PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH
ABDELHAMID IBN BADIS UNIVERSITY OF MOSTAGANEM

FACULTY OF FOREIGN LANGUAGES
DEPARTMENT OF ENGLISH

Towards an NLP Model for the Algerian Dialect: Building a Corpus for Machine Translation Processing

A dissertation submitted in partial fulfillment of the requirements
for the degree of **Doctorate LMD** in ***Linguistics***

by:

Abdelbasset Dou

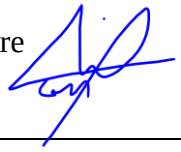
Board of Examiners:

Prof. Hanane Sarnou	University of Mostaganem	Chairperson
Prof. Khalida Kissi	Higher Teacher Training School of Oran	Supervisor
Prof. Nadia Idri	University of Bejaia	Examiner
Prof. Fatma Zaghar	University of Oran2	Examiner
Prof. Lynda Dali Youcef	Higher Teacher Training School of Oran	Examiner
Dr. Sabria Ould Si Bouziane	University of Mostaganem	Examiner

2025-2026

Declaration

I, Abdelbasset Dou, hereby declare that the work presented in this research work has not been submitted for any other degree or professional qualification, and that it is the result of my own independent work.

Signature  Date: 25 Oct 2025

Abstract

Neural Machine Translation (NMT) for Arabic dialects presents a significant challenge within Natural Language Processing due to a critical scarcity of digital resources. This issue is particularly acute for the Maghrebi continuum, where Algerian dialects like Oranian (ORN) are technologically underserved. This dissertation addresses this absolute void by tackling the dual problem of resource absence and the lack of MT benchmarks for the ORN-English language pair. The study details the creation of the novel Oranian-English Parallel Corpus (OEPC), an 8.5K sentence-pair dataset compiled through meticulous processes. OEPC was subsequently used to design and train a Sequence-to-Sequence (Seq2Seq) NMT system. The model's architecture is based on a two-layer stacked Long Short-Term Memory (LSTM) network, critically augmented with an additive attention mechanism. The final model achieves a robust BLEU score of 30.9 on a held-out test set, significantly outperforming an attention-less baseline. Notably, this result is highly competitive, surpassing even larger, semi-supervised systems for other Arabic dialects, which highlights the impact of a high-quality, linguistically consistent corpus. Qualitative error analysis confirms the model's competency with core grammatical structures while identifying classic low-resource challenges related to out-of-vocabulary words, idiomatic expressions, and semantic ambiguity. The findings validate a "depth-over-breadth" corpus strategy as an effective approach for initiating research on zero-resource dialects. This work provides a methodological blueprint for future work pointing towards Transformer architectures and semi-supervised learning.

Keywords: Algerian Arabic, Low-Resource Languages, Neural Machine Translation, OEPC, Oranian Dialect, Sequence-to-Sequence.

Acknowledgements

I would like to express my deepest gratitude to my advisor, *Prof. Khalida Kissi*, for her unwavering guidance, insightful feedback, and constant support throughout my doctoral journey. Your expertise and dedication were instrumental in shaping this dissertation and my growth as a researcher.

My appreciation goes to *Prof. Hanane Sarnou*, the head of DSPM Research Laboratory, for providing a supportive and intellectually stimulating environment.

I also extend my sincere thanks to the board of examiners, *Prof. Nadia Idri*, *Prof. Fatma Zaghar*, *Prof. Lynda Dali Youcef* and *Dr. Sabria Ould Si Bouziane*, for their time, valuable comments, and rigorous review of my work. Your critical insights would help to strengthen this research.

I am profoundly grateful to all the Algerian scholars and researchers in the field of Natural Language Processing whose pioneering work and publications served as a constant source of inspiration and motivation. Your contributions to the field have not only advanced our collective knowledge but also fueled my passion to contribute to this important area of study.

Finally, I want to thank my fellow colleagues and friends, without whose support this endeavor would have been much more difficult. Your camaraderie, encouragement, and willingness to listen provided invaluable support and perspective during challenging times.

Contents

Declaration i

Abstract ii

Acknowledgements iii

List of Tables viii

List of Figures x

List of Abbreviations, Acronyms and Initialisms xii

INTRODUCTION 1

1. Background 2

2. Research Gap 4

3. Motivation 7

4. Statement of the Problem 8

5. Research Objectives 9

6. Research Methods and Design 10

7. Contribution 12

8. Dissertation Structure 14

CHAPTER 1: MACHINE TRANSLATION 16

1. Artificial Intelligence Primer 17

2. Introducing Machine Translation 19

3. History of Machine Translation 20

    3.1. Pioneers and Precursors 22

    3.2. The Georgetown-IBM Experiment 26

3.3. The ALPAC Report 27

4. Machine Translation Paradigms and Approaches 28

4.1. Rule-Based Machine Translation 30

4.2. Knowledge-Based Machine Translation 33

4.3. Corpus-Based Machine Translation 34

4.3.1. Example-Based Machine Translation 34

4.3.2. Statistical Machine Translation 36

4.3.3. Hybrid Machine Translation 40

4.3.4. Neural Machine Translation 41

5. Summary of Chapter 43

CHAPTER 2: MACHINE TRANSLATION OF ARABIC DIALECTS 46

1. Arabic Dialects Processing 46

2. A Survey on Arabic Dialects MT 50

2.1. Arabic Dialects-MSA 53

2.2. Arabic Dialects-ENG 60

3. Processing the Algerian Dialects 69

3.1. Algerian Dialects Corpora 70

3.2. Machine Translation of Algerian Dialects 79

4. Summary and Discussion 84

CHAPTER 3: DATASET AND EXPERIMENT SETUP 87

1. The Target Algerian Dialect 87

1.1. Description of the Oranian Dialect 89

1.1.1. Major Phonological Features 89

1.1.2. Morphosyntactic Features 90
1.2. Challenges of Processing ORN 95
1.2.1. Scripting and Orthography 96
1.2.2. Morphosyntactic Aspects 97
1.2.3. Language Resources 98
2. Research Design and Procedure 98
2.1. Corpora Compilation Project 99
2.1.1. Building ORN 100
2.1.2. Building OEPC 104
2.1.3. Data Preprocessing 106
2.2. Components of the Proposed Model 108
2.2.1. The Encoder-Decoder Framework 109
2.2.2. Sequence-to-Sequence Model Implementation 110
2.2.3. LSTM Networks 111
2.2.4. The Attention Mechanism 114
2.3. Model Architecture and Hyperparameters 116
2.3.1 Tokenization and Input Representation 118
2.3.2. Core Seq2Seq Architecture 119
3. Process of the Experiment 121
3.1. Training Environment and Procedure 121
3.2. Model Evaluation Metrics 125
4. Summary 127

CHAPTER 4: RESULTS AND DISCUSSION 129

1. Results 130

1.1. Quantitative Analysis 130

1.1.1. Machine Learning Dynamics 130

1.1.2. Evaluating Test Data 132

1.1.3. Model Evaluation 133

1.2. Qualitative Analysis 138

1.2.1. Visualizing Translation with Attention 138

1.2.2. Error Analysis 150

1.2.3. Diagnosis of Intrinsic Complexities 154

1.2.3.1. Lexical Chasm of OOVs and NEs 154

1.2.3.2. Lexical and Structural Ambiguity 159

1.2.3.3. Idiomaticity and Cultural Meanings 164

2. Discussion 166

3. Summary 174

CONCLUSION 175

1. Synthesis of Findings 176

2. Implications 179

3. Limitations 181

4. Avenues for Future Research 183

REFERENCES 189

Abstract in Arabic 240

List of Tables

Table 1. Arabic Dialects Continuum 48

Table 2. Notable Morphological Analyzers for Arabic Dialects Processing 51

Table 3. Notable MT Research Works (Arabic Dialects-MSA) 59

Table 4. Notable MT Research Works (Arabic Dialects-ENG) 68

Table 5. Notable Algerian Dialects' Speech Datasets 70

Table 6. Notable Textual Corpora of Algerian Dialects 73

Table 7. MT Experiments on Algerian Dialects 84

Table 8. Pronouns in the Oranian Dialect 91

Table 9. Verb Conjugation in the Oranian Dialect 92

Table 10. Verb Negation in the Oranian Dialect 94

Table 11. Common Interrogatives in the Oranian Dialect 95

Table 12. Types of ORN Data Sources 101

Table 13. Non-Arabic Vocabulary in ORN 102

Table 14. (Non-)Inserted Lexis in ORN 103

Table 15. Normalized Written Forms in ORN 103

Table 16. Major Statistics of OEPC 105

Table 17. The Proposed Model's Components and Hyperparameters 120

Table 18. Training and Validation Metrics Per Epoch 131

Table 19. Evaluation Scores for Different Model Configurations 134

Table 20. Model Evaluation across Different Metrics 136

Table 21. Selected Test-Set ORN Sentences with Reference and Hypothesis Translations 150

Table 22. Illustrations of OOVs 155

Table 23. Illustrations of NEs Errors 157

Table 24. Types and Illustrations of Ambiguities 161

Table 25. Comparative Analysis of the Proposed Model's Performance Across Relevant Studies 167

List of Figures

Figure 1. Machine Translation Systems (1960s to 2020s)..... 29

Figure 2. The Vauquois Pyramid /Triangle (Vauquois, 1968) 32

Figure 3. Demonstration of ATG Process in Transfer-Based MT..... 32

Figure 4. The Main Steps in EBMT..... 36

Figure 5. Milestone Contributions throughout the History of MT 44

Figure 6. An Illustrative Linguistic Variation across the Arab World 47

Figure 7. Trends in Arabic Dialect NLP Tasks from 2014 to 2024..... 49

Figure 8. Map of Oran City..... 88

Figure 9. Examples of the Oranian Dialect's Rich Morphology 97

Figure 10. Procedure of Textual Corpora Compilation in ERAD 100

Figure 11. Examples from OEPC 105

Figure 12. Histogram of Sentences Length in OEPC 106

Figure 13. Subword Tokenization of Oranian Dialect..... 108

Figure 14. The Basic Encoder-Decoder Model 109

Figure 15. The Basic Seq2Seq Encoder-Decoder Model 111

Figure 16. LSTM with Interacting Layers..... 113

Figure 17. Seq2Seq Model with Attention Layer..... 115

Figure 18. The Proposed RNN Encoder-Decoder Model with Attention Mechanism..... 117

Figure 19. NMT Experimental Development and Pipeline 123

Figure 20. Model Training and Validation Performance..... 132

Figure 21. A Step-by-Step Visualization of the Attention Mechanism 139

Figure 22. Attention Score of an Exemplary 2-gram Sentence 142

Figure 23. Attention Score of an Exemplary 3-gram Sentence 143

Figure 24. Attention Score of an Exemplary 4-gram Sentence 144

Figure 25. Attention Score of an Exemplary 5-gram Sentence 146

Figure 26. Attention Score of an Exemplary 6-gram Sentence 148

Figure 27. Attention Score of an Exemplary 10-gram Sentence 149

List of Abbreviations, Acronyms and Initialisms

AI	Artificial Intelligence
AIDA	Automatic Identification for Dialectal Arabic
ALG	Algerian dialect(s) / Algiers dialect
ALGASD	Algerian Arabic Speech Database
ALGDC	Algerian Dialect Corpus
ALPAC	Automatic Language Processing Advisory Committee
AMCASC	Algerian Modern Colloquial Arabic Speech Corpus
ANB	Annaba dialect
ASR	Automatic Speech Recognition
BART	Bidirectional and Auto-Regressive Transformer
BERT	Bidirectional Encoder Representations from Transformers
BOLT	Broad Operational Language Translation
BP	Brevity Penalty
BPE	Byte Pair Encoding
BTG	Bracketing Transduction Grammar
CALYOU	Comparable spoken ALgerian extracted from YOUtube.
CAT	Computer-Aided Translation
CBMT	Corpus-Based Machine Translation
CNN(s)	Convolutional Neural Network(s)
CODA	Conventional Orthography for Dialectal Arabic
DARPA	The Defense Advanced Research Projects Agency
DDMT	Data Driven Machine Translation
DL	Deep Learning
EBMT	Example-Based Machine Translation
EGY	Egyptian dialect(s)
ENG	English Language
EOS	End of Sentence (sequence/string)
ERAD	Empowering and Resourcing Algerian Dialects
FACST	French Algerian Code-Switching Triggered corpus
FAHQT	Fully Automatic High-Quality Translation

FRE	French Language
GDSRTD	General Directorate for Scientific Research and Technological Development
GNMT	Google Neural Machine Translation
GLF	Gulf dialects
GPT	Generative Pre-trained Transformer
GPU	Graphic Processing Unit
GRU(s)	Gated Recurrent Unit(s)
HAMT	Human-Aided Machine Translation
HMT	Hybrid Machine Translation
HSMT	Hierarchical Statistical Machine Translation
HT	Hypothesis Translation
IBM	International Business Machines (corporation)
IRQ	Iraqi dialect(s)
JOR	Jordanian dialect(s)
KAB	Kabyle (Berber/Tamazight) dialect
KBMT	Knowledge-Based Machine Translation
LBY	Libyan dialect(s)
LEB	Lebanese dialect(s)
LEV	Levantine dialect(s)
LLM(s)	Large Language Model(s)
LM(s)	Language Model(s)
LRL(s)	Low Resource Language(s)
LSTM	Long Short-Term Memory
MADAMIRA	Morphological Analysis and Disambiguation for Arabic
MADAR	Multi Arabic Dialect Applications and Resources
MAGEAD	Morphological Analyzer and Generator for Arabic Dialects
MAHT	Machine-Aided Human Translation
mBART	Multilingual Bidirectional and Auto-Regressive Transformer
MDED	Moroccan Dialect Electronic Dictionary
MENA	Middle East and North Africa
METEOR	Metric for Evaluation of Translation with Explicit ORdering

MGB	Maghrebi dialects
MHA	Multi-Head Attention
MIT	Massachusetts Institute of Technology
ML	Machine Learning
MOR	Moroccan dialect(s)
MSA	Modern Standard Arabic
MT	Machine Translation
MTL	Multi-Task Learning
MuDMaT	Multi-Dialect Machine Translation
NE(s)	Named Entity(ies)
NER	Named Entity Recognition
NLP	Natural Language Processing
NMT	Neural Machine Translation
OEPC	Oranian English Parallel Corpus
OOV(s)	Out-of-Vocabulary (words)
ORN	Oranian dialect
OTAM	Offensive Tamazight corpus
PADIC	Parallel Arabic DIalect Corpus
PAL	Palestinian Dialect(s)
PBSMT	Phrase-Based Statistical Machine Translation
PoS	Part of Speech
PPL	Perplexity (metric)
QAT	Qatari dialect(s)
RBMT	Rule-Based Machine Translation
RNN(s)	Recurrent Neural Network(s)
RT	Reference Translation
SA	Sentiment Analysis
SAU	Saudi Arabia dialect(s)
SBSMT	Syntax-Based Statistical Machine Translation
Seq2Seq	Sequence-to-Sequence
SL	Source Language
SMT	Statistical Machine Translation

SOS	Start of Sentence (sequence/string)
STRCDA	Scientific and Technical Research Center for the Development of Arabic
SUD	Sudanese dialects(s)
SYR	Syrian dialect(s)
TL	Target Language
TM(s)	Translation Model(s)
TPU	Tensor Processing Unit
TTL	Transductive Transfer Learning
TUN	Tunisian dialect(s)
TWIFIL	TWItter proFILing corpus
USA	The United States of America
USSR	Union of Soviet Socialist Republics
WBSMT	Word-Based Statistical Machine Translation
WER	Word Error Rate
WWII	The Second World War
YEM	Yemeni dialect(s)

INTRODUCTION

For centuries, the study of human language was approached through theoretical frameworks. Even the generative grammar proposed by Chomsky (1965) revealed the profound, rule-based complexity of linguistic systems. However, while adept at describing abstract competence, these theories struggled to model the probabilistic reality of language performance. This challenge was inherited by the interdisciplinary field of 'computational linguistics', which emerged in the mid-20th century with the ambition to make language computationally tractable (Mitkov, 2022). Inherently linked to the broader goals of Artificial Intelligence (AI), this new frontier initially attempted to codify the rule-based systems of traditional linguistics into explicit algorithms (Lenci & Padó, 2022). This symbolic AI paradigm resulted in meticulously hand-crafted systems that, while moderately successful in constrained domains, were mostly brittle to scale or contend with the immense ambiguity and variability of real-world linguistic phenomena (Jurafsky & Martin, 2025).

The inherent limitations of the symbolic approach catalyzed a paradigm shift toward empirical, data-driven methodologies. The rise of Natural Language Processing (NLP) field, therefore, was propelled by the statistical revolution and Machine Learning (ML). The latter abandoned hand-crafted rules in favor of probabilistic models learned automatically from large corpora, embracing the inherent uncertainty of language (Fradkov, 2020). This era was, in turn, superseded by a more profound transformation driven by Deep Learning (DL), where models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) demonstrated an unprecedented ability to learn hierarchical feature representations directly from sequential data (DiPietro & Hager, 2020). This progression has culminated in the current era of Transformer-based

architectures, whose parallelizable self-attention mechanism has enabled the training of massive pre-trained Large Language Models (LLMs) like Bidirectional Encoder Representations from Transformer (BERT) and Generative Pre-trained Transformer (GPT), which now achieve state-of-the-art performance across a diverse array of NLP tasks, with machine translation standing as one of its most significant applications.

1. Background

Machine Translation (MT) is arguably one of the oldest and most ambitious goals of computational linguistics, dating back to Warren Weaver's influential memorandum in 1949, which proposed using techniques from cryptography and information theory to automate the process of language translation. At its core, MT is the task of automatically converting a text from a source language (SL) into an equivalent text in a target language (TL). A common misconception, particularly among those outside the field, is that translation is a straightforward process of word-for-word substitution, a mere lexical lookup. This simplistic view belies the profound complexity of the task. True automated translation requires a deep understanding of syntax, semantics, morphology, and pragmatics in both SL and TL. It necessitates resolving lexical ambiguity (e.g., the word "bank" in English), handling structural differences (e.g., subject-verb-object vs. subject-object-verb word order), and accurately conveying idiomatic expressions and cultural nuances that have no direct literal equivalent.

The history of MT reflects an evolving understanding of this complexity, moving through three major paradigms. Rule-Based Machine Translation (RBMT), the dominant approach for decades, relied on human experts to manually encode bilingual dictionaries and extensive sets of linguistic rules. These systems were transparent and predictable but astronomically expensive to build, difficult to maintain, and notoriously poor at handling the exceptions and irregularities that

pervade natural language. The rise of large-scale data and statistical methods gave way to Statistical Machine Translation (SMT), which learned probabilistic "translation models" and "language models" from parallel corpora. SMT represented a significant leap forward in quality but often produced disfluent or grammatically awkward output, as it translated text in small, localized phrases without a holistic understanding of the sentence's structure.

The current state-of-the-art is dominated by Neural Machine Translation (NMT). NMT models, typically based on the encoder-decoder architecture, read the entire SL sentence into a continuous vector representation (the "thought vector") and then generate the target translation, conditioned on both the source representation and the previously generated words. This end-to-end approach allows NMT systems to learn much more abstract representations of language. Bolstered by the attention mechanism, modern NMT models can produce remarkably fluent and accurate translations, often approaching human quality for high-resource language pairs like French-English or German-English.

However, the remarkable success of NMT comes with a critical dependency: data. These models are data-hungry, requiring millions, if not tens of millions, of high-quality, human-translated parallel sentences to learn their parameters effectively. This reality has created a significant digital divide. While a handful of languages with strong economic and political influence, such as English, Mandarin Chinese, and Spanish, are considered "high-resource," the vast majority of the world's over 7,000 languages are Low-Resource Languages (LRLs). These languages suffer from a scarcity of the digital resources (large corpora and annotated datasets) that are prerequisites for building modern language technologies.

The Arabic language presents a particularly complex case. While Modern Standard Arabic (MSA) has a considerable amount of text data (news, books, official documents), making it a

relatively high-resource language, the daily, lived reality for over 420 million speakers is dialectal Arabic (Yafooz, 2024). The latter encompasses a vast and diverse continuum of regional varieties (e.g., Egyptian, Levantine, Maghrebi) that differ significantly from MSA and from each other in terms of phonology, lexicon, and syntax (Habash, 2021). These dialects are the languages of social media, daily conversation, and cultural expression, yet they are overwhelmingly low-resource in the digital domain (Dahou et al., 2025). The need to translate LRLs, including these vital Arabic dialects, is not merely an academic exercise. It is a matter of digital inclusion, cultural preservation, and economic empowerment. Providing MT capabilities for these languages breaks down communication barriers, enables access to global information, allows speakers to participate in the digital economy in their native tongue, and ensures that their rich cultural heritage is represented and preserved in the digital age. The present inquiry squarely addresses this imperative by focusing on the development of MT capabilities for one such under-resourced variety: the Oranian dialect (henceforth, ORN) of Algerian Arabic.

2. Research Gap

The field of Arabic NLP has made significant strides, yet a persistent and critical research gap exists in the processing of dialectal varieties. As the comprehensive surveys by Elnagar et al. (2021) and Abdul-Nabi et al. (2024) underscore, the overwhelming focus of research and resource development has been on MSA. While valuable, this focus neglects the primary mode of communication for hundreds of millions of people and leaves a vast linguistic landscape largely unexplored by modern NLP technologies. Within this landscape, the dialects of the Maghreb, and Algerian dialects in particular, remain among the most under-resourced and under-studied. In this respect, these several specific, interlocking gaps in the existing body of research are identified: (1) the absence of a dedicated parallel corpus for the Algerian dialect(s) and English, (2) the lack of

prior MT experiments for any Algerian dialect into English, (3) the outdated methodologies in existing Algerian dialects research, (4) insufficient orthographic and preprocessing standards, and (5) the lack of benchmarks for evaluation.

First, **t**he most fundamental gap is the complete non-existence of a parallel corpus suitable for training an NMT model between ORN and English. A review of existing resources reveals that very few cater to Algerian dialects at all. The Parallel Arabic Dialectal Corpus (PADIC), while a notable contribution, contains Algerian data parallel to MSA, not English (Meftouh et al., 2015; Lichouri & Abbas, 2021). Its size is also moderate by NMT standards, making it insufficient for training DL models. There are no other publicly available corpora, either monolingual or bilingual, that focus specifically on ORN or Western Algerian dialects more broadly. This resource vacuum represents the single greatest barrier to any research in this domain.

Second, as a direct consequence of the resource gap, there have been no previous attempts to build an MT system, neural or otherwise, for the ORN-to-ENG language pair. The few MT experiments involving Algerian dialects have been restricted to translation between MSA and the dialect, or have been part of broader multi-dialect systems where Algerian is one of many varieties. This leaves the unique translation challenges posed by the ORN-ENG pair, such as the handling of code-switching, the translation of highly localized idioms, and the structural differences with English, completely uninvestigated.

Third, much of the existing work on Algerian dialects predates the widespread adoption of the most advanced neural technologies. While pioneering for their time, these studies often relied on traditional machine learning models or early NMT architectures without the benefit of sophisticated components like Transformer-based self-attention. Furthermore, critical details of the experimental setup, such as tokenizer design and hyperparameter tuning, were not optimized

for the specific morphological and orthographic challenges of the dialect. There is a clear gap in applying the current state-of-the-art NMT architectures and best practices to the task of Algerian dialect translation.

Fourth, Algerian dialects, including ORN, lack a standardized orthography. They are often written inconsistently on social media, frequently using Arabizi (Arabic chat alphabet) or containing a high degree of code-switching. Previous datasets have not systematically addressed this variation in a way that is leveraged by modern neural tokenizers. There is a need for a resource creation process that includes a rigorous normalization and preprocessing pipeline designed specifically to handle these orthographic challenges, which can then serve as a model for future work.

Fifth, without prior models or standardized test sets, there is no established benchmark against which to measure the performance of any MT system of Algerian dialect into English. Any new research in this area must not only build a model but also establish this crucial first benchmark, providing a quantitative baseline that future researchers can use to measure their improvements and contributions. In brief, the research landscape is characterized by a critical void: no dedicated parallel corpus, no prior translation models, no application of state-of-the-art NMT techniques, and no established evaluation benchmarks for the ORN-ENG language pair. Therefore, the application of DL models to this particular Algerian dialect represents the current cutting-edge of NLP research. The journey from abstract linguistic rules to data-driven DL models has been an odyssey of increasing complexity and capability, transforming the aspiration of computational language understanding into a tangible and rapidly evolving reality. This study situates itself within this vibrant context, seeking to leverage these advanced paradigms to address a long-standing challenge at the intersection of language, culture, and technology: an MT of an LRL.

3. Motivation

The motivation for this research is rooted in a commitment to addressing the digital marginalization of LRLs and empowering their speech communities through the development of essential language technologies. The "digital divide" is often conceived in terms of access to hardware or internet connectivity, but a more insidious divide exists at the level of language (Bella et al., 2023). When the vast majority of digital content, services, and opportunities are available only in a handful of high-resource languages, speakers of LRLs are effectively excluded from full participation in the global digital society. This research is driven by the conviction that building foundational NLP resources, such as MT systems, is a critical step toward a more equitable and inclusive digital future.

The Arabic-speaking world provides a potent example of this challenge. With MSA serving as the formal and institutional language, the rich and vibrant dialects spoken in everyday life are often relegated to a secondary status in the technological domain (Dahou et al., 2025). This is particularly true for the Maghrebi dialects spoken in North Africa, which are among the most divergent from MSA. Empowering these dialects with dedicated datasets and NLP tools is a crucial act of both linguistic preservation and community empowerment. It validates the primary language of its speakers, facilitates the creation of localized digital content and services, and enables better communication both within the dialect-speaking community and with the wider world.

Within this context, Algeria presents a unique and compelling case. As the largest country in Africa and the Arab world, it is home to a complex and dynamic linguistic environment. Algerian dialects, rich with influences from Berber, French, Spanish, and Ottoman Turkish, represent a fascinating and challenging subject for NLP research. This study is motivated by the desire to bring these dialects into the fold of modern computational linguistics, providing them

with the resources they deserve. Creating a robust dataset and a high-performing MT model for an Algerian dialect serves as a powerful proof-of-concept, demonstrating that these "low-resource" challenges are surmountable and encouraging further research into other Algerian and Maghrebi varieties.

The specific choice of ORN is motivated by its linguistic and cultural significance. Oran is the second-largest city in Algeria and a major economic, cultural, and intellectual hub for the western part of the country. Its dialect is not only widely spoken but also highly influential, serving as a representative for the entire Western Algerian dialect group. Characterized by a distinct phonetic inventory and a lexicon rich with borrowings, particularly from Spanish and French, ORN poses unique and interesting challenges for MT. This work is therefore driven by a dual motivation: a scholarly desire to tackle a novel and complex NLP problem, and a social commitment to using technology to empower the speakers of ORN and, by extension, to advance the field of Algerian and Maghrebi NLP as a whole.

4. Statement of the Problem

The advancement of modern computational linguistics, particularly in the domain of NMT, is fundamentally constrained by an acute scarcity of structured, high-quality data for the vast majority of the world's languages. State-of-the-art deep learning models are supervised systems that depend entirely on the availability of large-scale parallel corpora for their training, a requirement that leaves most languages, including the majority of Arabic dialects, technologically underserved. This global challenge is particularly pronounced within the Algerian linguistic context, where a complete resource vacuum exists for its prominent dialects. The specific problem addressed by the present research work is therefore the absolute absence of a foundational parallel corpus for any Algerian dialect and English, which has in turn created a critical, threefold void:

a resource void preventing any data-driven research; a technological void, as no NMT model or any other NLP tool exists to bridge this communication gap; and a scientific void, leaving the unique linguistic challenges of translating this dialect entirely unstudied from a modern computational perspective.

This research directly confronts this multifaceted problem by focusing on ORN, a high-impact case study due to its status as the second most prominent dialect in Algeria and the primary representative of the Western Algerian dialect group. The problem is thus framed as a two-part challenge: first, to overcome the foundational resource barrier through the systematic construction of a novel parallel corpus, and second, to leverage this new resource to design, train, and rigorously evaluate the first-ever NMT model for this language pair. By tackling this, the study aims not only to provide the first tangible solution for a specific and significant speech community but also to establish a groundbreaking and replicable methodological blueprint. This framework can then serve as a guide for addressing similar data scarcity challenges in other critically under-resourced Algerian -or Arabic- dialects, thereby contributing to a more linguistically inclusive digital landscape.

5. Research Objectives

To address the problem stated above, this dissertation sets out to achieve three primary, sequential objectives, forming a comprehensive research arc that moves from foundational resource creation to final empirical evaluation. The first objective is to compile and prepare the novel Oranian-English Parallel Corpus (OEPC), thereby addressing the foundational data-scarcity gap by creating the first publicly available, machine-readable parallel dataset for this language pair. Building upon this, the second objective is to design, train, and optimize an NMT system specifically tailored to the characteristics of the newly created OEPC. The final objective

is to conduct a rigorous and multi-faceted evaluation of the trained model's performance, which will serve to establish the first scientific benchmark for ORN-ENG translation and analyze the specific challenges inherent to this task.

These cumulative objectives guide the research from an initial state of zero resources to the final delivery of a novel dataset, a functional translation model, and a thorough analysis of its performance. This entire research endeavor is thus encapsulated in the following central research question:

To what extent can an NMT model, trained on a purpose-built parallel corpus, effectively and accurately translate the low-resource Oranian dialect into English?

6. Research Methods and Design

The methodological design of this study follows a quantitative, empirical paradigm built upon a necessary and significant constructive component. The research is constructive in that its primary prerequisite involves the creation of tangible artifacts: a new linguistic resource (OEPC) and a new NMT benchmark. It is empirical in that it employs controlled experiments and objective measurements to systematically assess the performance of the constructed system and draw data-driven conclusions. This methodology unfolds in a three-phase chronological pipeline that directly aligns with the research objectives.

The initial and most critical phase is corpus creation and preparation. The foundational data scarcity of ORN necessitates this preparatory work in corpus linguistics. This involves the careful collection of authentic ORN sentences from diverse contemporary sources, followed by a meticulous professional human translation process to create the parallel English counterparts. This newly created OEPC with 8.5K sentence pairs is then subjected to a rigorous data normalization

and pre-processing pipeline to standardize orthographic inconsistencies. Finally, the corpus is partitioned into training, validation, and test sets (80-10-10 ratio) to facilitate a methodologically sound development and evaluation workflow.

The second phase is model design and implementation. The rationale for the model architecture is based on proven, powerful components suited for sequence generation and linguistic complexity. The NMT system is designed as a Sequence-to-Sequence (Seq2Seq) model employing a two-layer stacked LSTM network for both the encoder and decoder, chosen for its demonstrated efficacy in modeling long-range sequential dependencies. Critically, an additive attention mechanism is integrated to overcome the information bottleneck of traditional encoder-decoder designs. This mechanism enables the model to dynamically focus on the most relevant segments of the source sentence while generating each target word, a vital feature for aligning syntactically divergent languages like ORN and ENG. This architecture is implemented with different relevant hyperparameters selected to balance model capacity with the risk of overfitting on the moderately-sized corpus.

The final phase consists of a comprehensive evaluation protocol. To holistically assess the model defined in the previous step, a dual-pronged approach is used. First, a rigorous quantitative analysis is performed on the held-out test set using a suite of standard automated metrics. Second, this empirical data is complemented by a detailed qualitative analysis, involving the manual inspection of translation examples, a diagnostic error analysis, and visualizations of the internal attention mechanism. This cohesive methodological design ensures that the study begins by constructing the necessary tools, proceeds by building a reliable technological system, and concludes with a thorough, evidence-based assessment of its performance.

7. Contribution

The current study makes several novel and significant contributions to the fields of NLP, computational linguistics, and Arabic dialectology. By addressing a critical and long-standing gap in the research literature, this study advances our understanding of LRL processing and provides tangible resources that will catalyze future work in the domain of Algerian and Maghrebi NLP. The primary contributions of this research are: (1) building a foundational linguistic resource, (2) achieving the first benchmark for ORN-ENG MT, (3) designing a blueprint for future NMT experiments on Algerian dialects, (4) bringing empirical insights into dialectal MT challenges, and (5) empowerment of ORN speech community.

First, the most significant and lasting contribution of this study is the creation and dissemination of OEPC. As the first parallel corpus of its kind, OEPC is a foundational resource for the scholarly community. It immediately enables new avenues of research not only in machine translation but also in contrastive linguistics, computational dialectology, and the development of other NLP tools for ORN. This contribution directly addresses the data scarcity problem and provides a valuable, reusable asset for future researchers.

Second, this research contributes the first-ever NMT model for the ORN-ENG language pair. The fully trained and evaluated model, along with its established performance scores on metrics serves as a crucial pioneering benchmark. All future research aiming to improve the MT of Algerian dialects into English will be able to measure its progress against the results established in this dissertation. This sets a baseline and formalizes the task for the NLP community.

Third, the present study provides a comprehensive, end-to-end methodological blueprint for tackling NMT in a low-resource dialect context. This blueprint covers data collection and

normalization strategies for an unstandardized dialect, model selection, hyperparameter tuning for a moderately-sized dataset, and multi-faceted evaluation The detailed documentation of the pipeline can serve as a practical guide for researchers working on other under-resourced Arabic dialects, particularly within the Maghrebi continuum (e.g., Moroccan, Tunisian, Libyan dialects), or even other LRLs globally.

Fourth, through its extensive quantitative and qualitative analysis, this research offers valuable empirical insights into the specific challenges of translating the Algerian dialects, and ORN in particular. The error analysis identifies common failure points for a state-of-the-art NMT architecture, such as the handling of code-switching, the translation of culturally specific idioms, and the disambiguation of morphologically complex verbs. The findings will contribute to the broader scientific understanding of the unique difficulties posed by dialectal and low-resource NMT.

Fifth, beyond its academic contributions, this work represents a meaningful step toward the technological empowerment of the Oranian and Algerian speech communities. By creating a foundational piece of language technology, this study helps to elevate the digital presence of the dialect and provides a tool that, with further development, can facilitate communication, information access, and greater participation in the global digital landscape. In conclusion, this dissertation makes a multi-faceted contribution that is at once linguistic, technological, methodological, and empirical. It moves the study of ORN from a state of data-zero to having its first corpus, its first NMT model, and its first set of performance benchmarks, thereby opening a new and promising subfield of research.

8. Dissertation Structure

This study is organized into four chapters. Each chapter is designed to build logically upon the previous one, guiding the reader from the broad research context to the specific contributions of this study. Chapter 1 situates the research within the broader academic fields of computational linguistics, NLP, and MT. It establishes the conceptual foundation by tracing the evolution of the field, delineating the major MT paradigms from early rule-based and knowledge-based systems to the subsequent data-driven statistical and neural approaches.

Chapter 2 provides a critical survey of the relevant scholarly literature concerning the computational processing of Arabic dialects. The review begins with a broad examination of resources and models developed for various Arabic varieties before systematically narrowing its focus to the Maghrebi region. It specifically scrutinizes the current state of Algerian dialect processing, with a detailed analysis of existing corpora, prior NLP tasks, and the few MT experiments that have been conducted. This chapter serves to substantiate the research gap, highlighting the absence of resources and prior work for the ORN-ENG language pair.

Chapter 3 reflects on the research gap concerning Algerian dialects, articulates the study's motivation, and formally sets its methodological design. It presents the detailed methodological framework employed in this study, providing a transparent and replicable account of the entire experimental process. The chapter begins by delineating the constructive research design. It then meticulously documents the multi-stage process of compiling and preparing the novel OEPC. Following the corpus creation, the chapter specifies the NMT system's technical design, detailing the Seq2Seq architecture, the two-layer LSTM networks, the integrated attention mechanism, and the complete configuration of the model's parameters.

Chapter 4 presents, analyzes, and interprets the empirical outcomes of the NMT experiment. The chapter begins with a comprehensive quantitative assessment of the model's performance on the held-out test set, reporting scores across a suite of standard evaluation metrics. This is followed by an in-depth qualitative analysis of the translation output, which includes an error analysis and a visual examination of the model’s attention mechanism. The chapter culminates in a detailed discussion that synthesizes these findings, contextualizes them by making comparisons to relevant studies in low-resource MT, and conclusively addresses the central research question.

CHAPTER 1: MACHINE TRANSLATION

The amassing technological boons in the last few decades have brought almost completely different communicative aspects into our lives. We started to depend more on machines and devices in many activities like reading, watching, driving, etc. Machine-aided and human-aided tasks have reached almost all areas of living. Today, life is full of technologically-based or -assisted products. It has been a must to mention 'Artificial Intelligence' (AI) in myriads of sorts and areas. AI is in the movies and news we watch, e-books we read, robots that assist human activities, cars that are self-driven, cameras and drones we use for filming, medical devices and systems, online platforms from which we buy all kinds of goods and utilities, and in many other technologies that influenced our present day. Indeed, there are several conceptions and misconceptions about AI; for that, most of us are buried in the AI pundits' words, introductions and opinions. Some consider AI replacing the human race, some others keep insisting on its fuzziness. Realizing AI explicitly and imagining it over-actively can sum up the struggle for its scientific identification.

In the midst of technological trends and AI boost during the last few decades, the automated translation has become a noteworthy field of interest. With the assistance of computers, and as a significant subfield of computational linguistics, 'Machine Translation' (MT) enables computers -or machines- to process the lexical, syntactic and semantic levels of texts or speeches in one natural language in order to reproduce them into another. MT systems, therefore, aim at minimizing the human aid within processes like translating or editing texts. As a subject about

which experts in sciences and humanities do not tend to be neutral, MT has had a long history that witnesses no shortage of vehement disparagers and supporters alike.

In the light of what has been mentioned, this chapter implicitly reconsiders the history and evolution of MT as it appraises the various approaches to and models of MT. For that, it is worth introducing AI and MT technologies briefly with their short historical frameworks. The major approaches of MT are further presented with relevance to the features they provided to the field. However, the chosen paradigm for this research (SMT) is thoroughly presented. Then, challenges of processing natural languages and automated translation are briefly discussed. Critically, the last section of this chapter provides the major challenges of processing and translating the Arabic language.

1. Artificial Intelligence Primer

To discern "intelligence" in its ordinary sense, Mueller and Massaron (2018, p. 8) listed some mental activities like obtaining and processing information, reasoning, understanding, grasping truths, considering and making meanings, separating facts from beliefs, etc. Such introduction is prone to explaining intelligence briefly. However, when it comes to understanding 'intelligence' in the sense of computational processes, it is necessary to categorize intelligence in order to decide what it particularly involves. As humans, we accomplish tasks through our reliance on multiple intelligences. Recognizing these intelligences helps us relate to the computational simulations that reflect how computers or machines can: (1) think humanly, (2) act humanly, (3) think rationally, and (4) act rationally (Russell & Norvig, 2016, pp. 2-5).

Since their existence, humans used their intelligence to create machines that help them. For that, a strong desire for intelligent machines triggered their "exciting new effort to make computers think ... machines with minds, in the full and literal sense" (Haugeland, 1985, p. 2). However, making intelligence features in the artificiality of entities was initiated by Alan Turing. In 1950, in order to determine whether an artificial entity is intelligent, Turing concretely tested, through an interrogator, a human and an artificial entity with a textual device. Based on posing questions, the interrogator was required to distinguish the other human from the artificial entity based on answers to questions posed by the interrogator (Neapolitan & Jiang, 2018, p. 2). Eventually, researchers started to develop intelligent systems. Dartmouth workshop in 1956 involved several experts, namely the so-called father of AI, John McCarthy, who coined the term 'Artificial Intelligence'. The latter emerged as a new field of interest that aimed at creating learning, reacting and decision-making computer systems (McCarthy, 1956).

AI has been defined differently. Kurzweil (1990), for instance, considers AI as an art where machines "perform functions that require intelligence when performed by people." Moreover, Winston (1992) posits AI as "the study of computations that make it possible to perceive, reason and act" (p. 5). Contending that it could make confusion with computational intelligence, Poole et al. (1998) referred to AI as "the study of the design of intelligent agents" (p. xv). During the 1970s and 1980s, AI was approached differently; cognitive, logical and probabilistic approaches as well as knowledge-based systems (Neapolitan & Jiang, 2018, pp. 4-6).

However, AI development witnessed a period of reduced funding commonly referred to as 'AI winter'. The hype period was significantly marked by teaching computers to learn from data without human operators. Such a technology became known as 'machine learning' (ML). At its

peak, in the last two decades, the technology of 'deep learning' (DL) strived to make the intelligent agents working like the human brain by virtue of available huge datasets, high-quality computers, smart algorithms, and the committed business investments (Mueller & Massaron, 2018, pp. 16-17).

The development of AI and the improvement of the computational processing of natural languages co-existed in a technological revolution era. Synchronously, the automated translation went through ups and downs since the 1950s. By virtue of the continuously-upgraded and fast computing infrastructures, AI rapidly helped creating bases for ML and making sophisticated algorithms that, in turn, helped mining bottomless troves of data. Therefore, the availability of big data and new techniques made it easier for humans to translate thousands of words and sentences in a few seconds through MT systems. In this respect, it is necessary to briefly review the history of MT in order to highlight its features and check the evolution of its models and approaches.

2. Introducing Machine Translation

The act of transferring meanings of a given text from one language to another, taking into account accuracy and style, is commonly referred to as 'translation'. The latter requires understanding the language of the original text (source language, SL) and expressing contents coherently in another language (target language, TL). There are two major methods of translation; literal and contextual. The first describes the translator's commitment to the meanings of the vocabulary of the original text. The second depends on understanding the original text's general meaning and words-in-context. Indeed, translation's major aim is understanding and communicating with the others.

Such definitions differ when the translator is a machine. Under different paradigms, and over decades, MT was defined and introduced differently. Hutchins (1994) refers MT, as a term, to "computerized systems responsible for the production of translations with or without human assistance" (p. 2322). Similarly, Whitelock and Kilby (1995) considered MT as an interdisciplinary process combining lexicography, linguistics, computational linguistics, computer science and language engineering (p. 2). The central core of MT, thus, is automating the process of transforming an input (SL) into an output (TL).

The long history and the variety of interpretations made the term "Machine Translation" misleading. Since "machine" is there, Sager (1994) accounted that MT refers to automatic systems with no human involvement (p. 326). Nonetheless, shifting from 'automated' to 'machine' terminology was based on the hypothesis that natural languages can be fully described, controlled and mathematically coded by machines, more precisely computers (Wilss, 1999, p. 140). At this point, one might ask: how has translation been an automated operation in the last few decades?

3. History of Machine Translation

The MT point of start has been considered hard to associate certain events in history. The thoughts of mechanizing translation can be traced back to the 17th century though only after three centuries that realistic attempts emerged. At that time, there was a need to break linguistic and communicational barriers. For that, Hutchins (2000a) reports that ideas like creating (a) universal, logical language(s) remained a philosophical issue shared among philosophers (like Descartes, Leibniz, Wilkins, etc.) and other scholarly figures (p. 5).

Hutchins (2000a) introduced the precursing idea of automating the translation by referring to Cartesian proposals in the 17th century. For him, Descartes thought of seeing languages as coded with symbols in which equivalent items from two or more languages would share the same code or symbol. In the 19th century, sociopolitical circumstances, though, motivated many attempts to overcome linguistic barriers by bringing out a universal language. Prominently, the two major ones are: Esperanto by Zamenhof in 1887, and Begriffsschrift by Frege in 1879 (Helaoui, 2007, p. 5).

The need for devices that help humans overcome language and communication barriers existed for two centuries before the technology revolution. The idea of creating a universal language arose for the sake of an internationally-open communication, particularly in the educational and scientific arenas. Until the middle of the 20th century, writers and scientists continued to suggest mechanical dictionaries that started to logically construct the genuine differences among languages. All these attempts and ideas were reinforced by the two world wars as they were foreshadowing some principles of the automated translation of languages, known at that time as mechanical translation.

Moreover, some researchers confirm marking MT beginning with the event of the British breaking of the German "Enigma". The latter was a tool that electro-mechanically encodes and decodes German military and diplomatic communications during World War II. Such success, along with breaking the Japanese Type A code by the American intelligence, brought optimism for the prospects of MT (Aransa, 2015, p. 10). After the war, a natural move towards checking computers' potential capabilities was marked. Soon, the automatization of language translation became a candidate for exploitation (Pugh, 1992, p. 15).

3.1. Pioneers and Precursors

In 1933, the idea of mechanizing the translation emerged by patents developed by Georges Artsrouni and Petr Trojanskij. The latter suggested to automatize a dictionary and schematize interlingual grammars using Esperanto (Hutchins, 2000a, p. 5). In addition, Hutchins (2004) described Artsrouni and Trojanskij's patents as containing "the first definite legitimate forerunners of machine translation" (p. 12). In France, Artsrouni's patent could be developed as a mechanical multilingual dictionary; and in Russia, Trojanskij's patent had the same feature with more grammar functions coded and interpreted through 'universal' symbols. Besides, such codes represented how MT would be basically configured.

However, such inventions were ignored by governments. Even Zhirkov (1956) commented that linguists were sceptic about them for being impractical. Panov et al. (1956) stated that Trojanskij was issued an author's certificate for a mechanized dictionary in 1933 (p. 14). Eventually, the USSR Academy of Sciences investigated his proposal and manuscripts from 1933 to 1947. In 1959, the collection was published showing how Trojanskij had anticipated the ideas of the early MT researchers in the West (Bel'skaya et al., 1959). Trojanskij's contribution to mechanical translation was thoroughly presented by Hutchins and Lovtskii (2000). Apart from being non-linguist, understanding Trojanskij's invention was challenging due to the unfamiliarity with the electromechanical equipment employed and the unconventional language used to describe the patent. Meanwhile, other pioneers of MT started to appear in the 1940s and 1950s.

MT researchers acknowledge Warren Weaver's quote when they provide history of the field. Poibeau (2017) claims that "the father of machine translation -and more generally of natural language processing- is unquestionably Warren Weaver" (p. 51). Weaver was one of the early

computer scientists in post WWII. After armies, such researchers started to look at the problem of translation solved by computers as the war codes. In March 1947, Weaver, in his letter to the mathematician and cyberneticist Norbert Wiener, claims “when I look at an article in Russian, I say this is really written in English but it has been coded in some strange symbols I will now proceed to decode” (Green et al., 2015, p. 48). For him, cryptanalysis techniques started to seem more congenial to be applied to translation.

Weaver’s memorandum proposals were applicable to many contexts. Coauthored with Claude Shannon, in 1949, the book “a mathematical theory of communication.” Weaver and Shannon suggested encoding a message from a source, sending or transmitting, then decoding it by a receiver. Cryptography was based on their model as would be the automated translation according to their theory (Shannon & Weaver, 1949). An encoded text in a language ***A*** should be comprehended, thus, decoded in another language ***B,*** yet decoding is proposed to be the machine task. The Weaver’s initiative and the decode-thought was then elaborated in a memorandum (Weaver, 1949). Further, he proposed new theories rather than the word-for-word approach (Weaver, 1970, p. 106).

Indeed, the credo that drove MT since the 1950s is viewing translation as decoding which is still used in the jargon of applied MT. At a time when computers were not carriable, some experts were ambitious enough to tackle problems like MT. Garg and Agarwal (2018) review that Warren Weaver’s proposal for the use of computers to build MT systems led to serious and extensive studies. The assistance of humans before and/or after this automation excludes MT from being fully computer-based. Nevertheless, human-aided machine translation (HAMT) and machine-aided human translation (MAHT) brought uncertainty to the boundaries between them.

For that, the uniting term, computer-aided translation (CAT), emerged as a bridge to avoid all confusions.

