension [Ghannay, 2017]. Si les mots apparaissent dans des contextes sémantiquement ou syntaxiquement proches, leurs vecteurs sont proches en termes de distance.

Nous avons utilisé le modèle CBOW (*Continuous Bag of Words model*) du *Word2vec* sur le corpus algérien que nous avons collecté (voir chapitre 2.9). Nous avons opté pour ce modèle parce que les tests que nous avons menés ont donné de meilleures performances avec l'utilisation de CBOW qu'avec le *Skip-gram*. Dans le processus d'apprentissage tous les commentaires d'une même vidéo sont concaténés afin d'associer un seul document par vidéo. Dans la figure 4.6 nous donnons une représentation graphique des mots proches de l'entrée arabe طبيب qui signifie en français *médecin*, obtenu avec la méthode CBOW du *Word2vec* en utilisant une fenêtre d'analyse de 100 mots avant et après le mot en cours de traitement et une couche cachée de 100 neurones. Ces paramètres ont été déterminés d'une manière expérimentale. Nous rappelons que la taille de corpus sur lequel l'apprentissage est effectué est de 17M de mots (voir section 2.9).

Ce modèle arrive donc à capturer pour un même mot, ses mots les plus proches y compris ceux qui sont exprimés dans une autre langue (*médecin et médecine*). L'analyse des corrélations obtenues a montré des relations pertinentes entre les mots et ceux qui

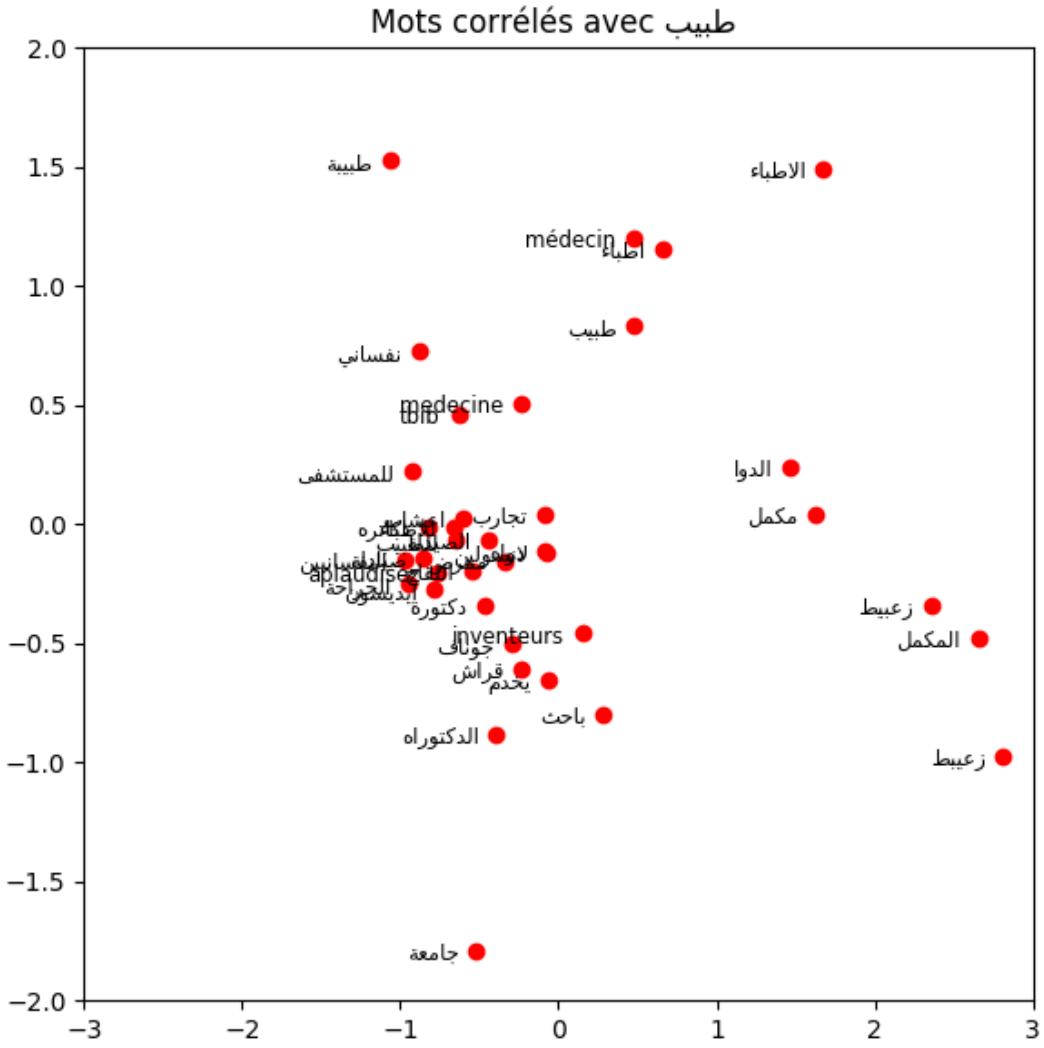


FIGURE 4.6 – Les mots corrélés avec le mot طبيب (médecin) obtenus avec CBOW.

leur sont très proches. C'est pourquoi, nous avons décidé de nous affranchir de l'utilisation d'un dictionnaire arabe-français pour aligner des textes. Au lieu de chercher les traductions possibles d'un mot dans un dictionnaire bilingue, on les cherche dans une liste de mots qui sont proches de celui-ci. Pour ce faire, pour chaque mot w_s écrit en caractère s, avec $s \in \{SA, SL\}$, où SA est le script arabe et SL est le script latin, nous cherchons ses n mots corrélés écrits en caractères $\bar{s}$. Par exemple, dans la liste de mots corrélés obtenues pour l'entrée طبيب, nous ne gardons que les mots proches écrits en script différent comme le montre la figure 4.7.

Partant de cette hypothèse, nous avons adapté la formule 4.1 puisqu'il ne s'agit plus de faire correspondre un mot avec un autre, mais de rechercher la correspondance dans une liste

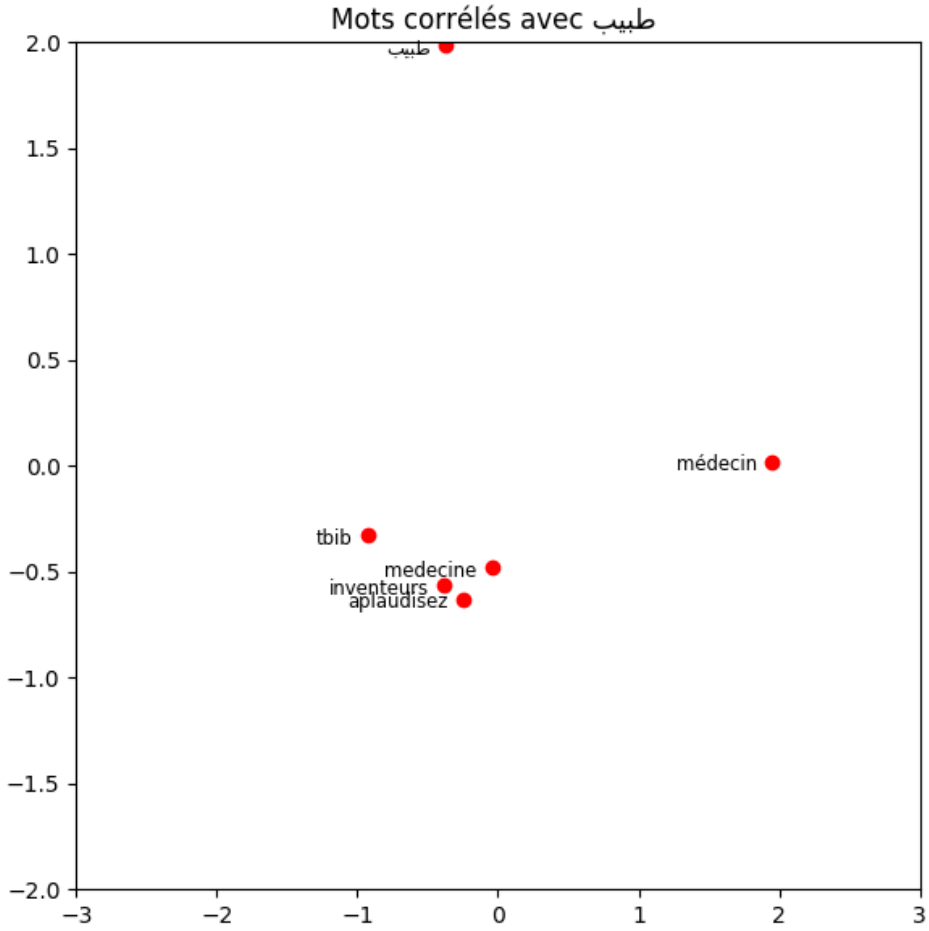


FIGURE 4.7 – Les mots corrélés avec le mot طبيب (*médecin*) écrits en caractères différents.

de mots corrélés comme le montre la formule suivante :

$$S_{W2V}(C_{LS}, C_{AS}) = \frac{\sum_{w \in C_{LS}} \sigma(w, C_{AS}) + \sum_{w \in C_{AS}} \sigma(w, C_{LS})}{|C_{LS}| + |C_{AS}|} \tag{4.5}$$

σ est une fonction qui renvoie 1 si un mot w appartenant à V_x (la liste des mots du document x) appartient à la liste des mots corrélés (notée $L(w)$) au mot w.

$$\sigma(w, V_x) = \begin{cases} 1 \ si \ L(w) \cap V_x \neq \emptyset \\ 0 \ sinon \end{cases} \tag{4.6}$$

4.3.4 Une approche itérative basée sur le *multilingual word embedding*

En utilisant les vecteurs produits par l'architecture CBOW du *Word2vec*, nous avons réussi à capter des mots qui sont sémantiquement proches. Cependant, nous avons remarqué que pour certains mots, particulièrement ceux qui ne sont pas fréquents dans le corpus, leurs

listes de mots corrélés comportent des mots qui ne devraient pas se retrouver dans ces listes. À titre d'illustration, prenons l'exemple précédent طبيب (*médecin*) les deux mots *inventeurs, aplaudisez*[24] ne sont proches ni sémantiquement ni syntaxiquement de طبيب. Par conséquent, il est important de raffiner cette liste afin d'améliorer la qualité de la comparabilité.

Comme nous l'avons mentionné précédemment, plusieurs travaux de recherche ont montré que la projection de mots de plusieurs langues dans le même espace vectoriel est efficace si le corpus utilisé pour faire l'apprentissage *word embedding* multilingue est parallèle. À part le corpus PADIC [Meftouh et al., 2018] dont une partie seulement comporte des phrases parallèles avec le français, il n'existe pas de corpus parallèle de taille importante comportant des phrases en dialecte et en français. Pour remédier à ce problème, il faudrait que le corpus sur lequel nous faisons l'apprentissage soit le plus comparable possible pour que l'apprentissage *word embedding* multilingue soit efficace. C'est pourquoi, nous avons décidé de répéter le processus d'alignement des commentaires du corpus d'apprentissage. De ce fait, à chaque itération nous utilisons le corpus comparable obtenu dans l'itération précédente pour apprendre les nouvelles listes de corrélation. Ce qui conduit à l'obtention d'un corpus dont les commentaires sont de plus en plus comparables et par conséquent à des listes de mots corrélés de plus en plus raffinées (voir la figure 4.8). Nous arrêtons les itérations une fois que le rappel calculé sur un corpus de développement commence à stagner.

Dans la figure 4.9, nous montrons l'évolution du mot طبيب (*médecin*) et ses mots corrélés sur quatre itérations de l'approche que nous venons d'exposer. Nous remarquons que l'itération 1 correspond au résultat obtenu dans la section 4.3.3. Lors de l'itération 2, on constate un rapprochement des mots liés au mot طبيب. Par ailleurs, de nouveaux mots proches ont fait leur apparition, par exemple : *médecins, patients, docteurs, généraliste, spécialité, temrad* (mot dialectal qui veut dire tomber malade), *diplôme, za3bit* (nom d'un médecin). De surcroît, des mots qui ne semblent pas être proches du mot étudié ont également fait leur apparition : *muslimah (musulmane)*. Nous pensons que les discussions ont porté sur une femme musulmane médecin, d'où l'introduction de ce mot. Pour les autres itérations on remarque l'introduction de nouveaux mots liés au mot طبيب.

24. applaudissez : on remarquera que ce mot est mal orthographié, nous l'avons reproduit tel qu'il a été écrit dans les commentaires de YouTube que nous avons collectés

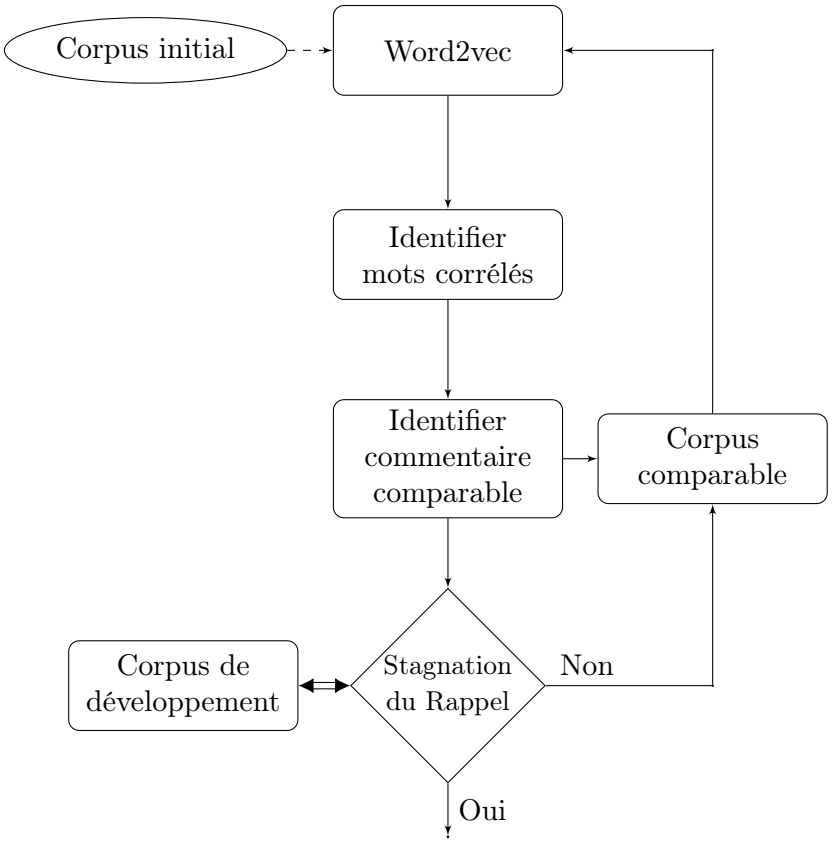


FIGURE 4.8 – Processus itératif de *mutlilingual word embedding*.

La question que l'on peut se poser est comment évaluer la qualité du dictionnaire construit de cette manière. Comme il n'existe pas de dictionnaire d'évaluation, nous avons vérifié à la main si les mots insérés à chaque itération étaient bien corrélés avec le mot طبيب. Ceci nous a permis de calculer un taux de mots correctement corrélés, taux que nous reportons dans la figure 4.10.

On remarque que lors de certaines itérations le taux de mots corrects baisse, mais que globalement ce taux augmente au fur à mesure et se stabilise à partir de la 17ème itération. Il est clair qu'il n'est pas possible de conclure à partir d'un exemple, mais cela nous a permis de nous faire une idée sur le processus d'insertion des mots corrélés.

4.3.5 Expérimentations

Nous évaluons ci-après les quatre méthodes présentées afin d'aligner les commentaires extraits de YouTube. Le test est effectué sur un corpus de YouTube aligné manuellement. Ce corpus est composé de 360 paires de commentaires comparables dont la partie source est composée de commentaires en script latin (français, anglais, ou arabizi) et la partie cible est constituée de commentaires en script arabe (dialecte ou MSA).

Le processus d'évaluation des méthodes proposées est réalisé comme suit : pour un commentaire source donné t_s, nous calculons la similarité avec les 360 commentaires cibles

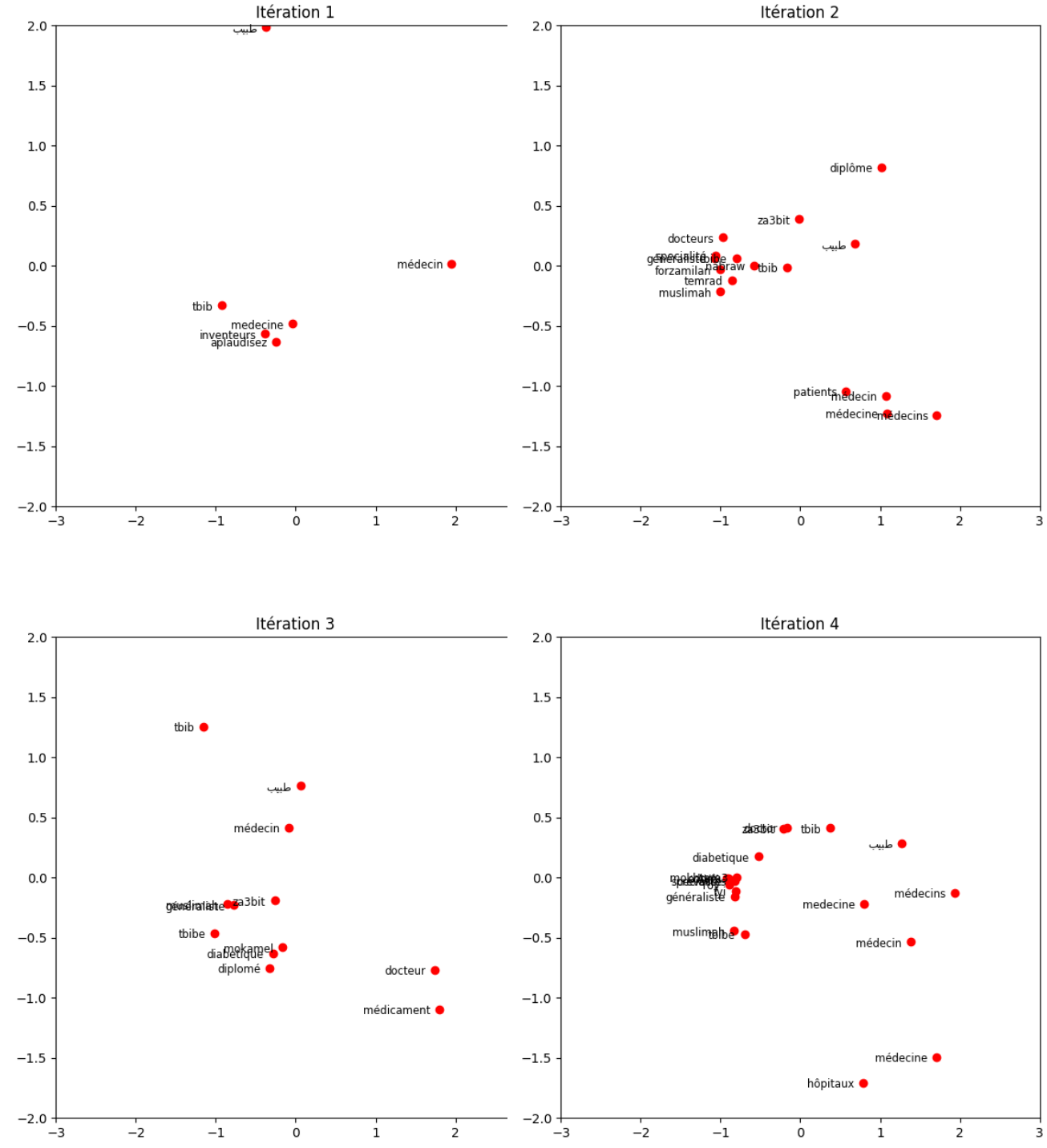


FIGURE 4.9 – Les listes de mots corrélés avec l'entrée طبيب obtenues lors des quatre premières itérations.

