



Traitement automatique du dialecte algérien : construction de ressources et d'outils linguistiques

Abdelhakim Benali, Lamia Hadrich Belguith, Mohamed Hedi Maaloul

► To cite this version:

Abdelhakim Benali, Lamia Hadrich Belguith, Mohamed Hedi Maaloul. Traitement automatique du dialecte algérien : construction de ressources et d'outils linguistiques. Informatique [cs]. Faculté des sciences économiques et de gestion Sfax, 2023. Français. <NNT : >. <tel-04732684>

HAL Id: tel-04732684

<https://hal.science/tel-04732684v1>

Submitted on 11 Oct 2024

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.



Copyright - All rights reserved

THÈSE DE DOCTORAT

*Une thèse présentée pour l'obtention du diplôme de
Docteur en philosophie
dans*

L'INFORMATIQUE

par

Abdelhakim BENALI

Traitement automatique du dialecte
algérien : construction de ressources
et d'outils linguistiques

Soutenue le 27/12/2023 devant le comité composé de :

Mme. Ikram Amous	Professeur, ENETCOM-Sfax	Président
Mme. Lamia Hadrich Belguith	Professeur, FSEG-Sfax	Directeur de thèse
Mme. Leila Ben Ayed	Professeur, ENSI-Tunis	Rapporteur
Mr. Mohamed Tarek Hamdani	Maitre de conférences, ISI-Mahdia	Rapporteur
Mr. Chafik Aloulou	Maître de conférences, FSEG-Sfax	Examineur

Année universitaire 2020-2021

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

وَقَدْ عَلِمْتُ أَنِّي

صَدَقَ اللَّهُ الْعَظِيمُ

*À l'âme de mon père, Abdenmour, dont l'amour et la sagesse
continuent à illuminer ma vie, même en son absence.*

*À ma mère, dont le courage et la tendresse ont été mes premières
leçons de vie.*

*À mon frère, compagnon de jeux et confident de toujours, qui a
partagé avec moi les joies et les défis de notre parcours.*

*À ma femme, mon amour et mon soutien indéfectible, dont la
présence dans ma vie m'apporte chaque jour bonheur et réconfort.*

*Et à ma petite-fille Lyne, dont les yeux brillent d'une curiosité
infinie et dont le sourire éclaire mes journées. Puisses-tu grandir
dans un monde empreint de savoir et de compréhension.*

*À ma belle-sœur, qui est devenue une partie précieuse de notre
famille, apportant sa chaleur et son amour.*

*Ce travail est dédié à vous tous, en reconnaissance de votre
amour, votre soutien indéfectible et votre foi en mes capacités.*

*Que ces pages écrites soient un témoignage humble de ma
gratitude éternelle.*

Votre dévoué,

Abdelhakim Benali



Remerciements

Tout d'abord, je voudrais louer et remercier Dieu, le Tout-Puissant, qui a accordé d'innombrables bénédictions, connaissances et opportunités à l'auteur, me permettant ainsi de finalement accomplir cette thèse.

Cette thèse a été réalisée dans le cadre d'une collaboration entre l'Algérie et la Tunisie, sous l'égide du Ministère de l'Enseignement Supérieur et de la Recherche Scientifique de l'Algérie et du Ministère de l'Enseignement Supérieur et de la Recherche Scientifique de la Tunisie, dans le but d'intensifier la coopération entre les deux pays dans le domaine de la science et de la technologie.

Je tiens à exprimer ma sincère et profonde gratitude envers ma directrice de thèse, Madame Lamia Hadrich Belguith, professeure à la FSEGS, pour tous les efforts déployés dans cette thèse, depuis ses conseils, le temps consacré au travail, jusqu'à ses orientations. Je suis très reconnaissant d'avoir Monsieur Mohamed Hedi Maaloul en tant que co-directeur de thèse. Il m'a été d'une aide immense tout au long de ma carrière universitaire.

Je tiens à remercier le Pr Leila Ben Ayed et le Dr Mohamed Tarek Hamdani d'avoir accepté de lire mon manuscrit et de m'avoir fourni des commentaires utiles pour améliorer la qualité de mon travail.

Un merci particulier au Pr Ikram Amous, professeur à l'ENETCOM, de m'avoir honoré en acceptant d'être le président du comité d'examen de ma thèse. Je tiens également à remercier le Dr Chafik Aloulou, maître de conférences à FSEGS, d'avoir accepté l'invitation à faire partie du comité d'examen.

Enfin, ce travail n'aurait pas été possible sans le soutien de ma mère, Rabia; mon frère, Badreddine; ma femme, Khawla; ma fille, Lyne; ma belle-sœur, Zineb, ma famille et ma belle-famille. Mes amis, Saidi Abdessamad, Mehdi, Lokmane, Amine et Seif également. Ils ont toujours été présents à mes côtés en temps de besoin, avec leur amour constant, leurs encouragements, leur soutien moral et leurs bénédictions.



Résumé

La thèse se situe au cœur du développement de la technologie du Traitement Automatique du Langage Naturel (TALN) avec un intérêt particulier pour la langue arabe et le dialecte algérien. L'étude commence par une exploration approfondie de la complexité de la langue arabe, mettant en lumière les nuances entre l'arabe standard et les divers dialectes, en mettant un accent particulier sur le rôle central du dialecte algérien dans la vie quotidienne en Algérie.

Le principal défi du dialecte algérien réside dans son absence de règles linguistiques définies. Pour surmonter cette complexité, un corpus de données a été constitué, rassemblant des données écrites et parlées, fournissant ainsi la base essentielle pour le développement d'outils de traitement automatisé spécifiques au dialecte algérien.

L'étude propose également une méthode innovante pour la détection des frontières de phrases dans le dialecte algérien, une étape cruciale pour le traitement automatique des langues. Cette approche, basée sur des critères linguistiques, vise à améliorer la précision de cette tâche spécifique dans le contexte du dialecte algérien.

Enfin, l'analyse syntaxique du dialecte algérien est examinée en détail, avec le développement d'un analyseur grammatical spécifique utilisant un Treebank adapté au dialecte tunisien. En résumé, cette thèse représente une contribution significative au domaine du TALN en se concentrant sur le dialecte algérien. Grâce à la création d'un corpus de données, à l'élaboration d'une méthode de détection des frontières de phrases et au développement d'un analyseur grammatical adapté, cette recherche ouvre de nouvelles perspectives pour une interaction plus fluide entre les machines et le dialecte algérien, promouvant ainsi l'utilisation de l'intelligence artificielle dans le contexte linguistique algérien.

Mots-clés : Dialecte Algérien, Traitement automatique du langage naturel, Segmentation, Analyse syntaxique.



Abstract

The thesis delves into the heart of Natural Language Processing (NLP) technology development, with a specific focus on the Arabic language and the Algerian dialect. The study commences with a comprehensive exploration of the complexity of the Arabic language, highlighting the nuances between standard Arabic and various dialects, with a special emphasis on the pivotal role of the Algerian dialect in daily life in Algeria.

The primary challenge of the Algerian dialect lies in its lack of defined linguistic rules. To overcome this complexity, a comprehensive dataset has been compiled, incorporating written and spoken data. This dataset forms the essential foundation for the development of specialized automated processing tools tailored to the Algerian dialect.

Additionally, the study introduces an innovative method for sentence boundary detection in the Algerian dialect, a crucial step in natural language processing. This approach, grounded in linguistic criteria, aims to enhance the accuracy of this specific task within the context of the Algerian dialect.

Lastly, the syntactical analysis of the Algerian dialect is meticulously examined, involving the development of a dedicated grammatical analyzer utilizing a Treebank adapted for the Tunisian dialect. In summary, this thesis constitutes a substantial contribution to the field of NLP by focusing on the Algerian dialect. Through the creation of a robust dataset, the formulation of a precise sentence boundary detection method, and the development of a tailored grammatical analyzer, this research paves the way for smoother interaction between machines and the Algerian dialect, thereby promoting the utilization of artificial intelligence within the Algerian linguistic context.

Keywords : Algerian dialect, Automated Natural Language Processing, Segmentation, Syntactic analysis.



Table des matières

Remerciements	iii
Résumé	iv
Abstract	v
Table des figures	ix
Liste des tableaux	xi
Liste des abréviations	xii
Introduction générale	1
1 Caractéristiques générales de la langue arabe : du littéral au dialectal algérien	8
1.1 Introduction	9
1.2 La langue arabe écrite	10
1.2.1 La langue arabe, histoire, évolution et controverses	10
1.2.2 Particularités de la langue arabe écrite	12
1.2.3 Les différences entre la langue arabe et ses dialectes	14
1.3 Le dialecte algérien	15
1.3.1 Historique du dialecte algérien	16
1.3.2 La linguistique de dialecte algérien	17
1.3.3 Répartition sociolinguistique	18
1.4 Le dialecte algérien vs la langue arabe	22
1.4.1 Les caractéristiques phonologiques	23
1.4.2 Les caractéristiques morphologiques	26

1.4.3	Les caractéristiques lexicales	32
1.4.4	Les caractéristiques syntaxiques	32
1.5	Conclusion	33
2	Exploration et construction du corpus Liée au dialecte algérien	35
2.1	Introduction	36
2.2	Les corpus existants	37
2.2.1	Les corpus existants pour l'arabe et ses dialectes	38
2.2.2	Les corpus existants pour le dialecte algérien	39
2.3	Construction du corpus d'étude	41
2.3.1	Acquisition des données audio pour le corpus algérien	42
2.3.2	Acquisition des données écrites pour le corpus algérien	53
2.4	Caractéristiques de l'ALGDC : un Aperçu approfondi	57
2.4.1	Exploration des données parlées	58
2.4.2	Exploration des données écrites	60
2.4.3	Récapitulatif et synthèse : découverte de l'ALGDC	62
2.5	Évaluation comparative des corpus du dialecte algérien	64
2.6	Conclusion	65
3	Segmentation automatique de textes algériens en arabe dialectal	67
3.1	Introduction	68
3.2	Travaux préalables sur la segmentation des textes arabes et dialectaux .	69
3.2.1	Travaux réalisés au niveau de la langue arabe	69
3.2.2	Travaux réalisés au niveau des dialectes arabes	72
3.2.3	Travaux réalisés au niveau de dialecte algérien	73
3.3	Défis de la segmentation automatique de texte	74
3.3.1	Difficultés de segmentation dans les textes arabe	75
3.3.2	Difficultés de segmentation dans le dialecte algérien	80
3.4	Notre méthode proposée pour la segmentation de texte algérien en phrases	84
3.4.1	Contexte de la méthode	84
3.4.2	Inspiration de travaux antérieurs	85
3.4.3	Incorporation des règles du dialecte tunisien	86
3.4.4	Présentation de la méthode proposée	87
3.4.5	Méthode détaillée de segmentation du dialecte algérien	88
3.5	Réalisation et évaluation	92
3.6	Analyse des erreurs de segmentation	95

3.6.1	Contexte de l'erreur	96
3.6.2	Exemple d'erreur	96
3.6.3	Actions correctives	96
3.7	Conclusion	97
4	Analyse syntaxique dans le dialecte algérien : méthodes et résultats	99
4.1	Introduction	99
4.2	Exploration approfondie de l'analyse syntaxique	100
4.2.1	Avancements dans l'analyse syntaxique de l'arabe standard . . .	101
4.2.2	Avancements dans l'analyse syntaxique des dialectes arabes . . .	102
4.2.3	Innovations dans l'analyse syntaxique du dialecte algérien . . .	103
4.3	Défis et complexités : L'analyse syntaxique du dialecte algérien	103
4.3.1	Défis de l'analyse syntaxique	104
4.3.2	Complexités syntaxiques du dialecte algérien	106
4.3.3	Avantages de l'analyse syntaxique du dialecte algérien	108
4.3.4	Analyse syntaxique : arabe standard vs. dialecte algérien	109
4.4	Analyse syntaxique du dialecte algérien : une méthode proposée	111
4.4.1	Collection et préparation des données	112
4.4.2	Création du Treebank dédié au dialecte algérien	113
4.4.3	Architecture de la méthode	114
4.5	Réalisation et évaluation	120
4.5.1	Réalisation de Stanford-DZ	120
4.5.2	Évaluation de Stanford-DZ	121
4.5.3	Discussion des résultats	123
4.5.4	Perspectives	124
4.6	Conclusion	125
	Conclusion et perspectives	127
	Liste des publications	133
	Bibliographie	134



Table des figures

1	Structure hiérarchique des dialectes algériens[21]	17
2	Répartition de locuteurs par sexe	59
3	Hiérarchie pour la construction de ALGDC	64
4	méthode proposé pour la segmentation du dialecte algérien	88
5	Algorithme exprimant la Règle 1	90
6	Algorithme exprimant la Règle 2	90
7	méthode proposé pour la création d'un analyseur syntaxique du dialecte algérien	115
8	Treebank après l'analyseur syntaxique du dialecte tunisien	118
9	Treebank après correction pour le dialecte algérien.	118
10	Exemple d'un arbre syntaxique pour le dialecte algérien.	120

Liste des tableaux

1	exemple de voyellation en MSA	13
2	Conjugaison du verbe 'aimer' au présent, passé et impératif	20
3	différence lexicale entre les trois régions d'algerie	22
4	Système vocalique du dialecte algérien	24
5	Système consonantique du dialecte algérien	25
6	conjugaison du dialecte algérien en aoriste (présent ou future)	27
7	conjugaison du dialecte algérien en prétérit (passé général)	27
8	Pronoms affixes possessifs	28
9	Pronoms affixes compléments d'un verbe	28
10	conjugaison du verbe avoir en présent	29
11	conjugaison du verbe être en présent	29
12	conjugaison du verbe être en imparfait	30
13	conjugaison du verbe être en futur	30
14	pronoms sujet de DAlg	33
15	Exemples de prononciation exceptionnelle du dialecte algérien	49
16	Tokenisation du mot وكنيناها لكم	50
17	échelle de classement de degré d'accord selon la valeur de Kappa [72]	52
18	Répartition des Locuteurs par Âge	59
19	Répartition d'enregistrement par domaine	59
20	répartition des vidéos téléchargées en fonction de réseaux sociaux	60
21	Répartition d'enregistrement par domaine	61
22	description des corpus prise	62
23	description des corpus ALGDC	63
24	Comparaison entre les Corpus Existant et ALGCD pour le Dialecte Algérien	65

25	Format de règles conçues pour la segmentation de texte	89
26	Les résultats de l'évaluation sur la segmentation des textes en dialecte algérien	94
27	Résultats de l'évaluation de Stanford-DZ	122



Liste des abréviations

TALN : Traitement Automatique du Langage Naturel

DA : Dialecte Arabe

ASM : Arabe Standard Moderne

DT : Dialecte Tunisien

DALG Dialecte Algérien

ALGDC ALGérien Dialect Corpus

SegDz Segmentaire du Dialecte Algérien

SegTn Segmentaire du dialecte Tunisien



Introduction générale

Contexte et motivation

L'histoire de la communication humaine est une saga riche et diversifiée, où l'humanité a utilisé des gestes, des langages et même des dessins pour transcender les barrières linguistiques et transmettre des idées et des émotions. Cette capacité à communiquer a été au cœur de notre développement en tant qu'espèce, permettant le partage de connaissances, la coordination sociale et le renforcement des liens culturels. À travers les continents et les siècles, différentes langues ont émergé, chacune porteuse de sa propre histoire et de sa propre identité culturelle. Cependant, cette diversité linguistique a également créé des défis considérables en matière de communication interculturelle [33]. Les difficultés à comprendre et à se faire comprendre dans un monde multilingue ont été un défi constant, mettant en évidence le besoin croissant de solutions pour surmonter ces obstacles et faciliter la communication entre les peuples.

Au cours de l'histoire, l'humanité a franchi une étape cruciale en passant de simples formes de communication à la création de machines capables de comprendre et d'utiliser le langage humain. Cette évolution remarquable a ouvert la voie à un vaste éventail d'applications technologiques, allant de la reconnaissance automatique de la parole à la traduction automatique et au résumé automatique. Ces avancées ont révolutionné notre manière de travailler, de comprendre et de partager l'information, nous plongeant dans un monde d'opportunités et de défis inédits. La recherche

continue dans ce domaine promet de repousser encore plus loin les limites de notre interaction avec les machines[48], offrant des perspectives passionnantes pour l'avenir de la communication humaine.

Au fil des siècles, la langue arabe s'est affirmée comme l'une des langues les plus anciennes et les plus parlées au monde, portant avec elle l'héritage riche et profond de l'histoire humaine. Dans les dernières décennies, elle a également été le théâtre d'une révolution remarquable dans le domaine du Traitement Automatique du Langage Naturel (TALN). Cette révolution a été alimentée par un intérêt croissant des chercheurs et des innovateurs pour les subtilités phonétiques, morphologiques et lexicales de la langue arabe. Cette complexité linguistique a incité à des études approfondies, donnant lieu à des avancées notables dans le domaine du TALN, permettant ainsi d'explorer et de comprendre la langue arabe de manière plus précise et sophistiquée.

La diversité linguistique de l'arabe, en tant que langue sémitique parlée dans de nombreux pays, a depuis longtemps captivé les linguistes et les chercheurs en informatique linguistique. Cette langue, considérée comme l'une des plus anciennes et influentes au monde, possède une richesse phonétique, morphologique et lexicale remarquable. L'arabe standard moderne, souvent enseigné dans les écoles et utilisé dans les médias, a établi un socle commun de communication dans le monde arabe[37]. Cependant, au-delà de cette norme, les nombreux dialectes arabes ont évolué de manière distincte dans diverses régions, reflétant les nuances culturelles et historiques spécifiques à chaque communauté. Ces dialectes, dont le dialecte algérien, présentent des différences linguistiques significatives, rendant ainsi leur traitement automatique un défi stimulant pour les chercheurs en linguistique informatique.

Le dialecte algérien, résultant d'une fusion complexe d'influences berbères autochtones, d'occupations coloniales française et espagnole[36], et de contacts avec d'autres langues méditerranéennes, se distingue par ses spécificités phonologiques, morphologiques et lexicales uniques. Cette diversité linguistique, reflétée dans son articulation distinctive des sons et son usage créatif du vocabulaire[21], offre un aperçu

captivant de l'histoire mouvementée de l'Algérie. Dans ce contexte, la coexistence complexe entre le dialecte algérien, l'arabe standard moderne (ASM) et les autres dialectes arabes crée un terrain linguistique complexe, où les normes éducatives et professionnelles se mêlent à la vitalité quotidienne du dialecte algérien. Cependant, malgré sa richesse, le dialecte algérien a été sous-étudié dans le domaine du Traitement Automatique du Langage Naturel (TALN), en grande partie en raison du manque d'outils spécifiques et de ressources adaptées. Cette lacune suscite l'innovation, incitant les chercheurs à concevoir des méthodes créatives pour comprendre et exploiter pleinement ce dialecte complexe. Dans cette quête, les chercheurs explorent les subtilités de la langue algérienne, cherchant à transcender les défis et à découvrir de nouvelles opportunités pour l'analyse linguistique avancée, offrant ainsi une perspective unique sur le riche patrimoine linguistique de l'Algérie et ouvrant des portes vers des applications pratiques innovantes dans le domaine en constante évolution du TALN.

Problématiques

Le cœur de notre problématique réside dans le traitement automatique du dialecte algérien, une tâche complexe entravée par le manque crucial de ressources, telles que des corpus et des données, nécessaires à l'avancement du Traitement Automatique du Langage Naturel (TALN) dans ce contexte spécifique. Cette pénurie de matériel linguistique représente un obstacle majeur pour les chercheurs, limitant leur capacité à concevoir des outils sophistiqués pour comprendre et interagir efficacement avec le dialecte algérien. L'absence de corpus de qualité, essentiel pour l'entraînement des modèles de TALN, entrave la création de technologies avancées de traitement du langage. De plus, le manque d'outils syntaxiques, morphologiques et sémantiques adaptés complique davantage le développement de machines capables de comprendre les subtilités de ce dialecte.

La rareté de l'usage du dialecte algérien dans les communications officielles et écrites aggrave ce problème. L'absence d'une norme linguistique formelle pour le dialecte, combinée à sa faible présence dans les contextes officiels, a réduit l'incitation à investir dans le développement de ressources spécifiques. En conséquence, les chercheurs se retrouvent dans un cercle vicieux, où le manque de données et d'outils pertinents limite les avancées dans le domaine du TALN, ce qui, à son tour, contribue à perpétuer le manque de ressources. Cette situation représente un défi de taille, car elle entrave non seulement la recherche académique mais aussi le développement de technologies cruciales telles que la reconnaissance vocale, la traduction automatique et l'analyse sémantique pour le dialecte algérien.

Ainsi, la résolution de ce problème exige non seulement un investissement accru dans la collecte et la création de ressources linguistiques dédiées au dialecte algérien, mais aussi un engagement envers la normalisation et la reconnaissance officielle de ce dialecte dans les contextes formels. Ces efforts conjoints sont essentiels pour surmonter ces défis, ouvrant ainsi la voie à des avancées significatives dans le domaine du TALN pour le dialecte algérien.

Objectifs et contributions

Cette thèse représente une exploration approfondie et une contribution significative à la compréhension et au traitement du dialecte algérien. Notre objectif principal était de jeter la lumière sur les multiples facettes du dialecte algérien, en détaillant ses caractéristiques phonologiques, morphologiques et lexicales, ainsi que les distinctions essentielles qui le séparent de la langue arabe standard moderne (ASM). Nous avons minutieusement examiné les spécificités linguistiques du dialecte, des sons distinctifs aux structures morphologiques et à l'usage lexical unique, pour offrir une compréhension complète de cette variante linguistique complexe.

Dans le cadre de notre recherche, nous avons entrepris la création d'une base de données exhaustive du dialecte algérien, comprenant à la fois des données écrites et

des enregistrements de parole. Cette base de données est sans précédent par sa taille et sa diversité, ce qui en fait une ressource inestimable pour la recherche en TALN sur le dialecte algérien. Nous avons également exposé en détail notre méthodologie pour nettoyer, normaliser et prétraiter ces données, afin de les rendre exploitables pour l'entraînement de modèles de traitement automatique du langage.

En outre, nous avons abordé un aspect critique du TALN pour le dialecte algérien en développant une méthode de segmentation. Cette méthode permet de déterminer précisément le début et la fin des phrases en dialecte algérien, facilitant ainsi le traitement automatique et l'analyse linguistique des textes. Nous avons expliqué en détail les règles linguistiques et les critères qui sous-tendent cette segmentation, contribuant ainsi à une meilleure compréhension de la structure linguistique du dialecte.

Pour aller encore plus loin dans notre contribution au domaine du TALN pour le dialecte algérien, nous avons développé une méthode d'analyse syntaxique, accompagnée de la création d'un Treebank spécifique pour le dialecte algérien. Ces avancées ont permis d'appliquer des analyses syntaxiques avancées aux textes algériens, ouvrant la voie à des résultats significatifs et stimulants.

En résumé, cette thèse représente un effort complet pour démystifier et valoriser le dialecte algérien dans le domaine du TALN. Nos contributions en matière de bases de données, de segmentation, d'analyse syntaxique et de ressources linguistiques spécifiques pour le dialecte algérien ont pour but de favoriser la recherche future dans ce domaine et de permettre le développement de technologies de traitement du langage plus précises et sophistiquées pour cette variante linguistique riche et complexe.

Structure du document

Dans cette thèse, nous avons structuré notre document en quatre chapitres distincts, chacun représentant un pilier essentiel dans notre exploration approfondie du dialecte

algérien et sa place dans le contexte plus large de la langue arabe et du Traitement Automatique du Langage Naturel (TALN).

- **Chapitre 1 :** Dans ce chapitre inaugural, nous nous plongeons dans l'univers vaste et fascinant de la langue arabe, un langage ancien qui a évolué au fil des siècles, donnant naissance à une multitude de dialectes uniques. Nous examinons en profondeur les spécificités de la langue arabe moderne standard et, plus précisément, nous nous concentrons sur le dialecte algérien. En scrutant ses caractéristiques phonologiques, morphologiques et lexicales, nous mettons en lumière les nuances subtiles qui le distinguent, révélant ainsi la richesse linguistique de cette variante. Nous explorons également l'influence historique et culturelle qui a façonné le dialecte algérien, plongeant dans les strates complexes de l'histoire méditerranéenne, des conquêtes arabes aux interactions avec les langues berbères autochtones, du legs colonial français et espagnol à l'émergence d'une identité linguistique algérienne unique.
- **Chapitre 2 :** Ce chapitre se présente comme un pilier central de notre recherche, dévoilant la genèse et l'évolution de notre Corpus ALGDC. Nous détaillons les méthodes minutieuses de collecte de données, du recueil des conversations parlées aux transcriptions précises des textes écrits. En nous penchant sur les techniques de transcription spécifiques au dialecte algérien, nous exposons les défis et les solutions que nous avons rencontrés lors de cette phase cruciale. Notre évaluation comparative approfondie, comparant notre corpus avec d'autres ressources linguistiques disponibles pour d'autres dialectes arabes, offre un aperçu inestimable de la richesse et de la diversité du dialecte algérien dans le paysage linguistique global.
- **Chapitre 3 :** La segmentation des phrases dans le contexte du dialecte algérien constitue un enjeu majeur dans le traitement automatique du langage. Dans ce chapitre, nous plongeons dans les complexités de ce processus, explorant les méthodes traditionnelles et soulignant leurs limites dans le contexte spécifique du dialecte algérien. Nous détaillons ensuite notre approche novatrice, basée

sur des règles linguistiques sophistiquées, pour résoudre ce problème délicat. Nous examinons en détail la réalisation de notre méthode et son évaluation rigoureuse, démontrant ainsi son efficacité dans la segmentation précise des phrases dialectales algériennes.

- **Chapitre 4** : Dans ce chapitre final, nous plongeons dans les méandres de l'analyse syntaxique du dialecte algérien. Nous parcourons les recherches antérieures dans ce domaine, mettant en évidence les avancées et les lacunes existantes. En dévoilant notre méthodologie novatrice, basée sur des règles syntaxiques spécifiques au dialecte algérien, nous démontrons notre capacité à surmonter les défis uniques posés par la structure syntaxique complexe de ce dialecte. Notre évaluation approfondie met en lumière l'efficacité de notre méthode, ouvrant ainsi la voie à des analyses syntaxiques plus précises et sophistiquées pour le dialecte algérien.

À travers ces quatre chapitres, notre thèse représente une plongée profonde et détaillée dans le monde du dialecte algérien, éclairant ses nuances linguistiques, explorant les complexités de sa structure, et proposant des solutions innovantes pour son traitement automatique. Chaque chapitre représente un chapitre crucial dans notre effort continu pour comprendre et célébrer la richesse linguistique du dialecte algérien, offrant ainsi une contribution significative à la recherche en TALN et jetant les bases d'une exploration future plus approfondie dans ce domaine passionnant.

Chapitre

1

Caractéristiques générales de la langue arabe : du littéral au dialectal algérien

Table des matières partielles

2.1	Introduction	36
2.2	Les corpus existants	37
2.2.1	Les corpus existants pour l'arabe et ses dialectes	38
2.2.2	Les corpus existants pour le dialecte algérien	39
2.3	Construction du corpus d'étude	41
2.3.1	Acquisition des données audio pour le corpus algérien	42
2.3.2	Acquisition des données écrites pour le corpus algérien	53
2.4	Caractéristiques de l'ALGDC : un Aperçu approfondi	57
2.4.1	Exploration des données parlées	58
2.4.2	Exploration des données écrites	60
2.4.3	Récapitulatif et synthèse : découverte de l'ALGDC	62
2.5	Évaluation comparative des corpus du dialecte algérien	64
2.6	Conclusion	65

1.1 Introduction

L'arabe, en tant que langue appartenant à la famille des langues sémitiques, occupe une place particulière dans le monde linguistique. Elle est non seulement la langue la plus parlée, mais elle revêt également une importance religieuse en tant que langue du Coran dans la religion islamique. L'arabe se divise essentiellement en trois principales catégories : l'arabe standard moderne, l'arabe classique et les dialectes arabes. Cette diversité linguistique découle de la pluralité de ses locuteurs à travers le monde et des évolutions dans les moyens de communication, tels que la télévision, la radiodiffusion et les réseaux sociaux, qui ont engendré une grande variété dans cette langue [52].

Le progrès constant de la recherche scientifique et du développement technologique a imposé la nécessité de maîtriser pleinement la langue arabe et de l'utiliser de manière automatisée dans divers domaines. Cela a donné lieu à de nombreuses recherches visant à intégrer la langue arabe dans la technologie moderne. À ce jour, plusieurs travaux ont été menés dans le domaine du traitement automatique de la langue arabe, avec la mise en place d'outils visant à faciliter le traitement de cette langue. Ce domaine a connu des avancées remarquables au cours des dernières années, avec de nombreux chercheurs cherchant à simplifier le traitement automatique de l'arabe.

Certains chercheurs se sont orientés vers des aspects spécifiques de la langue arabe, notamment ses dialectes, qui varient d'un pays à l'autre en raison des coutumes, des traditions, des civilisations et des influences coloniales. Ces dialectes posent un défi particulier au traitement automatique du langage en raison de l'absence de normes grammaticales.

Parmi ces dialectes, le dialecte algérien se distingue, étant un mélange complexe de l'arabe, du français et du berbère. Cette hybridation résulte de la longue période de colonisation française en Algérie, ainsi que de la diversité géographique et culturelle

du pays, marquée par une succession de civilisations. Cette variété dialectale est parfois incompréhensible pour certains arabophones [23].

Dans ce chapitre, nous examinerons de près la langue arabe, son évolution et les avancées en matière de traitement automatique. Nous aborderons également les particularités de la langue arabe écrite, les différences entre l'arabe classique et les dialectes arabes, la diversité des dialectes arabes, notamment le dialecte algérien. Nous explorerons également la répartition sociolinguistique, les caractéristiques phonologiques et les spécificités lexicales et grammaticales du dialecte algérien.

1.2 La langue arabe écrite

La langue arabe, l'une des plus anciennes au monde, est reconnue pour sa richesse et sa complexité. Elle compte plus de 400 millions de locuteurs à travers le globe [20]. Cette langue possède des caractéristiques distinctives, une grammaire élaborée et un lexique riche, mais présente également des particularités telles que l'absence de voyelles dans son écriture standard et l'utilisation de marques de ponctuation spécifiques. Dans cette section, nous explorerons ces caractéristiques de la langue arabe écrite et les défis qu'elles posent en matière de TALN et de traitement des dialectes arabes.

1.2.1 La langue arabe, histoire, évolution et controverses

La langue arabe, appartenant à la famille des langues sémitiques, suscite depuis longtemps des débats parmi les historiens concernant son origine. Certains associent la langue arabe à des figures religieuses, affirmant que le prophète Ismaël l'a parlée dès l'âge de quatorze ans, voire qu'elle remonte à la langue d'Adam au paradis. Cependant, il est généralement admis que la langue arabe est devenue prééminente au sixième siècle après J.-C. grâce à la poésie et à la littérature de la période pré-islamique, et grâce à l'avènement du Saint Coran au septième siècle. Suite à la propagation de

l'islam et des conquêtes arabes dans la péninsule arabique et en Afrique du Nord après la mort du prophète Mahomet, la langue arabe s'est répandue dans le monde entier et est devenue la langue de la civilisation et du progrès [28].

Avec le temps, la langue arabe a évolué pour devenir une forme de diglossie, en raison du mélange entre les arabes et les non-arabes pendant les conquêtes islamiques. Une distinction est ainsi apparue entre l'arabe classique et l'arabe standard moderne. L'arabe classique est la langue dans laquelle le Coran a été révélé et est demeurée la langue des textes arabes jusqu'à l'ère islamique [54].

Les chercheurs, en particulier les sociolinguistes, continuent de débattre de la classification de la langue arabe. Selon l'approche de Ferguson et Charles [74], la langue arabe est divisée en deux parties distinctes. La première partie englobe l'arabe classique et l'arabe moderne standard, tandis que la seconde partie est constituée des dialectes arabes. La première section est utilisée dans des contextes tels que les discours religieux, l'éducation et les médias, tandis que la seconde section est employée dans la communication quotidienne entre différents groupes de la société et au sein des familles.

Il convient de noter qu'il existe trois strates de la langue arabe qui coexistent malgré leurs différences et leur degré d'institutionnalisation et de compréhension :

L'arabe classique (AC) : Il s'agit de la langue ancestrale, qui a prospéré au sein des anciennes civilisations arabes, notamment la civilisation pré-islamique. Elle est caractérisée par une riche tradition poétique et littéraire, ainsi que par son statut en tant que langue du Coran [56].

L'arabe standard moderne (ASM) : Cette variante de l'arabe classique a émergé en réponse à l'influence des sociétés étrangères et d'autres langues. Elle a été modifiée pour s'adapter aux échanges sociaux et politiques, enrichissant ainsi le dictionnaire arabe de nouveaux termes. L'ASM est enseigné dans les écoles [56].

Le dialecte arabe (DA) : Les dialectes varient d'un pays à l'autre, ainsi qu'entre les villes, les États, les zones rurales, les populations bédouines, les individus éduqués et

analphabètes. Il s'agit de la langue naturelle de l'homme, acquise dès la naissance sans nécessiter d'enseignement formel. Les dialectes arabes sont couramment utilisés dans les interactions quotidiennes au sein des différentes strates de la société.

1.2.2 Particularités de la langue arabe écrite

Contrairement à d'autres langues, l'arabe se révèle être l'une des langues les plus complexes à traiter automatiquement en raison de ses caractéristiques morphologiques et grammaticales particulières. Initialement, le traitement de la langue arabe se limitait aux dictionnaires et à la morphologie de la langue. Toutefois, avec les avancées technologiques, les progrès de la recherche scientifique, et l'émergence du World Wide Web et de l'Internet, des efforts ont été entrepris pour explorer le traitement de la langue arabe. Ces dernières années, le domaine du traitement automatique de la langue arabe a connu d'importants développements en matière de recherche et de projets, notamment dans le domaine du résumé automatique de textes arabes, de la reconnaissance automatique de la langue arabe et de l'indexation automatique des documents. Plusieurs outils ont été créés pour faciliter le traitement automatique, tels que des analyseurs grammaticaux et morphologiques, parmi d'autres [45].