Meanwhile, Andrew D. Booth, a British physicist and crystallographer, was interested in developing a mechanical dictionary. Later, he worked with Richard H. Richens to produce word-for-word translations through a punched card (Richens & Booth, 1955). With the increased interests in NLP, research on data storage and access started to find solutions to the small storage capacity that early computers had. Booth improved the dictionary management and storage, and let to an MT global approach; yet, such propositions were still considered as simplistic. (Poibeau, 2017, p. 51). Weaver and Booth were considered the first researchers who proposed to use the computer to translate natural languages (Chan, 2004, p. 291). In addition, they stimulated other institutes and researchers such as Abraham Kaplan, Victor Oswald, and Erwin Reifler to conduct MT research.

Yehoshua Bar-Hillel, an Israeli philosopher, was hired by the Massachusetts Institute of Technology (MIT) Research Laboratory of Electronics in 1951. As influenced by Weaver, he visited the major American MT research centers to prepare for an inaugural MT conference in June 1952 (Hutchins, 1997b, p. 220). Further, Bar-Hillel anticipated using dissemination and assimilation for MT claiming that pure MT was "achievable only at the price of inaccuracy." His argument was in favor of mixed MT, that is "a translation process in which a human brain intervenes" (Bar-Hillel, 1951, pp. 233-234).

The first MT conference was held at the MIT in 1952. Although Bar-Hillel considered MT output possible to make sense, "but still be dead wrong" (Bar-Hillel, 1955, p. 191). In the proceedings of the same conference, Bull et al. (1955) affirmed that MT would pose a greater risk

if the output would still be inaccurate or deceptive (p. 91). After this event, the first MT book was a compilation of fourteen essays published as an anthology entitled *"Machine translation of languages: Fourteen essays"* edited by Locke and Booth (1955). The remainder of Bar-Hillel's survey proposed MT pre- and post-editing, exhibiting a pragmatic bent that would become clearer over the course of a decade.

After taking a hiatus, Bar-Hillel was asked by the US Office of Naval Research to carry out a further survey of MT research. In 1958, he traveled to several American and British research facilities to gather any publicly accessible data on Soviet Union advancements. His critical report was published in 1959. The MT advances, thus, were due to a set of heuristics that aid programs, which Bar-Hillel (1960) called "bags of tricks". Moreover, despite paying homage to linguistics, the majority of MT system developers, as viewed by Bar-Hillel, needed to revise how to apply any of its results when creating their systems (Nirenburg & Wilks, 2000, p. 163).

Bar-Hillel's contributions are meticulously chronicled by Hutchins (2000b). His main contention was that fully automatic high-quality translation (FAHQT) is unreasonable and that, despite predictions of impending triumph, he "could not be persuaded of their validity" (Bar-Hillel, 1960, p. 94). He offered an addendum that purportedly proved FAHQT to be impossible. The evidence was this challenging statement that requires extralinguistic knowledge: "Little John was looking for his toy box. Finally, he found it. *The box was in the pen*". Sixty-three years later, Google Translate is unable to accurately translate this sentence into several languages. Subsequently, most of those who supported MT research in the US government started to share Bar-Hillel's pessimism, which significantly decreased funding for MT experiments and improvements.

3.2. The Georgetown-IBM Experiment

In 1954, a famous experiment in Georgetown at Johnson university, Washington, was conducted. A machine could translate 250 words with only six grammar rules and 49 sentences though it was considered as toy system (Goshawke et al., 1987, p. 26). It was a collaborative work with IBM that employed word-for-word English-Russian translations without built-in linguistic components (Hutchins, 1995, p. 434). In fact, that was not a full-fledged translation system. Besides, spending much research money on MT was hardly looked at as necessary (Vasconcellos, 2000). Jointly with IBM, the Georgetown university experiment forecasted a promising solution for MT systems within few years. Besides, such optimistic results guided other MT experiments in other countries whose research experts required more funding.

After eight years, the Georgetown University MT project team had to resort to post-editing, as recommended earlier by Bar-Hillel (1960). However, post-edited translation "took slightly longer to do and was more expensive than conventional human translation" (Pierce, 1966, p. 19). Even Beyer (1965) stated, "I doubt if the edited translation reads as smoothly as one which I would have started from scratch" (p. 50).

Although the English-Russian demonstration showed a potential success, there were still fears from the Georgetown-IBM about the predictions of MT full capabilities. From this experiment onwards, more MT activities spread in many areas. Research continued in a number of countries, such as the USA and England (since 1948), the Soviet Union (since 1954), Japan (since 1956), Italy and France (since 1959), etc. Research teams worked on operational MT and on developing MT systems. Perhaps the main reason behind this research was the desire to check and evaluate scientific and technological information from the competing countries. As the

governments and militaries increased their funding plans during the decade of optimism, the bubble burst in 1966 by publishing the ALPAC report.

3.3. The ALPAC Report

In spite of unrealized capabilities of MT methods, the research community was hopeful and optimistic about the future of MT before 1966. Verily, the Automatic Language Processing Advisory Committee (ALPAC) report in 1966 had its influence on the progress in MT systems. Somers (2003) estimated 12 to 20 million dollars to be invested by the US government on MT experiments during the late 1950s and early 1960s (p. 4). Sponsors and the general public expected high-quality translations, yet this report caused their optimism to vanish. As it issued that there is no point in conducting MT research for 20 million dollars were spent in the US per year while all expensive computers would not make it cheaper.

Maximizing the power of computers and dealing with NLP systems would surely cost and be demanding. For that, the ALPAC (1966) concluded that:

- MT and post-edited MT is more costly than human translation.
- Since human translators are available, less attention should be drawn to automatization.
- No advantage is achieved from using MT over human translation.
- Funding should be directed to basic linguistic research (Church & Hovy, 1993; Hutchins, 2003).

Contrary to what is commonly believed, the report did not signal the end of MT research; rather, it recommended speeding up the human translation process and evaluating the relative speed and cost of machine-aided translation, as suggested earlier by Bar-Hillel (Pierce, 1966). Indeed, the ALPAC (1966) investigated the MT situation and recommended to stop supporting

MT in the mid-1960s. Such a report dispersed many research groups and caused, in one way or another, new considerations about MT experiments (Hutchins, 2001; Slocum, 1980). Funding, lack of cooperation and divergence in the methods were the major obstacles in the first decades of MT history. Although ambitious precursors came to give hope to applied and computational linguists, dispersion and disillusion occurred after every MT attempt (Somers, 1992).

Nevertheless, researchers worldwide kept working on finding ways to make machines acquire knowledge about natural languages. Considered as a failure for more than a decade awaited the MT revival till the late 1980s due to commercial and political conditions during the Cold War (Hutchins, 2000a, pp. 4-5). Emerging as an academic affair, MT became the concern of some universities and institutes' researchers. The openness was through mere exchange of reports and organizing conferences. The rise of interest was described by Pugh (1992) through stating that "the 1990s and beyond seem to herald a greater internationalization of MT efforts, and there is reason to hope that the ghost of ALPAC will finally be laid to rest" (p. 17).

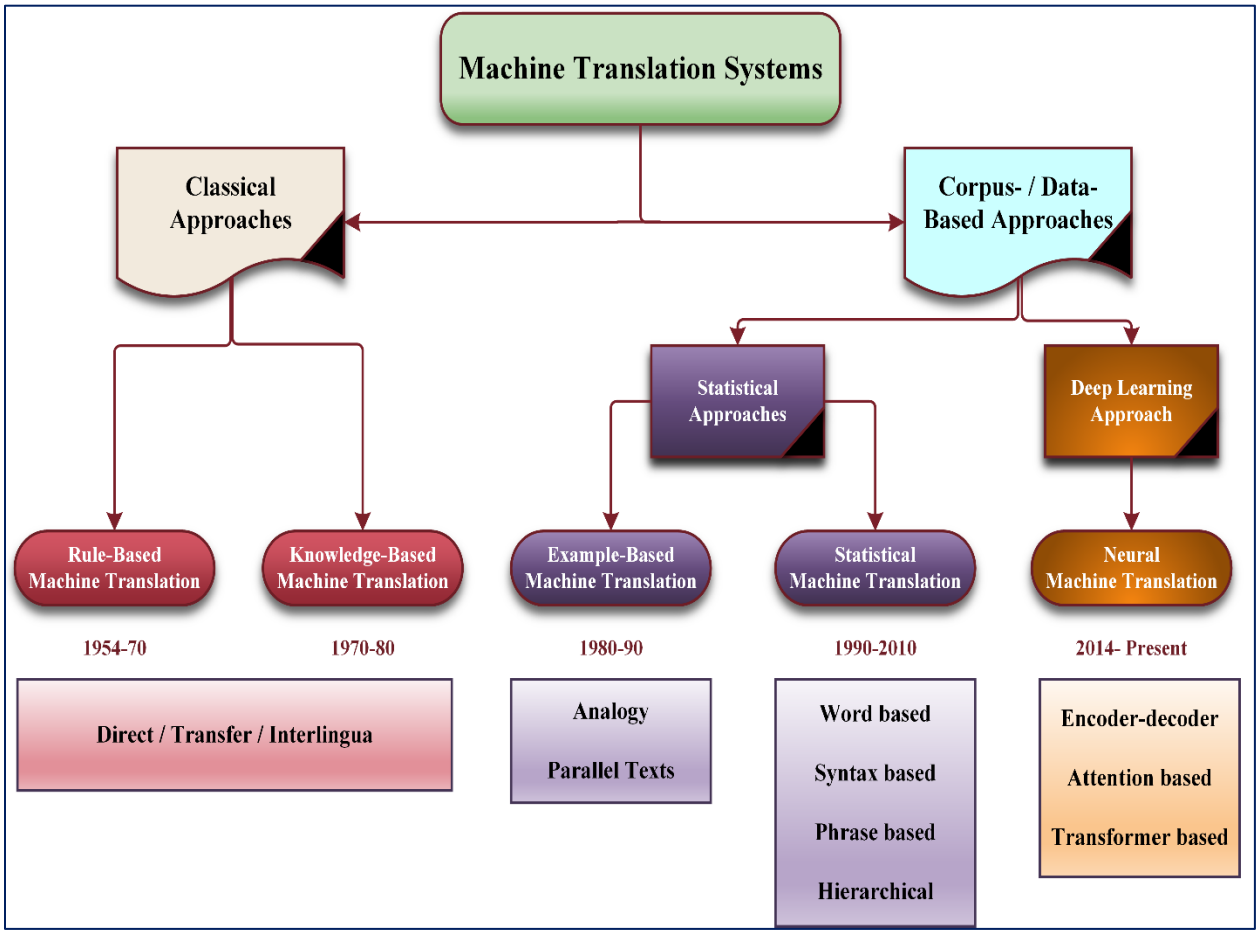
4. Machine Translation Paradigms and Approaches

The number of automatic translation types has multiplied since the 1960s till now. For that, terminological inconsistency and confusion are provoked by the variety of approaches within -the one or all- paradigms of MT (Slocum, 1985). Indeed, the MT paradigms differed in the perspectives of researchers' concerns and contributions. Bhattacharyya (2015), for instance, considers MT as having evolved under three major paradigms: rule-based (RBMT), example-based (EBMT), and statistical (SMT). However, most of the recent research (such as Mondal et al., 2023) categorize them into classical and corpus-based approaches. In this respect, MT systems

and their approaches are demonstrated in Figure 1 with references to their dates of emergence and major approaches.

Figure 1

Machine Translation Systems (1960s to 2020s)



Based on the fact that there is no clear agreed-upon unanimity on the categorization of MT systems, and for the purpose of the present research work, all of the emerging paradigms and approaches, as presented in Figure 1, will be introduced in this chapter. In line with their historical development from the 1960s till 2020s, three major MT paradigms, with their approaches, are presented: (1) rule-based, (2) knowledge-based and (3) corpus-based MT.

4.1. Rule-Based Machine Translation

As TL requires explanation, rules have their advantage of supplying. Rule-Based Machine Translation (RBMT) refers to applying SL rules as well as adapting them to TL engineered structures. RBMT is featured by linguistic rules and its models work as "a series of processes which analyze input text: morphological, syntactic and/or semantic analyses; and a process of generating text as a result of a series of structural conversions based on an internal structure or some interlingua" (Carl & Way, 2003, p. xviii). Lehrberger and Bourbeau (1988) contend that RBMT was the first commercial system that used a knowledge-driven approach. The most popular RBMT systems are SYSTRAN (Toma, 1977; Senez, 1994), OpenLogos MT (Scott & Barreiro, 2009), Apertium (Ramírez-Sánchez et al., 2006; Forcada et al., 2011) and Grammatical Framework (Ranta, 2011).

The generating criteria in RBMT are manually created using linguistic knowledge and relying on bi- or multi-lingual dictionaries. There are three representations in RBMT: direct, transfer and interlingua. With no intermediate representation, translation is direct. Based on dictionaries and word-for-word translation, the direct system uses simple rules on the surface form to (re)order words in TL. With more abstract representations, translation is a transfer process. As this system integrates a syntactic analysis, SL structure is exploited and TL structure is more accurately provided. With more formal and sound representations, an interlingua system is used. That is, contents of SL are transferred to a pivot language from which the translation of TL occurs (Poibeau, 2017, pp. 27-28). Indeed, semantics requires a translation stage between a natural language input and their logical representations in this system (Wilks, 2008, p. 35).

First, MT used arbitrary texts, or texts of interest to the military. Then, in the 1960s, the funding factor and the linguistic difficulties increased. The earliest MT system -the direct method- came as a direct processing strategy without any kinds of intermediate stages. Less primitive software designs led later to indirect approaches (Hutchins & Somers, 1992, p. 72). Linguistic naiveties were soon recognized; the internal analysis of SL was required. The indirect method came with interlingua approach and transfer approach. The former refers to analyzing SL and generating an output through its representation. It is interlingual due to the neutrality between SL and TL representations (p. 73). The latter refers to analyzing (interface) language-dependent, abstract representations via bilingual modules interposition (p. 75). All three approaches: direct, interlingua and transfer, were conducted through non-interventionist or interactive modes.

It is noteworthy that the transfer process relies on the chain of analysis-transfer-generation (ATG). The aforementioned strategies -or sub-systems- of RBMT, along with a clarification of ATG process, were demonstrated in the Vauquois Pyramid proposed by the French MT researcher Bernard Vauquois in the 1960s (Costa-jussà & Fonollosa, 2015, p. 4). As shown in Figure 2, the Vauquois triangle explains how an input is processed, then an output is generated within the varying strategies of direct, transfer and interlingua. Each of the ATG steps has certain rules. First, rules to analyze SL; second, rules to transfer any represented analysis; and third, rules to generate TL (see Figure 3). Rule-makers contribute to the strength or weakness of rules according to their linguistic expertise as well as their understanding of the SL and TL properties. In RBMT, the translation process is controlled by dictionaries and inspected by linguists. For that, this approach was considered as slow and time-consuming (Carl & Way, 2003, p. xviii).

Figure 2

The Vauquois Pyramid /Triangle (Vauquois, 1968)

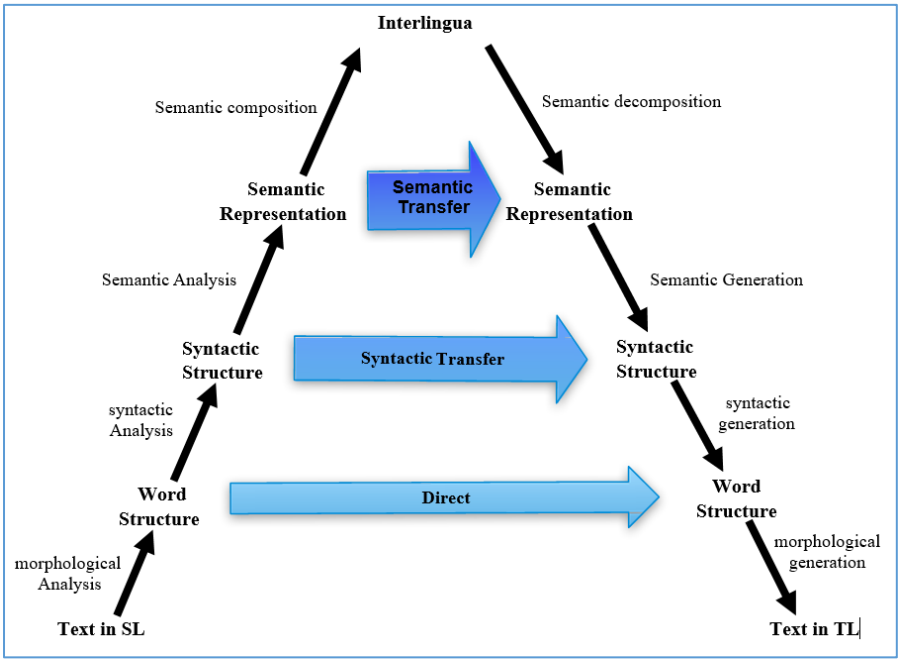
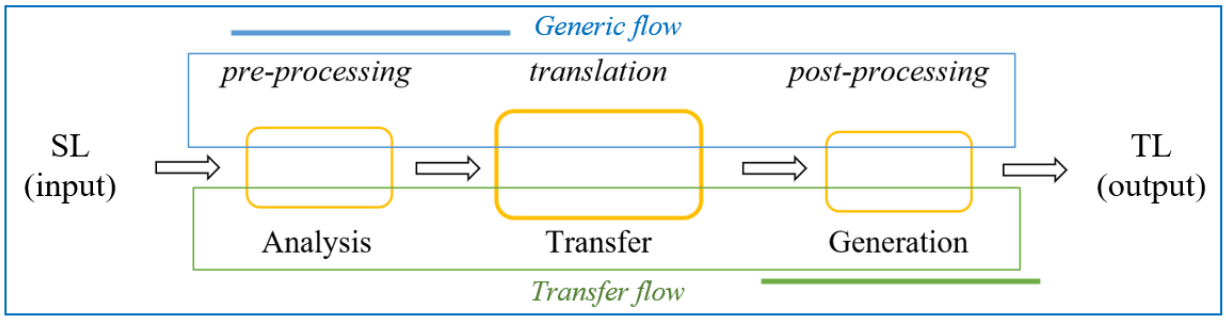


Figure 3

Demonstration of ATG Process in Transfer-Based MT



Although the transfer method was a sub-system in RBMT, most of the following MT systems came with similar flows of operations but with more sophisticated steps and processes. In spite of their high precision, rule-based models are featured with low recall. Besides, some situations have more than one applicable or suitable rule -along with rules conflict- and this

requires more careful rule ordering. For that, Bhattacharyya (2015) contends that, in RBMT, "the onus of empowering the MT system lies with the human expert" (p. 26). Apart from the technical obstacles in the course of developing MT systems, the semantic features of SL made it complicated for RBMT to process and engineer some languages (such as Arabic and Japanese). The proposal for a dataset on which the MT system relies -or should rely- came as a solution.

4.2. Knowledge-Based Machine Translation

Knowledge-Based Machine Translation (KBMT) refers to the limited domains' knowledge provided to MT processing of SL. The knowledge layers used in MT are (1) conceptual or ontological knowledge (such as animal, human and plant), (2) world knowledge (such as axioms and natural laws), and situational or factual knowledge (such as place, time and status). Linguists and programmers supply the KBMT system with linguistic and computational instructions (Nirenburg, Beale & Domashnev, 1994). They work on making data readable to the computer depending on terminological materials. In such system, the MT process may be unidirectional or bidirectional (from SL only, or from and to SL).

KBMT can be described as a RBMT system with extensive semantic and pragmatic knowledge of a given domain. As it was built upon the idea of RBMT, KBMT incorporated various forms of external knowledge to enhance the translation quality and address some of the limitations of pure rule-based systems. However, KBMT required handcrafted linguistic rules; therefore, it took much time, energy and funding. The three common projects that used KBMT systems are KBMT-89 (Nirenburg, 1989), Verbmobil (Wahlster, 1993) and DBR-MAT (Bontcheva, 1997).

4.3. Corpus-Based Machine Translation

Corpus-Based Machine Translation (CBMT) refers to extracting equivalents from existing data of parallel corpora. This approach has, by virtue of corpora availability, witnessed wide uses and practices starting from the 1990s (Yngve, 2000). As it was known as data-driven MT (DDMT), CBMT paved new directions to the field. Large scale texts proposed alternatives "to overcome the knowledge acquisition problem of RBMT systems". Instead, parallel corpora are "used and/or consulted to derive the required knowledge for new translations" (Carl & Way, 2003, p. xviii).

CBMT had developed in the 1990s through improving probability factors in its application. Therefore, the language patterns identification witnessed a thorough scrutiny in large data stores such as the web. Bilingual documents alignment had been realized through sub-alignment processes like segments, patterns, sentences, phrases and words alignments (Brown et al., 1993). For that, and because of data alignment and corpora availability, CBMT became the seed of other significant MT approaches that are: Example-Based Machine Translation (EBMT), Statistical Machine Translation (SMT) and Neural Machine Translation (NMT).

4.3.1. Example-Based Machine Translation

An initial direction within data-driven automated translations was Example-Based Machine Translation (EBMT). Translators within the example-based model "collect a bilingual corpus of translation pairs and then use a best match algorithm to find the closest example to the source phrase in question" (Arnold et al. 2001, p. 188). The mapped rules, thus, elevates the EBMT to be a process by analogy that produces translation templates. EBMT can also be referred to as

'memory-based', and it is mainly about emulating how human translators think (Nirenburg, Carbonell & Tomita, 1994).

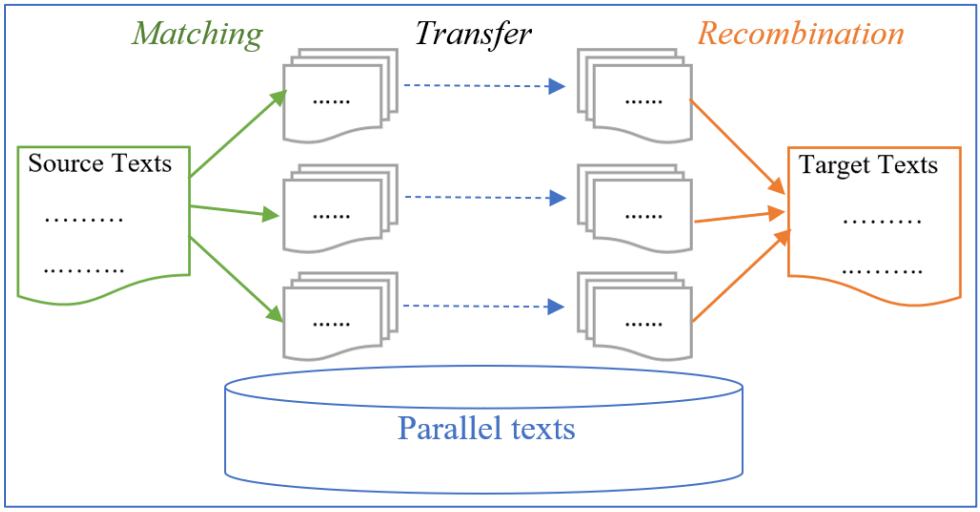
It is interesting how AI stood up on the opposing processes that are human knowledge-driven or/and data-driven. Then, how about a combination of such contrast features? Church (2011) rationally pointed towards the evolution of such a synergy of methodologies, namely in NLP. For that, the EBMT formalism depended on both human rules and data. That is, the identification process of translation patterns, which are brought from data, is rule-driven (Bhattacharyya, 2015, p. 27). As it is situated between RBMT and SMT, EBMT integrates databases techniques and rules; yet, more facets need to be explored in order to distinguish it from SMT (Somers, 1999).

EBMT is based on the idea of analogy. That is, preprocessing the parallel texts is optional (Dajun & Yun, 2015; Vandeghinste et al., 2013). There are three major steps in EBMT: matching, transfer and recombination. Matching is the starting point to find translations (TL) for input fragments (SL) within the data storage of the phrase-, sentence-, or paragraph-aligned bilingual parallel corpora. The following step, transfer, includes extraction from the examples of matched fragments. The final step is where the translated fragments are generated according to the fragments of SL (Brockett et al., 2002). These steps are demonstrated in Figure 4.

To illustrate an EBMT system, Nagao (1984, reprinted in 2003) initiated setting sentences in Japanese with their English translations as examples presented to the MT system in order to learn structures and correspondences between SL and TL. With the same language pairs, Brockett et al. (2002) conducted an EBMT on abstract linguistic representations. Eventually, manual rule-curation steps were replaced by the automated machine learning process.

Figure 4

The Main Steps in EBMT



Until the 1990s, the limitations of MT centered on speed, cost, subject matter restrictions, and issues with vocabulary accuracy and consistency (Gross, 1992, p. 102). Shortly after several experiments on EBMT, Brown et al. (1990) introduced a more mathematics-and-probability-based approach known as Statistical Machine Translation (SMT).

4.3.2. Statistical Machine Translation

Statistical methods had been primarily used in ASR, lexicography, and other NLP applications (Brown et al., 1990, p. 79). The pioneers of SMT are IBM researchers as the starting point was in the late 1980s. Their contributions were a move in computational linguistics; they could overwhelmingly dominate the field. The central function of SMT is "to learn how to translate from a large corpus of pairs of equivalent source and target sentences" (Goutte et al., 2009, p. 2). Hence, the machine is trained on such an input-output framework. The statistical approach is based on assigning that every sentence in SL (S) has a possible translation in TL (T). Hence, pairs of sentences are deduced from the probability (p) equation formulated as P(T|S). Components of sub-

formulas of SL and TL probabilities are referred to as 'language model' (LM) and translation model (TM) respectively. The more sentences are well-formed, the higher LM probability is and the more concentrating TM is (Koehn et al., 2003a).

Based on bilingual corpora, the SMT approach started with treating words in sentences as strings, that is, word-based models (Brown et al., 1990, p. 84). Notwithstanding their merits, word-based MT systems could not properly deal with homonymy and gender linguistic instances. Then, Och and Ney (2000; 2002) improved a phrase-based system that came to process units (words sequence) by giving wights (score of probability) to phrase-pairs. Progressively, the experiments continued to test systems that process sentences with more complex syntactic structures, particularly those introduced by Yamada and Knight (2001), and Marcu and Wong (2002). Then, the n-gram-based models evolved to bring the phrase-based approach (as in Marino et al., 2006).

It is noteworthy to mention that Brown et al. (1990) presented 'alignment' as a critical step in the SMT process. It is defined as linking the parts of (S,T) pairs through indicating the TL words' origins. Bhattacharyya (2015) claims that "one reason behind the stupendous success of SMT as the reigning paradigm of machine translation is the discovery of the expectation maximization-based word alignment algorithm for SMT" (p. 27). In fact, alignment was developed at three levels: word, phrase and sentence. Brown et al. (1990) employed probability models so that "every sentence in one language is a possible translation of any sentence in the other" (p. 79). In this model, for every sentence pair (S,T), there exists a probability, Pr($T|S$). Brown et al.'s equation provides that:

$$Pr(S|T) = \frac{Pr(S)\ Pr(T|S)}{Pr(T)}$$

To simplify, words in SL have the probability Pr(S) to produce words in TL; the language model probability Pr(T) suggests an order of source words provided for the translation probability Pr(S|T). In line with this, Xiong and Zhang (2016) posit, in a log-linear model, that the best translation ***ê*** among all possible translations is found by the probability of a target sentence ***e*** (p. 2). For them, the basic SMT architecture, given ***f*** as a source sentence, is as follows:

$$\hat{e} = \underset{e}{argmax}\, P\,(e \mid f\,)$$

Indeed, SMT depends on two probabilistic models: a language model (LM) and a translation model (TM). The former relies on monolingual corpora for a more fluent output. The latter relies on bilingual corpora. TMs help in estimating how probable a target sentence ***t*** is a translation for the source sentence ***s***. The conditional probability equation ***P(t|s)*** is used for finding ***t*** of the maximum probability for ***s***. Then, the best translation ***ê*** is weighed as follows:

$$\hat{e} = \underset{t}{argmax}\, P\,(t \mid s\,) = \underset{t}{argmax} \underbrace{P(t \mid s\,)}_{translation\ model} \text{ x } \underbrace{P\,(t\,)}_{Language\ model}$$

The SMT paradigm, as presented briefly, relies on certain models that the machine statistically learns from parallel corpora and on decoded technical rules -commonly known as algorithms- to automatically translate instances from one natural language into another (Xiong & Zhang, 2016). By virtue of computer science development and the availability of largescale parallel texts, SMT witnessed substantial progress that led to shifting from word-based to phrase- and syntax-based SMT in the early 1990s.

Several approaches were developed within the SMT paradigm. Sentence-related approaches targeted its units: word, phrase and syntax. The approach that operates at the level of

individual words is known as Word-Based Statistical Machine Translation (WBSMT). It aligns and translates linguistic items based on statistical patterns observed in bilingual corpora. However, it may struggle with handling word reordering and complex sentence structures. With larger units, Phrase-Based Statistical Machine Translation (PBSMT) is involved. Under this approach, phrases are extracted from parallel texts and aligned to find translation equivalents (Marcu & Wong, 2002). Although PBSMT can handle limited word reordering, it offers improved translation quality compared to word-based models (Koehn et al., 2003a).

In addition, integrating syntactic structures into the automated translation process and considering grammatical relationships between words is done by Syntax-Based Statistical Machine Translation (SBSMT). For more coherent and accurate translations, SBSMT captures complex sentence structures and word reordering more effectively (Yamada & Knight, 2001). The aforementioned approaches triggered more researchers to develop SMT systems. As a way to address the limitations of traditional PBSMT, Hierarchical Statistical Machine Translation (HSMT) emerged to incorporate more linguistic structures into the translation process (Li et al., 2010).

HSMT, as introduced by Chiang (2005), represents a progression beyond the flat, phrase-level TMs by incorporating linguistic hierarchies and syntactic information to capture more complex sentence structures and linguistic nuances. HSMT systems share several features with standard PBSMT systems (Chiang, 2007). However, they are distinctive in the way they guide their reordering model through effective synchronous context-free grammar; thus, more flexible lexicalization and reordering occur. The milestone works in HSMT are attributed to Imamura

(2001) and Watanabe et al. (2002). The most popular HSMT systems are HiFST (Iglesias et al., 2009), cdec (Dyer et al., 2010), Travatar (Neubig, 2013), and Joshua (Post et al., 2015).

Significant steps forward in the SMT evolution has been marked with various projects and open-source toolkits. Bracketing Transduction Grammar -or BTG-based SMT- is another popular SMT formalism that could compete HSMT in terms of BLEU-measured translation quality (He et al., 2008). BTG-based SMT proved quality at hierarchical and long-distance reordering as well as synchronous context-free grammars. Through using PBMT for translation units and SBMT for hierarchal structures construction, BTG-based SMT could connect both statistical methods effectively (Xiong & Zhang, 2016, p. 15). Indeed, the multiplicity of approaches that CBMT brought could motivate researchers to combine them with classical MT systems. This became known as hybrid MT.

4.3.3. Hybrid Machine Translation

Using more than one MT method is known as Hybrid Machine Translation (HMT). Apparently, integrating numerous methods within a single MT system is made for the purpose of improving outputs, remedying defects and enhancing translation quality (Costa-jussà & Fonollosa, 2015; Labaka et al. 2014). HMT was mostly a combination of RBMT and SMT. One major MT system that employed HMT is PROMT (Molchanov, 2012).

HMT holds promise in addressing the limitations of individual MT paradigms. By leveraging their complementary strengths, it offers potential benefits such as enhanced translation accuracy, improved handling of domain-specific terminologies, and robustness against data scarcity (Mondal et al., 2023, p. 10140). However, integration complexities, system compatibility,

and optimization of system selection strategies are critical hurdles. After two decades of machine learning for SMT and HMT improvements, a paradigm shift in MT occurred.

4.3.4. Neural Machine Translation

By the turn of the third millennium, works in AI became more ambitious, and the future of NLP tasks seemed prosperous to experts. MT systems' designs, in the same path, developed. Evaluating the shared parameters in LMs and TMs and the mechanism of predicting equivalencies revealed that models cannot adapt to non-existing phrases and that, in turn, can lead to phrases dissimilarities and sparsity (Garg & Agarwal, 2018, p. 7). Bengio and Senécal (2003) proposed a revolutionary MT system where words have continuous fixed dimensions known as word vectors. Through neural networks, promising sparsity solutions have been dealt with. The extensive exposure to words helps the system's decoder learn the morphosyntactic and lexicosemantic aspects of the words. Prominent works that contributed to NMT were conducted by Collobert and Weston (2008) and Sutskever et al. (2011).

As an approach to overcome the limitations of SMT, using neural networks for MT was suggested earlier (Allen, 1987; Pollack, 1990; Chrisman, 1991; Forcada & Ñeco, 1997; Castano & Casacuberta, 1997). However, adopting this idea demonstrated its potential only with the increased computational power (Kalchbrenner & Blunsom, 2013). Unlike SMT, which involves distinct building blocks, NMT employs a single neural network to handle the entire translation process. Both SMT and NMT systems can be seen as evolutions of transfer-based MT systems from the 1980s. In SMT, transfer rules for semantics and syntax are segmented, while in NMT, these rules are abstracted into high-dimensional vector spaces that may be language-independent (Johnson et al., 2017).

The NMT system aims at predicting the likelihood of a word sequence through building and training a large artificial neural network. This process is referred to as ‘deep learning’ (DL) (Bahdanau et al., 2015). NMT addresses traditional MT systems’ shortcomings. Its major merits lie in its architecture and direct, automated learning (Wu et al., 2016, p. 1). Moreover, NMT systems depends on a larger number of parameters, also known as network patterns. The two common patterns are Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs). These are a set of states consuming SL sequences and generating translated outputs.

The major merit of NMT is that SL and TL texts can be trained in a fast and more accurate course. Notably, the early NMT methods are motivated from networks integrated in SMT. There are four NMT sub-approaches where TMs are involved: the Seq2Seq model (Sutskever et al., 2014), encoder-decoder model (Bahdanau et al., 2015; Cho et al., 2014), the attention mechanism (Cui et al. 2016), and the transformer model (Vaswani et al. 2017; Devlin et al. 2018).

The progress and evolution in NMT methods and models have been reviewed and studied by several researchers such as Garg and Agarwal (2018), Stahlberg (2020), Tan et al. (2020) and Sharma et al. (2023). Indeed, NMT addressed the limitations of RBMT and SMT methods through DL. In 2016, Google initiated “Google Neural Machine Translation” (GNMT) project (Wu et al., 2016), and it is considered the MT state-of-the-art. Other worth-mentioning notable NMT applications are Microsoft Translator, Meta AI Translate, and OpenNMT (Klein et al., 2017). Given that the empirical component of the present study relies on a specific NMT architecture, a comprehensive explanation of these particular models and techniques will be provided in Chapter 3 to contextualize the methodological approach.

5. Summary of Chapter

This chapter has chronicled the remarkable historical and methodological evolution of MT, (synthesized in the chronological timeline presented in Figure 5). As demonstrated, the field has progressed through four distinct paradigms: RBMT, EBMT, SMT and NMT, each defined by its own philosophical approach and punctuated by seminal contributions. The timeline highlights the most influential research that defines each era, from the Georgetown-IBM Experiment to the Transformer architecture. It serves as a representation of a field shaped by countless visionary pioneers from diverse disciplines, including mathematics, linguistics, and computer science.

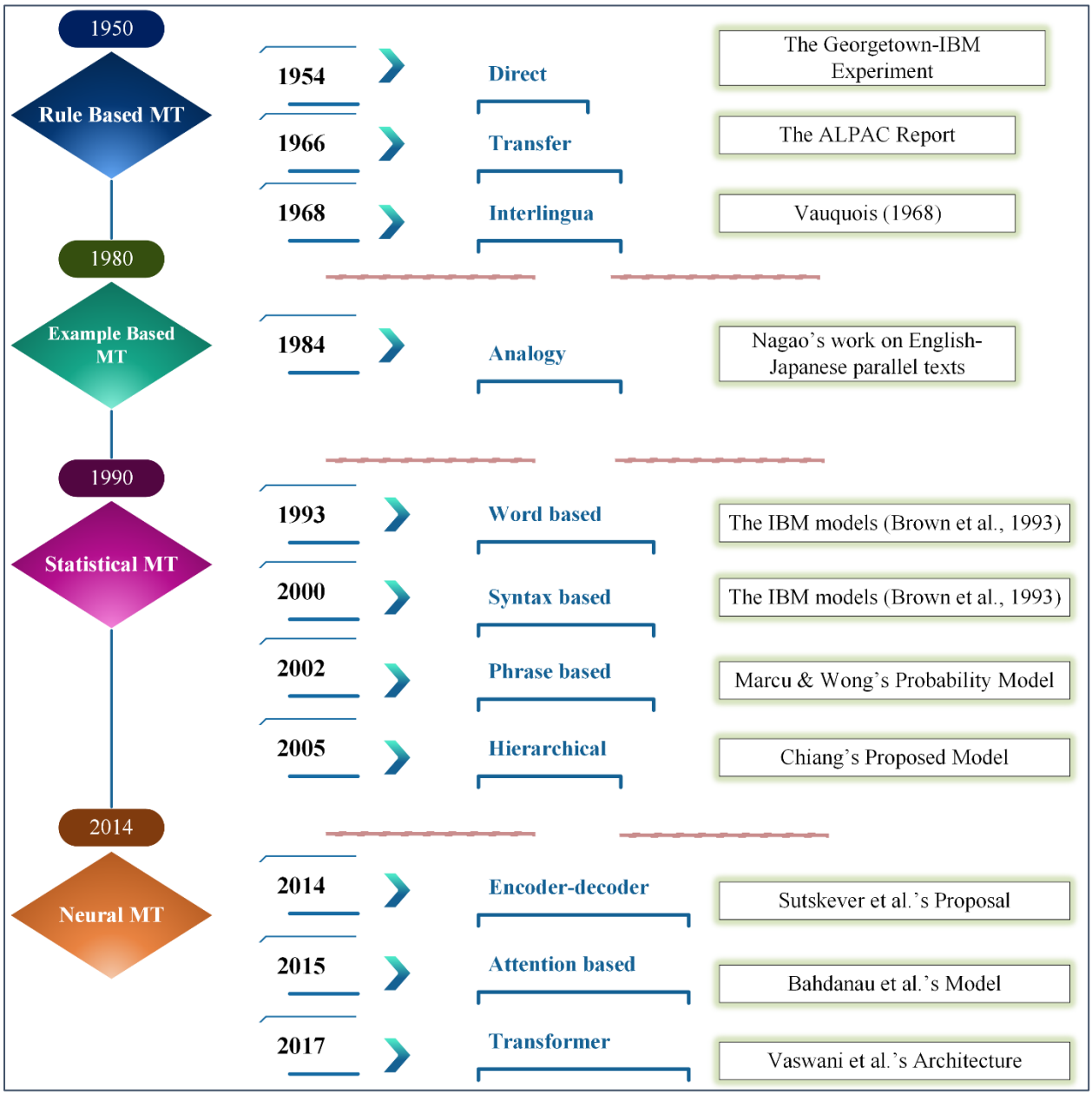
The narrative of MT can be distinctly partitioned into two major epochs demarcated by the pivotal ALPAC Report of 1966. The "first generation" of MT was dominated by the ambitious rule-based paradigm, which sought to encode human linguistic knowledge into direct, transfer, and interlingua systems. Following the ALPAC report's critical assessment, the "second generation" reflected a gradual shift towards more data-informed techniques, beginning with the example-based approach in the 1980s, which introduced the foundational concept of translation by analogy using parallel texts. This set the stage for the computational and technological enhancements of the 1990s, causing many early misconceptions about the MT limits to fade.

The subsequent eras represent a profound shift towards data-driven learning, commencing with the statistical paradigm. Pioneered by the IBM models, SMT matured through successive stages of increasing sophistication, evolving from word-based and syntax-based models to more powerful phrase-based and hierarchical systems. This trajectory culminated in the current state-of-the-art neural paradigm, which has itself undergone a rapid internal evolution from foundational encoder-decoder architectures to the introduction of the transformative attention mechanism and

the now-dominant Transformer. Both the SMT and NMT paradigms, though different in their technical underpinnings, represent the most linguistically informed and complex approaches to date, proving capable of handling intricate linguistic phenomena with unprecedented fluency.

Figure 5

Milestone Contributions throughout the History of MT



In tracing this history, this chapter has distinguished the core philosophies that have guided MT development, from early theoretical ambitions to the sophisticated, empirically-driven technologies of today. This evolutionary arc has recently turned towards one of the field's most pressing challenges: the effective automated translation of LRLs. Within this modern frontier, Arabic dialects emerge as a prominent and complex case study, primarily hampered by the scarcity of parallel corpora. It is in this context, building upon seven decades of innovation, that the present study is situated, aiming to contribute to the ongoing effort of extending the frontiers of MT to a historically underserved linguistic domain.

CHAPTER 2: MACHINE TRANSLATION OF ARABIC DIALECTS

Processing the natural Arabic classical and modern varieties is dated back to the initial EBMT experiments in the 1980s. With the progressive improvement in MT systems, various SMT and NMT models could raise the quality of the automated translation from and into Arabic. Nevertheless, MT experts worldwide were not limited to formal and standard languages. The extensive use of dialects, especially in online communication and content making, urged scholars to work on the improvement of dialects automatic translation. MT of Arabic dialects, for instance, can be considered as a recent activity whose age is around one decade. Earlier works focused more on processing and corpora building.

While progress has been made, achieving high-quality translations for all Arabic dialects remains a significant research area. It is a complex and challenging task due to the linguistic diversity and variations across different dialects. However, the present research work is concerned with dialectal Arabic, namely the Algerian dialect. Particularly, this chapter is devoted to briefly survey MT of Arabic dialects focusing on the Algerian case. Furthermore, the difficulties of processing Algerian dialects are discussed. The existing Algerian dialects corpora are described, and the MT experiments on them are reviewed.

1. Arabic Dialects Processing

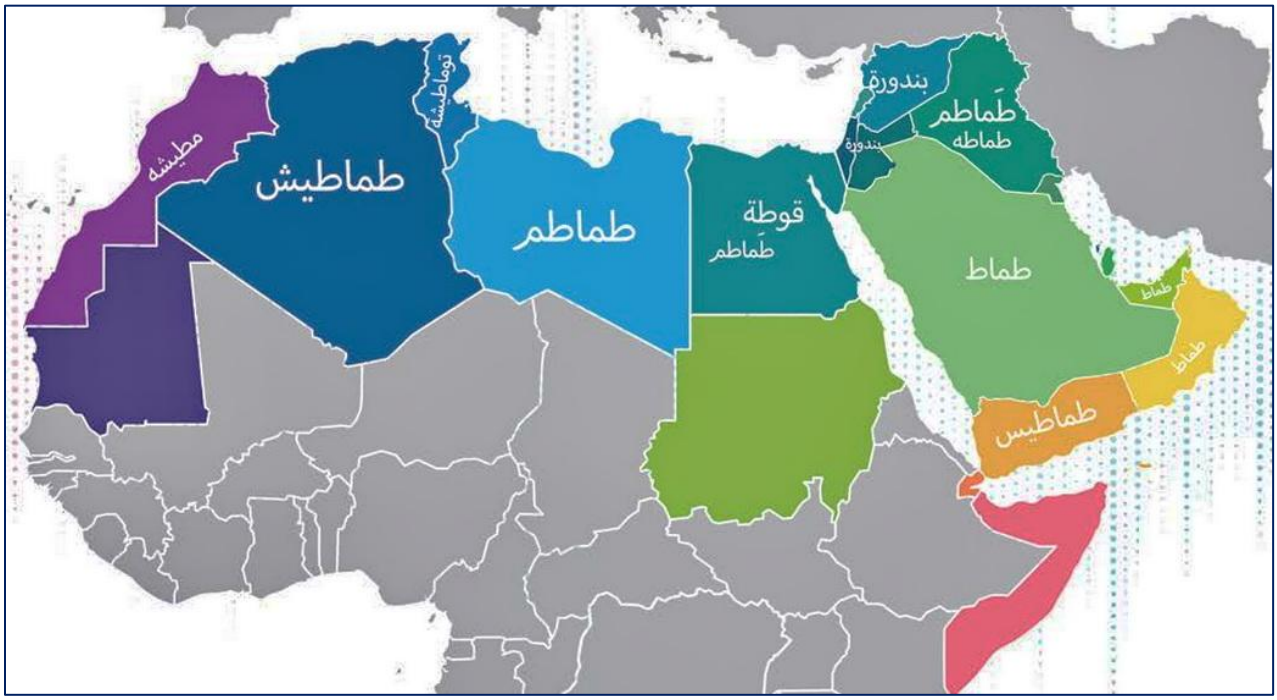
Arabic belongs to the Semitic language family. It is the fifth most spoken in the world by over 420 million people. As it is a diglossic language, the high or formal variety is Modern Standard Arabic (MSA), with a variety of spoken dialects. MSA is the official language of twenty-

two Arab countries situating in the common Middle East and North Africa (MENA) region. MSA is used in education, government, and the media whereas the spoken dialects -or vernaculars- are used in everyday communication (Suleiman, 2013).

Any differences that occur between standard and non-standard variations are based on vocabulary, syntax and pronunciation. MSA and its dialects propose a large number of differences that vary even across dialects and mainly by region. Figure 6 demonstrates the map of the Arab world with an illustration of linguistic variation across the major regional dialects.

Figure 6

An Illustrative Linguistic Variation across the Arab World



Note. The figure (adopted from Ali et al., 2017, p. 127) demonstrates the major lexical variations of the word 'tomato' (طماطم) from the east to the west regions/countries.

Indeed, the easter one moves within MENA, the closer the dialects are to MSA. The rules of these dialects differ according to regional, historical and sociolinguistic reasons. Researchers varied in their classification of Arabic dialects though Keleg et al. (2025) note that these regional distinctions are often fluid rather than discrete. According to Versteegh (2014), there are five major regional dialects in the Arab world. Table 1 presents these regions with country references.

Table 1

Arabic Dialects Continuum

	West ◄——————————————————————► **East**				
Regional Dialects	Maghrebi (MGB)	Egyptian (EGY)	Levantine (LEV)	Mesopotamian (IRQ)	Arabian Peninsula dialects (GLF)
Countries	Libya, Tunisia, Algeria, Morocco and Mauritania.	The dialects of the Nile valley: Egypt and Sudan.	Lebanon, Syria, Jordan and Palestine.	The dialects of Iraq.	Kuwait, KSA, Bahrain, Qatar, UAE, Oman and Yemen.