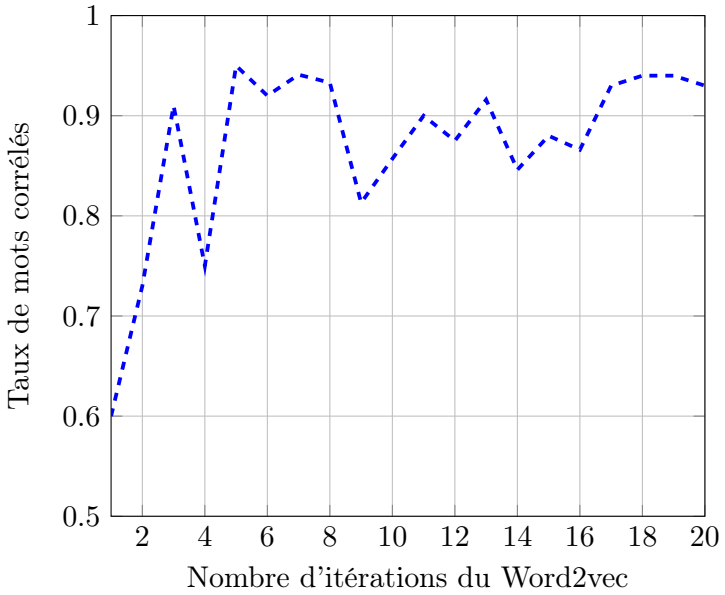


FIGURE 4.10 – Le taux de mots corrélés avec l'entrée طبيب durant le processus itératif du *Word2vec*.

t_c, puis nous sélectionnons les *n-top* commentaires cibles en fonction de la valeur de la similarité utilisée. Ensuite, nous vérifions si le texte cible correspondant est le premier dans la liste (R@1), dans les 5 premiers commentaires (R@5) et dans les 10 premiers commentaires (R@10). Dans le tableau 4.12, nous reportons les résultats en terme de Rappel.

Corpus	**R@1**	**R@5**	**R@10**
MLG	4	12	19
$MLG - Sound$	11	27	40
SC_{w2v}	23	42	51
$SC_{w2v} - Sound$	24	45	52
$Iterative\ \ SC_{w2v} + Sound$	**33**	48	54

TABLE 4.12 – Résultats obtenus par les méthodes proposées en terme de rappel R@1, R@5, et R@10 sur un corpus de développement construit manuellement.

On constate que la méthode la moins adaptée à l'alignement des commentaires de YouTube est la méthode classique basée sur le dictionnaire *MLG* puisqu'elle est fondée sur un dictionnaire multilingue (*Open multilingual WordNet*) utilisant l'arabe standard, le français et l'anglais. Comme les commentaires que nous avons collectés sont écrits en dialecte, en français et en anglais, les mauvais résultats étaient donc prévisibles.

La deuxième méthode *MLG-Sound* avait comme objectif d'améliorer les performances de la première en prenant en compte l'aspect phonétique des mots. En effet, la variabilité d'écriture des mots du dialecte nous a obligés à trouver une solution afin d'associer tous les mots correspondant à la même entrée. Cette méthode a permis d'améliorer le Rappel en rang 1 de 175%, mais les résultats demeurent très faibles.

En utilisant l'approche basée sur le *multilingual word embedding* nous avons considérablement amélioré les résultats (109% en rang 1). Ce qui correspond à un Rappel de 23% en rang 1.

Dans la quatrième expérience nous avons combiné l'algorithme Soundex avec le *Word2vec*, nous avons constaté une légère amélioration de 4% de R@1 et de 6% de R@5. On constate que l'impact de Soundex sur le *Word2vec* est limité puisque le *Word2vec* a cette capacité d'associer les mots morphologiquement proches d'un mot donné. Or, Soundex a été utilisé pour capter les similitudes phonétiques liées aux dérivations morphologiques.

Avec le processus itératif, nous avons amélioré significativement les résultats (33% en rang 1).

Dans l'annexe C nous présentons des exemples du corpus comparable que nous avons construit en utilisant le processus itératif *multilingual word embedding*.

4.4 Conclusion

Dans ce chapitre, nous avons décrit les méthodes que nous avons utilisées afin de construire automatiquement des corpus comparables à partir des réseaux sociaux. Partant de ce fait, deux corpus comparables ont été construits automatiquement : un corpus comparable de tweets pour les deux langues formelles (arabe standard et anglais) et un corpus comparable qui regroupe des commentaires de Youtube exprimés en langues formelle et en dialecte algérien.

Nous avons montré dans la première partie de ce chapitre que les données issues des réseaux sociaux notamment celles de Twitter nécessitent un pré-traitement, car contrairement aux données provenant des journaux, les tweets contiennent des données supplémentaires nécessaires à certains traitements comme les *hashtags*, les URL, les abréviations, etc. Pour utiliser les outils classiques de traitement automatique des langues, nous avons appliqué plusieurs pré-traitements sur les corpus collectés. Ces pré-traitements consistent en la réécriture de certains mots, le nettoyage de certains autres, l'homogénéisation de l'écriture des nombres, des dates, etc. Pour aligner les tweets comparables, nous avons utilisé une méthode basée sur le dictionnaire à laquelle nous avons ajouté des traitements particuliers concernant notamment : les noms propres, les dates, les nombres, etc. Ces informations ont permis d'améliorer les résultats de 5,6% par rapport à la méthode fondée seulement sur le dictionnaire.

Dans la deuxième partie de ce chapitre, nous avons abordé la question de l'alignement des données vernaculaires pour en faire un corpus comparable. Ce corpus concerne les commentaires qui peuvent être en : dialecte algérien, arabe standard, français et en anglais. Pour établir la correspondance entre les commentaires, nous avons testé la méthode basée sur un dictionnaire multilingue conçu pour les langues formelles. Malgré le fait que le dialecte algérien partage un nombre considérable de mots avec l'arabe standard, néanmoins les résultats ont été décevants. Pour améliorer les résultats nous avons décidé d'utiliser l'encodage phonétique des mots pour associer ceux qui sont proches au sens phonétique. L'utilisation de Soundex a permis d'améliorer les résultats de 7 points en terme de rappel. Cette performance est intéressante, mais néanmoins insuffisante. Ensuite, nous avons utilisé une approche fondée sur le principe du *Word2vec*. Cette méthode est intéressante parce

qu'elle permet de trouver pour un mot tous ceux qui lui sont proches du point de vue syntaxique, sémantique ou autres. En effet, cette liste permet d'associer plus facilement des commentaires possédant des mots provenant de cette même liste. Le résultat obtenu avec cette approche est meilleur comparé à ceux obtenus avec les deux méthodes précédentes. Nous avons ensuite itéré l'approche *Word2vec* sur le corpus de départ afin d'améliorer la qualité de la liste des mots corrélés et par conséquent, la qualité de la comparabilité du corpus de départ. En effet, cette méthode a permis une amélioration de 22 points en terme de rappel par rapport à la méthode basée sur le dictionnaire.

Nous avons finalement construit automatiquement un corpus comparable nommé CALYOU (*Comparable ALgerian harvested from YOUtube*) grâce à cette dernière méthode. Ce corpus sera disponible dans le site de l'équipe $SM_{ar}T$ [25].

25. https://smart.loria.fr/corpora/

5

La construction automatique de ressources lexicales pour les dialectes maghrébins

Sommaire

5.1 Introduction 78
5.2 Construction d'un lexique des différentes formes d'un même mot 78
5.2.1 La méthode d'extraction 79
5.2.2 Protocole d'évaluation 81
5.3 L'extraction de lexique de sentiments 85
5.3.1 La méthodologie proposée 85
5.3.2 Protocole d'évaluation 87
5.4 Conclusion 90

Dès qu'on s'intéresse au traitement automatique de n'importe quelle langue, la disponibilité de ressources lexicales présente une importance primordiale. De ce fait, de nombreux travaux de recherche ont proposé des méthodes automatiques ou semi-automatiques afin de créer ce type de ressources. Cependant, la plupart d'entre eux ont été limités à un nombre de langues occidentales comme la langue anglaise.

Dans ce chapitre, nous allons décrire notre démarche de création de deux lexiques pour les dialectes maghrébins. Le premier lexique présente une ressource regroupant les variabilités lexicales existant dans les dialectes, ce qui est l'objet de cette étude. Tandis que, le deuxième représente un lexique de sentiment dans lequel un degré de polarité est associé à chaque entrée.

5.1 Introduction

Dans de nombreuses applications du traitement automatique des langues, les ressources lexicales sont indispensables. La création de ces ressources est coûteuse parce qu'elle nécessite un travail de collecte, de regroupement et de vérification effectué par des spécialistes. Pour éviter la construction manuelle de ces ressources, plusieurs travaux de recherche ont entrepris d'automatiser leur construction. Comme on l'a vu dans le chapitre 1, ces travaux ont donné lieu à plusieurs lexiques, mais ils concernent particulièrement les langues formelles comme l'arabe standard, l'anglais, le français, etc. En revanche, pour les langues peu dotées comme les dialectes arabes, il existe très peu de lexiques.

Pour apporter notre pierre à l'édifice de construction des ressources, nous proposons dans ce chapitre d'en construire automatiquement deux pour trois dialectes du Maghreb. La première, regroupe les variabilités lexicales des mots du dialecte. Chaque entrée est composée d'un mot écrit en script arabe (MSA ou dialecte) ou en script latin (arabizi, français ou anglais) suivie de plusieurs variations d'écriture dans un script différent de celui de l'entrée. Ces variations d'écritures sont identifiées automatiquement en utilisant les corpus collectés précédemment. La deuxième ressource que nous proposons est destinée à l'analyse de sentiments où chaque entrée est composée d'un mot dialectal et de sa polarité.

5.2 Construction d'un lexique des différentes formes d'un même mot

L'avènement des réseaux sociaux a vu l'éclosion de l'utilisation de l'arabe dialectal. Les données textuelles produites constituent une mine pour le traitement automatique des langues. Cependant, comme l'expression est libre et que n'importe quel usager a la possibilité de s'exprimer, il en résulte des textes mal écrits, mal orthographiés et parce que le dialecte n'a pas de norme d'écriture, un même mot peut être écrit de plusieurs manières différentes. Afin d'illustrer ce problème, nous donnons dans le tableau 5.1 quelques exemples de mots extraits des trois corpus maghrébins présentés dans le chapitre 2. Il est à noter que, chaque ligne correspond aux différentes façons d'écrire le même mot en script arabe et latin.

	Écriture en script arabe	**Écriture en script latin**	**Traduction**
ALG	خوّويا, خويا, خوي, خوياا	khuya, khoya, 5oya, 5ouya, khouya, khoyà, khuya	Mon frère
MAR	لواليدة, الواليدة, الوليدة, الوالدة لوليدة, لوالدة	lwalida, lwlida, lwalyda, lewalida, lwaleda, lwalidaa, alwalida, elwalida, olwalida	La maman
TUN	ممصطك, ممصتك, ممستك, ممسطك	mamstek, mamestek, mamsteek, mamstk, maamstek	Comme tu es chiant

Table 5.1 – Quelques exemples montrant la variabilité lexicale existant dans les dialectes maghrébins.

Ces exemples montrent clairement la diversité d'écriture des mots dans les trois dialectes du Maghreb. On remarquera aussi à travers ces exemples que le nombre de possibilités

d'écriture en script latin est plus important qu'en script arabe. Cela est naturel parce qu'il y a plusieurs manières d'écrire une lettre arabe qui n'a pas de correspondant en script latin. Par ailleurs, lorsqu'on écrit de l'arabe avec le script latin on a tendance à ajouter les voyelles comme si l'on écrivait en arabe. Dans ce cas, les usagers les rajoutent chacun à sa manière. À titre d'exemple, prenons le mot لواليدة qui signifié *la mère* en dialecte marocain. On constate qu'il y a 15 manières d'écrire ce mot dont 9 en script latin.

De surcroît, cette variabilité ne se limite pas aux dialectes, elle concerne également les mots mal orthographiés des langues formelles utilisées avec les dialectes. Nous donnons, à titre d'illustration, quelques exemples de mots mal écrits dans le tableau 5.2.

Mots en langue formelle	**Écriture correcte**	**Langue**
commentair	commentaire	français
mester, mister, mstr	mister	anglais
film, flm	film	français
فليم *flym*, فيلم *fylm*, فلم *flm*	فلم *flm*	arabe standard

TABLE 5.2 – Quelques exemples de mots mal orthographiés.

Partant de ce constat, nous proposons une méthode permettant d'identifier les différentes formes d'écriture d'un mot à partir de nos corpus. À notre connaissance, ce type de lexique n'existe pas. L'avantage d'une telle ressource est d'améliorer les performances de plusieurs systèmes de traitement automatique des langues en homogénéisant l'écriture des mots.

Nous présentons brièvement dans ce qui suit, les processus d'extraction de ces formes graphiques à partir des trois corpus du Maghreb.

5.2.1 La méthode d'extraction

Le processus de construction du lexique des différentes formes d'un même mot est fondée sur la même méthode itérative ayant permis de construire le lexique qui a servi pour l'alignement des corpus comparables présentée dans le chapitre précédent (voir la section 4.3.4). Ce processus se déroule comme suit. À chaque itération, deux lexiques sont produits : un comprenant des mots corrélés qu'on utilise pour l'alignement des commentaires et un deuxième comportant les différentes formes d'un même mot. À chaque itération chacun de ces lexiques est affiné. Dans ce qui suit, nous détaillons juste la construction du deuxième.

Pour chaque mot w^s d'un corpus, n mots corrélés sont sélectionnés en gardant les mots les plus proches au sens de la distance cosinus entre leurs vecteurs représentatifs. Le mot w^s sera associé à des mots écrits en script arabe et en script latin. Nous avons sélectionné pour chaque mot, les 40 mots les plus proches. Pour ce deuxième lexique, ce qui nous intéresse est d'établir une liste de mots proches au sens graphie et non pas de garder la liste de tous les mots corrélés. Afin d'obtenir ce lexique, nous appliquons deux traitements à la liste des mots corrélés $Lc(w^s)$.

Traitement concernant les mots de même script

De la liste $Lc(w^s)$ nous gardons les mots de même script et vérifiant la contrainte suivante : $R(w^s) = R(w_j^s)$ avec $w_j^s \in Lc(w^s)$. $R(x)$ est une fonction qui supprime les voyelles de x. Pour le script arabe, la suppression concerne les voyelles longues représentées par les lettres suivantes : ي, ا, و. Les voyelles courtes ne sont généralement pas utilisées dans les corpus extraits des réseaux sociaux. Un exemple du résultat de ce traitement sur les différentes formes d'écriture de الوالدة est donné dans le tableau 5.3. Toutes ces formes sont encodées par (ل ل د ة)

Possibilités d'écriture	Encodage
لوليدة, لوالدة, الوالدة, الوليدة, الواليدة, لواليدة	ل ل د ة

TABLE 5.3 – Des exemples des mots après la suppression des voyelles arabes.

Pour les mots écrits en script latin, on procède à un encodage interne de certains chiffres ou combinaisons de lettres qui sont utilisés pour la représentation des sons arabes écrits en script latin avant la suppression des voyelles. Le tableau 5.4 donne la liste des entrées et l'encodage opéré. Par exemple, si l'on rencontre les trois caractères *9, k, c* qui correspondent au son représenté par la lettre arabe ق, on les remplace par lettre *q*.

Lettre arabe	Entrée	Encodage
ا	2, e, i	a
ب	p	b
ج	dj	j
خ	5	kh
ش	sh	ch
ط	6	t
ع	', 3,e	a
ف	v	f
ق	9, k,c	q
ك	c	k
ة	a	h
و	ou, u	w
ى	i, a	y

TABLE 5.4 – Encodage des caractères utilisés en arabizi.

Dans le tableau 5.5, nous donnons un exemple du résultat de ce traitement sur des mots écrits en script latin.

Possibilités d'écriture	Les consonnes communes
khuya, khoya, khouya, khoyà, khuya, 5oya, 5ouya	kh y

TABLE 5.5 – Des exemples des mots après le traitement.

Traitement concernant les mots de scripts différents

Dans le traitement précédent, nous nous sommes concentrés sur les mots de même script. Dans celui-ci nous allons proposer une procédure permettant d'associer un mot dans un script s à un ou plusieurs mots écrits dans un script différent $\bar{s}$. Un mot w^s sera associé à un mot $w_i^{\bar{s}}$ écrit dans un script différent si sa translittération $T(w_i^{\bar{s}})$ est identique à ce même mot.

$$\exists i,\ T(w_i^{\bar{s}}) = w^s \ et\ w_i^{\bar{s}} \in Lc(w^s)$$

La translittération est effectuée en utilisant la table en annexe B. Nous donnons ci-dessous le résultat de l'application de ce traitement au mot écrit en script arabe يرحمك :

يرحمك : yr7mk yr7mak yrhamak yarhamek yarhemak yarhamk yarhmek yr7mek yere7mek yarhamak yarhemek yrhmk yar7mak yarhmak yer7mak yarahmak yar7mek yarahmk yerhemek yarahmek yerehmek yerhamek yar7mik yare7mek yerhamak yer7mek yerehemek yarhmeke rahimaka yrahmek yrahmak irahmak irhmak irahmek irhmk yra7mk yerahmak yrehmak yera7mak yerehmk yrhmak yera7mek yrehmek yara7mak yarehmek yara7mek yerahmeke yrhmek yarehmak yarhmk yerhmk yarhmeek yra7mak yerahmek ir7mak yra7mek yrahmk yarhamoka yrehmk yar7mk yerhmk ira7mak irehmek yerhmek yarahemek yerahmk

TABLE 5.6 – Différentes façons d'écrire le mot يرحمك (*Dieu vous bénisse*) en script latin.

L'application des traitements expliqués précédemment nous a permis de produire des lexiques des différentes graphies d'un même mot et ce pour chacun des trois dialectes du Maghreb. Les tableaux 5.7, 5.8 et 5.9 font référence respectivement à quelques exemples construits automatiquement pour l'algérien, le marocain et le tunisien.

Entrée	**Écritures possibles**
خويا	khuya, khoya, 5oya, 5ouya, khouya, khoyà, khuya
mister	ميستر ,مستر
مانسوطيش	mansotich, mansotiiche, mansautich, mnsotich, mansoutich, mansotiche, mansotiwch, manesotich, manssotiche, mansoutiche, manssotich
فيلم	film, flm, film

TABLE 5.7 – Exemples de quelques entrées du lexique algérien.

Dans le tableau 5.10 nous donnons la taille des trois lexiques construits automatiquement pour les trois dialectes.

5.2.2 Protocole d'évaluation

Deux types de méthodes d'évaluation des lexiques existent dans la littérature : l'évaluation manuelle avec des experts et l'évaluation automatique en utilisant les mesures classiques : Rappel, Précision et F-mesure. À notre connaissance, il n'existe pas de lexique de ce type

Entrée	Écritures possibles
ماتديهاش	omatedihach, matedihach, mtdihach, omatdihach, matdihach matadihach, matdihache
كاتفاجي	katfewji katfawji katfwji
منتفرج	mantfarej, mantfaraj, mantfarj, mantfraj, mntfrj, mantfrej mantfrj, manatfaraj
zewine	زوين, زويّين

TABLE 5.8 – Exemples de quelques entrées du lexique marocain.