Malgré ces avancées dans le domaine du traitement automatique, il subsiste divers défis liés à la langue arabe écrite, notamment sa complexité morphologique, l'absence de prononciation dans la plupart des textes arabes, et d'autres aspects.

1.2.2.1 Absence des voyelles

L'un des problèmes les plus courants lors du traitement automatique de la langue arabe réside dans l'absence de voyelles, car elles confèrent un sens distinct aux mots ou aux phrases. Les voyelles sont généralement indiquées en tant que signes diacritiques positionnés au-dessus ou au-dessous des lettres, et leur présence est plus fréquente dans des textes littéraires tels que la poésie, la prose, et les textes du Coran. L'absence de ces voyelles peut avoir un impact négatif sur la compréhension automatique des

textes arabes par les machines. En effet, chaque modification de la position des signes diacritiques peut entraîner un changement significatif dans le sens d'un mot.

mot voyellé	pré-notion	notion d'écrire
كَتَبَ	kataba	il a écrit
كُتِبَ	kutiba	il a été écrit
كُتُب	kutub	Des livres
كَتَبَ	katob	un écrit
كَتَّبَ	kattaba	il a fait écrire
كُتِّبَ	kuttiba	faire écrire-forme factitive
كَتَبَ	katibbo	fais écrire
كَتَّبَ	katabba	comme trancher
كَتَّبَ	katabb	comme 'tranchement'
...	...	...

TABLE 1. *exemple de voyellation en MSA*

1.2.2.2 Agglutination

La langue arabe se distingue par le fait que les pronoms, les prépositions et les articles sont en réalité attachés au nom, à l'adjectif ou à la particule à laquelle ils se rapportent. Cette caractéristique ne se retrouve pas dans les autres langues latines. Par conséquent, lors de la traduction d'un mot arabe en français, cette particularité peut engendrer une certaine ambiguïté morphologique lors du traitement automatique. Par exemple, le mot arabe "أتذكروننا" signifie en français : "est-ce que vous souvenez-vous de nous?". Il est difficile en langue arabe de faire la distinction entre un proclitique ou un enclitique attaché au caractère original du mot. Prenons par exemple le caractère "و" dans le mot "وصل" (il est arrivé). Dans ce cas, "و" fait partie des lettres du mot. Cependant, dans le mot "وفتح" (il a ouvert), "و" est un proclitique [60].

1.2.2.3 Absence de ponctuation régulière

Contrairement à d'autres langues, la langue arabe ne dépend pas uniquement des signes de ponctuation et des marqueurs typographiques. Il est donc rare de trouver ces derniers utilisés dans les textes arabes. On peut ainsi rencontrer un texte arabe complet sans aucune marque, à l'exception du point à la fin du paragraphe. Cette particularité rend le traitement automatique de cette langue plus complexe que celui du français ou de l'anglais, dans lesquels les signes de ponctuation peuvent être utilisés pour segmenter les phrases. En langue arabe, la division des phrases repose principalement sur les particules, les conjonctions de coordination, les mots-clés de la langue arabe, etc [58].

1.2.3 Les différences entre la langue arabe et ses dialectes

Comme mentionné précédemment, il existe des différences significatives entre la langue arabe standard et les dialectes en termes de grammaire, de vocabulaire et de prononciation. Ces différences varient selon les régions géographiques. Par exemple, les dialectes du Golfe sont plus proches de la langue arabe standard, tandis que les dialectes du Maghreb présentent des écarts plus marqués.

En ce qui concerne les règles linguistiques, la langue arabe standard dispose de règles fixes et strictes, ainsi que de nombreux dictionnaires et ouvrages de grammaire qui en font une langue normée et bien référencée. En revanche, les dialectes arabes sont caractérisés par des règles linguistiques variant d'une région à l'autre, ce qui les rend plus difficiles à enseigner et à normaliser.

En ce qui concerne le vocabulaire, la langue arabe standard possède un riche lexique composé de mots fixes soumis à des règles strictes. Ces mots sont standardisés pour tous les arabophones. En revanche, les dialectes se distinguent par l'utilisation de mots populaires variables et instables, parfois empruntés à d'autres langues.

En ce qui concerne la prononciation, la langue arabe standard dispose d'une prononciation unifiée, quelle que soit la région ou le pays. En revanche, les dialectes varient considérablement d'un pays à l'autre, d'une ville à l'autre, voire d'une région à l'autre.

De plus, la langue arabe standard est écrite et enseignée dans les écoles et les universités. Elle est la langue principale de nombreux pays arabophones et est enseignée dès le plus jeune âge. En revanche, l'apprentissage des dialectes arabes se fait principalement par le biais de l'interaction quotidienne et de l'usage dans des contextes familiaux et sociaux, ainsi que sur les réseaux sociaux.

1.3 Le dialecte algérien

L'Arabe moderne standard, par sa vocation de langue officielle dans de nombreux États arabes, s'illustre à travers les colonnes des journaux, les ondes des stations de radio, les pages des ouvrages et les discours des médias institutionnels. Il se dresse comme la forme la plus éminente de la langue arabe à l'échelle planétaire, souverainement enseignée dans les établissements éducatifs et utilisée pour les correspondances administratives. Toutefois, lorsque surviennent les interactions quotidiennes entre les membres d'une société, il advient que l'on se tourne, le plus souvent, vers le dialecte propre à chaque région. Ces dialectes, marqués par leur variabilité selon les traditions, les civilisations et les héritages coloniaux de chaque pays arabe, trouvent leur ancrage linguistique dans la base de l'Arabe moderne standard, auquel s'adjoint un lexique emprunté à d'autres langues, donnant ainsi naissance à une langue hybride d'usage courant [7].

Parmi ces dialectes, le dialecte algérien se distingue en tant que patrimoine linguistique des Algériens. Réputé pour sa complexité, en grande partie en raison de l'abondance de mots empruntés à la langue française, il fait face à une lacune notable dans les pratiques d'écriture, ce qui pose des défis considérables pour son automatisé dans le cadre du traitement automatisé du langage [23].

1.3.1 Historique du dialecte algérien

L'Algérie est le plus grand pays africain par superficie, sa superficie est estimée à 2.381.741 km² et comprend 48 départements. La première langue de l'Algérie est la langue arabe moderne standard, et avec l'expansion de la région et la multiplicité des villes, on constate qu'il existe de nombreux dialectes différents en Algérie, et environ 80% des habitants de l'Algérie utilisent le dialecte arabe algérien. Il existe des dialectes non arabes tels que le dialecte kabyle, et cela est dû aux différentes civilisations qui se sont installées en Algérie au fil du temps [14].

Au début, le dialecte algérien était berbère, et avec les conquêtes islamiques et l'étape d'arabisation qui arriva aux septième et onzième siècles, l'Algérie se familiarisa avec la langue arabe. Premièrement, au septième siècle, le processus d'arabisation n'a touché que certaines villes, telles que Tlemcen et Constantine et ses environs ruraux, et le dialecte de ces villes s'appelait Pre-Hilal, signifiant le dialecte sédentaire. Quant au reste des villes, il restait la langue berbère [61]. Deuxièmement, et au XIe siècle, le processus d'arabisation s'est généralisé, car il a touché de nombreuses villes algériennes et a été appelé le dialecte bédouin, et ce dernier est divisé en quatre sous-dialectes :

- Le dialecte Ma'qilian, qui touchait toutes les villes proches de la frontière marocaine, et le dialecte de ces villes était influencé par le dialecte marocain, comme Oran, Naàma, etc.
- Dialecte Sulaymite, qui regroupe toutes les villes qui ont des frontières avec la Tunisie, on retrouve donc son dialecte très proche du dialecte tunisien, comme la ville d'Annaba, Souk Ahras, Tebessa, etc.
- Dialecte Hilali, qui comprend trois sous-dialectes :
 - Hilali du désert, qui porte le dialecte utilisé dans le désert du Sahara en Algérie.
 - Hilali Tellien, qui couvre toutes les villes du Tell et les plaines de Constantine, et du nord de la région de Hodna à la rivière de Seybouse.

- Un dialecte entièrement bédouin, il comprend Alger et les villes côtières surplombant la mer. En raison des différences linguistiques, nous avons divisé ce dernier en deux parties AlgerBlanks et Sahel-Tell [47].

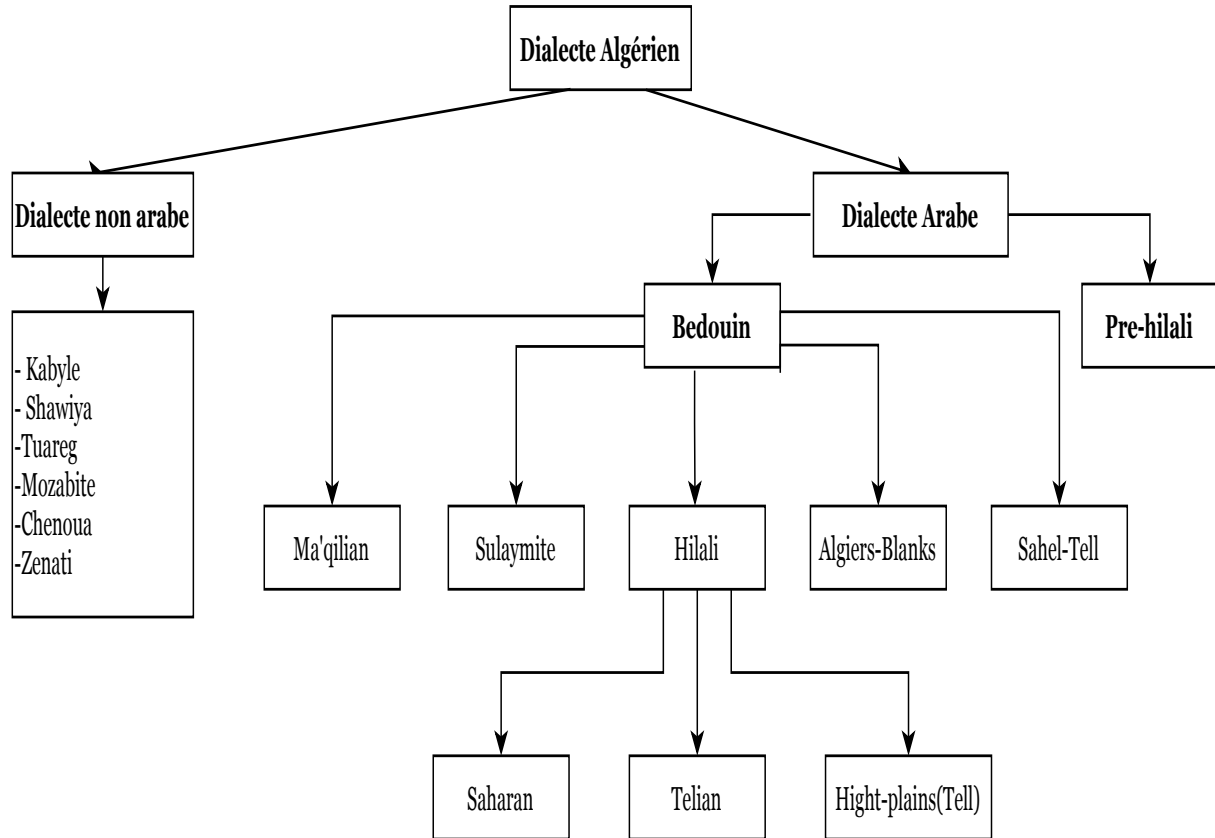


FIGURE 1. Structure hiérarchique des dialectes algériens[21]

1.3.2 La linguistique de dialecte algérien

Le dialecte algérien est couramment utilisé dans la vie quotidienne pour les communications entre les membres de la société. Il est présent dans les médias informels tels que la radio, les émissions de télévision, et les séries télévisées. De plus, il est largement répandu sur des plateformes en ligne telles que YouTube, grâce à des créateurs de contenu tels que des YouTubeurs et des Instagrammeurs. Contrairement à la langue arabe standard, le dialecte algérien n'est pas enseigné dans les écoles. Il est généralement acquis naturellement par les enfants lors de leur développement, car c'est le dialecte avec lequel ils grandissent. Cependant, il est peu présent dans les

communications écrites officielles. Ces dernières années, il a gagné en importance sur les médias sociaux, notamment à travers des messages écrits[23].

Le dialecte algérien présente également des caractéristiques orthographiques et morphologiques propres, tout comme la langue arabe standard.

1.3.3 Répartition sociolinguistique

Avec l'évolution linguistique en Algérie, le dialecte algérien est devenu la langue maternelle prédominante. La langue arabe standard moderne a acquis le statut officiel de la langue algérienne, se retrouvant dans les manuels scolaires, la législation présidentielle et les médias arabophones. Quant à la langue française, elle a occupé une place de choix dans la société algérienne en raison de son statut en tant que langue officielle pendant la période coloniale française. Après l'indépendance de l'Algérie dans les années 60, la langue arabe a été réaffirmée comme langue officielle dans la constitution algérienne.

L'utilisation de la langue arabe est prédominante dans l'enseignement primaire et secondaire, ainsi que dans certaines spécialités littéraires au niveau universitaire. En revanche, la langue française est la deuxième langue du pays et est utilisée pour l'enseignement de toutes les disciplines scientifiques et techniques[62].

Il existe une grande variation dans l'écriture du dialecte algérien, avec certains préférant les lettres latines et d'autres optant pour les caractères arabes, en fonction de leur niveau d'éducation et de leurs préférences linguistiques. Lorsque le dialecte algérien est utilisé à l'oral, il existe également des variations entre les locuteurs au sein de la communauté. Certains intègrent des mots provenant d'autres langues, notamment le français. Cela se traduit par des spécificités morphosyntaxiques et lexicales propres au dialecte algérien, résultant de l'influence historique de diverses civilisations sur la culture algérienne[11].

Le dialecte algérien est riche en emprunts linguistiques issus d'autres langues, principalement des langues des pays coloniaux ou des pays ayant influencé la vie du peuple algérien.

Parmi ces langues on trouve le turc, l'espagnol, l'italien et le français.

Par exemple, nous trouvons le mot سلام « Sallam » « salut » qui appartient à la langue arabe standard moderne .

De même, le mot فرملي « Farmliy » «infirmier» a une origine française.

Et le mot بابور «BaAbuwr » « bateau » qui est un mot turc.

Le terme شلاغم « šlaγ m » « moustaches » est un mot qui vient de la langue berbère.

Le mot سمانة « smAnh» «une semaine» est empruntée à la langue espagnole. [10].

1.3.3.1 Différences morphologiques

Le dialecte algérien se distingue par sa propre morphologie qui le différencie des autres dialectes. Tout d'abord, en ce qui concerne **la conjugaison des verbes**, contrairement à d'autres langues, la conjugaison dans le dialecte algérien repose soit sur l'ajout de préfixes au début du mot, soit sur l'ajout de suffixes à la fin, voire les deux, en fonction du verbe, de sa conjugaison, du pronom utilisé et du temps de conjugaison.

Le tableau 2 ci-dessous illustre la conjugaison du verbe 'aimer' dans le dialecte algérien, avec les temps les plus couramment utilisés, à savoir le présent, le passé et l'impératif.

Verbe	Présent	Traduction	Passé	Traduction	impératif	traduction
حب	أحب	J'aime	أحببت	J'ai aimé	——	——
	أحب	Tu aimes	أحببت	Tu as aimé	أحب	Aime
	أحبّي	Tu aimes	أحبّيتي	Tu as aimé	أحبّي	Aime
	أحب	Il aime	أحب	Il a aimé	——	——
	أحب	Elle aime	أحبت	Elle a aimé	——	——
	أحبو	Nous aimons	أحبينا	Nous avons aimé	أحبو	Aimons
	أحبو	Vous aimez	أحبّيتو	Vous avez aimé	أحبو	Aimez
	أحبو	Ils aiment	أحبو	Ils ont aimé	——	——

TABLE 2. *Conjugaison du verbe 'aimer' au présent, passé et impératif*

À partir du tableau 2, nous pouvons conclure que le dialecte algérien a les préfixes suivants " ن " et " ت " et " ي " Quant aux suffixes, ce sont les lettres suivantes " و ", " ي ", " ت ", " يت ", " ينا ", " يتي ", " يتو " Quant à la conjugaison future, elle est comme la conjugaison au présent avec l'ajout des mots dans le futur tel que أومبعد Awmζd راح rAH غدوة γdwh et qui signifient (après, aller, demain) [29].

Deuxièmement, la négation dans le dialecte algérien :

Il existe deux types de négation dans le dialecte algérien : la négation avec le verbe et la négation avec les noms et les adjectifs.

Dans la négation avec le verbe, on ajoute à la fois un préfixe et un suffixe. Prenons l'exemple du verbe "j'aime" qui, en dialecte algérien, se traduit par "محبو" au masculin et "محبها" au féminin. La négation de ce verbe est "je ne l'aime pas", ce qui se traduit par "ما محبوش". En comparaison avec la langue française, "ما" équivaut à "ne" et "ش" équivaut à "pas".

Quant au deuxième type de négation, qui concerne les noms et les adjectifs, on utilise le mot "ماشي" («mAshy»). Prenons, par exemple, la phrase : "Cette école est belle". En dialecte algérien, on dit "هذا المسيد شباب" («hAd Almsyd šbAb»). La négation de cette phrase serait "Cette école n'est pas belle", ce qui se traduit en dialecte algérien par

"هاد المسيد ماشي شباب" («hAd Almsyd mAšy šbAb»). Ainsi, dans ce cas, nous utilisons le mot "ماشي" («mAšy») pour exprimer la négation.

1.3.3.2 Différences phonologiques

Il existe des différences phonétiques entre les villes d'Algérie. Par exemple, la lettre arabe ق est prononcée différemment d'une région à l'autre.

Arrêt uvulaire ق : comme dans la langue arabe et le dialecte tunisien, on le trouve prononcé comme un q] dans la plupart des régions, en particulier dans les villes, comme Constantine et Alger.

Son palatin ق : on le trouve prononcé comme un g] dans de nombreuses régions d'Algérie, qui sont majoritairement bédouines, comme la ville d'Annaba, Sétif, etc.

Arrêt glottal ا : comme dans le dialecte égyptien, on le trouve en Algérie chez certains habitants de Tlemcen.

k postpalatal : ce son est propre au dialecte algérien, on ne le trouve pas dans les autres dialectes d'Afrique du Nord. Il est utilisé par les ruraux, comme dans les villes de Jijel et de Kabylie.

1.3.3.3 Différences lexicales

Le dialecte algérien présente des variations lexicales d'une région à l'autre, comme évoqué précédemment dans la section 1.3.1. Ces différences régionales entraînent des mots qui peuvent avoir plusieurs significations.

Dans le tableau 3, nous nous concentrerons sur trois régions distinctes de l'est, du centre et de l'ouest de l'Algérie pour illustrer la diversité des significations des mots et les variations des dictionnaires utilisés dans chaque région.

Dalg de l'est	Dalg de centre	Dalg de ouest	ASM	français
طرباقة	بليغة	نعالة	شيشب	Tong
زقرقوا	بوجغللوا	بابوش	حلزون	escargot
جرانة	قينقرا	قرقرة	ضفدع	grenouille
ياسر	بزاف	بزاف	كثير	beaucoup
نحكي	نهدر	نهدر	أتكلم	je parle
باهي	صح	نیشان	بالضبط	exacte
صحن	طبسي	طبسي	صحن	assiette
مغرف	مغيرفة	معيلقة	ملعقة	Cuillère
نحكم طاكسي	نشد طاكسي	نرفد طاكسي	أستقل سيارة اجرة	je prend un taxi
عضام	بيضة	بيضة	بيضة	un œuf
الجامع	المسيد	المدرسة	المدرسة	L'école
نحجم	نحفف	نحسن	أحلق شعري	je me coiffe
ازرب	غاؤل	خفف	تعالى بسرعة	Venir vite
البرمة	القدرة	القدرة	قدر	marmite
القرايج	الماعن	الماعن	الأطباق	la vaisselle

TABLE 3. *différence lexicale entre les trois régions d'algerie*

1.4 Le dialecte algérien vs la langue arabe

La langue arabe se distingue par sa grammaire fixe, qui la différencie des autres langues. Elle possède également des règles phonémiques, morphologiques et orthographiques strictes, composées de 28 lettres alphabétiques.

En ce qui concerne le dialecte algérien, sa grammaire dérive de la langue arabe, et il partage la plupart de ses lettres avec la langue arabe standard. Cependant, il présente quelques différences notables, notamment l'ajout de trois lettres spécifiques (ث) (پ) (ڤ).

Il existe de nombreuses différences entre le dialecte algérien et la langue arabe en ce qui concerne les phonèmes, la morphologie et l'orthographe.

Dans cette section, nous mettrons en lumière quelques-unes des différences significatives entre la langue arabe et ce dialecte.

1.4.1 Les caractéristiques phonologiques

Le dialecte algérien se distingue par plusieurs caractéristiques phonologiques qui le différencient des autres dialectes arabes, parmi lesquelles figurent son système vocalique et son système consonantique

1.4.1.1 Système vocalique

Le système vocalique dialectal algérien est un système hybride importé à la fois du système phonémique de la langue arabe et de la langue française.

Du côté de la langue arabe, ce système se compose de six phonèmes principaux : [a], [i], [o], [e], [ɛ], [u], ainsi que de leurs équivalents longs : [a:], [i:], [o:], [e:], [ɛ:], [u:].

En ce qui concerne la langue française, elle présente 16 voyelles : [a], [ɑ], [e], [ɛ], [i], [o], [ɔ], [u], [y], [ɑ], [œ], [ø], [ã], [ɔ̃], [ɛ̃], [œ̃].

Le système vocalique du dialecte algérien est le résultat d'une combinaison de la phonétique de la langue arabe et de la langue française, donnant ainsi naissance aux trois voyelles essentielles du dialecte algérien : [a], [u], [i], accompagnées de leurs équivalents longs [a:], [u:], [i:] [13].

Voyelles	exemple	Français	DAlg	ASM
[a]	bra	Il est guéri	برى	شفي
[u]	bi :ru	Un bureau	بيرو	مكتب
[i]	bqry	Bovin	بقري	بقري
[a :]	ba : k	Le bac	باك	البكالوريا
[u :]	ku :n	Sois	كون	كن
[i :]	bi :r	Un puits	بير	بئر

TABLE 4. *Système vocalique du dialecte algérien*

En comparant le système vocalique de la langue française et du DAlg, nous constatons qu'il y a certaines des voyelles françaises manquantes dans le dialecte Algérien, et dans ce cas elles sont remplacées par d'autres voyelles proches de celui-ci.

1.4.1.2 Système consonantique

Le système statique du dialecte algérien est essentiellement le système de la langue arabe avec quelques ajouts et modifications.

Le tableau 5 résume toutes les consonnes du système consonantique du DAlg.

Consonne API	Buckwalter	DAlg	MSA	Francais
[b]	[bunjɛ]	بنية	لكمة	Un coup de poing
[p]	[brtAbl]	برتابل	محمول	Un portable
[t]	[tlAH]	تلاح	بنفسه ألقى	Il s'est jeté
[θ]	[vwm]	ثوم	ثوم	ail
[ʒ]	[jbl]	جبل	جبل	Une montagne
[g]	[GAE]	فاع	الكل	Tout
[h]	[HSAn]	حصان	حصان	Un cheval
[χ]	[xAf]	خاف	خاف	Il a peur
[d]	[drAhm]	دراهم	نقود	argent
[*]	[*rAry]	ذراري	ذكور	Des garçons
[r]	[rAjl]	راجل	رجل	Un homme
[z]	[zyt]	زيت	زيت	Une huile
[s]	[smA]	سما	سما	ciel
[ʃ]	[ʃɛAf]	شاف	رأى	Il a vu
[s]	[sAm]	صام	صام	Il a jeuné
[D]	[Dw]	ضو	ضوء	Une lumière
[t]	[TAqh]	طاقة	نافذة	Une fenêtre
[ç]	[ç ɛ :m]	عام	عام	Un an
[gh]	[ghsl]	غسل	يغسل	Il a lavé
[f]	[fhɛm]	فهم	فهم لقد	Il a compris
[v]	[villeɛ]	فيلا	فيلا	Une villa
[q]	[qbɛl]	قبل	قبل	Il a accepté
[k]	[kbɛr]	كبر	كبر	Il a grandi
[l]	[lɛm]	لام	باللوم ألقى	Il a blâmé
[m]	[mi :nɛ]	مينا	منجم	Une mine
[n]	[nɛ]	نسي	نسي	oublié
[h]	[hbɛl]	هبل	مجنون اصبح	Il est devenu fou
[w]	[w ɛ :lu]	والو	شيئ لا	rien
[j]	[j ɛ mʃi]	يمشي	يمشي	Il marche
[z]	[yxzr]	يخزر	أنظر	regarde
[g]	[parkig]	پركينف	السيارات موقف	Parking

TABLE 5. *Système consonantique du dialecte algérien*

1.4.2 Les caractéristiques morphologiques

Le dialecte algérien se distingue par la présence de certaines caractéristiques morphologiques différentes de celles de la langue arabe, mais plutôt proches de l'accent nord-africain.

En ce qui concerne l'inflexion, dans le dialecte algérien, les terminaisons des noms et des verbes sont supprimées, ce qui est similaire à la plupart des dialectes arabes.

La conjugaison à l'indicatif est le temps le plus couramment utilisé par défaut, contrairement à d'autres modes qui ne se retrouvent pas dans le dialecte algérien. De même, le double pluriel et le pluriel féminin ont complètement disparu du dialecte algérien, étant remplacés par le pluriel masculin[36].

Ces détails concernent les caractéristiques morphologiques du dialecte algérien, qui englobent également d'autres aspects liés aux verbes, aux noms et à la catégorie grammaticale.

1.4.2.1 La morphologie verbale

Les verbes en DALg n'ont que deux temps principaux : l'aoriste, qui exprime des idées pour le présent et le futur ; et le prétérit, qui exprime toutes les nuances du passé. Pour conjuguer un verbe en dialecte algérien, il faut connaître le radical, qui est la troisième personne du singulier masculine au prétérit ou passé, comme **كتب، شرب** "šrb", "ktb", signifiant "il a bu", "il a écrit". Pour conjuguer un verbe en dialecte algérien, on ajoute les suffixes ou préfixes avant et après le verbe, en supposant que la racine soit exprimée avec un tiret (-).

	Français		Dalg
Singulier			
1 er pers	N————		ن
2 e pers	masc	T————	ت
	fém	T———— i	تي
3e pers	masc	Y————	ي
	fém	T————	ت
Pluriel			
1 er pers	N————ou		نوا
2 e pers	T————ou		توا
3e pers	Y————ou		يوا

TABLE 6. conjugaison du dialecte algérien en aoriste (présent ou future)

	Français		Dalg
Singulier			
1 er pers	———T		تْ
2 e pers	masc	———T	تْ
	fém	———Ti	تيْ
3e pers	masc	———	
	fém	———E'T	تْ
Pluriel			
1 er pers	———NA		نا
2 e pers	———Tou		تُوا
3e pers	———ou		وا

TABLE 7. conjugaison du dialecte algérien en prétérit (passé général)

En ce qui concerne les pronoms ou affixes complémentaires comprennent les sept syllabes suivantes : هم، كم، نا، ها، ه، ك، ي، -i/ek/ho ou hou/ha/na/koum/houm

Ces syllabes ne sont jamais réalisées isolément; elles sont ajoutées aux noms pour exprimer la possession, aux verbes et prépositions pour faire des pronoms des compléments directs ou indirects.

داري	Dar-i	Ma maison
دارك	Dar-ek	Ta maison
داره	Dar-hou	Sa maison (a lui)
دارها	Dar-ha	Sa maison (a elle)
دارنا	Dar-na	Notre maison
داركم	Dar-koum	Votre maison
دارهم	Dar-houm	Leur maison (pour les 2 genre)

TABLE 8. *Pronoms affixes possessifs*

شاف	Chaf	Il a vu
شافني	Chaf-ni	Il m'a vu ou vue
شافك	Chaf-ek	Il t'a vu ou vue
شافه	Chaf-hou	Il l'a vu
شافها	Chaf-ha	Il l'a vue
شافنا	Chaf-na	Il nous a vus ou vues
شافكم	Chaf-koum	Il vous a vus ou vues
شافهم	Chaf-houm	Il les a vus ou vues

TABLE 9. *Pronoms affixes compléments d'un verbe*

On retrouve aussi dans le dialecte algérien le verbe venu, et le mot عند "çnd" qui signifie «chez» suivre par des affixes.

عندي	and-i	chez moi ou j'ai
عندك	and-ek	chez toi ou tu as
عنده	and-hou	chez lui ou il a
عندها	and-ha	chez elle ou elle a
عندنا	and-na	chez nous ou nous avons
عندكم	and-koum	chez vous ou vous avez
عندهم	and-houm	chez eux, elles ou elle ont

TABLE 10. conjugaison du verbe avoir en présent

la conjugaison de verbe avoir en imparfait c'est la conjugaison en présent en ajoutant le mot كان (kan) (il était) au début. Concernent le futur s'exprime aussi par le présent et on ajoute le mot يكون (ikoun) (il sera) au début [75].

la conjugaison de verbe être en présent dans le Dalg s'exprime par syllabe (را) (ra) et on ajoute des affixes.

راني	ra-ni	je suis
راك	ra-k	tu es
راكي	ra-ki	tu es (fém)
راه	ra-hou ou ra-h	il est
راها / راهي	ra-ha ou ra-hi	elle est
رانا	ra-na	nous sommes
راكم	ra-koum	vous êtes
راهم	ra-houm	ils, elles sont

TABLE 11. conjugaison du verbe être en présent

كنت	kount	j'étais
كنت	kount	tu étais
كنتي	kounti	tu étais (fém)
كان	kan	il était
كانت	kanét ou elle	était
كنا	kounna	nous étions
كنتوا	kountou	vous étiez
كانوا	kanou	ils, elles étaient

TABLE 12. conjugaison du verbe être en imparfait

نكون	nkoun	je serai
تكون	tkoun	tu seras
تكوني	tkouni	tu seras (fém)
يكون	ikoun	il sera
تكون	tkoun ou elle	sera
نكونوا	nkounou	nous serons
تكونوا	tkounou	vous serez
يكونوا	ikounou	ils, elles seront

TABLE 13. conjugaison du verbe être en futur

1.4.2.2 Catégorie grammaticale

Concernant les catégories grammaticales, nous pouvons les diviser en deux parties : la section des nombres et la section des genres.

En ce qui concerne la section des nombres, dans le dialecte algérien, les suffixes et préfixes numériques de la langue arabe sont complètement supprimés et remplacés par des mots numériques. Par exemple, dans la langue arabe, on dit (بنتين), c'est-à-dire "deux filles". Dans le dialecte algérien, le mot s'écrit au pluriel avec l'ajout d'un mot

numérique avant ou après, tel que (زوج بنات), signifiant "deux filles". Ainsi, nous notons qu'il n'y a pas d'utilisation de suffixes et de préfixes indiquant le nombre dans le dialecte algérien, ceux-ci étant remplacés par des mots quantitatifs tels que (Bzaf, Zoj, Tlata ... etc.) (بزاف, ثلاثة, زوج)

Concernant la section du genre, il y a des mots que l'on peut associer au masculin et au féminin sans changer, contrairement à la langue arabe, car ils changent avec ce qui les précède ou les suit. Par exemple, "Beaucoup de filles" (فتيات كثيرات) et "Il y a beaucoup d'hommes" (رجال كثير). On remarque ici que le mot «beaucoup» (كثير) a changé avec le changement du mot en arabe. Dans le dialecte algérien, il ne change pas, par exemple, bzAf bnAt (بزاف بنات) et bzAf drAry (بزاف دراري).

1.4.2.3 Les nouveaux clitiques

Dans le dialecte algérien, on retrouve des clitiques nouveaux qui n'existent pas dans la langue arabe standard moderne. Par exemple, lors de l'interrogatoire pour exprimer des phrases interrogatives, il faut vérifier si la phrase contient une particule interrogative ou non [76].

Premièrement, si la phrase contient déjà une particule interrogative telle que (واش), wAš, signifiant "quoi", (وين), wyn, signifiant "où", (وعلاش), wɛlAš, signifiant "pourquoi", ou (شحال), šHAɪ, signifiant "combien". Dans ce cas, rien ne change dans la phrase, par exemple : واش قلت wAš qlt Qu'est-ce que tu as dit?? وين كنت wyn knt où étais-tu? وعلاش ماجيتش wɛlAš mAjytš pourquoi tu n'es pas venu? شحال البيض šHAɪ AlbyD Combien coûtent les œufs? Deuxièmement, s'il n'y a pas de particule interrogative dans la phrase, on ajoute (ما) avant le verbe et le suffixe (ش) après lui, ce qui représente le "ne, pas" en français. Par exemple : (ما كتبتش دوروسك), mA ktbtš drwswk?, ce qui signifie "N'a-t-il pas écrit ses leçons?"

1.4.3 Les caractéristiques lexicales

Dans le dialecte algérien, on trouve quelques mots lexicaux qui ne se retrouvent ni en arabe standard moderne ni en français, ni dans d'autres dialectes. Ces mots sont spécifiques au dialecte algérien, tels que :

- (اروح) : signifiant "viens ici".
- (عفايس ou عفسة) : signifiant "un truc".
- (برك) : signifiant "seulement".
- (كروسة) : signifiant "voiture".

Il y a également d'autres mots dérivés de la langue française et adaptés au dialecte algérien, tels que :

- (ديپايني) : signifiant "aidez-moi".
- (دوبلاني) : signifiant "il m'a doublé".
- (امپريميلي) : signifiant "imprime moi".

1.4.4 Les caractéristiques syntaxiques

La syntaxe du dialecte algérien diffère des caractéristiques syntaxiques de l'arabe standard moderne (ASM). Par exemple, nous constatons que l'arabe standard moderne dispose de treize pronoms nominaux, tandis que dans le dialecte algérien, nous en trouvons seulement huit, ce qui signifie que certains pronoms ont été supprimés.

Les pronoms duelle (deuxième personne duelle - اتما) et (troisième personne duelle - هما) ont été supprimés. De même, les pronoms de deuxième personne du pluriel féminin (انتن) et de troisième personne du pluriel féminin (هن) ont été supprimés. Quant aux pronoms de première personne duelle (نحن) et de première personne du pluriel (نحن), ils ont été remplacés par le mot (احنا).