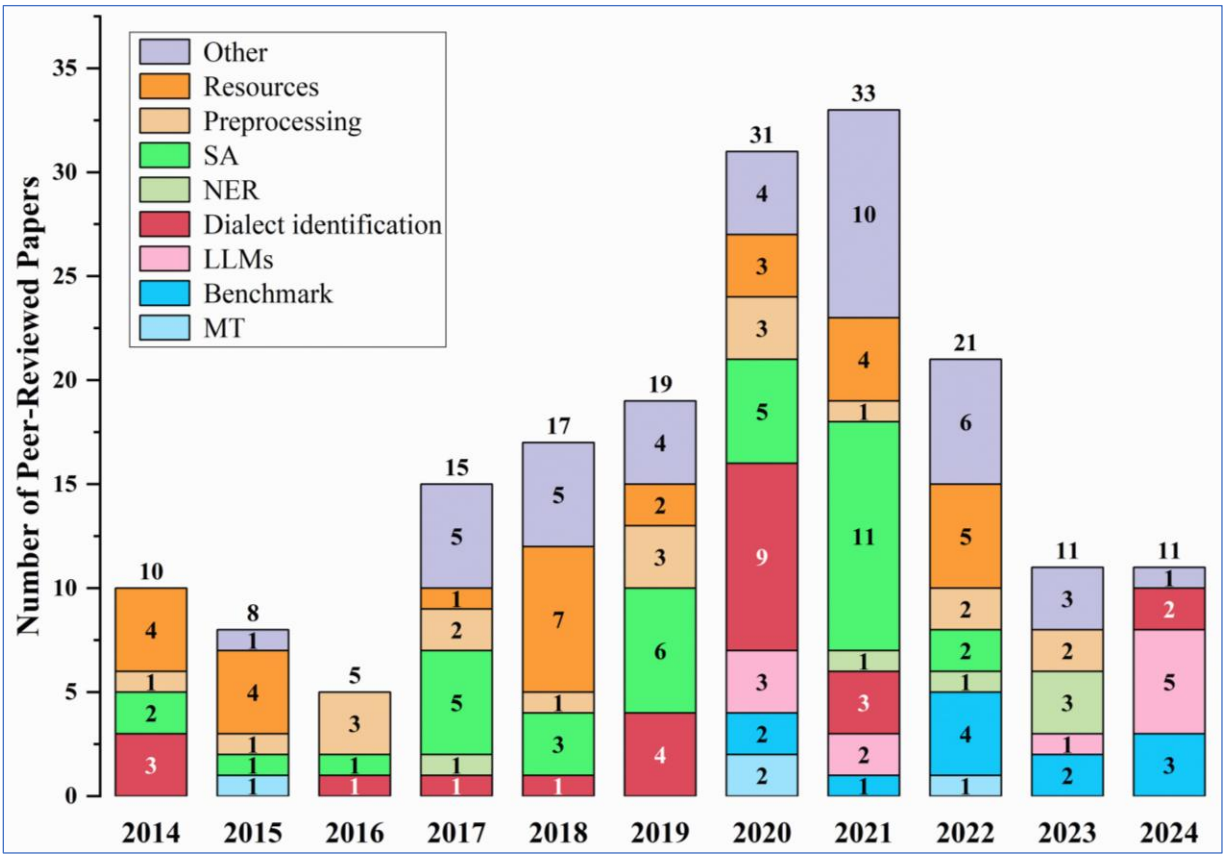
Such a diversity of dialects and sub-dialects represents an NLP challenge. That makes it difficult to develop MT systems that can translate across all the colloquiality of these dialects. In the last two decades, researchers devoted much efforts to improve Arabic dialect processing, build dialectal corpora, and enhance the qualities of experiments on Arabic dialects. Particularly, over the past decade, Arabic dialect processing has increasingly focused on a few dominant varieties and a core set of NLP tasks.

According to Dahou et al. (2025), mixed-dialect corpora, combining MSA with multiple dialects, have been the most frequently processed, reflecting a trend toward building broadly applicable models. Among individual dialects, MGB and EGY have been consistently represented,

with Algerian gaining notable prominence after 2018. LEV dialects appear regularly but less frequently, while GUL and IRQ have been addressed only sporadically. Figure 7 demonstrates the distribution of the prominent works on Arabic dialect processing highlighting the major NLP tasks.

Figure 7

Trends in Arabic Dialect NLP Tasks from 2014 to 2024



Note. Adopted from Dahou et al. (2025, p. 5). Selection criteria are as defined by the authors.

On the task level, SA has been the most researched application, peaking in 2020–2021, driven by interest in opinion mining from social media data. Other widely explored tasks include dialect identification, event detection, and PoS tagging, the latter gaining momentum after 2017. NER, preprocessing, and resource creation have been steady research areas, underpinning the

development of more advanced tools. More recently, LLMs have emerged as a growing focus, especially from 2018 onwards. By contrast, tasks such as hate speech detection, rumor detection, and text classification remain less common but show gradual growth, reflecting interest in socially aware and safety-related NLP applications. This distribution reveals a dual trend: a concentration on resource-building and sentiment-centric applications, and a gradual diversification into specialized tasks and underrepresented dialects.

While Dahou et al.'s (2025) survey offers a comprehensive and timely synthesis of research on Arabic dialect processing, its methodological scope is not without limitations. By prioritizing only highly cited and ostensibly "impactful" works, the review omits a number of prominent contributions that, although less widely cited, have nonetheless introduced innovative methodologies or addressed underexplored dialects. Furthermore, the exclusion of studies in which Arabic dialect processing is integrated with foreign-language data (cross-lingual alignment, code-switching analysis, or multilingual model training) results in an incomplete portrayal of the field's breadth. This omission is particularly significant given that cross-linguistic approaches play a pivotal role in tasks such as MT, where the interplay between Arabic dialects and other languages is central. These gaps underscore the need to examine research that explicitly engages with cross-dialectal as well as multilingual MT scenarios.

2. A Survey on Arabic Dialects MT

There has been significant progress in the automated translation of several Arabic dialects in recent years. The development of new corpora and MT techniques for Arabic dialects has been surveyed by Harrat et al. (2019a) and Babaali and Salem (2023). Particularly, the present section reviews the notable MT research works and linguistic resources for Arabic dialects. It is important

to note that research focused earlier on dialect-MSA MT, then on dialect-English MT. Nonetheless, in the course of improving Arabic dialects processing, some researchers worked on building tools that disambiguate their morphology.

Morphological analyzers, also known as segmenters, are designed tools to dissect the structure of words, breaking them down into their constituent morphemes and clitics, such as roots, affixes, and grammatical features (Al-Sughaiyer & Al-Kharashi, 2004). Understanding the morphology of words is of paramount importance in Arabic dialects due to the intricate nature of their linguistic features, where slight modifications can completely alter the meaning of a word. The significance of morphological analyzers in processing Arabic dialects lies in their ability to facilitate MT tasks. Several segmenters have been built and developed to process Arabic dialects (see Table 2); however, the majority of them targeted EGY, LEV and GLF.

Table 2

Notable Morphological Analyzers for Arabic Dialects Processing

Research works	Analyzers	Dialects
Habash and Rambow (2006)	MAGEAD	LEV
Habash, Eskandar and Hawwari (2012)	CALIMA	EGY
Almeman and Lee (2012)	AlKhalil	GLF
Zribi et al. (2013)	AlKhalil	TUN
Habash et al. (2013)	MADA-ARZ	EGY
Pasha et al. (2014)	MADAMIRA	EGY
Salloum and Habash (2014)	ADAM	Multi-dialectal
Hamdi et al. (2015)	MAGEAD	TUN
Khalifa et al. (2016)	YAMAMA	Multi-dialectal

Harrat et al. (2016)	BAMA	ALG
Al-Shargi et al. (2016)	ALMOR	MOR / YEM
Eskandar et al. (2016)	ALMOR	LEV / EGY
Khalifa et al. (2017)	CALIMA	GLF
Obeid et al. (2022)	CAMELIRA	Multi-dialectal

Critiques of existing Arabic dialect morphological analyzers highlight several limitations that affect their overall utility. Firstly, many of these analyzers suffer from a limited database (especially ALMOR), which restricts their ability to accurately analyze and generate morphological forms for a wide range of dialect words. Additionally, a significant drawback of such tools as CALIMA is their focus primarily on verbs, neglecting other parts of speech like nouns and adjectives, which limits their scope in analyzing the entirety of dialect text. Moreover, these analyzers often struggle with capturing the essential element of text context, resulting in lower accuracy and potential misinterpretations of words within sentences or discourse.

Furthermore, extending or updating these analyzers can be a time-consuming process due to manual annotations (as in MADAMIRA) and the need to add entries to the database, making them less adaptable to evolving dialects. Another critical limitation is the absence of a syntactic component (the case of MAGEAD, which hinders a deeper understanding of sentence structure and relationships between words. Lastly, the use of commercial components in such analyzers as AlKhalil can impose cost constraints and licensing issues. Other detailed critiques of these analyzers have been provided by recent reviews such as Alothman and Alsalman (2020) and Tachicart et al. (2022). Despite all the disadvantages, the need for further development and refinement is still motivating researchers to overcome common limitations and enhance the effectiveness of these analyzers.

2.1. Arabic Dialects-MSA

Arabic inter-dialectal MT has improved by virtue of segmenters. Yet, the efforts are needed for translating these dialects into MSA and ENG. There have been several research works that practically experimented Arabic dialects-MSA MT. For instance, Bakr et al. (2008) proposed a sentence-based MT from EGY into MSA. They used a hybrid approach: (1) SMT for sentence segmentation and tagging, and (2) RBMT for generating diacritized sentences relying on a bilingual dictionary. Through manual segmentation and tagging, their study was based on a corpus of 1K dialectal sentences.

A second prominent contribution was by Salloum and Habash (2012) who introduced Elissa. The latter is a RBMT system that translates EGY, LEV, IRQ, and to a lesser extent, GLF into MSA. Following an identification phase conducted by the user, Elissa converts dialectal words into paraphrases in MSA using the dialectal morphological analyzer ADAM (Salloum & Habash, 2011), morphological transfer rules, and dialect-standard Arabic dictionaries. The generated paraphrases form a lattice, from which the best hypotheses are selected through a 5-gram language model trained on a corpus of 200M words.

Further, Mohamed et al. (2012) presented a RBMT system that translates from MSA into Egyptian dialectal Arabic. The system was evaluated in terms of morpho-syntactic tagging and OOV coverage. In terms of tagging, the recall rate improved from 73.24% to 86.84%, while the OOVs rate decreased from 28.98% to 16.66%.

Due to the lack of parallel data, some researchers depended on monolingual corpora in their MT experiments. For instance, Almansor and Al-Ani (2017) proposed word embeddings to

capture semantic similarities between EGY and MSA. Such a method required monolingual data for MSA and EGY. Data preprocessing steps like tokenization and normalization were applied. However, the authors did not mention the size of trained data. For that, Alnassan (2023) proposed strategies for the automatic standardization of dialects to bridge the gap between the vast diversity of Arabic dialects and MSA.

Faheem et al. (2024) proposed a semi-supervised approach to deal with EGY-MSA MT. With the largest dataset for this language pair, over 15M EGY sentences were collected. As for MSA data, they reached 20M sentences. In addition, a parallel corpus of 40K EGY-MSA sentences was created. Similar to a previous paper (Faheem et al., 2023), the researchers employed three methods: (1) a Seq2Seq model for EGY-MSA shared vocabulary, (2) an unsupervised transformer model for EGY monolingual corpus, and (3) a hybrid system that trains the supervised EGY-MSA parallel data and incorporates the monolingual data subsequently. To regularize and handle OOVs, the authors used different techniques. The NMT hybrid semi-supervised approach achieved the best BLEU scores of 29.5. As a solid contribution to the application of semi-supervised NMT to dialectal Arabic, this research work demonstrates that even unsupervised monolingual corpora may lead to better MT performance.

Likewise, Yemeni (YEM) and Jordanian (JOR) dialects received attention from NLP experts to be automatically translated into MSA. For the Sana'ani dialect of Yemen, Al-Gaphari and Al-Yadoumi (2010) also adopted a rule-based approach that translates YEM into MSA. The system achieved a recall rate of 77.32% on a test corpus of 9,386 words.

As for the JOR-MSA case, Al-Ibrahim and Duwairi (2020) utilized the RNN encoder-decoder model for word-level (24.2K word pairs) and sentence-level (500 sentence pairs)

translations. The reported results proved an accuracy of 91.3% for word-to-word MT and an accuracy of 63.2% for sentence-to-sentence MT. The significance of the latter experiment by Al-Ibrahim and Duwairi (2020) lies in its exploration of a relatively underexplored area. Translating a specific Arabic dialect like JOR into MSA using DL techniques encourages the potential of the proposed framework in enhancing the translation accuracy.

Farhan et al. (2020) presented novel NMT systems for translating dialectal Arabic into MSA. Their key innovations were using shared word embeddings and leveraging subword information. The authors evaluated the systems in both supervised and unsupervised settings. The former involved JOR-MSA training whereas the latter involved training SAU and EGY datasets. Using shared word embeddings to enhance cross-dialect translation was novel and promising. Besides, the unsupervised experiments demonstrated the potential of the proposed approach. Indeed, the systems achieved high BLEU scores, especially in the unsupervised setting. However, Farhan et al. (2020) did not compare their results to other relevant multi-lingual NMT systems.

Furthermore, there were some experiments that targeted MGB (the Maghrebi dialects, namely MOR, TUN and ALG). As for the Moroccan dialect(s), Tachicart et al. (2014) built the Moroccan Dialect Electronic Dictionary (MDED) which was then experimented through an HMT approach by Tachicart and Bouzoubaa (2014). The latter employed an approach that combines rule-based methods with a language model. Their MT system was based on morphological analysis with Alkhalil morphological analyzer (Boudlal et al., 2010), adapted and enriched with the affixes of the Moroccan dialect, and a bilingual dictionary constructed from television scripts and web pages.

After an identification step that separates the dialectal content (MOR) from standard Arabic, the text is analyzed and segmented into annotated dialectal units. These units are then aligned with one or more corresponding MSA units using the bilingual dictionary. In the generation step, MSA sentences are generated and passed to the language model to produce the most accurate sentences.

According to TUN-MSA MT experiments, Hamdi et al. (2013) proposed an MT system to process verbal forms in Tunisian dialect. It was based on a deep morphological analysis using root and pattern. Their initial experiment on TUN-MSA was not mainly MT-guided. Rather, it was on generating MSA that is approximate to TUN. Similar to the approach employed by Sawaf (2010), Mohamed et al. (2012), and Salloum and Habash (2013), Hamdi et al. (2013) added the deep morphological analysis based on MAGEAD (Habash & Rambow, 2006). The system translates in both directions (MSA into TUN and vice versa). Its recall rate is 84% for MT from TUN and 80% for MT from MSA.

Within a dialect identification project, Sadat et al. (2014) conducted a TUN-MSA RBMT experiment. They collected 6K words from social media and blogs and the TUN-MSA manually translated parallel corpus contained fifty pairs of sentences for MT evaluation. However, the score of TUN-MSA translation experiments was low. A prominent MT project is the MuDMaT (Multi-Dialect Machine Translation) Project (2014-2017). It was devoted to the automatic translation of Arabic dialects, particularly the Maghrebi dialects (ALG, MOR and TUN), MSA and FRE. According to Sadat (2015), an initial RBMT system was developed to translate TUN into MSA and FRE.

In addition, Sghaier and Zrigui (2020) experimented TUN-MSA MT utilizing a rule-based approach. Their MT process went through morphological analysis and disambiguation, lexical and structural transfer, and morphological generation and spelling corrections, ultimately yielding the translated text in MSA. The test set was 100 sentences and the MT evaluation brought promising scores that suggest the effectiveness of their proposed system.

Recently, Torjmen and Haddar (2021) addressed the development of automatic Named-Entity Recognition (NER) and MT systems for the Tunisian dialect. Recognizing the challenges posed by the linguistic differences between TUN and MSA, the morphosyntactic analysis was performed thoroughly by Torjmen and Haddar (2022) who proposed a method based on a bilingual dictionary and a set of elaborated local grammars. These local grammars are transformed into finite-state transducers using the NooJ linguistic platform. The system was tested on a corpus of 20K words, yielding ambitious results. Both studies contribute to bridging the understanding gap for non-Tunisians by effectively recognizing named entities (NEs) in TUN.

More recently, Kchaou et al. (2023) utilized neural networks methodologies for TUN-MSA MT. They proposed an approach to create a parallel corpus with data augmentation methods to effectively increase the size and quality of the corpus. Indeed, this was initiated earlier by Kchaou et al. (2020) for an SMT experiment. The size of the augmented corpus reached 32K sentences in the dialect. Besides, Kchaou et al. (2023) compared the performance of CNN, RNN, and transformer models for the TUN-MSA translation task, revealing that the transformer model outperformed the others with a BLEU score of 60. Their MT experiment highlights the effectiveness of data augmentation, particularly for a low resource variety like TUN.

Alhammi and Haddar (2022) proposed a method for building a bilingual dictionary to translate words and phrases from the Libyan dialect (LBY) into MSA. They divided LBY words into shared words that also exist in MSA versus other specific words from the dialect to form the dictionary entries manually. The bilingual dictionary included over 2K entries with their equivalents in MSA. However, the researchers did not provide a practical experiment or evaluation for their dictionary.

Small-sized corpora could urge researchers to use multi-task learning (MTL) in their MT experiments. It is based on encoder-learning strategy across multiple language pairs. For that, it is adopted for inter- or cross-dialectal jointly datasets for a better generalization. The level of similarities between Arabic dialects increases the MTL effect on the MT quality (Moukafih et al., 2022, p. 584). Two notable research works that have adopted this approach in the literature. First, Baniata et al. (2018a) whose corpus contained 13805 pairs of LEV-MSA sentences along with 17736 pairs of MGB-MSA sentences. For the test sets, 2K pairs of sentences were used from each parallel corpus. They found that MTL models outperform the independently-trained models. In a consecutive study, Baniata et al. (2018b) examined the influence of PoS tagging and MSA-ENG MT tasks on the translation of dialects. Indeed, the MTL model with PoS tagging was found to improve the NMT of experimented dialects into MSA.

Second, Moukafih et al. (2022) explored MTL strategies to capitalize on the relationships between Arabic dialects, aiming to enhance the MT performance. Their experimental outcomes were particularly impressive and could exhibit notable improvements over traditional MT methods. Moukafih et al. (2022) underscored the effectiveness of MTL by reporting that, out of 42 MT experiments, this approach outperformed SMT in 88% of cases, even when utilizing a

relatively small corpus. The viability of such newly-employed approaches and strategies in Arabic dialects MT opens avenues for further exploration and development in the domain of processing Arabic dialects.

In this respect, the prominent contributions are classified by year in Table 3 along with the adopted MT approach(s) and corpus description.

Table 3

Notable MT Research Works (Arabic Dialects-MSA)

Research works	SL-TL	Approach(es)	Corpus
Bakr et al. (2008)	EGY-MSA	SMT + RBMT	1K sentences
Al-Gaphari and Al-Yadoumi (2010)	YEM-MSA	RBMT + stemming	9386 words from Sana'ani Dialect
Salloum and Habash (2012)	LEV/EGY/IRQ/GLF-MSA	RBMT + morphology analysis	-
Mohamed et al. (2012)	MSA-EGY	RBMT	292 comments
Hamdi et al. (2013)	TUN-MSA	RBMT + morphology analysis	1,5K pairs of tokens
Sadat et al. (2014)	TUN-MSA	RBMT + bilingual dictionary	50 sentences
Tachicart and Bouzoubaa (2014)	MOR-MSA	SMT + bilingual dictionary	MDED (18K pairs of words)
Baniata et al. (2018a)	LEV/MGB-MSA	NMT (through MTL)	15.8K LEV-MSA and 19.7K MGB-MSA sentences
Al-Ibrahim and Duwairi (2020)	JOR-MSA	NMT (RNN)	24.2K word pairs and 500 sentence pairs
Farhan et al. (2020)	JOR-SAU-EGY-MSA	NMT + shared word embeddings	309K sentences (all sets)
Sghaier and Zrigui (2020)	TUN-MSA	RBMT + morphology analysis	100 sentences

Torjmen and Haddar (2021)	TUN-MSA	RBMT + bilingual dictionary	20K words
Moukafih et al. (2022)	MGB/LEV-MSA	SMT + NMT (through MTL)	PADIC (6.1K parallel sentences)
Kchaou et al. (2023)	TUN-MSA	NMT (RNN, CNN and Transformer models)	175K (Baseline 34K) TUN-MSA parallel sentences
Faheem et al. (2024)	EGY-MSA	NMT (supervised and unsupervised)	Over 15M EGY sentences, 20M MSA sentences and 40K EGY-MSA sentences

2.2. Arabic Dialects-ENG

MT of Arabic dialects into English is as recent as MT of Arabic dialects into MSA. However, MT into English was not always direct. That is, some studies used MSA as a pivoting language. Due to several challenges inherent in dialects themselves, using MSA to pivot the translation task was a common workaround before advancements in technology and resources. In this section, major research works and projects that covered MT of Arabic dialects into English are reviewed.

Al-Mannai et al. (2014) utilized unsupervised morphological segmentation (using Morfessor tool) of Qatari dialect (QAT) to enhance the performance of the Qatari-English MT. dialect translation system into English. An improvement was realized at the MT system's performance compared to a system without any segmentation and to a system with ATB segmentation. A preliminary SMT system was presented using an unsegmented parallel corpus of 13K sentence pairs. Besides, a multi-dialect segmentation model was learned from a multidialectal corpus (QAT, EGY, LEV, and MSA). The BLEU score of this system improved by 1.5 points compared to the same system without segmentation.

Sawaf (2010) employed an HMT system that combines both statistical and rule-based approaches. This system translates texts in dialectal Arabic (web content and transcriptions of broadcasts) into ENG using MSA as a pivot language. Dialectal texts are normalized using character-level transfer rules. Sawaf (2010), moreover, used morphological analyzers for both dialectal and standard varieties. The study used a corpus of 14M sentences covering a wide range of dialects, including Levantine dialects (LEB, SYR, PAL, and JOR), Gulf Arabic dialects (IRQ, SAU, and the southern Arabian Peninsula), the Nile region (EGY and SUD), and Maghrebi dialects (LBY, MOR and TUN). The results demonstrated that the hybrid approach outperforms both the individual statistical and rule-based approaches. Additionally, the normalization of dialectal texts (training and test data) improves the BLEU score by 2% for web-extracted text and 1% for broadcasted content and information.

Salloum and Habash (2013) evaluated Elissa manually, and 93% of the Arabic sentences generated by Elissa were deemed correct. Furthermore, Elissa was utilized as a pivot tool for dialectal Arabic-English MT. Such a pivoting method raised the enhancement level of the MT system by 1.4%. They constructed a multi-dialectal (IRQ, LEV, GLF and EGY) corpus of 7345 sentences. Salloum and Habash (2013) validated the MSA outputs by Elissa, yet vocabularies are still inputs from the dialects, therefore, the MT system produced mistranslations.

A prominent EGY-ENG SMT system was presented by Sajjad et al. (2013). A parallel corpus of 38K sentence pairs was utilized for training, development and testing. The study employed morphological and phonological transfer rules to align EGY with MSA at the word level. These transformations reduced the OOVs rate from 5.2% to 2.6% and improved the BLEU

score by 1.87 points. Such an MT enhancement occurred because EGY was mapped through an automatic conversion method to be closer to MSA.

In the opposite direction, and using MSA as a pivot language, Jeblee et al. (2014) presented a system that translates ENG into EGY. Based on an English-Arabic SMT system learned from a corpus of five million sentence pairs, the researchers used domain adaptation techniques. For adaptation purposes, a tri-parallel corpus (ENG, MSA and EGY) of 100K sentences was created using transfer rules. In this research, the method for generating EGY from morphologically disambiguated MSA sentences by Mohamed et al. (2012) was extended to automatically obtain EGY sentences. The study demonstrated the adaptation between EGY and MSA as it highlighted how this adaptation improves the MT system's performance.

The quality of EGY-ENG MT was enhanced by reducing the OOVs rate in the study of Durrani et al. (2014). The adopted approach involved converting EGY to MSA by using an LLM to select the best standard Arabic alternatives for dialectal OOVs via a beam search algorithm. The results then are translated into English using an SMT system. It was found that spelling correction improves the BLEU score by 1.7 points compared to a basic system that translates EGY directly into English without editing.

Some MT experiments were parts of particular research projects. For instance, the Defense Advanced Research Projects Agency (DARPA) initiated the BOLT program (Broad Operational Language Translation). This project aimed to develop new techniques for the automatic translation and linguistic analysis applicable to informal genres of text and speech in online or personal communication. Indeed, the BOLT project focused on MT between English and other varieties, among which there is the Egyptian Arabic dialect. It encompassed three technical domains: the

development of algorithms and MT systems, data collection and corpus compilation, and evaluation.

As part of this program, Zbib et al. (2012) constructed two parallel corpora: LEV-ENG (1.1M words) and EGY-ENG (380K words) by translating parts extracted from a large web corpus of Arabic texts into English. Dialect classification and translation were conducted using crowdsourcing. Several experiments were performed with SMT systems trained on these corpora, along with a parallel MSA-ENG corpus (150M words in MSA). The researchers used the morphological analyzer MADA (Habash & Rambow, 2005) for segmentation, and that improved the MT quality.

Zbib et al. (2012) also demonstrated that translation systems trained on corpora combining dialectal and standard Arabic data tend to yield better results, which is natural given that dialectal texts are mostly a mixture of standard and dialectal Arabic. Regarding the use of MSA as a pivot language to translate LEV, the authors noted a 2.3-point improvement in BLEU score for the initial experiments with an extended dialectal corpus. However, after adding more dialectal data to the training corpus (400K words), direct MT brought better results despite the relatively low rate of OOVs when the system pivoted through MSA.

Aminian et al. (2014) addressed the issue of dialectal OOVs in the context of Arabic dialects-English MT. They adopted an approach that normalizes dialectal words into MSA words using AIDA (Automatic Identification for Dialectal Arabic) as a tool for dialect identification and glossing (Elfardy & Diab, 2012; Elfardy et al., 2014). They also used the morphological analyzer MADAMIRA (Pasha et al., 2014) for the replacement of OOVs with MSA equivalents. The training datasets contained 29M MSA and 5M dialectal Arabic tokenized words. This approach

improved the SMT quality by an absolute BLEU score increase of 0.4% and 0.3% for AIDA and MADAMIRA, respectively.

Furthermore, in the BOLT program, Aransa (2015) implemented various techniques for experimenting EGY-ENG MT. Since the BOLT project did not cover EGY only, the domain adaptation technique was adopted to train corpora. Therefore, the SMT system performance was enhanced by treating different dialects as separate domains. Adaptation techniques were applied to LMs and TMs. Aransa (2015) evaluated different morphological segmentation schemes and addressed the OOVs issue through proposing a technique involving the transliteration of rare words, such as proper nouns.

Indeed, the Egyptian dialect was a target for most of the MT experiments on Arabic dialects. Sajjad et al. (2016) handled the morphological analysis and Arabic Treebank (ATB) segmentation of EGY data. They addressed the unique challenges of translating this specific dialect into ENG using a PBSMT approach. The combination of several techniques such as segmentation, domain adaptation and transliteration models enhanced the MT quality while using MSA as a pivot language.

A notable research work in the Arabic dialects-English MT literature that targeted the Levantine dialect is conducted by Hassan et al. (2017). They proposed a novel approach to translate LEV into ENG through word embeddings. Besides, the researchers created synthetic parallel data from MSA; they generated LEV-ENG corpus from MSA-ENG parallel data of 3M sentences. Such an approach of effective data generation technique by exploiting word embeddings and existing parallel data in related languages gave over 2.8 BLEU score improvement for an NMT system. The resulting system even outperformed a baseline trained on human-translated test data.

Tawfik et al. (2019) tackled leveraging more abundant MSA resources to translate the dialects into English. They conducted NMT experiments on EGY, LEV and GLF through using a sufficiently accurate morphology-aware segmentation. The latter helped the NMT model better leverage the common stems/roots between MSA and dialectal words, while learning the dialectal prefixes/suffixes from the limited SL data. Based on some existing corpora, Tawfik et al. (2019) made a compelling case for the morphological analysis to improve MT for resource-constrained language varieties.

In addition, Abid (2020) addressed a critical gap in low-resource Arabic dialects MT research. He developed benchmark EGY and LEV datasets for SMT experiments. The researcher depended on bootstrapping to augment the data and adopted a multilingual training approach. For optimal unsupervised segmentation parameters for the aforementioned dialects, baselines for ENG into EGY and LEV MT were improved. Along with MSA, this research proved that multilingual datasets would achieve better scores.

Furthermore, Nagoudi et al. (2021) investigated MT from code-mixed EGY and MSA into ENG. Along with MSA-ENG parallel corpora, the collected data from different sources reached 56K EGY sentences with their ENG equivalents. They employed Seq2Seq transformer models trained from scratch and other pre-trained LMs. The latter performed well on the code-mixed MSA-EGY translation task. Eventually, the code-mixed EGY-ENG test set achieved a BLEU score of 25.72.

Recently, Gaser et al. (2023) explored segmentation approaches for a code-switched EGY-ENG NMT. As code-switching, mixing languages within a sentence, presents challenges for NMT due to data sparsity and morphological complexity. Segmentation aims to address these issues by

breaking down code-switched sentences into smaller, monolingual segments. They evaluated three categories of segmentation approaches: (1) supervised morphology-based using MADAMIRA, (2) unsupervised morphology-based using segmenters such as MorphA-Gram and Morfessor, and (3) unsupervised frequency-based. The NMT transformer architecture was trained on each segmentation setup to extensively analyze different data sizes and sentence types of high and low code-switching. Gaser et al. (2023) concluded that frequency-based segmentation performed better in low-resource settings as they proved that higher code-switching led to lower MT quality.

Similarly, Hamed et al. (2023) compared different data augmentation approaches to improve MT of code-switched EGY into ENG depending on the same datasets in the study of Gaser et al. (2023). However, three techniques were compared: (1) lexical replacements of words in the SL with their equivalents in TL, (2) linguistic rules to code-switch parts of sentences, and (3) back-translation which involves translating ENG texts to code-switched ENG. Augmentation approaches were implemented with varying degrees of reliance on parallel code-switched data. The NMT system performance was measured on both MT quality (BLEU score) and augmentation quality (human evaluation of fluency and naturalness). Although the researchers found that back-translation and predictive lexical replacement perform best for both MT and augmentation quality, the high percentages of code-switched words in augmented data did not lead to an improvement in the MT performance.

The need to develop MT systems that can handle Arabic dialects, along with a lack of datasets to support the development of Arabic multi-dialect translation systems, motivated Alzamzami and Saddik (2023) to construct a valuable resource for MT studies. Their proposed English-multidialectal Arabic dataset contains 15K English sentences translated into four major

Arabic dialects: GLF, YEM, IRQ and LEV (total size is 75K sentences). Transformer neural network and Seq2Seq models were employed. MT experiments showed superior performance of the former approach.

Recently, Kchaou et al. (2025) have presented a significant contribution to the NMT of TUN into ENG with a pivot approach using MSA as an intermediate language. Their primary contribution is the development of a comprehensive trilingual TUN-MSA-ENG corpus, which they leveraged to train their models. With data augmentation strategy, the study corpus reached 188K parallel sentences. As found by Kchaou et al. (2025), the direct TUN-ENG translation model achieved a BLEU score of 41.78. Moreover, the best performance, with a notable BLEU score of 67.89 was achieved with MSA as a pivot language. This demonstrates that deep neural network architectures can effectively handle the complexities of informal, dialectal languages with limited parallel data.

The provided comprehensive review of existing literature reveals a diverse landscape of scholarly contributions. However, EGY and LEV received the highest portion of concern. The review encapsulates a synthesis of various works, encompassing the expertise of different NLP experts over the last decade. Table 4 presents a structured overview, delineating pertinent details such as authors, publication years, characteristics of corpora employed, and the distinct approaches adopted in each study.

Table 4

Notable MT Research Works (Arabic Dialects-ENG)

Research works	SL-(pivot)-TL	Approach(es)	Corpus Description
Sawaf (2010)	LEV/GLF/EGY/MGB-ENG	SMT + RBMT	14M sentences with web data
Zbib et al. (2012)	LEV/EGY-(MSA)-ENG	SMT + morphology segmentation	8M MSA-ENG words and 380K LEV-EGY-ENG words
Salloum and Habash (2013)	LEV/EGY/IRQ/GLF-ENG	SMT + RBMT	7345 sentences (all sets)
Sajjad et al. (2013)	EGY-(MSA)-ENG	SMT + morphology analysis	38K EGY-ENG sentences and 200K MSA-ENG sentences
Jeblee et al. (2014)	ENG-(MSA)-EGY	SMT + dialect adaptation system	100K sentence tri-parallel corpus
Al-Mannai et al. (2014)	QAT-(MSA)-ENG	SMT + unsupervised segmentation	13K QAT sentences
Durrani et al. (2014)	EGY-(MSA)- ENG	SMT + mapping and convolution model	8.5K EGY-ENG sentences and 300K MSA-ENG sentences
Aminian et al. (2014)	EGY-(MSA)-ENG	SMT + morphology analysis	29M MSA words + 5M EGY words
Aransa (2015)	EGY-ENG	SMT	Different corpora (5M MSA words, 2M EGY words and 7.9M ENG words
Sajjad et al. (2016)	EGY-(MSA)-ENG	SMT + different models	9K sentences
Hassan et al. (2017)	LEV-ENG	NMT + word embeddings	160K sentences (baseline) / 6K synthetic LEV-ENG sentences
Tawfik et al. (2019)	GLF/LEV/EGY-ENG	NMT + word segmentation	38K (GLF) / 219K (LEV) / 502K EGY sentences
Abid (2020)	ENG-(MSA)-EGY-LEV	SMT + data augmentation	1.49M MSA sentences, 402K EGY sentences and 138K LEV sentences
Nagoudi et al. (2021)	Code Switched EGY-ENG	NMT	56K sentences
Gaser et al. (2023); Hamed et al. (2023)	Code Switched EGY-ENG	NMT (augmentation); NMT (segmentation)	6213 sentences (66.9% are code-mixed) and 309K EGY-ENG sentences

Alzamzami and Saddik (2023)	ENG-GLF/YEM/IRQ/LEV	NMT	75K sentences
Kchaou et al. (2025)	TUN-(MSA)-ENG	NMT + augmentation	188K sentences

While progress has been made in MT of Arabic dialects into MSA and ENG, challenges persist due to the limited availability of parallel corpora, essential training fuel for MT models. This resource scarcity is not exclusive to Arabic dialects, as similar hurdles exist for translating other standard languages and their dialects. Fortunately, researchers are devising innovative solutions, leveraging the inherent linguistic proximity between well-resourced languages and their less-resourced counterparts. This brings us to the specific case of Algerian dialects, where the following section delves deeper into the processing techniques employed and the effectiveness of MT models in this unique linguistic landscape.

3. Processing the Algerian Dialects

As a part of the Maghrebi dialects, the Algerian colloquial versions of the language are spoken by more than thirty million people in North Africa. The Maghrebi dialects are commonly described as different from the eastern varieties like GLF, LEV and EGY. Particularly, the Algerian dialects are more complex due to the sociolinguistic situation of Algeria, which has been described widely by linguistics and sociolinguistics experts (Morsly, 1986; Benrabah, 2014; Negadi, 2015; Moulay-Lahssan & Kerras, 2016).

In order to gradually raise the scores of processing the Algerian dialects, some research works have started NLP experiments for the Algerian dialect(s) on a small number of existing corpora and encouraged that additional data is constantly needed. In this respect, language resources for- and the MT of- the Algerian dialects are reviewed.

3.1. Algerian Dialects Corpora

Some survey studies have described the Arabic dialectal corpora used for NLP tasks such as Harrat et al. (2019a) and Babaali and Salem (2023). Younes et al. (2020) contextualized their review of speech and textual corpora representing the Maghrebi dialects (TUN, MOR, LBY and ALG). Within this section, the extant linguistic resources pertaining to Algerian dialects are elucidated. Prior to an in-depth exploration of textual corpora, a concise overview of noteworthy speech datasets is provided. Table 5 catalogs significant ALG speech corpora alongside their respective descriptions, offering a succinct reference point. Each dataset is introduced and reviewed afterwards.

Table 5

Notable Algerian Dialects' Speech Datasets

Research works	Corpus	Size	Description
Hamdani et al. (2010)	ALGSAD	100h	Recordings of 1080 sentences by 300 Algerian speakers from 11 different cities.
Djellab et al. (2017)	AMCASC	88h	Telephone recordings from five Algerian cities.
Bougrine et al. (2017)	KALAM'DZ	104h	Multimedia from YouTube, Online Radio and TV that demonstrate eight Algerian dialectal variations.
Amazouz et al. (2018)	FACST	7h30	Recordings of twenty bilingual speakers of Algerian dialects and French.

Hamdani et al. (2010) built the Algerian Arabic Speech Database, ALGASD. It is the first rich speech dataset capturing Algerian accents. Diverse recordings from 300 speakers across eleven Algerian provinces were collected. An ASR system was trained on ALGASD and achieved impressive accuracy (92%). That demonstrates its value for building Algerian-accented MSA recognition systems. The authors further highlighted potential applications beyond ASR, making

ALGASD a notable resource for studying and understanding Algerian dialects' variations through acoustic phonetic analysis.

Djellab et al. (2017) presented the Algerian Modern Colloquial Arabic Speech Corpus (AMCASC), a speech dataset collected to assist with audio forensic analysis and dialect recognition. After providing a background on language contact phenomena, they processed the data using current standard acoustic approaches. The AMCASC contains over 88 hours of telephonic and microphone recordings from nearly 700 male and over 40 female Algerian Arabic speakers from different regions. Speakers include both monolinguals and bilinguals (code-switching between Arabic and French). Code-switching, intra-group variability, and recording quality differences made the automatic dialect/accent recognition challenging.

Moreover, Bougrine et al. (2017) presented a methodology for building a web-based speech corpus for under-resourced languages. They constructed KALAM'DZ, a 104-hour speech corpus covering eight Algerian dialectal varieties with 4881 speakers. The sources were Algerian TVs, local radios, YouTube channels to ensure diversity of topics, speakers and recording conditions. The corpus included both automatic and manual annotations, providing valuable information about speakers, sub-dialects, and utterance content. This facilitates deeper analysis and exploration of the data.

Amazouz et al. (2018) created the French Algerian Code-Switching Triggered (FACST) audio corpus. It was designed to support linguistic studies in phonetics and prosody. FACST contained recordings from twenty Algerian bilingual speakers with both read speech and elicited spontaneous speech containing triggered spontaneous code-switching. Manual segmentation was done based on language change and speaker turns. Transcription was in orthographic French and

a transliteration for Algerian Arabic. The total duration of the speech set is 7.5 hours; French segments outnumber Arabic largely for most speakers, indicating it as the dominant language in the corpus.

Recently, Benali et al. (2023) constructed the Algerian Dialect Corpus (ALGDC) from spoken and written data using manual and automatic collection methods. The duration of speech collected is 117 hours (29,6K sentences). The rule-based method for sentence boundary detection in ALGDC was proposed based on rules extracted from Algerian and Tunisian dialects. Indeed, the segmentation achieved good results. However, the lack of fixed grammar rules represented a limitation in the processing of ALGDC. The latter remains a valuable resource for NLP experts to address the challenges of processing the Algerian dialects.

The speech datasets provide valuable audio samples that capture the phonetic qualities of Algerian dialect(s). However, textual corpora are also important for studying these dialects, as they allow computational analysis of vocabulary, grammar, and other linguistic patterns. As the current study deals with textual MT, detailing the textual resources that cover the Algerian dialects is worth reviewing. Table 6 outlines the existing textual corpora of Algerian dialects that have been used for NLP tasks. It is important to note that, in user-generated content, Arabic appears not only in different dialects but also written using Roman letters instead of the Arabic alphabet. This script is referred to as '*Arabizi*' (van der Wees et al., 2016, p. 16).

Cotterell et al. (2014) introduced a valuable resource for the study of ALG-FRE code-switching. The corpus was extracted from online newspaper comments and included a variety of code-switching patterns, reflecting real-world language use. Data in Arabic was annotated on Mechanical Turk (Cotterell & Callison-Burch, 2014). From 339K collected comments, 1K

comments were annotated and romanized (Algerian dialect re-scripted in Roman letters). The authors provided initial analyses of the corpus and addressed the growing interest in computational approaches to code-switching. However, the corpus was limited to online comments, which might not be representative of all speech situations. Besides, the size of the dataset is relatively small compared to some other code-switching corpora.

Table 6

Notable Textual Corpora of Algerian Dialects

Research works	**(Corpus) Size**	**Source**	**Script**	**Application(s)**
Cotterell et al. (2014)	1K comments	Newspaper comments	*Arabizi*, *FR.*	Code-switching.
Meftouh et al. (2015)	(PADIC) 6.4K parallel sentences	Recordings	*Arabizi*, *MSA.*	MT.
Abidi et al. (2017)	(CALYOU) 853K comments	YouTube comments	*Arabizi*, *MSA*, *FR.*	Comparable corpus building.
Abainia (2020)	(DZDC12) 2.4K sentences	Social media	*Arabizi*, *FR.*	Code-switching, Transliteration.
Moudjari et al. (2020)	(TWIFIL) 9K tweets	Twitter	*Arabizi*, *MSA*, *FR.*	SA.
Lichouri and Abbas (2021)	(PADIC 2) 6.4K	PADIC	*Arabizi*, *MSA*, *Arabized Kabyle*	MT.

Note. This table presents only the studies that introduced their own corpora. The studies that introduced lexicon data or dictionaries are not included. Research that depended on datasets from external sources -or did not build their corpora from scratch- are not counted either.

Abidi et al. (2017) presented CALYOU, a comparable corpus of spoken Algerian dialect extracted from YouTube comments. The authors emphasized the challenges of dealing with the diverse linguistic landscape of Algeria, where local dialects, MSA and French coexist and often intermingle in online communication. The comments were extracted from 22K videos, the dataset included 853K comments (12.7M words) after cleaning. The authors employed a word-to-vector approach to align comparable sentences, achieving significantly better results than simpler dictionary-based methods.

Besides, the harvesting process from YouTube allowed Abidi et al. (2017) for potential extension of the corpus. Indeed, Abidi and Smaïli (2017) extended CALYOU to reach 1.1M comments (17.7M words). CALYOU can be used for various NLP tasks such as adapting MSA language models to Algerian dialect, extracting parallel phrases for MT, and studying code-switching phenomena. Yet, it is important to acknowledge the limitations in the possible biases inherent in its origin. CALYOU is based on YouTube comments, which may not be representative of spoken Algerian in all contexts.

Moreover, Abainia (2020) introduced DZDC12, a valuable resource for studying Algerian Arabic and its interaction with French in online communication. The corpus outlined twelve different Algerian sub-dialects and it contained 2400 comments (over 47.3K words including 36.2K in Arabic, over 8K in French and 76 in English). DZDC12 focused on Algerian Arabizi-French code-switching, a complex phenomenon in Algerian online discourse. Furthermore, it included annotations for gender, region, city, transliteration into Arabic script, translation into MSA, emotion detection, and word-level segmentation. Thus, it is suitable for various NLP tasks, including SA, MT, code-switching analysis and dialect identification.

DZDC12 is a valuable contribution to the field of NLP, particularly for research on Algerian Arabic and code-switching phenomena. Its rich annotations and diverse applications make it a valuable resource for researchers and developers working on various NLP tasks. However, while sizable for its specific domain, DZDC12 may not be as extensive as some general-purpose corpora, potentially limiting its statistical power for some applications. As it contained online data, DZDC12 may not fully represent all aspects of Algerian Arabic spoken in other contexts.

Abainia et al. (2022) introduced a new corpus for Offensive Tamazight (OTAM). Collected from 6,250 Facebook comments, KAB lexicons were labeled as offensive or normal. OTAM included 12,680 offensive/abusive words and phrases in Tamazight. Through machine learning and DL models, Abainia et al. (2022) experimented offensive language detection on OTAM. The latter, however, is a moderate sized labeled corpus and offensive word lexicon. There is still room for improvement with more advanced models.

In addition, Moudjari et al. (2020) addressed the lack of resources for opinion and emotion analysis of North African Arabic dialects, focusing specifically on Algerian dialect. They used DL models to perform sentiment classifications of Algerian speakers' tweets. Moudjari et al. (2020) introduced a collaborative annotation platform called Twitter Profiling (TWIFIL). The latter generated a dataset of 9K ALG tweets annotated for sentiment, emotion, and extra-linguistic features. However, some details of the dataset, such as the weight of tweets and number of words, were not mentioned. Besides, it may not have elaborated on some specific examples or use cases of how the dataset and platform could be used.

The most notable dataset in the literature is the Parallel Arabic Dialect Corpus (PADIC). Meftouh et al. (2015) introduced PADIC, a parallel corpus of five Arabic dialects: two Algerian variants (ALG and ANB), TUN (the city of Sfax), SYR (the dialect of Damascus), and PAL (the dialect of Gaza). Each dialect includes 6400 sentences (273K tokens) aligned with MSA equivalents. The corpus was built from scratch through a combination of recordings of Algerian dialects which were transcribed and translated, as well as translating an MSA corpus into the other dialects. They conducted MT experiments on all language pairs, exploring the challenges and potential of translating between Arabic dialects and MSA. The best results were achieved between the ALG-ANB and PAL-SYR pairs, while the worst results were between the SYR-ALG pairs.

The seed for such a corpus was made by Meftouh et al. (2012). They constructed the first Algerian dialect-MSA parallel corpus by building a dictionary of words in ANB translated into MSA. Afterwards, the researchers attempted to signify the importance of such resources in bridging the gap between the colloquial Algerian Arabic dialects and MSA. Eventually, Harrat et al. (2014) outlined prominent steps through the creation and enrichment of ALG-ANB parallel corpus. Along with a morphological analysis model, Harrat et al. (2014) proposed diacritization and MT systems for the compiled dataset.

In terms of novelty, Meftouh et al. (2015) extended the dataset to PADIC as the first large-scale, multi-dialect Arabic corpus at the time, addressing the lack of resources for dialectal MT. As for the methodological rigor, Meftouh et al. (2015) clearly described the corpus creation process, experimental setup, and results. In addition, they employed various MT techniques and analyzed the impact of language model smoothing on MT quality. The MT evaluation provided valuable insights into the relative performance of different approaches for dialectal MT.

Though substantial for its time, PADIC is relatively small compared to modern multilingual corpora, potentially limiting the generalizability of the findings. Limited corpus data implies limited scope of linguistic patterns. The impact of language model smoothing techniques was analyzed but found to not make a significant difference on the small corpus. Interpolating with larger MSA corpora also did not improve performance. However, PADIC remains a valuable resource for research and development of Algerian dialect(s) processing.

After three years, Meftouh et al. (2018) presented an extended version of PADIC with the addition of MOR part (the dialect of Rabat). The latter was represented in an amount of 42587 words. Thus, PADIC contained six Arabic dialects, four of which are MGB. MT experiments were performed between all dialect pairs in PADIC. Again, the highest BLEU scores were achieved between the close dialects (ALG-ANB and PAL-SYR). In addition, inter-Maghreb dialects scored higher than ALG-LEV pairs.

Furthermore, a zero-resourced dialect, Kabyle (KAB), scripted in Arabic, has been recently added to PADIC (Lichouri & Abbas, 2021). The latest version of PADIC 2 addressed the key challenges of under-resourced dialects MT. Lichouri and Abbas (2021) incorporated a zero-resourced Berber dialect (KAB) alongside multiple Arabic dialects. Therefore, the corpus became a seven-dialect dataset in addition to MSA. Thorough RBMT and SMT experiments were conducted with and without transliteration, providing valuable insights into the effectiveness of different processing strategies. This resource opens doors for further NLP research on KAB.

The average BLEU score for RBMT was 42.68% and for SMT was 61.94%. Given its distance from Arabic dialects, the scores for KAB were lower. A major weakness in the study of Lichouri and Abbas (2021) is that the factors contributing to translation errors which could guide

future research efforts were not identified. Nonetheless, additional qualitative analysis or human evaluation could strengthen the findings. Although the authors mentioned existing resources for Arabic dialects, a more detailed comparison with PADIC 2 in terms of size, diversity, and domain coverage could further highlight its contribution.

It is noteworthy that ALG was a part in the Multi Arabic Dialect Applications and Resources (MADAR) project. Bouamor et al. (2018) collected basic traveling expressions from twenty-five Arabic cities, 2000 sentences for each. MADAR was in MSA, English and French scripts and aimed at improving the automatic identification and translation of Arabic dialects. Some datasets that included ALG have been constructed for the automatic identification of dialects. The most notable ones are proposed by Guellil and Azouaou (2016), Adouane et al. (2016), Adouane and Dobnik (2017).

In addition, there are some recent studies that proposed their datasets including ALG for different NLP applications. For instance, Dahou and Cheragui (2023a; 2023b) constructed a corpus of 21K ALG sentences from Facebook and YouTube comments for NER tasks. Indeed, the sources are mostly social media comments or common webpages where ALG may occur like newspapers' sites. Other significant datasets were trained for SA experiments as in Guellil et al. (2017), Soumeur et al. (2018), Rahab et al. (2019), Guellil (2021), Guellil et al. (2021), and Hamadouche et al. (2023). Besides, some dictionaries and lexicon-based corpora have been created for the automatic transliteration of Algerian dialect. The major research works are proposed by Abidi and Smaïli (2018) and Guellil et al. (2018).