Entrée	Écritures possibles
مسط	massit, msata, masta, masset, amsat, mastt, msataa, amset mast, mastaa, maset, masett, mssata, maseett, maaset
لغناية	lghnaya lghneya lghoneya laghneya
barbara	برابرا, باربرا, بربرا
ma7leha	محلاها, محلااه, محلااه, ماحلاها, مااحلاها, محلاه

TABLE 5.9 – Exemples de quelques entrées du lexique tunisien.

	Algérien	Marocain	Tunisien
Nombre d'entrées	6762	6715	3677

TABLE 5.10 – La taille des trois lexiques.

afin de l'utiliser comme un lexique de référence. Par ailleurs, il est difficile de trouver des experts en dialecte capable d'évaluer les lexiques que nous avons produits. Par conséquent, nous avons procédé à l'évaluation automatique qui consiste à comparer les entrées du lexique construit avec celles d'un lexique de référence créé manuellement.

En raison de la difficulté de création manuelle d'un lexique de référence, nous avons décidé de n'évaluer que le lexique algérien en utilisant un lexique de référence construit semi-automatiquement.

Dans les paragraphes qui suivent, nous détaillons le protocole d'évaluation de ce lexique.

La création d'un lexique de référence

Pour construire un lexique de référence pour le dialecte algérien, nous avons sélectionné aléatoirement une liste de mots en script arabe et en script latin à partir du corpus algérien. Pour chaque mot w de cette liste, nous recherchons tous les mots du corpus dont le code délivré par Soundex (voir section 4.2.3) est le même que celui de ce mot. Autrement dit, on construit une liste $L(w)$ avec $\forall w_i \in L(w) \Rightarrow S(w) = S(w_i)$, $S(w)$ est la fonction Soundex qui renvoie un code phonétique pour w. Les mots vérifiant cette contrainte sont mis dans une liste $L_S(w)$. Nous procédons ensuite au filtrage manuel de celle-ci pour ne garder que les mots qui correspondent réellement à des variantes d'écriture du mot w. Par ailleurs, nous ajoutons à la main les variantes d'écriture qui n'ont pas été récupérées par Soundex. En fait, cette

méthode nous permet de ne pas démarrer la création de cette liste de référence à partir de zéro.

Pour illustrer ce traitement, dans le tableau 5.11, nous donnons un exemple de mots candidats obtenus avec Soundex pour le mot يشفيها (*Il la guérit*).
Dans le tableau 5.11, les mots qui sont en gras ne correspondent pas à des variantes d'écriture

يشفيها	*ychafih* **yechf ychoufo ychouf yechfo** *yachfih* ychafiha yachfiha yechafiha ychfiha *ychafih* *yachefih ychafi*

TABLE 5.11 – Un exemple des candidats potentiels obtenus grâce à l'encodage de Soundex.

de cette entrée, ceux qui sont en rouge sont des mots de la même famille que l'entrée mais ne correspondant pas non plus à des variantes de l'entrée. Ils seront donc tous supprimés. Il n'y a que les mots en italique qui correspondent bien à l'objectif recherché. On ajoutera aussi d'autres variantes que nous jugeons utiles. Le tableau 5.12 indique les variantes d'écriture du mot يشفيها conservées.

يشفيها	ychafiha yachfiha yechafiha ychfiha yechfiha ychfeha yechfeha ychfha ychafeha

TABLE 5.12 – Un exemple d'une entrée du lexique de *Word2vec* référence.

Ce processus a conduit à la construction d'un lexique de référence composé de 560 entrées, avec un nombre moyen de formes graphiques par entrée de 6, un maximum de 17 et un minimum de 1.

L'évaluation de la méthode de construction du lexique

La figure 5.1 montre l'évolution des trois valeurs : Rappel, Précision et F-mesure en fonction du nombre d'itérations du processus décrit précédemment de création du lexique. Les courbes montrent clairement que les trois mesures progressent d'une manière continue. À partir de l'itération 17, les trois valeurs commencent à se rapprocher les unes des autres, et à partir de l'itération 20, la précision commence à faiblir par rapport au Rappel. Comme ce qui nous intéresse est de réduire les faux positifs, nous avons donc favorisé la Précision, c'est pourquoi nous avons décidé d'arrêter le processus itératif à la vingtième itération.

Nous constatons qu'au début du processus d'apprentissage, la F-mesure est de 21%. Cela est dû au fait que le corpus de départ utilisé pour l'apprentissage des vecteurs représentatifs des mots n'était pas comparable. On remarquera que lors des itérations suivantes la F-mesure s'améliore progressivement ce qui augure d'une meilleure qualité d'alignement du corpus d'apprentissage.

Dans la figure 5.2, nous étudions l'évolution des entrées ajoutées au lexique lors du processus itératif du *Word2vec*.

Cette figure montre que le nombre d'entrées du lexique augmente en fonction du nombre d'itérations du *Word2vec*. Nous remarquons que 85% des entrées de ce lexique ont été ajoutées lors des dix premières itérations du *Word2vec*. Ce qui montre que l'algorithme est

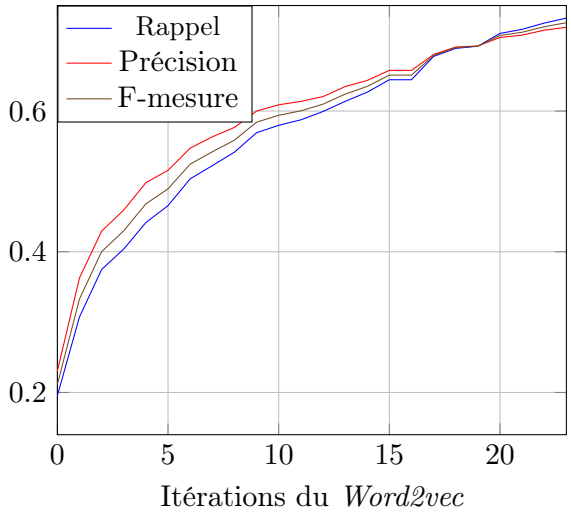


FIGURE 5.1 – l'évolution des rappels, de la précision et de la F-mesure en fonction des itérations *Word2vec*.

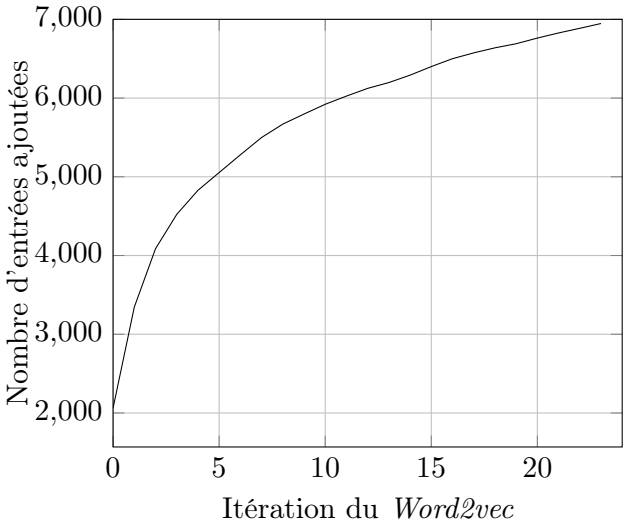


FIGURE 5.2 – L'évolution du nombre d'entrées du lexique en fonction des itérations.

capable de récupérer une majorité des mots en une dizaine d'itérations.

Dans l'expérience de la figure 5.3, l'objectif est de connaître la distribution du nombre d'entrées ajoutées à la première et à la dernière itération comportant n formes. On remarque que lors de la première itération du *Word2vec* lorsque le corpus n'est pas comparable le nombre d'entrées ayant n formes d'écriture est plus faible que le nombre d'entrées lors de la dernière itération. En effet, pour les entrées n'ayant qu'une forme d'écriture l'augmentation est de 28,5% et pour les entrées ayant entre 5 et 10 formes d'écriture la croissance est de 20,8%.

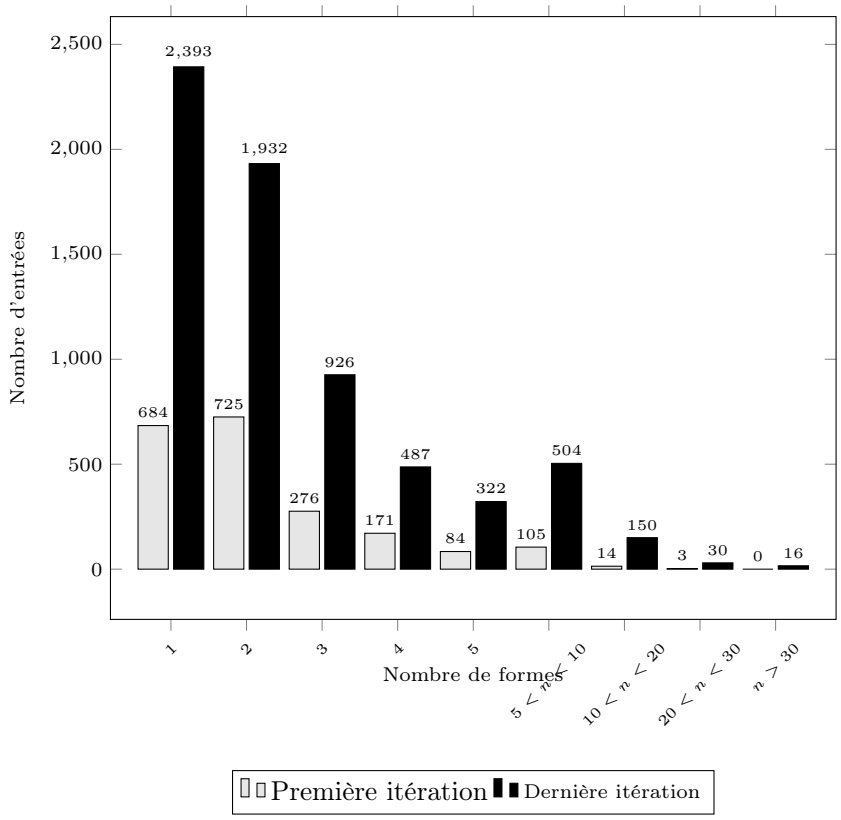


FIGURE 5.3 – La progression du nombre de formes ajoutées entre la première et la dernière itération.

5.3 L'extraction de lexique de sentiments

Dans cette section, nous décrivons notre approche pour construire des ressources lexicales destinées à l'analyse de sentiments pour les dialectes du Maghreb. Quelques travaux ont proposé de créer manuellement ou semi-automatiquement des lexiques de ce type pour certains dialectes du Maghreb [Ameur et al., 2016b],[Mataoui et al., 2016] et [Guellil et al., 2018]. Malgré la prédominance de l'arabizi dans les corpus des réseaux sociaux, la majorité des ces travaux est dédiée aux mots dialectaux écrits seulement en script arabe. Nous nous proposons de construire automatiquement un lexique de sentiments par dialecte (algérien, marocain et tunisien) comportant des mots en arabizi et en arabe. Cette approche est fondée sur la similarité entre les mots, comme dans les travaux de [Htait et al., 2017] et [Turney and Littman, 2003] pour les langues formelles.

5.3.1 La méthodologie proposée

Pour déterminer la polarité des mots des dialectes nous nous basons sur la proximité de leur orientation avec celle des mots de base que nous appellerons plus loin des mots-germes. À notre connaissance, il est difficile de trouver des travaux identifiant automatiquement la polarité des mots des dialectes arabes. Pour remédier à ce manque, nous proposons une méthode qui s'inspire des travaux de [Turney and Littman, 2003] et [Htait et al., 2017] qui ont été appliqués à la langue anglaise. La méthode que nous proposons est composée de deux

sous-tâches :

Identification des mots germes

L'idée consiste à estimer la polarité d'un mot en se basant sur celles des mots d'une liste établie en amont comme présentée dans [Turney and Littman, 2003]. Les mots de cette liste sont appelés *mots germes* ou (*seed words*). Dans [Turney and Littman, 2003], les mots germes sont ceux dont la polarité est évidente. Dans [Htait et al., 2017], en plus des mots identifiés par Turney, les auteurs ont ajouté une liste d'une quarantaine de mots-germes identifiés à partir des mots les plus fréquents. Ces mots ont été étiquetés manuellement par les auteurs.

Pour ce qui nous concerne, nous avons sélectionné, à partir d'une liste des mots les plus fréquents de chaque corpus collecté dans la chapitre 2, quatre-vingt mots-germes. Ces derniers correspondent aux mots pour lesquels il n'y a aucune ambiguïté de polarité. Nous leur avons donc affecté les polarités adéquates. Le nombre de mots-germes retenus est relativement élevé par rapport à l'expérience effectuée par les auteurs [Htait et al., 2017] parce que nos lexiques de dialecte comportent des mots écrits en script arabe et en script latin. Dans le tableau 5.13 nous donnons quelques exemples de ces mots-germes retenus.

Algérien		**Marocain**		**Tunisien**	
Positif	**Négatif**	**Positif**	**Négatif**	**Positif**	**Négatif**
chaba	شيات	روعة	tfou	barcha	ماسط
bravo	mok	hbiba	ينعل	ma7lek	جاهل
هايل	harki	tbarklah	himar	تهبل	حيوان
الصحة	جاهل	كنحماق	حرام	حلوه	تافه

TABLE 5.13 – Quelques exemples de mots germes positifs et négatifs pour les trois dialectes.

Estimation de la polarité des mots du lexique

Dans [Turney and Littman, 2003] les auteurs estiment qu'un mot est positif s'il est proche des mots-germes positifs. De la même manière un mot est négatif s'il est proche de mots germes négatifs. L'orientation d'un mot est calculée sur la base de la différence entre son association avec les mots-germes positifs et les mots-germes négatifs, comme le montre l'équation 5.1 :

$$SO(w) = \sum_{w_p \in MGP} sim(w, w_p) - \sum_{w_n \in MGN} sim(w, w_n) \tag{5.1}$$

Où MGP et MGN correspondent respectivement à la liste des mots-germes positifs et négatifs.
Le calcul de la similarité est effectué en utilisant l'information mutuelle. Pour ce qui nous concerne, les mots dont nous voulons identifier l'orientation sont représentés avec le principe du *word embedding*. C'est pourquoi, dans notre cas nous calculons la similarité cosinus entre les vecteurs représentatifs des mots.

La méthode proposée nous a permis de construire trois lexiques de sentiments pour les trois dialectes du Maghreb. Dans le tableau 5.14 nous donnons la taille de ces lexiques.

	Algérien	**Marocain**	**Tunisien**
Nombre d'entrées	11243	23405	10810

TABLE 5.14 – La taille des trois lexiques de sentiments.

Nous donnons dans le tableau 5.15 quelques exemples des entrées positives et négatives pour les trois lexiques.

Algérien		**Marocain**		**Tunisien**	
Positif	**Négatif**	**Positif**	**Négatif**	**Positif**	**Négatif**
انموت *"j'adore"*	terroristes	ne9iya *"propre"*	خبثاء *"salauds"*	to9tool *Une tuerie*	ياعفن *"espèce de pourriture"*
ywefe9 *"vous aide"*	جردان *"des rats"*	drayfa *"gentille"*	talba *"Mendicité"*	باهي *"bien"*	msaték *"tu es chiant"*
inawrek *"Illumine ton chemin"*	leklab *"des chiens"*	lghzalin *"les beaux"*	balkhl3a *"apeuré"*	damour *"d'amour"*	chmdakhlek *"ça ne te regarde pas"*
rawea *"splendide"*	roukhs *"sans valeur"*	كنبغيك *"je t'aime"*	جيفة *"charognard"*	تخبل *"magnifique"*	التعالب *"des malins"*
حفضك *"te protège"*	kref *"ordure"*	3aafak *"stp"*	الملعون *"Le maudit"*	maalemm *"chef"*	ياملهط
mignon	القزم *"nain"*	خموس ...	كيكدب *"il ment"*	ifadalek ...	المنحط *"inepte"*
twa7achtek *"tu me manques"*	khanez *"pourri"*	katfwej ...	b9ra *"vache"*	machaallah ...	7ramm *"péché"*
مفخرة *"fierté"*	شيات *"lèche bottes"*	حدكات *"dynamique"*	برهوش *"Gamin"*	bitawfik *"bon courage"*	mahboul *"malade mental"*
yabark *"bénisse"*	mosakh *"sale"*	kan7am9 *"j'adore"*	المافيات *"les mafias"*	محلاها *"trop belle"*	باشع *"horrible"*
ميرسي *"merci"*	نعجة *"faible"*	كتحمق *"elle est magnifique"*	كنحسد *"J'envie"*	بشفا *"bon appétit"*	همج *"J'envie"*

TABLE 5.15 – Quelques exemples de mots positifs ou négatifs extraits des trois lexiques de sentiments construits automatiquement.

5.3.2 Protocole d'évaluation

Dans la section précédente nous avons proposé une méthode qui nous a permis de construire automatiquement trois lexiques de sentiments pour les trois dialectes étudiés dans cette thèse. L'examen minutieux de ces lexiques a montré que les contenus sont très cohérents en terme de polarité positive ou négative, même si notre étude a montré que certains mots ne semblent pas avoir la bonne polarité. Par exemple le mot اتحوسو dans le dialecte algérien peut correspondre à deux acceptions. Il peut signifier *vous vous promenez* qui est plutôt positif, mais il peut aussi vouloir dire *vous cherchez*. Dans notre lexique ce mot a eu une polarité négative, c'est probablement son contexte d'utilisation qui a fait qu'il soit négatif. Par exemple, اتحوسو على المشاكل veut dire *vous cherchez des problèmes*. Dans ce cas, ce mot est bien négatif. Évidemment, on ne peut pas se contenter de cette analyse, il serait intéressant d'évaluer ces lexiques. Pour ce faire, il faut trouver des corpus dialectaux étiquetés en terme de polarité, une tâche qui est plutôt difficile puisqu'on travaille avec des dialectes. Nos recherches nous ont permis de trouver un corpus pour le dialecte marocain et

un pour le dialecte tunisien. Pour l'algérien nous n'avons malheureusement pas eu la même chance. Par conséquent nous avons dû en construire un. Dans la suite nous décrivons ces trois corpus.

— *ElecMorocco* : ce corpus a été collecté et annoté par les auteurs de [Elouardighi et al., 2018]. Il regroupe des commentaires en dialecte marocain publiés sur Facebook et concernant les élections marocaines. Ce corpus a été annoté manuellement et a donné 6431 commentaires négatifs et 3523 commentaires positifs. Notons que ce corpus contient seulement des commentaires écrits en script arabe.

— *TSAC* : ce corpus a été construit par les auteurs de [Medhaffar et al., 2017] à partir des commentaires postés sur des pages officielles Facebook des radios et des chaînes de télévision tunisiennes, à savoir Mosaique FM, JawhraFM, Shemes FM, HiwarElttounsi TV et Nessma TV. Ce corpus est composé de 7154 commentaires positifs et 6514 commentaires négatifs écrits en script arabe et latin.

— *SentAlg* : nous avons extrait de YouTube 750 commentaires en dialecte algérien que nous avons annotés manuellement ce qui a donné lieu à 390 commentaires positifs et à 360 commentaires négatifs.

Pour calculer la polarité d'un commentaire nous sommons la polarité de chacun de ses termes. Lorsqu'aucun mot d'un commentaire n'existe pas dans le lexique de sentiments, le commentaire se voit affecter l'étiquette *indécis*. Les résultats en terme de rappel et de précision sont donnés dans le tableau 5.16.