أنا	ana	moi
انت	enta	toi ou tu (masc)
انتي	enti	toi ou tu (fém)
هو	houwa	lui, il
هي	hiya	elle
احنا	ah'na	nous
انتم	entoum	vous
هم	houm	eux, ils, elles

TABLE 14. *pronoms sujet de DAlg*

1.5 Conclusion

Dans ce premier chapitre, nous avons exploré les caractéristiques générales de la langue arabe, en mettant l'accent sur son évolution vers le dialecte algérien. Nous avons examiné l'histoire complexe du dialecte algérien, en soulignant ses origines, son développement historique et les influences linguistiques qui l'ont façonné au fil des siècles.

La linguistique du dialecte algérien a été étudiée en détail, mettant en lumière ses caractéristiques phonologiques, morphologiques, lexicales et syntaxiques uniques. Nous avons examiné comment le dialecte algérien se démarque de la langue arabe standard moderne à travers des différences significatives dans sa phonologie, sa morphologie et son lexique.

En parallèle, nous avons également examiné la langue arabe elle-même, en explorant son histoire et en mettant en évidence ses caractéristiques phonologiques, morphologiques, lexicales et syntaxiques. Cette comparaison nous a permis de mieux comprendre les différences et les similitudes entre la langue arabe standard et le dialecte algérien.

En conclusion, ce chapitre nous a offert un aperçu complet des caractéristiques générales de la langue arabe et de son évolution vers le dialecte algérien. Les distinctions linguistiques entre ces deux variétés de la langue nous rappellent l'importance de l'étude approfondie de chaque dialecte pour une compréhension complète de la richesse linguistique de l'Algérie.

Dans les chapitres suivants, nous présenterons les ressources textuelles créées pour le dialecte algérien, ce qui contribuera à enrichir notre compréhension de cette variété linguistique unique et à mettre en lumière son importance dans le paysage linguistique algérien.

Chapitre

2

Exploration et construction du corpus Liée au dialecte algérien

Table des matières partielles

3.1	Introduction	68
3.2	Travaux préalables sur la segmentation des textes arabes et dialectaux	69
3.2.1	Travaux réalisés au niveau de la langue arabe	69
3.2.2	Travaux réalisés au niveau des dialectes arabes	72
3.2.3	Travaux réalisés au niveau de dialecte algérien	73
3.3	Défis de la segmentation automatique de texte	74
3.3.1	Difficultés de segmentation dans les textes arabe	75
3.3.2	Difficultés de segmentation dans le dialecte algérien	80
3.4	Notre méthode proposée pour la segmentation de texte algérien en phrases	84
3.4.1	Contexte de la méthode	84
3.4.2	Inspiration de travaux antérieurs	85
3.4.3	Incorporation des règles du dialecte tunisien	86
3.4.4	Présentation de la méthode proposée	87
3.4.5	Méthode détaillée de segmentation du dialecte algérien	88
3.5	Réalisation et évaluation	92

3.6	Analyse des erreurs de segmentation	95
3.6.1	Contexte de l'erreur	96
3.6.2	Exemple d'erreur	96
3.6.3	Actions correctives	96
3.7	Conclusion	97

2.1 Introduction

L'une des composantes cruciales du traitement automatique des langues naturelles (TALN) est le corpus linguistique, que nous pouvons définir comme une vaste collection de textes originaux, c'est-à-dire des échantillons de la langue produits à partir de situations de communication réelles [8].

Contrairement aux langues française et anglaise, les corpus de la langue arabe sont peu nombreux en comparaison avec d'autres langues. Cependant, ces dernières années, il y a eu une augmentation significative, en raison de l'intérêt croissant des chercheurs pour la langue arabe. Malgré cette augmentation, les corpus des dialectes arabes restent limités, en particulier le dialecte algérien. Cela s'explique par le peu d'utilisation du dialecte algérien dans les conférences, à la télévision ou à la radio, et par la prédominance de l'arabe standard et du français. De plus, la diversité des dialectes algériens et leurs variations régionales rendent la collecte de corpus scientifiques dans ce dialecte particulièrement difficile, contribuant ainsi à la rareté des corpus linguistiques.

La situation linguistique en Algérie est extrêmement complexe, en raison des influences ethniques, géographiques et de l'occupation coloniale. Ces facteurs ont créé un environnement social complexe, caractérisé par une diversité linguistique considérable qui influence le comportement linguistique quotidien des locuteurs en Algérie [22].

Ainsi, dans le cadre de ce chapitre, nous avons été amenés à construire notre propre corpus d'étude ALGDC, à la fois écrit et parlé, en expliquant les méthodes fondamentales utilisées pour sa construction. Nous détaillerons également les étapes de son nettoyage et de sa normalisation.

2.2 Les corpus existants

Rappelons que la compréhension automatique est étroitement liée à la disponibilité de corpus linguistiques robustes. Cependant, comme nous l'avons signalé dans le chapitre 1, les ressources linguistiques arabes et ses dialectes, notamment le dialecte algérien, sont très rares, voire même indisponibles. Néanmoins, la plupart des corpus sont collectés à partir de sites Internet et principalement à partir des médias sociaux et via des techniques d'exploration Web (*web crawling*) et l'utilisation d'interfaces de programmation d'application (API) pour des plateformes comme Facebook et Twitter (renommé récemment X) [6] qui ne sont généralement pas suffisants pour garantir une représentation linguistique exhaustive et précise des variétés régionales et informelles des dialectes arabes, y compris le dialecte algérien.

Ainsi, il existe deux méthodes pour la construction des corpus. La première méthode est manuelle, où les données sont collectées manuellement par des chercheurs. La seconde méthode est automatique, où des algorithmes sont utilisés pour collecter les données. En ce qui concerne le traitement automatique du langage, il se divise en deux catégories principales :

Le traitement automatique de la langue ou du dialecte écrit : Ce premier type implique la collecte de textes à partir de sites Web et des médias sociaux, comme mentionné précédemment. Ces corpus sont ensuite nettoyés et normalisés en suivant les conventions d'écriture propres à chaque langue ou dialecte.

le traitement automatique de la langue ou du dialecte parlé : Dans ce cas, le corpus de données se compose d'enregistrements audio des locuteurs, ou d'enregistrements

audio extraits d’Internet, de la radio ou de la télévision. Ces enregistrements sont ensuite convertis en textes à travers un processus appelé transcription. Cette transcription peut être réalisée manuellement ou à l’aide d’outils spécifiques pour faciliter le processus. Une fois transformés en textes, ces enregistrements audio sont traités de la même manière que les textes écrits, en passant par des étapes de nettoyage et de standardisation.

2.2.1 Les corpus existants pour l’arabe et ses dialectes

Un vaste éventail d’efforts de recherche a été déployé pour constituer des corpus de la langue arabe et de ses nombreux dialectes. Parmi les initiatives les plus notables figure le corpus **AWATIF**, présenté par Abdul-Mageed et al. [42]. Ce corpus imposant comprend pas moins de 10,723 phrases de la langue arabe, minutieusement extraites du web, avec une emphase particulière sur des sources comme l’encyclopédie Wikipédia.

De manière complémentaire, **TAGREED** [38] se distingue en rassemblant 3,015 tweets en arabe, offrant ainsi une perspective contemporaine sur l’évolution du langage sur les médias sociaux.

En explorant les spécificités dialectales, plusieurs corpus se démarquent également. **TSAC**[25], par exemple, englobe 17,060 publications sur Facebook, fournissant un instantané précieux des expressions linguistiques dans le dialecte tunisien sur les plateformes numériques.

STAC [37] va au-delà de l’écrit pour capturer l’essence du dialecte tunisien parlé spontanément, avec un impressionnant enregistrement audio de 4 heures et 50 minutes provenant de diverses chaînes de télévision et de radio tunisiennes.

Un autre corpus d’une nature différente est **TuDiCoI** [40]. Cet ensemble de données orales offre un aperçu captivant des dialogues quotidiens dans une gare, enregistrés minutieusement à l’aide d’un enregistreur numérique au guichet. Bien que 104 heures aient été enregistrées, seulement 52 heures ont été transcrites, soulignant la complexité

du processus de transcription et l'effort requis pour rendre ces données accessibles à la recherche.

Enfin, dans un contexte plus large et multilingue, le corpus **PADIC** [35] se démarque en intégrant 5 dialectes, dont l'algérien. Il englobe un impressionnant total de 6,412 phrases pour chaque dialecte, offrant ainsi une riche variété de données pour l'analyse comparative des dialectes arabes.

Élargissant encore la perspective, le corpus multilingue **MADAR** [16] embrasse 25 dialectes provenant de diverses villes arabes, en plus de l'ASM, du français et de l'anglais. Pour le dialecte algérien, ce corpus fournit une base solide de 2,000 phrases, constituant ainsi une ressource inestimable pour l'étude des variations linguistiques et culturelles dans la région.

Ces corpus, avec leur diversité et leur profondeur, constituent un pilier fondamental pour la recherche en traitement automatique des langues, ouvrant la voie à une compréhension approfondie de la richesse linguistique de l'arabe et de ses multiples dialectes.

2.2.2 Les corpus existants pour le dialecte algérien

Le dialecte algérien, avec sa diversité régionale et linguistique, a suscité un intérêt croissant parmi les chercheurs en linguistique et en traitement automatique des langues. Plusieurs corpus significatifs ont été créés pour documenter cette richesse linguistique. Parmi eux, **ALGSAD** [51] se distingue en tant que corpus vocal, rassemblant les voix de 300 locuteurs issus de 11 régions algériennes différentes. Chacun de ces locuteurs a récité un ensemble de 200 phrases, résultant en un total impressionnant de 1,080 lectures enregistrées.

Un autre corpus vocal notable est **KALAM'DZ** [21], un projet ambitieux qui a capturé huit principaux dialectes arabes algériens. Ce corpus a été construit à partir de matériel diffusé sur les chaînes de télévision et de radio algériennes, ainsi que sur

YouTube. Avec 4,881 locuteurs et plus de 104 heures d'enregistrements audio, il offre une plongée profonde dans les nuances linguistiques du dialecte algérien.

L'**AMCASC** [22] (Corpus Algérien de la Parole Arabe Familiale Moderne) se démarque comme un corpus verbal, comprenant les enregistrements de 735 locuteurs provenant de 20 régions différentes d'Algérie, collectés à travers des appels téléphoniques. Ces enregistrements, couvrant près de 88 heures, offrent un aperçu authentique et varié de l'usage quotidien du dialecte algérien.

Dans le domaine complexe de l'analyse des sentiments, plusieurs corpus écrits ont été élaborés. Le travail de Abdelli et al. [9] a impliqué l'utilisation d'un corpus écrit, rassemblé à partir de diverses sources arabo-algériennes via l'API Facebook. Ils ont également intégré un ensemble de données open source, **LABR** [39], centré sur les critiques de livres, comprenant un total de 49,864 phrases. Un autre effort significatif a été réalisé par mataoui et al. [30], qui a constitué un corpus écrit à partir de commentaires provenant des pages algériennes les plus influentes sur Facebook. Cette collection de données comprenait 7,698 commentaires rédigés dans divers styles, notamment l'arabe en alphabet arabe, en alphabet romain, ainsi que des commentaires en dialecte algérien, écrits dans les deux systèmes d'écriture.

Le travail de Dahou et al. [2] présente **DZ-SMS**, un corpus issu de diverses régions géographiques d'Algérie, représente une précieuse collection de 6 000 messages courts authentiques (SMS) en dialecte algérien, transcrits en caractères latins ou en symboles. La diversité géographique des SMS, provenant des zones du Moyen, de l'Est et de l'Ouest du pays, offre un aperçu riche et varié de la langue. La collecte méticuleuse de ces messages dans 42 départements administratifs et géographiques révèle la diversité linguistique unique de chaque région. Cette ampleur et cette diversité font de ce corpus une ressource exceptionnelle pour les chercheurs en traitement automatique du langage naturel, leur permettant d'explorer en détail les subtilités linguistiques du dialecte algérien dans le contexte contemporain des communications par SMS.

Zergat et al.[4] présente un corpus de parole algérienne à grande échelle, baptisé **"Sawt El-Djazair"**, spécialement créé pour répondre aux besoins de la reconnaissance vocale et de la comparaison vocale forensique. Face au manque de corpus appropriés pour les dialectes arabes algériens, les chercheurs ont collecté des enregistrements de célébrités provenant de diverses régions d'Algérie, ainsi que des participants utilisant des appels cellulaires et des applications de voix sur IP. Ce corpus de près de 50 heures couvre douze sous-dialectes algériens et offre une variété de scénarios et de sujets de parole. Il est désormais mis à disposition de la communauté de recherche, offrant ainsi une ressource précieuse pour les travaux en reconnaissance vocale et comparaison vocale forensique, contribuant ainsi à l'avancement de ces domaines.

L'examen attentif de ces corpus met en lumière la richesse et la complexité du dialecte algérien, offrant ainsi une ressource précieuse pour les chercheurs cherchant à comprendre et à analyser les subtilités linguistiques de cette variante régionale. Cependant, malgré ces recherches, il est important de noter que le dialecte algérien reste relativement pauvre en ressources comparé à d'autres dialectes arabes. Les travaux dans ce domaine restent souvent limités et nécessitent un investissement accru pour exploiter pleinement le potentiel de cette langue dynamique et variée.

2.3 Construction du corpus d'étude

Le dialecte algérien, dépourvu de règles linguistiques standardisées, présente un défi fascinant pour les linguistes et les chercheurs en traitement automatique des langues. Avant d'entamer toute analyse ou de proposer des ressources linguistiques, il est impératif de créer un corpus robuste. Ce corpus jouera un rôle crucial dans le développement ultérieur des outils de traitement automatique spécifiques au dialecte algérien. La collecte des données a été réalisée de manière méticuleuse, combinant trois méthodes différentes et deux genres distincts - parlé et écrit. Plusieurs étapes ont été rigoureusement suivies pour nettoyer, organiser et normaliser les bases de données.

Dans les prochaines sections, nous détaillerons les méthodes de collecte des données, les techniques employées ainsi que les étapes de préparation et de standardisation par lesquelles le corpus est passé. Cela nous permettra de disposer d'un corpus écrit exceptionnellement vaste et normalisé du dialecte algérien, une ressource précieuse pour la recherche linguistique et l'analyse automatisée de cette variante régionale riche et complexe.

2.3.1 Acquisition des données audio pour le corpus algérien

Nos efforts initiaux se sont concentrés sur la collecte de données orales, enregistrées et capturées dans des fichiers audio impliquant plusieurs locuteurs. Ces données sont cruciales dans le domaine du Traitement Automatique des Langues Naturelles (TALN). Il existe deux méthodes principales pour collecter ces données :

La première implique l'enregistrement vocal de locuteurs provenant de diverses catégories sociales, discutant de sujets variés. Ces enregistrements sont réalisés dans des studios spécialisés, ainsi que dans des environnements tels que les rues, les aéroports, et d'autres lieux publics, en utilisant des équipements d'enregistrement spécifiques. Cependant, cette méthode peut s'avérer coûteuse en termes de ressources financières et de temps.

La deuxième méthode, appelée "téléchargement et enregistrement", a été proposée par Waibel et al. [66], comme décrit par Zribi [33]. Elle consiste à effectuer des recherches sur Internet pour trouver des déclarations ou des fichiers audio pertinents. À notre époque, Internet représente une vaste source de ressources accessible à tous les chercheurs. Il contient des centaines de milliards de vidéos, d'images et de textes qui peuvent être extrêmement bénéfiques pour les chercheurs travaillant dans le domaine linguistique.

Cette approche innovante capitalise sur l'énorme base de données disponible en ligne, offrant aux chercheurs des opportunités considérables pour explorer et exploiter ces ressources riches en données linguistiques. Dans les sections suivantes, nous

détaillerons notre méthode de recherche en ligne, expliquant en détail comment nous avons identifié et utilisé les fichiers audio pertinents pour notre corpus du dialecte algérien.

2.3.1.1 Techniques de collecte des données orales

Dans notre démarche, nous avons adopté une approche hybride pour collecter des données et construire un corpus spécialisé pour le dialecte algérien à l'oral. Nous avons combiné les deux méthodes suivantes de manière intégrée.

Initialement, nous avons entrepris de collecter des enregistrements audio auprès de locuteurs variés, représentant diverses catégories sociales, dans des studios privés. L'utilisation d'outils d'enregistrement spécialisés, notamment des microphones magnétiques, a permis de capturer des discours portant sur divers sujets tels que la politique, le sport, la culture et la santé. Notre focalisation s'est particulièrement orientée vers les événements liés aux manifestations algériennes du 22 février 2019. Nous avons veillé à sélectionner des orateurs provenant de différentes régions du pays, de divers âges et sexes. Ces enregistrements sonores ont été supervisés par des experts spécialisés dans le domaine du son pour garantir leur qualité et leur pertinence.

Parallèlement, nous avons mis en œuvre la deuxième méthode, consistant à effectuer des recherches en ligne pour trouver des clips audio en dialecte algérien, que nous avons ensuite téléchargés et traités. Nous avons exploré la plateforme YouTube, une source immense de données visuelles et auditives. Nos recherches ont ciblé des entretiens de presse, des programmes télévisés, des émissions sportives et politiques en dialecte algérien. De plus, nous avons téléchargé des vidéos de YouTubeurs et de créateurs de contenu considérées comme des minutes de données dans le dialecte algérien. Après cette étape de recherche et de téléchargement, nous avons procédé à la conversion des vidéos en clips audio au format MP3. Pour ce faire, nous avons utilisé l'outil Format Factory¹, un logiciel gratuit de conversion de formats réputé pour son efficacité. Cette approche combinée nous a permis d'obtenir un ensemble

1. www.pcfreetime.com/formatfactory

de données oral diversifié et représentatif du dialecte algérien, prêt à être intégré dans notre corpus.

2.3.1.2 Transcription des données orales

Après avoir téléchargé les vidéos portant sur divers sujets en dialecte algérien et les avoir converties en fichiers audio, l'étape suivante consiste en la transcription. La transcription implique la représentation symbolique des données audio sous forme de texte, facilitant ainsi une analyse approfondie et un codage ultérieur [67]. Cette étape revêt une importance cruciale dans la construction d'un corpus, car elle impacte significativement les étapes ultérieures du traitement des données.

L'un des défis majeurs dans la transcription, en particulier pour les dialectes arabes, réside dans l'absence de règles orthographiques standardisées. En raison de la nature orale des dialectes, dépourvus de grammaire formelle, les variations de prononciation peuvent entraîner des ambiguïtés lors de la transcription. Il est souvent difficile de déterminer le format d'écriture idéal pour un mot particulier.

Pour surmonter ces défis, l'utilisation d'outils de transcription est indispensable. Plusieurs outils sont disponibles, parmi lesquels Praat², Transcriber³, Anvil⁴, Elan⁵, etc. Chacun de ces outils possède ses propres caractéristiques et avantages.

Dans notre travail, nous avons opté pour Praat [15], l'analyseur de son. Ce choix s'est justifié par plusieurs raisons : tout d'abord, Praat est un logiciel libre, offrant une solution économique pour notre projet. De plus, il a été utilisé avec succès dans notre laboratoire pour transcrire le corpus STAC [37], démontrant son efficacité. En outre, Praat se distingue par son interface graphique conviviale, simplifiant le processus de transcription. Il offre diverses fonctionnalités, telles que la transcription, l'analyse phonétique, le marquage et la segmentation des fichiers, facilitant ainsi le travail des experts supervisant le processus.

2. <http://www.fon.hum.uva.nl/praat/>

3. <http://trans.sourceforge.net/en/presentation.php>

4. <http://www.anvil-software.de/>

5. <http://icar.univ-lyon2.fr/projets/corinte/confection/elan.htm>

2.3.1.3 Défis de la transcription dans le dialecte algérien parlés

La transcription consiste à convertir chaque son en son équivalent écrit, un processus particulièrement complexe pour les dialectes parlés qui ne disposent pas de règles d'écriture standardisées. Dans le cas des dialectes arabes, et spécifiquement pour le dialecte algérien, l'absence de conventions orthographiques spécifiques engendre des difficultés à la fois liées à l'oralité du discours et aux règles grammaticales de l'écrit.

Dans notre travail de transcription du dialecte algérien, nous avons été confrontés à divers défis. La multiplicité de mots ayant le même sens, due à la diversité des dialectes en Algérie et aux variations de prononciation, complique la tâche. De plus, certaines variations dans l'écriture de certains mots, considérées comme étrangères au dialecte algérien, ont également posé des problèmes de transcription.

Naviguer à travers ces défis exigeait une compréhension approfondie des nuances linguistiques et une expertise dans les particularités du dialecte algérien. Malgré ces obstacles, Nous avons persévéré pour garantir une transcription précise et fidèle des enregistrements oraux, allant jusqu'à prendre en considération la structure des phrases ainsi que la transcription des mots tels que nous les entendons. Nous prenons en compte les hésitations, répétitions, reprises, etc. Cette transcription a contribué à la construction d'un corpus linguistique authentique et représentatif du dialecte algérien parlé.

A. Défis spécifiques liés à l'oralité

Les difficultés liées à la transcription manuelle dans le contexte des dialectes oraux, notamment le dialecte algérien, sont multiples et souvent complexes. Ces défis surgissent principalement des conditions d'enregistrement, parfois peu idéales, et des caractéristiques inhérentes à l'oralité des discours.

Lors de la transcription manuelle, nous sommes confrontés à des enregistrements de voix parfois altérés par des bruits de fond, des échos ou des musiques d'ambiance

provenant d'émissions télévisées ou de podcasts de YouTubers. Ces interférences rendent le processus de transcription manuelle ardu, rendant la transcription automatique impossible [34]. De plus, l'absence d'outils automatiques adaptés au dialecte algérien a compliqué davantage cette tâche.

Au-delà des conditions d'enregistrement, d'autres défis surgissent lors de la transcription. Parfois, le signal sonore n'est pas clairement décodé, et il est difficile de discerner la prononciation exacte des mots en raison de divers facteurs tels que la vitesse de parole du locuteur, le chevauchement des sons et les hésitations. De plus, les interruptions, les mots incomplets et les chevauchements de discours, surtout dans des contextes spontanés, sont des phénomènes courants. Chacun de ces éléments requiert des corrections minutieuses pour attribuer un sens à chaque mot prononcé [33].

Ces défis nous ont poussés à développer des techniques spécifiques pour surmonter ces obstacles et garantir une transcription précise et fidèle malgré la complexité du dialecte algérien parlé.

B. Complexités spécifiques du dialecte algérien

Le dialecte algérien, comme de nombreux dialectes arabes, présente des défis uniques en raison de ses différences linguistiques avec l'arabe standard moderne (ASM). Contrairement à l'ASM, qui possède des règles grammaticales et morphologiques bien établies et peut être étudié et traité automatiquement [64], les dialectes arabes, y compris le dialecte algérien, sont complexes en raison de leurs propres règles linguistiques et de leur richesse lexicale.

L'ASM est une langue complète avec des règles convenues, mais ces règles ne peuvent pas être directement appliquées aux dialectes arabes en raison de leurs différences intrinsèques. Le dialecte algérien se distingue par son vaste lexique, comprenant des mots empruntés à d'autres langues telles que l'arabe, le français, etc. En tentant de transcrire le dialecte algérien en utilisant les règles de l'arabe standard, on se heurte à des mots qui n'ont pas d'équivalent dans l'arabe standard

(براف، قبلا، فاع) et à l'utilisation fréquente de mots français dans le discours algérien (parce que, mais, droite, gauche, etc.). Cette diversité linguistique rend impossible la transcription précise du dialecte algérien en utilisant uniquement les conventions de l'arabe standard. Ces particularités ont exigé des approches de transcription spécifiques et nuancées pour capturer fidèlement la richesse linguistique du dialecte algérien.

2.3.1.4 Convention d'écriture pour le dialecte algérien

Au cours du processus de transcription manuelle du dialecte algérien, nous sommes souvent confrontés à des mots de sens similaire mais avec plusieurs variantes d'écriture. Dans de tels cas, il est crucial de suivre une règle orthographique uniforme pour assurer la cohérence dans l'écriture des mots. Pour ce faire, nous adoptons la Convention d'Orthographe des Dialectes Arabes (CODA) proposée par Habash et al.[43]. La CODA est un accord cohérent visant à standardiser l'écriture des dialectes arabes en utilisant l'alphabet arabe. Son objectif est d'établir un cadre unifié pour l'écriture de tous les dialectes arabes, facilitant ainsi le développement d'applications de Traitement Automatique des Langues (TAL). Chaque mot dans le dialecte algérien est associé à une forme orthographique unique, représentant sa phonologie et sa morphologie.

Le dialecte algérien possède sa propre carte CODA, spécifique à ses particularités phonologiques et morphologiques. Comme mentionné dans la section 1.3.1 du premier chapitre, l'Algérie comprend une diversité de dialectes. **Pour garantir l'uniformité, nous avons choisi d'utiliser le dialecte algérois pratiqué dans les médias, notamment à Alger, la capitale d'Algérie. Ce choix s'explique par sa large diffusion et sa compréhension étendue dans toutes les régions du pays.**

Saadane et Habash ont présenté un accord d'écriture spécifique pour le dialecte algérien, décrivant la carte CODA du dialecte[36]. Cet accord comprend diverses règles d'écriture normalisées, prenant en compte à la fois les particularités du dialecte

algérien et les règles de la langue arabe, assurant ainsi une transcription cohérente et précise du dialecte dans notre corpus.

A. Règles phonémiques :

Le dialecte algérien présente des particularités phonétiques distinctes par rapport à l'arabe moderne standard. Ces règles phonémiques sont cruciales pour comprendre la prononciation et l'orthographe spécifiques utilisées dans ce dialecte.

- **Voyelles longues :** En arabe moderne, il n'y a pas de voyelle / : e /, qui s'écrit ay ou iA, selon son homologue arabe. La solution adoptée pour le dialecte algérien est d'utiliser les mots apparentés avec aA dans l'ASM. Cette transformation est la plus appropriée pour préserver la forme non diacritique de base du mot, par exemple : دار daAr /da :r/ 'tourner' et diAr /de :r/ 'fais'.
- **Raccourcissement des voyelles :** Les voyelles longues du dialecte algérien sont généralement écrites sous forme longue, sauf dans quelques exceptions, elles sont raccourcies lors de l'ajout d'un suffixe ou d'une critique, même si elles sont écrites longues. Par exemple, تقول لهم tqawl lhm /tqullhum/ "tu leur dis" et pas تقلهم tqulhm. Le raccourcissement des voyelles peut également être observé dans les mots comportant deux voyelles longues. En phonétique, en DA, même si deux voyelles longues sont écrites, un seul mot est autorisé, c'est-à-dire que chaque mot phonémique ne doit accentuer qu'une syllabe. Par exemple, صائمين SaAymiyn "Jeûne" et pas صيمين saAymiyn.

Les exceptions phono-lexicales : Ces particularités phonétiques et lexicales spécifiques au dialecte algérien reflètent l'interconnexion complexe entre l'histoire linguistique et les influences culturelles. Comprendre ces exceptions est essentiel pour une interprétation précise et contextuelle de la langue dans le contexte algérien.

- **Le Qaf algérien :** La lettre qāf ق est utilisée dans le dialecte algérien pour représenter les consonnes suivantes : / q /, / g /, / k /. Le tableau 15 ci-dessous représente des exemples de prononciation d'exclusion de / g /

CODA		Pronunciation	Francais
بقرة	baqrah	/bagra/	vache
قاناتيك	qaAnaAtiyk	/ga :na :ti :k/	aussi
قاوري	qiAwriy	/ge :wriy/	étranger

TABLE 15. Exemples de prononciation exceptionnelle du dialecte algérien

- **Prononciation multiple des consonnes :** Dans le dialecte algérien, nous utilisons la forme ASM pour écrire des consonnes ayant des prononciations multiples. Si une forme ASM correspondante existe, elle doit être choisie en fonction de la proximité avec la consonne précédente. Le dialecte algérien semble avoir plus de prononciations que l'ASM, où l'ASM est orthographié de la même manière que l'ALG.
- **La lettre Hamza :** Dans le dialecte algérien, la prononciation de la lettre hamza est différente de la prononciation en ASM, c'est-à-dire que l'arrêt glottal sera épelée phonétiquement, mais quand Hamza sera prononcé en ALG, on appliquera les mêmes règles d'orthographe ASM. De plus, le phonème d'arrêt de la glotte qui apparaît dans de nombreux mots ASM a disparu dans le dialecte algérien, par exemple le mot كاس kAs 'verre' (pas comme ASM كَّاس kaAs) et aussi le mot بئر biyr 'Puit' (pas comme ASM بَّئر biyr). De plus, les mots commençant par Hamzat Alif n'apparaîtront pas dans ALG CODA, comme dans le mot الارض AlAarD /larD/ 'terre' (non لرض larD).
- **L'article défini :** Si le mot contient l'article Al (ال), vous devez faire la distinction entre les lettres du soleil et de la lune. Pour les lettres du soleil, le «L» est silencieux et les lettres suivantes sont doublées (gémation) en prononciation et en écriture par exemple النهار Alnnhar 'jour' (non انهار AnnhAr. Contrairement avec les lettres de la lune, le «A» n'est pas prononcé, le «L» de l'article est prononcé et la lettre qui suit n'est pas doublée, ni en prononciation ni en écriture. comme le cas dans le mot القمر Alqmar 'la lune' (non لقمر lqmar).
- **N de construction numérique :**

Le CODA algérien ajoute le phonème / n / après certains chiffres dans les cas de construction, par exemple سطاشن طابلة sTA\$n TaAbla «16 tables» alors que le nombre 16 est prononcé seul سطاش sTA\$. Cette exception est valable pour les formes Nombre Construit avec un nombre compris entre 11 et 19 précédant un nom au singulier .

B. Règles morphologiques :

Dans le cadre des règles morphologiques du dialecte algérien, plusieurs aspects distinctifs se démarquent. Ces règles dictent la structure des mots, les préfixes et les suffixes utilisés, ainsi que les modifications apportées aux racines lexicales. Comprendre ces règles est essentiel pour saisir la complexité morphologique unique de ce dialecte.

- **Clitiques attachés :** Le dialecte ALG, comme beaucoup d'autres dialectes, utilise presque tous les clitiques attachés à MSA, l'article défini + ال Al +, la future particule proclitique + ح Ha + (exprimée à l'est de l'Algérie comme la ville d'Annaba), la conjonction de coordination + و w +, la particule de négation enclitique + ش + + S. De plus, ALG utilise les nouvelles formes réduites de clitiques attachées du MSA, par exemple + ع c +, + م m +, + ه h +, + ف f +.

Le tableau 16 illustre quelques exemples de ces clitiques où nous considérons le mot وکلیناهاکم wikliynaAhaAlkum « et nous avons mangé votre nourriture ».

Enclitiques			suffixes	radical	Proclitiques
کم	ل	ها	نا	کلي	و
kum	l	haA	naA	Kliy	wi

TABLE 16. Tokenisation du mot وکلیناهاکم

- **Les Clitiques de séparation :** La règle d'orthographe pour les enclitiques d'objet indirect et le proclitique de négation ما mA est conservée dans la carte ALG

CODA. Cette carte met une séparation en utilisant un espace entre la particule de négation et l'objet indirect, par exemple, *ما جاب لكمش* mA jAb lkumš / ma + jab + lkum + š / 'il ne vous a pas donné'.

- **Exceptions lexicales :** Le CODA algérien contient une liste de mots du dialecte algérien qui ont une orthographe spécifique. Cette orthographe spécifique peut être incompatible avec la carte de CODA présentée ci-dessus et peut être orthographiée couramment de différentes manières. Ces exceptions incluent par exemple :
 - Les démonstratifs *هذوك* (h*wk) (et non *هاذوكة* (hA*wkah) «ces», *هكذا* (hk*A) «comme ça» (pas *هاكذا* (hkda), ou *هكدا* (hAkdA) ou *هاكدا* (haAkdaA)).
 - La préposition «Je sais» est exprimée par la phrase *على بالي* (ElY bAly) (pas *عمبالي* (EmbAly) , ou *عن بالي* (En bAly), ou *علبالي* (ElbAly)

2.3.1.5 Évaluation de la transcription manuelle

Afin d'assurer l'efficacité de notre processus de transcription, la compréhension entre les deux experts du domaine (Rédaction Créative en Dialecte Algérien) qui ont effectué la transcription manuelle doit être testée. Pour ce faire, une mesure du degré d'accord est nécessaire afin de déterminer la compatibilité de ces jugements d'experts du domaine.

Pour évaluer le taux d'accord, nous utilisons le coefficient kappa, tel que défini par Cohen [73]. Ce coefficient kappa, noté K, représente le pourcentage maximal de correction d'accord à partir d'effets accidentels mineurs et varie entre -1 et 1 [70].

L'accord observé entre les deux experts du domaine est calculé selon l'équation suivante :

$$k = \frac{P_a - P_e}{1 - P_e} \quad (2.1)$$

Où :

- P0 représente le taux de consensus combiné des deux experts.

- Pe représente le taux de concordance aléatoire, calculé en divisant la somme des produits de chaque type de norme de chaque expert par le carré du nombre total de réponses à recueillir.
- r est le nombre de modalités.
- n est le vecteur échelle d'observation du tableau de contingence avec r modalités.
- N est le nombre total de réponses à recueillir.

$$Pe = \frac{1}{N^2} \sum_{i=1}^r (n_i \cdot n_i) \quad (2.2)$$

Le degré d'accord est évalué en fonction de la valeur de Kappa. Plus cette valeur augmente, plus le degré d'accord entre les experts du domaine est élevé. L'échelle présentée par Landis et Koch [72] est utilisée pour classer le degré d'accord en fonction de la valeur de Kappa

Accord	Kappa
Excellent accord	$0.81 \leq K \leq 1.00$
Bon accord	$0.61 \leq K \leq 0.80$
Accord modéré	$0.41 \leq K \leq 0.60$
Accord médiocre	$0.21 \leq K \leq 0.40$
Mauvais accord	$0.00 \leq K \leq 0.20$
Très mauvais accord	$-1 \leq K \leq 0.00$

TABLE 17. échelle de classement de degré d'accord selon la valeur de Kappa [72]

Dans notre étude, nous avons obtenu une valeur moyenne de Kappa de 0,64 pour l'accord entre les experts du domaine en transcription avant l'application de la convention CODA pour le dialecte algérien. Cette valeur est considérée comme un bon accord selon l'échelle de classification de Landis et Koch [72]. Cette variation peut être attribuée au bruit présent dans les enregistrements ainsi qu'à la difficulté d'entendre certains mots par les experts, ou à des différences dans leur capacité auditive.