Having extensively surveyed the processing of the Algerian dialect(s), presenting the major speech and text datasets available, we now pivot our focus towards the application of any relevant

corpora in MT experiments. While the preceding sections shed light on the availability and characteristics of these foundational datasets, the forthcoming section delves into the endeavors aimed at harnessing these resources to enhance the automated translation capabilities for Algerian Arabic.

3.2. Machine Translation of Algerian Dialects

By bridging the gap between data acquisition and practical implementation, it is important to assess the current state of MT systems tailored to Algerian dialects and identify avenues for further advancement in this crucial domain. To the best of the researcher's knowledge, the sole research endeavor dedicated to MT of Algerian dialects is exemplified by the Torjman project[1]. Within the scope of this initiative, both corpus construction and MT experiments were directed towards Algerian Arabic and MSA. This national research project, inaugurated in 2011 under the auspices of the General Directorate for Scientific Research and Technological Development (GDSRTD), was spearheaded by the Scientific and Technical Research Center for the Development of Arabic (STRCDA) and financially supported by the Algerian Ministry of Higher Education and Scientific Research.

The Torjman team proffered innovative solutions through this comprehensive undertaking, with PADIC emerging as one of their refined products. Their vision encompassed facilitating linguistic fluidity for Algerian speakers, enabling seamless transition between MSA and the array of dialects in common parlance. The first research published under this project is by Meftouh et al. (2012). The ANB-MSA parallel corpus motivated the researcher, Salima Harrat, to extend the

[1] A PNR coded **24/u23/1902**. Details are published by Mourad Abbas at https://sites.google.com/site/mouradabbas9/national-research-projects-pnr Access date: 12 May 2025

corpus size and add more dialects for the enhancement of an SMT system. Harrat et al. (2015) presented a full description of PADIC construction and experimentation on an SMT system. In her doctoral dissertation, Harrat (2018) provided thorough details of MT experiments on PADIC. The BLEU scores of ALG-MSA. ANB-MSA and ALG-ANB were 13.55, 12.54, and 61.06, respectively (Harrat, 2018, p. 91).

PADIC represents the sole existing parallel corpus that contains two variants of Algerian Arabic: ALG and ANB. Lichouri et al. (2018) used PADIC for the enhancement of Arabic dialect identification at the word and sentence levels. Different classifiers were used by the authors. For Arabic dialects, the average accuracy was 92%, with the best performance for MGB and MSA pairs. For Algerian dialects, an additional dataset was created covering eight dialects. One-hundred ANB-ALG sentences from PADIC were translated into the dialects of Tenes, Constantine, Djelfa, Ain Defla, Tizi-Ouzou, and Batna. Eventually, the word-level approach worked better, achieving 76% accuracy. Such an approach shows promise for Algerian dialect identification.

Furthermore, Harrat et al. (2019b) attempted to enhance morphological segmentation of Arabic dialects spoken in the Maghreb region. The authors used both CALYOU and PADIC corpora and adapted the word segmentation tool, Morfessor, to handle words scripted in both Arabic and Latin. Morfessor could identify valid morphological segments with 70 to 79% accuracy. A new parallel corpus containing ALG, TUN, and MOR sentences aligned with FRE translations was created through augmenting an existing multi-dialect corpus with FRE. The size of ALG data is 40.75K words in the parallel corpora, and 412.93K words in monolingual corpora for unsupervised segmentation.

Indeed, morphological segmentation contributed to the improvement of dialects pre-processing for MT performance from those dialects into FRE, and to the reduction of OOVs rates by 35 to 50%. Harrat et al. (2019b) utilized manual annotation for semi-supervised segmentation to enhance segmenting nearly 80% of the data scripted in Arabic. The key contribution of developing morphological segmentation for under-resourced languages is to demonstrate its benefits for MT tasks. Challenges remain in fully handling orthographic variation and certain morphological patterns like negation.

Besides, scripting the Algerian dialect in Latin letters and numbers (Arabizi) was tackled by Guellil et al. (2020). Arabizi is commonly used by Algerians on social media; thus, the authors attempted to process it automatically using SMT and NMT approaches. 6,233 sentences were taken from PADIC dataset to be automatically transliterated into Arabizi using a rule-based algorithm. Although neural transliteration performed better than statistical, the final reviewed dataset consisted of 1300 sentences which limited the MT performance. Further, the authors did not compare previous benchmark results with their results. More analysis could be provided into error patterns and differences between the employed models. Based on PADIC (Meftouh et al., 2018) and other datasets, Guellil (2021) experimented thoroughly SA of ALG Arabizi-MSA through automatic transliteration.

Lichouri et al. (2023) evaluated statistical and neural methods for Arabic dialect identification using the parallel dataset, PADIC, and a monolingual dataset, AraDial. The authors used various classifiers, training on PADIC and testing on AraDial, and vice versa. They compared the impact of parallel and non-parallel corpora on the automatic identification of dialects. Results showed training on the non-parallel AraDial corpus gave better accuracy (92.08%) compared to

training on the parallel PADIC corpus (88.24%). Using a non-parallel corpus for training seems to improve results for some dialects like ALG.

Slim et al. (2021) employed an NMT approach to translate Algerian dialect comments on social media related to COVID-19 into MSA. The dataset for this study was a combination of sentences from PADIC with new COVID-19 Facebook comments. The authors first filtered the comments to identify those mentioning COVID-19 using an LSTM model, and then translated them into MSA. For the filtering, they achieved an accuracy of 91.96% in identification. For the translation task, the system achieved a BLEU score of 22.10. This shows the translation quality is reasonable, but still has room for improvement. The results demonstrate the feasibility of using NMT techniques to translate social media text into MSA.

Similarly, Slim et al. (2022) tackled the NMT for low-resource Algerian Arabic dialects using a Transductive Transfer Learning (TTL) strategy. The authors worked on both MADAR and PADIC corpora. Such an approach achieved significant improvements in BLEU scores from 16.5 to 35.87 for the attention Seq2Seq model. This demonstrates the effectiveness of the proposed approach. The NMT experiments by Slim et al. (2022) compared models with and without attention mechanisms before and after transfer learning, providing insights into the impact of transfer learning. However, error analysis and more examples should have been added to better understand where improvements occur.

As mentioned earlier in the review, Moukafih et al. (2022) adopted NMT models to process Arabic dialects. They addressed the challenges of Arabic dialects MT through proposing an MTL approach using PADIC to leverage the similarities between dialects. Indeed, employing MTL to capitalize on dialect similarities is a promising strategy. Besides, such a model outperformed SMT

and single-task learning in most cases. The BLEU scores achieved 35.06 for ALG-MSA and 30.81 for ANB-MSA under the MTL strategy. However, ANB-ALG and ALG-ANB performed better in SMT models.

Recently, Slim and Melouah (2024) proposed an incremental transfer learning approach to improve an NMT system for different low-resource Arabic dialects by transferring linguistic knowledge across dialects. Three Maghrebi dialects were targeted: Algerian, Tunisian and Moroccan using transformer and attentional Seq2Seq models. Data were extracted from both MADAR and PADIC corpora. The achieved BLEU score improvements for the Algerian dialect were 80% in the former model and 98% in the latter. One strength of their work lies in leveraging the relationship between Arabic dialects and shared linguistic features. However, the limited evaluation of Maghrebi dialects should be accompanied by further assessment of other dialects.

The exploration of MT for the Algerian dialect has predominantly relied on SMT methodologies, with some recent NMT experiments. PADIC emerged as the primary dataset utilized across the reviewed studies. The comprehensive review of existing literature reveals a consistent trend in research methodology, predominantly cross-dialectal MT approaches. Notably, among the reviewed experiments, PADIC stands out as the most prominent dataset employed. The existing MT experiments on Algerian dialects are summarized in Table 7, showing the trained dataset, employed approaches and the highest achieved BLEU scores for the MT quality.

Table 7

MT Experiments on Algerian Dialects

Research Works	Datasets	Approaches	BLEU Scores
Meftouh et al. (2015)	PADIC	PBSMT	67:31 (ANB-ALG) 61:06 (ALG-ANB)
Harrat (2018)	PADIC	PBSMT	13.55 (ALG-MSA) 12.54 (ANB-MSA)
Meftouh et al. (2018)	PADIC (extended)	PBSMT	15.10 (ALG-MSA) 14.45 (ANB-MSA)
Harrat et al. (2019b)	PADIC + CALYOU	SMT / Segmentation	7.31 (ALG- FRE)
Lichouri and Abbas (2021)	PADIC 2	RBMT + SMT / transliteration	71.71 (ALG-MSA) 70.76 (ANB-MSA) 25.89 (KAB-MSA)
Slim et al. (2021)	PADIC + 1.2K sentences	NMT	22.10
Slim et al. (2022)	PADIC + MADAR	NMT / TTL	35.87
Moukafih et al. (2022)	PADIC	NMT / MTL	27.71 (MSA-ALG) 27.27 (MSA-ANB) 35.06 (ALG-MSA) 30.81 (ANB-MSA)
Slim and Melouah (2024)	PADIC + MADAR	NMT / Transformer	10 to 11 (ALG-MSA)

4. Summary and Discussion

The present chapter reviewed the relevant literature in the field of MT of Arabic dialects. The key points of this review are summarized as follows:

- All the cited works focus on the processing of Arabic dialects. Most of them focus on MT of these dialects from and into English. Using dialects as the SL was more frequent than using them as the TL. In a considerable number of works, MSA was a pivot language in the MT experiments.

- Despite the number of studies that applied NLP tasks on Arabic dialects, covering dialects is still seen as limited. This, in turn, indicates that MT of Arabic dialects is still in its early stages.

- Among the processed dialects, Egyptian Arabic is the most studied variety in terms of MT. Maghrebi dialects are less represented, and certain dialects like Mauritanian, Kuwaiti, Omani, and Bahraini have not been studied in this context.

- Regarding methodology, hybrid approaches that combine SMT with RBMT are the most commonly employed. Handling dialect OOVs, moreover, was treated by MSA-pivoting or transfer rules strategies.

- Using MSA as a pivot language in Arabic dialects MT into English was critical. Most of the contributions agreed that this technique improves the MT quality. However, some contributions emphasized that this comes at the expense of a lower OOVs rate for MSA-pivoting MT systems.

- In terms of language resources and parallel corpora availability, there is still a significant lack of dialectal data. All the cited works faced this issue, and a substantial portion of the effort is dedicated to creating resources from scratch. Much efforts were directed to generating or augmenting dialectal corpora, preprocessing web data, or, in highly-funded projects, using crowdsourcing.

Despite their widespread use in the Arab world, Arabic dialects have a limited number of dedicated works for MT into and from these dialects. Yet, the results of these efforts are still in the early stages. Even though they are variants of the same language, these dialects require more

improvements in MT systems. Moreover, they are much more commonly used in online communication and the daily lives of Arabs than MSA.

In addition to the observed scarcity of dedicated works for MT of Arabic dialects into English, the availability of parallel datasets specifically tailored to Algerian dialects is notably sparse. Recognizing this gap, the present study endeavors to address this deficiency by pioneering the construction of a parallel corpus representative of one of the significant Algerian dialects. This initiative not only contributes to the advancement of processing Algerian Arabic but also holds the potential to catalyze further research and development in the domain of MT.

As for the Algerian context, Torjman focused on the cross-dialectal dataset construction. However, the project did not include foreign languages -or at least the lingua franca- in their works. In the context of SMT, the existing corpora of Algerian dialects lack bilingual texts and are still of smaller sizes. Thus, the efficiency of building an MT system requires more reasonable sizes. In this respect, there is an urge to build large-scale from-scratch datasets including English within parallel corpora that represent the Algerian dialect.

CHAPTER 3: DATASET AND EXPERIMENT SETUP

Having established the theoretical foundations of MT in Chapter 1 and critically reviewed the existing literature on Arabic dialect processing in Chapter 2, the focus then shifts from a theoretical to a practical and constructive basis. The preceding chapters substantiated the critical resource gap for the Oranian dialect, thereby creating the imperative for the methodological work detailed herein. This chapter provides a transparent and replicable blueprint of the entire experimental process, outlining the steps taken to address the research gap, from foundational data creation to the final design of the computational model.

Accordingly, this chapter begins by describing the creation of a foundational linguistic resource for the Oranian dialect (ORN), a critical first step in a zero-resource context. The specific methodology for compiling the Oranian-English Parallel Corpus (OEPC) is delineated, followed by an analytical overview of the final corpus's characteristics. The chapter then transitions to the NMT system, specifying its complete architecture, core components, and parameter configurations. Finally, the chapter concludes by defining the suite of evaluation metrics that will be employed in the subsequent chapter to quantitatively and qualitatively assess the model's performance.

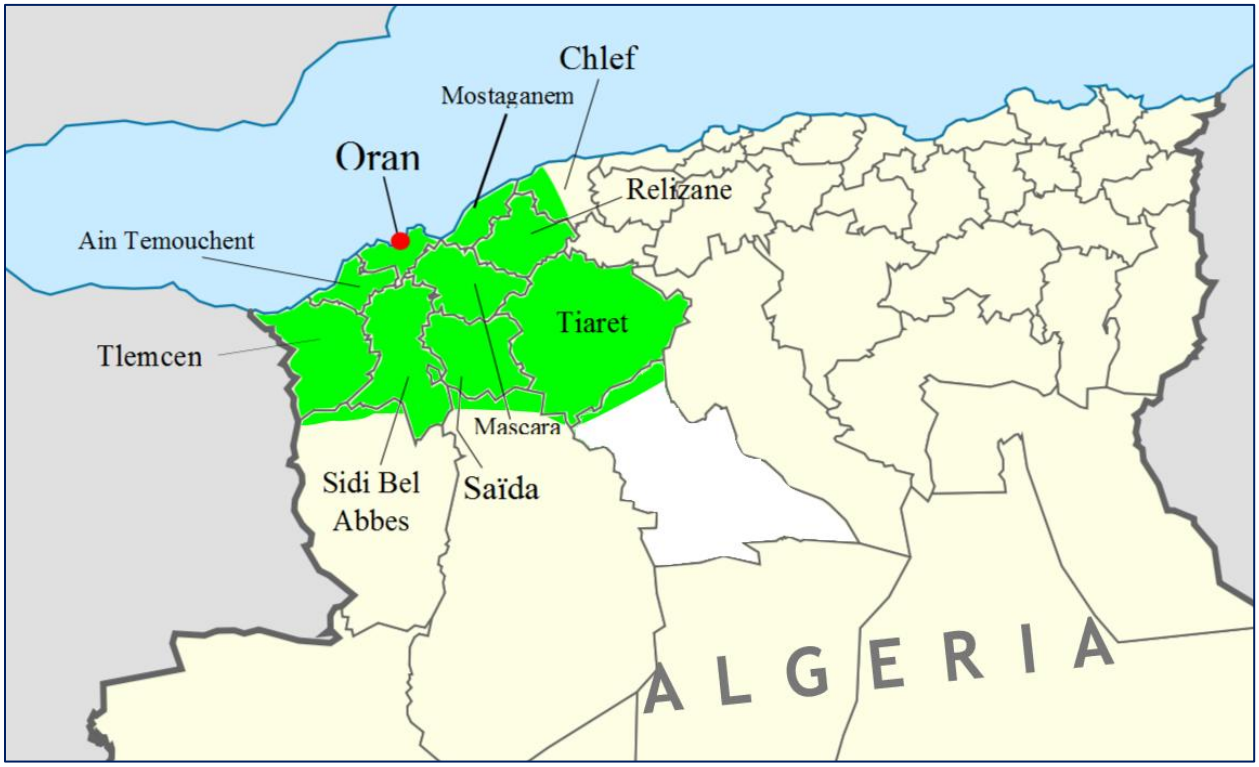
1. The Target Algerian Dialect

The spoken Algerian dialects -or Darija- represent rich and complex varieties that are used in daily life communication, social media, TV shows, telephone conversations, etc. The Oranian

dialect is spoken not only in the city of Oran, but in western Algeria, geographically from Ain-Temouchent to the city of Chlef (Mezzoudj et al., 2019, p.5). Oran is an Algerian city situated in the northwest (see Figure 8). As the second metropolis in Algeria after Algiers, it is known for its large population, maritime coastal attractivity, and industrial and commercial interests.

Figure 8

Map of Oran City



Note. The green area demonstrates the northwest Algerian cities and the geo-linguistic ecology of the Oranian dialect.

The verbal phonology and morphology of Oran's dialect have been thoroughly examined by Algerian researchers in dialectology and sociolinguistic variation such as Bouhadiba (1988) and Labed (2014; 2016). However, studying dialects from a computational linguistic perspective is different from presenting them through a sociolinguistic lens. That is, in the case of text-based

AI applications, it is not necessary to describe the phonological aspects of the target variety. In the present study, a text-based, that is one-modal, MT is carried out; therefore, in this section, the major morpho-syntactic features of the Oranian dialect are described. In addition, lexical and morphological particularities should be noted to make it easier for language resources' readers to understand and distinguish this dialect.

1.1. Description of the Oranian Dialect

Several historical factors demonstrate how Algeria is affected by Arabic, Latin and Turkish languages (Chami, 2009). For that reason, the Algerian dialects are different from MSA at the lexical, phonological, morphological and syntactic levels. Such complexity has been addressed by computational linguists (Souag, 2006; Saadane & Habash, 2015). Sociolinguists and linguists have classified Algerian dialects differently depending on geographical areas and ethnic factors (e.g., Queffélec et al., 2002; Meftouh et al., 2012; Harrat et al., 2014; Alalou, 2023). In this respect, and for the purposes of introducing the target Algerian dialect for this study, it is necessary to describe the major phonological and morphosyntactic features of the Oranian dialect.

1.1.1. Major Phonological Features

The Oranians' pronunciation is primarily influenced by Bedouin Arabic due to the migration factor. This is evident in specific phonological characteristics such as the velarization of **[q]**. In Oranian, the MSA letter *qāf* [q] is typically realized as a voiced velar [g] in most cases. For example: قال /qāl/ ('He said') becomes ڨال /gāl/. Moreover, non-Arabic phonemes are also adapted in this dialect. The labiodental [v] is used in words like فوطي /vōtī/ ('vote'), which comes from

French '*voter'*. In addition, the MSA interdental fricatives are typically realized as stops in the Oranian dialect as:

/θ/ (ث) becomes [t], for example, /ثلاثة/ (θalāθah) becomes /تلاتة/ (tlātah) ('three').

/ð/ (ذ) is pronounced as [d], so /ذَهَب/ (ðahab) becomes /دهَب/ (dhab) ('gold').

/ðˤ/ (ظ) shifts to a stronger [dˤ], merging with the sound of **ض**. (Charef, 2016).

While the major phonological features of the Oranian dialect are presented, it is important to emphasize that the primary focus remains on text processing using a textual dataset. Understanding the phonological tendencies, particularly how certain MSA sounds are merged or altered in spoken Oranian, helps clarify potential ambiguities in written representations of the dialect. This is crucial for accurate and effective text processing, as it provides context for variations in spelling. However, the latter is more influenced by morphosyntax.

1.1.2. Morphosyntactic Features

Like some other Algerian dialects, the Oranian dialect distinguishes between pronouns and has no dual (*Al Muthana*) in morphology. As for the feminine form, it does not exist in the plural pronouns. Although some pronouns may be spoken differently by some Oranian speakers, the commonly used forms are described and presented in the table below. First, Table 8 shows the personal pronouns which stand alone as the subject or object of a verb. Second, it shows the object pronouns which attach to verbs as suffixes indicating the object. Since they are similar to object pronouns, possessive pronouns -marked by suffixes attached to nouns- are presented together. However, the only difference is in the first singular pronoun which is -ني (*-ni*)/ as a dependent

pronoun as in ضربني /*drabni*/ (he hit me), and ي- (-i) as a possessive one as in داري /*dari*/ (my house).

Table 8

Pronouns in the Oranian Dialect

Person	Singular	Plural
Personal/Independent Pronouns		
1st	أنا (ana)	حنا (hna)
2nd	نتا (nta) / نتي (nti)	نتوما (ntuma)
3rd	هو (huwa) / هي (hiya)	هوما (huma)
Object/Dependent Pronouns and Possessive Pronouns		
1st	-ني (-ni) / -ي (-i)	-نا (-na)
2nd	-ك (-k) / -كي (-ki)	-كم (-kom)
3rd	-ه (-h) / -ها (-ha)	-هم (-hom)

Guerrero (2015; 2021) thoroughly explained the morphosyntactic features of this dialect, particularly how Oranians use a combination of prefixes and suffixes to mark tense, aspect, person, and gender in verb conjugation. Verb inflection, more commonly known as conjugation, is used to determine tense, voice (passive or active), person, number, and gender. Declension, the inflection of nouns, adjectives, and pronouns, also expresses person, number, and gender. Table 9 illustrates the conjugation of the regular triliteral verb دخل (to enter) in both past and present tenses.

Based on the verb structure, three active participles occur for the verb دْخَلْ /*dxal*/ which are: داخلين / داخلة / داخل /*daxel, daxla, daxlin*/. Moreover, it is obvious in the Oranian dialect that both the present and the imperative show gender differentiation in the masculine and feminine singular as

in: *nta təfṭăṛ*, *nti təffəṭri* (you have lunch) and *(ə)fṭăṛ*, *(əf)fəṭri* (have lunch!). Geminating the first syllable in the second feminine singular person and in the plural forms of the present is a salient feature of this and many other Algerian dialects.

Table 9

Verb Conjugation in the Oranian Dialect

Person	Singular	Plural
Past tense		
1st pr.	دخَلْت (*dxalt*)	دخَلْنا (*dxalna*)
2nd pr.	دخَلْت (*dxalt*) / دخَلْتي (*dxalti*)	دخَلْتو (*dxaltu*)
3rd pr.	دخَلْ (*dxal*) / دَخْلَتْ (*daxlet*)	دَخْلُو (*daxlu*)
Present Tense		
1st pr.	ندخل (*nudxul*)	ندخلو (*ndxlu*)
2nd pr.	تدخل (*tudxul*) / تدخلي (*tudxli*)	تدخلو (*tdxlu*)
3rd pr.	يدخل (*yudxul*) / تدخل (*tudxul*)	يدخلو (*ydxlu*)
Imperative		
2nd pr.	ادخل (*odxul*) / ادخلي (*odxli*)	ادخلو (*odxlu*)

Defective triliteral verbs display similar inflections with minor differences. For instance, the verb أكل (to eat) in MSA becomes كلا in the Oranian dialect and it is conjugated as: *Kilt, klīna, kilt, klīti, klītu, kla, klāt, klāw*, and the active participles of this verb are: كالي / كالية/ كاليين */kāli, kālya, kālĭyīn/*. In addition, the Oranian dialect is distinguished by generally omitting a present tense marker or preverb. However, the particle "*ṛa*–" can be employed to specifically highlight the present moment or to function as a copula (similar to "am/is/are" in English). It precedes a verb to describe a progressive action as in "*ṛāhi tḍūṛ*" (it is circulating), "*ṛāh yəʒri*" (he is running), and

sometimes precedes a noun (mostly active participles) as in: "*ṛāhum sāktin*" (they are silent), "*ṛāk fāhəm!"* (you see!).

As for the future tense, the Oranian speakers use the invariable particle "*ġādi*". Unlike some grammatical structures that change form depending on the subject, "ġādi" remains the same regardless of who's performing the future action as in: "ġādi tṛūḥi" (you will go) "*ġādi nŭḍṛŭbha*" (I will hit her). Sometimes, Oranians use "*bah*" or "*bei*" to refer to near future or intentions as in: "*bah nqra ktab*" (I'm going to read a book), "*bei nakul*" (I'm going to eat).

Negation in the Oranian dialect is expressed with the particles "ما" (*mā*) and "ما …ش" (*mā ... š*). The latter is specific to the Maghrebi dialects. Using these particles, the negative form is obtained by adding the affixes "ما" and "ش" to the conjugated verb; "ما" as a prefix and "ش" as a suffix (see Table 10). When the verb is preceded by the particle "*ṛa*–" which is equivalent to the auxiliary verb "to be" in the present indicative, negation is obtained by adding the affixes "ما" and "ش" to the particle as in: "*ma ṛāhumš sāktin*" (they are not silent), "*ma ṛākš təfhəm*" (you do not understand). In some cases, negating is obtained by the prefix "ma" and followed by negative markers like *walu*, *ghir* or *hatta*.

It is noted that imperatives (ما تدخلش /ما تدخليش/ما تدخلوش) are similar to the imperfectives when negated. For the future tense, the particle "*ġādi*" is negated as "*mā ġādiš"*. The Oranian dialect, like many Arabic dialects, exhibits distinctive features in its negation strategies. Oranians also negate pronouns, nouns, adverbs and adjectives by preceding them with "*meši*" (not/ un-) as in "*meši ana*" (not me), "*meši lyum*" (not today) and "*meši kbira*" (not big).

Table 10

Verb Negation in the Oranian Dialect

Person	Singular	Plural
Past tense		
1st pr.	ما دخَلْتش(*ma dxaltš*)	ما دخَلْناش(*ma dxalnaš*)
2nd pr.	ما دخَلْتش(*ma dxaltš*) / ما دخَلْتيش(*ma dxaltiš*)	دخَلْتوش ما(*ma dxaltuš*)
3rd pr.	ما دخَلْش(*ma dxalš*) / ما دَخْلَتْش(*ma daxletš*)	ما دَخْلُوش(*ma daxluš*)
Present Tense		
1st pr.	ما ندخلش(*ma nudxulš*)	ما ندخلوش(*ma ndxluš*)
2nd pr.	ما تدخلش(*ma todxulš*) / ما تدخليش(*ma todxliš*)	ما تدخلوش(*ma tdxluš*)
3rd pr.	ما يدخلش(*ma yudxulš*) / ما تدخلش(*ma tudxulš*)	ما يدخلوش(*ma ydxluš*)

The interrogative words in Oranian Arabic provide a fascinating glimpse into how this dialect forms questions. The common question word is "شاوالا" which has an alternative for feminine "شايالا". However, the latter form is also used as the conjunction "but". Moreover, the use of "*šah*" is also common for objects and, less frequently, reasons. Also, "*wīnta*" is used to ask about broad time ranges while "*qăysāš*" is used for specific time. As for the particle "*ki–*," it connects with verbs or phrases to express or question the manner (how) as in "*ki lqitni?*" (how did you find me?), and sometimes functions as the conjunction "like" as in "*ki ħālti"* (like my state). The contexts in which these interrogatives are used determine their syntactic behavior in sentences. Table 11 demonstrates the common question words in the Oranian dialect with their English translations.

The provided description focuses on the major structural elements of the Oranian dialect, specifically focusing on morphology (word formation) and syntax (sentence structure). Other

features of the target dialect will be encountered and discussed in the processing and results sections. Clarifying the polysynthetic features of this dialect raises some questions regarding its processibility by the machine.

Table 11

Common Interrogatives in the Oranian Dialect

ORN	Transcription	English
وينتا	*winta*	When
قيساش	*qăysāš*	At what time
وين	*win*	Where
منين / مين	*mnin / min*	From where
شاوالا / شايالا	*šāwāla*	What
شا	*šā*	What
علاش	*ʕlāš*	Why
شكون	*škun*	Who
شحال	*šhal*	How much/many
كي	*ki*	How

1.2. Challenges of Processing ORN

Several research works addressed the Arabic -and its dialects- processing challenges (Althobaiti, 2020; Zakraoui et al., 2021; Almaaytah & Alzobidy, 2023). The machine identifies the language if it has sufficient pre-trained models or linguistic rules about it (Zaidan & Callison-Burch, 2014). There are three major challenges of processing the dialects: orthography, morphosyntactic aspects and the lack of language resources. They are briefly presented in this section with reference to the Oranian dialect.

1.2.1. Scripting and Orthography

Among the major works conducted on Arabic dialects scripting is Maamouri et al. (2004) who developed a set of rules defining the conversational Levantine dialect. They used Arabic transcription guidelines and annotation conventions. In addition, Habash, Diab and Rambow (2012) proposed a Conventional Orthography for Dialectal Arabic (CODA). This work is inspired by the Linguistic Data Consortium (LDC) guidelines for transcribing. However, CODA allows many abstracts from the variations, whereas the LDC guidelines focus more on phonological variations in sub-dialects.

Algerian dialects lack codified written forms, leading to diverse and often inconsistent representations in informal settings, particularly online. While some Algerians write these dialects using Arabic script, many utilize a Latin-based transliteration called Arabizi. This lack of standardization poses challenges for defining "correct" written forms, leaving dialect users to navigate conventions independently (Bies et al., 2014). Nevertheless, researchers and developers often adopt specific rules to represent dialectal features in writing, aiming to balance consistency and authenticity.

The speakers of the Oranian dialect tend to employ different spellings based on pronunciation and personal preferences (Charef, 2016). For instance, they sometimes remove long vowel letters that are essential when they write assuming that they are short as in "منروحش" or they add them assuming they are long as in "عالامة" (instead of "علامة"). In addition, the final feminine-indicating letter "ة" is misused in the dialect scripting as in "ورقه" or "ورقا" (instead of "ورقة"). Other examples will be discussed in this study throughout the process of data collection and the results. Thus, the inconsistencies of scripting the Oranian dialect create significant hurdles for

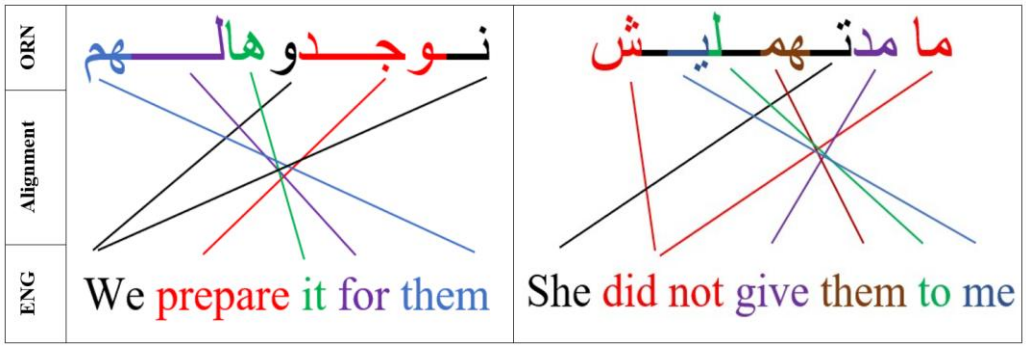
computational linguists; however, practical approaches to setting guidelines for the dialect will reduce such hurdles.

1.2.2. Morphosyntactic Aspects

The inflectional nature of the Arabic and its dialects makes a challenge for the processing of its morphology. For instance, some prepositions are written attached to the noun as in "فالدار" (at home) and "مالدار" (from home). As noted in the previous section, conjugation forms in the Oranian dialect also represent a complex morphosyntax. In this respect, Figure 9 demonstrates two examples of morphologically-rich sentences from the Oranian dialect translated into English (ما مدتهمليش /*mā mdthmlish*/ and نوجدو هملها /*nwjdwhalhum*/)

Figure 9

Examples of the Oranian Dialect's Rich Morphology



When highly complex verbs are placed at the heart of sentence structure, information about the subject, object, tense, aspect, mood, and more are incorporated within their morphology. This verb-centricity often leads to long-distance dependencies, where elements related to the verb might be separated by prepositions, words or phrases in the sentence. Capturing these complex relationships and dependencies is crucial for accurate machine processing. Such challenges have been addressed by many studies that have brought high-score lemmatization (Habash et al., 2013; Zalmout &

Habash, 2019; Hashem, 2021). Interestingly, Habash et al. (2022) have recently proposed a new approach to modeling Arabic morphology with an eye towards multi-dialectness.

Another morphosyntactic challenge for processing the Oranian dialect in the free word-order. Flexible sentence structures, thus, make grammatical relations between words primarily conveyed through morphology rather than their position in the sentence. This free word order, while expressive, makes it challenging for traditional parsing algorithms that rely on fixed patterns and linear relationships between words. To accurately process and analyze the Oranian sentences, sophisticated parsing techniques that can account for long-distance dependencies and the interplay between morphology and syntax are required.

1.2.3. Language Resources

The scarcity of linguistic resources for the Oranian dialect, such as annotated corpora, dictionaries, and morphological analyzers, further exacerbates the challenges posed by its polysynthetic nature. Without sufficient data, it becomes even harder to train accurate and robust NLP models, particularly for tasks like morphological segmentation, PoS tagging, and syntactic parsing. This lack of resources highlights the need for innovative approaches that can leverage data from related languages or employ unsupervised and semi-supervised learning techniques to compensate for the absence of extensive labeled data.

2. Research Design and Procedure

The methodological framework of this research is grounded in a hybrid paradigm that integrates both constructive and empirical approaches. The choice of this approach is a direct response to the foundational challenge of the study: the zero-resource status of ORN. This status

mandates a sequential procedure where the necessary linguistic resources must first be created before any computational modeling can be undertaken. Accordingly, this section details the systematic workflow followed in this research. It begins by describing the compilation and preparation of the Oranian-English Parallel Corpus (OEPC), outlining the entire process from data sourcing to final normalization. Following this, the section delineates the design of the proposed NMT system, providing the rationale for the selected architecture and its specific components.

2.1. Corpora Compilation Project

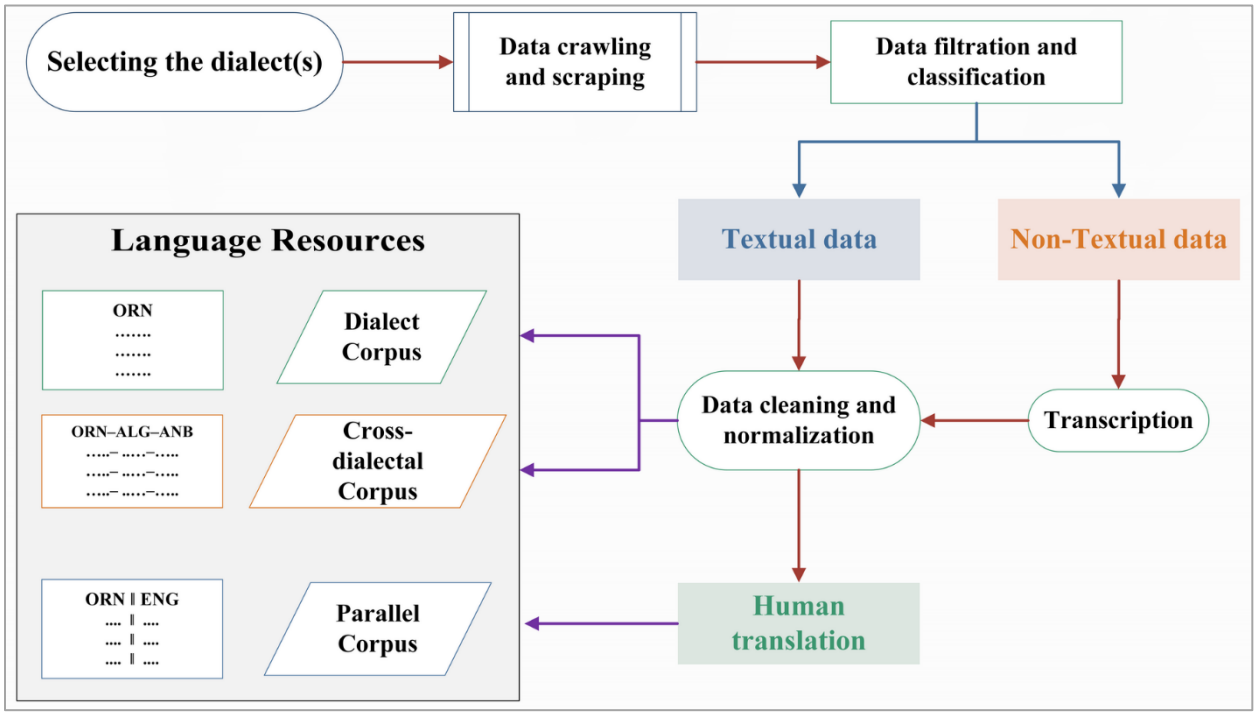
Recognizing the critical absence of datasets for the Oranian dialect and other Algerian varieties, this work is developed under the umbrella of the Empowering and Resourcing Algerian Dialects (ERAD) project[2]. ERAD aims to address this gap by constructing language resources for Algerian dialects from the ground up. The full description and prioritized objectives are provided by Dou and Kissi (2024; 2025). The ERAD project draws inspiration from the work of Harrat et al. (2017) on Arabic dialect corpora creation, taking into account potential pitfalls and implementing procedural safeguards to ensure high-quality resource development. A rigorous three-pronged approach focusing on scripting conventions, normalization processes, and corpus compilation methods will be implemented (see Figure 10).

The outcomes of ERAD are three different types of corpora: mono-dialectal, cross-dialectal and parallel. Preliminary procedures of data collection and filtration is shared. The following sub-sections describe the creation of the first products in this project.

[2] The project website is accessible: https://sites.google.com/view/eradproject/

Figure 10

Procedure of Textual Corpora Compilation in ERAD



2.1.1. Building ORN

The corpus data are taken from naturally-occurring linguistic instances of textual, audio, and audiovisual language practices. The seed corpus created within the framework of this dissertation concerns Oran's dialect labeled as 'ORN'. After assuring the native Oranian dialect occurring in the sources, several portions from movies, sitcom scenarios, daily-life conversations, TV shows, Radio streaming, YouTube videos, podcasts, social media communications and audio recordings are transcribed. Texts are collected from social media posts and comments. The types of ORN featured sources are demonstrated in Table 12.

Table 12

Types of ORN Data Sources

	Text	Audio	Video
Duration	-	≈ 8 hours	≈ 50 hours
Raw sentences	2500	3000	4500
Corpus items	2110	2620	3770
Total	8500 sentences		

Raw sentences are all the transcribed data with opening credits, ads, non-native Oranian speech, full French sentences, lengthy French phrases and other unnatural linguistic instances. The inclusion-exclusion criteria for any linguistic data to be in ORN corpus are primarily and purely orthological. The corpus contents are scripted in Arabic letters with conventional writing structures. Moreover, individual preferences and social media idiosyncrasies are treated accordingly.

The scripting of words that are not originally Arabic brings some challenges. Since Maghrebi dialects and sub-dialects are distinguished for being more affected by multilingual features than other Arabic dialects (Harrat et al., 2019b; Younes et al., 2020), several words are originated from Latin, Turkish and Berber words (see examples in Table 13), leading to codeswitching. Particularly, borrowing plays a significant role in shaping the Oranian dialect's vocabulary.

The Oranian dialect draws loanwords from various sources, reflecting the historical contact and cultural exchange between Arabic and other linguistic communities in the region. The vocabulary-based inclusion criterion in ORN is based on the common use of the words. Non-Arabic words that Oranians use for hundreds of years with less or no contemporary alternatives are included.

Table 13

Non-Arabic Vocabulary in ORN

Language	Words in ORN (origin)	English
Spanish	كوزينة *(Cucina)*	Kitchen
	كوردا *(Cuerda)*	Cord
	بورسة *(Bolsa)*	Plastic bag
	شنقلة *(Chancla)*	Flip flops
French	لوطو *(L'auto)*	Car
	كازيرنا *(Caserne)*	Barracks
	سنتورة *(Ceinture)*	Belt
	صاي *(ça y est)*	That's it
Turkish	تزدام *(tezdām)*	Wallet
	بشْكير *(bèşkîr)*	Towel
	بلّار *(bellār)*	Glass
	دوزان *(dūzān)*	Materials
Berber	زكرم (ẓəkṛūm)	Bolt
	مازوزي (māẓūẓi)	Boy of the family
	الغاشي (l-ġāši)	People

Despite the influence of non-Arabic lexicon, treating the SL items is carefully taken in terms of codification and orthography throughout the compilation of ORN. That is, the researcher prefers to avoid French words and expressions that have alternatives in the dialect. Some French - or Latin- lexis is, however, considered in the corpus for they have less or no alternatives in the Oranian dialect (examples are provided in Table 14). The recognition of all potentially used French expressions in the corpus will make more likely a code-switching instance of the language.

When it comes to spelling, and because of reduced Latin words, ORN items are transcribed based on Arabic letters regardless of how Oranians pronounce them. Moreover, to enhance consistency and minimize out-of-vocabulary (OOV) occurrences within the corpus, a

normalization process is implemented on the raw data. This involves standardizing written forms to establish a unified representation. However, datasets that are built from scratch are commonly normalized during transcription. Indeed, dialect processing systems may be confounded if the raw texts are not normalized (Darwish et al., 2012).

Table 14

(Non-)Inserted Lexis in ORN

	Latin Lexis	**Arabized Latin Lexis**
	catastrophe, *clef*, *preparer*, *manger*, *exactement*, etc. (catastrophe, key, prepare, eat, exactly)	لوطو، بوليسي، سبيطار، كارطاب، ميكرو، كوزينة، إلخ. (car, policeman, hospital, schoolbag, computer, kitchen)
Choice	***Non-Inserted***	***Inserted***
Reasons	Oranians have alternatives to these words. They may appear in Arabizi.	Frequently used, with less or no alternative, and commonly written in Arabizi.

Throughout the creation of ORN, some Arabic letters are normalized; however, special symbols, non-Arabic characters and diacritics are removed. ORN sentences, therefore, consist only of mono-lingual coherent data. The normalization process of ORN treats different writing issues as demonstrated in Table 15.

Table 15

Normalized Written Forms in ORN

Source text issues	**Definition**	**Examples**
Elongations	repeating vowel and consonants forms	rbḥuuu ربحووو / bzāāāāf بزاااف
Spelling mistakes	dropped / transposed letters	rwāḥu رواحو / mrāsāk مراسك
Abbreviations	(borrowed) short forms	CEM سويام / CV سيفي
Special tokens	Emotional / ejaculative expressions	Lol لول / hahaha هاهاها

In order for the corpora users to search for or use accurate spelling, the normalization of how linguistic units should be spelled or written is a necessary process. Moreover, normalizing the instances "نيشاااان" and "نيييشان" into the word "نيشان" (Right) will not impact the dialectic distinctiveness of the target variety.

Although there are no conventional writing guides for the Arabic dialects, several researchers depended on using MSA-conformed normalization features (Dahou & Cheragui, 2023a; 2023b). ERAD project has been initiated with the potential to transcribe the Algerian dialects, normalize and conform their written forms to MSA spelling rules. For the manipulation of dialect orthography use by non-experts, several linguistic items need to be written in the corpus differently and with more than one form according to the common use frequencies of their appearance by the dialect's natives.

2.1.2. Building OEPC

Following the established criteria for corpus construction, items in ORN dataset are transcribed using Arabic script, irrespective of regional pronunciation variations. Given that the ORN dataset was compiled de novo and is intended for inclusion in a parallel corpus, human translation was conducted by native Algerian speakers who reviewed the ORN transcriptions and provided corresponding ENG equivalents. The initial OEPC was subsequently compiled by aligning these sentence pairs. A representative sample is provided in Figure 11.

Figure 11

Examples from OEPC

ORN	ENG
راها معمرة ما خدمناش بيها	it is full we did not use it
ما تنساوش توصلو خالتكم لدارها	do not forget to drive your aunt to her house
راهم عندي مفاتح تبغي قاع نروح معاك	I have keys if you want I go with you
مكاش مشكل عيطلي وقت لي تبغي	it is ok call me whenever you want
نروحو دروك إيلا بغيت	we go now if you want
شغل راها باينتله باللي نتايا راك تبروفيتي	as if it seems to him that you are profitting
هاوالا قدامك و اهدر معاه	here he is and talk to him
بلع عليها كي البونيشا في الدار	he locked her up like a maid at home
الجيلالي بغا بنته تروح وحدها	Djilali wanted his daughter to go alone
زكرمي عليه فالدار خليه يلعب شويا	lock him in the house let him play a bit
خليتها تسيل ما علابالیش ایلا زیرتو علیها	I left it flowing I do not know if you tightened it

The statistical analysis of OEPC reveals insights into sentence structure and lexical diversity across both languages. Sentence lengths in the corpus range from single words to extended constructions. On average, ORN sentences contain 5.2 words, while their corresponding English translations have an average length of 7.3 words. Furthermore, the proportion of OEPC unique items stands at 31.14% for the ORN component and 31.09% for the ENG component, as demonstrated in Table 16. Moreover, this difference in distribution is demonstrated in the density plot in Figure 12. ENG sentence length curve indicates the prevalence of longer ENG sentences.

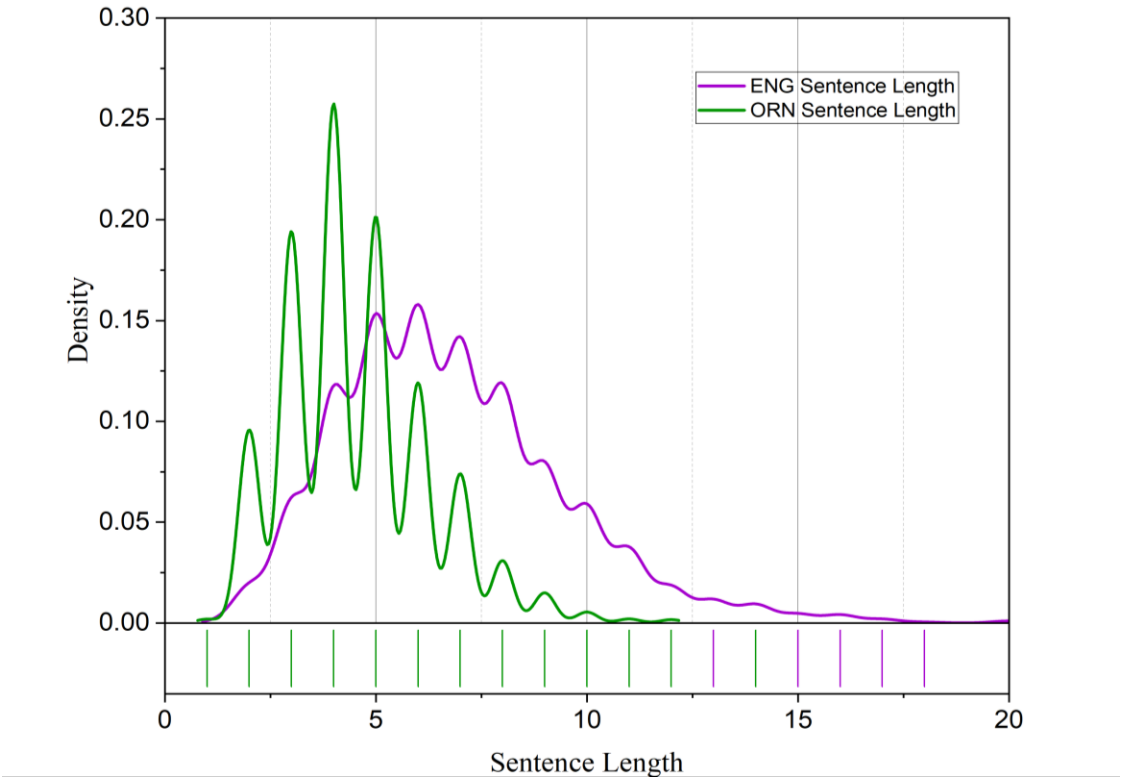
Table 16

Major Statistics of OEPC

Dataset	Sentences (average)	#Words	#Distinct
ORN	8.5K (5.2)	53,255	16,584
ENG	8.5K (7.3)	62,361	19,392

Figure 12

Histogram of Sentences Length in OEPC



Following the compilation and description of OEPC, the next critical step in preparing the dataset for NMT involves preprocessing and tokenization. This stage plays a pivotal role in transforming raw text into a structured format that can be efficiently processed by machine learning models. In order to deal with an LRL such as ORN and mitigate issues like OOVs and morphological variations, the selection of tokenization methods becomes crucially important.