SentAlg		**ElecMorocco**		**TSAC**	
Rappel	**Précision**	**Rappel**	**Précision**	**Rappel**	**Précision**
88,11%	88,64%	59,29%	63,23%	64,03%	63,78%

TABLE 5.16 – Résultats expérimentaux sur les corpus des dialectes du Maghreb.

Même si la comparaison est difficile à faire, nous constatons que les résultats pour l'algérien en termes de rappel et de précision sont élevés. Ce corpus de faible taille par rapport aux deux autres a permis d'obtenir de très bons résultats, néanmoins il est difficile de les situer par rapport aux deux autres. Les résultats pour le tunisien sont assez bons en terme de rappel et de précision. Quant au corpus marocain, le rappel est moyen et la précision est assez bonne. Une analyse de ce corpus nous a montré que ce dernier contient beaucoup de phrases en arabe standard (MSA) et aussi beaucoup sont code-switchées (dialecte et MSA). En effet, *ElecMoroco* est un corpus dont le thème est la politique marocaine, il est donc naturel que les commentaires soient rédigés en arabe standard. Par ailleurs, le lexique marocain que nous avons développé ne comporte pas beaucoup de mots en arabe standard. Ceci est dû au fait que le corpus d'apprentissage ayant servi à la création de ce lexique ne comporte pas beaucoup de commentaires en arabe standard comme nous l'avons montré dans la section 2.4. En effet, seulement 27,3% du corpus global est écrit en MSA.

Afin d'analyser plus finement ces résultats, nous avons décidé d'étudier les erreurs de notre système de classification de sentiments comme le montre le tableau 5.17. Pour le marocain, notre système de classification a réussi à identifier 32% des commentaires positifs. Les faux positifs représentent plus de 75% des vrais positifs. Autrement dit, sur ce corpus

notre système de classification est permissif quant à l'identification des commentaires positifs. Concernant les commentaires négatifs, notre système est plus performant puisqu'il arrive à identifier plus de 86% des commentaires négatifs. Quant aux commentaires identifiés comme négatifs alors qu'ils ne le sont pas, ils représentent 42% des vrais négatifs. Le taux des faux négatifs est donc plus faible que le taux des faux positifs. Nous remarquons que notre système de classification est indécis pour seulement 0,32% des commentaires. Autrement dit, il est incapable d'attribuer une étiquette positive ou négative.

Pour le corpus tunisien, notre classifieur identifie 54% des commentaires positifs. Les faux positifs représentent 17% des commentaires réellement positifs. Quant aux commentaires négatifs, le classifieur arrive à identifier 74% de ces commentaires. Alors que pour les faux négatifs, ils représentent 39% des commentaires réellement négatifs. Quant aux indécis, ils représentent 17%.

En comparant ces derniers résultats à ceux obtenus pour le corpus marocain, nous constatons un meilleur rappel des positifs et un taux beaucoup plus faible des faux positifs. Pour les négatifs, le rappel est élevé comme pour le marocain et le taux des faux négatifs est similaire. En conclusion, en utilisant la même méthode nous obtenons de meilleurs résultats, ce qui renforce l'hypothèse que le corpus *ElecMoroco* n'est pas adapté à ce type de test. Nous pouvons formuler une dernière remarque concernant les indécis du corpus tunisien, en effet le taux est très élevé par rapport à celui de marocain.

Pour le petit corpus du dialecte algérien, le rappel des commentaires positifs est de 92% avec un taux de faux positifs de 16%. Le rappel des commentaires négatifs est de 84% avec un taux de faux négatifs de 10%. Nous remarquons que le taux des indécis est de 0%.

Il semblerait que nos classifieurs sont meilleurs pour l'identification des commentaires négatifs. Cela est dû probablement aux lexiques que nous avons construits automatiquement à partir des commentaires des réseaux sociaux dans lesquels naturellement il y a énormément d'échanges houleux.

Corpus	**Positif**		**Négatif**		**Indécis**	
	Vrais positifs	Faux positifs	Vrais négatifs	faux négatifs	Positif	Négatif
SentAlg	358	57	303	30	0	0
ElecMorocco	1132	852	5560	2373	18	19
TSAC	3865	674	4823	1916	1373	1017

TABLE 5.17 – Une analyse détaillée des résultats d'analyse de sentiment sur les trois corpus des dialectes du Maghreb.

Afin de réaliser une comparaison sur des corpus de tests plus équilibrés en terme de taille, nous avons sélectionné aléatoirement à partir des corpus *ElecMorocco* et *TSAC* un nombre de commentaires positifs et négatifs similaire à celui de *SentAlg*. Les résultats sont donnés dans le tableau 5.18 et dans le tableau 5.19. On constate que les classifieurs se comportent de la même manière que pour les corpus d'origine. Autrement dit, les résultats de l'algérien sont devant le tunisien et ceux du marocain derrière le tunisien. On constatera une baisse considérable des indécis pour le tunisien et le marocain qui ne représentent que 0,4%.

SentAlg		ElecMorocco		TSAC	
Rappel	**Précision**	**Rappel**	**Précision**	**Rappel**	**Précision**
88,11%	88,64%	57,99%	59,79%	74,36%	78,64%

TABLE 5.18 – Résultats expérimentaux sur des corpus de même taille de trois dialectes.

Corpus	**Positif**		**Négatif**		**Indécis**	
	Vrais positifs	Faux positifs	Vrais négatifs	faux négatifs	Positif	Négatif
SentAlg	359	57	303	30	0	0
ElecMorocco	139	70	290	249	2	1
TSAC	210	17	343	179	2	1

TABLE 5.19 – Une analyse détaillée des résultats d'analyse de sentiment sur des corpus de même taille.

5.4 Conclusion

L'objectif de ce chapitre a consisté à proposer des ressources lexicales pour les dialectes : algérien, marocain et tunisien. Pour ce faire, nous avons proposé deux méthodes fondées sur le *word embedding* afin de les construire automatiquement.

La première méthode a consisté à produire un lexique pour chaque dialecte comportant pour chaque entrée, les différentes formes d'écriture de celle-ci. Ces lexiques ont été produits à partir des corpus d'apprentissage extraits de YouTube. Ceci a conduit à l'obtention de trois lexiques comportant respectivement 6,7k, 6,7k et 3,6k entrées pour l'algérien, le marocain et le tunisien. Pour valider l'approche, nous avons testé la qualité du lexique algérien par rapport à un lexique de référence construit semi-automatiquement en utilisant les mesures de rappel et de précision.

À notre connaissance, ce genre de lexique n'existe pas, il pourrait avoir de nombreuses applications en traitement automatique des langues, comme par exemple dans l'identification des segments parallèles, dans l'amélioration des résultats de la traduction automatique des dialectes, etc.

La deuxième méthode a consisté à créer des lexiques de sentiments pour les trois dialectes. Pour ce faire, nous nous sommes basés sur 80 mots germes pour chaque dialecte que nous avons sélectionnés à partir des mots les plus fréquents extrait de nos corpus. Ensuite, pour déterminer l'orientation d'un mot donné, nous calculons sa polarité grâce à une fonction qui combine les polarités positives et négatives des mots germes. Ce qui a conduit à l'obtention de trois lexiques de sentiments comportant respectivement 11,2k, 23,4k et 10,8k entrées pour l'algérien, le marocain et le tunisien.

6

Contribution au projet AMIS : extraction automatique de corpus comparables et analyse de sentiments multilingues

Sommaire

6.1 Introduction 92
6.2 Corpus AMIS 92
6.3 Schéma d'analyse de sentiments 93
6.4 L'identification des vidéos comparables 93
6.4.1 Méthode fondée sur le dictionnaire bilingue 94
6.4.2 Méthode basée sur le *word embedding* 94
6.4.3 Expérimentations 94
6.5 Analyse de sentiments multilingues à granularité fine 95
6.5.1 La théorie *appraisal* 96
6.5.2 Travaux sur l'*appraisal* 97
6.5.3 La construction automatique d'un lexique d'*appraisal* 98
6.5.4 Modèle de prédiction de sentiments à granularité fine 100
6.5.5 Évaluation des vidéos du projet AMIS 102
6.6 Conclusion 104

Dans ce chapitre nous présentons nos contributions au projet AMIS (Access to Multilingual Information and Opinions). Ce projet vise à développer un système d'aide à la compréhension de l'information multilingue. Pour ce qui nous concerne, nous nous focaliserons sur l'établissement d'une revue d'analyse de sentiments à partir de deux vidéos dans deux langues différentes et portant sur le même sujet.

6.1 Introduction

L'émergence de nouvelles technologies et des outils de communication tels que l'Internet ont favorisé la diffusion et la croissance de l'information sur le Web. Par conséquent, des dizaines de milliers d'émissions et d'actualités sont disponibles en différentes langues. Malheureusement, l'usager n'a accès qu'à une quantité faible de ces émissions, cela est dû à la barrière de la langue. Par ailleurs, la masse impressionnante d'émissions et d'actualités fait que l'usager zappe rapidement d'une information à une autre sans que l'on se donne la peine d'aller au bout d'une émission. Il est donc utile de proposer aux utilisateurs un moyen permettant de résumer l'information. Pour ce faire, le projet AMIS (*Access to Multilingual Information and Opinions*) propose de développer un système d'aide à la compréhension de l'information multilingue sans aucune intervention humaine. Ce qu'on entend par compréhension dans ce contexte consiste à construire un résumé pertinent dans une langue cible à partir d'une vidéo exprimée dans une langue étrangère. De surcroît, ce projet permet aux utilisateurs, non seulement, de visionner l'information dans sa propre langue, mais aussi de la comparer à une autre vidéo portant sur le même sujet et diffusée dans une langue étrangère. La réalisation de ce projet exige l'utilisation de plusieurs composantes nécessitant une interaction judicieuse entre les différents modules [Smaïli et al., 2018]. Les architectures développées dans le cadre de ce projet font appel aux composantes suivantes : analyse de la vidéo, système de reconnaissance automatique de la parole, segmenteur de transcription, traduction automatique, résumé de textes et analyse de sentiments.

Dans ce qui suit, nous présenterons seulement ce qui est nécessaire à la réalisation de la tâche dont on était responsable dans ce projet à savoir la comparaison de sentiments de deux vidéos diffusées dans deux langues différentes.

6.2 Corpus AMIS

Afin d'atteindre notre objectif nous avons besoin de vidéos portant sur des sujets similaires et diffusées dans des langues différentes. Dans le projet AMIS, une centaine d'heures de vidéos a été collectée pour chacune des langues suivantes : arabe, français et anglais. Afin de récupérer ces vidéos, les auteurs dans [Kozbiał and Leszczuk, 2019] se sont basés sur une liste de *hashtags* portant sur des sujets polémiques : *#syria, #animalright, #deathpenalty,* etc. Ces sujets sont intéressants pour établir une revue de sentiments ou d'opinions contradictoires. Dans le tableau 6.1 nous donnons pour chaque langue le nombre de vidéos collectées.

Langue	**Nombre des vidéos**
Anglais	1874
Arabe	1503
Français	2046

TABLE 6.1 – Le nombre de vidéos par langue.

6.3 Schéma d'analyse de sentiments

Dans la figure 6.1, nous présentons un aperçu global du modèle permettant d'analyser les sentiments sous-jacents de deux vidéos portant sur un même sujet et exprimées dans deux langues différentes. Nous signalons que la comparaison de sentiments ne se fait pas comme habituellement sur deux textes puisque en entrée nous disposons de deux vidéos. Il est donc nécessaire de transformer le corpus de vidéos collectées en textes et de les aligner. En conclusion, deux étapes sont nécessaires :

1- L'alignement des vidéos : il est question d'identifier dans le corpus des vidéos, celles portant sur le même sujet. Autrement dit, nous souhaitons associer à chaque vidéo d'une langue A une vidéo d'une langue B. Pour ce faire, nous nous basons sur les méthodes que nous avons proposées pour créer des corpus comparables (voir le chapitre 4). Ces méthodes fondées sur les dictionnaires ou le *word embedding* agissent sur des textes. C'est pourquoi, nous utilisons un système de reconnaissance automatique de la parole par langue pour transformer le signal de la vidéo en texte.

2- Analyse d'opinions à granularité fine : une fois que les vidéos sont alignées, cette deuxième étape consiste à comparer les opinions sous-jacentes en terme de polarité et en terme d'opinion à granularité fine fondée sur la théorie de l'*appraisal*. Cette théorie sera détaillée dans les sections suivantes. Le résultat de cette étape produit une évaluation quantitative et qualitative pour chaque vidéo source et pour chaque vidéo cible.

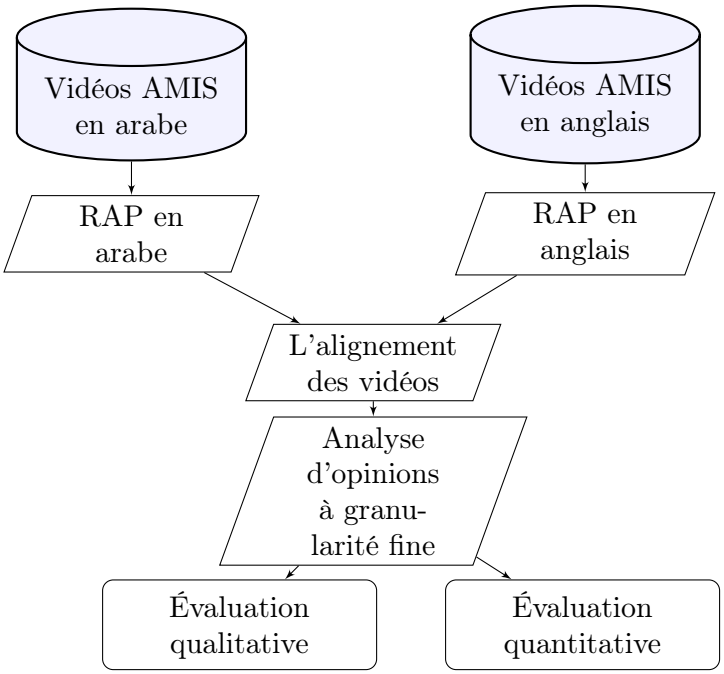


FIGURE 6.1 – Une vue globale du schéma d'analyse de sentiments.

6.4 L'identification des vidéos comparables

L'entrée du processus d'alignement est constituée de deux ensembles V_a et V_e. V_a est composé des transcriptions obtenues par le système de reconnaissance automatique de la parole

pour l'arabe. V_e est composé des transcriptions obtenues par le système de reconnaissance automatique de la parole pour l'anglais. Nous cherchons la vidéo v_a^r dans l'ensemble des vidéos V_a qui peut être appariée à la vidéo v_e^n dans l'ensemble des vidéos V_e. Pour ce faire, nous utilisons les deux méthodes de comparabilité que nous avons utilisées dans le chapitre 4 : la méthode fondée sur l'utilisation de dictionnaire [Li and Gaussier, 2010] et la méthode basée le principe du *word embedding*.

6.4.1 Méthode fondée sur le dictionnaire bilingue

Cette méthode consiste à estimer le degré de comparabilité entre deux documents bilingues en fonction de la proportion de mots de la première vidéo possédant au moins une traduction dans la deuxième vidéo et vice-versa. Cette méthode nécessite l'utilisation d'un dictionnaire bilingue, pour ce faire, nous en avons utilisé deux : *Open Multilingual WordNet* (OMWN) [26] comportant 17k entrées et une table de traduction [Menacer et al., 2017b] composée de 297k paires d'entrées (arabe et anglais). De la même manière que pour le traitement des tweets arabes (voir chapitre 4) nous avons lemmatisé les textes en arabe en utilisant l'analyseur morphologique BAMA (*Buckwalter Arabic Morphological Analyzer*). Par ailleurs, même si l'anglais n'a pas les mêmes contraintes morphologiques que l'arabe, nous avons également utilisé un analyseur morphologique (l'outil TreeTagger) pour réduire les formes fléchies des mots.

6.4.2 Méthode basée sur le *word embedding*

La méthode que nous utilisons ici, est fondée sur la même hypothèse retenue pour aligner des données vernaculaires dans la section 4.3.3. L'idée de base est d'utiliser la représentation distribuée des mots afin de trouver les mots sémantiquement proches les uns des autres dans deux documents où chacun d'eux est écrit dans une langue différente. Nous prenons comme hypothèse que les mots sémantiquement proches provenant de deux langues permettent de renforcer le degré de comparabilité entre deux documents. Pour ce faire, nous avons appris une représentation bilingue distribuée des mots en utilisant le modèle CBOW (*Continuous Bag of Words model*) sur le corpus parallèle de MultiUN [Eisele and Chen, 2010b]. Ce dernier est composé de 9 millions paires de phrases anglais-arabe.

6.4.3 Expérimentations

Les méthodes présentées précédemment sont évaluées sur un corpus de test composé de 123 paires de vidéos comparables extraites du site Web d'Euronews [Jouvet et al., 2017]. Toutes les vidéos ont été transcrites en utilisant les systèmes de reconnaissance automatique de la parole développés dans le projet AMIS pour les deux langues : arabe [Menacer et al., 2017c] et anglais [Jouvet et al., 2017].

Comme mentionné dans le chapitre 4, les résultats de la comparabilité sont souvent estimés par le rappel. Ce dernier est calculé à différents rangs (top1, top5, top10). Dans le tableau 6.2 nous donnons les résultats obtenus en terme de rappel pour les deux méthodes testées : celle basée sur les dictionnaires avec ces deux variantes (d'une part

26. http://compling.hss.ntu.edu.sg/omw/

DicMA fondée sur OMWN et l'utilisation d'un analyseur morphologique et d'autre part une table de traduction *DicTT*) et celle basée sur la représentation distribuée des mots (*CBOW*).

Rappel	**Top1**	**Top5**	**Top10**
DicMA	43,5	65,3	76,6
DicTT	**70**	**90**	**92**
CBOW	39	62	75

TABLE 6.2 – La performance de différentes méthodes de comparabilité en termes Top1, Top5, et Top10 sur un corpus de référence d'Euronews.

Ce tableau montre que le meilleur résultat est obtenu en utilisant la méthode basée sur le dictionnaire et plus précisément celle utilisant la table de traduction *DicTT*. Au rang 5 (top5), le rappel est de 90%. Pour la même méthode en utilisant la ressource externe OMWN, les résultats sont faibles au rang 1. Ceci est dû au nombre limité d'entrées de ce dictionnaire qui ne permet pas de couvrir tous les domaines des vidéos du corpus de référence extrait d'Euronews. Pour le corpus de référence utilisé, cette méthode donne un rappel égal à 70% de top1, puis il augmente jusqu'à 92% de Top10. Ce résultat montre que cette méthode dans la majorité de cas permet de récupérer la bonne vidéo cible dans les 10 meilleures candidates. Ce tableau montre également que la mesure basée sur le *word embedding CBOW* donne presque le même résultat que la méthode *DicMA* qui se base sur le dictionnaire OMWN. Ce résultat est très intéressant, car sans ressource externe (un dictionnaire bilingue et un analyseur morphologique), cette méthode donne presque la même performance que la méthode *DicMA*. Par conséquent, cette méthode pourrait être intéressante pour les langues peu dotées en ressources bilingues.

Nous avons utilisé la méthode qui donne le meilleur Rappel pour aligner les vidéos d'AMIS. Par conséquent, nous avons récupéré toutes les paires de vidéos comparables de la base de données AMIS, ce qui a conduit à 360 paires de vidéos comparables arabe-anglais. Nous rappelons que le nombre total de vidéos en arabe est de 1542, elles concernent plusieurs sujets qui n'ont pas nécessairement la même correspondance en anglais. De plus, nous avons sélectionné uniquement les paires de vidéos pour lesquelles les scores de comparabilité étaient élevés.