Après l'application de la convention CODA, la moyenne de Kappa a augmenté à 0,71, ce qui est également considéré comme un bon accord selon la même échelle[72].

La moyenne du coefficient Kappa pour les transcriptions, qu'elles aient été effectuées sans l'utilisation de CODA-ALG ou conformément à cette convention, a été

calculée en prenant la moyenne des Kappas obtenus lors de l'application de l'équation 2.1 à la transcription de l'ensemble des données parlées, réalisée par les experts, en tenant compte de leur accord sur l'application des règles CODA.

2.3.2 Acquisition des données écrites pour le corpus algérien

Comme précédemment mentionné 2.3, deux méthodes distinctes sont utilisées pour collecter les données. La première méthode consiste à enregistrer directement les locuteurs à l'aide de divers moyens et supports spécifiques, soit dans des endroits dédiés à cet effet, soit en extrayant des enregistrements d'Internet, la principale source de données, grâce à ce que nous appelons la méthode de "téléchargement et enregistrement". Cela signifie télécharger soit le fichier audio directement, soit la vidéo puis la convertir en fichier audio.

En ce qui concerne la deuxième méthode, qui concerne la collecte de données écrites, elle repose principalement sur l'extraction d'informations à partir du Web et des sites de réseaux sociaux. Différentes méthodes sont utilisées pour cette extraction, suivies d'une phase de nettoyage des données. Au final, nous obtenons un corpus de type écrit, uniformisé grâce à un accord orthographique pour le dialecte en question.

Dans les prochains paragraphes, nous explorerons les différentes méthodes de collecte de données écrites pour le dialecte algérien.

2.3.2.1 Techniques de collecte des données écrites

La collecte de données écrites était auparavant une tâche quasi impossible. Cependant, avec l'avènement de la technologie, Internet est devenu la meilleure et la plus simple source de données, offrant également un gain de temps considérable [18].

Avec l'essor des médias sociaux au cours de la dernière décennie, les chercheurs ont commencé à explorer ces plateformes comme source complémentaire aux médias traditionnels. Le terme "médias sociaux" fait référence aux "applications basées sur Internet qui reposent sur les principes idéologiques et technologiques du Web 2.0".

Le Web 2.0 signifie que le contenu et les applications ne sont plus seulement créés et publiés par des individus, mais sont constamment modifiés par tous les utilisateurs de manière collaborative et participative [50]. En raison de cette définition large, les médias sociaux ont diverses applications.

Notre méthode de collecte de données écrites se décompose en deux étapes distinctes : une étape dépendante principalement des médias sociaux, et une autre étape basée sur l'extraction de corpus de données préexistantes du dialecte algérien.

A. Collecte des données des réseaux sociaux

Notre méthode de collecte de données écrites se divise en deux étapes. Tout d'abord, nous avons exploité les réseaux sociaux en tant que source de données. Twitter (X), bien qu'employé par les Algériens (avec plus de 625 000 utilisateurs), n'est pas utilisé de manière excessive. Malgré cela, nous avons réussi à extraire quelques tweets rédigés en dialecte algérien en arabe. Twitter (X) revêt une importance particulière pour plusieurs raisons : son format de message court oblige les rédacteurs à adopter un style concis tout en leur permettant d'inclure des liens vers des ressources originales. De plus, il facilite la publication et la diffusion d'informations. Twitter (X) offre également un accès aisé à son contenu, ce qui permet de cibler les sujets et de collecter les données à rechercher [46].

Twitter (X) abrite de nombreux commentaires dans diverses langues et domaines. Pour effectuer nos recherches, nous avons utilisé l'API⁶ suggérée par Twitter (X), qui repose sur l'utilisation de mots-clés pour la recherche.

Nous avons également exploité Facebook, l'un des sites de réseau social les plus fréquentés. Selon le rapport numérique de l'Algérie⁷ pour l'année 2021, 97,9 % des internautes algériens accèdent à Facebook via divers appareils mobiles et tablettes numériques. D'après ce même rapport, le nombre d'utilisateurs de ce réseau social

6. L'API Twitter (X) est une interface de programmation utilisée pour se connecter aux données Twitter (X) de manière automatisée.

7. <https://datareportal.com/reports/digital-2021-algeria>

en Algérie atteignait en janvier 2021 plus de 23 millions, soit 71,8 % de la population de plus de 13 ans. Nos recherches sur Facebook se sont également basées sur l'API Facebook Graph⁸ [17]. Nous avons extrait des commentaires rédigés en arabe dans des dialectes algériens, en ciblant des sujets liés au mouvement national, aux élections présidentielles de décembre 2019 et à la pandémie de la Covid-19.

De manière similaire aux réseaux sociaux, les forums et les pages algériens constituent une ressource importante pour les données écrites en dialecte algérien. On y trouve fréquemment des histoires, des blagues et des énigmes dans ce dialecte, notamment sur des forums populaires comme le Djelfa Forum, classé comme l'un des forums les plus consultés par les Algériens en 2019, selon les statistiques d'Amazon Alexa⁹.

B. Collecte de données à partir de corpus existant

La deuxième étape de notre collecte de données écrites implique l'extraction d'informations à partir de corpus dédiés au dialecte algérien. Comme précédemment mentionné dans les recherches sur le dialecte algérien, plusieurs corpus écrits en algérien sont disponibles. Nous avons sélectionné deux de ces corpus pour enrichir notre ensemble de données.

Le premier corpus provient de PADIC¹⁰ de Meftouh et al. [35], un corpus multilingue. Le deuxième corpus est MADAR¹¹ de Bouamor et al. [16]. Notre choix s'est porté sur ces deux sources en raison de leur accessibilité et de leur disponibilité en ligne.

Notre travail sur ces corpus ne se limite pas à leur téléchargement. Nous avons entrepris l'extraction spécifique du dialecte algérien. Ces deux bases de données

8. Facebook Graph API

9. <https://www.amazon.com/>

10. https://www.researchgate.net/publication/316463706_PADIC_A_Parallel_Arabic_Dialect_Corpus

11. <https://camel.abudhabi.nyu.edu/madar/CorpusViewer>

comprennent diverses langues et dialectes, et notre première tâche a été de filtrer uniquement les données algériennes. Ensuite, un processus de prétraitement a été appliqué, un sujet que nous aborderons ultérieurement. Ce prétraitement est essentiel, car il influence directement les résultats de nos analyses.

Après avoir collecté les données écrites en deux étapes - d'abord à partir de sites de réseaux sociaux tels que Facebook et Twitter (X), puis en extrayant des données arabes du dialecte algérien à partir des deux corpus multilingues existants - nous passons maintenant à l'étape cruciale du prétraitement. Cette phase joue un rôle central dans l'impact de nos résultats d'analyse.

2.3.2.2 Étape de pré-traitement

L'étape de pré-traitement consiste à nettoyer les données écrites et à les préparer pour une utilisation ultérieure dans le processus de traitement.

Dans un premier temps, nous avons appliqué les règles de la convention CODA pour le dialecte algérien, comme expliqué en détail dans la section 2.3.1.4, à l'ensemble de nos données écrites. Cette opération a été menée avec l'aide de deux experts du domaine (Rédaction Créative en Dialecte Algérien). Le coefficient Kappa pour l'accord entre ces experts s'est élevé à 0,78, une valeur considérée comme solide selon les critères de Landis et Koch [72], indiquant une bonne compatibilité entre les experts.

Dans un autre volet du pré-traitement, nous avons éliminé les phrases courtes composées de moins de 4 mots, car elles n'apportent pas une réelle valeur dans les résultats du traitement. De plus, tous les mots incompréhensibles ont été retirés, afin de garantir un flux de traitement fluide. Comme indiqué précédemment dans la section 2.3.2.1, la plupart de nos données proviennent d'Internet et des sites de réseaux sociaux, nécessitant un filtrage rigoureux.

Au cours de cette étape, nous avons également normalisé l'écriture des mots, ce que nous appelons l'étape de normalisation. Sur les médias sociaux, les utilisateurs ont la liberté d'écrire les mots de différentes manières. Cette variabilité dans l'orthographe

a été standardisée pour assurer la cohérence des données. Naturellement, cette normalisation a entraîné une réduction du nombre de phrases et de mots dans notre corpus.

2.4 Caractéristiques de l'ALGDC : un Aperçu approfondi

Dans cette section, nous allons présenter le Corpus ALGDC, également connu sous le nom de "Corpus of the Algerian Dialect". Cette ressource linguistique a été conçue dans le but de faciliter le traitement automatique du dialecte algérien. ALGDC se distingue par sa taille imposante, sa diversité, et la variété des sujets qu'il aborde. Il se caractérise également par la pluralité des locuteurs représentés, ce qui en fait un outil adapté à une multitude d'applications de traitement automatique pour le dialecte arabe algérien.

Ce corpus est unique en raison de son inclusion de données parlées et écrites, ce qui élargit considérablement sa portée et sa pertinence. De plus, il a été enrichi grâce à l'intégration d'autres corpus, renforçant ainsi sa richesse linguistique et culturelle. Ce qui distingue particulièrement ALGDC est le processus minutieux de purification manuelle auquel il a été soumis. Des experts du domaine ont collaboré pour éliminer toute impureté et garantir la qualité optimale des données.

Le processus de Construction d'ALGDC a débuté avec une collection initiale de plus de 50 000 phrases en dialecte algérien. À travers des étapes rigoureuses de prétraitement, comprenant l'élimination des phrases courtes de moins de quatre mots ainsi que la correction des erreurs de transcription et d'enregistrement audio, le corpus a été affiné pour aboutir à 35 000 phrases prêtes pour le traitement ultérieur. Cette sélection minutieuse garantit la qualité et la fiabilité des données présentes dans ALGDC.

2.4.1 Exploration des données parlées

Dans cette section, nous décrivons en détail la méthodologie que nous avons suivie pour la collecte des données parlées en dialecte algérien. Notre approche s'est basée sur deux méthodes distinctes : l'enregistrement direct et l'extraction audio à partir de vidéos. Ces méthodes ont permis d'obtenir un total impressionnant de 117 heures et 38 minutes de conversations en dialecte algérien, impliquant différents locuteurs des deux sexes et couvrant une variété de sujets.

2.4.1.1 Méthode d'enregistrement direct

Dans cette phase, nous avons rassemblé des données parlées auprès de 200 locuteurs de différents sexes, dont l'âge variait de 20 à 40 ans. Les sujets de conversation englobaient divers domaines tels que la culture, le sport, la santé, etc. Cependant, il est important de noter que le nombre d'orateurs était relativement limité. Cette limitation résulte du manque de bénévoles ainsi que des coûts élevés associés à cette méthode, incluant l'achat d'équipement spécialisé, la location de studios d'enregistrement et les frais des intervenants. Malgré ces défis, nous avons réussi à enregistrer un total de 39 heures et 50 minutes, avec une moyenne de 20 minutes par orateur. Par la suite, ces enregistrements ont été transcrits, donnant initialement lieu à 31 476 mots, soit environ 5 246 phrases en dialecte algérien écrites en arabe. La Figure 2 illustre la répartition des locuteurs par sexe, et le Tableau 18 présente la répartition des locuteurs par tranche d'âge.

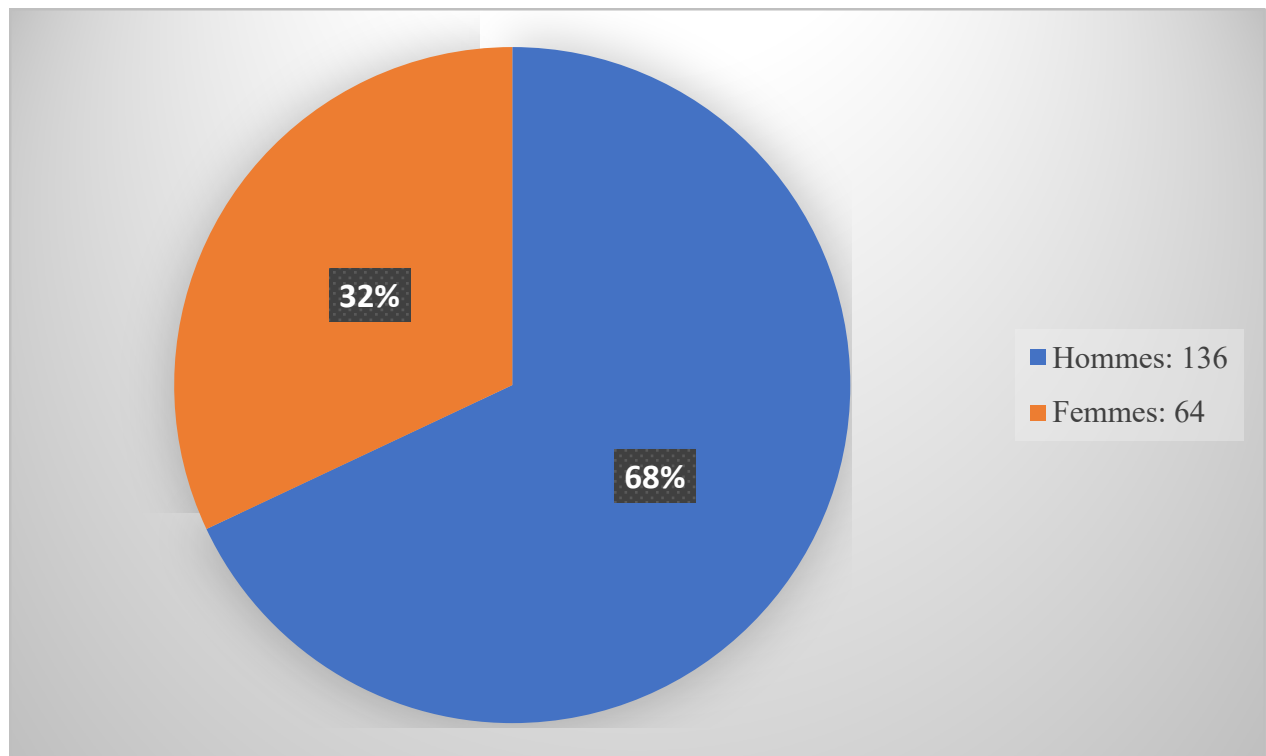


FIGURE 2. Répartition de locuteurs par sexe

Age	20-25	25-30	30-35	35-40
Nb locuteurs	45	72	57	26

TABLE 18. Répartition des Locuteurs par Âge

En outre, les conversations enregistrées ont été classées en fonction de leur domaine thématique, comme indiqué dans le Tableau 21. Ces domaines comprennent la politique, la santé, le sport, la culture, les sujets religieux et divers autres thèmes. Le Tableau 21 présente également la durée totale d'enregistrement pour chaque domaine.

Domaine	Politique	Santé	Sport	Culture	Réligieux	Autre	Total
Durée	06 :57 :13	07 :52 :48	07 :26 :50	06 :13 :36	03 :36 :52	07 :42 :47	39 :50 :06

TABLE 19. Répartition d'enregistrement par domaine

2.4.1.2 Méthode de téléchargement

Dans la méthodologie de téléchargement que nous avons précédemment décrite, nous avons collecté des clips vidéo en dialecte algérien provenant de diverses émissions de télévision, programmes télévisés et actualités diffusées sur différentes chaînes algériennes publiques et privées. Nous avons également inclus des contenus d'influenceurs présents sur des plateformes telles que YouTube, Instagram et Facebook. Ces vidéos ont été converties en fichiers audio pour un traitement ultérieur.

Réseau sociaux	YouTube	Facebook	Instagram	Tik-Tok	Twitter (X)
Nombre	117	52	31	17	14

TABLE 20. *répartition des vidéos téléchargées en fonction de réseaux sociaux*

Au total, ce processus nous a permis de rassembler 78 heures d'enregistrements audio, impliquant 395 locuteurs. Chaque segment de locuteur a été extrait dans son propre clip audio, facilitant ainsi le processus de transcription. Nous avons téléchargé environ 230 vidéos au cours de cette étape.

Suite à la transcription, nous avons obtenu un total de 82 896 mots, avec une moyenne de 8 mots par phrase, soit environ 10 362 phrases au total.

2.4.2 Exploration des données écrites

Dans la section précédente (2.3.2), nous avons abordé en détail la collecte de données écrites, en mettant l'accent sur diverses méthodes telles que l'extraction à partir de plateformes de réseaux sociaux et l'utilisation de corpus existants pour enrichir notre ensemble de données écrites.

Au terme de cette étape, nous avons rassemblé environ 36 958 phrases du dialecte algérien rédigées en arabe, provenant de divers groupes de la société et couvrant un

large éventail de sujets. Cela équivaut à un total de 221 748 mots, avec une moyenne de 6 mots par phrase.

2.4.2.1 Méthode des réseaux sociaux

À ce stade, nous avons collecté des données écrites en dialecte algérien sur Internet, exploitant une source de données ouverte immense. Nous nous sommes particulièrement concentrés sur les sites de réseaux sociaux, qui nous ont permis d'effectuer des recherches ciblées grâce à la fonction de recherche par mot-clé. Cette approche nous a permis d'acquérir des données écrites sur des sujets d'actualité au sein de la société algérienne, tels que le mouvement populaire, les élections, l'épidémie de Corona, la Coupe d'Afrique, etc. Cette diversité de sujets inclut des domaines variés tels que le sport, la politique, la culture, la santé, etc.

Source	Journaux en ligne	Facebook	Twitter	Forums	Autre	Total
Nb de phrases	2200	7392	9721	5031	4202	28546

TABLE 21. Répartition d'enregistrement par domaine

À l'issue de cette phase, nous avons rassemblé 28 546 phrases en dialecte algérien, écrites en arabe et provenant de diverses sources telles que Facebook, Twitter (X), les forums en ligne, les journaux en ligne, etc. En moyenne, chaque phrase contenait 7 mots, totalisant ainsi 199 822 mots.

2.4.2.2 Méthode d'intégration des corpus existants

À ce stade, nous avons utilisé les corpus existants en dialecte algérien pour enrichir notre propre corpus. Après avoir examiné tous les ensembles de données disponibles écrits dans le dialecte algérien, nous avons sélectionné deux corpus : PADIC et MADAR. Notre choix s'est basé sur leur accessibilité gratuite pour tous les chercheurs ainsi que sur leur pureté, exempts de mots étrangers au dialecte algérien. Le corpus

PADIC, selon l'auteur [35], contient 38 707 mots, répartis en 6 412 phrases, avec une moyenne de 7 mots par phrase. En ce qui concerne le deuxième corpus, MADAR, selon [16], il comprend 2 000 phrases, totalisant environ 13 000 mots, avec une moyenne de 6,5 mots par phrase.

En combinant ces deux corpus, nous avons enrichi notre ensemble de données avec 8 412 phrases authentiques, totalisant plus de 51 707 mots écrits en arabe dans le dialecte algérien, avec une moyenne de 6 mots par phrase.

Corpus	Nb phrases	Nb mots
PADIC	6412	38707
MADAR	2000	13000
Totale	8412	51707

TABLE 22. *description des corpus prise*

2.4.3 Récapitulatif et synthèse : découverte de l'ALGDC

Dans cette section, nous consolidons notre exploration du Corpus ALGDC, un vaste ensemble de données en dialecte algérien écrit en arabe, créé dans le but de l'étudier et de le rendre accessible à l'automatisation dans le domaine du traitement automatique des dialectes arabes, relevant du domaine du traitement automatique des langues naturelles.

Notre recherche a initialement visé les données parlées, impliquant la collecte d'un corpus oral à partir de diverses sources. Ces données orales ont ensuite été transcrites en format écrit. Nous avons utilisé deux méthodes pour obtenir ces données parlées : la méthode d'enregistrement direct, impliquant l'enregistrement de locuteurs bénévoles dans des studios équipés, et la méthode de téléchargement, où nous avons extrait des clips audio de vidéos téléchargées depuis les réseaux sociaux et autres plateformes similaires. Les deux méthodes, directes ou indirectes, ont impliqué une étape de transcription supervisée par des experts du domaine.

Cependant, en raison de diverses contraintes telles que des limitations financières, temporelles, et un manque de bénévoles dû à d'autres engagements, notre

méthodologie de recherche s'est orientée vers les données écrites. Nous avons exploité Internet comme principale source de données écrites, en téléchargeant tout contenu écrit en arabe en dialecte algérien à partir de divers sites, en particulier les réseaux sociaux largement utilisés en Algérie. Cette phase, appelée normalisation, a impliqué le nettoyage des données de toute impureté, ajout ou incohérence linguistique, et la standardisation des mots pour faciliter leur traitement ultérieur. En complément, notre ensemble de données a été enrichi par l'incorporation de deux corpus existants disponibles sur Internet, choisis pour leur accessibilité et leur gratuité. Ces corpus ont également subi un processus de normalisation pour garantir leur cohérence avec le reste des données.

Au terme de ce processus, notre ensemble de données a atteint un total de 52 566 phrases et 388 447 mots en dialecte algérien écrit en arabe. Après un nettoyage minutieux, nous avons obtenu 29 600 phrases, avec une moyenne d'environ 6 mots par phrase, ainsi que 183 520 mots, formant ainsi la base solide de notre Corpus ALGDC. Pour une représentation graphique de ce processus, veuillez vous référer à la Figure 3. Ce corpus, résultat d'un travail rigoureux de collecte, de transcription et de normalisation, sert de pierre angulaire à notre recherche sur le traitement automatique du dialecte algérien.

	Nb phrases	Nb mot
Les données parlé		
Phase d'enregistrement	5246	31476
Phase de téléchargement	10362	82896
Les données écrite		
Phase de réseaux sociaux	28546	199822
Phrase de corpus existants	8412	51707
Corpus ALGDC avant nettoyage	52566	365901
Corpus ALGDC après nettoyage	29600	183520

TABLE 23. *description des corpus ALGDC*

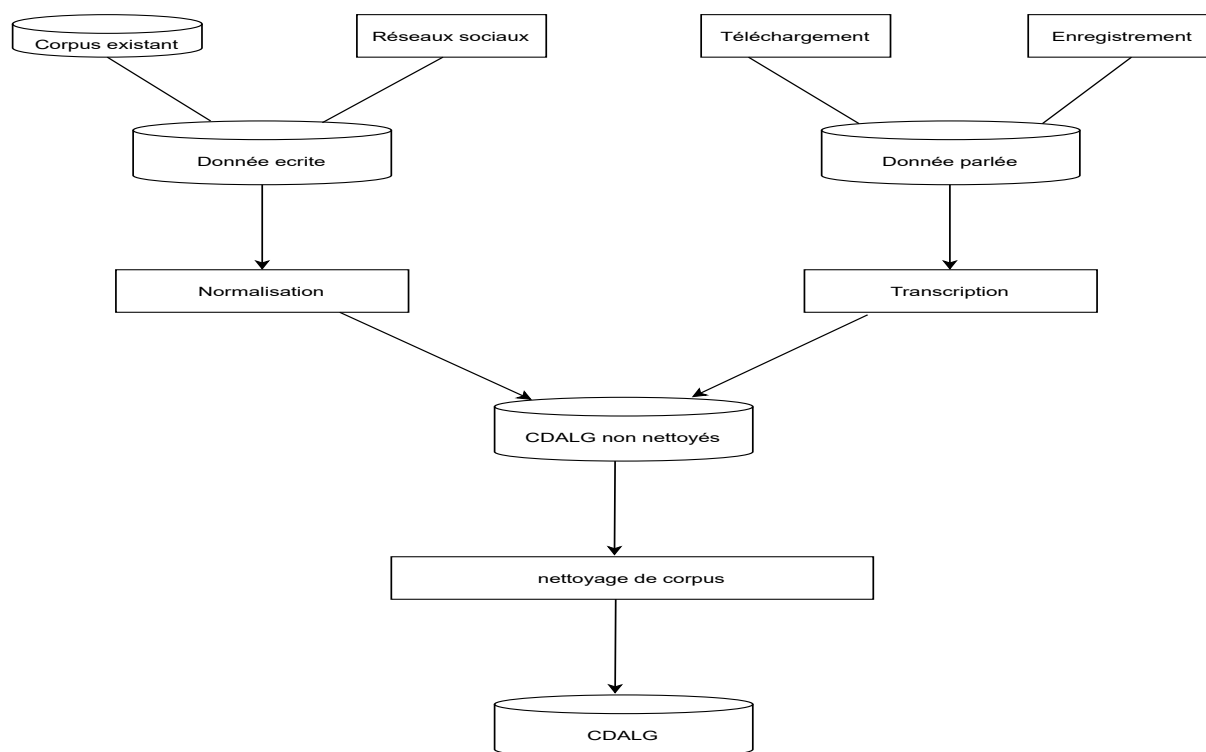


FIGURE 3. Hiérarchie pour la construction de ALGDC

2.5 Évaluation comparative des corpus du dialecte algérien

Dans cette nouvelle section, nous procéderons à une évaluation comparative approfondie de notre corpus ALGDC avec les autres corpus du dialecte algérien disponibles dans le domaine de la recherche en traitement automatique des langues. Nous avons élaboré un tableau comparatif (voir Tableau 24) qui met en lumière divers aspects tels que le type de données collectées, le langage utilisé, la taille par heure d'enregistrement, et la taille par phrase moyenne.

Notre corpus ALGDC se distingue par sa diversité, comprenant à la fois des données parlées et écrites. Les enregistrements parlés ont une taille moyenne de 6 207 mots par heure, tandis que la taille moyenne par phrase est de 6 mots. Ces données offrent une représentation authentique du dialecte algérien dans divers contextes. En comparaison, les autres corpus varient en termes de type de données, de langage utilisé

et de taille moyenne par phrase, les plaçant à différents niveaux d'utilité pour les applications de traitement automatique des langues.

Cette évaluation comparative démontre la richesse et la variété de l'ALGDC en tant que ressource linguistique pour la recherche et le développement dans le domaine du traitement automatique du dialecte algérien.

	Type de données	Langue	Taille (Heures)	Taille (Phrases)
ALGSAD	Parlé	Dialecte algérien	100 H	/
KLAMDZ	Parlé	Dialecte algérien	104 H	/
AMCASC	Parlé	Dialecte algérien	88 H	/
Sawt El-Djazaïr	Parlé	Dialecte algérien	50 H	/
TWIFIL	Écrit	Dialecte algérien	/	9000 Tweets
PADIC	Écrit	Multilingue	/	6412
MADAR	Écrit	Multilingue	/	2000
ALGDC	Écrit/Parlé	Dialecte algérien	117 H	29600

TABLE 24. Comparaison entre les Corpus Existant et ALGCD pour le Dialecte Algérien

2.6 Conclusion

Ce chapitre a exploré en profondeur la construction du ALGDC. Nous avons souligné l'importance d'un corpus bien préparé dans le domaine du traitement automatique des langues, en mettant en lumière sa signification et son rôle central dans la garantie de résultats précis. En examinant les corpus et les recherches existants sur la langue arabe et ses dialectes, nous avons identifié les corpus arabes les plus notables. Parmi ces corpus, nous avons examiné spécifiquement les ensembles de données algériens les plus importants, tels que PADIC et MADAR.

Dans les sections précédente, nous avons détaillé notre méthodologie de collecte de données, qui s'est articulée autour de deux étapes clés. La première étape consistait à collecter des données orales, soit par enregistrement direct auprès des locuteurs, soit par téléchargement depuis Internet. La deuxième étape impliquait la collecte de données écrites en langue arabe, à partir des médias sociaux ou de bases de données préexistantes.

Alors que nous exposons les différents modes de collecte de données, nous avons mis en évidence l'importance cruciale de l'étape de transcription, où les données parlées étaient converties en données écrites en suivant la convention spécifique du dialecte algérien, la convention CODA. Nous avons également abordé les étapes de normalisation, de nettoyage et de pré-traitement, soulignant ainsi notre engagement envers la qualité des données.

Enfin, nous avons présenté une vue d'ensemble détaillée de notre corpus à chaque étape du processus. Cela comprenait le nombre total de phrases et de mots, ainsi que la diversité des sujets et des sources utilisées pour la collecte des données, tant orales qu'écrites. Cette diversité a été illustrée à travers un diagramme clair, mettant en lumière la richesse de notre corpus ALGDC.

Ce chapitre a posé les bases solides pour notre recherche en fournissant un aperçu complet de la construction de notre corpus. Dans les chapitres suivants, nous exploiterons ce corpus dans le cadre de notre étude sur le traitement automatique des dialectes arabes. Nous utiliserons les outils proposés pour analyser et comprendre en profondeur le dialecte algérien, tirant parti de la richesse et de la diversité de l'ALGDC pour produire des résultats significatifs et approfondir notre compréhension de cette langue dynamique.

Chapitre

3

Segmentation automatique de textes algériens en arabe dialectal

Table des matières partielles

4.1	Introduction	99
4.2	Exploration approfondie de l'analyse syntaxique	100
4.2.1	Avancements dans l'analyse syntaxique de l'arabe standard . .	101
4.2.2	Avancements dans l'analyse syntaxique des dialectes arabes . .	102
4.2.3	Innovations dans l'analyse syntaxique du dialecte algérien . . .	103
4.3	Défis et complexités : L'analyse syntaxique du dialecte algérien	103
4.3.1	Défis de l'analyse syntaxique	104
4.3.2	Complexités syntaxiques du dialecte algérien	106
4.3.3	Avantages de l'analyse syntaxique du dialecte algérien	108
4.3.4	Analyse syntaxique : arabe standard vs. dialecte algérien . . .	109
4.4	Analyse syntaxique du dialecte algérien : une méthode proposée	111
4.4.1	Collection et préparation des données	112
4.4.2	Création du Treebank dédié au dialecte algérien	113
4.4.3	Architecture de la méthode	114
4.5	Réalisation et évaluation	120
4.5.1	Réalisation de Stanford-DZ	120

4.5.2	Évaluation de Stanford-DZ	121
4.5.3	Discussion des résultats	123
4.5.4	Perspectives	124
4.6	Conclusion	125

3.1 Introduction

Il est vrai que l'arabe, qu'il soit classique, moderne ou dialectal, fait partie des langues difficiles à traiter automatiquement pour diverses et nombreuses raisons (voir Chapitre 1). Cependant, la nécessité d'outils de TAL en arabe est nécessaire. Dans le domaine du traitement automatique des langues naturelles, la segmentation de texte en phrase est l'une des étapes les plus cruciales dans l'analyse automatique de textes arabes. la nécessité de cette étape intervient dans divers domaines tels que le résumé automatique, la traduction automatique, la reconnaissance automatique des langues et des dialectes, entre autres. L'importance fondamentale de cette étape réside dans son pouvoir d'extraire l'information recherchée, délimitant ainsi les frontières des phrases et des paragraphes [3].

Analyser un texte sans passer par l'étape de segmentation engendre inévitablement des résultats médiocres. De plus, des segmentations incorrectes peuvent entraîner des erreurs dans le traitement automatique des textes, soulignant ainsi l'importance critique d'une segmentation précise basée sur des règles linguistiques spécifiques.

Dans ce chapitre, nous explorerons la détection des frontières textuelles en vue de leur utilisation dans le traitement automatisé. Nous examinerons les défis inhérents à la segmentation des textes arabes, en mettant l'accent sur les particularités des dialectes. Notre attention sera particulièrement portée sur le dialecte algérien, mettant en lumière les difficultés spécifiques auxquelles il est confronté. Nous aborderons également les méthodes employées pour la détection des frontières textuelles dans le contexte complexe du dialecte algérien. Ce chapitre servira de fondement essentiel

pour la compréhension approfondie des enjeux liés à la segmentation linguistique dans ce dialecte particulier.

3.2 Travaux préalables sur la segmentation des textes arabes et dialectaux

Avant d’explorer les défis spécifiques liés à la segmentation des textes en dialecte algérien, il est pertinent d’examiner les travaux antérieurs menés dans le domaine de la segmentation des textes arabes en général. Cette revue offre un aperçu des avancées significatives et des stratégies innovantes développées pour aborder les nuances linguistiques complexes présentes à la fois dans l’arabe standard et les dialectes régionaux, y compris le dialecte algérien.

Au cours des dernières décennies, les chercheurs se sont efforcés de surmonter les obstacles posés par la richesse morphologique, l’absence d’espaces, les différences dialectales et d’autres caractéristiques uniques de la langue arabe et de ses dialectes. Les travaux réalisés ont englobé divers aspects de la segmentation, y compris l’exploration des méthodes basées sur des règles, l’application de techniques d’apprentissage automatique, et l’intégration de connaissances linguistiques approfondies pour améliorer la précision de la segmentation.

3.2.1 Travaux réalisés au niveau de la langue arabe

Dans le domaine du traitement automatique des langues, la segmentation des textes représente une étape cruciale pour diverses applications, notamment l’analyse de texte, l’extraction d’informations et la création de résumés automatiques. L’identification précise des segments contenant les informations recherchées est essentielle. Une analyse de texte débutant sans une segmentation préalable en phrases peut conduire à des résultats peu fiables. De même, l’utilisation d’un segmenteur inapproprié peut entraîner l’accumulation d’erreurs lors du traitement automatique du

texte [45]. Il est impératif de suivre un ensemble de règles et de normes pour effectuer ce processus, ces règles variant selon la langue traitée. Cependant, la segmentation des textes n'est pas une tâche facile, en particulier lorsqu'il s'agit de langues hybrides et de dialectes ne suivant pas des règles linguistiques spécifiques.

Parmi les défis majeurs auxquels la langue arabe standard est confrontée, on peut citer l'absence de majuscules et de signes de ponctuation, l'ambiguïté phonétique des mots, l'ambiguïté structurelle et l'agglutination. Malgré l'importance de cette étape dans l'analyse automatique du texte, il est étonnant de constater qu'il existe peu d'applications de segmentation pour l'arabe. Parmi les quelques travaux réalisés dans ce domaine, nous pouvons citer les quatre plus importants que nous connaissons.