2.1.3. Data Preprocessing

Tokenization is a fundamental step in the data preprocessing pipeline for NMT. It involves segmenting raw text into manageable units, typically words or subword tokens, helps reduce vocabulary size and handle the complexities of language morphology and word formation. Proper

tokenization is crucial for ensuring that the model can process diverse word forms and unseen words during inference (Sennrich et al., 2016), especially in LRLs such as the Oranian dialect. For the latter, where word-level tokenization may fail due to limited training data or high morphological complexity, subword tokenization techniques, such as Byte Pair Encoding (BPE), have become essential.

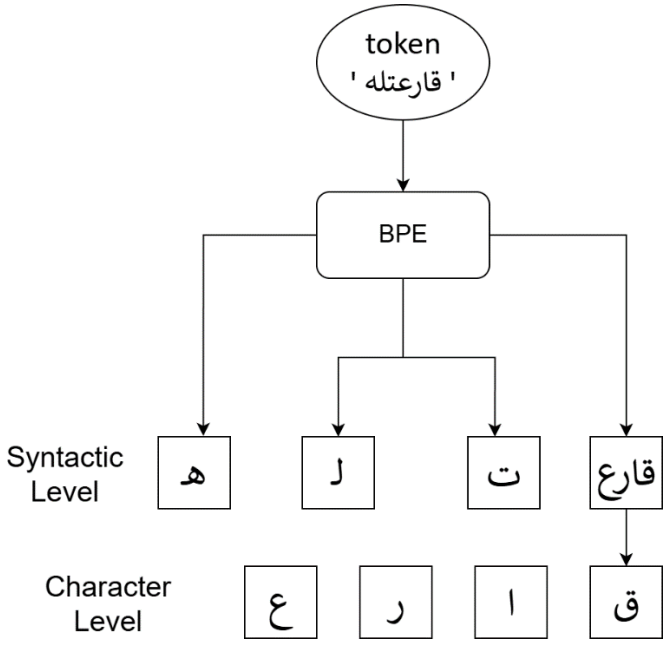
BPE is a data-driven approach that iteratively merges the most frequent pairs of characters or subword units in a corpus to build a vocabulary of subwords (Gage, 1994). This method helps tackle issues of sparsity by breaking down rare or unseen words into smaller, more frequent components, allowing for better generalization during translation. In the context of ORN, where the dialect's orthography is not standardized and word forms vary widely, BPE serves as an effective initial tokenization step to handle diverse spelling variations and morphological phenomena (Sennrich et al., 2016). Given the relatively small size of the ORN dataset, BPE ensures that even words that are rarely seen in the corpus can be represented by subword tokens. The Oranian word example 'قارعتله' (Translated "*I waited for him*") in Figure 13 demonstrates how the subword tokenization is executed.

Following BPE, WordPiece tokenization, which is inherently integrated into the pre-trained DziriBERT model (Abdaoui et al., 2021), is used to further refine the tokenization process. WordPiece, a subword tokenization method introduced in the BERT architecture, optimizes the subword vocabulary based on the likelihood of subword units in a large corpus (Wu et al., 2016). DziriBERT, which is specifically trained on Algerian dialects, uses WordPiece to tokenize input text into a vocabulary learned from this dialect. By leveraging DziriBERT's pretrained embeddings, the model can effectively capture contextual relationships between subwords, even

for dialectal words or phrases not explicitly seen in the training data. This integration of WordPiece tokenization ensures that ORN benefits from both fine-grained segmentation and the rich knowledge embedded in the pre-trained model, improving translation quality and robustness.

Figure 13

Subword Tokenization of Oranian Dialect



Having prepared the tokenized data, the next focus shifts to the integration of this data within the NMT framework. This involves detailing the proposed model components and architecture, emphasizing how they are structured to achieve optimal performance for the ORN-ENG translation task.

2.2. Components of the Proposed Model

The core methodology employed in this study for translating the Oranian dialect into English leverages NMT whose models have demonstrated state-of-the-art performance across

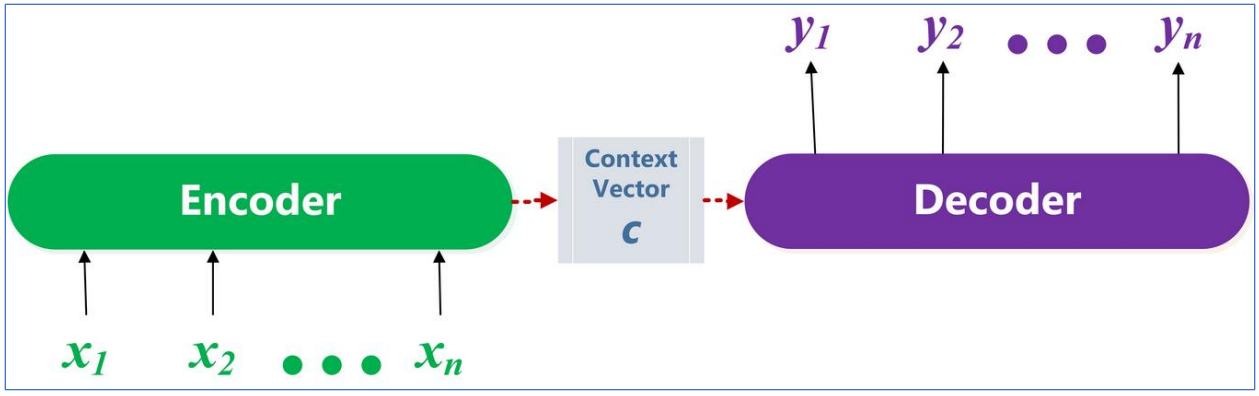
various language pairs by treating translation as an end-to-end Seq2Seq learning problem (Sutskever et al., 2014; Bahdanau et al., 2015). Specifically, the proposed model for the present study is based on the widely successful encoder-decoder framework augmented with an attention mechanism. This section details the architecture and its components, contextualized for OEPC.

2.2.1. The Encoder-Decoder Framework

At its heart, the NMT model consists of two main components: an encoder and a decoder. The basic structure of both RNNs (encoder and decoder) are demonstrated in Figure 14. This architecture is designed to map an input sequence of variable length (ORN) to an output sequence of potentially different variable length (ENG).

Figure 14

The Basic Encoder-Decoder Model



The encoder's role is to process the input sentence, token by token (where tokens are typically words or subword units). It reads the entire source sequence and compresses the information, capturing its semantic meaning and contextual dependencies, into a fixed-size vector representation *c*. This representation is often referred to as the "context vector" or "thought vector" (Sutskever et al., 2014; Cho et al., 2014). For an Oranian input sentence $X = (x_1, x_2, \ldots x_n)$, the

encoder produces a context vector *c*. The decoder takes the generated vector by the encoder as input. Its task is to generate the target English translation, $Y = (y_1, y_2, \ldots y_n)$, one token at a time.

2.2.2. Seq2Seq Model Implementation

In practical NMT implementations, both the encoder and the decoder are typically realized using RNNs (Kalchbrenner & Blunsom, 2013), giving rise to the Seq2Seq model (Sutskever et al., 2014). RNNs, particularly variants like LSTMs (Hochreiter & Schmidhuber, 1997) or Gated Recurrent Units (GRUs) (Cho et al., 2014), are well-suited for processing sequential data like language due to their ability to maintain an internal hidden state that captures information from previous timesteps. The activation function of Seq2Seq model is:

$$P(y \mid x) = P(y_1 \mid x)P(y_2 \mid y_1, x)P(y_3 \mid y_1, y_2, x) \ldots, P(y_T \mid y_1, \ldots, y_{T-1}, x)$$

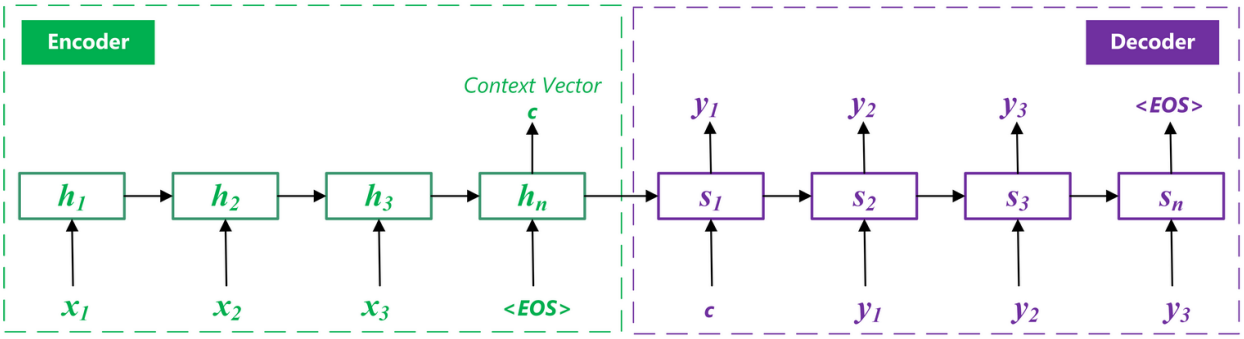
In other words, the probability of generating the entire target sequence $\boldsymbol{y}$ (ENG), given the source sequence *x* (ORN), is calculated as the product of the conditional probabilities of generating each target token y_T one by one, where each token's probability depends on the original source sequence *x* and all the target tokens generated before it (y_1 to y_{T-1}). In this respect, the choice of Seq2Seq model supports this study by handling variable input (ORN) and output (ENG) lengths.

Sutskever et al. (2014) explains the Seq2Seq NMT by stating that one LSTM reads an input sequence *x*, one timestep *t* at a time, to create a vector *c* and pushes another LSTM to extract an output sequence *y* from that vector (p. 2). A tutorial for using this model was elaborated by Neubig (2017). As visualized in Figure 15, the encoder RNN processes the Oranian input sequence *x* sequentially. At each timestep *t*, it updates its hidden state h_t based on the current input token x_i and the previous hidden state h_{t-1}, and the final hidden state h_x (or a transformation thereof) serves as the context vector *c*. The decoder RNN is initialized with the context vector *c* (often as its initial

hidden state s_0). At each timestep t, it takes the previously generated English word y_{t-1} and its previous hidden state s_{t-1} to generate the current hidden state s_t and predict the next English word y_t.

Figure 15

The Basic Seq2Seq Encoder-Decoder Model



The review conducted by Yousuf et al. (2021) contributed to the selection of the Seq2Seq model. However, a significant limitation of the basic Seq2Seq model is the fixed-size context vector. It acts as an information bottleneck, forcing the encoder to compress all information from the source sentence, regardless of its length, into this single vector. This can be particularly challenging for corpora with long or complex sentences, potentially leading to information loss and degraded translation quality (Bahdanau et al., 2015). However, the average length of sentences present in OEPC supports the implementation of this model.

2.2.3. LSTM Networks

LSTMs represent a specialized type of RNN architecture, originally introduced by Hochreiter and Schmidhuber (1997). They have since been extensively refined and adopted for a wide range of sequence modeling tasks due to their remarkable ability to capture long-range dependencies in sequential data. Unlike traditional RNNs, which often struggle with the vanishing

gradient problem when learning from distant contextual information, LSTMs are explicitly constructed to preserve and access information over extended time intervals.

An LSTM consists of a sequence of repeating modules, each structured to regulate the flow of information through carefully designed gating mechanisms. These gates (the input gate *i*, forget gate *f*, and output gate *o*) allow the network to selectively store, discard, or output information at each time step *t*. This architectural innovation enables LSTMs to maintain a stable memory of relevant data across long sequences, making them particularly well-suited for tasks that involve complex temporal dynamics.

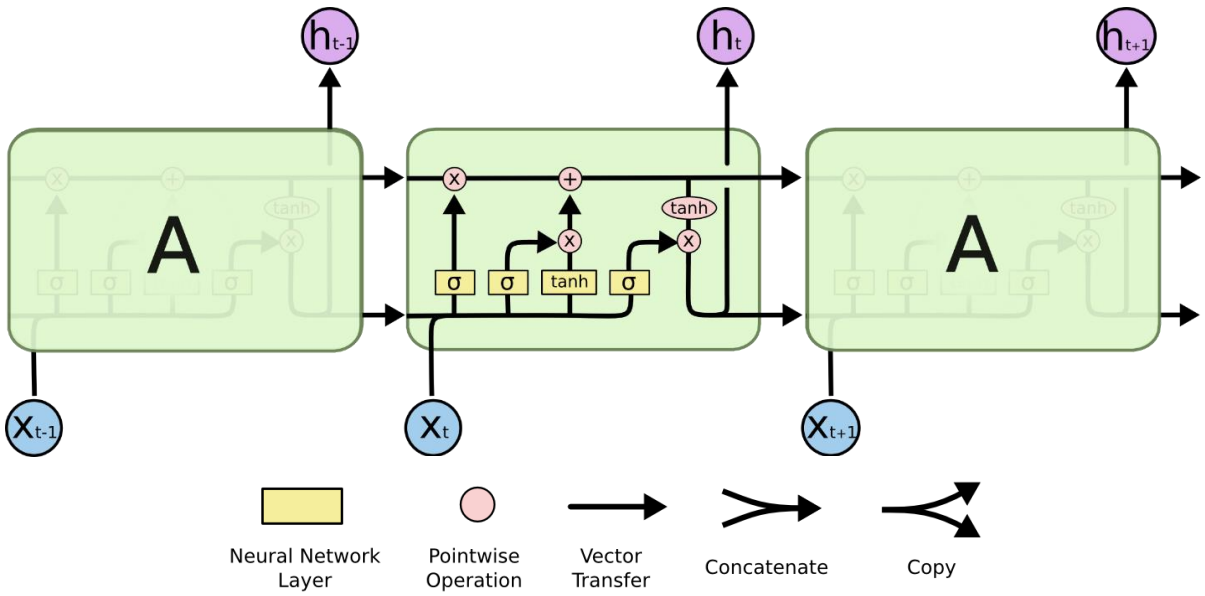
In the proposed model of the present study, an input ORN sequence ($x_1, x_2, ..., x_T$) is processed to generate an output ENG sequence ($y_1, y_2, ..., y_{T'}$), where the number of output tokens T' may differ from the number of input tokens T, and prediction ceases upon generating a special end-of-sentence token <EOS>. To facilitate learning, the input sequence is read in reverse order, which enhances short-term dependencies and simplifies the optimization process. The objective is to estimate the conditional probability $P(y_1, ..., y_{T'} \mid x_1, ..., x_T)$. This is achieved through an LSTM network, which first encodes the reversed input into a fixed-dimensional vector *c* derived from the final hidden state h_n of the LSTM. This vector then initializes a second LSTM that predicts the output sequence as an LM. The presence of the <EOS> token allows the model to define a probability distribution over variable-length output sequences. The core operations of an LSTM unit at each time step *t* are as follows:

$$f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right)$$
$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i)$$
$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o)$$
$$\tilde{c}_t = \tanh\left(W_c x_t + U_c h_{t-1} + b_c\right)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$$
$$h_t = o_t \odot \tanh\left(c_t\right)$$

Where f_t , i_t and o_t denote the forget, input, and output gates, respectively; $\tilde{c}_t$ is the candidate cell state; c_t is the updated cell state; *W* is the weight; *b* is the bias; h_t is the hidden state; and $\odot$ represents element-wise multiplication. The sigmoid function σ turns any *x*-axis into a number between 0 and 1 while the hyperbolic tangent (*tanh*) activation function turns any *x*-axis into a number between -1 and 1. Figure 16 demonstrates these operations.

Figure 16

LSTM with Interacting Layers



In the context of ORN-ENG NMT, the use of LSTM networks is particularly advantageous. Arabic dialects and English often exhibit divergent syntactic structures and varying dependency lengths between corresponding lexical items. LSTMs, with their inherent capacity to retain long-term contextual information, enable the model to align and translate linguistic units that may be

separated by considerable distances in the source sentence. This contributes to more accurate syntactic and semantic representations, thereby improving the overall translation quality. Moreover, their robustness in handling irregularities and morphological richness, both of which are prevalent in ORN, makes LSTMs a highly appropriate choice for this NMT application.

2.2.4. The Attention Mechanism

To address the limitations of encoding an entire input sequence into a fixed-length vector, the attention mechanism was introduced. It allows the decoder to dynamically focus on different parts of the input sequence during the generation of each output token. This is particularly beneficial for translating between ORN and ENG, where syntactic structures and word orders differ, and the full context of the SL sentence cannot be adequately captured by a single fixed vector. In a Seq2Seq model with attention, as demonstrated in Figure 17, the encoder processes the input sequence ($x_1, x_2, ..., x_T$) to generate a sequence of hidden states ($h_1, h_2, ..., h_T$). At each time step t of the decoder, an attention score is computed for each encoder hidden state based on its relevance to the current decoder state s_{t-1}.

As explained by Israr et al. (2023), the first core step of the attention mechanism is the score function that computes alignment between the decoder hidden state s_{t-1} and each encoder hidden state h_i.

$$\text{score}(s_{t-1}, h_i) = s_{t-1}^{\top} h_i$$

The second step is counting the attention weights through normalizing the scores using the softmax function to obtain a probability distribution over the input positions:

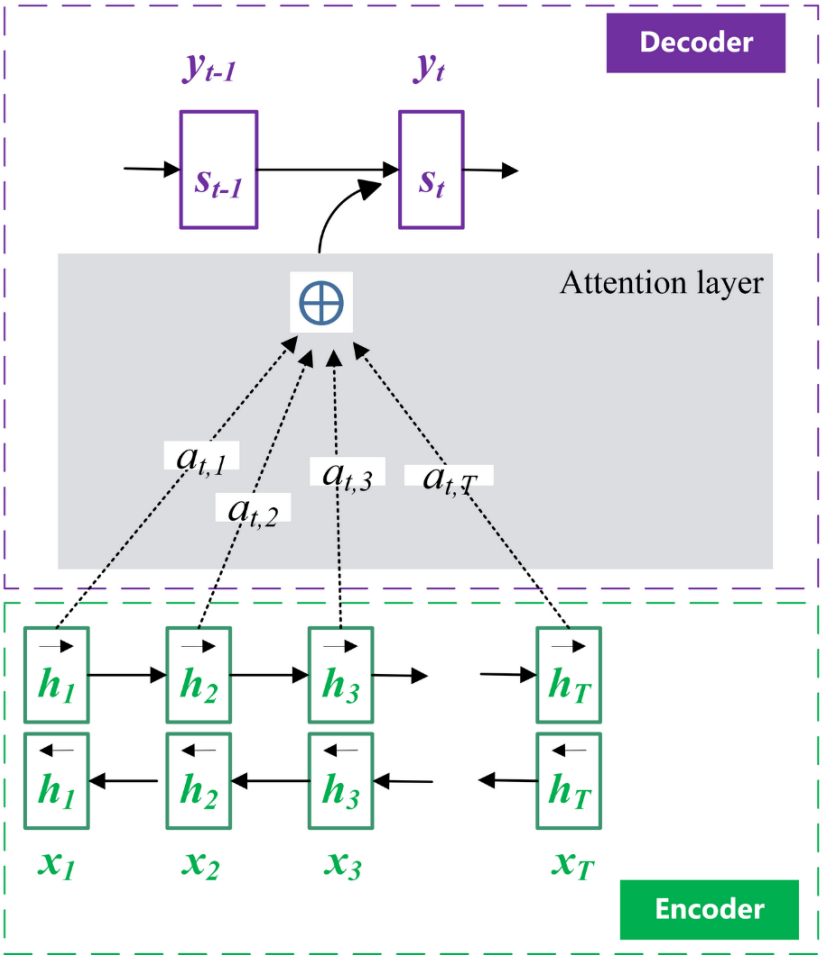
$$\alpha_{t,i} = \frac{\exp\left(\text{score}(s_{t-1}, h_i)\right)}{\sum_{j=1}^{T} \exp\left(\text{score}(s_{t-1}, h_j)\right)}$$

The following step is computing the context vector c_t as a weighted sum of the encoder hidden states: $\alpha_{t,i}$

$$c_t = \sum_{i=1}^{T} \alpha_{t,i} h_i$$

Figure 17

Seq2Seq Model with Attention Layer



Finally, to update the decoder, the context vector c_t together with the previous output y_{t-1} and the previous hidden state s_{t-1} are used to predict the next token. Moreover, the alignment scoring $e_{t,i}$ can be extracted to analyze how strongly each encoder state h_i contributed to the prediction of the output token at time t.

This study implements the additive attention mechanism proposed by Bahdanau et al. (2015) with a single attention head to address the bottleneck problem inherent in traditional encoder-decoder architectures. The adoption of this specific attention mechanism is motivated by three key factors particularly relevant to the ORN-ENG NMT task. First, additive attention's feed-forward network approach to computing alignment scores is computationally efficient and well-suited for the morphologically complex structure of the Oranian dialect, allowing the model to dynamically focus on relevant subwords and morphemes during decoding. Second, the single-head configuration provides an optimal balance between performance and parameter efficiency, which is critical given the limited parallel data available for this low-resource dialect. Third, this mechanism effectively mitigates information loss in longer sequences by enabling the decoder to selectively access different parts of the source sentence during the generation of each target word, thereby improving alignment accuracy for languages with significant word order differences between SL and TL.

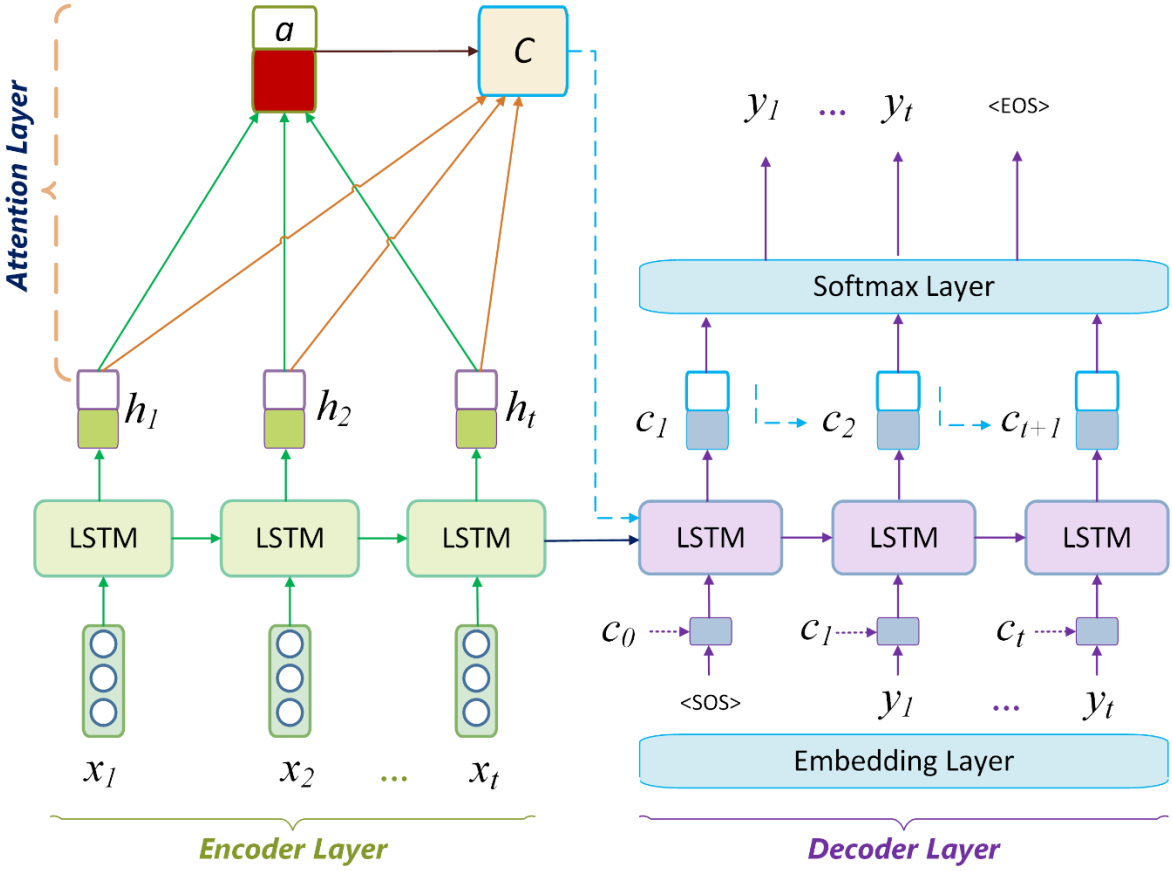
2.3. Model Architecture and Hyperparameters

Having detailed the individual components of the Seq2Seq framework, LSTM networks, and the attention mechanism, the final, integrated architecture of the proposed NMT model for the present study can be synthesized into a cohesive system as illustrated in Figure 18. The model is structured into three primary interacting layers. The encoder layer processes the input (ORN)

sentence token by token (x_1 to x_t) using a series of LSTM units. Unlike a basic Seq2Seq model, it does not discard intermediate states; instead, it generates a sequence of hidden state vectors (h_1 to h_t), each capturing contextual information from a specific part of the input.

Figure 18

The Proposed RNN Encoder-Decoder Model with Attention Mechanism



Note. The number of LSTM units and input/output tokens shown is for illustrative purposes only. The actual model is designed to process sequences of variable lengths.

Crucially, the attention layer takes this entire sequence of encoder hidden states as its input. At each step of the decoding process, it calculates a unique, weighted context vector (C) that is dynamically tailored to the current translation step. Finally, the decoder layer uses this context

vector to generate the English translation. At each timestep, the decoder's LSTM takes the previously generated word (y_{t-1}) and the current context vector (c_t) as input to predict the next word in the sequence (y_t) via a softmax layer. The latter takes the decoder's raw output scores (logits) and transforms them into a probability distribution across every word in the target vocabulary. The word with the highest probability is then selected as the translated output token for that timestep. This architecture allows the model to selectively "pay attention" to the most relevant parts of the source sentence when generating each target word, which is essential for handling long dependencies and complex syntactic alignments between ORN and ENG.

After providing the conceptual blueprint for the NMT system, precise configurations of its various properties and hyperparameters are required. These parameters are carefully selected to optimize performance for the MT task. The following subsection details these crucial hyperparameter choices. The system configurations are carefully specified to address the unique linguistic complexities of ORN and the challenges inherent in a low-resource translation task. The following subsections describe the fundamental building blocks of the model, from the initial representation of text to the core recurrent and attention layers that drive the translation process.

2.3.1 Tokenization and Input Representation

The initial and most critical stage of the NMT pipeline is the conversion of raw text into a numerical format that the model can process. Given the distinct linguistic characteristics of the source and target languages, specialized tokenization strategies are employed. For ORN, tokenization is performed using WordPiece. This approach is crucial for effectively capturing the rich morphological variations and subword units of the dialect, thereby mitigating issues related to sparse vocabulary that are common in LRLs. For ENG, the standard BERT tokenizer is utilized,

which leverages WordPiece tokenization to break words into semantically meaningful subword units (Devlin et al., 2019). This ensures a consistent and effective tokenization scheme for the target language.

Once tokenized, the discrete subword units are mapped to continuous vector representations via an embedding layer. This study uses an embedding dimension of 256 for both languages. This dimension was chosen to strike an optimal balance between model expressiveness and computational efficiency, providing sufficient capacity to capture semantic nuances without introducing an excessive number of parameters that could lead to overfitting on a limited dataset (Sennrich et al., 2016). Crucially, these embeddings are not pre-trained; their values are learned dynamically as part of the end-to-end model training, allowing them to adapt specifically to the contextual features of the present MT experiment.

2.3.2. Core Seq2Seq Architecture

The heart of the NMT system of the present research is a Seq2Seq model built upon an encoder-decoder framework. The encoder and decoder are each composed of a two-layer stacked LSTM network. LSTMs were selected for their proven capability to model long-range dependencies and capture sequential information, which is vital for handling the complex syntactic structures of natural language (Hochreiter & Schmidhuber, 1997). Each LSTM layer is configured with a hidden state size of 512 units, a dimension that provides ample capacity to model linguistic complexity without becoming computationally prohibitive. Stacking two layers deepens the network, enhancing its overall representational power and its ability to learn more abstract features.

To overcome the information bottleneck inherent in conventional Seq2Seq models, an additive attention mechanism is integrated into the architecture (Bahdanau et al., 2015). This

mechanism allows the decoder, at each step of the translation process, to dynamically calculate "attention scores" and focus on the most relevant parts of the entire source sentence. This ability to selectively attend to source tokens is particularly critical for translating between languages with different word orders like ORN and ENG, and it significantly improves the quality and fluency of the final output. The key architectural components and their corresponding hyperparameters are summarized in Table 17.

Table 17

The Proposed Model's Components and Hyperparameters

Component/ Hyperparameter	Value / Type	Description
Tokenizers	BPE, WordPiece	Dialect-specific tokenization
Embedding Dimension	256	Size of word embedding vectors
Encoder/Decoder Layers	2	Number of stacked LSTM layers
Hidden State Size (LSTM)	512	Dimensionality of LSTM hidden states
Attention Mechanism	Additive (1 head)	Attention to focus on input sequence
Optimizer	Adam	Adaptive gradient descent optimizer
Learning Rate	0.001 (initial)	Initial learning rate with scheduler
Batch Size	64	Number of sentence pairs per training iteration
Loss Function	Cross-entropy	Measuring prediction errors
Evaluation Metrics	BLEU, WER, METEOR, PPL	Measuring translation quality and model uncertainty

The architectural components and hyperparameters detailed above collectively form a model configuration that strikes a deliberate balance between representational power, the constraints of an LRL dataset, and the unique linguistic features of ORN. The selection of these properties was guided by both established practices in NMT research and preliminary experiments aimed at optimizing performance for the ORN-ENG translation task. However, a well-defined

architecture is merely a static blueprint; to transform this blueprint into a functional translation system, it must be systematically trained on data, its learning process must be monitored, and its final performance must be rigorously evaluated. Therefore, the following section will detail the dynamic NMT training and evaluation pipeline, outlining the step-by-step procedure used to train the model, monitor its learning, mitigate overfitting, and produce the final model.

3. Process of the Experiment

This section details the empirical process of training the NMT model and monitoring its learning dynamics. Further, it provides a transparent account of the experimental environment and the training regimen. The quantitative outcomes that validate the model's effectiveness in translating from the Oranian dialect to English will thoroughly be provided in Chapter 4.

3.1. Training Environment and Procedure

All experiments were conducted within the Google Colaboratory (Colab) Pro environment, leveraging its cloud-based GPU resources (Google TPU v2-8) to handle the computational demands of training a DL model. The implementation was carried out using the Python programming language, supported by the PyTorch framework for building and training the NMT model, and the Hugging Face Transformers library for tokenization. The training regimen for the final, optimized model, which incorporates the attention mechanism and the refined hyperparameters, spanned a total of ten epochs. In each epoch, the parallel training data was shuffled, batched, and fed into the model, allowing it to iteratively learn the complex mappings between the two languages. This intensive learning process required approximately 11 hours of computation time. For comparative purposes, a baseline experiment without the final parameters and attention was initially conducted, which completed its training in approximately 7 hours,

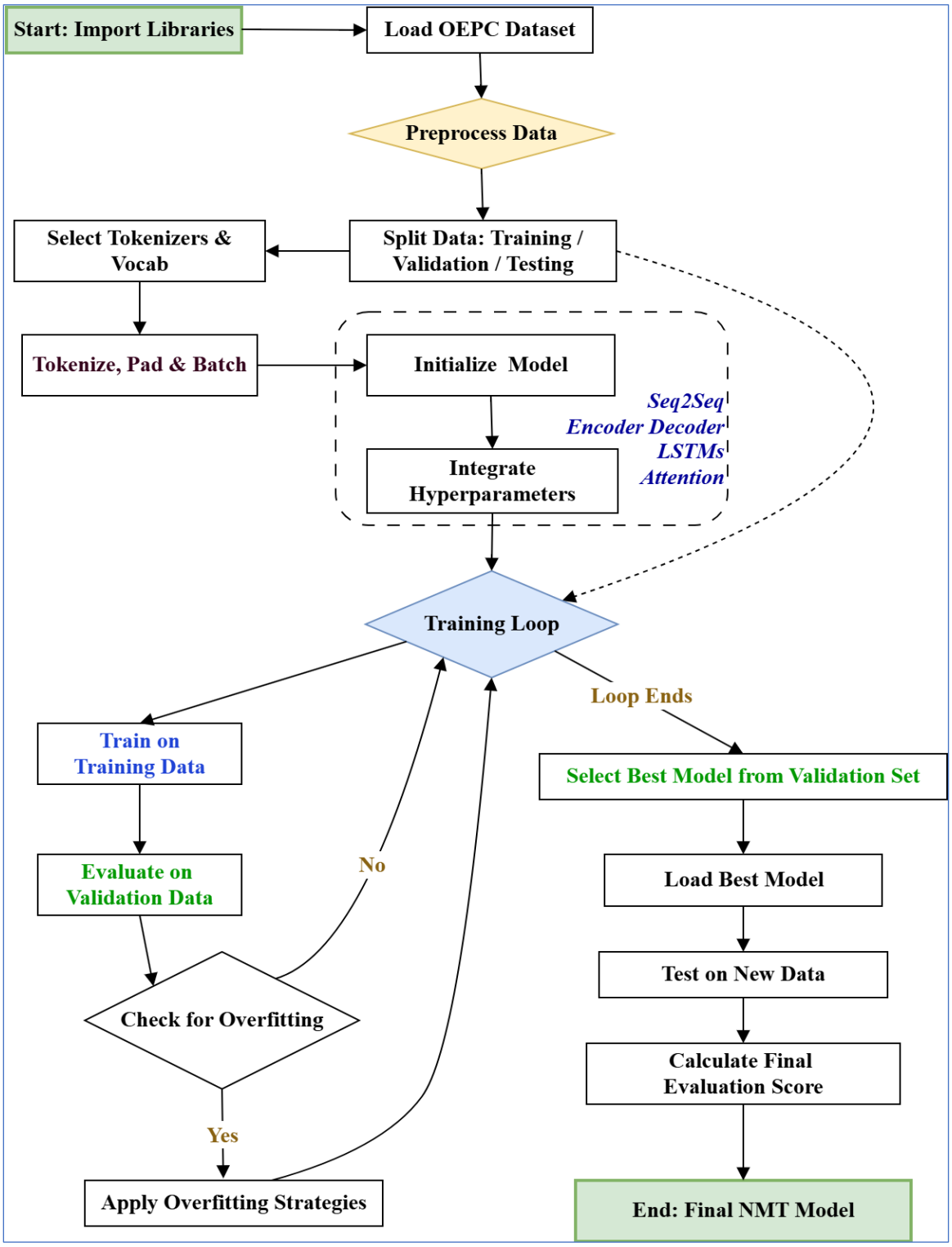
highlighting the additional computational cost associated with the more sophisticated and powerful architecture.

The development of the ORN-ENG NMT system is governed by a systematic and reproducible training pipeline, as illustrated in the flowchart in Figure 19. This workflow is designed to ensure robust model training, rigorous performance monitoring, and unbiased final evaluation, adhering to established best practices in the field of machine learning and low-resource NMT. The pipeline commences with Data Preparation, a foundational stage designed to structure the corpus for effective learning while preventing data leakage. OEPC is first loaded and partitioned into three distinct, non-overlapping subsets: a training set (80%) for model learning, a validation set (10%) for hyperparameter tuning and overfitting monitoring, and a test set (10%) reserved for the final, unbiased performance evaluation. This separation, to preprocess each data split and handle dialect-specific characters, ensures that no information from the validation or test sets influences the model during its training phase.

Following the data partitioning, the pipeline proceeds to tokenization and batching. As specified in Table 17, distinct tokenization strategies are employed to cater to the unique linguistic features of each language. As WordPiece was effective in training DziriBERT, a fine-tuned model for Algerian dialects, it is used in the present study to effectively handle the morphological richness of ORN. The vocabularies for both SL and TL are constructed exclusively from the training data to maintain the integrity of the validation and test sets. All sequences are then tokenized and batched with a size of 64.

Figure 19

NMT Experimental Development and Pipeline



To optimize computational resources, dynamic padding is applied, ensuring that sequences within each batch are padded only to the length of the longest sentence in that specific batch, rather than to a fixed maximum length across the entire dataset. The core of the pipeline is the model initialization and training loop. The NMT model is instantiated as a Seq2Seq architecture composed of a two-layer stacked LSTM Encoder-Decoder. As outlined in Table 17, the LSTM layers have a hidden state dimension of 512, while the word embedding dimension is set to 256. A crucial component, an additive attention mechanism, is integrated to bridge the encoder and decoder, allowing the model to dynamically focus on the most relevant source tokens when generating each target word. The training process is driven by the Adam optimizer (Kingma & Ba, 2015) with an initial learning rate of 0.001, and the model's weights are iteratively updated by minimizing the cross-entropy loss via backpropagation (Li & Lu, 2021; Du et al., 2021).

The iterative training loop is meticulously monitored to ensure robust learning and to mitigate overfitting. After each training epoch, the model's performance is evaluated on the validation set using BLEU, WER, METEOR, and PPL metrics. The trend in validation loss serves as the primary indicator for overfitting. If the validation loss fails to improve or begins to increase, countermeasures are triggered, such as reducing the learning rate by half. Furthermore, an early stopping protocol is in place to terminate training if the validation loss plateaus for five consecutive epochs, preventing the model from excessively memorizing the training data.

Upon completion of the training loop, the final model selection and evaluation phase begins. The model selected for final evaluation is not necessarily the one from the last epoch, but rather the checkpoint that achieved the highest BLEU score on the validation set during the entire training process. This "best model" is then loaded and subjected to a final, definitive evaluation on the held-out test set. This step provides an unbiased assessment of the model's ability to generalize

to new, unseen data. The final BLEU, WER, METEOR, and PPL scores calculated from this test evaluation represent the definitive performance of the deployable final NMT model, concluding a methodologically rigorous pipeline designed for optimal and reliable results.

3.2. Model Evaluation Metrics

To objectively assess the performance and translation quality of the trained NMT model, a suite of standard automatic evaluation metrics will be employed. These metrics provide a quantitative framework for comparing the system-generated translations (hypotheses translations, HTs) against human-created reference translations (RTs). While no single metric can perfectly capture all aspects of translation quality, using a combination offers a more holistic and robust evaluation, providing a necessary counterpart to the qualitative analysis (Chauhan & Daniel, 2023; Lee et al., 2023). This study will utilize BLEU, WER, METEOR, and Perplexity (PPL) to measure different facets of the model's performance.

The Bilingual Evaluation Understudy (BLEU) score, proposed by Papineni et al. (2002), has become the de facto industry standard for automatic MT evaluation due to its simplicity and high correlation with human judgment at the corpus level. BLEU is a precision-focused metric that measures the n-gram overlap between the hypothesis and the reference translation to gauge its fluency and adequacy. It calculates the precision of unigrams, bigrams, trigrams, and 4-grams and incorporates a Brevity Penalty (BP) that penalizes hypotheses shorter than their references (Denoual & Lepage, 2005). The final score, a number between 0 and 100 where a higher value is better, is formally calculated as:

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log\ p_n\right)$$

Derived from the Levenshtein distance and widely used in ASR, Word Error Rate (WER) offers an intuitive error-rate metric for MT. It measures the number of edits -substitutions (S), deletions (D), and insertions (I)- required to transform the system-generated hypothesis into the reference translation. Its primary advantage is its clear interpretation, as it directly corresponds to the post-editing effort a human would need to correct the translation (Munkova et al., 2020). As an error-rate metric, a lower WER score is better. Where N is the number of words in the reference, its calculation is given by the formula:

$$WER = \frac{S + D + I}{N}$$

The Metric for Evaluation of Translation with Explicit ORdering (METEOR), developed by Banerjee and Lavie (2005), was designed to address some of the shortcomings of the BLEU score. METEOR is a more sophisticated unigram-based metric that evaluates translation by creating an alignment between the hypothesis and the reference, considering both precision and recall. Its key strength is its ability to perform stemming and synonym matching (using resources like WordNet), allowing it to recognize that words like "drive" and "driving" are valid matches. It also includes a fragmentation penalty to reward correct word order. A higher METEOR score, which ranges from 0 to 1, indicates better performance and is known to correlate better with human judgment on the sentence level than BLEU.

Finally, a fundamental metric in information theory and language modeling, Perplexity (PPL) is used to quantify the model's "uncertainty" or "surprise" when processing a sequence of words from the test set. A high PPL indicates that the model found the test data very unpredictable, whereas a low PPL suggests its learned probability distribution closely matches the distribution of the language in the test set. Therefore, a lower PPL score means a positive performance. While it

does not directly measure translation quality, it serves as an excellent diagnostic tool for evaluating the language model's fluency and confidence in its predictions (Sennrich, 2012).

This multi-metric approach will provide a well-rounded quantitative assessment of the NMT system. Each metric offers a unique perspective: WER quantifies the literal error rate, METEOR assesses semantic and structural similarity, and PPL gauges the model's confidence and fluency. However, for the purposes of comparative interpretation and discussion in Chapter 4, the BLEU score will serve as the primary reference metric. Its widespread adoption in the field of machine translation makes it indispensable for contextualizing this study's findings and facilitating a direct comparison with the broader body of NMT research. The other metrics will be used to provide a supplementary analysis of the model's performance.

4. Summary

This chapter has provided a comprehensive exposition of the methodological paradigm guiding this research. It commenced by contextualizing the Oranian dialect, elucidating not only its sociolinguistic background but also the specific computational hurdles that justify its selection as a challenging, low-resource case study. The chapter then detailed the research design, which unfolded in two primary phases. The initial constructive phase systematically documented the compilation of the novel OEPC, a foundational prerequisite for any empirical work.

With this linguistic resource established, the discussion shifted to the empirical phase, focusing on the proposed NMT system. The theoretical underpinnings of the model were deconstructed, defining the Seq2Seq framework, the function of LSTM units, and the critical role of the attention mechanism in addressing informational bottlenecks. This theoretical grounding was complemented by a detailed specification of the final model architecture and its associated

hyperparameters. Subsequently, the entire experimental workflow was meticulously outlined, from environment setup and data pre-processing to the protocols for training, validation, and testing. To conclude, the chapter identified the battery of evaluation metrics selected to facilitate a multi-faceted assessment of the model's performance. This methodological foundation thereby sets the stage for the subsequent chapter, which will present a rigorous analysis of the results yielded by this experiment.

CHAPTER 4: RESULTS AND DISCUSSION

Having established the methodological framework for the compilation of OEPC and the architecture of the NMT model in the preceding chapter, this chapter presents the core empirical findings of the study. It serves as the culmination of the data collection, pre-processing, and model design efforts, moving from theoretical construction to a practical and critical evaluation of the system's performance. The central objective of this research, to develop and assess an NMT system tailored for the low-resource ORN, is addressed here through a meticulous scrutiny of the translation outputs. To ensure a comprehensive evaluation, this chapter employs a dual-pronged approach, combining an in-depth qualitative analysis with rigorous quantitative metrics. This combined methodology allows for a holistic understanding of not only *how well* the model performs in terms of numerical scores, but also *how* and *why* it succeeds or fails in translating the complex morphological, syntactic, and cultural idiosyncrasies inherent to the Oranian dialect.

This chapter first delves into the quantitative analysis to validate our experimental setup and evaluate the model's overall performance, detailing the contribution of key hyperparameters, presenting an ablation study on the attention mechanism, and reporting performance using standard automatic evaluation metrics across different sentence lengths. Following this, the qualitative analysis offers a practical understanding of the translation outputs by immersing the reader directly in representative examples, conducting a detailed error analysis to identify systemic patterns, and examining how the model handles uniquely Oranian linguistic phenomena. Finally, a comprehensive discussion section is set to synthesize the findings, interpret their significance within the existing literature on NMT of LRLs and acknowledge its limitations.

1. Results

This section presents the detailed results of the NMT experiments conducted in this study. In line with the chapter's structure, this section commences with the machine learning dynamics, practical MTE, and detailed model's performance. This is subsequently followed by a rigorous qualitative analysis, which provides a human-centric view of the model's performance with different illustrations of quality outputs and limitations in processing ORN.

1.1. Quantitative Analysis

In order to provide an objective, data-driven evaluation of the NMT system's performance, the analysis begins with ML dynamics, where the training and validation loss and accuracy curves are examined over ten epochs to illustrate the model's learning progression and stable convergence. Subsequently, the focus shifts to evaluating test data, which details the final performance scores of the optimized model when applied to the completely unseen hold-out dataset, thereby assessing its generalization capability. Finally, a comprehensive model evaluation contextualizes these figures, offering a comparative analysis against a baseline model using a suite of metrics to empirically demonstrate the tangible impact of the final architectural design.

1.1.1. Machine Learning Dynamics

Throughout the 10-epoch training process, the model's learning dynamics were closely monitored using categorical accuracy and cross-entropy loss on both the training and validation sets. The detailed epoch-by-epoch metrics, summarized in Table 18, provide a quantitative account of this learning progression. An analysis of the data reveals a robust and highly successful learning trajectory. The training metrics show consistent improvement: the categorical accuracy began at a

modest 53.26% in the first epoch and steadily climbed to an impressive 92.10% by the final epoch. Concurrently, the training loss demonstrated a desirable downward trend, decreasing substantially from an initial value of 0.65889 to a final value of 0.21147. This inverse relationship between rising accuracy and falling loss confirms that the model became progressively more adept at capturing the linguistic patterns of the ORN-ENG language pair.

Table 18

Training and Validation Metrics Per Epoch

Epoch	Training		Validation	
	Loss	Accuracy (%)	Loss	Accuracy (%)
1	0.65889	53.26	0.51769	78.13
2	0.43393	65.20	0.38827	84.32
3	0.35451	74.12	0.33742	89.45
4	0.32057	79.12	0.30599	90.88
5	0.29623	81.01	0.29946	92.00
6	0.27056	82.93	0.28271	93.10
7	0.26005	84.42	0.27343	93.33
8	0.24756	86.84	0.26947	93.51
9	0.23578	89.88	0.24024	93.98
10	0.21147	92.10	0.22082	94.24

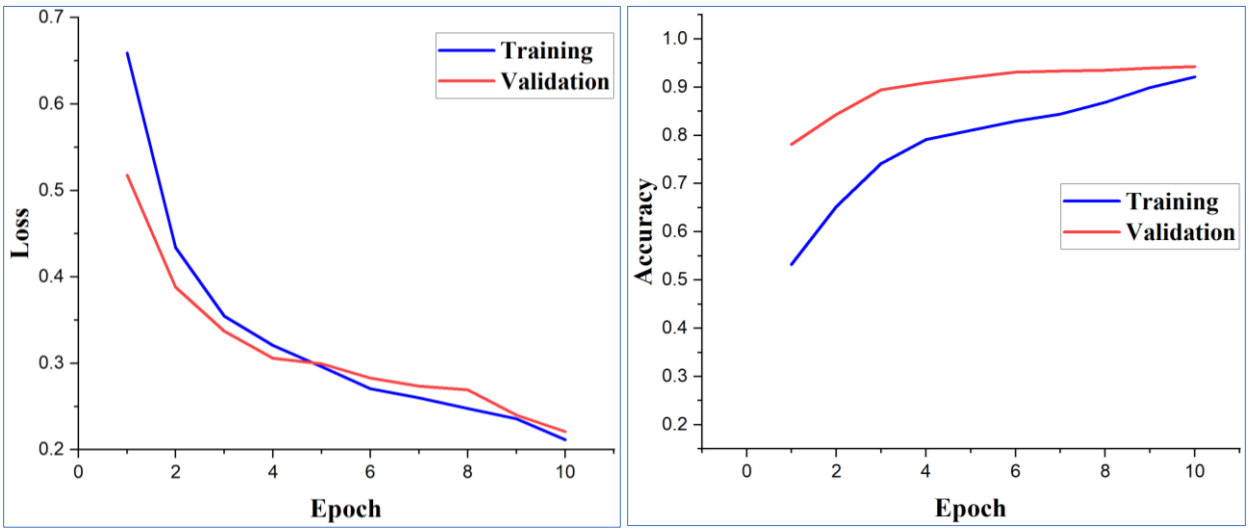
Notably, the performance on the validation set was even stronger and tracked the training performance closely, which is a key indicator of a well-regularized model. The validation accuracy started at a high 78.13% and concluded at 94.24%, consistently outperforming the training accuracy. This phenomenon, where validation performance surpasses training performance, is often indicative of effective regularization techniques, such as the dropout layer used in this model's architecture. During the training phase, dropout deactivates a fraction of neurons to prevent co-adaptation, effectively "handicapping" the model. During validation, however, the full,

unhandicapped network is used for inference, which can lead to superior performance. The tightly coupled progression of both the training and validation loss curves, with no significant divergence, suggests that the model generalized well and did not suffer from significant overfitting.