6.5 Analyse de sentiments multilingues à granularité fine

Comme mentionné dans l'état de l'art, d'une manière générale, les méthodes consacrées à l'analyse de sentiments visent à classer des textes selon leur polarité (négative ou positive). Dans d'autres études, des critères supplémentaires d'analyse sont utilisés pour effectuer une étude plus fine en utilisant des émotions telles que : la colère, le dégoût, la peur, la joie, la tristesse et la surprise [Strapparava and Mihalcea, 2007]. D'autres études plus poussées s'inspirent de la théorie linguistique de l'*appraisal* [Whitelaw et al., 2005], [Korenek and Šimko, 2014] et [Alamsyah et al., 2015], une théorie qui permet une analyse plus fine que nous détaillons ci-dessous. En effet, nous avons adopté cette approche dans nos travaux.

6.5.1 La théorie *appraisal*

La théorie *appraisal* a été développée par White et Martin [Martin and White, 2005] dans le cadre du principe de la linguistique fonctionnelle systémique [Halliday, 1994]. L'idée est d'effectuer une analyse de sentiments plus fine que celle effectuée par l'approche fondée sur la polarité en utilisant des attributs additionnels d'opinions comme : l'attitude, la graduation et l'engagement. Cette théorie est représentée par un graphe qui regroupe quelques catégories de sentiments exprimés par un être humain. Le schéma de la figure 6.2 illustre ces catégories.

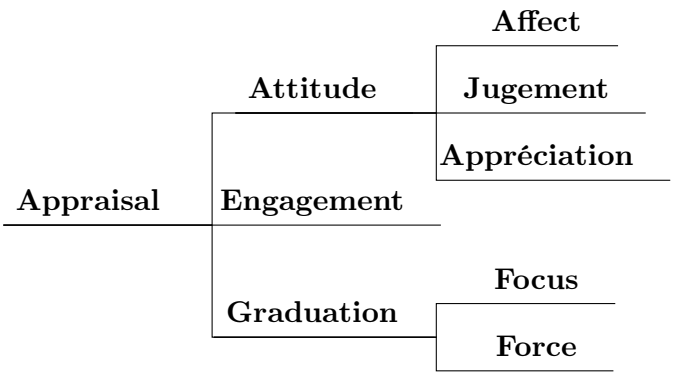


FIGURE 6.2 – Taxonomie de l'*appraisal*.

La catégorie attitude

L'attitude est selon Martin et White une catégorie de sentiments qui exprime l'état d'une personne au moment où elle écrit son texte. Cette catégorie est subdivisée en trois sous-catégories d'émotions.

- *L'affect :* concerne l'état émotionnel d'une personne, tel que : *la tristesse, la surprise, la colère,* etc.
- *Le jugement :* est relatif au comportement d'une personne par exemple : *courageux, populaire, fameux, brave,* etc.
- *L'appréciation :* concerne l'opinion d'une personne sur les qualités intrinsèques d'un objet, par exemple : *beau, novateur, étonnant,* etc.

La catégorie engagement

L'engagement est une catégorie qui détermine la position du texte par rapport a un événement. Il reflète la probabilité ou la possibilité d'une action comme dans les mots : peut-être, il semblerait que, etc.

La catégorie graduation

Cette catégorie accentue l'attitude et l'engagement. Selon Martin et White on dénombre deux sous-catégories : force et focus. Généralement, la graduation est exprimée à l'aide de

modifieurs linguistiques.

La Force exprime l'intensité de l'émotion. Par exemple, le mot *très* accentue l'adjectif qui le suit et par conséquent il intensifie la polarité de cet adjectif.

Le Focus permet de rendre le sens d'une assertion plus précis ou moins précis. Par exemple, *c'est un vrai défi* met un peu plus l'accent sur la complexité du défi par rapport à l'assertion *c'est un défi*.

6.5.2 Travaux sur l'*appraisal*

Les auteurs de [Whitelaw and Garg, 2005] ont construit un lexique composé de 1329 termes étiquetés selon les sous-catégories de l'*appraisal* en partant des exemples extraits du livre de [Martin and White, 2005] et en se basant sur les synonymes de *WordNet*. Les auteurs ont utilisé ce lexique pour classer un corpus de commentaires de films composé de 1000 documents positifs et 1000 documents négatifs et ont comparé leur méthode, mais en remplaçant leur lexique par celui de *Bing* [Hu and Liu, 2004] et ils montrent que les performances sont meilleures avec le lexique fondé sur l'*appraisal*. D'autres approches similaires ont été proposées par [Argamon et al., 2007], [Gardin, 2009], et [Korenek and Šimko, 2014].

Dans [Zhang and Ferrari, 2010], les auteurs ont construit un lexique fondé sur la théorie de l'*appraisal*, à partir d'un dictionnaire chinois *HowNet*[27]. Pour ce faire, les auteurs ont adapté certaines de leurs catégories pour les mettre en correspondance avec les catégories de la théorie de l'*appraisal*. Ils ont ajouté aussi des entrées à leur lexique pour couvrir les catégories manquantes. L'analyse d'un texte est basé sur ce lexique et quelques règles syntaxique locales. Le résultat de l'analyse de chaque énoncé se présente sous la forme d'une fiche regroupant les informations suivantes : la catégorie, la polarité, la force/le focus, et la cible (voir la figure 6.3).

Texte : "J'ai vraiment aimé votre site"
Catégorie : émotion
Polarité : positive
Focus : avivé
Cible : le site

FIGURE 6.3 – Un exemple d'une fiche produite par la méthode de [Zhang and Ferrari, 2010] .

Cette théorie a été utilisée également pour construire des bases de connaissances et des ontologies comme dans les travaux [Balahur et al., 2012] et [Dragos et al., 2018].

Dans [Balahur et al., 2012], les auteurs ont proposé une nouvelle base de connaissances appelée *EmotiNet* composée de chaînes d'actions modélisant les réactions affectives trouvées dans le texte. La méthode utilisée est fondée sur la théorie de l'*appraissal* pour analyser ces textes. Pour construire cette base, les auteurs ont utilisé des situations affectives

27. http://www.keenage.com

recensées dans le corpus *ISEAR : International Survey on Emotion Antecedents and Reactions*. Pour enrichir les connaissances, les auteurs ont utilisé d'autres ressources telles que VerbOcean [Chklovski and Pantel, 2004], ConceptNet [Speer et al., 2016] et SentiWordNet [Baccianella et al., 2010]. L'évaluation de l'approche a montré l'intérêt d'utiliser *EmotiNet* pour analyser les émotions correspondant aux chaînes d'actions.

Dans [Dragos et al., 2018], les auteurs ont construit une ontologie d'émotions fondée sur la taxonomie de l'*Appraissal* et en y ajoutant d'autres concepts supplémentaires. Certaines catégories de l'*Appraissal* ont été décrites plus finement en introduisant plusieurs sous-catégories. Le résultat est composé de 46 concepts représentés par une hiérarchie à 6 niveaux. Une fois que les concepts et les relations ont été modélisés par l'ontologie, des instances linguistiques ont été ajoutées à l'aide de WordNet. Cette ressource a été utilisée ensuite pour analyser une collection de tweets récupérés à l'aide de mots-clés comme : haine, racisme, extrême droite.

6.5.3 La construction automatique d'un lexique d'*appraisal*

Dans la suite, pour des raisons de commodité, nous n'utiliserons que les catégories Attitude et Graduation. Pour pouvoir effectuer une analyse basée sur la théorie de l'*appraisal*, nous avons besoin de construire un lexique dans lequel chaque mot est associé à sa catégorie *appraisal*. À notre connaissance ce type de lexique n'existe pas. Pour en construire un, nous avons utilisé un lexique de polarité proposé par Minqing Hu et Bing Liu [Hu and Liu, 2004] comme un lexique de départ que nous enrichirons avec des catégories *appraisal*. Ce lexique est composé de 4913 mot négatifs et 2718 mots positifs. Nous nommerons ce lexique *Bing*. À côté de ce lexique, nous avons construit à la main une liste comportant 363 mots avec leur catégorie Attitude en se basant sur les exemples existant dans le livre de Martin et White [Martin and White, 2005]. Nous nommerons cette liste *MW363*. Cette liste désormais contient pour chaque mot sa catégorie *appraisal* et sa polarité. Nous donnons quelques exemples de la liste dans le tableau 6.3.

Entrée	**Sous-categorie de l'Attitude**	**Polarité**
Lucky	Judgement	Positive
Obscure	Judgement	Négative
Confident	Affect	Positive
Love	Affect	Positive
Helpful	Appréciation	Positive

TABLE 6.3 – Quelques exemples de la liste *MW363*.

Ensuite, nous avons décidé de construire un lexique, comprenant des étiquettes *appraisal*, plus grand que *MW363*. C'est pourquoi, nous avons associé, d'une manière automatique, à chaque entrée du lexique *Bing* la sous-catégorie correspondante d'Attitude en utilisant la méthode *Word2vec* et la liste *MW363*. Pour ce faire, chaque mot de *MW363* et de *Bing* est représenté par un vecteur *Word2vec* produit par Google [28]. Pour chaque mot x du lexique *Bing*, nous déterminons les n mots les plus proches dans *MW363*. Chaque mot de cette liste de n mots les plus proches est associé à une des sous-catégories d'Attitude. Nous

28. https://code.google.com/archive/p/word2vec/

affectons ensuite à x la sous-catégorie dominante d'Attitude. Comme les mots du lexique *Bing* possèdent une polarité, nous leur affectons une sous-catégorie d'Attitude ayant la même polarité. Par exemple, si un mot qui se voit attribuer la sous-catégorie Affect, celle-ci pourra être soit positive, soit négative. Dans la figure 6.4, nous donnons une représentation graphique des mots de *MW*363 proches avec l'entrée *criminal* du lexique *Bing*. Nous remarquons que dans cet exemple, le mot *criminal* est proche de 11 mots de type *jugement négatif* (Jud_N), de 6 mots de type *jugement positif* (Jud_P), de un mot de type *appréciation négative* et un mot de type *appréciation positive* (App_P, App_N).

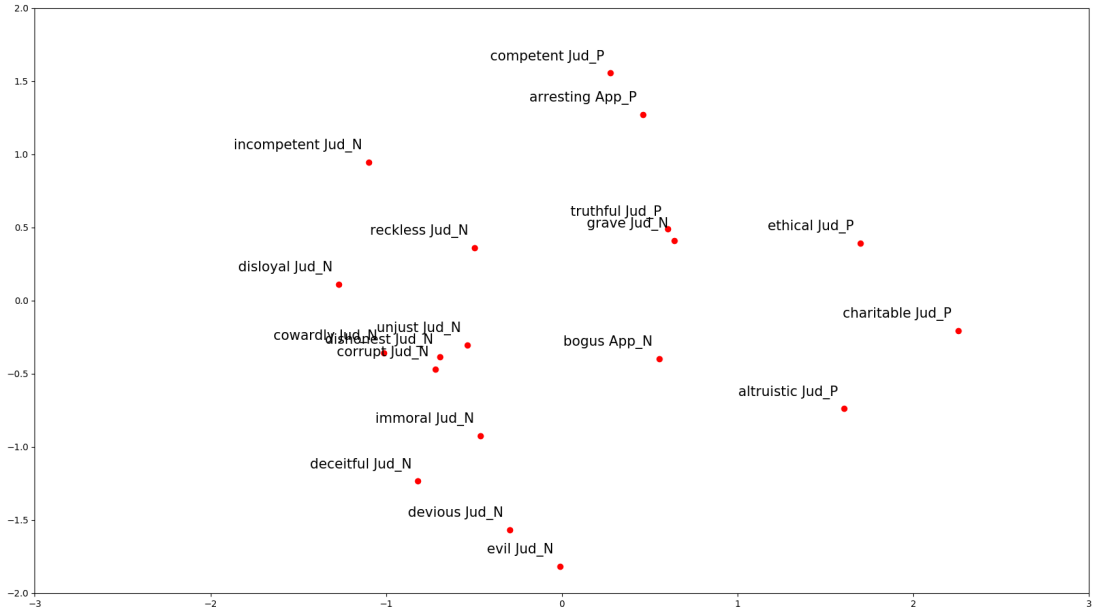


FIGURE 6.4 – Les mots proches avec le mot *criminal* obtenus en utilisant les vecteurs *Word2vec* de *Google*.

On affectera à chaque mot de *Bing* un score (S_{app}) *appraisal* positif ou négatif calculé selon la formule 6.1.

$$S_{App}(X) = \frac{1}{d_n} \sum_{i=1}^{d_n} cos(X, W_i) * P_{W_i} \tag{6.1}$$

où :

- d_n : le nombre de mots de la sous-catégorie dominante dans la liste des n mots proches de x.
- W_i : un mot appartenant à la liste des mots de la sous-catégorie dominante.
- x : un mot du lexique de *Bing*.
- $P_{W_i} = \begin{cases} +1 & \text{Si } W_i \text{ est positif} \\ -1 & \text{sinon.} \end{cases}$

Un score positif (respectivement négatif) indique à quel point le mot en question est positif (respectivement négatif) par rapport à une des sous-catégories d'Attitude. Le résultat

de cette méthode a permis d'enrichir le lexique initial *Bing* par des opinions plus fines telles que : Affect, Appréciation et Jugement. Nous nommons *BingApp* ce nouveau lexique.

Nous rappelons que notre objectif est d'analyser deux vidéos, une en anglais et l'autre en arabe, en termes d'opinions. Afin de travailler avec les mêmes ressources en arabe et en anglais, nous avons traduit BingApp en arabe et nous avons conservé pour chaque mot arabe la même sous-catégorie et le même score que celui du mot anglais. Dans le tableau 6.4, nous donnons quelques exemples du lexique final.

Entrée	**Traduction arabe**	**Catégorie/Sous-catégorie** *appraisal*	S_{App}	**Polarité**
Criminal	مجرم	Attitude/Judgment	-0,45	N
Attentive	منتبه	Attitude/Judgment	0,41	P
Worried	قلق	Attitude/Affect	-0,45	N
Satisfied	راض	Attitude/Affect	0,24	P
Harmonious	متناغم	Attitude/Appreciation	0,63	P

TABLE 6.4 – Quelques exemple du lexique BingApp.

6.5.4 Modèle de prédiction de sentiments à granularité fine

Afin d'analyser finement l'opinion des vidéos ou plus exactement des transcriptions correspondant à ces vidéos, il est nécessaire de se pencher sur au moins deux phénomènes linguistiques. Pour la clarté de nos propos, examinons l'exemple suivant : *Ce gâteau n'est pas très bon.* Cette phrase contient une marque de négation qui peut inverser complètement l'opinion de cette phrase. Dans ce même exemple, l'adverbe *très* est utilisé pour renforcer l'adjectif *bon.* En d'autres termes, il modifie son intensité en ajoutant de la force à cet adjectif. Le gâteau n'est pas juste mauvais, il est très mauvais. Par conséquent, ce phénomène doit être pris en compte surtout que la *Force* est une sous-catégorie de la catégorie *Graduation* de l'approche *appraisal.*

Traitement de la négation pour l'analyse d'opinions

Nous avons ajouté une nouvelle catégorie à la taxonomie de l'*appraisal* que nous avons appelée *Inversion*, avec la sous-catégorie *Négation.* Ensuite, nous avons ajouté au lexique *BingApp* les mots de négation : *Not, No, Neither, Nor, etc.* et nous leur avons affecté la catégorie *Inversion.* Pendant le traitement de l'analyse de l'opinion, si la catégorie *Inversion* est repérée dans le texte, alors la polarité du mot qui suit la négation est inversée.

Traitement de la *Force*

Afin de prendre en considération la *force* dans l'analyse d'opinions, nous avons ajouté manuellement dans le dictionnaire *BingApp* 40 mots qui jouent le rôle de modifieurs. Ces mots sont affectés à la sous-catégorie *force* de la catégorie *Graduation.* Ces modifieurs sont répartis en quatre classes. Chaque classe indique l'intensité des modifieurs et se voit attribuer un score proportionnel à son degré d'intensification du mot. Ces poids ont été fixés manuellement. Dans

le tableau 6.5, nous donnons quelques exemples de ces classes de *force* ainsi que les mots qui leur appartiennent.

Les classes de *force*	**Modifieurs**
Extrême	hardly, scarcely, barely, very, greatly, etc.
Élevé	large, less, distant, more, etc.
Modéré	somewhat, relatively, rather, reasonably, many, etc.
Faible	slightly, least, small, etc.

TABLE 6.5 – Les quatre classes de modifieurs de l'intensité.

Dans la figure 6.5, nous illustrons la nouvelle taxonomie de la théorie de l'*appraisal*.

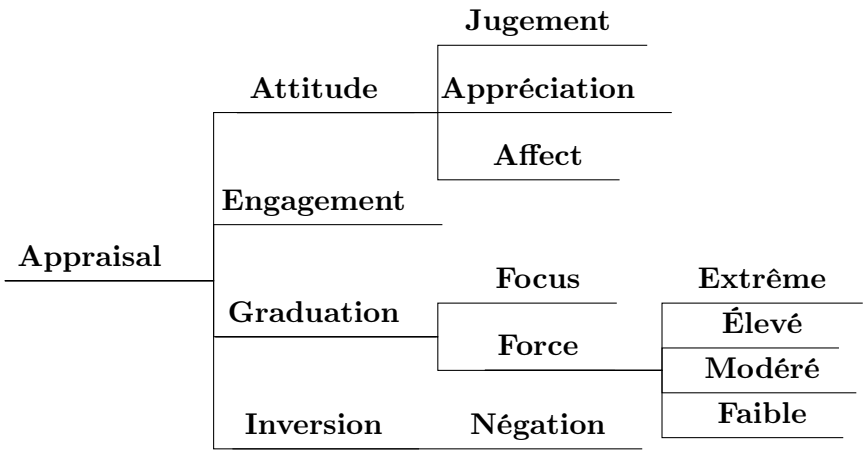


FIGURE 6.5 – Nouvelle taxonomie de l'*appraisal*.

Expérimentation et évaluation

Pour évaluer la qualité du lexique que nous avons créé (*BingApp*), nous avons décidé de l'évaluer sur un corpus public de sentiments concernant les avis des Internautes sur des films [Pang and Lee, 2004]. Ce corpus est composé de 1000 avis positifs et de 1000 avis négatifs. Dans la tableau 6.6 nous donnons les résultats en termes de rappel et précision en utilisant le lexique de départ *Bing* et en utilisant le lexique que nous avons créé *BingApp*.

Méthode	**Rappel**	**Précision**
Bing	69,3%	69,4%
BingApp	**70**%	**71**%

TABLE 6.6 – Les résultats de l'analyse d'opinions en utilisant un lexique de polarité (*Bing*) et un lexique d'*appraisal* (*BingApp*).

Ces résultats montrent une légère amélioration lorsqu'on utilise le dictionnaire incluant des opinions plus fines que celles exprimées par la seule polarité. Nous avons donc décidé dans la suite d'effectuer une évaluation à granularité fine de nos vidéos en utilisant le lexique d'opinions *BingApp* fondé sur la théorie *appraisal*.

6.5.5 Évaluation des vidéos du projet AMIS

Notre objectif est de proposer une évaluation de deux vidéos portant sur le même sujet. Pour ce faire, nous utiliserons le lexique *BingApp* et nous produirons une évaluation quantitative et une évaluation qualitative.