- A) belguith et al ont proposé le système STAR est un segmenteur de texte arabe basé sur l'analyse contextuelle des signes de ponctuation et de certaines particules [63]. Il est spécifiquement conçu pour les textes arabes non vocalisés, en se basant principalement sur les signes de ponctuation et les conjonctions de coordination. Ce système prend en entrée un texte écrit en arabe au format TXT et produit en sortie un texte segmenté en phrases et paragraphes, en se basant sur un corpus de livres sur l'éducation tunisienne. Cette méthode est inspirée des travaux de [69], qui ont introduit des indicateurs linguistiques de déclencheurs, appelés indicateurs, ainsi que des index supplémentaires liés à ces indicateurs, accompagnés d'un ensemble de règles. Ces règles se divisent en trois types de déclencheurs : les signes de ponctuation, les particules et les mots de liaison. Avant de segmenter les textes arabes selon la méthode proposée par belghith et al [63], il est essentiel d'étudier le contexte droit et gauche de chaque mot ou particule jouant le rôle de séparateur. Après évaluation sur deux corpus, l'un composé de livres et l'autre d'articles de journaux, les mesures de rappel et de précision pour le premier corpus sont respectivement de 88,26% et 80,65%, tandis que pour le second corpus, elles sont de 75,81% et 65,66%, indiquant que les articles de presse contiennent des erreurs typographiques. Toutefois, il

est important de noter que cette méthode se limite à l'arabe standard moderne (MSA).

- B) Un autre travail, mentionné par Mouelhi et Zoubeir [57], propose une méthode de segmentation basée sur le modèle d'analyse des mots arabes, en se concentrant sur l'articulation et le vocabulaire. Cette méthode distingue trois types de segmentation : la segmentation lexicale, qui divise le texte en mots ou tokens (unités lexicales); la segmentation morphologique, qui divise le texte en petites unités distinctes (morphèmes); et enfin, la segmentation syntaxique, qui divise le texte en unités supérieures aux mots (syntagmes, propositions, etc.). Le système AraSeg est un segmenteur basé sur un dictionnaire où les données sont stockées dans la mémoire du logiciel, et la segmentation se fait en interrogeant le lexique généré. Cette technique repose sur l'analyse lexicale des mots arabes et comprend trois étapes :
 - i) l'étape de séparation, qui isole les constituants immédiats du mot à vérifier
 - ii) la vérification des correspondances pour ne conserver que les couples compatibles, et enfin.
 - iii) l'étape d'interrogation du lexique généré. Après évaluation, le taux moyen d'erreur est de 11,28% et le taux d'efficacité est de 88,72%.
- C) Un autre travail mené par keskes et al [44] traite de la segmentation des textes arabes et décrit une méthode basée sur des règles pour segmenter les textes arabes en clauses. La méthode proposée repose sur une analyse approfondie de nombreux indices lexicaux et de ponctuation. Deux types différents de corpus ont été analysés : des articles de journaux et des manuels scolaires du primaire. Les auteurs proposent un algorithme de segmentation en trois étapes : d'abord, ils utilisent uniquement la ponctuation, puis ils s'appuient uniquement sur l'index lexical, et enfin, ils utilisent à la fois la typologie et l'index de vocabulaire. Les résultats ont été comparés à un ensemble de données segmentées manuellement par des experts. L'évaluation de la méthode a été réalisée sur les deux corpus. Le taux de rappel et le taux de précision sont respectivement de 85% et 86%

pour le premier corpus de manuels scolaires. Pour le deuxième corpus composé d'articles de journaux, les taux de rappel et de précision sont de 67% et 69% respectivement.

3.2.2 Travaux réalisés au niveau des dialectes arabes

Dans le domaine du traitement du langage naturel, la segmentation des mots en leurs éléments constitutifs représente une étape cruciale. Pour la segmentation textuelle du dialecte tunisien (DT), Zribi et al [32] proposent trois stratégies distinctes pour reconnaître les limites des phrases en arabe tunisien. Dans ce travail, ils ont exploré une approche basée sur des règles, une approche statistique et une approche hybride. Ces différentes stratégies s'appuient sur deux signaux prosodiques simples ainsi que sur des indices linguistiques. La stratégie proposée a donné des résultats prometteurs, ouvrant ainsi des perspectives intéressantes pour la segmentation des dialectes arabes.

Par ailleurs, Samih et al [27] démontrent comment un segmenteur peut être entraîné à l'aide de réseaux neuronaux en utilisant seulement 350 tweets étiquetés, sans recourir à la normalisation ni à l'utilisation de caractéristiques lexicales ou de ressources linguistiques pour les dialectes arabes. Au niveau des caractères, ils traitent la segmentation comme un problème d'étiquetage de séquence. Leur approche, basée sur ces fondements, se révèle compétitive face aux méthodes de pointe qui dépendent fortement de ressources supplémentaires, comme le démontre une expérience comparative. Ces avancées sont cruciales dans le contexte des dialectes arabes, où les particularités linguistiques ajoutent un niveau de complexité supplémentaire à la segmentation automatique des textes. Ces travaux sont non seulement des contributions significatives à la recherche dans ce domaine, mais ils ouvrent également la voie à de nouvelles méthodes et applications dans le traitement des dialectes arabes.

3.2.3 Travaux réalisés au niveau de dialecte algérien

Le dialecte algérien représente un domaine de recherche unique et peu exploré dans le contexte du traitement automatique du langage naturel. Contrairement à d'autres langues ou dialectes arabes, il existe une pénurie marquée de travaux consacrés spécifiquement à la segmentation automatique du dialecte algérien. Jusqu'à présent, la majorité des efforts dans ce domaine ont impliqué des méthodes de segmentation manuelle, où les experts linguistiques ont été directement impliqués dans le processus de délimitation des unités linguistiques.

Cette dépendance de la segmentation manuelle a entraîné des limitations significatives. Tout d'abord, le processus est intensif en termes de temps et de ressources, nécessitant un investissement considérable en termes de main-d'œuvre qualifiée. De plus, en raison de la complexité du dialecte algérien, caractérisé par ses spécificités phonétiques, morphologiques et lexicales, la segmentation manuelle peut parfois être sujette à des erreurs humaines et à des interprétations subjectives.

L'absence de méthodes de segmentation automatique spécifiques pour le dialecte algérien pose un défi majeur pour la recherche en traitement automatique du langage dans le contexte algérien. Cela limite non seulement le développement d'applications linguistiques précises pour ce dialecte mais complique également l'analyse linguistique approfondie, entravant ainsi le potentiel de recherche dans ce domaine.

Dans le cadre de ce travail, nous nous sommes intéressés à la segmentation automatique du dialecte algérien. En élaborant des méthodes spécifiques et en exploitant les dernières avancées technologiques dans le domaine du traitement automatique du langage, notre objectif est de surmonter les limitations existantes et de contribuer de manière significative à l'avancement de la segmentation du dialecte algérien.

3.3 Défis de la segmentation automatique de texte

Pour entreprendre toute analyse de texte automatisée, il est impératif de délimiter les phrases afin de faciliter la tâche. Ce processus exige le suivi de règles et de normes spécifiques, lesquelles varient selon la langue traitée. La segmentation des textes se révèle complexe, particulièrement lorsqu'elle concerne des langues hybrides ou des dialectes dépourvus de règles linguistiques précises. Dans ces cas, l'improvisation s'avère essentielle, impliquant la création de règles spécifiques pour orienter la segmentation.

Les difficultés liées à la segmentation des textes peuvent être catégorisées en deux parties distinctes : celles relatives aux textes écrits et celles concernant les textes transcrits et spontanés.

Dans les textes écrits, la segmentation se heurte à divers défis, surtout lorsqu'elle repose sur des signes de ponctuation parfois ambigus. Par exemple, l'utilisation du point peut être délicate. Dans des contextes particuliers, ce signe est interprété comme une séparation entre les lettres d'une abréviation (comme dans T.A.L.N, U.S.A). De même, la virgule est présente dans des nombres décimaux (comme 3.14) et dans les liens web (comme <http://www.miracl.rnu.tn/>), sans marquer la fin d'une phrase.

Les majuscules posent également des problèmes. Bien qu'elles soient généralement considérées comme le début d'une nouvelle phrase, ce n'est pas une règle invariable. Par exemple, dans les noms propres, la majuscule est utilisée sans signifier nécessairement le début d'une nouvelle phrase.

La transcription écrite de la parole orale introduit des complexités supplémentaires dues à sa nature spontanée. Ces difficultés diffèrent significativement de celles rencontrées dans la segmentation des textes écrits, principalement en raison de la présence de disfluences. Les discours spontanés impliquent des mots incomplets, des hésitations et des interruptions, ce qui rend la segmentation plus délicate.

Dans le cadre de nos recherches, nous mettrons en lumière les obstacles auxquels les chercheurs sont confrontés lorsqu'ils définissent les frontières des phrases, que ce soit dans le domaine du traitement automatique des langues, comme l'arabe, ou dans celui des dialectes, tels que le dialecte algérien. Ces défis complexes exigent une approche méticuleuse pour garantir une segmentation précise et significative des textes.

3.3.1 Difficultés de segmentation dans les textes arabe

La langue arabe, parlée par plus de 330 millions de personnes en tant que langue maternelle, s'étend sur une vaste zone géographique, de la région du golfe Arabo-Persique à l'est jusqu'à l'océan Atlantique à l'ouest [53]. Profondément structurée, l'arabe se caractérise par une morphologie complexe qui joue un rôle central dans sa construction [68][59].

La segmentation automatique de la langue arabe se heurte à d'énormes difficultés et obstacles, principalement dus aux caractéristiques intrinsèques de cette langue. Nous examinerons ci-dessous certains des défis les plus cruciaux rencontrés lors de l'analyse de cette langue.

3.3.1.1 L'absence des lettres majuscule

Contrairement à l'anglais et à de nombreuses autres langues basées sur l'alphabet latin, l'arabe ne possède pas de signaux orthographiques spécifiques indiquant le début d'une nouvelle phrase. Dans ces langues, la première lettre en majuscule signale qu'un mot ou une séquence de mots constitue une entité nommée ou marque le début d'une phrase. Cependant, en arabe, l'absence de lettres majuscules crée une difficulté majeure pour délimiter les frontières des phrases. Cette absence de signalisation claire complique le processus de segmentation des textes en unités significatives [55].

exemple :

عرفت الجزائر في منتصف سنة ٢٠٢١ ارتفاعا كبيرا في نسبة الإصابات والوفيات من جائحة كورونا
كبلد افريقي تعتبر الجزائر الأكبر من حيث انتشار فيروس كورونا
*À la mi-2021, l'Algérie a connu une augmentation significative du taux de cas et de décès dus
à la pandémie de Corona. En tant que pays d'Afrique, c'est l'un des plus importants en termes
de propagation du virus Corona.*

Dans l'exemple ci-dessus, on peut observer la distinction entre l'écriture latine et l'écriture arabe. Cette dernière ne comporte pas de majuscules, contrairement au français ou à l'anglais, où les majuscules jouent le rôle de balises signalant le début des phrases.

En français et en anglais, l'utilisation des majuscules aide considérablement à la segmentation des textes, car elles marquent généralement le début d'une nouvelle idée ou d'une nouvelle proposition. Cette absence de distinction formelle dans l'écriture arabe rend le processus de segmentation beaucoup plus complexe. En l'absence de ces indicateurs visuels clairs, les outils de traitement automatique du langage arabe doivent s'appuyer sur des techniques plus sophistiquées et des règles linguistiques complexes pour identifier les frontières des phrases de manière précise.

L'un des aspects les plus saillants de l'écriture arabe qui fascine les non-Arabophones est la présence des voyelles dans les mots. Un simple changement dans l'accentuation des voyelles peut totalement transformer le sens d'un mot.

De nos jours, dans l'usage courant de l'arabe moderne, les textes sont généralement écrits sans les voyelles, ce qui crée une ambiguïté. Cela entraîne la perte de plusieurs caractéristiques morphologiques, rendant ainsi complexe le processus de segmentation automatique. En l'absence de voyelles, les mots peuvent avoir plusieurs significations, ce qui complique la tâche des systèmes de traitement automatique du langage en essayant de déterminer précisément les limites des mots et des phrases. Cette ambiguïté vocale constitue l'un des défis majeurs dans la segmentation automatique des textes arabes.

Prenons par exemple le mot كُتِبَ "ktb" Et cela peut signifier des livres "kutub" كُتُب ou signifie aussi il a écrit "ktaba" كُتِب ou aussi "a été écrit" "kut ib" كُتِب

3.3.1.2 L'ambiguïté structurelle

La langue arabe se distingue par sa richesse lexicale et sa capacité à composer des phrases complexes. Une particularité notable réside dans la construction des phrases, souvent composées de nombreux mots. Prenons, par exemple, la phrase suivante qui illustre cette complexité avec ses 78 mots :

كان فيروس كورونا و لا يزال الى حد الساعة من بين اهم الاسباب التي تؤدي بموت اعز الناس الينا و اقربهم الى قلوبنا بسبب غفلتنا عن تدابير السلامة الموصى بها من منظمة الصحة العالمية من التباعد و لبس الكمامات و استعمال المعقم و التي كان لها اثر سلبي على سلامة المواطن و صحتها رغم جهود الدولة المسخرة في اطار مجابهة هذا الوباء و التدابير المسلطة في ارض الواقع و حملة التلقيحات المجانية التي اصبحت في كل اقطار الوطن.

L'arabe offre la flexibilité de structurer des phrases de différentes manières sans altérer le sens. Par exemple, la phrase "L'enfant a mangé une pomme" peut être écrite de trois manières différentes en arabe, simplement en modifiant la position des mots, tout en conservant le sens de la phrase :

أكل الطفل التفاحة / الطفل اكل التفاحة / التفاحة أكلها الطفل

Cette diversité structurelle représente un défi lors de la segmentation automatique des textes arabes, car il est nécessaire de comprendre la construction profonde des phrases pour réaliser une segmentation précise.

3.3.1.3 La ponctuation dans les dialectes arabes

La ponctuation revêt une importance cruciale dans la langue arabe ainsi que dans les dialectes algériens. Elle sert à transmettre le sens, l'intonation et la structure des

phrases. Alors que l'arabe standard utilise divers signes de ponctuation tels que le point, le point d'interrogation, le point d'exclamation, le point-virgule et les guillemets, ces signes sont également présents dans le dialecte algérien. Cependant, leur utilisation peut être plus souple dans les dialectes, variant d'une région à l'autre.

Dans le dialecte algérien, le point (.) est utilisé pour indiquer la fin d'une phrase ou d'une déclaration, tout comme en arabe standard. Le point d'interrogation (?) est employé pour poser des questions, et le point d'exclamation (!) exprime la surprise ou une exclamation passionnée. Le point-virgule (;) marque une pause plus longue qu'une virgule, souvent utilisé dans des phrases complexes.

Il est important de noter que l'utilisation précise des signes de ponctuation dépend du contexte et du style d'écriture. Dans une conversation informelle, les règles peuvent être plus flexibles, tandis que dans un texte littéraire ou formel, le respect des conventions est généralement plus strict. Cette adaptabilité témoigne de la richesse et de la flexibilité de la langue arabe et des dialectes algériens, permettant aux locuteurs de communiquer efficacement et avec précision.

3.3.1.4 L'agglutination

Les particules, qu'elles soient prépositions, conjonctions ou coordonnants, jouent un rôle crucial dans toutes les langues, particulièrement en arabe. Ces éléments sont considérés comme l'épine dorsale de toute phrase, car écrire ou parler sans eux est quasiment impossible. Leur utilisation constitue un art de la rhétorique. Si ces particules sont employées de manière appropriée, elles enrichissent le langage, mais si elles sont mal utilisées, elles peuvent provoquer l'effet inverse, obscurcissant le discours.

Il existe neuf conjonctions de coordination, chacune ayant un rôle spécifique dans la construction des phrases :

- واو العطف (waaw al-'atf) : et, utilisée pour lier des éléments.
- أو (aw) : ou, utilisée pour présenter une alternative.

- ثمَّ (thumma) : puis, indique la séquence dans le temps.
- الفاء (al-faa) : et, souvent utilisée pour exprimer la conséquence ou la justification.
- حتَّى (hatta) : jusqu'à, indique la limite temporelle ou spatiale.
- أمَّ (am) : ou, utilisée dans les questions.
- لا (laa) : non, indique la négation.
- بل (bal) : mais, introduit une opposition.
- لكنَّ (lakin) : cependant, utilisée pour marquer un contraste.

Ces conjonctions assurent la cohésion dans le discours. Elles peuvent également servir à combiner des mots, comme dans l'exemple :

أكل الطفل التفاحة والموزة

Dans la phrase "L'enfant a mangé la pomme et la banane", la conjonction "et" relie les deux objets.

En matière de segmentation de texte, les conjonctions de coordination jouent un rôle crucial en marquant le début et la fin des phrases. Cependant, il est important de noter que ces conjonctions sont attachées au mot suivant, ce qui signifie qu'elles ne sont pas toujours évidentes. Par exemple :

قد يصل وزن الطفل عند الولادة إلى ٤ كغ

Dans cette phrase "Le poids de l'enfant à la naissance peut atteindre 4 kg", la lettre و (waaw) est la première lettre du mot "poids". Il est donc crucial de comprendre le contexte pour segmenter correctement les phrases et éviter toute confusion.

En résumé, la segmentation des textes en arabe ne peut pas uniquement reposer sur les signes de ponctuation en raison de leur rareté dans la langue arabe. Elle dépend principalement de ces mots de liaison et des conjonctions, qui jouent un rôle majeur en clarifiant les limites des phrases. Pour éviter toute erreur, il est essentiel d'analyser attentivement les mots adjacents à ces particules.

3.3.2 Difficultés de segmentation dans le dialecte algérien

Le dialecte algérien, l'une des variantes des dialectes arabes, présente des particularités qui le distinguent des règles strictes de la langue arabe standard. La définition des limites de ses phrases revêt une importance cruciale dans le domaine du Traitement Automatisé du Langage Naturel (TALN). Cependant, cette tâche se révèle ardue en raison des nombreuses complexités propres au dialecte algérien. Ces difficultés émanent à la fois de ses caractéristiques spontanées et de celles spécifiques au dialecte, rendant ainsi le processus de traitement automatisé particulièrement complexe.

3.3.2.1 Complexités de segmentation liées au dialecte algérien

Le dialecte algérien, lorsqu'il est utilisé dans un discours spontané, présente divers obstacles qui rendent la segmentation en phrases distinctes particulièrement complexe. Parmi les défis les plus importants, on peut citer :

- **Des pauses remplies** : Ces mots sans signification, tels que *أيه، عني، أهه، آاذ، أمم، etc.*, sont utilisés pour marquer des pauses dans la parole, souvent lorsqu'on cherche ses mots. Ils peuvent apparaître au début, au milieu ou à la fin d'une phrase, ne fournissant pas de repères fiables pour déterminer les frontières des phrases.
- **Des répétitions** : Dans les conversations spontanées, les locuteurs ont tendance à répéter des mots ou des parties de phrases pour réfléchir ou pour compléter leur discours. Ces répétitions peuvent rendre la segmentation difficile et altérer le sens des phrases.

Exemple 1 :

آلوكان مجاتش - لوكان مجاتش الكرونا راني مقعدتش قاع فالدار

"Si ce n'était pas Corona - Si ce n'était pas Corona, je ne resterais pas à la maison"

lwkAn mjAtš - lwkAn mjAtš AlkrwnA rAny mqçdtš gAç fAldAr

Exemple 2 :

نقول - نقول للجزائريين الكل يروحو يلحقوا علجال الكورونا زادت بزاف

nqwl - nqwl lljzAyrn Alkl yrwHw ylqHwA çljAl AlkwrwnA zAdt bzAf

Je dis - je dis à tous les Algériens d'aller se faire vacciner car le taux de corona est
en augmentation

- **Des révisions :** Parfois, le locuteur modifie involontairement certaines parties de son discours, affectant ainsi la structure de la phrase et son sens.

Exemple :

بزاف عفائس ماهومش ملاح في الشعب - فينا و لازم يتبدلوا

bzAf çfAys mAhwmš mlAH fy Alšçb - fynA w lAzm ytbdlwA

Beaucoup de choses ne sont pas bonnes chez les gens - en nous et nous devons
changer pour le bien.

- **Des reprises :** parfois, nous constatons que le locuteur recourt à l'abandon d'une partie de son discours pour commencer une nouvelle phrase. Où nous trouvons qu'il y a une première phrase inachevée et une autre nouvelle énoncé.

Exemple :

عندنا منتخب قوي بزاف يرجح ... - بلماضي هو لخلانا نولو هكدا

çndnA mntxb qwy bzAf yrbH ... - blmADy hw lxlAnA nwlw hkda

Nous avons une équipe forte qui gagne... - Belmadi est celui qui a fait cette
équipe forte

- **Des hésitations** : Dans le contexte de la conversation et qu'il s'agisse d'un discours spontané, on retrouve souvent de l'hésitation dans les propos du locuteur, et cela affecte le format de la phrase extraite, et cela revient négativement plus tard dans l'étape de traitement automatique.

Exemple :

نقول بلي بوتفليقة ... اه ماندخلش في لعفايس تاع السياسة

ان شاء الله لفيه الخير ربي يقدموا و برك

nqwl bly bwtflyqa ... Ah mAndxl\$ fy lEfAys tAE AlsyAsah An\$A'Allh lfyh
Alxyr rby yqdmwA w brk

Je dis que Bouteflika... Non, je ne m'immisce pas dans les affaires politiques du
pays. Je souhaite de tout mon coeur le meilleur pour l'Algérie

- **Des interruptions** : Le locuteur peut insérer d'autres phrases au milieu de son discours, ce qui peut rendre la phrase initiale incompréhensible.

Exemple :

حنا يا خو وين تروح تلقى براج تحس روحك في حبس كبير أنا الشهرية متكمليش الشهر
وليت نخم نحرق. نطلب من السلطات يرخفوشوية في البراجات راهي سيركيلاسيون كبيرة
HnA yA xw wyn trwH tlqY brAj tHs rwHk fy Hbs kbyr OnA Al\$hr yp mtkmly\$
Al\$hr wlyt nxmm nHrq. nTlb mn AlsITAt yrxfw \$wyA fy AlbrAjAt rAhy
srkylAsywn kbyrp

Nous, mon frère, où vas-tu ? Tu trouves le poste de contrôle de la police. Tu as l'impression d'être dans une grande prison. Mon salaire ne dure pas un mois. J'ai commencé à penser à l'immigration clandestine. Je demande aux autorités de réduire l'inspection. procédures aux contrôles de sécurité car il y a beaucoup d'embouteillages.

- **Des bruis** : Les enregistrements spontanés sont souvent affectés par divers bruits tels que des rires, des applaudissements, des fonds musicaux, des disputes et des chuchotements, qui peuvent altérer la qualité de l'enregistrement et compliquer la segmentation.

Ces défis inhérents au dialecte algérien soulignent la nécessité d'approches de segmentation spécifiques pour traiter avec précision ce type de discours spontané et dynamique.

3.3.2.2 Complexités de segmentation liées au caractère spontané du dialecte algérien

La segmentation du dialecte algérien présente une difficulté particulière en raison de sa nature spontanée et de son absence de standardisation formelle. Contrairement à l'arabe standard, qui bénéficie d'une structure grammaticale plus rigide et est principalement utilisé dans les écrits formels, le dialecte algérien est principalement employé dans les conversations informelles et les interactions quotidiennes.

Le caractère spontané du dialecte algérien se traduit par une fluidité dans la prononciation des mots et des phrases. Les locuteurs natifs ont tendance à fusionner les mots, à omettre ou à modifier certains sons, et à adopter des variations régionales dans la prononciation. Cela rend la tâche de segmentation plus complexe pour les non-natifs ou ceux qui ne sont pas familiers avec les particularités du dialecte.

Une autre difficulté réside dans l'absence de règles strictes concernant la séparation des mots. Contrairement à l'arabe standard, où chaque mot est clairement délimité, dans le dialecte algérien, il peut être difficile de déterminer où un mot commence et où

il se termine. Les locuteurs natifs peuvent fusionner plusieurs mots en un seul son ou séparer un seul mot en plusieurs segments en fonction du contexte et de l'accentuation.

De plus, l'utilisation d'expressions idiomatiques, de dialectes spécifiques à certaines régions et de mots empruntés à d'autres langues, tels que le français, peut rendre la segmentation encore plus délicate. Ces éléments ajoutent une couche supplémentaire de complexité lorsqu'il s'agit de diviser les phrases en unités distinctes.

Il est important de noter que malgré ces défis, la segmentation du dialecte algérien n'est pas impossible. Les locuteurs natifs ont souvent une intuition naturelle pour comprendre où se trouvent les limites des mots, grâce à leur exposition continue à la langue depuis leur enfance. Cependant, pour les apprenants ou les non-natifs, cela peut nécessiter une pratique et une familiarité accrues avec le dialecte pour développer une meilleure compréhension de la segmentation et de la structure du dialecte algérien.

3.4 Notre méthode proposée pour la segmentation de texte algérien en phrases

Après avoir discuté des travaux existants dans la section 3.2, il est important de noter que, à notre connaissance, aucun travail de segmentation automatique n'a été réalisé pour le dialecte algérien. Dans cette section, nous présentons notre méthode pour déterminer les limites des phrases dans le dialecte algérien écrit en arabe. Notre méthode est la première du genre pour ce dialecte, et elle repose principalement sur des règles linguistiques que nous avons adaptées et extraites manuellement. Cette méthode nous permet de segmenter les phrases et de montrer leurs limites.

3.4.1 Contexte de la méthode

Pour cerner les limites des phrases dans le contexte complexe du dialecte algérien, nous avons élaboré une méthode fondée sur des règles linguistiques précises, élaborées

en étroite collaboration avec des experts. L'objectif initial était de répondre à un besoin crucial dans le domaine de la linguistique computationnelle : la segmentation précise des textes algériens. Cette nécessité émergeait d'une lacune évidente dans la recherche existante : l'absence de méthodes spécifiques pour le dialecte algérien. En considérant cette lacune, nous avons plongé dans l'univers des dialectes arabes, en particulier le dialecte algérien et tunisien, afin de concevoir une approche novatrice.

Notre laboratoire MIRACL, pionnier dans l'étude des dialectes arabes, avait déjà réalisé des travaux préliminaires sur la segmentation en arabe et en dialecte tunisien. Forts de cette expertise, nous avons entrepris d'adapter et d'étendre ces connaissances à la complexité spécifique du dialecte algérien. La décision d'inclure des éléments du dialecte tunisien n'était pas arbitraire. Elle découle de la proximité géographique et linguistique entre les deux dialectes, facilitant l'identification de schémas et de structures linguistiques communes.

Cette méthode, élaborée dans un contexte de collaboration interdisciplinaire, a été conçue pour transcender les barrières géographiques et linguistiques. Notre objectif était de créer un outil précis et fiable, capable de segmenter efficacement les phrases dans le dialecte algérien, tout en s'inscrivant dans le continuum des travaux sur les dialectes arabes. En résumé, notre démarche s'enracine dans la reconnaissance d'un besoin spécifique, soutenue par une expertise préexistante et façonnée par l'exploration approfondie des nuances linguistiques entre les dialectes arabes.

3.4.2 Inspiration de travaux antérieurs

Pour élaborer notre méthode proposée pour la segmentation de texte en dialecte algérien, nous nous sommes plongés dans un vaste éventail de travaux antérieurs dans le domaine de la linguistique computationnelle. L'examen attentif de ces recherches a révélé des schémas et des approches fructueuses que nous pourrions intégrer dans notre méthode. Parmi ces sources d'inspiration, l'étude approfondie réalisée par [63] a retenu particulièrement notre attention.

Le travail de Belguith et al [63] a jeté les bases d'une méthode de segmentation pour les textes en arabe. Leur approche, basée sur des règles linguistiques spécifiques, a démontré une précision remarquable dans la délimitation des phrases. Nous avons puisé dans leur méthode pour créer un point de départ solide. Cependant, nous avons pris soin de personnaliser et d'étendre ces principes pour s'adapter aux subtilités particulières du dialecte algérien.

L'inspiration tirée de travaux antérieurs ne se limite pas à un simple copier-coller des méthodes existantes. Au contraire, nous avons procédé à une analyse approfondie de ces approches pour en extraire les éléments pertinents tout en tenant compte des spécificités du dialecte algérien. En fusionnant ces concepts éprouvés avec notre expertise locale, nous avons pu développer une méthode unique et adaptable, capable de résoudre les défis spécifiques de la segmentation dans le contexte linguistique complexe de l'Algérie.

Cette phase d'inspiration et d'adaptation nous a permis de construire sur les réussites antérieures tout en repoussant les limites de la segmentation des textes arabes. Notre approche intègre ainsi les meilleures pratiques de travaux antérieurs tout en les adaptant de manière innovante pour répondre aux besoins spécifiques du dialecte algérien.

3.4.3 Incorporation des règles du dialecte tunisien

L'intégration des règles spécifiques au dialecte tunisien dans notre méthode de segmentation a été une démarche stratégique motivée par plusieurs considérations clés. Tout d'abord, notre laboratoire, fort de son expertise en dialectologie arabe, avait déjà réalisé des avancées significatives dans le domaine de la segmentation en dialecte tunisien. Cette expérience préalable nous a offert une base solide pour explorer les similitudes et les différences entre le dialecte tunisien et le dialecte algérien.

La proximité géographique et linguistique entre le Tunisien et l'Algérien a constitué un facteur déterminant dans notre choix d'intégrer des éléments du dialecte

tunisien. Ces dialectes partagent des racines linguistiques communes, mais présentent également des variations subtiles dans leur structure et leur grammaire. En étudiant ces nuances, nous avons pu identifier des schémas récurrents qui transcendent les frontières nationales.

L'intégration de ces règles spécifiques au dialecte tunisien dans notre méthode ne s'est pas limitée à une simple traduction mécanique. Au contraire, nous avons procédé à une analyse comparative minutieuse, mettant en lumière les particularités du dialecte algérien et tunisien. Cette approche comparative nous a permis de sélectionner et d'adapter judicieusement les éléments pertinents du dialecte tunisien, les intégrant de manière organique dans notre méthode pour le dialecte algérien.

En combinant l'expertise préexistante de notre laboratoire dans le domaine du dialecte tunisien avec notre connaissance approfondie du dialecte algérien, nous avons créé une méthode de segmentation riche et nuancée. Cette fusion des dialectes a abouti à une méthode holistique, capable de saisir les subtilités linguistiques propres à chaque dialecte tout en les unissant dans un cadre cohérent.

3.4.4 Présentation de la méthode proposée

La Figure 4 illustre notre méthode de segmentation pour le dialecte algérien. Cette méthode se fonde sur un ensemble complexe de règles linguistiques, combinant les spécificités du dialecte algérien et les éléments clés du dialecte tunisien adaptés à l'arabe algérien. Ces règles nous permettent de détecter avec précision les limites des phrases dans le contexte linguistique unique du dialecte algérien.

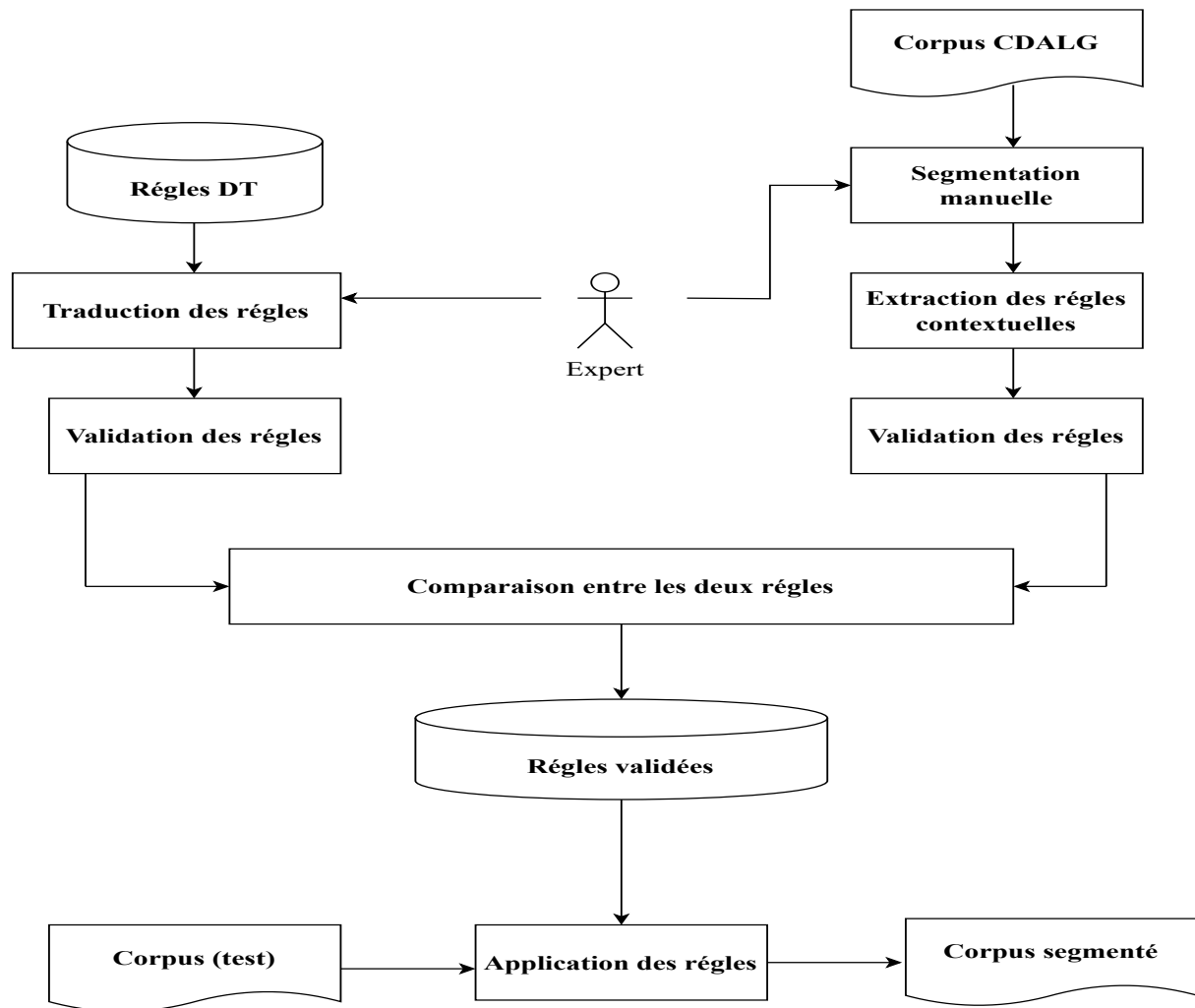


FIGURE 4. méthode proposé pour la segmentation du dialecte algérien

3.4.5 Méthode détaillée de segmentation du dialecte algérien

Dans cette section, nous plongeons profondément dans les mécanismes de notre méthode de segmentation pour le dialecte algérien. Après avoir posé le contexte et exploré les travaux antérieurs, nous sommes prêts à présenter en détail notre approche innovante. Cette méthode détaillée expose les règles linguistiques spécifiques que nous avons développées, ainsi que les adaptations des concepts tunisiens que nous avons intégrées. À travers cette exploration minutieuse, nous démontrerons comment notre méthode se distingue, offrant une compréhension exhaustive de son fonctionnement et de son potentiel dans le domaine du traitement automatique du langage naturel pour le dialecte algérien.