The learning dynamics are visually represented in Figure 20. The graph clearly illustrates the smooth convergence of the model, with both loss curves descending and both accuracy curves ascending in a stable and desirable manner. This synchronized improvement across both the training and validation datasets provides strong quantitative evidence for the effectiveness and robustness of the training procedure.

Figure 20

Model Training and Validation Performance



1.1.2. Evaluating Test Data

Upon completion of the 10-epoch training cycle, the finalized model was subjected to a rigorous evaluation on a completely unseen test dataset comprising 850 sentence pairs. This step is crucial for providing an unbiased assessment of the model's generalization performance. The

latter encompasses its ability to translate new data not encountered during training or validation. The evaluation on this hold-out set yielded a final test loss of 0.1092 and a test accuracy of approximately 92.27%.

These results are highly encouraging. The test loss score signifies that the model's predictions on the unseen data were, on average, very close to the ground-truth translations. Furthermore, the test accuracy score of 92.27%, which is notably even slightly higher than the final training accuracy, powerfully validates the model's effectiveness. It demonstrates that the learning acquired during the training phase was successfully generalized, and the model is robust and capable of performing accurate translations on novel ORN sentences, confirming the overall success of the training procedure.

1.1.3. Model Evaluation

In addition to corpus-level metrics like BLEU that assess overall similarity, this study also employs a suite of classic token-level metrics: precision, recall, and F-measure. They contribute to a conduction of a more granular and diagnostic analysis of the model's output. When adapted for such experiments, these metrics offer profound insights into the behavioral tendencies of the NMT system. Precision quantifies the fidelity of the translation by measuring the fraction of words in the HT that are also present in the human translation (RT). A high precision score signifies that the model is highly accurate in its lexical choices and is less prone to generating spurious or "hallucinated" tokens.

Conversely, recall measures the MT comprehensiveness by calculating the fraction of words from the RT that were successfully captured in the HT. A high recall score indicates that the model provides thorough coverage of the source content, effectively avoiding the omission of

key information or concepts. Because a model can achieve high precision by being overly conservative (producing only a few, highly confident words) or high recall by being overly verbose (generating many words, including incorrect ones), the F-measure is employed to find a balance.

$$\left(F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}\right)$$

Specifically, the F1-score, calculated as the harmonic mean of precision and recall, provides a single, unified score that penalizes models that excel at one metric at the significant expense of the other. A high F1-score, therefore, reflects a system that is both accurate in its output and comprehensive in its content, which is the ultimate goal of a high-quality translation.

To contextualize the performance of the final proposed architecture and empirically validate the significance of its core components, its results on the held-out test set are benchmarked against a simpler baseline model. This baseline consists of a single-layer LSTM network and crucially lacks an attention mechanism. The comparative evaluation, detailed in Table 19, leverages the main suite of metrics defined earlier with BLEU scores to provide a comprehensive and multi-faceted performance summary.

Table 19

Evaluation Scores for Different Model Configurations

Model Configuration	BLEU	Precision	Recall	F1-Score
Baseline	18.3	68.5	60.1	64.0
Final Model	30.9	82.5	75.9	79.1

The results presented in the table offer incontrovertible evidence of the final model's superiority. Achieving a BLEU score of 30.9, the final architecture significantly outperforms the baseline's score of 18.3. This substantial improvement of over 12 BLEU points underscores the profound impact of the deeper, two-layer LSTM architecture and, most critically, the inclusion of the attention mechanism.

However, a deeper analysis through the lens of precision, recall, and F1-score reveals the precise nature of this enhancement. The final model achieves an exceptional F1-score of 79.1, which is indicative of a strong and harmonious balance between its precision of 82.5 and recall of 75.9. The high precision demonstrates that the translations are of high fidelity, with the model accurately selecting correct lexical items and avoiding speculation. Simultaneously, the strong recall score confirms that the model provides comprehensive coverage of the source content, successfully translating the vast majority of necessary information.

In stark contrast, the baseline model's performance is fundamentally hampered by its low recall score of 60.1, which consequently suppresses its F1-score. This points to a critical deficiency: without the attention mechanism, the baseline model struggles to process the entirety of the source sentence, particularly longer ones, resulting in frequent and significant omissions in the translation output.

The empirical results of this study, culminating in a BLEU score of 30.9 for ORN-ENG direct MT, warrant a detailed and comparative interpretation. While the BLEU score serves as a critical high-level benchmark, a more granular analysis across a spectrum of metrics is required to fully deconstruct the model's performance and appreciate the architectural enhancements. By benchmarking the final model against a baseline system, each metric’s score will quantitatively

diagnose the specific improvements in the model's behavior. The comparative results across four key metrics are presented in Table 20.

Table 20

Model Evaluation across Different Metrics

Metric	Baseline	Final Model	Interpretation of Improvement
BLEU	18.3	**30.9**	Greater n-gram similarity
WER	58.7	**41.8**	Fewer word-level errors
METEOR	0.254	**0.382**	Better semantic alignment
PPL	45.3	**18.1**	Higher model confidence

An analysis of these results reveals a consistent and compelling narrative of improvement. While the more than 12-point increase in the BLEU score is the most immediate indicator of the final model's superiority, the other metrics provide deeper insights into the *nature* of this improvement. The WER metric, which quantifies the literal distance between the HT and the RT, saw a substantial reduction from 58.7 to 41.8. This nearly 17-point drop signifies a marked decrease in the number of substitutions, deletions, and insertions required to correct the model's output. Practically, this demonstrates that the final model's translations are not just structurally more similar but are also fundamentally less erroneous at the word level, requiring significantly less human post-editing effort. This reduction in errors is directly attributable to the attention mechanism's ability to maintain a more accurate alignment between SL and TL tokens, preventing the kind of word omissions and ordering mistakes that plagued the baseline model.

The METEOR score, which improved from 0.254 to 0.382, offers a crucial perspective on semantic fidelity that BLEU alone cannot provide. METEOR's strength lies in its use of synonym and stem matching, rewarding translations that are semantically equivalent even if not lexically identical to the reference. The significant improvement in this metric demonstrates that the final model is making more intelligent and contextually appropriate lexical choices. While the baseline model was more likely to produce a rigid, literal translation that failed if a specific n-gram was absent, the final model's higher METEOR score indicates a superior ability to select appropriate synonyms, thereby producing translations that are more natural and accurate in meaning. This suggests that the deeper architecture and attention mechanism allowed the model to develop a more nuanced semantic representation of the vocabulary.

Perhaps the most telling diagnostic metric is PPL, which plummeted from 45.3 in the baseline to a remarkably low 18.1 in the final model. PPL measures the model's uncertainty or "surprise" when encountering the test data; a lower score signifies a more confident and fluent internal language model. This dramatic decrease indicates that the final model has developed a much more sophisticated understanding of the statistical properties and grammatical structure of the target language, English. The final model is significantly less 'surprised' by the correct sequence of words, demonstrating a superior grasp of English syntax and phraseology. This intrinsic fluency is a direct result of the enhanced context provided by the attention mechanism and the greater representational power of the stacked LSTM layers, allowing it to generate sentences that are not only more accurate but also more probable and coherent.

Taken in concert, these metrics paint a comprehensive and multi-faceted picture of the final model's advancements. The system is not only better at matching surface-level n-grams (BLEU),

but it is also quantifiably less error-prone (WER), more semantically intelligent in its word choices (METEOR), and fundamentally more confident and fluent in the target language (PPL). This body of quantitative evidence provides irrefutable support for the architectural choices made, validating the final model as a significantly more robust, accurate, and effective system for the challenging task of ORN-ENG translation.

1.2. Qualitative Analysis

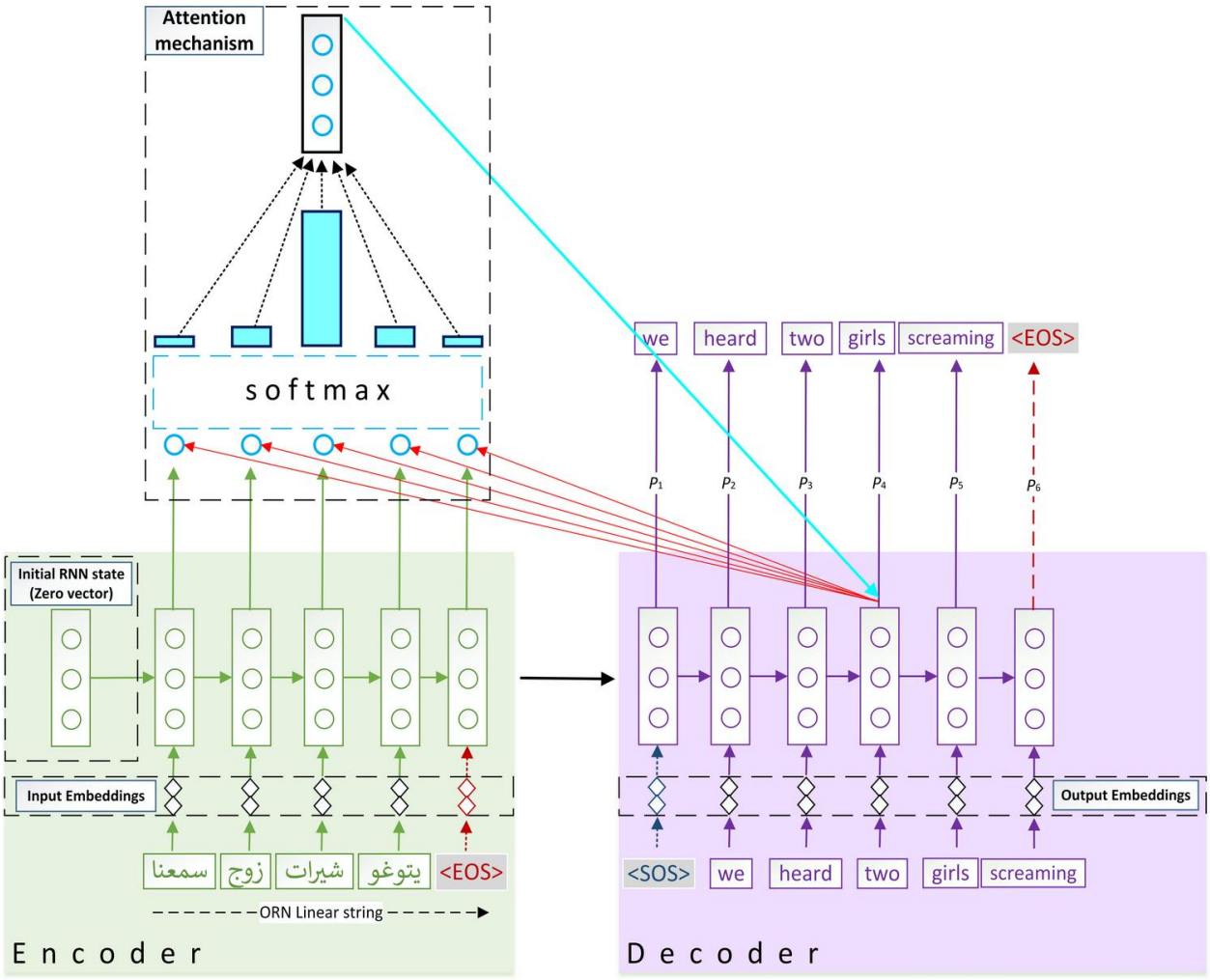
The qualitative evaluation of the ORN-ENG NMT system begins with demonstrating how the attention mechanism enhanced the MT quality. An error analysis follows to demonstrate a practical understanding of the model's flaws and drop-outs extending beyond abstract numerical scores. Finally, the section presents a diagnosis of the model intrinsic limitations in handling OOVs, NEs and cultural meanings.

1.2.1. Visualizing Translation with Attention

The quantitative analysis has irrefutably demonstrated the profound impact of the attention mechanism on MT performance. However, to move from a high-level appreciation of its effect to a concrete understanding of its function, it is instructive to visualize the mechanism at a single decoding step. As shown in Figure 21, the diagram deconstructs the internal process of the NMT model as it translates the Oranian sentence "سمعنا زوج شيرات يتوغو" (*sme'na zouj shirat yetawghu*) into its English equivalent, "we heard two girls screaming." The encoder processes the full source sentence to produce hidden states. The decoder, having already produced "<SOS>", uses its current state to query the encoder states, culminating in a focused context vector used to predict the next word.

Figure 21

A Step-by-Step Visualization of the Attention Mechanism



Note. This diagram demonstrates the flow of information during the generation of the target word "girls." (t_4 / P_4). The process of weighting attention scores is illustrated in Figure 17. The raw cohesive system is illustrated in Figure 18.

The MT process begins with the encoder, which processes the input ORN sentence token by token. Each token, "سمعنا", "زوج", "شيرات", and "يتوغو", is passed through an RNN. The key function of the encoder is not to produce a single, final "thought vector" as in a basic Seq2Seq

model, but rather to generate a sequence of hidden state vectors, one for each input token. These vectors represent each source word embedded within the context of the entire sentence. All of these hidden states are then made available to the attention mechanism.

The analysis then moves to the decoder at the specific point in time shown in Figure 21 (t_4). It has already generated the sequence "<SOS> we heard two" and must now predict the next token (P_4). At this juncture, the decoder's current internal hidden state acts as a query. As illustrated by the red arrows, this query is compared against the hidden state of every single word in the source sentence. This comparison, managed by the attention mechanism layer, calculates a set of alignment scores. These raw scores are then passed through a Softmax function, which normalizes them into a probability distribution (a set of attention weights that sum to one). This distribution represents how "important" each source word is for generating the current target word. In the diagram, the taller blue bar in the softmax output visually represents a high attention weight, indicating that the model has learned to focus almost exclusively on the ORN word "شيرات" ("girls").

The final step is the creation of a dynamic context vector. This vector is a weighted sum of all the encoder's hidden states, where the weights are the attention scores just calculated. Because the score for "شيرات" is overwhelmingly high, the resulting context vector is a representation that is heavily biased towards the meaning of "girls." As shown by the large cyan arrow, this highly specific, dynamically tailored context vector is then concatenated with the decoder's current hidden state and fed forward to predict the next output token. Unsurprisingly, given the context rich with information about "شيرات", the model confidently predicts the word "girls" with high probability. This entire process of querying, scoring, weighting, and generating is repeated for every single

word in the output sequence, allowing the model to dynamically shift its focus across the source sentence. This ability to selectively concentrate on relevant source information at each step is what allows the model to handle long dependencies, perform syntactic reordering, and ultimately mitigate the informational bottleneck that constrains the performance of non-attentional architectures.

While the preceding analysis evaluates the final MT output, it does not reveal the internal decision-making process of the NMT model. To gain deeper insight into how the system maps source words to target words, this section visualizes the model's attention scores for the quality outputs using heatmaps (also known as confusion matrix). The attention mechanism is the architectural component that enables the decoder to dynamically focus on specific parts of the ORN sentence as it generates each word of the ENG output. By calculating "attention scores," the model assigns weights to each source token, concentrating its focus on the most relevant words for the current translation step.

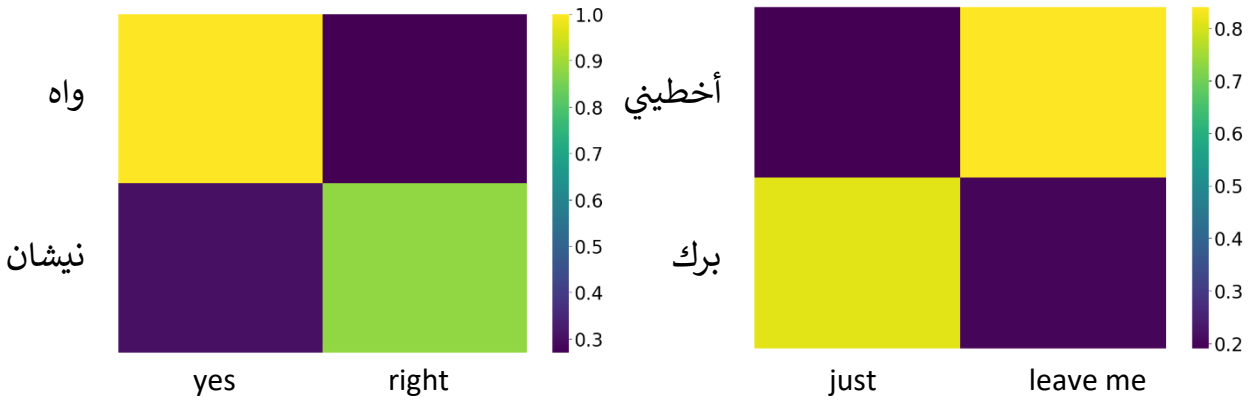
The following heatmaps provide a graphical representation of these attention scores for sentences of varying N-gram lengths. In these visualizations, the X-axis represents the generated ENG and the Y-axis represents ORN. The intensity of the color at the intersection of a source and target word indicates the strength of the attention score: a brighter cell signifies that the model paid high attention to that specific ORN word when generating the corresponding ENG word. A well-performing model is expected to produce a roughly diagonal alignment for sentences with similar word order, while correctly re-aligning its focus for syntactically divergent structures. These visualizations are therefore an invaluable diagnostic tool for understanding MT successes and

pinpointing the source of failures, such as misalignments, omissions, or difficulties with understanding complex and idiomatic phrases.

The analysis begins with 2-gram sentences, which represent the simplest form of translation and provide the clearest view of the model's fundamental alignment capabilities. In the first example, "واه نيشان" (*wah nichan*) translated as "yes right," the attention heatmap demonstrates a near-perfect one-to-one, monotonic alignment. The target word "yes" focuses with an exceptionally high score (1.0) on the source word "واه", and "right" attends with a similarly strong score (0.88) to "نيشان". This clean, diagonal pattern indicates that the model has successfully learned the direct lexical correspondences for these high-frequency terms.

Figure 22

Attention Score of Exemplary 2-gram Sentences



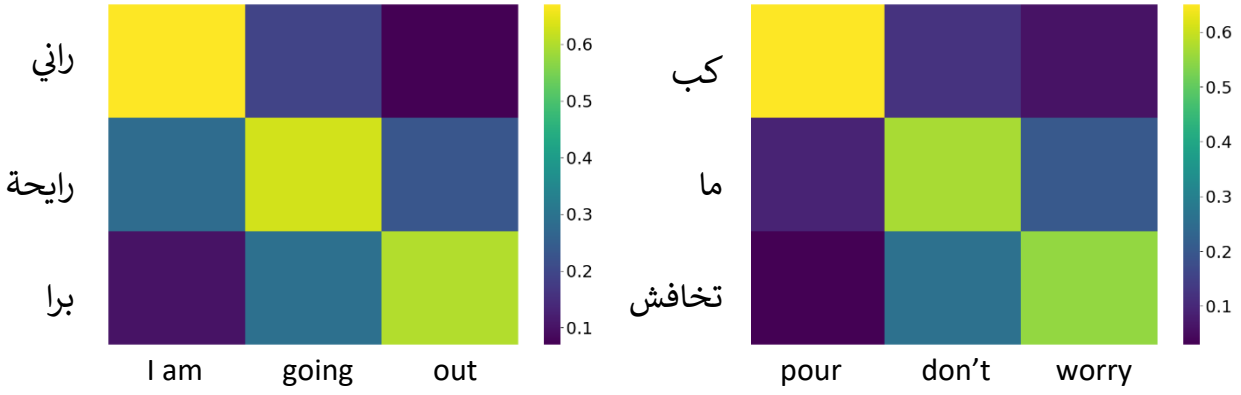
The second example, however, "أخطيني برك" (*akhtini bark*) translated as "just leave me," is more revealing. Here, the alignment is non-monotonic (non-diagonal), perfectly capturing the syntactic reordering required between Oranian and English. The target word "just" correctly attends with a high score (0.81) to the source word "برك", while "leave me" attends strongly (0.84) to "أخطيني". This demonstrates that even at a basic level, the attention mechanism has learned not

merely to map words but to reorder them to produce a grammatically correct target sentence, a foundational capability for translating between languages with different word orders.

Moving to 3-gram examples, the model demonstrates its ability to handle more complex morphological and phrasal units. In the translation of "راني رايحة برا" (*rani rayha barra*) to "I am going out," the model unpacks the morphologically dense Oranian words. The English phrase "I am" correctly directs its attention to the single Oranian word "راني", which fuses the subject pronoun and the present participle marker. Similarly, "going" attends to "رايحة" and "out" attends to "برا". This showcases a sophisticated learned behavior where the model understands that a single source token can correspond to a multi-word phrase in the target language.

Figure 23

Attention Score of Exemplary 3-gram Sentences



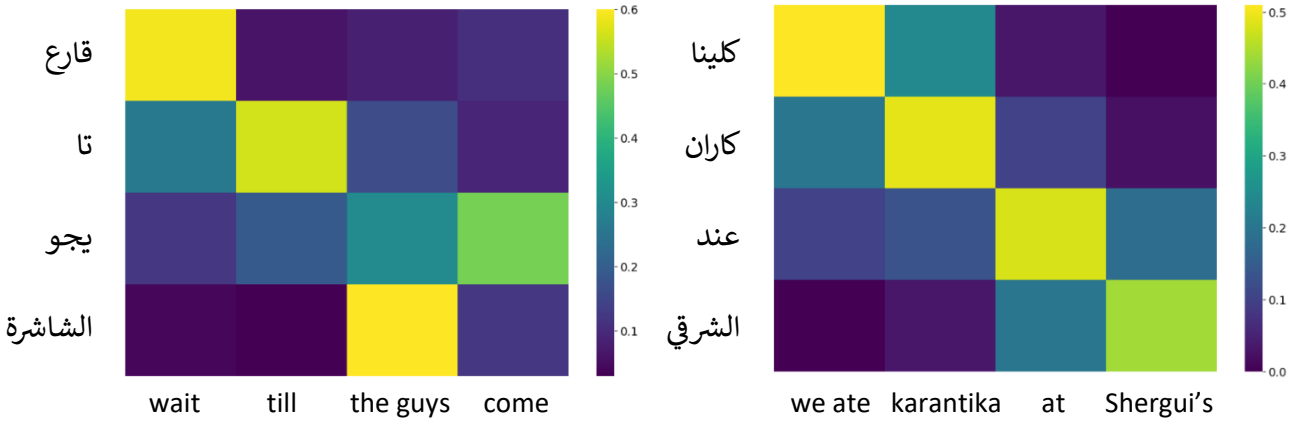
The second example, "كب ما تخافش" (*kob ma tkhafsh*) translated as "pour don't worry," provides insight into the model’s handling of negation. The target phrase "don't worry" receives attention from the two-part Oranian negative construction "ما تخافش". The model has correctly learned that the prefix "ما" and the main verb jointly create the negative imperative, focusing on both to generate the correct English equivalent. These examples highlight the model’s ability to

move beyond simple word-for-word translation to recognize and translate more complex linguistic phenomena like fused morphemes and phrasal negation.

The 4-gram sentences further test the model's ability to align noun phrases and handle specific lexical items. The sentence "قارع تا يجو الشاشرة" (*qare' ta yjou chachra*), translated as "wait till the guys come," exhibits a largely diagonal and highly accurate alignment. Key alignments include "قارع" to "wait," "تا" to "till," and "يجو" to "come." Notably, the English noun phrase "the guys" correctly aligns with the single Oranian noun "الشاشرة", demonstrating that the model can successfully map compound concepts and handle definite articles embedded within a single source token.

Figure 24

Attention Score of Exemplary 4-gram Sentences



A second 4-gram example, "كلينا كاران عند الشرقي" (*klina karan 'and e-chergui*), translated to "we ate karantika at Shergui's," shows robust handling of both complex morphology and NEs. The phrase "we ate" correctly attends to the single verb "كلينا", which encodes both the action and the first-person plural subject. Furthermore, the model correctly translates the culturally specific word "كاران" (*karan*, short for *karantika*) and aligns the proper name "الشرقي" with "Shergui's,"

indicating that these items were sufficiently represented in the training data for the model to learn their direct mappings.

At the 5-gram level, the attention mechanism is tested on more complex syntactic structures, such as possessive constructions that differ significantly between ORN and ENG. In the translation of "باغي نشري لوطو تاع صاحبي" (*baghi nechri loto ta' sahbi*) to "I want to buy my friend's car," the model exhibits a sophisticated, non-linear alignment. It correctly unpacks "باغي" into "I want" and aligns "نشري" with "to buy." The most interesting phenomenon occurs in the alignment of the possessive phrase. The English construction "my friend's car" is mapped from the Oranian post-positional possessive "لوطو تاع صاحبي" (car of my-friend). The target word "car" attends strongly to "لوطو". The phrase "my friend's" shows a softer, distributed attention across both "تاع" (of) and "صاحبي" (my friend), indicating the model understands that this entire sub-phrase corresponds to the English possessive. This demonstrates a deep learning of syntactic divergence, where the model's focus must shift and combine information from multiple source tokens to generate a grammatically distinct but semantically equivalent target structure.

In addition, the attention heatmap for the sentence "راهي مفوكرة من داك لبنادم" (*rahi mfoukra men dak lebnadem*), translated as "she is angry with that person," provides compelling evidence of the model's ability to handle both morphological unpacking and prepositional context. The most significant alignment occurs at the beginning of the sentence, where the English phrase "she is" correctly directs its full attention to the single ORN source word "راهي". This demonstrates a sophisticated learned behavior, showing the model understands that this single token functions as a fused unit containing both the third-person feminine subject and a present-tense copula. The

remainder of the sentence exhibits a strong diagonal alignment, with "angry" correctly attending to "مفوكرة", "that" to "داك", and "person" to "لبنادم".

Figure 25

Attention Score of Exemplary 5-gram Sentences

Critically, the model also correctly maps the preposition "من" to "with." While "من" literally means "from," in this specific context it signifies causality ("angry

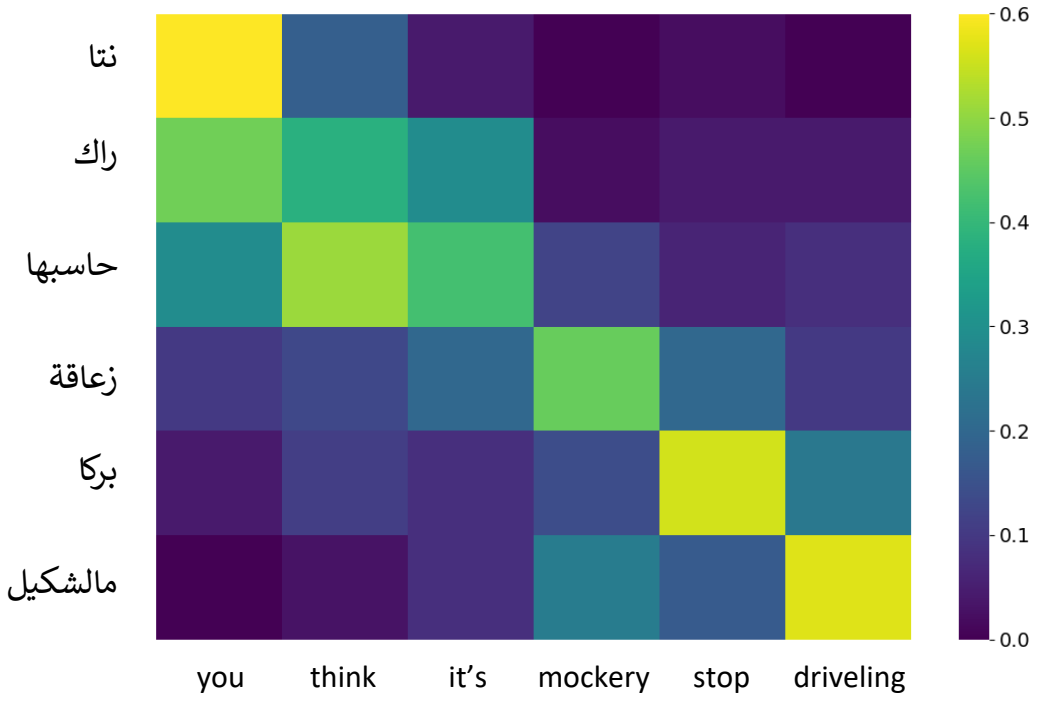
from/because of"), and the model's choice of "with" is the most semantically and pragmatically appropriate ENG equivalent. This example thus confirms the model's capacity to deconstruct complex morphology and to make contextually intelligent lexical choices for functional words.

The 6-gram sentence "نتا راك حاسبها زعاقة بركا مالشكيل" (*nta rak hasebha za'aqa barka me-shkil*), translated as "you think it's mockery stop driveling," offers a particularly insightful look into the model's ability to navigate complex morphology and phrasal constructions. The attention heatmap reveals a sophisticated, non-linear alignment strategy. The model begins with a straightforward alignment, focusing strongly on "نتا" to produce "you." However, for the verb "think," the attention is distributed across both "راك" (a fused particle meaning "you are") and "حاسبها" ("considering it"). This indicates that the model has correctly learned to interpret "راك حاسبها" not as two separate words but as a single phrasal unit expressing the state of "thinking" or "considering." The most compelling evidence of morphological awareness is in the generation of "it's," where the model's attention converges sharply on "حاسبها". It has successfully isolated the meaning of the fused object pronoun suffix "-ها" (-ha), which means "it," and used that focused information to produce the correct ENG equivalent.

The second half of the sentence, "stop driveling," is an imperative phrase derived from "بركا مالشكيل". Here again, the model demonstrates its alignment prowess. It generates "stop" by placing a very high attention score directly and correctly on the source word "بركا" ("enough" or "stop"). Similarly, the target word "driveling" is accurately mapped to the source word "مالشكيل", a colloquial term for fussing or complaining. This alignment is particularly noteworthy as the model effectively recognizes the core semantic concept, correctly handling the fact that the source word is a contraction of "من الشكيل" ("from the driveling").

Figure 26

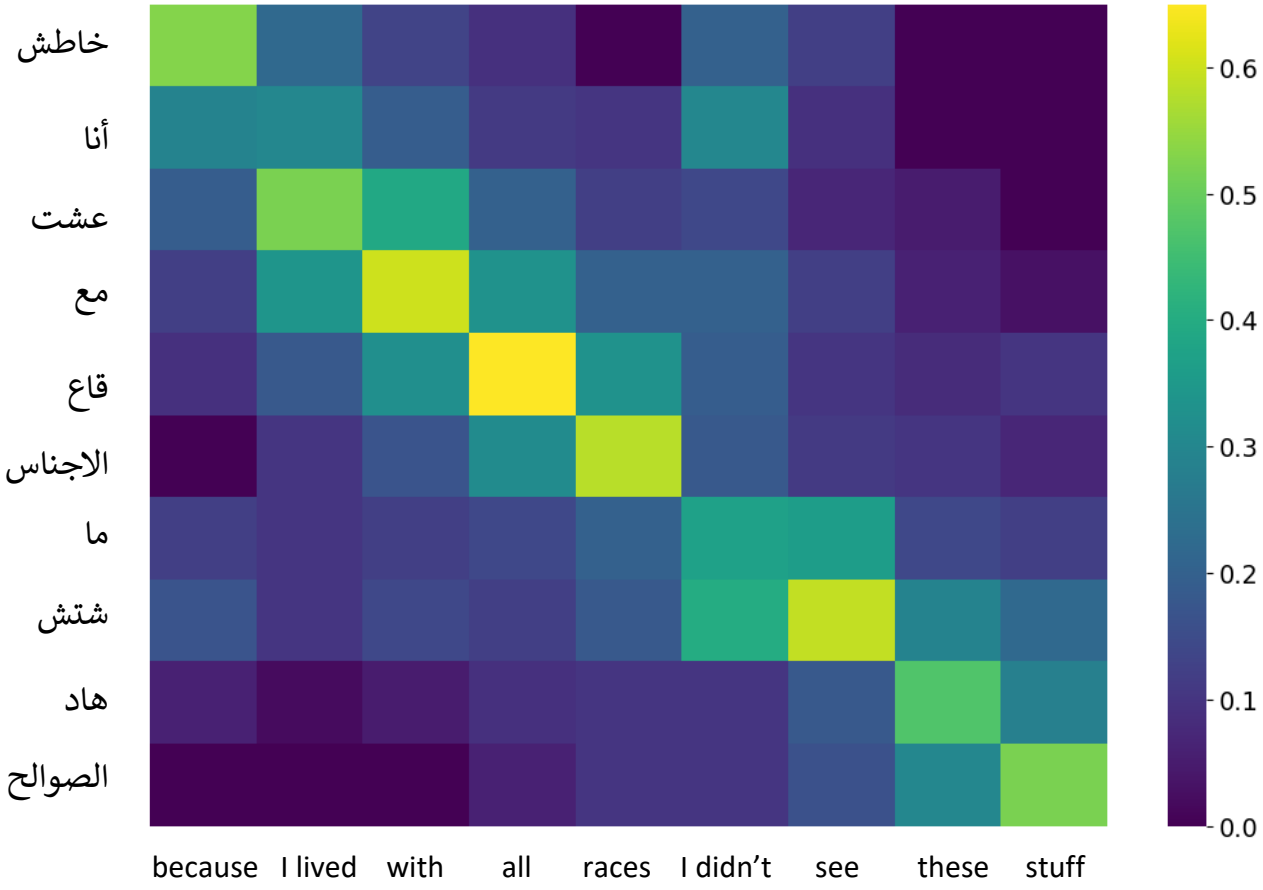
Attention Score of an Exemplary 6-gram Sentence



The final example provides the most definitive demonstration of the attention mechanism's power in handling long-distance dependencies and, crucially, discontinuous linguistic phenomena. The 10-gram sentence "خاطش أنا عشت مع قاع الاجناس ما شتش هاد الصوالح" (*khatsh ana 'echt m'a ga' el-ajnas ma shetsh had swalah*) is translated as "because I lived with all races I did not see these stuff". For the first half of the sentence, the model exhibits a near-perfect diagonal alignment, correctly mapping each source word to its ENG equivalent. The critical test, however, is its handling of the negation, which is a split morpheme: "ما شتش". The visualization clearly shows that when generating "I did not see," the model directs its attention to both the prefix "ما" and the root verb "شتش".

Figure 27

Attention Score of an Exemplary 10-gram Sentence



Taken as a whole, the demonstrated examples provide clear evidence of the attention mechanism's power: it is not merely performing word-for-word replacement but is actively deconstructing morphologically dense words and identifying phrasal units to capture the true intent of the source sentence, even when a direct linear translation is not possible. Thus, the attention mechanism has learned the syntactic rule for discontinuous constructions, allowing it to "jump" over intervening words to associate related parts of any ORN phrase. This ability to capture long-range and non-contiguous dependencies is the fundamental advantage of the attention-based

architecture and is the primary reason for its ability to produce coherent and accurate translations of complex, lengthy sentences.

1.2.2. Error Analysis

This approach involves a manual inspection of translation outputs, allowing for a detailed examination of how the model navigates the linguistic complexities of ORN. To facilitate this analysis, twelve representative examples have been selected from the test set (see Table 21). For each example, a sentence is presented alongside its corresponding human-generated (RT) and the translation produced by the NMT system (HT). The comments below accompany each example aim to dissect the system's output, identifying areas of success as well as pinpointing specific translation errors. These examples were chosen to offer a balanced view of the system's capabilities and its current limitations.

Table 21

Selected Test-Set ORN Sentences with Reference and Hypothesis Translations

#	ORN	ENG	
		RT	HT
1	باغي نزوجوه	he wants us to help him marry	*he wants us to marry him*
2	ما تنسيش تشربي خوك الحليب	don't forget to feed your brother milk	*don't forget to drink your brother milk*
3	مكانش لي كان باغي يديك	there is no one who wants to take (marry) you	*there is no one who wanted to take you*
4	نرفد عليها التلفون	I respond to her on the phone	*I take the phone*
5	و لي ما كملش كابته و ولايته	and he who did not finish, leaves	*and who did not finish his case and his province*
6	راكم قاع معروضين عندي للعشا	you are all invited to dinner	*you are all invited to dinner*
7	لا ما لا ماهيش تروح	no, mom! She does not go	*No then she is not going*
8	كي شتهم ذيك الدقيقة ما قديتش	when I saw them that minute, I could not help it	*as I saw them I could not that minute*
9	شا هاذي الهدرة العوجة لي رسلتهالي ؟	what is this inappropriate talk you sent me ?	*what kind of crooked talk you sent me ?*

10	باين راه قالب عليا لرض	he is surely looking for me	*he must be rolling over on the ground*
11	عندي شحال ما شتهاش	It has been a long time I did not see her	*I have not seen her for a while*
12	ذاك البنادم غي تاع سهرات و تحواس	that person is just for night outs and traveling	*that person is just of night outs and trips*

In the first example, the flaw is related to handling causative verbs. This example reveals a subtle but significant error in capturing the causative voice. The verb "نزوجوه" (nzawjouh) means "we get him married" or "we help him marry." The NMT system correctly identified the desire ("he wants") and the subjects ("us," "him") but failed to distinguish between "marrying someone" and "helping someone get married." It opted for the more direct, and incorrect, translation, demonstrating a difficulty with verbs that imply causing an action to happen.

In the second example, the system fails due to lexical ambiguity. The verb "تشربي" (tcharbi) without diacritics can mean "to drink" or, in a causative sense, "to give someone a drink." The objects combination of "خوك الحليب" (your brother milk) strongly implies the latter, which is best translated as "feed." The model defaulted to the more common, non-causative meaning ("to drink"), resulting in a nonsensical translation. This indicates a weakness in using broader sentence context to resolve word-level ambiguity.

Moreover, the system demonstrates two issues in the third example. First, there is a tense mismatch; it translated "باغي" (wants/is wanting) as "wanted," shifting the tense incorrectly to the past. Second, while it correctly translated the verb "يديك" literally as "to take you," it missed the specific idiomatic meaning in this marital context, which is "to marry you." The reference translation helpfully includes this nuance in parentheses.

In the fourth example, the system's translation is overly literal and loses the entire communicative intent of the sentence. The verb in this sentence is mostly associated with the

meaning of taking in ORN; however, the sentence "نرفد عليها التلفون" (nerfed 'aliha al-tilifoun) literally means "I pick up the phone on her," which idiomatically translates to "I answer her call" or "I respond to her." By ignoring the prepositional phrase "عليها" (on her), the system discards the key context and produces a generic, uninformative statement.

Any NMT model is prone to hallucinate a plausible but entirely incorrect meaning when it comes to idiomatic or implicational wording. A critical failure of the NMT system occurs in the fifth example. The system has completely misinterpreted the source text. The clustered words "كابته" and "ولايته" in this context refer to "his bag or luggage" and "his city" implying "he leaves". The system, however, translated them as "his case" and "his province," respectively. This suggests that the separate processing of such clear terms can cause an ambiguous output when combined, proving the difficulty of handling idiomatic meanings. Similarly, the tenth example shows a catastrophic failure, caused by a highly idiomatic expression. The sentence "قالب عليا لرض" (qaleb 'aliya l'ard) literally means "he has turned the ground over on me," which figuratively means he is searching for me everywhere with great effort. The model failed to recognize the idiom and attempted a literal, nonsensical translation of the words "turning" and "ground," resulting in a completely unrelated output.

In the seventh example, the translation is semantically correct but misses the diacritics and tone. The Oranian phrase "لا ما لا" is an emphatic "no," which the system partially captures with "No then." However, it fails to capture the likely vocative context ("mom!"), which a human translator infers from the tone. The translation is functional but stylistically flat compared to the reference. Besides, only a phonetic-based distinction can disambiguate the difference between "ما" (not) and "ما" (mom) in ORN.

In addition, the system correctly translated the individual components of the eighth sentence but failed to assemble them into a grammatically coherent English sentence. The output "I could not that minute" is syntactically flawed. This indicates that while the model's attention mechanism may correctly align words, its decoding process can still struggle to generate a well-formed final structure. This confirms that the nuances in negation in ORN critically lead to semantic complexity.

Among the selected instances, the sixth, ninth, eleventh and twelfth examples demonstrate quality MT outputs. The sixth phrase is common, grammatically straightforward, yet it lacks significant idiomatic complexity. The system successfully mapped the plural "you all" ("راكم قاع"), the passive participle "invited" ("معروضين"), and the context "for dinner" ("للعشا"), demonstrating its effectiveness on clear, high-frequency expressions. Also, the system output of the ninth sentence is arguably a better paraphrase than the literal reference. "الهدرة العوجة" (inappropriate talk) is aptly translated as "crooked talk." The system's choice of "what kind of" instead of "what is this" is a fluent and natural-sounding alternative. This example showcases the model's ability to generate excellent translations that capture the full semantic weight of the source text.

Furthermore, the meaning of the eleventh sentence was successfully captured, expressing the passage of time since last seeing someone. The output, "I have not seen her for a while," is a very natural and fluent English equivalent, demonstrating strong performance on common temporal expressions. In the last example, the system made excellent lexical choices, translating "سهرات" as "night outs" and "تحواس" as "trips." The only minor flaw is the use of the preposition "of" instead of "for," making the final output slightly less fluent than the human reference. This represents a near-success with a minor grammatical slip.

1.2.3. Diagnosis of Intrinsic Complexities

While the preceding sections identified specific instances of MT failure, and the attention visualizations revealed the internal dynamics of word alignment, a deeper diagnostic inquiry is necessary to understand the underlying sources of these errors. The challenges in translating ORN are not merely random artifacts of an imperfect model; rather, they are systemic issues rooted in the intrinsic linguistic and cultural properties of a low-resource, morphologically complex language. This section moves from cataloging errors to explaining their origins, deconstructing the most critical pitfalls that characterize the ORN-ENG NMT task. The examination of these complexities in detail delineates the pathways for future improvement. The following analysis will focus on four key areas of difficulty: the lexical chasm of OOVs, NEs, the hazards of semantic and lexical ambiguity, and the profound challenge of translating idiomatic and culturally-specific meanings.

1.2.3.1. Lexical Chasm of OOVs **and N**Es

The most immediate and pervasive challenge in any low-resource NMT setting is the problem of the unknown word. An NMT model, fundamentally a pattern-matching system, is constrained by the vocabulary of its training data. For the 8.5K-sentence OEPC, the lexical window is inherently narrow. This limitation manifests in two critical ways. First is the general OOV problem, which is significantly exacerbated in ORN by the lack of a standardized orthography. The same word can appear in various forms, potentially being treated as distinct, rare tokens by the tokenizer, thus increasing the likelihood of one being an OOV at inference time. Furthermore, the dialect's rich use of borrowings from French, Spanish, and Berber introduces a vast set of lexemes that are unlikely to be covered comprehensively.

The analysis of the model's handling of OOVs reveals a multi-layered set of challenges that extend beyond simple lexical absence. The failures can be deconstructed into distinct categories that reflect issues of orthographic variation, register mismatch, and fundamental data sparsity. The diagnostic analysis focuses on major instances of sense OOV (due to spelling or ambiguity) and complete OOV (unseen in the training or frequently used in MSA). Although the unstandardized, polysemous nature of ORN may bring uncountable OOV examples, Table 22 deconstructs the main representative OOV failures, with a subsequent interpretation elaborating on the patterns revealed therein.

Table 22

Illustrations of OOVs

OOVs	RT	HT	Type / Explanation
قوشحونا	they beat us	*saw us from afar*	Sense OOV/ Orthography
ستحفد	take care of	*(omitted)*	Sense OOV/ Orthography
شعبوية	populism	*people*	Complete OOV/ Not colloquial
عنيد	stubborn	*(omitted)*	Complete OOV/ Not colloquial
نحاس	brass	*(jinx)*	Complete OOV/ Unseen
مستكحل	close to black	*(omitted)*	Complete OOV/ Unseen

A primary challenge, unique to unstandardized spoken dialects, is the Sense OOV caused by orthography. The word "قوشحونا" exemplifies this perfectly. The informal, phonetic spelling used in the test set was not in the model's vocabulary. In response, the model appears to have defaulted to the orthographically closest known word, producing the translation "saw us from afar." This is not a random hallucination but a systematic failure: the model substitutes an unknown token with a known lookalike, actively introducing a severe semantic error. The case of "ستحفد" is a phonetic script of the formal spelling "استحفظ" resulted in a more straightforward omission.

Together, these examples show that orthographic variance is a dual threat: it can either render a word completely invisible to the model (omission) or cause it to be misinterpreted as another word (substitution).

A second critical category of failure is the complete OOV due to a non-colloquial register. This highlights the specific linguistic domain of the training corpus. The model's failure to translate formal terms like "عنيد" and "شعبوية" is not merely a sign of a small vocabulary, but a reflection of the fact that OEPC is overwhelmingly composed of spoken, informal language. In daily ORN conversation, the concept of "stubborn" is idiomatically-expressed as "راسه خشين" rather than with the aforementioned formal MSA word. The model, therefore, learned the common colloquial form and was never exposed to its formal synonym. Moreover, the translation output "people" is a classic example of a subword tokenization artifact; the model did not recognize the full word, but its tokenizer likely identified the familiar root "شعب" ("people") and translated that fragment, failing to comprehend the abstract suffix.

Finally, the fully-unseen category represents the most foundational challenge of data sparsity. Words like "مستكحل" and "نحاس" were simply absent from the corpus. While the omission of the former is a straightforward case of an unseen adjective, the translation of the latter is another fascinating substitution error. The source word for "jinx" or "bad luck" in Arabic dialects is often "نحس" (*nahs*), which is almost orthographically-identical to "نحاس". The model appears to have defaulted to the most similar-looking token in its memory, again producing a confident but semantically incorrect translation. The analysis of OOVs may paint a detailed picture of the model's boundaries, where failures are caused not just by lexical gaps but by a complex interplay of scripting chaos, register mismatch, and the model's tendency to fall back on superficial

orthographic similarities when its knowledge base is exhausted. The vulnerability to data sparsity, which underlies the OOV errors, manifests in a particularly acute form in the translation of NEs, which constitute a significant portion of the vocabulary's "long tail."

NEs constitute a significant and well-documented challenge for NMT systems, acting as a direct probe of a model's lexical knowledge and its ability to handle OOV phenomena. Table 23 presents a diagnostic summary of the trained model's performance on a representative sample of NEs found in ORN, classified with specific error types. The analysis of these results reveals a clear and predictable pattern: the model's success is almost exclusively determined by an entity's high frequency or global ubiquity within the training data, while its failures provide a precise taxonomy of the pitfalls inherent to low-resource translation.