L'évaluation quantitative attribuera une valeur numérique à chaque vidéo selon la formule 6.2.

$$S = \sum_{i=1}^{N} \alpha(w_{i-k}^{i-1}) * S_{App}(w_i) \tag{6.2}$$

Avec N est la taille de la vidéo en nombre de mots. α est le poids qui dépend des sous-catégories *Inversion* ou *Force* des k mots précédents le mot w_i (k a été fixé à 2). Il s'agit de la taille du cache dans lequel la *Force* ou la *Négation* est recherchée.

La deuxième évaluation est de type qualitative dans laquelle nous résumons les opinions exprimées dans la vidéo. Notre idée est de faciliter l'interprétation des opinions sous-jacentes et de ne pas se contenter d'attribuer un score global qui peut être positif ou négatif. Autrement dit, on fournit une sorte de revue d'opinions pour chacune des vidéos à comparer. Comme nous comparons deux vidéos portant sur le même sujet en deux langues différentes, l'usager aura à sa disposition deux revues d'opinions écrites dans la même langue. Ce qui lui permettra de se faire sa propre opinion sur le sujet. Chaque évaluation sera présentée dans un canevas respectant la forme de la figure 6.6.

The sentiment of the video is positive with a score $\mathbf{[}+X_p\mathbf{]}$ and negative with a score $\mathbf{[}-X_n\mathbf{]}$. X_{aff}% of the video concerns emotional reactions. X_{Jug}% of the video concerns the human behaviour according to social norms and $1-(X_{aff}+X_{Jug})$% of the video is about the appreciation of no human being entities.

FIGURE 6.6 – Le canevas utilisé pour produire une évaluation qualitative.

Cette forme correspond à la revue présentée à l'utilisateur et indique la proportion de polarité négative ainsi que la proportion de polarité positive de la vidéo. Elle indique également le pourcentage de chaque sous-catégorie de la catégorie désignant l'opinion *Attitude*. On donnera ainsi le pourcentage X_{aff} d'*Attitude* de type *Affect*, le pourcentage X_{Jug} de type *Jugement* et le reste de type *Appréciation* qui correspond à $1-(X_{aff}+X_{Jug})$.

Dans la figure 6.7 nous donnons un exemple d'une paire de vidéos comparables avec leurs évaluations quantitative et qualitative.

إيران "تخلّ" بالاتفاق النووي.. وتطالب ترمب باحترامه

إيران تجاوزت قليلا الحد المسموح به لا مخزون الماء الثقيل للمرة الثانية هكذا رأت الوكالة الدولية للطاقة الذرية في تقريرها و حد ما وصفه خبراء بأنه انتهاك للاتفاق النووي بيد أن الخارجية الأمريكية أكدت أن طهران لم تخفي ذلك و تتخذ خطوات لنقل الماء الثقيل الذي انتجته مما يزيد عن الكمية المحددة و كانت ايران قد تجاوزت قبل ذلك سقف ال منة و ثلاثين طنا متري من الماء الثقيل بنسبة واحد من عشرة في المئة الذي يستخدم لتبريد مفاعلات نووية مثل المفاعل الإيراني الذي لم يكتمل العمل فيه في اراك منذ يناير الماضي و تخشى دوائر عالمية عدة لمواصلة جهود إيجاد بيئة عملية تمكنها من تصنيع اسلحة نووية و كل ذلك في وقت ينتظر فيه موقف الرئيس الأمريكي المنتخب جرام من الاتفاق النووي بعد إعلانه أثناء الحملة الانتخابية عزمه إعادة النظر فيه و كان الرئيس الإيراني و مسؤولون أوروبيون نوهوا بأن الاتفاق النووي وثيقة دولية ملزمة و متصلة بالسلام الدولي و لا يمكن لتر بالتصرف و فيه لكن خبراء قانونيين دوليا رأوا أن الإدارة الأمريكية الجديدة في واشنطن لديها الحق في النظر في تعهدات سابقة إذا كان لها تداعيات مستقبلية على أمن الولايات المتحدة و مصالحها كتيب تركستان

Score positif = 1,56, **Score négatif** = -2,15
The sentiment of the video is positive with a score **1.56** and negative with a score **-2.15**. **50**% of the video concerns the human behaviour according to social norms and **50**% of the video is about the appreciation of no human being entities.

What will Trump's presidency mean for the Iran nuclear deal?

for what s will donald trump s presidency main for the iran nuclear deal he s castigated the packs tehran agreed with the u s and several global powers on a visit to prague iran s foreign minister said he expected the new administration to accept the agreement it on has been implementing it s part of the bargain unfortunately a the united states even during the current administration a has lot a fully a implemented it s part of the bargain but we believe that jaycee puree is a reasonable agreement good for everybody that s our preference doesn t mean that we do not have other options the deal remove sanctions on tehran in return for curbs on its suspected drive to develop nuclear weapons it was opposed by republicans and during the campaign trump called a disastrous too soft ironic the agreement already it is a few days after all worst heels i ve ever seen negotiated at any level by august the presidential candidate a temperate his rhetoric saying he wouldn t rip up the deal that would forcefully police it but with iran s leadership divided some observers predict its demise

Score positif = 1,42, **Score négatif** = -3,27
The sentiment of the video is positive with a score **1.42** and negative with a score **-3.27**. **7.14**% of the video concerns emotional reactions. **78.57**% of the video concerns the human behaviour according to social norms and **14.29**% of the video is about the appreciation of no human being entities.

FIGURE 6.7 – Un exemple d'évaluation de sentiments pour deux vidéos comparables dans deux langues différentes.

6.6 Conclusion

Nous avons dans ce chapitre présenté nos contributions au projet AMIS. Nous avons été confrontés à deux défis majeurs. Le premier a consisté à rendre les transcriptions de vidéos obtenues par deux systèmes de reconnaissance de la parole comparables. Pour ce faire, nous avons testé et comparé deux méthodes. La méthode fondée sur la table de traduction a donné le meilleur Rappel (70% au rang 1), c'est pourquoi nous avons retenu cette approche pour aligner l'ensemble des vidéos en anglais et en arabe du projet AMIS.

Le deuxième défi a consisté à comparer les paires de vidéos comparables en termes d'opinion. Pour ce faire, nous avons proposé une méthode fondée sur la théorie linguistique de l'*appraisal*. Pour résoudre ce problème, à partir d'un lexique de polarité de Minqing Hu et Bing Liu, nous avons développé une méthode pour l'enrichir avec des catégories et sous-catégories de l'*appraisal*. Pour y arriver, nous avons utilisé la méthode CBOW pour affecter à chaque entrée du lexique Bing la catégorie *appraisal* correspondante en se basant sur un lexique de catégories de 363 entrées comportant chacune une étiquette *appraisal*. Cela a conduit à la construction d'un lexique d'*appraisal* comportant 7,6k entrées.

Pour l'évaluation en termes d'opinions de deux vidéos, nous avons proposé une évaluation quantitative et une évaluation qualitative. Cette dernière se présente sous forme d'un canevas dans lequel on restitue l'ensemble des opinions à granularité fine que nous avons identifiées dans chacune des vidéos.

Conclusion et perspectives

Le traitement automatique des dialectes arabes présente plusieurs verrous scientifiques et en plus il souffre d'un manque crucial d'outils linguistiques et de ressources langagières. Dans cette thèse, nous avons tenté, à notre échelle, de remédier à ces insuffisances. Nous avons proposé des méthodes permettant de construire automatiquement des ressources pour les dialectes arabes. Les approches que nous avons proposées peuvent être appliquées à n'importe quel dialecte arabe à condition de collecter suffisamment de données du dialecte concerné à partir des réseaux sociaux. Dans cette thèse, nous nous sommes focalisés sur les dialectes : algérien, marocain et tunisien.

Il n'existe que très peu de corpus standards des dialectes du Maghreb, en revanche les réseaux sociaux regorgent de données dialectales. Par conséquent, nous avons exploité les forums de discussions afin de développer des ressources nécessaires au traitement des dialectes. Nous avons ainsi collecté au moins 16M de mots pour chacun des dialectes étudiés dans cette thèse. Ces données ont été nettoyées avant leur utilisation. Une étude analytique de ces corpus a été très bénéfique en termes de compréhension des différences entre les dialectes, elle a ainsi permis de mieux appréhender les difficultés inhérentes au traitement des dialectes. Nous avons constaté que les utilisateurs préfèrent écrire en dialecte plutôt qu'en arabe standard ou en français. Un des constats surprenants est que les utilisateurs choisissent plutôt le script latin que le script arabe pour écrire l'arabe dialectal. Par exemple, 46% du corpus tunisien est composé de commentaires en arabe dialectal écrit en script latin.

Il est connu que la langue française est très utilisée au Maghreb, l'analyse de nos corpus nous a confirmé cela, en effet, le dialecte algérien est celui qui est le plus impacté par cette langue, puisque 6,4% du corpus est écrit dans cette langue. Concernant l'utilisation de l'arabe standard dans nos corpus, nous avons constaté que le dialecte marocain est celui qui en comporte le plus (27,3%).

Dans cette thèse, nous avons traité une des caractéristiques importantes des dialectes arabes, il s'agit du code-swhitching. En effet, ce phénomène est très répandu dans ces dialectes. Cela nous a conduit à nous interroger sur la possibilité de mesurer la qualité des corpus collectés. En effet, dans certains travaux de recherche, on pourrait être intéressé par la collecte de données dialectales ne comprenant aucun mot en langue étrangère. C'est pourquoi, nous avons proposé une nouvelle mesure *CESAR* (*CodE-Switching According to a Reference language*) permettant d'estimer le degré de code-switching d'un corpus par rapport à une langue de référence. L'hypothèse de base est de permettre d'évaluer le bruit existant dans les corpus collectés à partir des réseaux sociaux par rapport à une langue de référence. CESAR est une métrique bornée qui associe 0 à un document dans lequel le texte

est entièrement dans la langue de référence (le dialecte dans notre cas). Une valeur maximale égale à 1 est attribuée à tout document ne contenant aucun mot de la langue de référence. Lorsque le texte comporte plusieurs langues, CESAR lui attribue une valeur comprise entre 0 et 1 proportionnelle au degré de code-switching dans le texte. Nous avons utilisé cette mesure pour quantifier le code-swtching dans les différents corpus. Les résultats ont montré que le corpus marocain comporte moins de code-switching que les deux autres selon la mesure CESAR qui lui a attribué une valeur de 0,23.

Ces corpus nous ont servi à proposer des méthodes statistiques et neuronales permettant de développer de nouvelles ressources nécessaires au traitement des dialectes étudiés dans cette thèse.

— **La construction de corpus comparables multilingues** : la question de l'alignement des données vernaculaires est un problème important. Pour extraire des documents comparables à partir de nos corpus, nous avons testé plusieurs méthodes. L'approche classique fondée sur le dictionnaire et sur l'encodage phonétique n'a pas donné des résultats satisfaisants. C'est pourquoi, nous avons proposé une nouvelle méthode qui n'utilise pas de dictionnaire externe. Cette approche est fondée sur le principe du *word embedding*. Elle est intéressante parce qu'elle permet de mettre en correspondance deux documents bilingues en se basant sur les mots sémantiquement proches. L'alignement est réalisé au bout d'un certain nombre d'itérations d'utilisation du *word embedding* qui permet d'affiner à chaque itération la liste des mots proches servant à la mise en correspondance des documents. En effet, cette méthode a permis une amélioration de 22 points en terme de rappel par rapport à la méthode basée sur le dictionnaire et elle nous a permis de construire un corpus comparable pour l'algérien *CALYOU* comportant 10k commentaires comparables. Ce corpus est composé d'une partie source qui contient des commentaires écrits en script latin et une partie cible composée des commentaires en script arabe. Signalons que la partie source peut être composée de commentaires entièrement en français, en anglais ou d'un mélange des deux langues précédentes et de l'arabizi.

— **La construction de dictionnaires de variabilité lexicale** : comme nous l'avons vu dans cette thèse le dialecte n'a pas de norme d'écriture, un même mot peut être écrit de plusieurs manières différentes. Cette variabilité de formes introduit une ambiguïté lexicale dans les traitements automatiques de la langue que l'on gère généralement en faisant appel à un analyseur morphologique, un analyseur syntaxique, à l'utilisation d'un dictionnaire ou autres. Dans ce travail, nous avons proposé une méthode permettant l'identification automatique des différentes possibilités d'écriture d'un même mot. Le résultat de ce travail a conduit à la construction de trois lexiques comportant pour chaque entrée, les différentes formes d'écriture possibles de celle-ci. Ils sont composés respectivement de 6,7k, 6,7k et 3,6k entrées pour l'algérien, le marocain et le tunisien.

  À notre connaissance, ce genre de lexique n'existe pas, il pourrait avoir de nombreuses applications en traitement automatique des langues, comme par exemple l'identification des segments parallèles, l'amélioration des résultats de la traduction automatique des dialectes, etc.

— **La construction de lexiques de sentiments** : une autre ressource importante destinée à l'analyse de sentiments des trois dialectes a été proposée dans cette thèse. Il s'agit d'un lexique où chaque entrée est associée à un degré de polarité. La méthode

utilisée pour déterminer automatiquement l'orientation sémantique des entrées se base sur la proximité sémantique de ces dernières avec une liste de mots (mots germes) dont l'orientation sémantique est fixée à la main. Le résultat de ce travail a conduit à l'obtention de trois lexiques de sentiments pour l'algérien, le marocain et le tunisien comportant respectivement 11,2k, 23,4k et 10,8k entrées.

Ces trois ressources ont été mises à la disposition de la communauté (voir le site de l'équipe $SMarT$ [29]).

Même si nos travaux ont porté particulièrement sur trois dialectes du Maghreb, nous avons également développé des méthodes et des ressources pour l'arabe standard.

— **La construction d'un corpus comparable de tweets** : nous avons proposé une méthode permettant d'aligner des paires de tweets écrits en arabe standard et en anglais. L'alignement des tweets est fondé sur une méthode à base de dictionnaire à laquelle nous avons ajouté un certain nombre de traitements. En effet, étant donnée la brièveté des messages de Twitter, des traitements particuliers ont été nécessaires pour assurer un maximum de mise en correspondance entre les tweets en arabe et les tweets en anglais. Citons par exemple, le traitement des *hashtags* qui sont souvent composés de mots-clés pertinents concaténés. La segmentation de ces *hashtags* est donc nécessaire pour augmenter le nombre de mots susceptibles d'être appariés. Un autre exemple de traitement particulier concerne les noms propres qui sont fréquents dans les tweets. En arabe, il est difficile de repérer les noms propres dans un texte, il a fallu donc procéder à une translittération de certains mots en utilisant un encodage phonétique. Ce qui a permis d'augmenter les chances de l'algorithme d'associer des tweets similaires.
Le résultat de ce travail a conduit à la contruction d'un corpus comparable arabe-anglais comportant 11,5k tweets.

— **La construction d'un corpus comparable à partir de vidéos** : certains de nos travaux ont été effectués dans le cadre du projet AMIS où nous nous sommes intéressés plus particulièrement à l'établissement d'une revue d'analyse de sentiments à partir de deux vidéos dans deux langues différentes et portant sur le même sujet. Pour atteindre cet objectif, il a donc fallu d'abord rendre les vidéos comparables. Les vidéos ont été transcrites à l'aide des systèmes de reconnaissance de la parole développés dans le cadre du projet pour la langue arabe et anglaise. Les méthodes de comparabilité que nous avons développées traitent donc des résultats de transcription automatique des vidéos. Pour ce faire, nous avons testé et comparé deux méthodes, une fondée sur le dictionnaire et une autre sur le principe du *Word embedding*. La méthode fondée sur le dictionnaire en utilisant une table de traduction a donné de meilleurs résultats. Cette méthode a été retenue afin d'identifier les paires de vidéos comparables à partir de l'ensemble des vidéos collectées dans le cadre du projet AMIS. Le résultat de ce travail a conduit à l'alignement de 24% des vidéos en arabe avec 19% des vidéos en anglais de l'ensemble des vidéos disponibles.

— **L'analyse de sentiments fondée sur la théorie de l'***appraisal* : afin d'analyser en terme de sentiments les vidéos comparables que nous avons identifiées, nous avons proposé une méthode fondée sur la théorie linguistique de l'*appraisal* que nous avons étendue. Cette méthode consiste à analyser un texte selon des caractéristiques plus fines que celles utilisées habituellement.

29. Disponible dans https://smart.loria.fr/corpora/

— **Proposition d'un lexique de sentiments fondé sur l'***appraisal* : pour ce faire, nous avons enrichi le lexique de polarité de Minqing Hu et Bing Liu avec des catégories d'*appraisal* en les identifiant d'une manière automatique. Le lexique résultat de ce traitement est composé de 7684 entrées comportant chacune une catégorie *appraisal* positive ou négative.

— **Évaluation des vidéos comparables en terme de sentiments** : pour l'évaluation du contenu des paires de vidéos comparables en terme de sentiments, nous avons proposé une évaluation quantitative et qualitative. Cette dernière se présente sous forme d'un canevas dans lequel on restitue l'ensemble des opinions à granularité fine que nous avons identifiées dans chacune des vidéos.

Les ressources que nous avons développées nous ouvrent plusieurs perspectives. Les corpus comparables peuvent être désormais exploités afin d'extraire des segments parallèles pouvant servir à créer des dictionnaires bilingues et des corpus parallèles qui seront utiles pour la traduction automatique des dialectes. En effet, les commentaires alignés constituent une base d'information extrêmement riche dans laquelle un même sujet est traité dans plusieurs langues dont le dialecte. Par conséquent, il existe vraisemblablement des segments de part et d'autre qui sont des traductions les uns des autres. La question est donc de développer de nouvelles approches permettant de les identifier d'une manière précise afin de les exploiter dans des dictionnaires bilingues ou des corpus parallèles.

Étant donnée la proximité des trois dialectes étudiés dans cette thèse, il serait intéressant d'exploiter les recouvrements grâce aux corpus que nous avons collectés. En effet, ce noyau commun pourrait servir à un système de reconnaissance de la parole développé pour l'arabe standard afin d'améliorer ses performances dans la reconnaissance de l'arabe code-switché avec ces dialectes. Cette proposition ne permet évidemment pas de résoudre le problème d'adaptation de l'arabe standard aux trois dialectes, mais représente une solution moins coûteuse que d'adapter un système de reconnaissance de la parole à chaque dialecte.

Par ailleurs, les corpus des trois dialectes pourraient être fusionnés pour n'en former qu'un seul et utiliser le principe du *multi-task learning* pour apprendre un seul modèle permettant de tirer profit des caractéristiques de chaque dialecte. Autrement dit, ce qui est appris pour un dialecte peut aider à l'apprentissage du modèle supportant un autre dialecte. En effet, cette approche permettra de mieux généraliser l'apprentissage en utilisant une représentation partagée des données.

L'approche que nous avons développée pour la construction d'un dictionnaire arabe de sentiments fondé sur la théorie de l'*appraisal* peut être envisagée pour les dialectes. À partir des lexiques de sentiments que nous avons développés pour les trois dialectes, il faudrait affecter automatiquement à chaque entrée, la catégorie *appraisal* adéquate, soit à partir des lexiques arabes ou français comportant ce genre de catégories en déterminant, par exemple, les mots dialectaux proches des mots arabes ou français disposant de ces mêmes catégories.

A

Nous présentons ici les mots-clés utilisés pour collecter les commentaires appropriés à chaque dialecte.