3.4.5.1 Extraction et organisation des règles

À ce stade, nous avons procédé à la segmentation manuelle du corpus en collaboration avec des experts du domaine. L'accord entre ces experts a été évalué en utilisant le coefficient kappa, obtenant ainsi une valeur élevée de 0,83. Cette étape initiale a été essentielle pour établir une base solide. De plus, il est important de noter que notre corpus est ALGDC, ce qui a influencé notre analyse. Par la suite, après une analyse minutieuse du corpus, nous avons pu extraire des règles linguistiques spécifiques que nous avons ensuite intégrées dans notre processus de segmentation. Ces règles se révéleront précieuses lors de la segmentation automatique.

Notre méthode débute par l'identification des mots et des signes caractéristiques du dialecte algérien, accompagnée de l'examen de leur contexte à gauche et à droite. Nous avons minutieusement sélectionné les règles les plus fiables, constituant ainsi le fondement de notre méthode. Pour garantir une segmentation précise et cohérente, nous avons organisé ces règles dans une séquence logique. Chaque règle a été conçue de manière à pouvoir corriger la segmentation issue de la règle précédente, ou des règles précédentes. Ces règles contextuelles nous permettent de reconnaître avec précision le début d'une nouvelle phrase, marquant ainsi la conclusion de la phrase précédente.

Après avoir recueilli les règles linguistiques qui nous ont aidés dans la segmentation automatique du dialecte algérien, soit un total de 22 règles linguistiques, nous présenterons à ce stade les plus importantes d'entre elles. Le format des règles contextuelles est le suivant :

Contexte à gauche	Marqueur	Contexte à droite
X+1	X	X-1

TABLE 25. *Format de règles conçues pour la segmentation de texte*

Règle 1 :

Dans la règle 1, si le marqueur est une question commençant par "dis-moi" (قولي، قوليلي، قولولي، قولنا) et suivi du mot "quoi" (واش), cela signifie que le déclencheur marque le début d'une nouvelle phrase. En effet, dans le dialecte algérien, la phrase "dis-moi quoi" indique systématiquement le début d'une nouvelle phrase et ne peut pas être utilisée au milieu ou à la fin d'une phrase.

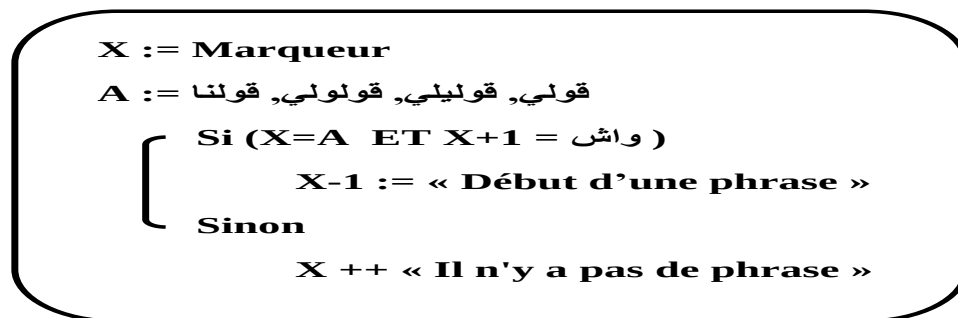


FIGURE 5. Algorithme exprimant la Règle 1

Règle 2 :

Dans la règle 2 que nous avons développée, si le mot est un mot de la liste des "mots de coordination" tels que "mais, de plus, et, aussi, et après" (ولا، بـصح، و، زید، تانا، و من بعد) cela signifie que le mot juste avant le marqueur indique la fin d'une phrase, et le marqueur indique le début d'une nouvelle phrase.

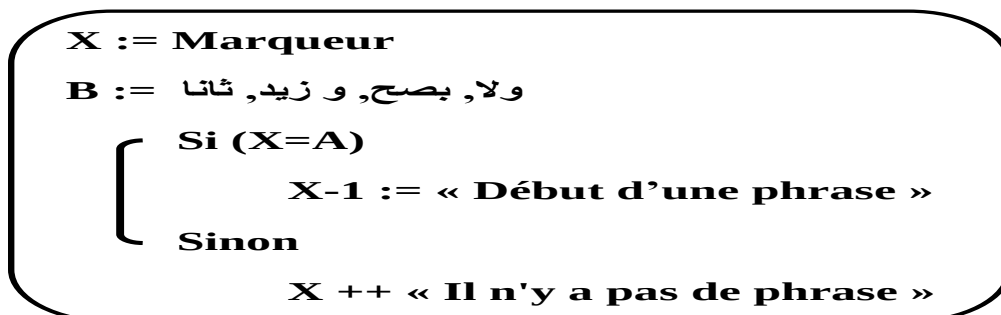


FIGURE 6. Algorithme exprimant la Règle 2

3.4.5.2 Enrichissement des règles de segmentation algérien

Rappelons que plusieurs mots sont en commun entre le DALG et le DT. Et suite à une étude minutieuse sur notre corpus, nous avons pu remarquer que plusieurs règles d'exploration contextuelle pour le Segdt, réalisées dans le travail de Zribi et al., peuvent être adaptées et utilisées au niveau de la segmentation du DALG. Dans ce principe, nous avons enrichi notre base de règles, et cette étape a été corrigée et validée par des experts du domaine.

À cette étape de notre méthode, nous avons considérablement enrichi notre base de règles linguistiques en incorporant des éléments spécifiques du dialecte tunisien, que nous avons traduits avec précision en DALG. Ces règles, sélectionnées avec soin, ont été extraites à l'aide d'une méthode similaire à la nôtre, comme présenté dans l'étude de référence par Zribi et al [32]. La traduction a été effectuée sous la supervision attentive de deux experts, garantissant un taux de précision exceptionnel de 0,87.

Cette étape marque un point tournant dans notre méthode, nous permettant d'élargir considérablement notre compréhension des nuances linguistiques régionales. L'intégration harmonieuse des éléments du dialecte tunisien dans notre cadre algérien renforce la robustesse de notre méthode, nous offrant ainsi une perspective plus riche et nuancée pour la segmentation du dialecte algérien.

3.4.5.3 Comparaison entre les deux ensembles de règles

Après avoir extrait les règles linguistiques basées sur les marqueurs spécifiques du dialecte algérien dans notre première étape, et suite à l'étape de traduction des règles du dialecte tunisien sous la supervision d'experts, nous avons procédé à une comparaison minutieuse des deux ensembles de règles obtenus. Les règles en double ont été éliminées, et de nouvelles règles ont été ajoutées suite à cette comparaison approfondie.

Ce processus de comparaison nous a permis de créer un ensemble de règles unifié et validé, prêt à être appliqué sur un corpus de test. Cette étape de raffinement a renforcé la qualité et la précision de nos règles, garantissant ainsi une segmentation plus précise et cohérente du dialecte algérien dans notre méthode.

3.5 Réalisation et évaluation

Notre approche de segmentation pour le dialecte algérien repose sur l'intégration de deux méthodes robustes : STAR et SegDT. Le système STAR, tel que présenté par [63], utilise une approche contextuelle basée sur la ponctuation et certaines particules pour effectuer une segmentation précise et efficace du texte en phrases. Il identifie la ponctuation et les particules indiquant la fin d'une phrase en utilisant des règles, et analyse le contexte pour garantir leur utilisation appropriée, même dans des cas où la ponctuation et les particules sont utilisées de manière non conventionnelle [1].

En parallèle, nous avons intégré la méthode de segmentation automatique du dialecte tunisien SegDT ([33]), qui combine des approches linguistiques, statistiques et hybrides. Cette méthode hybride s'appuie à la fois sur l'apprentissage et les règles linguistiques. Nous avons adapté cette méthode au dialecte algérien en y incorporant la section de segmentation basée sur des règles contextuelles. La fusion de ces deux méthodes a permis de créer un système de segmentation robuste et efficace pour le dialecte algérien.

Il est à noter que les deux systèmes, STAR et SegDT, utilisent le langage Perl, nécessitant ainsi notre maîtrise de ce langage en plus de Java. En outre, comme mentionné précédemment, nous avons principalement utilisé le corpus ALGDC pour extraire les règles linguistiques de segmentation, ainsi que pour mener nos expérimentations.

Pour évaluer l'efficacité de notre système SegDz, nous l'avons comparé à diverses méthodes d'évaluation, notamment la segmentation du dialecte algérien à l'aide du

système STAR, la segmentation du dialecte tunisien à l'aide de SegDt, ainsi que la segmentation utilisant des règles contextuelles extraites et traduites respectivement du dialecte algérien et tunisien. Ces évaluations ont été réalisées sur le corpus du dialecte algérien, et les performances ont été mesurées en termes de rappel, de précision et d'exactitude. Les résultats détaillés sont présentés dans le tableau 26.

Évaluation des méthodes de segmentation

Pour évaluer la performance de diverses méthodes de segmentation sur le dialecte algérien, nous avons adopté plusieurs méthodes :

- *Segmentation avec STAR* : Nous avons testé l'outil de segmentation STAR, conçu pour l'arabe standard moderne (ASM), sur le Corpus du dialecte algérien. Cette évaluation visait à mesurer l'efficacité de STAR dans la segmentation du dialecte algérien.
- *Segmentation avec SegDt* : Nous avons évalué les performances de SegDt, un segmenteur initialement conçu pour le dialecte tunisien, sur notre corpus de dialecte algérien. Cette évaluation portait sur l'exactitude de SegDt dans la segmentation du texte en unités significatives, telles que les mots et les phrases, pour le dialecte algérien.
- *RC-DALG (Règles Contextuelles du Dialecte Algérien)* : Cette méthode implique l'application des règles de segmentation extraites spécifiquement du corpus du dialecte algérien. Ces règles ont été détaillées dans la section 3.4.5.1 de l'article et ont été utilisées pour évaluer la fiabilité du système de segmentation du dialecte algérien.
- *RC-DTT (Règles Contextuelles du Dialecte Tunisien Traduites)* : Ici, nous avons utilisé les règles contextuelles du dialecte tunisien, traduites en dialecte algérien, et les avons adaptées au contexte algérien pour évaluer leur efficacité dans la segmentation du dialecte algérien. Ces règles traduites ont été évaluées sur le corpus du dialecte algérien pour mesurer leur précision.
- *SegDz (RC-DALG+DTT)* : Cette méthode novatrice combine les règles contextuelles extraites du dialecte algérien (RC-DALG) avec les règles

contextuelles traduites du dialecte tunisien (RC-DTT). Nous avons évalué les performances de cette méthode hybride sur un corpus contenant des données en dialecte algérien.

Cette évaluation approfondie nous a permis de comparer ces différentes méthodes de segmentation et de déterminer celle qui offre la meilleure précision et cohérence dans le contexte spécifique du dialecte algérien. Les résultats de ces évaluations sont résumés dans le tableau 26, fournissant ainsi un aperçu clair de l'efficacité de chaque méthode dans notre étude.

Méthode	Rappel	Précision	F-Mesure
Star	37.22	78.43	50.48
SegDT	45.88	82.76	59.03
RC-DALG	68.39	91.12	78.14
RC-DTT	70.93	92.78	80.40
SegDz	73.41	93.36	82.19

TABLE 26. *Les résultats de l'évaluation sur la segmentation des textes en dialecte algérien*

Discussion des résultats

Dans notre étude, nous avons évalué différentes méthodes de segmentation pour le dialecte algérien, chacune ayant ses propres avantages et défis. Les résultats présentés dans le tableau 26 offrent un aperçu clair des performances de chaque méthode, évaluées en termes de rappel, de précision et de F-mesure.

1. La méthode STAR, basée sur une approche contextuelle utilisant la ponctuation et les particules, montre une précision décente à 78.43%. Cependant, son faible rappel à 37.22% indique qu'elle manque de cohérence dans l'identification des limites de phrases, entraînant ainsi une F-mesure modérée de 50.48

2. SegDT, bien que conçu pour le dialecte tunisien, offre des performances respectables sur le dialecte algérien. Avec un rappel de 45.88%, une précision de 82.76% et une F-mesure de 59.03%, cette méthode démontre une amélioration par rapport à STAR, mais avec des marges d'amélioration.

3. Les règles contextuelles spécifiques extraites du corpus algérien (RC-DALG) présentent des performances nettement meilleures, avec un rappel de 68.39%, une

précision de 91.12% et une F-mesure de 78.14%. Ces règles, élaborées sur la base des spécificités du dialecte algérien, montrent une compréhension approfondie des caractéristiques linguistiques locales.

4. Les règles contextuelles traduites du dialecte tunisien (RC-DTT) maintiennent également des performances solides, avec un rappel de 70.93%, une précision de 92.78% et une F-mesure de 80.40%. Cette traduction réussie indique une similarité remarquable entre les deux dialectes.

5. Notre méthode hybride, SegDz, combinant les règles contextuelles du dialecte algérien et celles traduites du dialecte tunisien, présente les meilleurs résultats. Avec un rappel de 73.41%, une précision de 93.36% et une F-mesure de 82.19%, SegDz offre une segmentation précise et cohérente du dialecte algérien. Cette approche hybride capitalise sur les forces des deux dialectes, renforçant ainsi son efficacité.

En somme, notre étude démontre que les règles contextuelles élaborées spécifiquement pour le dialecte algérien, combinées à une adaptation intelligente des règles du dialecte tunisien, conduisent à une segmentation précise et fiable. SegDz, en particulier, représente une avancée significative dans la segmentation automatique du dialecte algérien. Ces résultats mettent en lumière l'importance d'une approche ciblée et adaptative pour la segmentation des langues dialectales, ouvrant ainsi la voie à des applications plus précises dans le domaine du traitement automatique du langage naturel pour le dialecte algérien et d'autres langues similaires.

3.6 Analyse des erreurs de segmentation

Notre analyse des erreurs se penche sur les situations où notre système de segmentation automatique du dialecte algérien a montré des faiblesses, afin de cerner les problèmes sous-jacents. Cette investigation approfondie a éclairé les domaines nécessitant des améliorations pour assurer la fiabilité de notre système. Ci-dessous, nous examinons un exemple concret illustrant une erreur courante :

3.6.1 Contexte de l'erreur

Nous avons établi une règle stipulant que le mot "ET" (و) doit précéder un verbe en dialecte algérien pour marquer le début d'une phrase. Cette règle sert à identifier les phrases en dialecte algérien dans un texte donné. Cependant, notre liste de verbes, utilisée comme déclencheur pour cette règle, n'est pas exhaustive. Certains verbes peuvent ne pas être inclus dans cette liste.

3.6.2 Exemple d'erreur

Considérons la phrase : "Chaque jour vient avec son propre salaire, mon frère. Une personne a besoin de travailler le matin et le soir pour bien vivre," e,"
كل يوم وبراکتوا یا خو و نافیقی صباح و عشية باش نعیش

Segmentation manuelle :

[S1] كل يوم وبراکتوا یا خو

[S2] و نافیقی صباح و عشية باش نعیش

Segmentation automatique

[S1] كل يوم وبراکتوا یا خو و نافیقی صباح و عشية باش نعیش

Dans cet exemple, la segmentation automatique a regroupé la phrase en une seule unité, tandis que la segmentation manuelle l'a divisée en deux parties distinctes. Cette erreur découle du fait que notre système ne reconnaît pas les verbes qui ne figurent pas dans notre liste, entraînant ainsi une segmentation incorrecte.

3.6.3 Actions correctives

Pour améliorer la précision de notre système, nous devons élargir notre liste de verbes pour inclure un ensemble plus complet de termes du dialecte algérien. Cela permettra à notre système de reconnaître un plus grand nombre de déclencheurs valides pour marquer les débuts de phrases. Travailler en collaboration avec des

locuteurs natifs et des experts linguistiques peut également contribuer à enrichir notre base de données verbales.

En conclusion, en augmentant notre base de données verbales et en améliorant la sensibilité de notre système aux nuances des verbes dialectaux algériens, nous pouvons corriger cette erreur spécifique et assurer une segmentation plus précise et cohérente des textes en dialecte algérien.

3.7 Conclusion

En conclusion de ce chapitre, nous avons mis en lumière l'importance cruciale de la segmentation des textes dans le domaine du traitement automatique des langues naturelles, notamment dans le contexte complexe du dialecte algérien. La segmentation est une étape fondamentale utilisée dans divers domaines tels que le résumé automatique, la traduction automatique et la reconnaissance des langues et dialectes. Cependant, les dialectes arabes, y compris le dialecte algérien, présentent des défis spécifiques dus à leurs caractéristiques linguistiques uniques et à l'absence de règles claires.

Dans ce chapitre, nous avons présenté une méthode de segmentation novatrice basée sur des règles linguistiques spécifiques au dialecte algérien. Ces règles ont été extraites manuellement à partir d'un corpus de texte algérien et ont été enrichies en incorporant des règles traduites du dialecte tunisien. L'intégration de deux méthodes, STAR et SegDT, offrant des approches contextuelles et des règles linguistiques, a conduit au développement du système SegDz, un système de segmentation robuste et efficace pour le dialecte algérien.

L'évaluation de SegDz a été réalisée en comparant ses performances avec d'autres méthodes de segmentation telles que STAR, SegDT, ainsi que des règles contextuelles spécifiques au dialecte algérien et traduites du dialecte tunisien. Malgré les améliorations apportées, certaines limitations subsistent, identifiées lors de

l'analyse des erreurs. Ces faiblesses soulignent la nécessité d'améliorations continues pour accroître la précision et la pertinence des résultats obtenus par notre système.

En résumé, la segmentation des textes en dialecte algérien représente un défi de taille dans le domaine du traitement automatique des langues. Bien que des progrès significatifs aient été réalisés grâce à l'approche novatrice présentée dans ce chapitre, des efforts continus sont essentiels pour surmonter les limitations actuelles. Ces améliorations permettront non seulement d'offrir une expérience utilisateur optimale dans les applications de traitement automatique des langues en dialecte algérien, mais contribueront également à l'avancement global de la recherche dans ce domaine complexe et fascinant.

Chapitre

4

Analyse syntaxique dans le dialecte algérien : méthodes et résultats

4.1 Introduction

L'analyse syntaxique revêt une importance cruciale dans le processus d'analyse automatique du langage, permettant de spécifier les structures syntaxiques des phrases. L'étude de l'arabe et de ses dialectes constitue un domaine passionnant, mais complexe, en raison des nombreux obstacles rencontrés. La complexité et la richesse de l'arabe standard moderne (MSA) ainsi que des dialectes arabes (DA) ajoutent à cette difficulté.

Le dialecte algérien se distingue comme l'un des dialectes les plus importants et les plus distincts par rapport à la langue standard. Il présente des particularités phonétiques, grammaticales et lexicales qui lui sont propres. L'utilisation d'un corpus tel que le Treebank, annoté avec une structure syntaxique pour chaque phrase, se révèle essentielle. Cet outil, comme celui proposé par [65], facilite la compréhension et l'analyse de cette langue, particulièrement dans le contexte du dialecte algérien. Dans ce chapitre, en plus de notre exploration approfondie de l'analyse syntaxique du dialecte algérien, nous mettrons en lumière les travaux préexistants dans le domaine de l'analyse syntaxique de la langue arabe, du dialecte arabe, et plus spécifiquement du

dialecte algérien. En examinant ces travaux antérieurs, nous pourrions contextualiser notre méthode d'analyse et identifier les avancées ainsi que les défis spécifiques rencontrés dans l'analyse syntaxique des dialectes arabes.

Cette contextualisation nous permettra de mieux comprendre les particularités du dialecte algérien par rapport à l'arabe standard et aux autres dialectes arabes, facilitant ainsi le développement d'approches plus précises et adaptées à cette variante linguistique spécifique. En examinant les progrès réalisés dans l'analyse syntaxique de l'arabe standard et des autres dialectes arabes, nous pourrions identifier les similitudes et les différences avec le dialecte algérien, offrant ainsi des perspectives précieuses pour notre propre méthodologie.

Ce chapitre se veut donc une exploration complète des efforts antérieurs dans le domaine de l'analyse syntaxique de la langue arabe, des dialectes arabes et du dialecte algérien, fournissant ainsi un cadre solide pour notre propre méthode d'analyse syntaxique spécifique au dialecte algérien. En combinant ces connaissances préalables avec notre propre approche novatrice, nous visons à contribuer de manière significative à la compréhension et à l'analyse des structures grammaticales complexes du dialecte algérien.

4.2 Exploration approfondie de l'analyse syntaxique

L'analyse syntaxique, en tant que discipline fondamentale du Traitement Automatique du Langage Naturel (TALN), a fait l'objet de nombreuses recherches approfondies. Cette section examine les progrès significatifs réalisés dans ce domaine, en se concentrant d'abord sur l'analyse syntaxique de la langue arabe dans son ensemble, puis en se focalisant sur les recherches spécifiques concernant le dialecte arabe, avec une attention particulière portée au dialecte algérien.

4.2.1 Avancements dans l'analyse syntaxique de l'arabe standard

Dans le domaine du traitement automatique des langues (TAL), l'analyse syntaxique revêt une importance cruciale dans de nombreuses opérations telles que la compréhension du texte, l'extraction d'informations ou la traduction [5].

Un analyseur syntaxique peut construire la structure syntaxique d'un texte d'entrée donné. En d'autres termes, sa tâche est de déterminer la fonction syntaxique de chaque élément de la phrase, ainsi que la dépendance syntaxique des éléments fondamentaux de la phrase, tels que le sujet-verbe ou le verbe-objet. En raison de la complexité et de la richesse de la langue arabe standard moderne, il s'agit d'une tâche délicate [41]. En identifiant les différents composants de la phrase et les rôles entre eux, l'analyse syntaxique peut relier la phrase à sa structure syntaxique possible. En essence, l'analyse syntaxique prend le résultat de l'analyse lexicale et fournit une structure hiérarchique de regroupements structurés et de relations fonctionnelles, unissant les groupes de sortie.

Il existe peu de travaux sur les analyseurs syntaxiques pour l'arabe. Nous citons ci-dessous les plus importants :

L'analyseur syntaxique de Stanford proposé par Green et Manning [49] est un analyseur basé sur la grammaire. Cet outil Stanford est un projet open source¹². Il résout l'analyse syntaxique en utilisant une grammaire stochastique non contextuelle. Il a été développé en utilisant l'Arabic Treebank [65]. Comme cet analyseur n'a pas d'outil de tokenisation intégré, le corpus doit déjà être tokenisé avant de pouvoir être segmenté en pronoms clitiques, prépositions, conjonctions, etc. La segmentation des déterminants clitiques "ال", "Al", "the" n'est pas requise par cet analyseur. L'évaluation de l'analyseur de Stanford avec le modèle de base a donné un score F de 79,86%, et avec le modèle GoldPos, 80,67%.

Il y a aussi l'outil CamelParser [31], qui est un système de pointe pour l'analyse de la dépendance syntaxique arabe associée à des traits morphologiques

12. <https://nlp.stanford.edu/software/lex-parser.shtml>

contextualisés. L'outil CamelParser utilise un désambiguïseur morphologique de pointe, qui est amélioré avec des caractéristiques syntaxiques. Les caractéristiques offertes par le système comprennent la dépendance de base avec des caractéristiques morphologiques, deux modes d'affichage arborescent, et l'analyse syntaxique arabe classique [31]. En termes de précision d'attachement étiqueté, d'attachement non étiqueté et de précision d'étiquette, CamelParser atteint respectivement 83,8%, 86,4% et 93,2%.

4.2.2 Avancements dans l'analyse syntaxique des dialectes arabes

L'analyse syntaxique implique de déterminer la plus petite composante syntaxique d'une énonciation pour construire son arbre syntaxique associé. Il existe quelques travaux concernant l'analyse syntaxique des dialectes arabes, parmi lesquels nous pouvons citer le travail de [26] comme exemple. Ils ont proposé une méthode pour générer un Treebank pour l'arabe tunisien qui a été intellectualisé, basé sur la constitution tunisienne.

En effet, ils ont commencé par prétraiter leur corpus pour le normaliser en utilisant le modèle orthographique "CODA-TUN". La constitution a ensuite été soumise à une étape de segmentation des phrases pour corriger la segmentation du corpus, qui était incorrecte en raison de la longueur des phrases. Ensuite, la constitution a été tokenisée pour compléter l'étape de prétraitement. Suite à cela, le corpus prétraité a été soumis à l'analyseur syntaxique de la version arabe du "Stanford Parser" pour créer un "Tunisian Treebank". Ils ont recommandé l'utilisation du Stanford Parser pour produire un modèle qu'ils ont testé à l'aide de leur Treebank, étant donné que l'arabe tunisien est une variante de l'arabe standard. Les taux de rappel et de précision sont respectivement de 66,77% et 64,43% [26].

4.2.3 Innovations dans l'analyse syntaxique du dialecte algérien

L'analyse syntaxique d'un dialecte aussi riche et complexe que l'algérien est une entreprise essentielle dans le domaine de la linguistique computationnelle. Elle ouvre la porte à une compréhension approfondie de la structure grammaticale de ce dialecte, offrant ainsi des perspectives précieuses pour la traduction automatique, la compréhension du langage naturel et d'autres applications dans le domaine du traitement automatique des langues. Cette section se concentre sur les travaux novateurs dans le domaine de l'analyse syntaxique du dialecte algérien.

Dans [24], les auteurs décrivent un processus de création et de mise en œuvre d'un analyseur syntaxique ASDA (Analyseur Syntaxique pour le Dialecte Algérien) qui étiquette les mots dans un corpus. Pour identifier les différents éléments grammaticaux, notamment l'étiquetage des parties du discours (POS), les chercheurs ont utilisé une table d'étiquetage contenant la racine de chaque phrase ainsi qu'un certain nombre de préfixes et de suffixes. Ce dialecte sera ensuite traité sémantiquement, ou traduit automatiquement, en utilisant ces étiquetages. De plus, les émotions représentées dans les textes écrits dans ce dialecte seront analysées [24].

4.3 Défis et complexités : L'analyse syntaxique du dialecte algérien

L'analyse syntaxique, en tant que branche fondamentale de la linguistique, explore en profondeur la structure grammaticale des langues, offrant un aperçu précieux des mécanismes complexes qui sous-tendent la communication humaine. Cependant, lorsqu'elle est appliquée à des dialectes spécifiques, notamment le dialecte algérien, elle révèle des défis uniques et des opportunités intrigantes.

Dans le monde vaste et diversifié des langues, chaque dialecte porte en lui une richesse culturelle et linguistique propre. Cependant, cette diversité linguistique

engendre également des défis considérables pour les linguistes et les chercheurs en traitement automatique du langage naturel (TALN). En effet, l'analyse syntaxique, qui cherche à décoder les règles et les structures grammaticales d'une langue, est confrontée à des variations subtiles mais significatives dans les dialectes arabes, y compris le dialecte algérien.

La langue arabe elle-même est célèbre pour sa complexité, avec ses multiples dialectes et son arabe standardisé. Cette diversité présente des défis majeurs pour l'analyse syntaxique, car les variations dans les structures de phrases, les conjugaisons verbales et l'utilisation des prépositions exigent une compréhension approfondie de chaque dialecte. Dans le contexte spécifique du dialecte algérien, les particularités phonétiques, lexicales et grammaticales ajoutent une couche de complexité supplémentaire, nécessitant des outils et des méthodes d'analyse spécialement adaptés.

Cependant, malgré ces défis, l'analyse syntaxique offre des avantages considérables dans le traitement automatique du dialecte algérien. En décomposant les structures grammaticales complexes en éléments compréhensibles, l'analyse syntaxique facilite non seulement la compréhension de la langue, mais aussi son utilisation dans des applications modernes telles que la traduction automatique, l'extraction d'informations et l'interaction homme-machine. En d'autres termes, une analyse syntaxique précise est la clé qui ouvre la porte à un traitement linguistique automatisé efficace et précis du dialecte algérien, permettant ainsi une communication plus fluide et une intégration accrue de cette variante linguistique riche dans le domaine en constante évolution de la technologie linguistique.

4.3.1 Défis de l'analyse syntaxique

L'analyse syntaxique, en particulier lorsqu'elle est appliquée à l'arabe standard et à ses dialectes, dont le dialecte algérien, se trouve face à un défi complexe. Cette complexité réside dans la diversité intrinsèque de la langue, tant au niveau

de l'arabe standard que des dialectes régionaux comme celui d'Algérie. Examinons attentivement ces défis pour mieux comprendre les subtilités qui se cachent derrière l'analyse syntaxique de ces textes.

4.3.1.1 Défis de l'analyse syntaxique de l'arabe standard

L'arabe standard est déjà réputé pour sa structure linguistique sophistiquée. Ses règles grammaticales complexes, telles que la conjugaison verbale, les cas grammaticaux et les structures de phrase, rendent l'analyse syntaxique arabe extrêmement exigeante. Les variations régionales dans l'arabe standard ajoutent un niveau supplémentaire de complexité. Par exemple, les nuances dans les prépositions, les pronoms et les verbes peuvent différer considérablement d'une région à l'autre, ce qui complique l'interprétation syntaxique.

Exemple : La construction de la phrase passive en arabe standard implique souvent une réorganisation complexe des mots, modifiant ainsi l'ordre syntaxique standard sujet-verbe-objet.

4.3.1.2 Défis de l'analyse syntaxique du dialecte algérien :

Le dialecte algérien représente un défi unique en matière d'analyse syntaxique. Les variations phonétiques, lexicales et syntaxiques d'une région algérienne à l'autre sont frappantes, créant une mosaïque linguistique complexe. Les expressions idiomatiques locales, les particularités dans la syntaxe des phrases et les variations dans l'emploi des prépositions et des conjonctions rendent l'analyse syntaxique algérienne particulièrement délicate.

Exemple concret : Dans certaines régions d'Algérie, les prépositions peuvent être omises dans certains contextes, créant des structures de phrases non standard qui défient les conventions grammaticales. De plus, les variations dans l'ordre des mots, les conjonctions et l'utilisation des prépositions peuvent différer considérablement

d'une région à l'autre, ajoutant une couche de complexité supplémentaire à l'analyse syntaxique.

L'analyse syntaxique du dialecte algérien est un défi pour plusieurs raisons. Tout d'abord, il existe une variabilité régionale importante dans ce dialecte, ce qui signifie qu'il peut y avoir des différences significatives dans les structures grammaticales et les usages locaux d'une région à l'autre. Cela rend difficile l'établissement de règles grammaticales universelles qui s'appliquent à toutes les régions.

En outre, le dialecte algérien présente des particularités par rapport à l'arabe littéraire qui peuvent rendre l'analyse syntaxique plus complexe. Par exemple, il y a des différences dans les formes verbales et les pronoms personnels, ainsi que dans la manière dont les mots sont reliés entre eux pour former des phrases.

Un autre défi majeur réside dans le manque de données et de ressources pour l'analyse syntaxique du dialecte algérien, rendant difficile l'application de méthodes informatiques ou automatiques. La langue algérienne est un dialecte parlé qui est souvent utilisé dans des contextes informels, ce qui rend ardue la collecte de données pour l'analyse.

Enfin, l'analyse syntaxique du dialecte algérien est complexe en raison des variations dans les usages et les contextes culturels dans lesquels il est utilisé, ce qui rend difficile l'application de règles universelles pour l'analyse. Ces défis, bien que complexes, ajoutent une couche de fascination à l'étude de l'analyse syntaxique du dialecte algérien, mettant en lumière la diversité et la richesse de cette variante linguistique.

4.3.2 Complexités syntaxiques du dialecte algérien

Le dialecte algérien, en tant que reflet de l'histoire, de la diversité culturelle et des influences linguistiques uniques de la région, présente des complexités syntaxiques fascinantes et distinctives. Explorons les spécificités de sa syntaxe, les comparant à l'arabe standard, pour comprendre la richesse de ce dialecte.

Spécificités de la syntaxe du dialecte algérien :

Une des caractéristiques marquantes du dialecte algérien réside dans ses règles de concordance sujet-verbe-objet. Contrairement à l'arabe standard, le dialecte algérien peut avoir des structures de phrases uniques, où l'ordre des éléments peut varier en fonction du contexte et de l'émotion exprimée. Cette flexibilité syntaxique permet au dialecte algérien de capturer des nuances subtiles, créant ainsi des communications riches en significations.

Comparaison avec l'arabe standard :

Comparativement à l'arabe standard, le dialecte algérien présente des particularités syntaxiques qui se démarquent. Les pronoms, par exemple, sont souvent intégrés de manière fluide dans la structure de la phrase, créant des constructions syntaxiques différentes de celles de l'arabe standard. Les prépositions et les conjonctions peuvent également être utilisées de manière unique pour exprimer des relations spatiales et temporelles, conférant au dialecte algérien une qualité poétique et expressive.

Complexités syntaxiques au niveau du dialecte algérien :

La complexité syntaxique du dialecte algérien va au-delà de la simple structure des phrases. Elle réside également dans la richesse des expressions idiomatiques et des métaphores utilisées. Ces éléments ajoutent des dimensions supplémentaires à l'analyse syntaxique, car ils nécessitent une compréhension profonde du contexte culturel et social dans lequel ils sont utilisés.

Exemple : Dans certaines régions d'Algérie, des expressions idiomatiques spécifiques peuvent modifier la signification d'une phrase de manière subtile mais significative, soulignant ainsi l'importance de la connaissance contextuelle dans l'analyse syntaxique du dialecte algérien.