Table 23

Illustrations of NEs Errors

NEs	Recognized	Not Recognized	Error Type
Person	بختة Bakhta الهواري Houari	بونجاح Bounedjah رامي Rami	Unseen Entity Ambiguity
Place	تليلات Tlelat مانيج ميرامار Miramar Park	عين البيضاء Ain Beida بتي لاك Petit Lac	Ambiguity Orthography
Organization	يوتوب YouTube آيفون iPhone	الفيفا FIFA كوسيدار Cosider	Ambiguity Unseen Entity
Products	بانادول Panadol أومو Omo	أسبيرين Aspirin فينوس Venus	Ambiguity Unseen Entity
Drink / Food	كوكا Coca بيتزا Pizza	حمود Hamoud بايلا Paella	Frequency bias Indecision
Money/Currency	فرنك Franc دورو Doro	جنيه Pound ريال Ryal	Unseen Entity Unseen Entity

The "Recognized" column showcases the model's capabilities under ideal data conditions. Its success with entities like *Bakhta*, *Houari*, *Tlelat*, and *Miramar Park* is attributable to their high frequency as common local names and places, which provided a strong and repeated learning

signal within the OEPC. More revealing is the model's consistent success with globally recognized entities. Brands like *YouTube*, *iPhone*, *Panadol*, *Omo*, and *Coca-Cola* have become part of a global lexicon. Their presence in the training data, even if minimal, is sufficient for the model to learn a stable mapping, proving that it is fully capable of translating NEs when the data provides a clear signal.

The "Not Recognized" column, interpreted through the assigned error types, deconstructs the specific reasons for failure with greater clarity. First, the most direct failure is the absence of entities in data training, a quintessential symptom of data sparsity. Entities like the footballer *Bounedjah*, the local company *Cosider*, the product *Venus*, and the currencies *Pound* and *Ryal* were simply absent or too rare in the 8.5K sentence corpus. Lacking any prior exposure, the model has no representational basis upon which to perform a translation. Moreover, the second error type -ambiguity- manifests in several forms. For the personal name *Rami*, the model likely confused it with a more common, local drink name and failed to disambiguate it as a proper person name. For place names like *Ain Beida*, the failure is a classic case of compositional ambiguity; the model translated the component parts ("eye" and "white") literally instead of recognizing the string as a single, indivisible entity. Similarly, for global entities like *FIFA* and *Aspirin*, the model failed to resolve the transliteration ambiguity.

Third, orthography and indecision are related errors stem from the dialect's frequent code-switching, unconventional spelling, and/or loanwords. The French place name *Petit Lac* introduces an orthographic challenge (بتيلاك / بتلاك / بتي لاك / بيتي لاك), as its non-Arabic script representation within ORN sentences creates confusion. The model's failure indicates an inability to establish a consistent rule for handling such loanwords as illustrated by the indecision over the

Spanish food name *Paella*. Faced with a new form (بايلا), the model is uncertain whether to transliterate it (*paila / payla / pila / peyla*, *etc.*), translate a potential meaning, or pass it through, resulting in an incorrect or omitted output.

Finally, the bias error highlights how a corpus may not reflect real-world local prevalence. While *Coca-Cola* (global) was recognized, *Hamoud* (a very popular local Algerian brand) was not. Rather, *Hamoud Boualem* is the full known name for the drink brand. This suggests that within the broad domains of the OEPC, the brand appeared, biasing the model's learning process away from the local entity. NEs, in synthesis, are part of the "long tail" of ORN lexical distribution and are frequently absent from the training corpus. Thus, the model's performance on NEs serves as a microcosm of the broader challenges of this NMT task. It confirms that successes are contingent on data exposure, while failures are not random but are systematic consequences of data sparsity, linguistic ambiguity, and the complex, multilingual reality of ORN.

1.2.3.2. Lexical and Structural Ambiguity

Beyond the straightforward challenge of unknown words, the model's most sophisticated failures occur in the domain of lexical and semantic ambiguity. These errors are particularly insightful because they demonstrate the limits of a purely statistical, frequency-based understanding of language. The model may recognize a word or phrase, but it lacks the contextual reasoning or world knowledge to correctly disambiguate its meaning. Table 24 demonstrates the major types of ambiguities found in the SL that caused inaccuracies in the model’s outputs.

Despite the contribution of the attention mechanism, some consistent patterns of high-frequency default are revealed. That is, the model systematically selects the most common, literal,

or compositional meaning of an ORN word or phrase, ignoring less frequent but contextually correct interpretations. This is evident in multiple forms of polysemy. For "بخير عليا" and "يسمى", the model translates the most common senses of "fine" and "called," failing to recognize their secondary, more nuanced meanings of "wealthy" and the discourse marker "so." Similarly, with "وحد" translated literally as "one," missing its functional role as an approximator ("around" or "quite"). This bias is a direct artifact of the training data, where the literal senses of these words were likely far more frequent, leading the model to learn a strong but shallow correlation.

This literalist tendency extends to the misinterpretation of non-compositional and figurative language. For the expressions "هدري على روحك" and "الحمار تاع خوك", the model performs a word-for-word compositional translation, resulting in the nonsensical outputs "talk about yourself" and "the dumb of your brother." It is fundamentally unable to recognize that the meaning of these phrases is holistic and non-literal. The fact that "dumb" was accurate reflects the attention interference, yet the possessive structure in ORN was not processed as descriptive, showing the lack of ability to process figurative meaning, a key feature that distinguishes human language comprehension from statistical pattern matching.

Finally, the system is vulnerable to both homonymy and syntactic ambiguity. For instance, the model incorrectly interprets "العوام" as "years" instead of "the swimmer". This reflects the phonetic and diactric significance in making meanings in ORN. Contextual disambiguation is, thus, realized by opting for the contextually plausible word even if it is less frequent. The translation of "هاجر خطبها واحد" is a more subtle failure of syntactic mapping. The model correctly identifies the passive voice but fails to generate the crucial agentive preposition "by," slightly altering the meaning and demonstrating its fragility in replicating complex grammatical structures

perfectly. Collectively, these examples demonstrate that the model's primary weakness lies in its inability to move beyond statistical frequency and compositional logic, leaving it brittle when faced with the polysemy, idiomaticity, and syntactic complexity of a rich human dialect.

Table 24

Types and Illustrations of Ambiguities

ORN	RT	HT	Type
بخير عليا	I am wealthy	*I am fine*	Polysemy / High-Frequency Default
يسمى	so / this means	*called*	Polysemy (Pragmatic vs. Lexical)
البحر ياكل غي العوام	the sea swallows only the swimmer	*the sea eats only years*	Contextual Disambiguation Failure
هاجر خطبها واحد	Hadjer got proposed to by someone	*Hadjer got proposed to someone*	Syntactic Ambiguity (Passive Voice Mapping)
هدري على روحك	talk as much as you can	*talk about yourself*	Idiomatic Ambiguity (Non-Compositional Phrase)
وحد العشر دقايق	quite ten minutes	*one ten minutes*	Functional Ambiguity (Numeral vs. Approximator)
الحمار تاع خوك	your dumb brother	*the dumb of your brother*	Metaphorical Ambiguity (Figurative Language)

When confronted with a lexical item whose function is not lexical but pragmatic, its translation is determined almost entirely by discourse context. The Oranian discursive, multifunctional particle '*yadra'* serves as an archetypal example of this phenomenon and therefore warrants a dedicated micro-analysis."يادرا" (*yadra*) is indeed ORN's equivalent of "يا ترى" (*ya tura*) in MSA, and it carries very similar nuances. The core meanings still revolve around wonder, speculation, curiosity, or a rhetorical question. As it is a highly contextual expression, the best translation will always depend on the specific sentence and the speaker's tone and intention. The common potential translations of "يادرا" (*yadra*) in ORN are listed below with inconsistent HTs:

When expressing wonder or curiosity:

- **I wonder if...** (The most common and versatile)

- RT: "يادرا يجي اليوم؟" (*yadra yji l'youm?*) -> "I wonder if he'll come today?"
- HT: *will he come today?*

- **Do you think...?** (When seeking an opinion or guess from someone)
  - RT: "يادرا ينجح ؟" (*yadra ynejah?*) -> "Do you think he'll pass?"
  - HT: *does he succeed?*
- **What if...?** (When contemplating a hypothetical scenario)
  - RT: "يادرا شا يصرا كون ما جيناش؟" (*yadra sha ysra koun ma jinash?*) -> "What if we hadn't come?"
  - HT: *so what happens if we did not come?*

When expressing doubt or uncertainty:

- **I wonder...** (Often used with a slight pause or a thoughtful tone)
  - RT: "يادرا صح هاد الهدرة؟" (*yadra sah hadra?*) -> "I wonder if this talk is true"
  - HT: *how true is this talk?*
- **Maybe... / Perhaps...** (When speculating about a possibility without certainty)
  - RT: "يادرا كان مريض" (*yadra kan mrid*) -> "Maybe he was sick"
  - HT: *you think he was sick*

When posing a rhetorical question or expressing mild exasperation/helplessness:

- **I wonder...** (Still fits, conveying a sense of "I have no idea" or "It's anyone's guess")
  - RT: "يادرا وين راهو رايح؟" (*yadra win raho rayeh?*) -> "I wonder where he's going?"
  - HT: *so where is he going?*
- **What on earth...? / What in the world...?** (More emphatic for rhetorical questions, often with a hint of surprise or frustration)
  - RT: "يادرا كي صرالهم؟" (*yadra ki sralhom?*) -> "What on earth happened to them?"

- HT: *so what happened to them?*

The model's grasp of lexical meaning is further tested by words whose primary function is not to denote a concept but to convey an emotional or pragmatic state. A paradigmatic case of this is the Oranian word 'حبوبة' (hbouba). Etymologically, this term can be traced to skin ailments like pimples or boils, a meaning connected to the common Algerian Arabic dysphemism 'مرض' ('sickness') for something frustrating. However, in contemporary Oranian speech, its literal meaning is almost entirely bleached; instead, 'حبوبة' functions as a dysphemistic placeholder, used to replace a noun or a set of events when the speaker is in a state of anger or frustration. Much like the English word "shit" can be used to refer to any object of contempt (e.g., "get this shit off my desk"), an Oranian speaker might refer to a misplaced, broken, or annoying object as 'hbouba'. Its translation is therefore entirely context-dependent and emotive, a nuance impossible for the model to capture from text alone.

This challenge is magnified by its deep cultural and cross-dialectal parallels. The use of a seemingly random noun as a placeholder for an object of frustration is a common feature in other Arabic dialects, such as 'نيلة' (indigo dye) in EGY or 'بطيخ' (watermelon) in LEV. For a model to correctly translate 'hbouba', it would need to understand not just a word but a pragmatic category of expression. The complexity is further compounded by the use of intensifiers. Adding the adjective 'كحلا' (black) to form 'حبوبة كحلا' does not describe a "black pimple" but escalates the emotional intensity, invoking a sense of pessimism, trouble or impending catastrophe. This non-compositional, culturally-grounded intensification is entirely opaque to the NMT system. The latter failed due to the complexity of processing pragmatic meanings used to express human emotions. This challenge is not an isolated phenomenon; it extends systematically from the

individual word to the phrasal level, leading into the even more complex and formidable domain of idiomatic and culturally-specific expressions.

1.2.3.3. Idiomaticity and Cultural Meanings

The most profound and categorical failures of the NMT system occur at the intersection of language and culture. Idiomatic expressions, which are non-compositional by nature, represent a cliff-edge for translation models. The meaning of an idiom cannot be derived from the sum of its constituent words, and a system trained on literal mappings will almost invariably fail. The catastrophic mistranslation of the phrase "قالب عليا لرض" (*qaleb 'aliya l'ard*), literally, "he must be rolling over on the ground" (see Table 21), into a nonsensical statement about "ground" and "rolling" is a quintessential example. The true meaning, "he is searching for me everywhere," is a holistic semantic unit that has no connection to its literal components, a concept the model is architecturally and data-wise ill-equipped to grasp.

This failure mode is not an isolated occurrence but a systemic weakness, evident across a wide spectrum of idiomatic constructions that range from the metaphorical to the grammaticalized. Consider the phrase "طاح عليهم", whose literal rendering, "fell on them," is comprehensible but entirely misses the idiomatic meaning of "he surprised them" or "he caught them red-handed." The model, lacking the 'world knowledge' that this common phrasal verb has a non-literal sense, defaults to the physical action, thereby losing the entire communicative intent. The challenge intensifies with metaphorical idioms grounded in imagery.

The expression "النو راهي خيط مالسما" literally translates to "the rain is a thread from the sky." A human listener immediately interprets this poetic image as "it is raining heavily" or "it's

pouring." The NMT system, however, is incapable of this imaginative leap; it reproduces the literal components, creating a translation that is quaintly poetic but functionally useless as a weather report. Even seemingly simple phrases can be deeply idiomatic. Moreover, "ما عندهاش بزاف" literally means "it does not have much." Its actual meaning, however, is temporal: "it's recent" or "not long ago." This is a highly abstract idiom where the semantic meaning has become fully divorced from the constituent words, and the model's inevitable literal translation ("it does not have a lot") completely misses the intended temporal context.

This inability to process non-literal meaning is part of a larger failure to process any information that requires extra-linguistic cultural context. Indeed, references to specific local foods, social customs, or historical events have no direct one-to-one equivalents. Translating such terms requires not just linguistic conversion but cultural explanation, a task far beyond the capability of our sentence-level NMT model. These failures are not merely inaccuracies; they represent a fundamental breakdown in the transmission of communicative intent and highlight the vast gulf that remains between statistical text generation and true human-level language understanding. They underscore that without a massive repository of contextually rich, culturally-grounded data, any NMT system will remain, in a sense, a permanent outsider to the language it attempts to translate.

The inventory of such idiomatic expressions in ORN is effectively uncountable, and many share a conceptual basis with idioms found in MSA and other dialects. However, the primary barrier for the NMT system arises when these concepts are realized using lexical items unique to ORN. When an idiom is built from words that are themselves culturally specific, the layers of non-literal meaning become deeply intertwined with local context. This renders the phrase entirely

opaque to a model that lacks this specific dialectal and cultural grounding, representing the ultimate challenge for MT of a living, evolving language.

2. Discussion

The primary objective of this research was to establish a foundational baseline for NMT from the Oranian dialect of Algerian Arabic to English. The culmination of this effort was through a Seq2Seq model incorporating LSTM and attention mechanisms, trained on the bespoke 8.5K-sentence parallel dataset and achieved a BLEU score of 30.9. This result is not merely a quantitative benchmark but a rich data point that encapsulates the opportunities and limitations inherent in translating a low-resource, morphologically complex, and orthographically unstandardized dialect. To fully interpret this outcome, it is essential to situate it within the broader landscape of contemporary NMT research, particularly concerning other Arabic dialects and LRLs.

This discussion will systematically deconstruct the performance of the ORN-ENG NMT system. It will begin by comparing the findings to the most contextually relevant studies on Algerian and Maghrebi dialects to assess regional viability. Following this, the analysis will pivot to an architectural and parametric comparison with a study on EGY-MSA, which employed a similar model. The discourse will then broaden to evaluate the strategic implications of our direct-corpus approach against state-of-the-art methodologies like pivot translation used in TUN-ENG MT. Finally, the specific challenges encountered in this research will be contextualized within the universal problems faced in LRL translation, drawing parallels with work on languages such as Bengali, Indonesian, and Urdu. Through this comparative lens, the score of the present study’s

experiment will emerge, highlighting both the significant contribution of this study and the clearly demarcated pathways for future research.

The present section focuses on the parametric and contextual features of relevant experimental studies. To frame the subsequent discussion, Table 25 synthesizes the key metrics and characteristics of the NMT systems from the most pertinent research, juxtaposed with the present study's achievements.

Table 25

Comparative Analysis of the Proposed Model's Performance Across Relevant Studies

Work	Language Pair	Dataset Size	Model Architecture	BLEU Score(s)	Relevance Factor
This Study (2026)	ORN-ENG	OEPC: 8.5K	Seq2Seq RNN (LSTM) + Attention	30.9	Primary Subject, LRL, Algerian.
Harrat (2018) based on Meftouh et al. (2015)	MAG (ALG, ANB, TUN, MOR) & LEV (SYR, PAL)	PADIC: ~40K words per dialect (~2-3K sentences)	PBSMT	15.1 (ALG→MSA), 25.99 (TUN→MSA)	Algerian Dialect, Foundational Multi-Dialect Corpus
Lichouri and Abbas (2021)	ALG, TUN, MOR, SYR, PAL, KAB.	PADIC v2.0 (~6.4K sentences)	SMT & RBMT	57.56 (ALG→MSA - SMT), 69.56 (ALG→MSA - RBMT)	Algerian, Multi-Dialect Corpus Extension
Faheem et al. (2024)	EGY-ENG	40K + 20M monolingual	Seq2Seq RNN (LSTM) + Attention / Transformer	Semi-supervised: 25 (RNN, LSTM, Attention), 29.5 (Transformer)	Similar Model & Architecture, LRL.
Kchaou et al. (2025)	TUN-(MSA)-ENG	188K	Transformer (Fairseq, JoeyNMT)	41.78 (Direct), 67.89 (Pivot)	LRL, Arabic Dialect-English

Abujar et al. (2021)	ENG-Bengali	1.8K	Seq2Seq RNN (LSTM) + Attention	22.3	LRL, Similar Model.
Puspitaningrum (2021)	Indonesian-ENG	~2.5K	Seq2Seq RNN (LSTM), Convolutional Seq2Seq, MHA	42.9 (RNN, lower than Convolutional Seq2Seq)	LRL, Comparison of Architectures
Israr et al. (2023)	Urdu-ENG	90K	Bi-RNN, LSTM, GRU, Transformer	47.14 (LSTM)	LRL, Similar Architecture & Hyperparameters

The most immediate and relevant points of comparison for this study are the foundational works that have sought to build resources for Algerian and Maghrebi dialects, notably the PADIC corpus and its subsequent extensions. The first PADIC version, followed with its extension (Meftouh et al., 2015; 2018) introduced a multidialectal parallel corpus featuring two Algerian dialects (ALG and ANB) alongside TUN, MOR, SYR, and PAL variants. The extension documented by Lichouri and Abbas (2021) further enriched this corpus with the Berber dialect (KAB). BLEU scores of studies on PADIC are more likely affected by the model and approach. These initiatives represent monumental efforts in resource creation, providing a broad overview of the Arabic dialectal landscape. However, their very breadth introduces a level of heterogeneity that can inadvertently dilute a model's ability to learn the fine-grained specifics of a single, localized dialect.

Herein lies the primary merit of this study’s approach. By constructing OEPC, linguistic depth and consistency were prioritized over dialectal breadth. While PADIC includes sentences from ALG and ANB, its lexical and phonological features differ from those of ORN. The latter, with its distinct prosody and deeper substratal influences from Spanish and French, possesses a

unique linguistic identity. A model trained on a pan-Algerian or pan-Maghrebi corpus must learn to generalize across these variations, which can lead to a "mean-average" translation that fails to capture the idiosyncratic flavor of any single dialect. The 30.9 BLEU score achieved in this study is therefore highly significant; it demonstrates that a focused, albeit smaller, corpus can empower an NMT model to achieve a commendable level of competency for a specific linguistic variety. This finding strongly advocates for the "depth-over-breadth" strategy in the initial stages of resource development for under-resourced dialects. It proves that creating a clean, consistent, single-dialect corpus is a viable and effective pathway to building a robust baseline translation system, sidestepping the "dialectal noise" that can confound models trained on more heterogeneous datasets.

However, this focus is not without its demerits. The smaller scale of the OEPC (8.5K sentences) inevitably restricts the lexical and syntactic diversity available to the model. The PADIC v2.0 corpus, at ~6.4K sentences per dialect pair, is comparable in size but is part of a larger, interconnected resource pool. The broader resource ecosystem allows for cross-dialectal learning and transfer learning experiments, an avenue not immediately available to our isolated OEPC. The qualitative analysis in our study revealed that while the model excels at translating structurally common and simple sentences, it falters significantly with rare vocabulary and complex syntactic constructions. This limitation is a direct consequence of the dataset's size and underscores a critical trade-off: the focused approach yields higher fidelity for the core language but sacrifices the lexical robustness that a larger, more diverse (even if noisier) dataset might provide. The 30.9 BLEU score should thus be seen as a high floor for common ORN usage patterns, but one that does not fully capture the model's brittleness at the lexical fringes.

To assess the architectural efficacy of the present study's model, the work of Faheem et al. (2024) on EGY-MSA translation provides an almost ideal parametric parallel. Their study also employed a Seq2Seq RNN with LSTM and attention, allowing for a direct comparison of performance while controlling for architectural variables. Faheem et al. (2024) report a BLEU score of 25 for RNN model trained on a 20M monolingual corpus (augmented) in a semi-supervised framework (Scored 29.5 with Transformer settings). Crucially, this reveals that the proposed model of the present study not only exceeds their RNN result but also outperforms their best Transformer model (29.5 BLEU), an achievement made with a purely supervised methodology on a parallel corpus less than a quarter of the size, which strongly suggests a high degree of linguistic consistency and learnability within the curated OEPC.

Several critical insights emerge from this comparison. Firstly, their substantially larger parallel dataset (40K vs. 8.5K) resulted in a lower BLEU score (22 vs. 30.9 in a supervised framework). This counter-intuitive result could be attributed to several factors: differences in the complexity of the test sets, the inherent linguistic distance between SLs and TLs, or the quality and consistency of the corpus creation process. A key merit of this study, therefore, is the demonstration of highly efficient learning with low-noise resource. This reinforces the value of meticulous data collection and cleaning, a process that can be more impactful than data volume alone, especially in low-resource settings.

Secondly, Faheem et al. (2024) powerfully illustrate the limitations of a purely supervised approach, a category into which the present study falls. Their ability to boost performance from a BLEU score of 22 to 25 by incorporating monolingual data highlights the single most significant demerit of this study's approach. Understanding ORN in the current study is confined entirely to

the 8.5K sentences of the OEPC. It has no exposure to the vast expanse of monolingual ORN texts from other sources. This severely curtails its ability to learn a robust language model for the source side, leading directly to the observed failures in handling OOV words and idiomatic expressions. The semi-supervised approach directly learns better word representations and contextual understanding from unlabeled data. Therefore, while ORN-ENG model establishes an exceptional supervised baseline, it is almost certainly a performance sub-ceiling. This comparison makes it clear that the most promising and immediate path to improving the model is to adopt a semi-supervised learning strategy, leveraging monolingual ORN data to enhance the encoder's language understanding.

The recent work by Kchaou et al. (2025) on TUN-ENG translation offers a strategic and state-of-the-art benchmark. Their study is significant for two reasons: its use of a Transformer architecture and its novel approach to data acquisition via pivot translation. Architecturally, their work aligns with the broader consensus in the NMT community that Transformers have largely superseded RNN-based models. Their final BLEU score of 67.89 is a powerful testament to the superior ability of the self-attention mechanism to capture long-range dependencies and complex contextual relationships within a sentence.

The model of the present study, based on a sequential LSTM architecture, is inherently more limited in this regard. While the attention mechanism allows it to focus on relevant parts of the SL sentence, it lacks the parallel processing and global contextual view that defines the Transformer. The merit of this choice is its relative computational efficiency and its proven effectiveness in establishing a strong baseline. The 30.9 BLEU score is an efficient achievement for an RNN. The demerit, however, is clear: the architectural ceiling of an RNN is significantly

lower than that of a Transformer. This comparison solidifies the conclusion that transitioning to a Transformer-based model is an essential next step to push beyond the current performance plateau; however, the dataset size should be considerable.

A strategic comparison with the results of Kchaou et al. (2025) can be more profound as they faced the same fundamental problem in LRL context: a scarcity of direct TUN-ENG parallel data. Instead of undertaking the laborious task of manual corpus creation, they employed a highly effective pivot strategy. By first translating a large, 188K-sentence TUN-MSA corpus into English using a pre-trained MSA-ENG model, they were able to semi-automatically generate a massive trilingual dataset. Their astonishing BLEU score of 67.89 was achieved by virtue of using this pivoted data.

This presents a stark contrast to the methodology in the present research work. The primary merit of creating the OEPC through direct, careful curation is quality. OEPC represents a "gold standard" of direct ORN-ENG translations, free from the cascading errors and semantic drift that can be introduced by an intermediate pivot language. The 30.9 BLEU score is a reflection of a model trained on authentic, direct mappings. The demerit of our approach is its profound lack of scalability. Creating the 8.5K sentence corpus was a time-intensive and resource-heavy process. Scaling it to over 100K sentences would be an almost insurmountable task. The pivot methodology by Kchaou et al. (2025), while potentially introducing noise and translationese, offers a pragmatic and scalable solution to the data bottleneck problem.

The challenges and outcomes observed in this study are not isolated phenomena but are emblematic of the wider field of low-resource NMT. Parallels can be drawn with research on languages as diverse as Bengali, Indonesian, and Urdu, which reveals a set of common, recurring

themes. The study by Abujar et al. (2021) on ENG-Bengali translation, for instance, used a very similar architecture (Seq2Seq with attention) on an even smaller dataset of 1.8K sentences, achieving a BLEU score of 22.3. This suggests a reasonable correlation between dataset size and performance for this class of model. Moreover, they identified data scarcity as the principal impediment to achieving higher accuracy, reaffirming that the core challenge in LRL translation is often data engineering rather than model innovation alone.

Similarly, Puspitaningrum (2021), in a comparative study on Indonesian-ENG translation, explored multiple architectures, including RNN, Convolutional Seq2Seq, and MHA. Their findings showed Convolutional Seq2Seq outperforming other models on their dataset, underscoring that there is no one-size-fits-all architecture; performance is often a function of the interplay between the model's inductive bias and the specific linguistic characteristics of the language pair. Our choice of an LSTM-based RNN was a well-justified starting point, but the findings of Puspitaningrum (2021) confirm that architectural exploration is a critical part of optimizing performance for any given LRL.

Finally, the extensive work by Israr et al. (2023) on Urdu-ENG NMT offers a powerful point of comparison. Using a much larger 90K parallel corpus, their LSTM-based model achieved a high BLEU score of 47.14. In a recent experiment, Israr et al. (2024), enhanced the BLEU score on CNN and GRU models with over 2 points. This result clearly demonstrates the performance that can be unlocked with a larger dataset, even without resorting to a Transformer architecture. Crucially, they highlight issues of perplexity and failed word alignment, and they proposed an attention-based dropout layer to refine the attention mechanism and improve generalization on unseen data. The qualitative analysis of ORN-ENG model revealed similar issues with fuzzy

alignments for complex phrases. The techniques proposed by Israr et al. (2023), such as specialized regularization within the attention layer, could be directly applicable to the model if it was trained on larger data for enhancement that is complementary to simply switching architectures.

In synthesis, the 30.9 BLEU score achieved in the ORN-ENG NMT experiment represents a significant and successful first step in the field of Arabic LRL MT. The discussion reveals that this performance is highly promising and competitive, especially considering the constraints of the dataset and the chosen architecture. The construction of the OEPC stands as a primary contribution, demonstrating that a meticulously curated, single-dialect corpus provides a high-quality learning signal, enabling our model to outperform larger, supervised systems for other Algerian or Arabic dialects.

3. Summary

In conclusion, this chapter has provided a multi-faceted analysis of the NMT system's performance, painting a dual picture of significant accomplishment and profound, systemic limitations. The investigation began with the quantitative results, establishing a promising BLEU score of 30.9, while a subsequent visual analysis of the attention mechanism confirmed its critical role in enabling the model to handle complex syntactic and morphological alignments. However, the core of the chapter pivoted to a deep qualitative diagnosis of the model's failure points. This chapter has thoroughly contextualized the quantitative success of the model within a qualitative framework that clarifies its limitations, setting the stage for the conclusion, which will provide implications and avenues for future research.

CONCLUSION

This dissertation embarked upon a research journey motivated by a critical and conspicuous lacuna in the field of computational linguistics: the complete absence of resources and MT systems for ORN. Positioned at the intersection of LRL processing and Arabic dialectology, this study was premised on the fundamental challenge that without a parallel corpus, no empirical investigation into NMT for the ORN-ENG language pair could proceed. The overarching ambition was therefore twofold: first, to address this "data-zero" problem through a constructive methodological paradigm by creating the first-ever OEPC; and second, to leverage this novel resource within an empirical paradigm to design, train, and rigorously evaluate the first NMT system for this dialect, thereby establishing a crucial first performance benchmark.

The present part of the study serves as a final synthesis of this entire endeavor. It moves beyond the granular presentation of results to offer a holistic interpretation of their collective meaning and significance. It begins by systematically recapitulating the principal findings of the investigation, providing a direct response to the research question that guided the dissertation. Subsequently, the conclusion elaborates on the broader implications of these findings, considering their contributions to theoretical, methodological, and practical domains within NLP. Moreover, a transparent and critical discussion of the study's inherent limitations is provided. Finally, these limitations are reframed as opportunities, forming the basis for a comprehensive set of recommendations for future research. This conclusion thus aims not only to summarize what has been accomplished but also to delineate the scholarly contribution of this dissertation to the field and to chart a clear path forward for the continued advancement of MT for Arabic dialects and other LRLs.

1. Synthesis of Findings

The research presented in this dissertation unfolded across a deliberate, two-phase trajectory designed to logically address the zero-resource status of the Oranian dialect. The initial, and most foundational, phase of this research was constructive. It confronted the primary barrier to entry by undertaking the meticulous process of corpus creation. This involved sourcing authentic linguistic data from a diverse array of contemporary media, followed by a rigorous pipeline of transcription, normalization, and professional human translation. The tangible result of this phase was OEPC, a novel dataset comprising 8,5K curated sentence pairs. This artifact is not merely a prerequisite for this study but stands as one of its primary, reusable contributions to the scholarly community.

The second phase was empirical. It focused on the design, implementation, and evaluation of an NMT system. Guided by established practices in LRL contexts, a Seq2Seq architecture was selected, built upon a two-layer stacked LSTM network for both the encoder and decoder. Recognizing the informational bottleneck of a basic Seq2Seq framework, this architecture was crucially augmented with an additive attention mechanism. This design choice was made to better handle the complexities of linguistic divergence between Oranian, a morphologically rich Arabic dialect, and English. The model was trained, validated, and tested following a methodologically sound pipeline to ensure the robustness and generalizability of its performance.

The evaluation of the trained NMT model yielded a rich set of quantitative and qualitative data, providing a multi-faceted view of its performance. The model’s performance, benchmarked against an attention-less baseline, demonstrated a significant leap in efficacy. The final, optimized model achieved a BLEU score of 30.9 on the held-out test set, a substantial improvement of over

12 points from the baseline score of 18.3. This result serves as the headline figure, establishing a strong and credible first benchmark for the ORN-ENG translation task. Further metrics corroborated this narrative of improvement. WER indicated a marked decrease in the literal errors requiring post-editing. The METEOR score confirmed that the model's lexical choices were not only more accurate but also more contextually and semantically appropriate. Perhaps most diagnostically, the model's PPL plummeted to signify a vast improvement in the model's confidence and the fluency of its internal language model for English.

The manual inspection of MT outputs provided crucial insights to the quantitative scores. The analysis revealed that the model demonstrates a high degree of competency in handling the core grammatical and syntactic structures of ORN. It successfully translates sentences with standard subject-verb-object constructions, manages common lexical items accurately, and preserves the general semantic intent of high-frequency expressions. However, the qualitative analysis also illuminated the precise nature of the model’s failures, which align closely with the well-documented challenges of low-resource NMT.

At the lexical level, performance was most directly constrained by the finite scope of the 8.5K sentence corpus, leading to predictable failures on OOVs and NEs. This issue of lexical sparsity was compounded by challenges with semantic ambiguity, where the model frequently struggled to leverage broader sentence context to correctly disambiguate ORN terms with multiple potential ENG equivalents. Beyond individual words, the system encountered difficulties with structural complexities. While the BPE tokenization and LSTM architecture could process some inflectional forms, the highly agglutinative and polysynthetic nature of ORN morphology and verbs occasionally resulted in misinterpreted conjugations or the omission of critical clitics. The

model’s most profound fragility, however, was exhibited when confronted with idiomatic and culturally-specific expressions; faced with non-compositional language, it consistently defaulted to nonsensical literal translations, betraying a lack of the deep semantic understanding that is a direct consequence of data scarcity. Apparently, these distinct error categories are not isolated deficiencies but are interconnected symptoms of this foundational data constraint, which collectively prevented the model from acquiring a truly comprehensive representation of the SL.

Based on the synthesized findings, a direct answer to the research question can be formulated. The empirical evidence demonstrates that an NMT model, under these conditions, can translate the Oranian dialect to a significant and foundational extent. The achievement of a BLEU score of 30.9 is not trivial; it represents a system that has successfully learned the primary linguistic mappings between the two languages and can produce useful, and often accurate, translations for a considerable portion of common-use dialectal sentences. It moves the dialect from a state of complete technological obscurity to one with a tangible, functional, and quantitatively benchmarked translation capability. The system's effectiveness is most pronounced in non-idiomatic, structurally straightforward contexts, which constitute a large part of everyday communication.

However, the extent of its effectiveness is bounded by predictable and well-defined limitations inherent to its low-resource training paradigm. The model's accuracy degrades substantially when confronted with lexical rarity, idiomaticity, and deep semantic ambiguity. Therefore, while it is *effective* in establishing a robust baseline and handling the core of the dialect, its *accuracy* is not yet sufficient for high-stakes or efficient applications without human intervention. The study thus concludes that the approach is highly successful as a proof-of-concept

and a foundational step, proving the viability of the task, but simultaneously underscores the significant journey that remains toward achieving fully automatic high-quality translation for the Oranian dialect.

2. Implications

The outcomes of this dissertation extend beyond the immediate context of Oranian, carrying significant implications for several domains of computational linguistics. These contributions are at once theoretical, methodological, and practical, offering valuable insights for scholars and practitioners in the field.

First, in the realm of low-resource dialectology, a persistent debate revolves around resource creation strategies. One approach favors breadth, creating multi-dialect corpora (like PADIC) to capture wider linguistic variation. The alternative, "depth-over-breadth," advocates for focusing on a single, clean, and linguistically consistent dialectal corpus. The high competitiveness of the 30.9 BLEU score, achieved on the bespoke 8.5K OEPC, provides strong empirical validation for the latter. It implies that for bootstrapping an MT system for a truly zero-resource dialect, prioritizing the depth and quality of -an even small- single-dialect corpus is a highly effective, and perhaps superior, initial strategy. It avoids the "dialectal noise" that can confound models, allowing them to achieve a more profound understanding of a specific variety's unique patterns.

Second, for many of the world's thousands of LRLs, the path to developing NLP tools is unclear. This dissertation provides a comprehensive, end-to-end methodological blueprint that is both transparent and replicable, at least for Arabic dialects. It details every stage, from the principles of data sourcing to the specifics of model architecture selection, training, and multi-

faceted evaluation. This serves as a practical guide for researchers wishing to tackle other underserved languages or dialects, demonstrating a clear and tested pathway from a state of zero resources to a functional and benchmarked NMT system.

Third, in an era overwhelmingly dominated by Transformer models, this research provides a crucial data point on the continued viability and utility of LSTM-based RNNs. While acknowledging the superior performance ceiling of Transformers, the achievement of a strong 30.9 BLEU score underscores that a well-implemented RNN with attention remains a powerful and computationally less intensive option for establishing a robust baseline, particularly when data is scarce. This has important implications for researchers in resource-constrained environments, suggesting that a successful NMT project does not exclusively hinge on access to the massive computational power required by large Transformer models.

Fourth, the most direct and lasting contribution of this work is OEPC. As a publicly documented and structured resource, OEPC will catalyze future research far beyond the scope of this dissertation. It can now be used for a wide range of scholarly inquiries, including contrastive linguistics, computational dialectology, development of other NLP tools (such as NER or PoS tagging), and as a high-quality development and test set for more advanced NMT systems. In addition, by creating the first MT system for ORN, this research represents a tangible step toward the technological empowerment of its speech community. It elevates the digital presence of the dialect, validating it as a language suitable for modern technology. While the current system is a baseline, it lays the groundwork for future tools that can facilitate communication, enable access to global information in the native dialect, and foster the creation of localized digital content, thereby helping to bridge the digital divide at the linguistic level.

Fifth. prior to this work, ORN-ENG MT was not a task that could be empirically studied. By creating the first dataset and establishing the first benchmark, the present research effectively formalizes a new, measurable subfield of research. All future work on Algerian or Arabic dialects, whether it involves different architectures, larger datasets, or novel pre-processing techniques, can now be quantitatively compared against the results presented herein. This act of formalization is a critical step in transforming a linguistic challenge into a sustainable and progressive scientific endeavor.

3. Limitations

A crucial component of any rigorous scholarly work is a transparent acknowledgment of its limitations. These boundaries, dictated by the scope, resources, and specific choices of the research design, are not weaknesses but rather essential contexts for interpreting the findings and for guiding future inquiry. The limitations of this study can be broadly categorized into two areas: those related to the corpus and those related to the model.

While meticulously curated, OEPC at 8.5K sentence pairs is, by the standards of data-hungry NMT models, a small dataset. This limitation is the root cause of many of the model’s qualitative failures, particularly its poor handling of OOVs and rare lexical items. The lexical diversity of the corpus is inherently finite, and it cannot possibly encompass the full vocabulary of the living ORN. This bottleneck on vocabulary is the single most significant constraint on the current model's performance ceiling. In addition, OEPC was constructed from a range of contemporary sources, including online videos, social media, and transcribed conversations. While this captures a significant slice of informal, colloquial ORN, it also means the corpus has a specific domain profile. The NMT model trained on this data is unlikely to generalize well to other genres,

such as traditional clothing, songs or folklore-related entities, should they exist in ORN. Its expertise is therefore narrow to common interactional practices.

Moreover, OEPC is a collection of sentence pairs, curated and translated in isolation. The model, therefore, learns to translate at the sentence level only. It has no access to broader document-level context, which is often crucial for resolving anaphora (e.g., pronoun references) and ensuring discourse cohesion. This limits its ability to produce translations that are coherent across multiple sentences. In this respect, language is dynamic and constantly evolving; yet, OEPC represents a snapshot of ORN at a particular point in time. It does not capture neologisms, evolving slang, or gradual shifts in usage. A static resource will always lag behind the living language, a fundamental challenge for any corpus-based linguistic research.

As for the model-related limitations, the major challenge was the purely supervised-learning paradigm. This means that the model's knowledge of ORN is derived exclusively from the 8.5K sentence pairs in the OEPC. It does not leverage the vast amounts of monolingual (Oranian-only) data that could potentially be scraped from the web, nor does it employ transfer learning from high-resource languages or pre-trained multilingual models. As powerfully demonstrated by cognate research in other Arabic dialects, semi-supervised techniques like back-translation that utilize monolingual data are proven methods for significantly boosting performance. The exclusion of such methods is a major limitation on the model's potential.

Furthermore, while the LSTM-based model proved effective for establishing a baseline, it is architecturally sub-optimal compared to the current state-of-the-art. Nonetheless, such a limitation is more apparent with small datasets. Therefore, the sequential nature of RNNs makes them less efficient at processing long sequences and, more importantly, less adept at modeling the

complex, long-range contextual relationships that the parallel self-attention mechanisms of Transformer architectures excel at. The choice of an LSTM model therefore places an inherent, and relatively low, ceiling on the maximum performance achievable with this dataset.

Another model-related limitation is the tokenization process. The study employed a standard subword tokenization pipeline. However, given the rich and non-concatenative morphology of the source dialect, more sophisticated, morphologically-aware segmentation or tokenization strategies could potentially yield better results by providing a more linguistically meaningful input representation to the model. The current approach may be sub-optimal for deconstructing complex Oranian verb forms. In this vein, a morphological analyzer for the dialect requires efficient trainings on lemmatizing and tokenizing its complex structures.

4. Avenues for Future Research

The limitations detailed above organically give rise to a rich and well-defined roadmap for future research. Each limitation points directly to a prospective avenue of inquiry that can build upon the foundation laid by this dissertation. The recommended points revolve around (1) expanding linguistic resources, (2) advancing model architectures and training paradigms, and (3) deepening linguistic and evaluative analysis.

The most immediate and impactful avenue for future work is the continued expansion of OEPC. Efforts should focus not only on increasing the quantity of sentence pairs to mitigate data sparsity but also on diversifying the domains and genres represented. Adding data from sources such as social media, news commentary, radio programs, series transcripts, or different types of narrative content would create a more generalizable corpus.

Inspired by the success seen in research on LEV, EGY and TUN, a highly promising strategy would be to employ a pivot-language approach. This would involve creating a larger ORN-MSA corpus (a linguistically closer task) and then using a high-quality, pre-existing MSA-English NMT system to semi-automatically generate a much larger, synthetic ORN-ENG corpus. The existing OEPC would then serve as a "gold-standard" fine-tuning and validation set. A parallel effort should be dedicated to compiling a large monolingual corpus of ORN. Such a resource would be invaluable for training more powerful language models and would be the essential prerequisite for employing a range of semi-supervised NMT techniques.

A critical next step is to transition from the LSTM-based model to a Transformer-based architecture. The dominance of the Transformer architecture is further exemplified by advanced pre-trained models such as Bidirectional and Auto-Regressive Transformers (BART) and mBART (Lewis et al., 2020; Liu et al., 2020), which excel in Seq2Seq tasks by learning to reconstruct corrupted text. Concurrently, the frontier of low-resource translation is being redefined by massive multilingual models, most notably Meta AI's NLLB (No Language Left Behind) project, which has established new state-of-the-art benchmarks for hundreds of historically underserved language pairs.

Perhaps the most pragmatic and empirically validated avenue for substantially elevating the performance of the ORN-ENG system is the implementation of semi-supervised learning techniques, with back-translation standing as the preeminent methodology to prioritize. This approach is designed to directly mitigate the primary limitation of data scarcity by leveraging vast quantities of easily accessible monolingual data to synthetically generate new parallel training pairs. The process, while computationally intensive, is methodologically straightforward and has

been shown to yield significant improvements in LRL settings. To illustrate, its application in the Oranian context would unfold in a three-stage process:

1- An inverse, or "backward," TM model (ENG→ORN) would be trained exclusively on the existing 8.5K sentence-pair OEPC. This model's purpose is not to be a high-quality ENG-ORN translator, but rather to function as a "synthetic data generator." Concurrently, a large monolingual corpus of the target language, potentially comprising millions of sentences scraped from the web (e.g., from news articles, encyclopedias, or public domain books.

2- The core of the strategy is executed: the backward ENG→ORN model is used to translate this entire massive monolingual English corpus into ORN. For instance, an authentic English sentence like "*The meeting will end and the team will leave tomorrow afternoon*" would be fed into this model, producing a machine-generated ORN equivalent, such as "*الاجتماع غادي يكمل و اليكيب يشوروا غدوا عشية*". While this "pseudo-Oranian" output may contain minor grammatical infelicities or unnatural phrasing compared to a native speaker's utterance, it remains a linguistically plausible and semantically coherent source sentence. Repeating this process millions of times would yield a massive synthetic parallel corpus where TL (ENG) side is authentic and SL (ORN) side is machine-generated.

3- This large synthetic corpus is concatenated with the original, high-quality 8.5K OEPC. This combined, multi-million-sentence dataset is then used to train a new, definitive "forward" NMT model (ORN→ENG). The power of this technique is twofold. On one hand, the inherent noise and slight imperfections in the pseudo-Oranian source text serve as a powerful form of data augmentation and regularization, forcing the model to learn more robust and generalized representations of the dialect rather than simply memorizing the patterns of the small, clean OEPC. On the other hand, and more importantly, it exposes the model's decoder to an immense

variety of authentic English vocabulary and syntax, dramatically enhancing its target-side language model and leading to far more fluent and natural-sounding output. By synthetically amplifying the training data in this manner, back-translation directly confronts the foundational limitation of data scarcity, offering the most potent pathway to improving lexical coverage, enhancing decoder fluency, and achieving the next echelon of translation quality for the Oranian dialect.

As for deepening linguistic and evaluative analysis, future work should investigate more advanced, linguistically-motivated segmentation and normalization strategies specifically designed for the morphological and orthographic challenges of ORN. This could involve developing a dedicated morphological analyzer or designing a multi-input model that can process both the original script and a phonetically-normalized representation. In this vein, the Seq2Seq architectures proposed by Saeed and Habash (2025) offer a promising pathway, as they have proven resilient in processing diverse Arabic dialects even in the absence of standardized orthography or fully-developed morphological tools.

Future research should significantly deepen the linguistic and evaluative analysis of ORN by concentrating on advanced segmentation and normalization strategies. These techniques are essential for tackling the unique morphological and orthographic challenges inherent in ORN. Specifically, for OOV, the strategies that incorporate approaches like those proposed by Alimi et al. (2025) and Affi and Latiri (2025) are recommended. This might involve developing a dedicated morphological analyzer for ORN or creating a multi-input model that can effectively process both the original ORN script and a phonetically-normalized representation, thereby improving robustness against unseen word forms.

Moreover, subsequent work should incorporate a stronger focus on NER within ORN processing, drawing inspiration from the techniques outlined by Kessentini et al. (2025). Effective NER is crucial for extracting meaningful, structured information from the text. Implementing advanced NER strategies alongside sophisticated OOV handling will allow for more precise and contextually aware linguistic analysis. By following the guidance of these proposed methods, researchers can develop systems that are far more adept at overcoming the linguistic complexity of ORN, leading to more accurate and insightful evaluations.

While automatic metrics are indispensable for model development, they are an imperfect proxy for human judgment. A crucial future study would be to conduct a comprehensive human evaluation of the system's output, assessing translations based on the metrics of adequacy and fluency. This would provide a more nuanced understanding of the model's true performance and help validate the correlation of automated metrics for this specific language pair. Nevertheless, research should move toward more targeted analyses focusing on specific linguistic challenges. For instance, dedicated studies could investigate the translation of negation, the handling of code-switching between Oranian and French, or the modeling of specific verb aspect systems, leading to more focused and linguistically-informed model improvements.

This dissertation set out to illuminate and address a dark spot on the map of modern language technology. It began with a simple yet profound challenge: the Oranian dialect, the mother tongue for millions, had no voice in the world of NLP and MT. By systematically constructing a novel linguistic resource and engineering the first NMT system for this variety, this dissertation has transformed the status of the Oranian dialect from data-zero to data-scarce, and from technologically non-existent to empirically benchmarked. The resulting BLEU score of 30.9

is more than a number; it is a proof of existence, a validation of viability, and a firm foundation upon which a future of innovation can now be built.

The journey undertaken herein has reaffirmed a fundamental truth of low-resource NLP: that the initial and most arduous challenges often lie not in algorithmic novelty but in the painstaking work of data creation and linguistic curation. OEPC stands as a testament to this, and its creation is arguably the most enduring contribution of this work. While the limitations of this initial step are clear, they do not diminish its significance. They are not endpoints, but rather signposts, pointing the way toward more sophisticated models, richer datasets, and a deeper understanding of the unique computational properties of this complex dialect. This work has taken the crucial first step, bringing the Oranian dialect into the global conversation of computational linguistics, and in doing so, opens a new and promising chapter in the ongoing quest for a more linguistically inclusive digital world.