ALG	Abdelmalek Sellal, Larbi dida, rym hakiki, عثمان عريوات, الطاهر ميسوم, خليدة تومي, naima salhi, Dzjoker chemsou, Zarouta youcef, Zanga crazy officiel, Dihya beauty, Ryma beautéAddict, Manelth, Oum walid, ناس سطح, Habib zeggat, قهوة القوسطو, سبيسفيك, Mourad oudia, l'inspecteur Tahar, hdidwan, عمارة حاج لخضر, bilahdoud, Sultan Achour, بوضو, ناس ملاح سيتي , swilah, جمعي فاميلي, Anes Tina 3okacha, USM Harrach,abdelkader chaou, Kamel Messaoudi, Mourad Djaafri, Dahman el harrachi, Hadj el hachemi, Guerouabi, Naima Dziria, عايلة كي ناس, حسان طاكسي, Carnaval fi dachra, Noujoum saf, nadia dziria, Nouri Koufi,elmazouni.
MAR	سكينة درابيل, ليلي حديوي, Hassan almaghribi, Nass el ghiwan, Eko Artiste, Dima drama, Amine filali TV, Karba zizo, Simo Sedraty, Simo daher, جميلة الهوني, مريم الزعيمي, نجاة عتابو, مطبخ خوخة, Abdeljabbar louzir, Lalla Fatema, Mohammed Nassib, Rimane beauty, Souhlifa, Said Naciri, Hassan Gonzalez, قناة زينة, Black moussiba, كبور و حلب, سامية اقريو, صفا حبيركو , نعيمة بوحمالة, Hassan el fad, Ma9tou3 men chajara, فتيحة اوتيلي, عويشة الدويبة, رمانة و برطال, بنت الفشوش, Min dar ldar xald omo, الخدامة النعاسة, Zman Kenza, كي كنتي كي وليتي, Wouja'e trabe, ابتسام لشكر, الشيخ سار.
TUN	Awled moufida, لوطفي بوشناق, Shems FM, Cuisine leila ben lazher, شباب تونس, Zomra, El zazouet, العزازة, حمة الهمامي, نوفال الورتاني, Les fellagas, عبد العزيز العروي, Habib Bourguiba, Salem monsieur, Hakek taaref, Taoufik ben brik, جغفور, راشد الغنوشي, Aveyro, قايد السبسي, كوجينتي, Leila ben khalifa, Alaa chebbi, عزيزة بولبيار, Fathia khairi, شكري الواعر, شمس الدين باشا, حسان الدوس, AkramMag, Tunis 2050, Captain 5obza, CapfmTunisie, Kayes Hurricane, M3kky sinacer, Walid tounssi, اسماعثماني, Nermine sfar, بن غريبة,

B

Corpus des tweets parallèles utilisé pour l'évaluation

Pour évaluer les méthodes d'alignement des tweets, nous avons utilisé le corpus parallèle des tweets alignés automatiquement par les auteurs de [Ling et al., 2013]. Dans le tableau B.1, nous donnons quelques exemples de paires de tweets considérées comme parallèles. Le corpus complet est disponible à l'adresse : http://www.cs.cmu.edu/~lingwang/microtopia/#twitter.

Tweet source	**Tweet Cible**
اعرف منين امال ماهر	a3raf menen amal maher
★∗○○○∗★∗★∗○○○★★♡ – ★	★∗○○○∗★∗★∗○○○★★
الدراما هدي جمّمّمّمال$ $___$ $ فيها المجدول الّي ب ٤٩ يوم	$, $
قطع الماس الماس!!!	diamond cut diamond!!.. diamante taglio
تم العثور على شخص فاهم خطاب مرسى جاري التحقق معه	C'est la vie
ستار اكاديمي ٩ في المسبح	فضاءيح ستار اكاديمي ٩
يا دين امى	yah my mother religion
بيت حلو بيت	home sweet home
مجاني لفترة محدودة	Raiding Company is now free for limited time

TABLE B.1 – Exemple de tweets parallèles extrait du corpus développé par [Ling et al., 2013].

Table de translittération

Dans le tableau B.2, nous donnons la table de translittération utilisée par certaines de nos méthodes.

Lettre	Nom	translittérations
ا	ALEF	2 a e i
ب	BEH	b p
ت	TEH	t
ث	THEH	th
ج	JEEM	j dj
ح	HAH	h
خ	KHAH	5 kh
د	DAL	d
ذ	THAL	d
ر	REH	r
ز	ZAIN	z
س	SEEN	s
ش	SHEEN	sh ch
ص	SAD	s
ض	DAD	d
ط	TAH	t 6
ط	TAH	d
ظ	ZAH	d
ع	AIN	’ 3 a e
غ	GHAIN	gh
ف	FEH	f v
ق	QAF	9 q k c
ك	KAF	k c
ل	LAM	l
م	MEEM	m
ن	NOON	n
ه	HEH	n
ة	TEH MARBOUTA	h a
و	WAW	w u ou
ي	YEH	y i a

Table B.2 – Table de translittération utilisée.

C

CALYOU

Nous présentons quelques exemples des commentaires comparables extraits du CALYOU.

Commentaire source	**Commentaire cible**
vive tahar misoum rabi yahafdak	تحيا طاهر ميسوم
Merci chemsou je t'adore	نحبك شمسو بزاف
recette réussite je l'ai essayé merci bcq	وصفتها ناجحة سيتها
thé song of thé musique	اسم موسقة لستعملتها
the best youtube video ever good luck Dz djoker	فيديو اكتر من رايع تشوفو 100 خطرة ما تكرهش
now we are free gladiator theme song	شكون يعرف اسم الاغنية عجبتني هاد موسيقة
10 mai 2017 mazalni nchouf la vidéo chkoun kima ana	لي مزال يعاود ف فيديو يكليكي جام علبالي رانا بزاف
bien dit wlh 3andek sah ya3rfou ghir yahdrou haja mata3jebhom	ولّله لنيشان عندك صح انا راني زعفان مانسوطيش

TABLE C.1 – Exemples de commentaires comparables de CALYOU.

D

Liste des publications

D.1 Revue internationale

1. **CESAR : A new metric to measure the level of Code-Switching in corpora Application to Maghrebian dialects**
   Karima, Abidi, Smaïli, Kamel, Natural Language Engineering, Cambridge University Press. to appear in 2019.

D.2 Conférences internationales

1. **Measuring the comparability of multilingual corpora extracted from Twitter and others**
   Karima, Abidi, Smaïli, Kamel, The Tenth International Conference on Natural Language Processing (HrTAL2016).
2. **How to match bilingual tweets ?**
   Karima, Abidi, Smaïli, Kamel, 66th NLP 2017 - Computer Science Conference Proceedings in Computer Science (ICIT 2017).
3. **CALYOU : A Comparable Spoken Algerian Corpus Harvested from YouTube**
   Karima, Abidi, Mohammed Amine Menacer, Smaïli, Kamel, 18th Annual Conference of the International Speech Communication Association (Interspeech), Stockholm Sweden August 20-24 2017.
4. **An empirical study of the Algerian dialect of Social network**
   Karima, Abidi et Kamel, Smaïli, International Conference on Natural Language, Signaland Speech Processing (ICNLSSP). Casablanca, Morocco, 2017.
5. **An Automatic Learning of an Algerian Dialect Lexicon by using Multilingual Word Embeddings**
   Karima, Abidi et Kamel, Smaïli 11th edition of the Language Resources and Evaluation Conference (LREC), 7-12 May 2018, Miyazaki (Japan).
6. **Automatic identification methods on a corpus of twenty five fine-grained Arabic dialects**
   Salima Harrat, Karima Meftouh, Karima Abidi, Kamel Smaïli, The 7_{th} International Conference on Arabic Language Processing (ICALP). 16-17 October 2019.

7. **Extractive Text-Based Summarization of Arabic videos : Issues, Approaches and Evaluations**
   M.A. Menacer, C.E. González-Gallard, K. Abidi, D. Fohr, D. Jouvet, D.Langlois, O. Mella, F. Sadat, J.M. et K.Smaïli. The 7_{th} International Conference on Arabic Language Processing (ICALP), 16-17 October 2019.
8. **A Fine-grained multilingual analysis based on the Appraisal theory : Application to Arabic and English videos**
   K. Abidi, D. Fohr, D. Jouvet, D. Langlois, O. Mella et K. Smaïli, The 7_{th} International Conference on Arabic Language Processing (ICALP),16-17 October 2019.
9. **The SMarT's classifier for Arabic fine-grained dialect identification**
   Karima Meftouh, Karima Abidi, Salima Harrat, Kamel Smaïli, Proceedings of the Fourth Arabic Natural Language Processing Workshop co-located with ACL, Association for Computational Linguistics, 2019.

Bibliographie

[Abdaoui et al., 2017] Abdaoui, A., Azé, J., Bringay, S., and Poncelet, P. (2017). FEEL : a French Expanded Emotion Lexicon. *Language Resources and Evaluation*, 51(3) :833–855.

[AbduI-Rauf and Schwenk, 2009] AbduI-Rauf, S. and Schwenk, H. (2009). On the use of comparable corpora to improve smt performance. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09.

[Afli et al., 2012] Afli, H., Barrault, L., and Schwenk, H. (2012). Traduction automatique à partir de corpus comparables : extraction de phrases parallèles à partir de données comparables multimodales. *Traitement automatique de langue Naturelle.*

[Aker et al., 2012] Aker, A., Kanoulas, E., and Gaizauskas, R. J. (2012). A light way to collect comparable corpora from the web. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 15–20.

[Alamsyah et al., 2015] Alamsyah, A., Rahmah, W., and Irawan, H. (2015). Sentiment analysis based on appraisal theory for marketing intelligence in indonesia ' s mobile phone market.

[Ameur et al., 2016a] Ameur, D., David, L., and Kamel, S. (2016a). Genetic-based decoder for statistical machine translation. In *Springer LNCS series, Lecture Notes in Computer Science.*

[Ameur et al., 2016b] Ameur, H., Jamoussi, S., and Ben Hamadou, A. (2016b). Exploiting emoticons to generate emotional dictionaries from facebook pages. In Czarnowski, I., Caballero, A. M., Howlett, R. J., and Jain, L. C., editors, *Intelligent Decision Technologies 2016.*

[Amiri et al., 2015] Amiri, F., Scerri, S., and Khodashahi, M. (2015). Lexicon-based sentiment analysis for Persian text. In *Proceedings of the International Conference Recent Advances in Natural Language Processing.*

[Andoni Azpeitia and Garcia, 2018] Andoni Azpeitia, T. E. and Garcia, E. M. (2018). Extracting parallel sentences from comparable corpora with stacc variants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).*

[Aqeel et al., 2006] Aqeel, S. U., Beitzel, S. M., Jensen, E. C., Grossman, D. A., and Frieder, O. (2006). On the development of name search techniques for arabic. *JASIST*, 57(6) :728–739.

[Aransa, 2015] Aransa, W. (2015). *Statistical Machine Translation of the Arabic Language.* Theses, Université du Maine.

[Argamon et al., 2007] Argamon, S., Bloom, K., Esuli, A., and Sebastiani, F. (2007). Automatically determining attitude type and force for sentiment analysis. In *Human Language Technology. Challenges of the Information Society, Third Language and Technology Conference, LTC 2007, Poznan, Poland, October 5-7, 2007, Revised Selected Papers*, pages 218–231.

[Asher et al., 2008] Asher, N., Benamara, F., and Mathieu, Y. Y. (2008). Distilling opinion in discourse : A preliminary study. Manchester, Royaume Uni.

[Auer, 1999] Auer, P. (1999). From codes-switching via language mixing to fused lects : Toward a dynamic typology of bilingual speech. *International Journal of Bilingualism*, 3(4) :309–332.

[Baccianella et al., 2010] Baccianella, S., Esuli, A., and Sebastiani, F. (2010). SentiWordNet 3.0 : An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*.

[Balahur et al., 2012] Balahur, A., Hermida, J. M., and Montoyo, A. (2012). Building and exploiting emotinet, a knowledge base for emotion detection based on the appraisal theory model. *IEEE Transactions on Affective Computing*, 3(1).

[Berardi et al., 2015] Berardi, G., Esuli, A., and Marcheggiani, D. (2015). Word embeddings go to italy : A comparison of models and training datasets. In *IIR*.

[Bestgen, 2002] Bestgen, Y. (2002). Détermination de la valence affective de termes dans de grands corpus de textes.

[Bestgen, 2006] Bestgen, Y. (2006). Déterminer automatiquement la valence affective de phrases : Amélioration de l'approche lexicale.

[Bo et al., 2011] Bo, L., Gaussier, É., and Aizawa, A. N. (2011). Clustering Comparable Corpora For Bilingual Lexicon Extraction. In *ACL-HLT 2011*, pages 473–478, Portland, Oregon, United States. Association for Computational Linguistics.

[Bouamor et al., 2014] Bouamor, H., Habash, N., and Oflazer, K. (2014). A multidialectal parallel corpus of arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014.*, pages 1240–1245.

[Bouamor et al., 2018] Bouamor, H., Habash, N., Salameh, M., Zaghouani, W., Rambow, O., Abdulrahim, D., Obeid, O., Khalifa, S., Eryani, F., Erdmann, A., and Oflazer, K. (2018). The MADAR arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018.*

[Chklovski and Pantel, 2004] Chklovski, T. and Pantel, P. (2004). VerbOcean : Mining the web for fine-grained semantic verb relations. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 33–40, Barcelona, Spain. Association for Computational Linguistics.

[Darwish, 2013] Darwish, K. (2013). Arabizi detection and conversion to arabic. *CoRR*, abs/1306.6755.

[Das and Gambäck, 2014] Das, A. and Gambäck, B. (2014). Identifying languages at the word level in code-mixed indian social media text. In *Proceedings of the 11th International Conference on Natural Language Processing, ICON 2014, Goa, India, December 18-21, 2014*, pages 378–387.

[Déjean and Gaussier, 2007] Déjean, H. and Gaussier, E. (2007). Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. In Teubert, W., Krishnamurthy, and R., editors, *Corpus Linguistics : Critical Concepts in Linguistics*. Routledge, England.

[Dragos et al., 2018] Dragos, V., Battistelli, D., and Kelodjoue, E. (2018). Beyond sentiments and opinions : Exploring social media with appraisal categories. *2018 21st International Conference on Information Fusion (FUSION)*, pages 1851–1858.

[Dumais et al., 1998] Dumais, S. T., Littman, M. L., and Landauer, T. K. (1998). Automatic cross-language retrieval using latent semantic indexing. In *In Bibliography 149 Grefenstette, G., editor, Cross-Language Information Retrieval, volume 2 of The Springer International Series on Information Retrieval, pages 51–62. Springer US.*

[Duwairi et al., 2015] Duwairi, R., Ahmed, N. A., and Al-Rifai, S. Y. (2015). Detecting sentiment embedded in arabic social media - a lexicon-based approach. *Journal of Intelligent and Fuzzy Systems*, 29 :107–117.

[Eisele and Chen, 2010a] Eisele, A. and Chen, Y. (2010a). Multiun : A multilingual corpus from united nation documents. In *LREC*.

[Eisele and Chen, 2010b] Eisele, A. and Chen, Y. (2010b). MultiUN : A multilingual corpus from united nation documents. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*.

[El-Halees, 2011] El-Halees, A. (2011). Arabic opinion mining using combined classification approach.

[Elouardighi et al., 2018] Elouardighi, A., Maghfour, M., Hammia, H., and Aazi, F. Z. (2018). Analyse des sentiments à partir des commentaires facebook publiés en arabe standard ou dialectal marocain par une approche d'apprentissage automatique. In *Extraction et Gestion des Connaissances, EGC 2018, Paris, France, January 23-26, 2018*, pages 329–334.

[Farghaly and Shaalan, 2009] Farghaly, A. and Shaalan, K. F. (2009). Arabic natural language processing : Challenges and solutions. *ACM Trans. Asian Lang. Inf. Process.*, 8(4) :14 :1–14 :22.

[Fung, 1997] Fung, P. (1997). Finding terminology translations from non-parallel corpora. In *Fifth Workshop on Very Large Corpora*.

[Fung and Cheung, 2004a] Fung, P. and Cheung, P. (2004a). Mining very-non-parallel corpora : Parallel sentence and lexicon extraction via bootstrapping and e. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*.

[Fung and Cheung, 2004b] Fung, P. and Cheung, P. (2004b). Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04.

[Gafaranga and Torras, 2002] Gafaranga, J. and Torras, M.-C. (2002). Interactional otherness : Towards a redefinition of code-switching. *International Journal of Bilingualism*, 6(1) :1–22.

[Gardin, 2009] Gardin, P. (2009). Application de la théorie de l'Appraisal à l'analyse d'opinion . In *Actes de la 17ème conférence sur le traitement automatique des langues naturelles (TALN 2010)*.

[Ghannay, 2017] Ghannay, S. (2017). *A study of continuous word representations applied to the automatic detection of speech recognition errors*. PhD thesis, Université du Maine.

[Ghosh et al., 2017] Ghosh, S., Ghosh, S., and Das, D. (2017). Complexity metric for code-mixed social media text. *CoRR*, abs/1707.01183.

[Go, 2009] Go, A. (2009). Sentiment classification using distant supervision.

[Goeuriot, 2009] Goeuriot, L. (2009). *Characterization and Compilation of Specialized Comparable Corpora*. PhD thesis, Université de Nantes.

[Guellil et al., 2018] Guellil, I., Adeel, A., Azouaou, F., and Hussain, A. (2018). Sentialg : Automated corpus annotation for algerian sentiment analysis. In Ren, J., Hussain, A., Zheng, J., Liu, C.-L., Luo, B., Zhao, H., and Zhao, X., editors, *Advances in Brain Inspired Cognitive Systems*.

[Halliday, 1994] Halliday, M. A. K. (1994). *An Introduction to Functional Grammar*. Hodder Arnold, third edition.

[Hamon et al., 2015] Hamon, T., Fraisse, A., Paroubek, P., Zweigenbaum, P., and Grouin, C. (2015). Analyse des émotions, sentiments et opinions exprimés dans les tweets : présentation et résultats de l'édition 2015 du défi fouille de texte (DEFT). In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2015)*, Caen, France.

[Harris, 1954] Harris, Z. S. (1954). Distributional structure. 10(2-3) :146–162.

[Hatzivassiloglou and McKeown, 1997] Hatzivassiloglou, V. and McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*.

[Hazem and Morin, 2018] Hazem, A. and Morin, E. (2018). Leveraging meta-embeddings for bilingual lexicon extraction from specialized comparable corpora. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 937–949.

[Honnet et al., 2018] Honnet, P.-E., Popescu-Belis, A., Musat, C., and Baeriswyl, M. (2018). Machine translation of low-resource spoken dialects : Strategies for normalizing swiss German. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

[Htait et al., 2017] Htait, A., Fournier, S., and Bellot, P. (2017). Identification semi-automatique de mots-germes pour l'analyse de sentiments et son intensité. In *COnférence en Recherche d'Informations et Applications - CORIA 2017, 14th French Information Retrieval Conference, Marseille, France, March 29-31, 2017. Proceedings.*, pages 415–424.

[Hu and Liu, 2004] Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04.

[Ismail and Manandhar, 2010] Ismail, A. and Manandhar, S. (2010). Bilingual lexicon extraction from comparable corpora using in-domain terms. In *Proceedings of the 23rd International Conference on Computational Linguistics : Posters*, COLING '10.

[Joshi, 1982] Joshi, A. K. (1982). Processing of Sentences with Intra-sentential Code-switching. In *Proceedings of the* 9th *Conference on Computational Linguistics - Volume 1*, COLING '82, pages 145–150.