En explorant ces complexités syntaxiques, nous sommes témoins de la créativité et de l'ingéniosité linguistique des locuteurs du dialecte algérien. Ces nuances reflètent l'âme même de la région et enrichissent notre compréhension des subtilités linguistiques de ce dialecte unique et diversifié.

4.3.3 Avantages de l'analyse syntaxique du dialecte algérien

L'analyse syntaxique du dialecte algérien offre une pléthore d'avantages, non seulement pour les chercheurs et les linguistes, mais aussi pour les applications pratiques dans le domaine du traitement automatique du langage naturel. Explorons ces avantages qui témoignent de l'importance de cette analyse approfondie.

Précision dans la communication :

En comprenant les subtilités syntaxiques du dialecte algérien, les locuteurs peuvent communiquer de manière plus précise et claire, évitant ainsi les malentendus et les interprétations incorrectes. L'analyse syntaxique permet de saisir les nuances et les intentions derrière chaque phrase, facilitant ainsi une communication plus efficace au sein des communautés algériennes.

Amélioration des applications de Traitement Automatique du Langage Naturel (TALN) :

Dans le domaine du TALN, l'analyse syntaxique précise du dialecte algérien est cruciale pour le développement d'applications telles que les traducteurs automatiques et les systèmes de reconnaissance vocale. En comprenant la structure syntaxique unique de ce dialecte, les ingénieurs en TALN peuvent créer des outils plus efficaces et précis, améliorant ainsi l'expérience utilisateur pour les locuteurs du dialecte algérien.

Préservation de l'héritage culturel :

Le dialecte algérien est porteur d'une richesse culturelle et historique unique. En analysant sa syntaxe, nous préservons cet héritage linguistique précieux. Une compréhension approfondie de la structure grammaticale du dialecte algérien contribue à maintenir vivante cette forme spécifique de la langue arabe, préservant ainsi l'identité culturelle de la région.

Adaptation aux besoins éducatifs :

Dans le domaine de l'éducation, une analyse syntaxique précise du dialecte algérien est essentielle pour l'enseignement des langues. Les enseignants peuvent utiliser cette analyse pour élaborer des programmes éducatifs adaptés aux particularités linguistiques des élèves algériens, favorisant ainsi un apprentissage plus efficace et une meilleure compréhension des concepts linguistiques.

En conclusion, l'analyse syntaxique du dialecte algérien transcende les simples frontières linguistiques. Elle ouvre des portes à une communication plus précise, à des avancées technologiques dans le domaine du TALN, à la préservation culturelle et à une éducation adaptée aux besoins des élèves algériens. Cette analyse, riche de son histoire et de sa diversité, est un témoignage de l'importance de comprendre les subtilités linguistiques des dialectes, préservant ainsi le tissu culturel qui les entoure.

4.3.4 Analyse syntaxique : arabe standard vs. dialecte algérien

L'analyse syntaxique est un domaine clé de l'étude des langues, et cela est particulièrement vrai pour la langue arabe et le dialecte algérien. Les travaux réalisés dans ce domaine ont permis de mieux comprendre les structures grammaticales de ces deux langues, mais ils ont également mis en évidence des défis uniques liés à l'analyse syntaxique de chacune d'elles.

En ce qui concerne la langue arabe standard, de nombreux travaux ont été réalisés pour décrire sa grammaire et ses structures syntaxiques. Ces travaux ont permis de développer des outils informatiques pour l'analyse syntaxique de la langue arabe, tels que des analyseurs syntaxiques automatiques. Cependant, il existe des variations régionales importantes dans la langue arabe, ce qui rend difficile l'établissement de règles grammaticales universelles.

Le dialecte algérien, d'autre part, a été moins étudié que la langue arabe standard. Les travaux réalisés sur l'analyse syntaxique du dialecte algérien ont mis en évidence des différences significatives par rapport à la langue arabe standard, notamment en ce qui concerne les formes verbales et les pronoms personnels. De plus, il existe des variations régionales importantes dans le dialecte algérien, ce qui rend difficile l'établissement de règles grammaticales universelles.

Malgré ces défis, des travaux récents ont commencé à utiliser des techniques d'apprentissage automatique pour l'analyse syntaxique du dialecte algérien, avec des résultats prometteurs. Il y a aussi des efforts pour collecter des corpus de données pour étudier le dialecte algérien, ce qui devrait permettre de développer des outils et des méthodes d'analyse syntaxique plus efficaces.

En résumé, les travaux réalisés dans le domaine de l'analyse syntaxique de la langue arabe et du dialecte algérien ont permis de mieux comprendre ces deux langues, mais ils ont également mis en évidence des défis uniques liés à l'analyse syntaxique de chacune d'elles. Les variations régionales et les différences entre la langue arabe standard et le dialecte algérien rendent difficile l'établissement de règles grammaticales universelles. Cependant, des travaux récents ont commencé à utiliser des techniques d'apprentissage automatique pour l'analyse syntaxique du dialecte algérien, avec des résultats prometteurs.

4.4 Analyse syntaxique du dialecte algérien : une méthode proposée

L'analyse syntaxique, pierre angulaire du traitement automatique du langage naturel, se décline en différentes approches dans la littérature spécialisée. Trois méthodes émergent en tant que piliers fondamentaux : l'approche linguistique, l'approche numérique et l'approche hybride. L'approche numérique, tout d'abord, repose sur des modèles statistiques élaborés à partir de vastes corpus d'entraînement. Ces modèles, fruit d'une automatisation ingénieuse, identifient les structures syntaxiques au sein des phrases et des textes. D'autre part, l'approche linguistique se fonde sur l'application rigoureuse de règles grammaticales définies. Ces règles décrivent les combinaisons potentielles de phrases, accompagnées d'une stratégie de parsing méticuleuse, permettant ainsi de déterminer l'appartenance d'une phrase donnée à la langue en question.

Finalement, l'approche hybride, en sage fusion des méthodes symboliques et numériques, explore le meilleur des deux mondes. Elle marie subtilement les méthodes traditionnelles basées sur les règles grammaticales avec les puissantes capacités des modèles statistiques. Cette union harmonieuse promet des avancées significatives dans l'analyse syntaxique en offrant une compréhension plus profonde et nuancée des langues, y compris le dialecte algérien.

Dans le cadre de cette recherche, notre objectif est d'introduire une méthode novatrice d'analyse syntaxique spécifiquement adaptée au dialecte algérien. En alliant la richesse des règles linguistiques propres à ce dialecte avec la précision des modèles statistiques, notre approche vise à surmonter les défis uniques que présente l'analyse syntaxique de cette langue. Dans les sections à venir, nous détaillerons notre méthodologie, depuis la collecte et la préparation des données jusqu'à l'évaluation minutieuse de notre approche. Cette méthode, à la croisée des chemins entre tradition

et modernité, ouvre de nouvelles perspectives passionnantes dans le domaine en constante évolution du traitement automatique du langage naturel.

4.4.1 Collection et préparation des données

Au cœur de notre méthodologie d'analyse syntaxique du dialecte algérien réside la qualité des données que nous utilisons. Notre source principale est le corpus ALGDC, minutieusement présenté dans le chapitre 2 de ce document. Cette précieuse ressource, spécialement élaborée pour capturer les subtilités et les particularités du dialecte algérien, a été le point de départ essentiel de notre étude.

Avant d'entrer dans le vif du sujet, nous avons entrepris un processus méticuleux de nettoyage et de prétraitement des données. Cette étape cruciale visait à garantir la cohérence, la qualité et la fiabilité de nos données. Tout d'abord, nous avons écarté les données aberrantes et les incohérences, garantissant ainsi un ensemble homogène. Ensuite, nous avons normalisé le texte, harmonisant les diverses variantes linguistiques et les caractéristiques spécifiques au dialecte algérien. Ce processus, rigoureux et systématique, a contribué à créer un corpus homogène et de haute qualité, formant ainsi la base solide sur laquelle repose notre analyse syntaxique.

De plus, nous avons procédé à une étape essentielle de tokenisation, découpant les textes en unités linguistiques significatives, que ce soit des mots, des phrases ou des paragraphes, facilitant ainsi l'analyse et la compréhension. Le découpage des phrases a été effectué à l'aide de notre méthode de segmentation du dialecte algérien, présentée en détail dans le chapitre 3. Les données ainsi préparées sont devenues notre matériau brut, prêt à être soumis à notre méthode d'analyse syntaxique novatrice.

Cette phase de collecte et de préparation des données, essentiellement humaine dans son exécution, a été guidée par notre souci de rigueur et d'authenticité. La qualité de nos données reflète notre engagement à capturer la richesse du dialecte algérien de manière fidèle et précise, ce qui, à son tour, assure la pertinence et la fiabilité de nos résultats d'analyse syntaxique. Dans les sections suivantes, nous expliquerons en

détail notre méthode d'analyse, construite sur cette base solide de données préparées avec soin et expertise.

4.4.2 Création du Treebank dédié au dialecte algérien

L'élaboration d'un Treebank spécifiquement adapté au dialecte algérien représente un jalon essentiel de notre méthodologie. Cette étape, divisée en deux phases distinctes, vise à fournir un ensemble de données annotées et structurées, permettant ainsi d'alimenter notre analyse syntaxique avec des informations linguistiques précieuses et fiables.

4.4.2.1 Phase semi-automatique : modélisation basée sur le dialecte tunisien

Dans un premier temps, nous avons opté pour une approche semi-automatique. Cette phase débute par la création manuelle d'un Treebank fondé sur l'analyseur du dialecte tunisien. Notre choix s'est porté sur le dialecte tunisien en raison de sa proximité avec le dialecte algérien, facilitant ainsi la création d'un modèle apte à générer automatiquement le dialecte algérien. De plus, l'analyseur du dialecte tunisien a été développé au sein de notre laboratoire Miracl¹³, nous procurant un accès privilégié à ses fonctionnalités.

L'objectif principal de cette étape est de concevoir un modèle robuste, capable de générer automatiquement la structure syntaxique du dialecte algérien. En combinant l'expertise linguistique avec les données du dialecte tunisien, nous créons un socle fiable pour la génération ultérieure du Treebank algérien.

4.4.2.2 Phase automatique : construction du Treebank algérien

Dans cette phase cruciale de notre méthodologie, les résultats obtenus lors de la première étape deviennent notre point de départ. Suite à la création manuelle ou semi-automatique de Treebank d'une partie de notre corpus ALGDC, effectuée grâce

13. <http://www.miracl.rnu.tn/>

à la méthode d'analyse syntaxique du dialecte tunisien, et après la correction et la validation méticuleuses par nos experts, nous utilisons le modèle généré lors de la première étape.

Ce modèle devient le moteur de notre méthode automatisée. Il nous permet d'entraîner notre système à analyser syntaxiquement le reste de notre corpus. Cette démarche aboutit à la création d'un vaste Treebank pour le Dialecte Algérien (DALG), minutieusement corrigé et validé. Cette base de données étendue et précise représente le résultat tangible de notre travail, prête à être explorée et exploitée dans diverses applications avancées du traitement automatique du langage naturel.

4.4.3 Architecture de la méthode

Notre approche pour développer un analyseur syntaxique pour le dialecte algérien repose sur un processus méticuleux et stratégique. Dans cette section, nous détaillerons les étapes clés de notre méthodologie intégrée, fusionnant l'expertise linguistique avec les avancées technologiques pour aboutir à un outil d'analyse syntaxique précis et fiable.

Nous commençons par le Nettoyage et le Prétraitement des Données, où les données extraites du corpus ALGDC subissent un processus rigoureux. Ce nettoyage, supervisé par des experts du domaine, élimine les imperfections et prépare les données pour l'analyse syntaxique.

Ensuite, nous abordons la Création du Treebank, un processus en deux phases intégrées. Dans la première phase semi-automatique, nous combinons l'expertise humaine avec l'assistance de l'analyseur du dialecte tunisien pour créer manuellement une partie du Treebank. Cette phase capitalise sur les similitudes linguistiques entre les dialectes tunisien et algérien, assurant une compréhension approfondie des nuances linguistiques spécifiques à l'algérien.

Dans la deuxième phase entièrement automatisée, les modèles et les algorithmes développés dans la première phase sont déployés pour créer le reste du Treebank

spécifique au dialecte algérien. Cette automatisation assure une analyse syntaxique cohérente et précise sur l'ensemble du corpus ALGDC.

L'ensemble de notre méthodologie est ancré dans un équilibre délicat entre l'expertise humaine et l'efficacité des technologies de pointe. Cette synergie entre le savoir-faire linguistique et l'innovation technologique garantit la création d'un outil d'analyse syntaxique robuste, capable de démystifier les subtilités du dialecte algérien. Dans les sections suivantes, nous explorerons les résultats fascinants de cette méthodologie, mettant en lumière son efficacité dans le domaine complexe du traitement automatique du langage naturel.

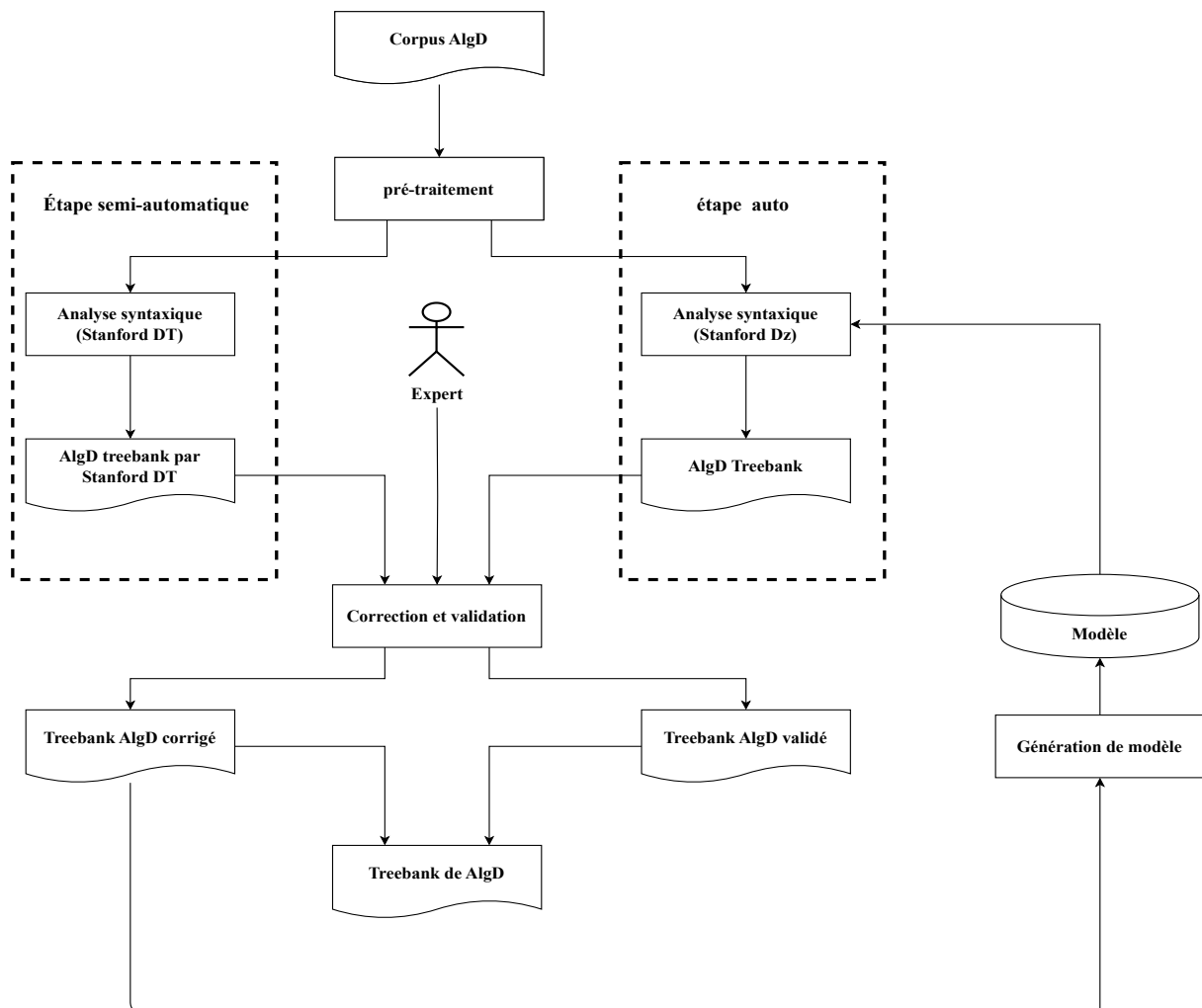


FIGURE 7. méthode proposé pour la création d'un analyseur syntaxique du dialecte algérien

En analysant la figure 7, nous pouvons préciser le cheminement du corpus ALGDC à travers notre méthodologie. Initialement, le corpus ALGDC, composé de

diverses phrases en dialecte algérien, est soumis à deux voies distinctes : l'étape semi-automatique et l'étape automatique.

À l'étape semi-automatique, une partie du corpus est acheminée vers le parseur du dialecte tunisien (Stanford DT). Ce parseur segmente les mots de manière syntaxique, fournissant ainsi une base pour créer le modèle utilisé dans le traitement automatisé. Cependant, cette étape nécessite une correction minutieuse des résultats pour garantir la précision et la cohérence, le tout supervisé par des experts du domaine.

Dans l'étape automatique, le modèle généré à partir de la première étape est employé pour analyser le reste du corpus via le parseur pour le dialecte algérien, Stanford Dz. Ce parseur automatique crée le Treebank pour le dialecte algérien de manière autonome.

Une fois ces étapes d'analyse syntaxique achevées, le corpus subit une étape de correction et de validation. Sous la supervision attentive d'experts, cette phase implique la rectification du Treebank résultant de l'étape semi-automatique, passé précédemment par le parseur du dialecte tunisien. De plus, le Treebank généré automatiquement lors de l'étape précédente est également scruté et validé. Ce processus a été mené avec un taux d'accord remarquable de 0,81 entre les experts, soulignant ainsi la fiabilité de notre méthodologie.

Cette démarche méthodique, mêlant habilement l'intervention humaine et la puissance des technologies de pointe, est le pilier de notre succès dans la création d'un analyseur syntaxique robuste pour le dialecte algérien. Ces résultats seront exposés et explorés en profondeur dans les sections suivantes, mettant en évidence l'efficacité et la précision de notre méthodologie innovante dans le domaine complexe du traitement automatique du langage naturel.

4.4.3.1 Présentation et conception du Treebank

Comme évoqué précédemment, le treebank occupe une place centrale dans notre analyse syntaxique. Il se matérialise en une collection de phrases décomposées en

arbres syntaxiques, révélant ainsi la structure interne des phrases et les relations entre les mots. Ces treebanks sont des outils essentiels pour des tâches complexes telles que l'analyse syntaxique automatique, la génération automatique de phrases, et l'entraînement de modèles d'apprentissage automatique en grammaire. Dans notre travail, nous avons créé notre treebank selon deux méthodes distinctes, mettant l'accent sur la précision et la fiabilité.

La création d'un treebank pour l'analyse syntaxique est un processus délicat exigeant une minutie exceptionnelle. La première étape implique l'utilisation de l'analyseur Stanford Parser spécialement conçu pour le dialecte tunisien. Cet outil génère des arbres syntaxiques pour les phrases du corpus. Cependant, il est important de noter que cette étape initiale peut parfois engendrer des erreurs ou des incohérences. C'est ici que notre deuxième étape intervient : la correction de ces erreurs et incohérences grâce à l'expertise de spécialistes du domaine. Ce processus garantit que notre treebank atteint un niveau de qualité optimal, prêt à être utilisé pour des analyses syntaxiques automatiques du dialecte algérien.

Il est crucial de souligner que l'analyseur Stanford Parser que nous utilisons est spécifiquement calibré pour générer des arbres syntaxiques du dialecte tunisien. Il prend en compte les subtilités syntaxiques et grammaticales propres à cette variante de la langue arabe. Cette spécialisation assure des résultats précis et fiables, surpassant ceux que pourrait offrir un analyseur généraliste.

Une fois le treebank créé et minutieusement corrigé, il devient un outil essentiel pour les analyses syntaxiques automatiques du dialecte algérien. Cette ressource permet d'accélérer les études linguistiques et les recherches sur le dialecte tunisien, augmentant ainsi l'efficacité des chercheurs. De plus, le treebank constitue une base solide pour l'entraînement de modèles de traitement automatique du langage, améliorant ainsi notre compréhension de la syntaxe spécifique à cette variante linguistique.

Après avoir créé le treebank pour la première partie du corpus, nous avons extrait le modèle, facilitant ainsi l'analyse syntaxique automatique de la seconde partie. Le parseur Stanford suggère des ajustements pour les corpus d'entrée, simplifiant ainsi le processus d'adaptation lors de la phase suivante. Ces exigences d'entrée précises, conjuguées à une sortie normalisée du parseur, simplifient grandement le travail des chercheurs. Pour illustrer ce processus, nous présentons un exemple du treebank d'une phrase après son analyse par le parseur pour le dialecte tunisien, suivi du résultat après correction.

Dans l'exemple suivant, nous avons utilisé la phrase "عرضونا للعشا و كان لعشا يكفي ليزانفتي قاع", qui signifie "Nous avons été invités à dîner, et il y avait assez de nourriture pour tous les invités". Après les étapes de prétraitement, la phrase devient "عرضو نا ل لعشا و كان لعشا يكفي قاع ليزانفتي".

```
(ROOT
(S
(VP ( VBP عرضو
(NP (PRP$ نا))
(PP (IN ل))
(NP (DTNN لعشا))))
(CC و)
(S
(PRT (RP كان))
(VP (VBP لعشا)
(S
(VP (VBP يكفي)
(NP (NN ليزانفتي)
(JJ قاع))))))))))
```

FIGURE 8. Treebank après l'analyseur syntaxique du dialecte tunisien

```
(ROOT
(S
(S
(VP (VBD عرضو)
(NP (PRP$ نا))
(PP (IN ل))
(NP (DTNN لعشا))))))
(CC و)
(S
(VP (VBD كان)
(NP (DTNN لعشا)
(VP (VBP يكفي))))
(NP (DTNNS ليزانفتي)
(NP (NOUN_QUANT قاع))))))
```

FIGURE 9. Treebank après correction pour le dialecte algérien.

Les figures 8 et 9 présentent un exemple de corrections du treebank de phrases du dialecte tunisien au dialecte algérien.

Initialement, la phrase a été divisée en deux parties, séparées par la conjonction (الواو), lors de la tokenisation de la phrase avec l'étiquette S. De plus, le mot (كان) a été corrigé, passant d'une particule à un verbe au passé, car il s'agit d'un verbe incomplet (فعل ناقص), et non d'une préposition ou d'une particule (حرف جر). Le libellé a été modifié de (PRT (RP كان)) à (VP (VBD كان)), avec VP représentant une phrase verbale, VBD représentant un verbe au passé, RP une particule et PRT représentant le label parent de RP.

Une autre modification a été apportée aux étiquettes pour le mot (لعشا/ dîner). À l'origine, l'analyse syntaxique indiquait que le mot était un verbe, ce qui est incorrect car il s'agit d'un nom singulier défini par un déterminant. La correction a consisté à remplacer (VP (VBP لعشا)) par (NP (DTNN لعشا)), où NP signifie phrase nominale, DTNN signifie déterminant + nom, singulier, et VBP signifie verbe à l'infinitif présent.

De plus, pour le mot (ليزانفتي), nous avons ajouté le déterminant car il s'agit d'un mot français. Lors de la traduction en arabe, le mot (الضيوف) est utilisé avec un déterminant. Dans l'analyse grammaticale (الإعراب), le mot (الضيوف) est considéré comme le sujet (الفاعل), et comme le verbe est à l'infinitif présent, le sujet est toujours déterminé.

Enfin, le mot (قاع/ tous) semble initialement être un adjectif, mais en réalité, dans le dialecte algérien, il signifie nombre ou quantité. Par conséquent, le libellé a été modifié de (JJ قاع) à (NOUN_QUANT قاع), où NOUN_QUANT signifie quantité nominale et JJ est un adjectif.

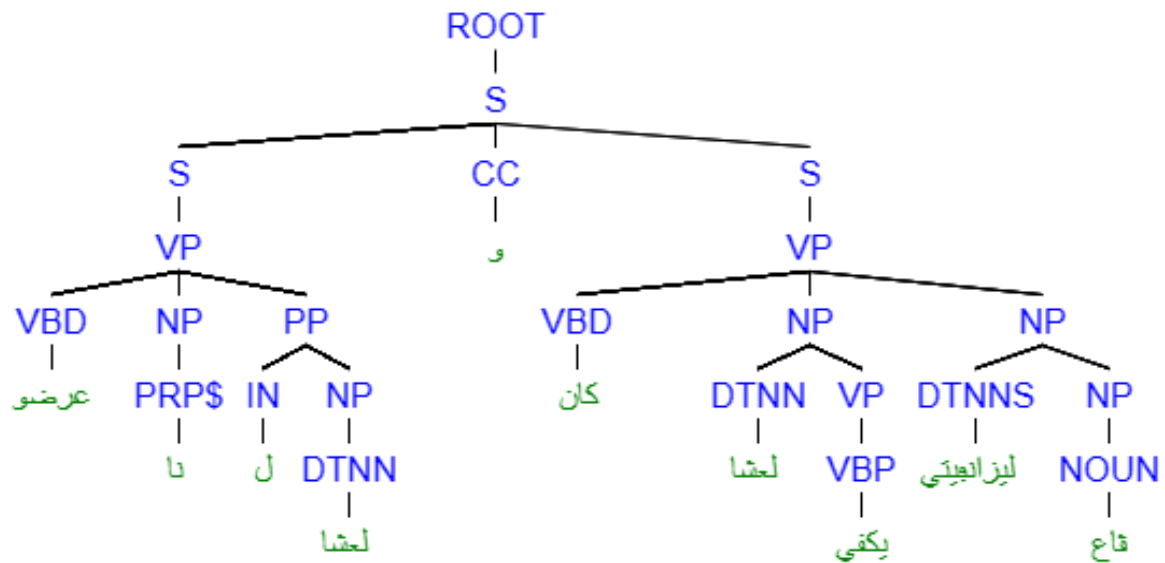


FIGURE 10. Exemple d'un arbre syntaxique pour le dialecte algérien.

Nous avons remarqué qu'il y avait de nombreuses corrections à apporter dans une seule phrase, ce qui rendait la création du Treebank coûteuse en termes de temps et d'argent. Par conséquent, nous avons opté pour une méthode entièrement manuelle pour une partie du corpus, tandis que la plus grande partie est générée automatiquement.

4.5 Réalisation et évaluation

Dans cette section, nous procédons à l'évaluation approfondie de notre méthode d'analyse syntaxique du dialecte algérien, ainsi que de notre système d'analyse, Stanford-DZ, résultant de cette méthodologie innovante. Après avoir exposé notre méthode d'analyse syntaxique, qui repose sur deux étapes.

4.5.1 Réalisation de Stanford-DZ

L'élaboration de notre méthode d'analyse syntaxique pour le dialecte algérien a été un processus complexe et méticuleux. Au cœur de cette entreprise, se trouve la création de Stanford-DZ, notre propre système d'analyse syntaxique spécialement conçu pour

répondre aux subtilités linguistiques du dialecte algérien. Cette réalisation innovante a exigé une réflexion approfondie et une expertise pointue. Nous avons traversé deux étapes cruciales pour donner vie à Stanford-DZ, notre outil phare dans cette entreprise ambitieuse.

Dans un premier temps, nous avons entrepris l'adaptation du Stanford parser, une plateforme en open-source reconnue pour sa flexibilité et son potentiel d'ajustement. Cette démarche nous a permis de personnaliser le système en fonction des spécificités linguistiques du dialecte algérien. L'aspect modifiable de ce parser nous a offert la liberté nécessaire pour l'ajuster précisément au contexte algérien, un élément essentiel pour garantir des analyses syntaxiques précises et pertinentes. Notre choix stratégique du Stanford parser s'est avéré fructueux, non seulement en raison de sa souplesse d'utilisation, mais aussi grâce à sa validation préalable au sein de notre Laboratoire Miracl, lui conférant ainsi une crédibilité scientifique indéniable.

Dans cette entreprise d'adaptation, la version tunisienne du Stanford parser a joué un rôle clé. Son utilisation judicieuse nous a permis de créer un Treebank spécifique au dialecte algérien. Cet ardu processus impliquait la correction minutieuse des résultats issus de l'analyse du corpus algérien par le Stanford-TUN, garantissant ainsi l'intégrité et la précision de notre Treebank. La collaboration entre la flexibilité du Stanford parser et notre expertise linguistique a abouti à la naissance de Stanford-DZ, un système d'analyse syntaxique taillé sur mesure pour le dialecte algérien. Ce travail rigoureux et novateur sert de fondement solide à notre évaluation et à notre discussion approfondie des résultats obtenus.

4.5.2 Évaluation de Stanford-DZ

Pour évaluer rigoureusement notre système d'analyse syntaxique du dialecte algérien, nous avons adopté une méthodologie comparative ingénieuse. Face à l'absence d'accès à d'autres analyseurs syntaxiques spécifiques au dialecte algérien, nous avons soumis une partie de notre corpus du dialecte algérien non analysé

à des tests exhaustifs. Trois analyseurs syntaxiques distincts ont été utilisés : l'analyseur syntaxique de l'arabe proposé par Stanford, une ressource open source universellement reconnue, l'analyseur syntaxique du dialecte tunisien développé au sein de notre laboratoire, Miracl, et enfin, notre propre Stanford-DZ, conçu spécifiquement pour le dialecte algérien.

Nous avons calculé trois mesures d'évaluation fondamentales pour évaluer la performance de Stanford-DZ : le rappel, la précision et la F-mesure. Ces indicateurs essentiels nous ont permis de comprendre la capacité de notre système à identifier précisément les structures syntaxiques dans le dialecte algérien. Les résultats détaillés de cette évaluation méticuleuse sont résumés de manière claire et concise dans le Tableau 27, offrant ainsi un aperçu instantané des performances de Stanford-DZ.

Cette approche comparative nous a offert une perspective multidimensionnelle sur les capacités de notre système. En mettant en parallèle les résultats obtenus avec ces trois outils, nous avons pu évaluer les performances de Stanford-DZ, identifier ses points forts et définir les domaines nécessitant des améliorations. Cette méthodologie a validé l'efficacité de Stanford-DZ et renforcé notre confiance en sa capacité à traiter les spécificités complexes du dialecte algérien.

Cette étude comparative approfondie a enrichi notre compréhension des nuances subtiles du dialecte algérien, tout en validant l'efficacité de Stanford-DZ dans un contexte diversifié. Les résultats captivants de cette évaluation serviront de base solide pour notre discussion ultérieure, où nous explorerons en détail les implications et les découvertes de notre méthodologie d'analyse syntaxique novatrice.

	Rappel	Précision	F-mesure
Stanford-ASM	32.21	39.56	35.51
Stanford-TUN	63.48	65.43	64.44
Stand-DZ	71.19	72.33	71.76

TABLE 27. *Résultats de l'évaluation de Stanford-DZ*

4.5.3 Discussion des résultats

Dans l'évaluation de notre système d'analyse syntaxique du dialecte algérien, nous avons comparé les performances de Stanford-DZ avec deux autres analyseurs syntaxiques : Stanford-ASM et Stanford-TUN. Les résultats, présentés dans le Tableau 27, mettent en lumière les différences significatives dans les performances des trois systèmes.

Stanford-ASM, bien qu'étant un outil reconnu, affiche un rappel relativement bas de 32.21%, indiquant que seulement environ un tiers des structures syntaxiques correctes ont été identifiées. La précision et la F-mesure de Stanford-ASM sont légèrement meilleures, mais elles restent en dessous des 40%, ce qui signifie que de nombreuses prédictions sont incorrectes.

Stanford-TUN montre une amélioration significative avec un rappel de 63.48%, indiquant une identification plus précise des structures syntaxiques. La précision et la F-mesure de Stanford-TUN sont également nettement meilleures que celles de Stanford-ASM, dépassant les 65%. Cependant, il existe encore des marges d'amélioration.

En revanche, notre système, Stanford-DZ, se démarque avec des performances impressionnantes. Il affiche un rappel de 71.19%, dépassant les deux autres systèmes. La précision de Stanford-DZ est également la plus élevée à 72.33%, ce qui se traduit par une F-mesure robuste de 71.76%. Ces résultats soulignent l'efficacité de Stanford-DZ dans l'identification précise des structures syntaxiques du dialecte algérien.

Cette comparaison met en évidence l'importance de notre approche personnalisée. En adaptant un analyseur syntaxique existant au contexte spécifique du dialecte algérien, nous avons réussi à surpasser les systèmes généralistes. Ces résultats prometteurs encouragent notre confiance dans l'efficacité de Stanford-DZ pour l'analyse syntaxique du dialecte algérien, ouvrant ainsi la voie à des applications futures dans des domaines variés du traitement automatique du langage naturel.

4.5.4 Perspectives

L'analyseur syntaxique que nous avons développé, Stanford-DZ, représente une avancée significative dans le domaine de l'analyse du dialecte algérien. Fort de ses performances exceptionnelles, ce système ouvre la voie à diverses perspectives et opportunités de recherche passionnantes.

1. Amélioration continue : Nous envisageons d'investir dans la recherche continue pour améliorer la précision et la couverture de Stanford-DZ. Cela implique l'exploration de nouvelles méthodes d'analyse syntaxique, l'intégration de l'apprentissage automatique pour une meilleure adaptation aux spécificités du dialecte algérien, ainsi que l'optimisation des ressources lexicales et grammaticales.
2. Expansion vers d'autres dialectes : Bien que notre analyseur soit spécifique au dialecte algérien, nous chercherons à étendre son applicabilité à d'autres dialectes arabes. Cette expansion nécessitera une recherche approfondie sur les particularités linguistiques de chaque dialecte et l'ajustement de notre système en conséquence.
3. Applications pratiques : Nous explorons également les applications pratiques de Stanford-DZ dans divers domaines, notamment la traduction automatique, l'assistance à la rédaction, la correction grammaticale et même l'enseignement des langues. L'intégration de Stanford-DZ dans ces domaines peut considérablement améliorer la qualité des résultats et l'efficacité des processus.
4. Collaboration internationale : Nous sommes ouverts à des collaborations internationales avec d'autres chercheurs et institutions travaillant sur des langues et des dialectes similaires. Ces collaborations peuvent favoriser l'échange de connaissances, les comparaisons interculturelles et la création de systèmes d'analyse syntaxique plus robustes et polyvalents.
5. Sensibilisation et éducation : En plus de notre travail de recherche, nous nous engageons à sensibiliser le public à l'importance de la préservation des dialectes et des langues régionales. Nous prévoyons d'organiser des ateliers, des séminaires et des

programmes éducatifs pour sensibiliser les communautés à la richesse linguistique et culturelle du dialecte algérien et d'autres dialectes similaires.

Ces perspectives illustrent notre engagement continu envers le développement de Stanford-DZ et sa contribution positive à la compréhension et à l'analyse du dialecte algérien, ainsi qu'à l'enrichissement global du domaine de la linguistique computationnelle.

4.6 Conclusion

Dans ce chapitre, nous avons plongé dans l'analyse syntaxique du dialecte algérien et de la langue arabe standard, deux domaines complexes et cruciaux pour comprendre les nuances linguistiques au sein des communautés algériennes. L'analyse syntaxique, qui explore la structure grammaticale des phrases, revêt une importance particulière dans le contexte du dialecte algérien en raison de ses variations régionales et de ses différences phonétiques, lexicales et syntaxiques.

Notre approche d'analyse syntaxique se distingue par son caractère méthodique et précis. Divisée en deux étapes principales, notre méthodologie combine habilement l'expertise manuelle et les capacités automatisées. La première étape, semi-automatique, repose sur la création manuelle d'un Treebank à partir de l'analyseur du dialecte tunisien. Cette démarche sert de base solide pour la modélisation ultérieure du dialecte algérien, permettant ainsi la génération automatique du Treebank dans la deuxième étape.

La deuxième phase, entièrement automatique, est le cœur de notre méthodologie. À l'aide du modèle élaboré lors de la première étape, nous avons automatisé le processus de création du Treebank pour le dialecte algérien. Cette étape a impliqué des analyses morphologiques approfondies, la segmentation en phrases et l'identification des structures syntaxiques. Notre corpus ALGDC, préalablement préparé avec soin

à travers des étapes de validation, de normalisation et de segmentation, a été le fondement de notre travail.

Notre méthodologie ne s'arrête pas là. Une étape cruciale de correction et de validation, supervisée par des experts du domaine, a été intégrée pour garantir la qualité et l'exactitude de notre Treebank. Cette rigueur a été attestée par un remarquable taux d'accord de 0,81 entre les experts, soulignant ainsi la robustesse de notre approche.

En résumé, notre chapitre nous a guidés à travers les méandres de l'analyse syntaxique du dialecte algérien. En combinant méthodologies manuelles et automatisées, nous avons réussi à surmonter les complexités linguistiques, offrant ainsi un outil précieux pour les études linguistiques avancées, la recherche et la compréhension approfondie du dialecte algérien dans son contexte régional et national. Ce travail ouvre des portes pour de futures applications dans le domaine du traitement automatique du langage naturel, renforçant ainsi notre compréhension de la richesse linguistique de l'Algérie.



Conclusion et perspectives

La langue arabe et le dialecte algérien jouent un rôle central dans la société algérienne, étant utilisés quotidiennement dans les interactions informelles et les médias sociaux. Cependant, le traitement automatique de ces langues pose des défis en raison de leurs particularités linguistiques. Au cours de cette thèse, nous avons exploré en détail la langue arabe et le dialecte algérien, ainsi que notre contribution à la création de ressources et d'outils pour le traitement automatique du dialecte algérien.

Le premier chapitre de notre thèse était consacré à une étude approfondie de la langue arabe et du dialecte algérien. Nous avons examiné leurs caractéristiques, leur évolution historique et leur utilisation dans la société algérienne.

Dans ce chapitre, nous avons commencé par présenter la langue arabe en tant que l'une des langues les plus anciennes et complexes, avec une riche grammaire et un vaste vocabulaire. Cependant, nous avons également souligné certaines de ses lacunes, telles que l'absence de voyelles et de signes de ponctuation réguliers, ce qui rend la langue arabe difficile à traiter automatiquement.

Ensuite, nous avons exploré l'histoire de la langue arabe en Algérie, en mettant l'accent sur les différentes couches linguistiques qui coexistent dans le pays. Ces couches comprennent l'arabe classique, l'arabe standard moderne et les dialectes arabes, dont le dialecte algérien. Nous avons noté l'importance du dialecte algérien

dans la vie quotidienne des Algériens, utilisé dans les communications informelles et les médias sociaux, bien qu'il ne soit pas enseigné dans les écoles.

Dans notre recherche historique, nous avons également examiné les influences des différentes civilisations qui ont contribué à la formation du dialecte algérien, notamment les influences berbères, arabes, ottomanes et françaises. Cette richesse culturelle a façonné le dialecte algérien et lui a donné ses particularités uniques.

Par la suite, nous avons abordé les différences entre la langue arabe standard et le dialecte algérien en termes de phonologie, de morphologie et d'orthographe. Les différences phonologiques comprennent le système vocalique et consonantique distinct du dialecte algérien par rapport à la langue arabe standard. Les différences morphologiques se manifestent dans la conjugaison des verbes et la négation.

Nous avons également souligné les particularités du dialecte algérien en matière de nombres et de genres, ainsi que ses spécificités syntaxiques. Ces différences rendent le dialecte algérien unique et lui confèrent une identité linguistique forte dans le contexte des dialectes arabes.

En résumé, le premier chapitre de notre thèse nous a permis d'acquérir une meilleure compréhension de la langue arabe et du dialecte algérien, ainsi que de leur histoire et de leur utilisation dans la société algérienne. Nous avons pris conscience de la complexité de la langue arabe et des particularités uniques du dialecte algérien, ce qui nous a préparés pour les chapitres suivants de notre recherche sur la création de ressources et les outils pour le traitement automatique du dialecte algérien.

Le deuxième chapitre de notre thèse a été dédié à la constitution du corpus de données pour le dialecte algérien et à la méthodologie de collecte des données. Nous avons souligné l'importance d'un corpus bien préparé et assemblé correctement pour garantir de bons résultats dans le traitement automatique.

Dans ce chapitre, nous avons mis en évidence la signification d'un corpus et son rôle essentiel dans le développement de modèles de traitement automatique du langage. Pour notre étude sur le dialecte algérien, nous avons pris en compte plusieurs corpus

et recherches existantes sur la langue arabe, en citant certaines des plus pertinentes. Parmi les corpus arabes les plus connus, nous avons relevé le PADIC (Arabic Dialects Parallel Web Corpus) et le MADAR (Multilingual Arabic Dialect Applications and Resources).

Notre méthodologie de collecte de données pour le corpus du dialecte algérien s'est déroulée en deux étapes principales :

Collecte de données orales : Nous avons enregistré directement des locuteurs natifs du dialecte algérien dans divers contextes conversationnels, et nous avons également recueilli des enregistrements audio existants à partir de sources disponibles en ligne.

Collecte de données écrites : Nous avons collecté des textes en langue arabe écrite représentant le dialecte algérien à partir de sources diverses, notamment les médias sociaux et des bases de données existantes.

Par la suite, nous avons expliqué en détail le processus de transcription des données orales, qui consiste à convertir les enregistrements audio en texte écrit en suivant une convention spécifique, dans notre cas, la convention CODA du dialecte algérien. Le processus de normalisation et de nettoyage des données a également été abordé, afin de rendre les données prêtes pour le traitement automatique.

Le chapitre a également présenté des statistiques sur notre corpus en termes de nombre total de phrases et de mots, ainsi que la diversité des sujets et des sources utilisées pour collecter les données orales et écrites. Enfin, nous avons fourni un diagramme décrivant notre méthodologie de collecte de données.

En résumé, le deuxième chapitre a mis en lumière l'importance d'un corpus bien constitué et de données de haute qualité pour la recherche sur le dialecte algérien et le traitement automatique de ce dernier. Notre méthodologie de collecte de données a permis de rassembler un corpus riche et diversifié, prêt à être utilisé pour nos travaux de recherche dans les chapitres suivants.

Dans le troisième chapitre, nous avons souligné l'importance de la segmentation dans le contexte du dialecte algérien et présenté une méthode basée sur des règles linguistiques pour la segmentation des textes.

La segmentation des textes dans les dialectes arabes, notamment le dialecte algérien, présente des défis en raison des particularités propres à ces dialectes et de l'absence de règles linguistiques claires. Notre approche consistait à extraire manuellement des règles à partir d'un corpus de texte en dialecte algérien, en les enrichissant également avec certaines règles traduites du dialecte tunisien.

En combinant deux méthodes, STAR et SegDT, qui utilisent des règles linguistiques et une approche contextuelle, nous avons développé un système de segmentation robuste et efficace pour le dialecte algérien, que nous avons appelé SegDz. Ce système a été évalué en comparant ses performances avec d'autres méthodes de segmentation telles que STAR et SegDT, ainsi qu'avec des règles contextuelles spécifiques au dialecte algérien traduites du dialecte tunisien.

Les mesures de rappel, de précision et d'exactitude ont été utilisées pour évaluer les performances de SegDz. Malgré les améliorations apportées, des faiblesses ont été identifiées, nécessitant des améliorations supplémentaires pour augmenter la précision et la pertinence des résultats. Cependant, notre système de segmentation SegDz constitue une avancée significative dans le traitement automatique du dialecte algérien, offrant des résultats prometteurs.

La segmentation des textes en dialecte algérien représente un défi important dans le domaine du traitement automatique des langues. Grâce à notre méthode basée sur des règles linguistiques et une approche contextuelle, nous avons pu développer le système SegDz, qui offre une segmentation robuste et efficace pour le dialecte algérien. Toutefois, des améliorations continues sont nécessaires pour surmonter les limitations et fournir une segmentation précise dans le dialecte algérien, offrant ainsi une meilleure expérience utilisateur dans les applications de traitement automatique

des langues. Ce travail ouvre également des perspectives pour la recherche future dans le domaine de la segmentation des textes dans les dialectes arabes.

L'analyse syntaxique est essentielle pour comprendre la structure grammaticale des phrases dans une langue donnée. Dans Le quatrième chapitre, nous avons abordé l'analyse syntaxique du dialecte algérien, qui s'avère être plus complexe que celle de la langue arabe standard, en raison de ses variations régionales et de ses particularités phonétiques, lexicales et syntaxiques.

Notre approche pour l'analyse syntaxique du dialecte algérien s'est décomposée en deux étapes majeures : une étape semi-automatique et une étape automatique. Dans un premier temps, nous avons créé manuellement un Treebank en nous basant sur l'analyseur du dialecte tunisien, ce qui nous a permis de développer un modèle pour générer automatiquement des structures pour le dialecte algérien. Cette première étape nous a demandé un effort important, mais elle s'est avérée cruciale pour construire une base solide pour l'analyse syntaxique automatique du dialecte algérien.

La deuxième étape a utilisé le modèle développé pour générer automatiquement un Treebank spécifique au dialecte algérien. Ce processus a impliqué plusieurs sous-étapes telles que l'analyse morphologique, la segmentation en phrases et l'identification de la structure syntaxique. Notre corpus d'entrée, le corpus ALGDC, a été soigneusement préparé et traité pour garantir des résultats cohérents et de haute qualité.

Cependant, l'analyse syntaxique du dialecte algérien n'est pas sans difficultés. Les variations régionales et les particularités linguistiques rendent le processus complexe et exigent une approche minutieuse. Malgré les efforts déployés pour améliorer l'efficacité de notre modèle, des limitations ont été identifiées lors de l'évaluation. Cela souligne la nécessité d'une attention continue à la recherche de solutions plus performantes pour l'analyse syntaxique du dialecte algérien. Ce chapitre sur l'analyse syntaxique a constitué une étape importante dans notre thèse pour comprendre comment les phrases sont structurées dans le dialecte algérien par rapport à la langue

arabe standard. Nous avons réussi à développer un modèle semi-automatique pour générer des structures syntaxiques spécifiques au dialecte algérien, mais il est évident qu'il reste des défis à relever. Ce travail ouvre la voie à des recherches futures visant à améliorer l'analyse syntaxique du dialecte algérien et à renforcer les capacités du traitement automatique des langues dans ce contexte spécifique.

Dans l'ensemble, notre recherche a contribué à une meilleure compréhension du dialecte algérien et à la création de ressources et d'outils pour son traitement automatique. Nous avons mis en évidence les particularités du dialecte algérien par rapport à la langue arabe standard, ainsi que les défis spécifiques liés à sa segmentation et à son analyse syntaxique.

Ces travaux ouvrent la voie à de futures recherches et applications dans le domaine du traitement automatique des langues, visant à améliorer la précision et la pertinence des outils de traitement du dialecte algérien. Notre contribution permettra de faciliter la communication et l'analyse de contenu dans le dialecte algérien, contribuant ainsi au développement des technologies linguistiques adaptées à cette variante linguistique spécifique.



Liste des publications

[**Abdelhakim Benali**, Mohamed Hédi Maâloul and Lamia Hadrich Belguith, “ Identification of the Algiers dialect using the linguistic rules”, ICCA-TICET-GECO 2018 : Joint Conference on Computing’1-2018 : 12th International Computing Conference on Arabic, 5th International Conference on ICT in Education, 1st International Conference on Geocomputation, Tunis, Tunisia, November 9-11, 2018.]

[**Abdelhakim Benali**, Mohamed Hédi Maâloul and Lamia Hadrich Belguith, " Corpus of the Algerian dialect CDalg : linguistic rules of decision", Language Processing and Knowledge Management International Conference (LPKM’2018), FSEGS, Sfax, Tunisia, October 17th - 18th, 2018.]

[**Abdelhakim Benali**, Mohamed Hédi Maaloul and Lamia Hadrish Belguith “Automatic Processing of Algerian Dialect : Corpus Construction and Segmentation, Sn computer Science, 2023. (<https://doi.org/10.1007/s42979-023-02097-1>)]



Bibliographie

- [1] Abdelhakim BENALI et al. "Automatic Processing of Algerian Dialect : Corpus Construction and Segmentation". In : *SN Computer Science* 4.5 (2023), p. 597.
- [2] Brahim DAHOU et al. "DZ-SMS : an authentic corpus of Algerian SMS". In : *ACM Transactions on Asian and Low-Resource Language Information Processing* 22.8 (2023), p. 1-21.
- [3] A MEKKI et al. "Named Entity Recognition of Tunisian Arabic Using the Bi-LSTM-CRF Model". In : *International Journal on Artificial Intelligence Tools* (2023).
- [4] Kawthar Yasmine ZERGAT et al. "The voice as a material clue : a new forensic Algerian Corpus". In : *Multimedia Tools and Applications* (2023), p. 1-19.
- [5] Md MAMUN et al. "Classification of textual sentiment using ensemble technique". In : *SN Computer Science* 3.1 (2022), p. 1-13.
- [6] Amira BARHOUMI. "Une approche neuronale pour l'analyse d'opinions en arabe". Thèse de doct. Le Mans, 2020.
- [7] Mohamed Amine MENACER et al. "Analyse de sentiments des vidéos en dialecte algérien". In : *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*

- (*RÉCITAL*, 22e édition). Volume 2 : *Traitement Automatique des Langues Naturelles*. ATALA ; AFCP. 2020, p. 296-304.
- [8] Anatol STEFANOWITSCH. *Corpus linguistics : A guide to the methodology*. Language Science Press, 2020.
- [9] Adel ABDELLI et al. "Sentiment Analysis of Arabic Algerian Dialect Using a Supervised Method". In : *2019 International Conference on Intelligent Systems and Advanced Computing Sciences (ISACS)*. IEEE. 2019, p. 1-6.
- [10] Salima HARRAT et al. "Machine translation for Arabic dialects (survey)". In : *Information Processing & Management* 56.2 (2019), p. 262-273.
- [11] Nassima KERRAS et al. "L'arabe standard et l'algérien : une approche sociolinguistique et une analyse grammaticale". In : *Íkala* 24.3 (2019), p. 521-535.
- [12] Khyati MAHENDRU. *How to determine the optimal K for K-Means?* en. Juin 2019. URL : <https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-k-means-708505d204eb> (visité le 03/09/2020).
- [13] MOUMNI YAÂKOUB. " : INTEGRATION PHONOLOGIQUE ET MORPHOLOGIQUE DES EMPRUNTS AU FRANÇAIS EN ARABE DIALECTAL ALGERIEN (région des Ziban)". Thèse de doct. faculté des lettres et des langues, 2019.
- [14] Abdel Hakim BENALI et al. "Identification of the Algiers dialect using the linguistic rules". In : *2018 JCCO Joint International Conference on ICT in Education and Training, International Conference on Computing in Arabic, and International Conference on Geocomputing (JCCO : TICET-ICCA-GECO)*. IEEE. 2018, p. 1-5.
- [15] Paul BOERSMA et David WEENINK. "Praat : Doing phonetics by computer (Version 6.0. 37)[Computer software]". In : *Amsterdam : Institute of Phonetic Sciences* (2018).
- [16] Houda BOUAMOR et al. "The madar arabic dialect corpus and lexicon". In : *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 2018.

- [17] Abdeljalil ELOUARDIGHI et al. "Analyse des sentiments à partir des commentaires Facebook publiés en Arabe standard ou dialectal marocain par une approche d'apprentissage automatique." In : *EGC*. 2018, p. 329-334.
- [18] Stefan STIEGLITZ et al. "Social media analytics—Challenges in topic discovery, data collection, and data preparation". In : *International journal of information management* 39 (2018), p. 156-168.
- [19] Wafia ADOUANE et Simon DOBNIK. "Identification of languages in Algerian Arabic multilingual documents". In : *Proceedings of the Third Arabic Natural Language Processing Workshop*. 2017, p. 1-8.
- [20] Ayesha Jumaa Salem AL MUKHAITI et al. "Dataset built for Arabic sentiment analysis". In : *International Conference on Advanced Intelligent Systems and Informatics*. Springer. 2017, p. 406-416.
- [21] Soumia BOUGRINE et al. "Toward aWeb-based Speech Corpus for Algerian Arabic Dialectal Varieties". In : *The Third Arabic Natural Language Processing Workshop (WANLP)*, pages 138–146, Valencia, Spain, 2017.
- [22] Mourad DJELLAB et al. "Algerian Modern Colloquial Arabic Speech Corpus (AMCASC) : regional accents recognition within complex socio-linguistic environments". In : *Language Resources and Evaluation* 51.3 (2017), p. 613-641.
- [23] Imane GUELLIL et al. "Une approche fondée sur les lexiques d'analyse de sentiments du dialecte algérien". In : *Traitement Automatique des Langues* (2017).
- [24] Imène GUELLIL et Faïçal AZOUAOU. "ASDA : Analyseur Syntaxique du Dialecte Alg {\'e} rien dans un but d'analyse s {\'e} mantique". In : *arXiv preprint arXiv :1707.08998* (2017).
- [25] Salima MDHAFFAR et al. "Sentiment analysis of tunisian dialects : Linguistic ressources and experiments". In : *Third Arabic Natural Language Processing Workshop (WANLP)*. 2017, p. 55-61.
- [26] Asma MEKKI et al. "Syntactic Analysis of the Tunisian Arabic." In : *LPKM*. 2017.

- [27] Younes SAMIH et al. "A neural architecture for dialectal Arabic segmentation". In : *Proceedings of the Third Arabic Natural Language Processing Workshop*. 2017, p. 46-54.
- [28] Ibrahim AL-HURI. *Arabic Language : Historic and Sociolinguistic Characteristics*. en. 2016. DOI : 10.13140/RG.2.2.16163.66089/1. URL : <http://rgdoi.net/10.13140/RG.2.2.16163.66089/1>.
- [29] Salima HARRAT et al. "An algerian dialect : Study and resources". In : *International journal of advanced computer science and applications (IJACSA)* 7.3 (2016), p. 384-396.
- [30] M'hamed MATAOUI et al. "A proposed lexicon-based sentiment analysis approach for the vernacular Algerian Arabic". In : *Research in Computing Science* 110 (2016), p. 55-70.
- [31] Anas SHAHROUR et al. "CamelParser : A system for Arabic syntactic analysis and morphological disambiguation". In : *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : System Demonstrations*. 2016, p. 228-232.
- [32] Inès ZRIBI et al. "Sentence boundary detection for transcribed Tunisian Arabic". In : *Bochumer Linguistische Arbeitsberichte* 323 (2016).
- [33] Inès ZRIBI. "Traitement automatique du dialecte tunisien : construction de ressources linguistiques". Thèse de doct. Université de Sfax - Tunisie, 2016.
- [34] Marwa GRAJA. "Compréhension automatique de la parole en dialecte tunisien dans le cadre des systmes de dialogue". Thèse de doct. Phd dissertation. Fsegs, Tunisie, 2015.
- [35] Karima MEFTOUH et al. "Machine translation experiments on PADIC : A parallel Arabic dialect corpus". In : *The 29th Pacific Asia conference on language, information and computation*. 2015.
- [36] Houda SAADANE et Nizar HABASH. "A Conventional Orthography for Algerian Arabic". In : 2015.

- [37] Inès ZRIBI et al. "Spoken Tunisian Arabic corpus "STAC" : transcription and annotation". In : *Research in computing science* 90 (2015), p. 123-135.
- [38] Nawaf ABDULLA et al. "Automatic lexicon construction for arabic sentiment analysis". In : *2014 International Conference on Future Internet of Things and Cloud*. IEEE. 2014, p. 547-552.
- [39] Mohamed ALY et Amir ATIYA. "Labr : A large scale arabic book reviews dataset". In : *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*. 2013, p. 494-498.
- [40] Marwa GRAJA et al. "Discriminative framework for spoken tunisian dialect understanding". In : *International Conference on Statistical Language and Speech Processing*. Springer. 2013, p. 102-110.
- [41] Elsa. "Analyse syntaxique à l'aide des tables du Lexique-Grammaire du français". In : *Linguisticæ Investigationes* 35.1 (2012), p. 147-151.
- [42] Muhammad ABDUL-MAGEED et Mona T DIAB. "AWATIF : A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis." In : *LREC*. T. 515. 2012, p. 3907-3914.
- [43] Nizar HABASH et al. "Conventional Orthography for Dialectal Arabic." In : *LREC*. 2012, p. 711-718.
- [44] Iskandar KESKES et al. "Clause-based Discourse Segmentation of Arabic Texts." In : *LREC*. 2012, p. 2826-2832.
- [45] Mohamed Hedi MAALOUL. "Approche hybride pour le résumé automatique de textes. Application à la langue arabe." Date Soutenance 18-12-2012. Theses. Université de Provence - Aix-Marseille I, déc. 2012. URL : <https://tel.archives-ouvertes.fr/tel-00756111>.
- [46] Alina STOICA et al. "Twitter : Extraction, Regroupement et Visualisation pour la Veille Stratégique". In : *Actes de la conférence Veille Stratégique et Technologique, Ajaccio, France*. 2012.

- [47] Christophe PEREIRA et Stefan WENINGER. "Arabic in the North African region". In : *The Semitic Languages* (2011), p. 954-969.
- [48] Younès BAHOU et al. "Un gestionnaire de dialogue oral arabe Homme-machine : Réalisation et évaluation." In : CORIA. 2010, p. 19-34.
- [49] Spence GREEN et Christopher D MANNING. "Better Arabic parsing : Baselines, evaluations, and analysis". In : *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. 2010, p. 394-402.
- [50] Andreas M KAPLAN et Michael HAENLEIN. "Users of the world, unite! The challenges and opportunities of Social Media". In : *Business horizons* 53.1 (2010), p. 59-68.
- [51] Sid Ahmed SELOUANI et Malika BOUDRAA. *Algerian Arabic speech database (ALGASD) : corpus design and automatic speech recognition application*. 2010.
- [52] Lamia Hadrich BELGUITH. "Analyse et résumé automatiques de documents : Problèmes, conception et réalisation". In : *Habilitation universitaire en informatique, Faculté des Sciences Économiques et de Gestion de Sfax (FSEG-SFAX)* (2009), p. 53.
- [53] Ali FARGHALY et Khaled SHAALAN. "Arabic natural language processing : Challenges and solutions". In : *ACM Transactions on Asian Language Information Processing (TALIP)* 8.4 (2009), p. 1-22.
- [54] Kees VERSTEEGH et al. "ENCYCLIC PEDIA () P ARABIC LANGUAGE AND LINGUISTICS". In : *African Studies* 8 (2009), p. 57-263.
- [55] Yassine BENAJIBA et al. "Arabic named entity recognition : An svm-based approach". In : *Proceedings of 2008 Arab International Conference on Information Technology (ACIT)*. Citeseer. 2008, p. 16-18.
- [56] Tej K BHATIA et William C RITCHIE. *The handbook of bilingualism*. John Wiley & Sons, 2008.
- [57] Zoubeir MOUELHI. "AraSeg.*. : un segmenteur semi-automatique des textes arabes". In : *Actes des JADT 2008* (2008), p. 867-878.

- [58] Lamia Hadrich BELGUITH et al. "MASPAR : De la segmentation à l'analyse syntaxique de textes arabes". In : *CÉPADUÈS-Editions, éditeur, Revue Information Interaction Intelligence I 3* (2007), p. 9-36.
- [59] Abdelhadi SOUDI et al. "Arabic computational morphology : knowledge-based and empirical methods". In : *Arabic computational morphology*. Springer, 2007, p. 3-14.
- [60] L BELGUITH et Nouha CHAABEN. "Analyse et désambiguïsation morphologiques de textes arabes non voyellés". In : *Actes de la 13ème conférence sur le Traitement Automatique des Langues Naturelles* (2006), p. 493-501.
- [61] Heikki PALVA. "Dialects : classification". In : *Encyclopedia of Arabic languages and linguistics, v. 1, A-Ed*. Brill, 2006, p. 604-613.
- [62] Khaoula TALEB IBRAHIMI. "L'Algérie : coexistence et concurrence des langues". In : *L'Année du Maghreb I* (2006), p. 207-218.
- [63] Lamia Hadrich BELGUITH et al. "Segmentation de textes arabes basée sur l'analyse contextuelle des signes de ponctuations et de certaines particules". In : *Actes de la 12ème conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*. 2005, p. 451-456.
- [64] Karin C RYDING. *A reference grammar of modern standard Arabic*. Cambridge university press, 2005.
- [65] Mohamed MAAMOURI et al. "The penn arabic treebank : Building a large-scale annotated arabic corpus". In : *NEMLAR conference on Arabic language resources and tools*. T. 27. Cairo. 2004, p. 466-467.
- [66] Alex WAIBEL et al. "Towards language portability in statistical speech translation". In : *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*. T. 3. IEEE. 2004, p. iii-765.
- [67] Blake D POLAND. "Transcription quality". In : *Handbook of interview research : Context and method* 629 (2002).

- [68] Mohamed ATTIA. "A large-scale computational processor of the Arabic morphology, and applications". In : *Faculty of engineering, Cairo university* (2000).
- [69] Jean-Pierre DESCLÉS. "Systèmes d'exploration contextuelle". In : *Co-texte et calcul du sens* (1997), p. 215-232.
- [70] Jean CARLETTA. "Assessing agreement on classification tasks : the kappa statistic". In : *arXiv preprint cmp-lg/9602004* (1996).
- [71] Philippe MARÇAIS. "Algeria". In : *Leiden, E.J. Brill* (1986.).
- [72] J Richard LANDIS et Gary G KOCH. "The measurement of observer agreement for categorical data". In : *biometrics* (1977), p. 159-174.
- [73] Jacob COHEN. "A coefficient of agreement for nominal scales". In : *Educational and psychological measurement* 20.1 (1960), p. 37-46.
- [74] Charles A FERGUSON. "Diglossia". In : *word* 15.2 (1959), p. 325-340.
- [75] Belkasssem Ben SEDIRA. *Dictionnaire Français Arabe (dialecte d'Algérie)*. Sous la dir. d'adolphe Jourdan librair-editeur ALGER. 1882.
- [76] Alexandre BELLEMARE. *Grammaire arabe (idiome d'Algérie) a l'usage de l'armée et des employés civils de l'Algérie*. A. Dubes, 1865.



Annexe : Étiquettes d'Analyse Syntaxique

Étiquette	Description
ADJ_NUM	Adjectif numérique
ADJP	Adjectif phrase
ADVP	Adverbe phrase
CC	Conjonction de coordination
CD	Nombre cardinal
DTADJ_NUM	Déterminant + adjectif numérique
DTCD	Déterminant + nombre cardinal
DTNN	Déterminant + nom, singulier
DTNNP	Déterminant + nom propre, singulier
DTNNPS	Déterminant + nom propre, pluriel
DTNNS	Déterminant + nom, pluriel
DTNOUN_QUANT	Déterminant + nom quantifier
DTIN	Déterminant + préposition ou conjonction de subordination
DTJJ	Déterminant + adjectif
DTJJR	Déterminant + adjectif comparatif
DTJJS	Déterminant + adjectif superlatif

DTRB	Déterminant + adverbe
DTRP	Déterminant + particule
FRAG	Fragment
IN	Préposition ou conjonction de subordination
INTJ	Interjection (correspond à l'étiquette parente UH
JJ	Adjectif
JJR	Adjectif comparatif
JJS	Adjectif superlatif
NN	Nom, singulier
NNP	Nom propre, singulier
NNPS	Nom propre, pluriel
NNS	Nom, pluriel
NOUN_ QUANT	Nom quantifier
NP	Phrase nominale
PP	Groupe Prépositionnel
PRN	Parenthetical
PRP	Pronom Personnel
PRP\$	Pronom Possessif
PRT	Particule (étiquette parente de RP)
PUNC	Ponctuation
RB	Adverbe
RBR	Adverbe comparatif
RBS	Adverbe superlatif
RP	Particule
S	Phrase
SBAR	Clauses relatives
SBARQ	Clauses relatives interrogatives
SQ	Phrase interrogative

VB	Verbe, forme basique
VBD	Verbe au passé
VBG	gérondif ou participe présent
VCN	Verbe participe passé
VBP	Verbe au présent
VP	Phrase verbale
WHADVP	Wh-adverbe phrase
WHADJP	Wh-adjectif phrase
WHNP	Wh-phrase nominale
WHPP	Wh-phrase prépositionnelle
WP	Wh-pronom
WRB	Wh-adverbe

Cette thèse s'inscrit dans le contexte du développement de la technologie de Traitement Automatique du Langage Naturel, avec une focalisation particulière sur la langue arabe, et plus précisément sur le dialecte algérien. Notre objectif principal consiste à contribuer à la compréhension et à l'efficacité du traitement du dialecte algérien en présentant des ressources et des outils dédiés. Le premier chapitre a exposé la complexité de la langue arabe standard et de ses divers dialectes, mettant en évidence les particularités spécifiques du dialecte algérien, ainsi que les défis linguistiques associés et le manque de règles formelles. Le deuxième chapitre a introduit des ressources spécifiques au dialecte algérien, notamment notre Corpus de données, recueilli à partir de données écrites et parlées, avec une transcription et une normalisation conformes à la convention d'écriture du dialecte algérien. La méthode de segmentation des phrases du dialecte algérien en arabe, basée sur des règles d'exploration contextuelle, a été présentée dans le troisième chapitre. Enfin, le quatrième chapitre a détaillé une méthode de construction d'un analyseur syntaxique pour le dialecte algérien, s'appuyant sur un arbre de décision spécialement créé. Ces contributions fournissent des bases solides pour les futurs travaux de traitement automatique du langage naturel axés sur le dialecte algérien, ouvrant ainsi de nouvelles perspectives dans ce domaine en constante évolution.

Mot clé : TALN, DALG, ALGDC, Segmentation du DALG, Analyse syntaxique du DALG

This thesis is situated in the context of developing Natural Language Processing (NLP) technology, with a specific focus on the Arabic language, particularly the Algerian dialect. Our main goal is to contribute to the understanding and effectiveness of processing the Algerian dialect by presenting dedicated resources and tools. The first chapter exposed the complexity of standard Arabic and its various dialects, highlighting specific characteristics of the Algerian dialect, along with associated linguistic challenges and the lack of formal rules. The second chapter introduced resources specific to the Algerian dialect, including our data corpus collected from written and spoken sources, with transcription and normalization following the writing convention of the Algerian dialect. The method of segmenting sentences in the Algerian dialect written in Arabic, based on contextual exploration rules, was presented in the third chapter. Finally, the fourth chapter detailed a method for constructing a syntactic analyzer for the Algerian dialect, relying on a decision tree specifically created. These contributions provide a solid foundation for future natural language processing work focused on the Algerian dialect, thus opening new perspectives in this constantly evolving field.

Keywords: NLP, Algerian dialect , Algerian corpora, Segmentation, Syntax Analysis.

هذه الرسالة تندرج ضمن سياق تطوير تقنية معالجة اللغة الطبيعية، مع التركيز الخاص على اللغة العربية، وتحديدًا على اللهجة الجزائرية. هدفنا الرئيسي هو المساهمة في فهم وتعزيز فعالية معالجة اللهجة الجزائرية من خلال تقديم موارد وأدوات مخصصة. كشف الفصل الأول عن تعقيد اللغة العربية الفصحى ومتنوعات لهجاتها، مسلطًا الضوء على الخصائص المحددة للهجة الجزائرية، إضافةً إلى التحديات اللغوية المرتبطة ونقص القواعد الرسمية. في الفصل الثاني، تم تقديم موارد مخصصة للهجة الجزائرية، بما في ذلك مجموعة بياناتنا، التي تم جمعها من البيانات المكتوبة والمنطوقة، مع التحويل والتوحيد وفقًا لاتفاقية كودا للهجة الجزائرية. في الفصل الثالث، عُرضت طريقة تقسيم الجمل للهجة الجزائرية في اللغة العربية، استنادًا إلى قواعد استكشاف السياق. وفي النهاية، في الفصل الرابع، تم تفصيل طريقة بناء محلل تركيب للهجة الجزائرية، باستخدام شجرة قرار مصممة خصيصًا. توفر هذه المساهمات أساسًا قوية للأعمال المستقبلية في مجال معالجة اللغة الطبيعية المركزة على اللهجة الجزائرية، وتفتح بذلك آفاقًا جديدة في هذا المجال المتطور باستمرار.

كلمات مفتاحية: المعالجة الآلية للغات الطبيعية، اللهجة الجزائرية، مجموعات بيانات جزائرية، تجزئة نصوص اللهجة الجزائرية، تحليل بناء الجملة