[Jouvet et al., 2017] Jouvet, D., Langlois, D., Menacer, M. A., Fohr, D., Mella, O., and Smaïli, K. (2017). About vocabulary adaptation for automatic speech recognition of video data. In *ICNLSSP'2017 - International Conference on Natural Language, Signal and Speech Processing*, pages 1–5, Casablanca, Morocco.

[Kadri and Nie, 2006] Kadri, Y. and Nie, J.-Y. (2006). Effective stemming for arabic information retrieval. *The Challenge of Arabic for NLP/MT, International Conference at the British Computer Society (BCS) (pp. 68–74). London.*

[Koehn, 2005] Koehn, P. (2005). In *Conference Proceedings : the tenth Machine Translation Summit*, Phuket, Thailand.

[Korenek and Šimko, 2014] Korenek, P. and Šimko, M. (2014). Sentiment analysis on microblog utilizing appraisal theory. *World Wide Web.*

[Koźbiał and Leszczuk, 2019] Koźbiał, A. and Leszczuk, M. (2019). Collection, analysis and summarization of video content. In Choroś, K., Kopel, M., Kukla, E., and Siemiński, A., editors, *Multimedia and Network Information Systems*, pages 405–414, Cham. Springer International Publishing.

[Lavecchia et al., 2007a] Lavecchia, C., Smaïli, K., and Langlois, D. (2007a). Building a bilingual dictionary from movie subtitles based on inter-lingual triggers. In *Translating and the Computer*, Londres, United Kingdom.

[Lavecchia et al., 2007b] Lavecchia, C., Smaïli, K., and Langlois, D. (2007b). Building Parallel Corpora from Movies. In *The 4th International Workshop on Natural Language Processing and Cognitive Science - NLPCS 2007*, Funchal, Madeira, Portugal.

[Li and Gaussier, 2010] Li, B. and Gaussier, É. (2010). Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *COLING 2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2010, Beijing, China*, pages 644–652.

[Li and Gaussier, 2013] Li, B. and Gaussier, É. (2013). Exploiting comparable corpora for lexicon extraction : Measuring and improving corpus quality. In *Building and Using Comparable Corpora.*, pages 131–149.

[Ling et al., 2013] Ling, W., Xiang, G., Dyer, C., Black, A., and Trancoso, I. (2013). Microblogs as parallel corpora. In *Proceedings of the 51st Annual Meeting on Association for Computational Linguistics*, ACL '13. Association for Computational Linguistics.

[Liu et al., 2018] Liu, J., Morin, E., and Saldarriaga, P. (2018). Towards a unified framework for bilingual terminology extraction of single-word and multi-word terms. In *Proceedings of the 27th International Conference on Computational Linguistics.*

[Ma, 1999] Ma, X. (1999). Parallel text collections at linguistic data consortium.

[Mahyoub et al., 2014] Mahyoub, F. H., Siddiqui, M. A., and Dahab, M. Y. (2014). Building an arabic sentiment lexicon using semi-supervised learning. *Journal of King Saud University - Computer and Information Sciences*, 26(4) :417 – 424. Special Issue on Arabic NLP.

[Martin and White, 2005] Martin, J. R. and White, P. R. R. (2005). *The language of evaluation : appraisal in English / J.R. Martin and P.R.R. White.* Palgrave Macmillan Basingstoke.

[Mataoui et al., 2016] Mataoui, M., Zelmati, O., and Boumechache, M. (2016). A proposed lexicon-based sentiment analysis approach for the vernacular algerian arabic. *Research in Computing Science*, 110 :55–70.

[Medhaffar et al., 2017] Medhaffar, S., Bougares, F., Estève, Y., and Belguith, L. H. (2017). Sentiment analysis of tunisian dialects : Linguistic ressources and experiments. In *Proceedings of the Third Arabic Natural Language Processing Workshop, WANLP 2017@EACL, Valencia, Spain, April 3, 2017*, pages 55–61.

[Meftouh et al., 2019] Meftouh, K., Abidi, K., Harrat, S., and Smaïli, K. (2019). The SMarT Classifier for Arabic Fine-Grained Dialect Identification. In *The Fourth Arabic Natural Language Processing Workshop co-located with ACL*, Florence, Italy.

[Meftouh et al., 2015] Meftouh, K., Harrat, S., Jamoussi, S., Abbas, M., and Smaïli, K. (2015). Machine translation experiments on PADIC : A parallel arabic dialect corpus. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation, PACLIC 29, Shanghai, China, October 30 - November 1, 2015*.

[Meftouh et al., 2018] Meftouh, K., Harrat, S., and Smaïli, K. (2018). PADIC : extension and new experiments. In 7th *International Conference on Advanced Technologies ICAT*, Antalya, Turkey.

[Menacer et al., 2017a] Menacer, M., Mella, O., Fohr, D., Jouvet, D., Langlois, D., and Smaïli, K. (2017a). Development of the Arabic Loria Automatic Speech Recognition system (ALASR) and its evaluation for Algerian dialect. In *Third International Conference On Arabic Computational Linguistics, Dubai*.

[Menacer et al., 2017b] Menacer, M. A., Langlois, D., Mella, O., Fohr, D., Jouvet, D., and Smaïli, K. (2017b). Is statistical machine translation approach dead ? In *ICNLSSP 2017 - International Conference on Natural Language, Signal and Speech Processing*, pages 1–5, Casablanca, Morocco. ISGA.

[Menacer et al., 2017c] Menacer, M. A., Mella, O., Fohr, D., Jouvet, D., Langlois, D., and Smaili, K. (2017c). An enhanced automatic speech recognition system for arabic. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 157–165.

[Meng et al., 2012] Meng, X., Wei, F., Xu, G., Zhang, L., Liu, X., Zhou, M., and Wang, H. (2012). Lost in translations ? building sentiment lexicons using context based machine translation. In *COLING*.

[Mihalcea et al., 2007] Mihalcea, R., Banea, C., and Wiebe, J. (2007). Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 976–983, Prague, Czech Republic. Association for Computational Linguistics.

[Mikolov et al., 2013a] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In *ICLR (Workshop)*.

[Mikolov et al., 2013b] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26 : 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119.

[Mikolov et al., 2013c] Mikolov, T., Yih, W.-t., and Zweig, G. (2013c). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 746–751. Association for Computational Linguistics.

[Mohammad et al., 2009] Mohammad, S., Dunne, C., and Dorr, B. J. (2009). Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In *EMNLP*.

[Mohammad and Turney, 2013] Mohammad, S. and Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *CoRR*, abs/1308.6297.

[Mohammad et al., 2016] Mohammad, S. M., Salameh, M., and Kiritchenko, S. (2016). How translation alters sentiment. *J. Artif. Int. Res.*, 55(1).

[Morin et al., 2008] Morin, E., Daille, B., Kageura, K., and Takeuchi, K. (2008). Brains, not Brawn : The Use of ''Smart" Comparable Corpora in Bilingual Terminology Mining.

[Munteanu et al., 2004] Munteanu, D. S., Fraser, A., and Marcu, D. (2004). Improved machine translation performance via parallel sentence extraction from comparable corpora. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics : HLT-NAACL 2004*.

[Otero and Lopez, 2011] Otero, P. G. and Lopez, I. G. (2011). Measuring comparability of multilingual corpora extracted from wikipedia. *on Iberian Cross-Language Natural Language Processings Tasks (ICL 2011), page 8*.

[Ott et al., 2018] Ott, M., Edunov, S., Grangier, D., and Auli, M. (2018). Scaling neural machine translation. *CoRR*, abs/1806.00187.

[Pang and Lee, 2004] Pang, B. and Lee, L. (2004). A sentimental education : Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*.

[Parker et al., 2011] Parker, R., Graff, D., Kong, J., Chen, K., and Maeda, K. (2011). English gigaword fifth edition. *LDC2011T07. Web Download. Philadelphia : Linguistic Data Consortium, 2011*.

[Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. D. (2014). Glove : Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543.

[Prochasson, 2009] Prochasson, E. E. (2009). *Multilingual alignment from specialised comparable corpora*. PhD thesis, Université de Nantes.

[Ramesh and Sankaranarayanan, 2018] Ramesh, S. H. and Sankaranarayanan, K. P. (2018). Neural machine translation for low resource languages using bilingual lexicon induced from comparable corpora. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Student Research Workshop*.

[Rapp, 1995] Rapp, R. (1995). Identifying word translations in non-parallel texts. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, ACL '95.

[Redouan, 2005] Redouan, R. (2005). Linguistic Constraints on Code-switching and Code-mixing of Bilingual Moroccan Arabic-French Speakers in Canada . In *Proceedings of the 4th International Symposium on Bilingualism, MA : Cascadilla Press*.

[Rose et al., 1998] Rose, T., Kilgarriff, A., Kilgarriff, A., and Rose, T. (1998). Measures for corpus similarity and homogeneity. In *Proceedings of the 3rd conference on Empirical Methods in Natural Language Processing*, pages 46–52. ACL-SIGDAT.

[Rouces et al., 2018] Rouces, J., Tahmasebi, N., Borin, L., and Eide, S. R. (2018). Sensaldo : Creating a sentiment lexicon for swedish. In *LREC 2018, Eleventh International Conference on Language Resources and Evaluation, 7-12 May 2018, Miyazaki (Japan)*. ELRA.

[Rouvier and Favre, 2016] Rouvier, M. and Favre, B. (2016). Building a robust sentiment lexicon with (almost) no resource. *CoRR*, abs/1612.05202.

[Saad et al., 2013] Saad, M., Langlois, D., and Smaïli, K. (2013). Extracting Comparable Articles from Wikipedia and Measuring their Comparabilities. *Procedia - Social and Behavioral Sciences*, 95 :40–47. Corpus Resources for Descriptive and Applied Studies. Current Challenges and Future Directions : Selected Papers from the 5th International Conference on Corpus Linguistics (CILC2013).

[Saad et al., 2014] Saad, M., Langlois, D., and Smaïli, K. (2014). Cross-Lingual Semantic Similarity Measure for Comparable Articles. In *Advances in Natural Language Processing - 9th International Conference on NLP, PolTAL 2014, Warsaw, Poland, September 17-19, 2014. Proceedings*, pages 105–115, Warsaw, Poland. Springer International Publishing.

[Saralegi and Alegria, 2007] Saralegi, X. and Alegria, I. (2007). Similitud entre documentos multilinges de carácter científico-técnico en un entorno web. *Procesamiento del Lenguaje Natural*, 39.

[Saralegi et al., 2008] Saralegi, X., Vicente, I. S., and Gurrutxaga, A. (2008). Automatic extraction of bilingual terms from comparable corpora in a popular science domain.

[Shakery and Zhai, 2013] Shakery, A. and Zhai, C. (2013). Leveraging comparable corpora for cross-lingual information retrieval in resource-lean language pairs. *Inf. Retr.*, 16(1).

[Smaïli et al., 2018] Smaïli, K., Fohr, D., González-Gallardo, C., Grega, M., Janowski, L., Jouvet, D., Komorowski, A., Kozbial, A., Langlois, D., Leszczuk, M., Mella, O., Menacer, M. A., Mendez, A., Linhares Pontes, E., Sanjuan, E., Swist, D., Torres-Moreno, J.-M., and Garcia-Zapirain, B. (2018). A First Summarization System of a Video in a Target Language. In *MISSI 2018 - 11th edition of the International Conference on Multimedia and Network Information Systems*, pages 1–12, Wroclaw, Poland.

[Smith et al., 2010] Smith, J. R., Quirk, C., and Toutanova, K. (2010). Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10.

[Speer et al., 2016] Speer, R., Chin, J., and Havasi, C. (2016). Conceptnet 5.5 : An open multilingual graph of general knowledge. *CoRR*, abs/1612.03975.

[Stone et al., 1962] Stone, P. J., Bales, R. F., Namenwirth, J. Z., and Ogilvie, D. M. (1962). The general inquirer : A computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science*, 7(4) :484–498.

[Strapparava and Mihalcea, 2007] Strapparava, C. and Mihalcea, R. (2007). Semeval-2007 task 14 : Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 70–74, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Strapparava and Valitutti, 2004] Strapparava, C. and Valitutti, A. (2004). WordNet affect : an affective extension of WordNet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*.

[Su and Babych, 2012] Su, F. and Babych, B. (2012). Measuring comparability of documents in non-parallel corpora for efficient extraction of (semi-)parallel translation equivalents. In *Proceedings of the Joint Workshop on Exploiting Synergies Between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*, EACL 2012.

[Tamura et al., 2012] Tamura, A., Watanabe, T., and Sumita, E. (2012). Bilingual lexicon extraction from comparable corpora using label propagation. In *Proceedings of the 2012*

Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL '12.

[Tawwab and Eldin, 2014] Tawwab, A. A. and Eldin, S. (2014). Socio Linguistic Study of Code Switching of the Arabic Language Speakers on Social Networking. *International Journal of English Linguistics*, 4(6).

[Teubert, 1996] Teubert, W. (1996). Comparable or Parallel Corpora ? *International Journal of Lexicography*, 9(3) :238–264.

[Tian et al., 2018] Tian, L., Lai, C., and Moore, J. D. (2018). Polarity and intensity : the two aspects of sentiment analysis. *CoRR*, abs/1807.01466.

[Tim, 2004] Tim, B. (2004). Buckwalter arabic morphological analyzer version 2.0. *LDC catalog number LDC2004L02, Technical report, ISBN 1-58563-324-0.*

[Turney, 2002] Turney, P. (2002). Thumbs up or thumbs down ? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.*

[Turney and Littman, 2003] Turney, P. D. and Littman, M. L. (2003). Measuring praise and criticism : Inference of semantic orientation from association. *CoRR*, cs.CL/0309034.

[Wang et al., 2019] Wang, B., Wang, A., Chen, F., Wang, Y., and Kuo, C. J. (2019). Evaluating word embedding models : Methods and experimental results. *CoRR*, abs/1901.09785.

[Whitelaw and Garg, 2005] Whitelaw, C. and Garg, N. (2005). Using appraisal taxonomies for sentiment analysis.

[Whitelaw et al., 2005] Whitelaw, C., Garg, N., and Argamon, S. (2005). Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, CIKM '05.

[Wiebe et al., 2005] Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation.*

[Zhang and Ferrari, 2010] Zhang, L. and Ferrari, S. (2010). Analyse d'opinion : annotation sémantique de textes chinois. In *Actes de la 17ème conférence sur le traitement automatique des langues naturelles (TALN 2010)*, page 6pages, Montréal, Canada.

Résumé

Le traitement automatique des langues est fondé sur l'utilisation des ressources langagières telles que les corpus de textes, les dictionnaires, les lexiques de sentiments, les analyseurs morpho-syntaxiques, les taggers, etc. Pour les langues naturelles, ces ressources sont souvent disponibles. En revanche, lorsqu'il est question de traiter les langues peu dotées, on est souvent confronté au manque d'outils et de données. Dans cette thèse, on s'intéresse à certaines formes vernaculaires de l'arabe utilisées au Maghreb. Ces formes sont connues sous le terme de dialecte que l'on peut classer dans la catégorie des langues peu dotées. Exceptés des textes brutes extraits généralement des réseaux sociaux, il existe très peu de ressources permettant de traiter les dialectes arabes. Ces derniers, comparativement aux autres langues peu dotées possèdent plusieurs spécificités qui les rendent plus difficile à traiter. Nous pouvons citer notamment l'absence de règles d'écriture de ces dialectes, ce qui conduit les usagers à écrire le dialecte sans suivre des règles précises, par conséquent un même mot peut avoir plusieurs graphies. Les mots en arabe dialectal peuvent s'écrire en utilisant le script arabe et/ou le script latin (écriture dite arabizi).

Pour les dialectes arabes du Maghreb, ils sont particulièrement influencés par des langues étrangères comme le français et l'anglais. En plus de l'emprunt de mots de ces langues, un autre phénomène est à prendre en compte en traitement automatique des dialectes. Il s'agit du problème connu sous le terme de code-switching. Ce phénomène est connu en linguistique sous le terme de diglossie. Cela a pour conséquence de laisser libre cours à l'utilisateur qui peut écrire en plusieurs langues dans une même phrase. Il peut ainsi commencer en dialecte arabe et au milieu de la phrase, il peut "switcher" vers le français, l'anglais ou l'arabe standard. En plus de cela, il existe plusieurs dialectes dans un même pays et a fortiori plusieurs dialectes différents dans le monde arabe. Il est donc clair que les outils NLP classiques développés pour l'arabe standard ne peuvent être utilisés directement pour traiter les dialectes.

L'objectif principal de ce travail consiste à proposer des méthodes permettant la construction automatique de ressources pour les dialectes arabes en général et les dialectes du Maghreb en particulier. Cela représente notre contribution à l'effort fourni par la communauté travaillant sur le traitement automatique des dialectes arabes. Nous avons ainsi produit des méthodes permettant de construire des corpus comparables, des ressources lexicales contenant les différentes formes d'une entrée et leur polarité. Par ailleurs, nous avons développé des méthodes pour le traitement de l'arabe standard sur des données de Twitter et également sur les transcriptions provenant d'un système de reconnaissance automatique de la parole opérant sur des vidéos en arabe extraites de chaînes de télévisions arabes telles que *Al Jazeera, France24, Euronews*, etc. Nous avons ainsi comparé les opinions des transcriptions automatiques provenant de sources vidéos multilingues différentes et portant sur le même sujet en développant une méthode fondée sur la théorie linguistique dite *appraisal*.

Mots-clés: Dialectes du Maghreb, Arabe standard, Comparabilité, Analyse de sentiments, *Word emebdding*, *appraisal*, Transcriptions de parole.

Abstract

Automatic language processing is based on the use of language resources such as corpora, dictionaries, lexicons of sentiments, morpho-syntactic analyzers, taggers, etc. For natural languages, these resources are often available. On the other hand, when it comes to dealing with under-resourced languages, there is often a lack of tools and data. In this thesis, we are interested in some of the vernacular forms of Arabic used in Maghreb. These forms are known as dialects, which can be classified as poorly endowed languages. Except for raw texts, which are generally extracted from social networks, there is not plenty resources allowing to process Arabic dialects. The latter, compared to other under-resourced languages, have several specificities that make them more difficult to process. We can mention, in particular the lack of rules for writing these dialects, which leads the users to write the dialect without following strict rules, so the same word can have several spellings. Words in Arabic dialect can be written using the Arabic script and/or the Latin script (arabizi).

For the Arab dialects of the Maghreb, they are particularly impacted by foreign languages such as French and English. In addition to the borrowed words from these languages, another phenomenon must be taken into account in automatic dialect processing. This is the problem known as code-switching. This phenomenon is known in linguistics as diglossia. This gives free rein to the user who can write in several languages in the same sentence. He can start in Arabic dialect and in the middle of the sentence, he can switch to French, English or modern standard Arabic. In addition to this, there are several dialects in the same country and a fortiori several different dialects in the Arab world. It is therefore clear that the classic NLP tools developed for modern standard Arabic cannot be used directly to process dialects.

The main objective of this thesis is to propose methods to build automatically resources for Arab dialects in general and more particularly for Maghreb dialects. This represents our contribution to the effort made by the community working on Arabic dialects. We have thus produced methods for building comparable corpora, lexical resources containing the different forms of an input and their polarity. In addition, we developed methods for processing modern standard Arabic on Twitter data and also on transcripts from an automatic speech recognition system operating on Arabic videos extracted from Arab television channels such as *Al Jazeera, France24, Euronews,* etc. We compared the opinions of automatic transcriptions from different multilingual video sources related to the same subject by developing a method based on linguistic theory called Appraisal.

Keywords: Maghreb Dialects, Modern Standard Arabic, Comparability, Sentiment Analysis, Word emebdding, Appraisal, Speech transcriptions.