REFERENCES

Abainia, K. (2020). DZDC12: A new multipurpose parallel Algerian Arabizi–French code-switched corpus. *Language Resources and Evaluation*, *54*(2), 419-455. https://doi.org/10.1007/s10579-019-09454-8

Abainia, K., Kara, K., & Hamouni, T. (2022). A new corpus and lexicon for offensive Tamazight language detection. In *Proceedings of the 7th International Workshop on Social Media World Sensors*, Barcelona, Spain. (pp. 1-6). Association for Computing Machinery. https://doi.org/10.1145/3544795.3544852

Abdaoui, A., Berrimi, M., Oussalah, M., & Moussaoui, A. (2021). DziriBERT: A pre-trained language model for the Algerian dialect. *arXiv preprint arXiv:2109.12346*. https://doi.org/10.48550/arXiv.2109.12346

Abdul-Nabi, R., Obeidat, R., & Bsoul, A. (2024). A survey on machine translation of low-resource Arabic dialects. In *2024 15th International Conference on Information and Communication Systems (ICICS)*, Irbid, Jordan. (pp. 1–6). IEEE. https://doi.org/10.1109/ICICS63486.2024.10638285

Abid, W. (2020). The SADID evaluation datasets for low-resource spoken language machine translation of Arabic dialects. In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 6030–6043). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.530

Abidi, K., & Smaïli, K. (2017). An empirical study of the Algerian dialect of social network. In K. Smaïli, M. Diouri, & K. Benzakour (Eds.), *Proceedings of the International Conference*

on Natural Language, Signal and Speech Processing, ICNLSSP 2017 (pp. 85–89). International Science and General Applications (ISGA). https://inria.hal.science/hal-01659997v1

Abidi, K., & Smaïli, K. (2018). An automatic learning of an Algerian dialect lexicon by using multilingual word embeddings. In N. Calzolari et al. (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (pp. 832-838). European Language Resources Association (ELRA). https://aclanthology.org/L18-1133/

Abidi, K., Menacer, M. A., & Smaili, K. (2017). CALYOU: A comparable spoken Algerian corpus harvested from YouTube. In *Proceedings of the 18th Annual Conference of the International Communication Association (Interspeech 2017)*. Stockholm, Sweden. (pp. 3742-3746). https://doi.org/10.21437/Interspeech.2017-1305

Abujar, S., Masum, A. K. M., Bhattacharya, A., Dutta, S., & Hossain, S. A. (2021). English to Bengali neural machine translation using global attention mechanism. In J. M. R. S. Tavares, S. Chakrabarti, A. Bhattacharya, & S. Ghatak (Eds.), *Emerging technologies in data mining and information security* (pp. 359–369). Springer. https://doi.org/10.1007/978-981-15-9774-9_35

Adouane, W., & Dobnik, S. (2017). Identification of languages in Algerian Arabic multilingual documents. In *Proceedings of the third Arabic Natural Language Processing Workshop,* Valencia, Spain. (pp. 1-8). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1301

Adouane, W., Semmar, N., Johansson, R., & Bobicev, V. (2016). Automatic detection of Arabicized Berber and Arabic varieties. In P. Nakov, M. Zampieri, L. Tan, N. Ljubešić, J. Tiedemann, & S. Malmasi (Eds.), *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)* (pp. 63–72). The COLING 2016 Organizing Committee. https://aclanthology.org/W16-4809/

Affi, M., & Latiri, C. (2025). Named entity recognition for Algerian Arabic dialect using multi-dialect-Arabic-BERT based architectures. In I. Czarnowski, R. J. Howlett, & L. C. Jain (Eds.), *Intelligent Decision Technologies, KESIDT 2024. Smart Innovation, Systems and Technologies, Vol 411*, (pp. 449–459). Springer. https://doi.org/10.1007/978-981-97-7419-7_39

Alalou, A. (2023). The sociolinguistic situation in North Africa: Recognizing and institutionalizing Tamazight and new challenges. *Annual Review of Linguistics*, *9*(1), 155-170. https://doi.org/10.1146/annurev-linguistics-030521-054916

Al-Gaphari, G. H., & Al-Yadoumi, M. (2010). A method to convert Sana'ani accent to modern standard Arabic. *International Journal of Information Science and Management (IJISM)*, *8*(1), 39-49.

Alhammi, H., & Haddar, K. (2022). Toward building a bilingual dictionary for Libyan dialect-modern standard Arabic machine translation. In D. Aveiro, J. Dietz, & J. Filipe (Eds.), *Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, Vol 2: IC3K* (pp. 75-81). Science and Technology Publications, Lda. https://doi.org/10.5220/0011560700003335

Ali, A., Chowdhury, S., Afify, M., El-Hajj, W., Hajj, H., Abbas, M., Hussein, A., Ghneim, N., Abushariah, M., & Alqudah, A. (2021). Connecting Arabs: Bridging the gap in dialectal speech recognition. *Communications of the ACM*, *64*(4), 124-129. https://doi.org/10.1145/3451150

Al-Ibrahim, R., & Duwairi, R. M. (2020). Neural machine translation from Jordanian dialect to modern standard Arabic. In *2020 11th International Conference on Information and Communication Systems (ICICS)*, Irbid, Jordan. (pp. 173-178). IEEE. https://doi.org/10.1109/ICICS49469.2020.239505

Alimi, T., Boujelbene, R., & Belguith, L. H. (2025). ARDIAL-BERT: Advancing multidialectal Arabic named entity recognition through continual pretraining. *IEEE Transactions on Artificial Intelligence*, 1–10. https://doi.org/10.1109/TAI.2025.3562162

Allen, R. B. (1987). Several studies on natural language and back-propagation. In *Proceedings of the IEEE First International Conference on Neural Networks Vol. 2*, San Diego, CA, USA, 335-341. IEEE.

Almaaytah, S. A., & Alzobidy, S. A. (2023). Challenges in rendering Arabic text to English using machine translation: A systematic literature review. *IEEE Access*, *11*, 94772–94779. https://doi.org/10.1109/ACCESS.2023.3309642

Al-Mannai, K., Sajjad, H., Khader, A., Al Obaidli, F., Nakov, P., & Vogel, S. (2014). Unsupervised word segmentation improves dialectal Arabic to English machine translation. N. Habash & S. Vogel (Eds.), *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)* (pp. 207–216). Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-3628

Almansor, E. H., & Al-Ani, A. (2017). Translating dialectal Arabic as low resource language using word embedding. In R. Mitkov & G. Angelova (Eds.), *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017* (pp. 52–57). INCOMA Ltd. https://doi.org/10.26615/978-954-452-049-6_008

Almeman, K., & Lee, M. (2012). Towards developing a multi-dialect morphological analyser for Arabic. In *4th International Conference on Arabic Language Processing*, Rabat, Morocco, 19-25.

Alnassan, A. (2023). Automatic standardization of Arabic dialects for machine translation. In *Proceedings of the International Conference on Ummah 2022*, Kelantan, Malaysia, 1013-1025. https://doi.org/10.48550/arXiv.2301.03447

Alothman, A., & Alsalman, A. (2020). Arabic morphological analysis techniques. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *11*(2), 214-222. https://doi.org/10.14569/IJACSA.2020.0110229

ALPAC (Automatic Language Processing Advisory Committee). (1966). *Languages and machines: Computers in translation and linguistics*, A Report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council, Washington, DC. https://doi.org/10.17226/9547

Al-Shargi, F., Kaplan, A., Eskander, R., Habash, N., & Rambow, O. (2016). Morphologically annotated corpora and morphological analyzers for Moroccan and Sanaani Yemeni Arabic. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association (ELRA), 1300–1306. https://aclanthology.org/L16-1207/

Al-Sughaiyer, I. A., & Al-Kharashi, I. A. (2004). Arabic morphological analysis techniques: A comprehensive survey. *Journal of the American society for information science and technology*, *55*(3), 189-213. https://doi.org/10.1002/asi.10368

Althobaiti, M. J. (2020). Automatic Arabic dialect identification systems for written texts: A survey. *arXiv preprint arXiv:2009.12622*. 1-41. https://doi.org/10.48550/arXiv.2009.12622

Alzamzami, F., & Saddik, A. E. (2023). OSN-MDAD: Machine translation dataset for Arabic multi-dialectal conversations on online social media. *arXiv preprint arXiv:2309.12137*. https://doi.org/10.48550/arXiv.2309.12137

Amazouz, D., Adda-Decker, M., & Lamel, L. (2018). The French-Algerian code-switching triggered audio corpus (FACST). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA), 1468-1473. https://aclanthology.org/L18-1233/

Aminian, M., Ghoneim, M., & Diab, M. (2014). Handling OOV words in dialectal Arabic to English machine translation. In P. Nakov, P. Osenova, & C. Vertan (Eds.), *Proceedings of the EMNLP'2014 Workshop on Language Technology for Closely Related Languages and Language Variants* (pp. 99–108). Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-4213

Aransa, W. (2015). *Statistical machine translation of the Arabic language*. [Doctoral Dissertation. Université du Maine]. https://theses.hal.science/tel-01316544v1

Arnold, D., Balkan, L., Humphreys, R. L., Meijer, S., & Sadler, L. (2001). *Machine translation: An introductory guide*. NCC Blackwell Ltd.

Babaali, B., & Salem, M. (2023). Survey of the Arabic machine translation corpora. In S. Chikhi, G. Diaz-Descalzo, A. Amine, A. Chaoui, D. E. Saidouni, & M. K. Kholladi (Eds.), *Modelling and Implementation of Complex Systems. MISC 2022. Lecture Notes in Networks and Systems, Vol. 593,* (pp. 205-219). Springer. https://doi.org/10.1007/978-3-031-18516-8_15

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In Y. Bengio & Y. LeCun (Eds.), *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015*, San Diego, CA, USA. (pp. 1–15). https://doi.org/10.48550/arXiv.1409.0473

Bakr, H. A., Shaalan, K., & Ziedan, I. (2008). A hybrid approach for converting written Egyptian colloquial dialect into diacritized Arabic. In *The 6th International Conference on Informatics and Systems, INFOS2008.* Cairo, Egypt. 27-33.

Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In J. Goldstein, A. Lavie, C. Lin, & C. Voss (Eds.), *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization* (pp. 65–72). Association for Computational Linguistics. https://aclanthology.org/W05-0909/

Baniata, L. H., Park, S., & Park, S. B. (2018a). A neural machine translation model for Arabic dialects that utilizes multitask learning (MTL). *Computational Intelligence and Neuroscience*. 7534712, 1-10. https://doi.org/10.1155/2018/7534712

Baniata, L. H., Park, S., & Park, S. B. (2018b). A multitask-based neural machine translation model with part-of-speech tags integration for Arabic dialects. *Applied Sciences*, *8*(12), 2502. 1-18. https://doi.org/10.3390/app8122502

Bar-Hillel, Y. (1951). The present state of research on mechanical translation. *American Documentation, 2*(4), 229-237. https://doi.org/10.1002/asi.5090020408

Bar-Hillel, Y. (1955). Idioms. In W. N. Locke & A. D. Booth (Eds.), *Machine translation of languages: Fourteen essays* (pp. 183-193). MIT Press.

Bar-Hillel, Y. (1959). Report on the state of machine translation in the United States and Great Britain. *Technical report no 1*. Hebrew University Jerusalem.

Bar-Hillel, Y. (1960). The present status of automatic translation of languages. *Advances in Computers 1,* 91-163. https://doi.org/10.1016/S0065-2458(08)60607-5

Bel'skaya, I. K., Korolev, L. N., & Panov, D. Y. (Eds.). (1959). *The translating machine of P. P. Trojanskij: Collection of materials on a translating machine for the translation from one language into others, proposed by P. P. Trojanskij in 1933*. CCCP, Moscow.

Bella, G., Helm, P., Koch, G., & Giunchiglia, F. (2023). Towards bridging the digital language divide. *arXiv preprint arXiv:2307.13405*. https://doi.org/10.48550/arXiv.2307.13405

Benali, A., Maaloul, M. H., & Belguith, L. H. (2023). Automatic processing of Algerian dialect: Corpus construction and segmentation. *SN Computer Science*, *4*(5), 597. https://doi.org/10.1007/s42979-023-02097-1

Bengio, Y., & Senécal, J. S. (2003). Quick training of probabilistic neural nets by importance sampling. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics* (pp. 17–24). PMLR. https://proceedings.mlr.press/r4/bengio03a.html

Benrabah, M. (2014). Competition between four "world" languages in Algeria. *Journal of World Languages*, *1*(1), 38-59. https://doi.org/10.1080/21698252.2014.893676

Beyer, R. T. (1965). Hurdling the language barrier. *Physics Today*, *18*(1), 46-52. https://doi.org/10.1063/1.3047154

Bhattacharyya, P. (2015). *Machine translation*. CRC Press. https://doi.org/10.1201/b18004

Bies, A., Song, Z., Maamouri, M., Grimes, S., Lee, H., Wright, J., Strassel, S., Habash, N., Eskander, R., & Rambow, O. (2014). Transliteration of Arabizi into Arabic orthography: Developing a parallel annotated Arabizi-Arabic script SMS/chat corpus. In N. Habash & S. Vogel (Eds.), *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)* (pp. 93–103). Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-3612

Bontcheva, K. (1997). Generation of multilingual explanations from conceptual graphs. In R. Mitkov & N. Nicolov (Eds.), *Recent Advances in Natural Language Processing: Selected Papers from RANLP '95* (pp. 365-374). John Benjamins Publishing Company. https://doi.org/10.1075/cilt.136.30bon

Bouamor, H., Habash, N., Salameh, M., Zaghouani, W., Rambow, O., Abdulrahim, D., Obeid, O., Khalifa, S., Eryani, F., Erdmann, A., & Oflazer, K. (2018). The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language*

Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA), 3387-3396. https://aclanthology.org/L18-1535/

Boudlal, A., Lakhouaja, A., Mazroui, A., Meziane, A., Bebah, M. O. A. O., & Shoul, M. (2010). Alkhalil morpho sys1: A morphosyntactic analysis system for Arabic texts. In *Proceedings of International Arab Conference on Information Technology ACIT2010* (pp. 1-6). The Colleges of Computing and Information Society, Zarqa University, Jordan.

Bougrine, S., Chorana, A., Lakhdari, A., & Cherroun, H. (2017). Toward a web-based speech corpus for Algerian dialectal Arabic varieties. In N. Habash, M. Diab, K. Darwish, W. El-Hajj, H. Al-Khalifa, H. Bouamor, N. Tomeh, M. El-Haj, & W. Zaghouani (Eds.), *Proceedings of the Third Arabic Natural Language Processing Workshop* (pp. 138–146). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1317

Bouhadiba, F. A. (1988). *Aspects of Algerian Arabic verb, phonology and morphology*. [Doctoral Dissertation, University of Reading]. http://ethos.bl.uk/383612

Brockett, C., Aikawa, T., Aue, A., Menezes, A., Quirk, C., & Suzuki, H. (2002). English-Japanese example-based machine translation using abstract linguistic representations. In *Proceedings of the 2002 COLING Workshop on Machine Translation in Asia - Volume 16* (pp. 1–7). Association for Computational Linguistics. https://doi.org/10.3115/1118794.1118798

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

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

Brunello, M. (2014). *Domain and genre dependency in statistical machine translation* [Doctoral dissertation, University of Leeds]. https://etheses.whiterose.ac.uk/id/eprint/8420/

Bull, W. E., Africa, C., & Teichroew, D. (1955). Some problems of the 'word'. In W. N. Locke & A. D. Booth (Eds.), *Machine translation of languages: Fourteen essays* (pp. 86-103). MIT Press.

Carl, M., & Way, A. (2003). *Recent advances in example-based machine translation*. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-010-0181-6

Castano, A., & Casacuberta, F. (1997). A connectionist approach to machine translation. In *Fifth European Conference on Speech Communication and Technology*. (pp. 1-4). https://doi.org/10.21437/Eurospeech.1997-50

Chami, A. (2009). A historical background of the linguistic situation in Algeria. *Almawaqif Journal*, *4*(1), 387-395.

Chan, S. (2004). *A dictionary of translation technology*. Chinese University Press.

Chan, S. (2014). The development of translation technology: 1967–2013. In S. Chan (Ed.), *Routledge Encyclopedia of Translation Technology* (pp. 3-31). Routledge. https://doi.org/10.4324/9781315749129

Charef, A. (2016). Linguistic peculiarities of Oran's dialect. *Al Kalim Journal*, *1*(2), 4-14.

Chauhan, S., & Daniel, P. (2023). A comprehensive survey on various fully automatic machine translation evaluation metrics. *Neural Processing Letters*, *55*(9), 12663-12717. https://doi.org/10.1007/s11063-022-10835-4

Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd annual meeting of the association for computational linguistics (acl'05)* (pp. 263-270). Association for Computational Linguistics. https://doi.org/10.3115/1219840.1219873

Chiang, D. (2007). Hierarchical phrase-based translation. *Computational Linguistics*, *33*(2), 201-228. https://doi.org/10.1162/coli.2007.33.2.201

Cho, K., Van Merriënboer, B., Gülçehre, C. Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1724–1734). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1179

Chomsky, N. (1965). Aspects of the Theory of Syntax. MIT Press. https://doi.org/10.21236/AD0616323

Chrisman, L. (1991). Learning recursive distributed representations for holistic computation. *Connection Science*, *3*(4), 345-366. https://doi.org/10.1080/09540099108946592

Church, K. (2011). A pendulum swung too far. *Linguistic Issues in Language Technology*, *6*(5), 1-27. https://doi.org/10.33011/lilt.v6i.1245

Church, K. W., & Hovy, E. H. (1993). Good applications for crummy machine translation. *Machine Translation*, *8*, 239-258. https://doi.org/10.1007/BF00981759

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning* (pp. 160–167). Association for Computing Machinery. https://doi.org/10.1145/1390156.1390177

Costa-jussà, M. R., & Fonollosa, J. A. (2015). Latest trends in hybrid machine translation and its applications. *Computer Speech & Language*, *32*(1), 3-10. https://doi.org/10.1016/j.csl.2014.11.001

Cotterell, R., & Callison-Burch, C. (2014). A multi-dialect, multi-genre corpus of informal written Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. (pp. 241-245).

Cotterell, R., Renduchintala, A., Saphra, N., & Callison-Burch, C. (2014). An Algerian Arabic French code-switched corpus. In *Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools.* Reykjavik, Iceland. (pp. 34-37).

Cui, Y., Chen, Z., Wei, S., Wang, S., Liu, T., & Hu, G. (2016). Attention-over-attention neural networks for reading comprehension. *arXiv preprint arXiv:1607.04423*. https://doi.org/10.48550/arXiv.1607.04423

Dahou, A. H., & Cheragui, M. A. (2023a). DzNER: A large Algerian named entity recognition dataset. *Natural Language Processing Journal*, *3*, 100005. https://doi.org/10.1016/j.nlp.2023.100005

Dahou, A. H., & Cheragui, M. A. (2023b). Impact of normalization and data augmentation in NER for Algerian Arabic dialect. In S. Chikhi, G. Diaz-Descalzo, A. Amine, A. Chaoui, D. E. Saidouni, & M. K. Kholladi (Eds.), *Modelling and implementation of complex systems* (pp. 250–261). Springer. https://doi.org/10.1007/978-3-031-18516-8_18

Dahou, A. H., & Cheragui, M. A. (2023c). Named entity recognition for Algerian Arabic dialect in social media. In M. R. Laouar, V. E. Balas, B. Lejdel, S. Eom, & M. A. Boudia (Eds.), *12th International Conference on Information Systems and Advanced Technologies "ICISAT 2022"* (pp. 143–153). Springer. https://doi.org/10.1007/978-3-031-25344-7_13

Dahou, A., Dahou, A. H., Cheragui, M. A., Abdedaiem, A., Al-qaness, M. A. A., Abd Elaziz, M., Ewees, A. A., & Zheng, Z. (2025). A survey on dialect Arabic processing and analysis: Recent advances and future trends. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *24*(8), 84. https://doi.org/10.1145/3747290

Dajun, Z., & Yun, W. (2015). Corpus-based machine translation: Its current development and perspectives. *International Forum of Teaching and Studies, 11*(1-2), 90-95.

Darwish, K., Magdy, W., & Mourad, A. (2012). Language processing for Arabic microblog retrieval. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management CIKM'12* (pp. 2427–2430). Association for Computing Machinery. https://doi.org/10.1145/2396761.2398658

Denoual, E., & Lepage, Y. (2005). BLEU in characters: Towards automatic MT evaluation in languages without word delimiters. *Companion Volume to the Proceedings of the Second International Joint Conference on Natural Language Processing*, 79-84. Association for Computational Linguistics. https://aclanthology.org/I05-2014/

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. https://doi.org/10.48550/arXiv.1810.04805

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

DiPietro, R., & Hager, G. D. (2020). Deep learning: RNNs and LSTM. In S. K. Zhou, D. Rueckert, & G. Fichtinger (Eds.), *Handbook of medical image computing and computer assisted intervention* (pp. 503-519). Academic Press. https://doi.org/10.1016/B978-0-12-816176-0.00026-0

Djellab, M., Amrouche, A., Bouridane, A., & Mehallegue, N. (2017). Algerian modern colloquial Arabic speech corpus (AMCASC): Regional accents recognition within complex socio-linguistic environments. *Language Resources and Evaluation*, *51*, 613-641. https://doi.org/10.1007/s10579-016-9347-6

Dou, A., & Kissi, K. (2024). Building the Oranian-English parallel corpus: Methodology and compilation Process. *Journal of Languages and Translation*, *4*(2), 161-174. https://doi.org/10.70204/jlt.v4i2.358

Dou, A., & Kissi, K. (2025). Building an Oranian-English parallel corpus for automated translation training. *Digital Scholarship in the Humanities, 40*(1), 87-95. https://doi.org/10.1093/llc/fqae089

Du, C., Tu, Z., & Jiang, J. (2021). Order-agnostic cross entropy for non-autoregressive machine translation. In *Proceedings of the 38th International Conference on Machine Learning*, (pp. 2849-2859). PMLR. https://proceedings.mlr.press/v139/du21c.html

Durrani, N., Al-Onaizan, Y., & Ittycheriah, A. (2014). Improving Egyptian-to-English SMT by mapping Egyptian into MSA. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing. CICLing 2014. Lecture Notes in Computer Science (Vol. 8404*, (pp. 271-282). Springer. https://doi.org/10.1007/978-3-642-54903-8_23

Dyer, C., Lopez, A., Ganitkevitch, J., Weese, J., Ture, F., Blunsom, P., Setiawan, H., Eidelman, V., & Resnik, P. (2010). cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In S. Kübler (Ed.), *Proceedings of the ACL 2010 System Demonstrations* (pp. 7–12). Association for Computational Linguistics. https://aclanthology.org/P10-4002/

Elfardy, H., & Diab, M. (2012). AIDA: Automatic identification and glossing of dialectal Arabic. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT)*, Trento, Italy. 83.

Elfardy, H., Al-Badrashiny, M., & Diab, M. (2014). AIDA: Identifying code switching in informal Arabic text. In M. Diab, J. Hirschberg, P. Fung, & T. Solorio (Eds.), *Proceedings of the First Workshop on Computational Approaches to Code Switching* (pp. 94–101). Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-3911

Elnagar, A., Yagi, S. M., Nassif, A. B., Shahin, I., & Salloum, S. A. (2021). Systematic literature review of dialectal Arabic: Identification and detection. *IEEE Access*, *9*, 31010-31042. https://doi.org/10.1109/ACCESS.2021.3059504

Eskander, R., Habash, N., Rambow, O., & Pasha, A. (2016). Creating resources for dialectal Arabic from a single annotation: A case study on Egyptian and Levantine. In Y. Matsumoto & R. Prasad (Eds.), *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 3455-3465). The COLING 2016 Organizing Committee. https://aclanthology.org/C16-1326/

Faheem, M. A., Wassif, K. T., Bayomi, H., & Abdou, S. M. (2024). Improving neural machine translation for low resource languages through non-parallel corpora: A case study of Egyptian dialect to modern standard Arabic translation. *Scientific Reports*, *14*(1), 2265. https://doi.org/10.1038/s41598-023-51090-4

Faheem, M. A., Wassif, K. T., Mobarz, H. B., & Abdou, S. M. (2023). Enhancing neural machine translation for languages with limited resources utilizing non-parallel texts: An examination of translation from Egyptian dialect to modern standard Arabic. In *2023 33rd International Conference on Computer Theory and Applications (ICCTA)* (pp. 190–198). IEEE. https://doi.org/10.1109/ICCTA60978.2023.10969230

Farhan, W., Talafha, B., Abuammar, A., Jaikat, R., Al-Ayyoub, M., Tarakji, A. B., & Toma, A. (2020). Unsupervised dialectal neural machine translation. *Information Processing & Management*, *57*(3), 102181. https://doi.org/10.1016/j.ipm.2019.102181

Forcada, M. L., & Ñeco, R. P. (1997). Recursive hetero-associative memories for translation. In J. Mira, R. Moreno-Díaz, & J. Cabestany (Eds.), *Biological and artificial computation: From*

neuroscience to technology (pp. 453–462). Springer Berlin Heidelberg. https://doi.org/10.1007/BFb0032504

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., & Tyers, F. M. (2011). Apertium: A free/open-source platform for rule-based machine translation. *Machine Translation*, *25*, 127-144. https://doi.org/10.1007/s10590-011-9090-0

Fradkov, A. L. (2020). Early history of machine learning. *IFAC-PapersOnLine*, *53*(2), 1385-1390. https://doi.org/10.1016/j.ifacol.2020.12.1888

Gage, P. (1994). A new algorithm for data compression. *C/C++ Users Journal*, *12*(2), 23-38.

Garg, A., & Agarwal, M. (2018). Machine translation: A literature review. *arXiv preprint arXiv:1901.01122*. 1-17. https://doi.org/10.48550/arXiv.1901.01122

Gaser, M., Mager, M., Hamed, I., Habash, N., Abdennadher, S., & Vu, N. T. (2023). Exploring segmentation approaches for neural machine translation of code-switched Egyptian Arabic-English text. *arXiv preprint arXiv:2210.06990*. 1-16. https://doi.org/10.48550/arXiv.2210.06990

Goshawke, W., Kelly, I. D., & Wigg, J. D. (1987). *Computer translation of natural language*. John Wiley & Sons.

Goutte, C., Cancedda, N., Dymetman, M., & Foster, G. (2009). *Learning machine translation*. MIT Press. https://doi.org/10.7551/mitpress/9780262072977.001.0001

Green, S., Heer, J., & Manning, C. D. (2015). Natural language translation at the intersection of AI and HCI: Old questions being answered with both AI and HCI. *Communications of the ACM, 58*(9), 46-53. https://doi.org/10.1145/2791301.2798086

Gross, A. (1992). Limitations of computer as a translation tool. In J. Newton (Ed.), *Computers in translation: A practical appraisal* (pp. 96-130). Routledge.

Guellil, I. (2021). *A sentiment analysis approach for Arabic dialects texts analysis based on automatic translation: Application to the Algerian dialect.* [Doctoral dissertation, Higher National School of Computer Science, Algeria]. https://doi.org/10.13140/RG.2.2.18651.90402

Guellil, I., & Azouaou, F. (2016). Arabic dialect identification with an unsupervised learning (based on a lexicon). application case: Algerian dialect. In *2016 IEEE Intl Conference on Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (DCABES)* (pp. 724–731). IEEE. https://doi.org/10.1109/CSE-EUC-DCABES.2016.268

Guellil, I., Adeel, A., Azouaou, F., Benali, F., Hachani, A. E., Dashtipour, K., Cogate, M., Ieracitano, C., Kashani, R., & Hussain, A. (2021). A Semi-supervised approach for sentiment analysis of Arab(ic+izi) messages: Application to the Algerian dialect. *SN Computer Science, 2*(118). https://doi.org/10.1007/s42979-021-00510-1

Guellil, I., Adeel, A., Azouaou, F., Benali, F., Hachani, A. E., & Hussain, A. (2018). Arabizi sentiment analysis based on transliteration and automatic corpus annotation. In A. Balahur, S. M. Mohammad, V. Hoste, & R. Klinger (Eds.), *Proceedings of the 9th Workshop on*

Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (pp. 335–341). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6249

Guellil, I., Azouaou, F., Benali, F., Hachani, A. E., & Mendoza, M. (2020). The role of transliteration in the process of Arabizi translation/sentiment analysis. In M. Abd Elaziz, M. A. A. Al-qaness, A. A. Ewees, & A. Dahou (Eds.), *Recent advances in NLP: The case of Arabic language* (pp. 101-128). Springer. https://doi.org/10.1007/978-3-030-34614-0_6

Guellil, I., Azouaou, F., Saâdane, H., & Semmar, N. (2017). Une approche fondée sur les lexiques d'analyse de sentiments du dialecte algérien. *Revue TAL: traitement automatique des langues*, *58*(3), 41-65.

Guerrero, J. (2015). Preliminary notes on the current Arabic dialect of Oran (Western Algeria). *Romano Arabica*, *15*, 219-233.

Guerrero, J. (2021). On interdental fricatives in the first-layer dialects of Maghrebi Arabic. *Brill's Journal of Afroasiatic Languages and Linguistics*, *13*(2), 288-308. https://doi.org/10.1163/18776930-01302005

Habash, N. (2021). Arabic dialect processing. In M. Zampieri & P. Nakov (Eds.), *Similar languages, varieties, and dialects: A computational perspective* (pp. 279–303). Cambridge University Press. https://doi.org/10.1017/9781108565080.017

Habash, N., & Rambow, O. (2005). Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In K. Knight, H. T. Ng, & K. Oflazer (Eds.), *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)* (pp. 573–580). Association for Computational Linguistics. https://doi.org/10.3115/1219840.1219911

Habash, N., & Rambow, O. (2006). MAGEAD: A morphological analyzer and generator for the Arabic dialects. In N. Calzolari, C. Cardie, & P. Isabelle (Eds.), *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics* (pp. 681–688). Association for Computational Linguistics. https://doi.org/10.3115/1220175.1220261

Habash, N., Diab, D., & Rambow, O. (2012). Conventional orthography for dialectal Arabic. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12),* Istanbul, Turkey. 711-718. European Language Resources Association (ELRA). https://aclanthology.org/L12-1328/

Habash, N., Eskander, R., & Hawwari, A. (2012). A morphological analyzer for Egyptian Arabic. In L. Cahill & A. Albright (Eds.), *Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology* (pp. 1–9). Association for Computational Linguistics. https://aclanthology.org/W12-2301/

Habash, N., Marzouk, R., Khairallah, C., & Khalifa, S. (2022). Morphotactic modeling in an open-source multi-dialectal Arabic morphological analyzer and generator. In G. Nicolai & E. Chodroff (Eds.), *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology* (pp. 92–102). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.sigmorphon-1.10

Habash, N., Roth, R., Rambow, O., Eskander, R., & Tomeh, N. (2013). Morphological analysis and disambiguation for dialectal Arabic. In L. Vanderwende, H. Daumé III, & K. Kirchhoff (Eds.), *Proceedings of the 2013 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies (pp. 426–432). Association for Computational Linguistics. https://aclanthology.org/N13-1044/

Hamadouche, K., Bousmaha, K. Z., Bekkoucha, M. A., & Hadrich-Belguith, L. (2023). AlgBERT: Automatic construction of annotated corpus for sentiment analysis in Algerian dialect. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *22*(12), 1-17. https://doi.org/10.1145/3632948

Hamdani, G. D., Selouani, S. A., & Boudraa, M. (2010). Algerian Arabic speech database (ALGASD): Corpus design and automatic speech recognition application. *Arabian Journal for Science and Engineering*, *35*(2), 157-166.

Hamdi, A., Boujelbane, R., Habash, N., & Nasr, A. (2013). The effects of factorizing root and pattern mapping in bidirectional Tunisian-standard Arabic machine translation. In A. Way, K. Sima'an, & M. L. Forcada (Eds.), *Proceedings of Machine Translation Summit XIV: Papers* (pp. 125-134). European Association for Machine Translation. https://aclanthology.org/2013.mtsummit-papers.16/

Hamdi, A., Nasr, A., Habash, N., & Gala, N. (2015). POS-tagging of Tunisian dialect using standard Arabic resources and tools. In N. Habash, S. Vogel, & K. Darwish (Eds.), *Proceedings of the Second Workshop on Arabic Natural Language Processing* (pp. 59–68). Association for Computational Linguistics. https://doi.org/10.18653/v1/W15-3207

Hamed, I., Habash, N., & Vu, N. T. (2023). Data augmentation techniques for machine translation of code-switched texts: A comparative study. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 140–154).

Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.11

Harrat, S. (2018). *Statistical machine translation: Application to low resourced languages.* [Doctoral dissertation, Higher National School of Computer Science, Algeria]. https://theses.hal.science/tel-03186940

Harrat, S., Meftouh, K., Abbas, M., & Smaïli, K. (2014). Building resources for Algerian Arabic dialects. In *Proceedings of the 15th Annual Conference of the International Communication Association Interspeech*, Singapore. 2123-2127. ISCA https://doi.org/10.21437/Interspeech.2014-481

Harrat, S., Meftouh, K., Abbas, M., Hidouci, W. K., & Smaïli, K. (2016). An Algerian dialect: Study and resources. *International Journal of Advanced Computer Science and Applications (IJACSA), 7*(3), 384-396. https://doi.org/10.14569/IJACSA.2016.070353

Harrat, S., Meftouh, K., Abbas, M., Jamoussi, S., Saad, M., & Smaïli, K. (2015). Cross-dialectal Arabic processing. In A. Gelbukh (Ed.), *International Conference on Intelligent Text Processing and Computational Linguistics CICLing 2015*, (pp. 620-632). Springer. https://doi.org/10.1007/978-3-319-18111-0_47

Harrat, S., Meftouh, K., & Smaïli, K. (2017). Creating parallel Arabic dialect corpus: Pitfalls to avoid. In A. Gelbukh (Ed.), *Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing CICLing 2017*, (pp. 1-11). https://hal.science/hal-01557405v1

Harrat, S., Meftouh, K., & Smaïli, K. (2018). Maghrebi Arabic dialect processing: An overview. *Journal of the International Science and General Applications, 1*(1).1-5.

Harrat, S., Meftouh, K., & Smaïli, K. (2019a). Machine translation for Arabic dialects (survey). *Information Processing and Management*, *56*(2), 262-273. https://doi.org/10.1016/j.ipm.2017.08.003

Harrat, S., Meftouh, K., & Smaïli, K. (2019b). Script independent morphological segmentation for Arabic Maghrebi dialects: An application to machine translation. *Computación y sistemas*, *23*(3), 979-989. https://doi.org/10.13053/cys-23-3-3267

Hashem, R. A. (2021). The quest for NLP applications and tools: The case of Standard Arabic and the dialects. In *2021 International Conference on Asian Language Processing (IALP)*, Singapore. 222-228. IEEE. https://doi.org/10.1109/IALP54817.2021.9675225

Hassan, H., Elaraby, M., & Tawfik, A. Y. (2017). Synthetic data for neural machine translation of spoken-dialects. In S. Sakti & M. Utiyama (Eds.), *Proceedings of the 14th International Conference on Spoken Language Translation* (pp. 82–89). International Workshop on Spoken Language Translation. https://aclanthology.org/2017.iwslt-1.12/

Haugeland, J. (1985). *Artificial intelligence: The very idea*. MIT Press.

He, Z., Liu, Q., & Lin, S. (2008). Improving statistical machine translation using lexicalized rule selection. In D. Scott & H. Uszkoreit (Eds.), *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)* (pp. 321–328). Coling 2008 Organizing Committee. https://doi.org/10.3115/1599081.1599122

Helaoui, R. (2007). *Building an English-to-Arabic statistical machine translation's system for the ISL's lecture translator*. Language Technologies Institute: Carnegie Mellon University.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hutchins, W. J., & Somers, H. L. (1992). *An introduction to machine translation*. Academic Press.

Hutchins, W. J. (1993). Latest developments in machine translation technology: Beginning a new era in MT research. In *Proceedings of Machine Translation Summit IV* (pp. 11–34). International Association for Machine Translation. https://aclanthology.org/1993.mtsummit-1.2/

Hutchins, W. J. (1994). Machine translation: History and general principles. In R. E. Asher & J. M. Y. Simpson (Eds.), *The Encyclopedia of language and linguistics* (pp. 2322-2332). Pergamon Press.

Hutchins, W. J. (1995). Machine translation: A brief history. In E.F.K. Koerner & R.E. Asher (Eds.), *Concise history of the language sciences: From the Sumerians to the cognitivists* (pp. 431–445). Pergamon Press. https://doi.org/10.1016/B978-0-08-042580-1.50066-0

Hutchins, W. J. (1997a). Translation technology and the translator. In *Proceedings of the Eleventh Conference of the Institute of Translation and Interpreting,* London. 113-120.

Hutchins, W. J. (1997b). From first conception to first demonstration: The nascent years of machine translation, 1947–1954. A chronology. *Machine Translation*, *12*, 195-252. https://doi.org/10.1023/A:1007969630568

Hutchins, W. J. (Ed.). (2000a). *Early years in machine translation: Memoirs and biographies of pioneers*. John Benjamins Publishing. https://doi.org/10.1075/sihols.97

Hutchins, W. J. (2000b). Yehoshua Bar-Hillel: A philosopher's contribution to machine translation. In W. J. Hutchins (Ed.), *Early years in machine translation: Memoirs and biographies of pioneers* (pp. 299-312). John Benjamins. https://doi.org/10.1075/sihols.97.26hut

Hutchins, W. J. (2001). Machine translation over fifty years. *Histoire épistémologie langage*, *23*(1), 7-31. https://doi.org/10.3406/hel.2001.2815

Hutchins, W. J. (2003). ALPAC: The (in)famous report. *Readings in machine translation, 14*, 131-135. https://doi.org/10.7551/mitpress/5779.003.0014

Hutchins, W. J. (2004). Two precursors of machine translation: Artsrouni and Trojanskij. *International Journal of Translation, 16*(1), 11-31.

Hutchins, W. J., & Lovtskii, E. (2000). Petr Petrovich Troyanskii (1894–1950): A forgotten pioneer of mechanical translation. *Machine translation*, *15*(3), 187-221. https://doi.org/10.1023/A:1011653602669

Iglesias, G., de Gisper, A., Banga, E. R., & Byrne, W. (2009). The HiFST system for the EuroParl Spanish-to-English task. *Procesamiento del lenguaje natural, 43*, 207-214.

Imamura, K. (2001). Hierarchical phrase alignment harmonized with parsing. *In Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium NLPRS 2001*, Tokyo, Japan. 377-384.

Israr, H., Khan, S.A., Tahir, M.A., Shahzad, M.K., Ahmad, M., & Zain, J. M. (2023). Neural machine translation models with attention-based dropout layer. *Computers, Materials & Continua, 75(2),* 2981–3009. https://doi.org/10.32604/cmc.2023.035814

Israr, H., Shahzad, M. K., & Anwar, S. (2024). Improved Urdu-English neural machine translation with a fully convolutional neural network encoder. *International Journal of Mathematical, Engineering & Management Sciences*, *9*(5), 1067-1088. https://doi.org/10.33889/IJMEMS.2024.9.5.056

Jeblee, S., Feely, W., Bouamor, H., Lavie, A., Habash, N., & Oflazer, K. (2014). Domain and dialect adaptation for machine translation into Egyptian Arabic. In N. Habash & S. Vogel (Eds.), *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)* (pp. 196–206). Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-3627

Johnson, M., Schuster, M., Le, Q.V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., & Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics, 5,* 339-351. https://doi.org/10.1162/tacl_a_00065

Jurafsky, D., & Martin, J. H. (2025). *Speech and language processing* (3rd ed.). Stanford University Press.

Kalchbrenner, N., & Blunsom, P. (2013). Recurrent continuous translation models. In D. Yarowsky et al. (Eds.), *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 1700–1709). Association for Computational Linguistics. https://aclanthology.org/D13-1176/

Kchaou, S., Boujelbane, R., & Belguith L. H. (2025). A deep neural network machine translation approach from a low-resource language to English. *Computación y Sistemas, 29*(2), 675-685. http://doi.org/10.13053/CyS-29-2-4705

Kchaou, S., Boujelbane, R., & Belguith, L. H. (2020). Parallel resources for Tunisian Arabic dialect translation. In I. Zitouni, M. Abdul-Mageed, H. Bouamor, F. Bougares, M. El-Haj, N. Tomeh & W. Zaghouani (Eds.), *Proceedings of the Fifth Arabic Natural Language*

Processing Workshop (pp. 200–206). Association for Computational Linguistics. https://aclanthology.org/2020.wanlp-1.18/

Kchaou, S., Boujelbane, R., & Hadrich, L. (2023). Hybrid pipeline for building Arabic Tunisian dialect-standard Arabic neural machine translation model from scratch. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *22*(3), 1-21. https://doi.org/10.1145/3568674

Keleg, A., Goldwater, S., & Magdy, W. (2025). Revisiting common assumptions about Arabic dialects in NLP. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 3309–3327). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.166

Kessentini, B., Alimi, T., Boujelben, R., & Belguith, L. H. (2025). Assessing BERT models for Arabic named entity recognition in a multi-dialectal context. In L. H. Belguith & K. Shaalan (Eds.), *Advancements in machine learning and natural language processing: Innovations and applications* (pp. 194-203). Springer. https://doi.org/10.1007/978-3-031-85067-7_19

Khalifa, S., Hassan, S., & Habash, N. (2017). A morphological analyzer for Gulf Arabic verbs. In N. Habash, M. Diab, K. Darwish, W. El-Hajj, H. Al-Khalifa, H. Bouamor, N. Tomeh, M. El-Haj, & W. Zaghouani (Eds.), *Proceedings of the Third Arabic Natural Language Processing Workshop* (pp. 35–45). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1305

Khalifa, S., Zalmout, N., & Habash, N. (2016). YAMAMA: Yet another multi-dialect Arabic morphological analyzer. In H. Watanabe (Ed.), *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations* (pp. 223–227). The COLING 2016 Organizing Committee. https://aclanthology.org/C16-2047/

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In Y. Bengio & Y. LeCun (Eds.), *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015*. San Diego, CA, USA. 1-15. https://doi.org/10.48550/arXiv.1412.6980

Klein, G., Kim, Y., Deng, Y., Senellart, J., & Rush, A. M. (2017). OpenNMT: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*. https://doi.org/10.48550/arXiv.1701.02810

Koehn, P. (2020). *Neural machine translation*. Cambridge University Press. https://doi.org/10.1017/9781108608480

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R. Dyer, C. Bojar, O., Constantin, A., & Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. *Annual Meeting Association for Computational Linguistics, 2(*2). 177–180. https://doi.org/10.3115/1557769.1557821

Koehn, P., Och, F. J., & Marcu, D. (2003a). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1* (pp. 48–54). Association for Computational Linguistics. https://doi.org/10.3115/1073445.1073462

Koehn, P., Och, F. J., & Marcu, D. (2003b). *Statistical phrase-based translation*. University of Southern California Marina Del Rey Information Sciences Inst. https://doi.org/10.21236/ADA461156

Kurzweil, R. (1990). *The age of intelligent machines*. MIT Press.

Labaka, G., España-Bonet, C., Màrquez, L., & Sarasola, K. (2014). A hybrid machine translation architecture guided by syntax. *Machine Translation, 28*, 91-125. https://doi.org/10.1007/s10590-014-9153-0

Labed, Z. (2014). *Genealogical koineisation in Oran speech community: The case of young university Oranees*. [Doctoral Dissertation. University of Es-senia, Oran, Algeria]. https://theses.univ-oran1.dz/these.php?id=42201446t

Labed, Z. (2016). Lexical variation in Oran dialect. *Al-Kalim Journal, 1*(1), 1-23.

Lee, S., Lee, J., Moon, H., Park, C., Seo, J., Eo, S., Koo, S., & Lim, H. (2023). A survey on evaluation metrics for machine translation. *Mathematics, 11*(4), 1006. https://doi.org/10.3390/math11041006

Lehrberger, J., & Bourbeau, L. (1988). *Machine translation: Linguistic characteristics of MT systems and general methodology of evaluation* (Vol. 15). John Benjamins Publishing. https://doi.org/10.1075/lis.15

Lenci, A., & Padó, S. (2022). Perspectives for natural language processing between AI, linguistics and cognitive science. *Frontiers in Artificial Intelligence, 5*, 1059998. https://doi.org/10.3389/frai.2022.1059998

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7871–7880). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.703

Li, H., & Lu, W. (2021). Mixed cross entropy loss for neural machine translation. In M. Meila & T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning* (Vol. 139, pp. 6425–6436). PMLR. https://proceedings.mlr.press/v139/li21n.html

Li, Z., Callison-Burch, C., Dyer, C., Ganitkevitch, J., Irvine, A., Khudanpur, S., Schwartz, L., Thornton, W., Wang, Z., Weese, J., & Zaidan, O. (2010). Joshua 2.0: A toolkit for parsing-based machine translation with syntax, semirings, discriminative training and other goodies. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR* (pp. 133-137). Association for Computational Linguistics. https://doi.org/10.3115/1626431.1626459

Lichouri, M., & Abbas, M. (2021). Machine translation for zero and low-resourced dialects using a new extended version of the dialectal parallel corpus (Padic v2. 0). In M. Abbas & A. A. Freihat (Eds.), *Proceedings of the 4th International Conference on Natural Language and Speech Processing (ICNLSP 2021)* (pp. 33–38). Association for Computational Linguistics. https://aclanthology.org/2021.icnlsp-1.4/

Lichouri, M., Abbas, M., & Lounnas, K. (2023). Impact of parallel vs. non parallel corpora on the identification of Arabic dialects. *OSF Preprint.* https://doi.org/10.31730/osf.io/grxt6

Lichouri, M., Abbas, M., Freihat, A. A., & Megtouf, D. E. H. (2018). Word-level vs sentence-level language identification: Application to Algerian and Arabic dialects. *Procedia Computer Science*, *142*, 246-253. https://doi.org/10.1016/j.procs.2018.10.484

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., & Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics, 8,* 726–742. https://doi.org/10.1162/tacl_a_00343

Liu, Y., Liu, Q., & Lin, S. (2006). Tree-to-string alignment template for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics* (pp. 609–616). Association for Computational Linguistics. https://doi.org/10.3115/1220175.1220252

Locke, W. N., & Booth, A. D. (Eds.). (1955). *Machine translation of languages: Fourteen essays*. MIT Press.

Maamouri, M., Buckwalter, T., & Cieri, C. (2004). Dialectal Arabic telephone speech corpus: Principles, tool design, and transcription conventions. In *NEMLAR International Conference on Arabic Language Resources and Tools*, Cairo, Egypt. 1-6.

Marcu, D., & Wong, W. (2002). A phrase-based, joint probability model for statistical machine translation. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10* (pp. 133–139). Association for Computational Linguistics. https://doi.org/10.3115/1118693.1118711

Marino, J. B., Banchs, R. E., Crego, J. M., de Gispert, A., Lambert, P., Fonollosa, J. A., & Costa-jussà, M. R. (2006). N-gram-based machine translation. *Computational Linguistics, 32*(4), 527-549. https://doi.org/10.1162/coli.2006.32.4.527

McCarthy, J. (1956). Measures of the value of information. *Proceedings of the National Academy of Sciences, 42*(9), 654-655. https://doi.org/10.1073/pnas.42.9.654

Meftouh, K., Bouchemal, N., & Smaïli, K. (2012). A study of a non-resourced language: An Algerian dialect. In *The third International Workshop on Spoken Language Technologies for Under-resourced Languages SLTU'12*, Cape Town, South Africa. 125-132. ISCA.

Meftouh, K., Harrat, S., & Smaïli, K. (2018). PADIC: Extension and new experiments. In I. Saritas & O. F. Bay (Eds.), *Proceedings of the 7th International Conference on Advanced Technologies ICAT'18* (pp. 559-564). SN Bilgi Teknolojileri.

Meftouh, K., Harrat, S., Jamoussi, S., Abbas, M., & Smaïli, K. (2015). Machine translation experiments on PADIC: A parallel Arabic dialect corpus. In H. Zhao (Ed.), *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation* (pp. 26–34). Association for Computational Linguistics. https://aclanthology.org/Y15-1004/

Mezzoudj, F., Loukam, M., & Belkredim, F. Z. (2019). Arabic Algerian Oranee dialectal language modelling oriented topic. *International Journal of Informatics and Applied Mathematics, 2*(2), 1-14.

Mitkov, R. (Ed.). (2022). *The Oxford handbook of computational linguistics*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199573691.001.0001

Mohamed, E., Mohit, B., & Oflazer, K. (2012). Transforming standard Arabic to colloquial Arabic. In H. Li, C.-Y. Lin, M. Osborne, G. G. Lee, & J. C. Park (Eds.), *Proceedings of the 50th*

Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 176–180). Association for Computational Linguistics. https://aclanthology.org/P12-2035/

Molchanov, A. (2012). PROMT DeepHybrid system for WMT12 shared translation task. In C. Callison-Burch, P. Koehn, C. Monz, M. Post, R. Soricut, & L. Specia (Eds.), *Proceedings of the Seventh Workshop on Statistical Machine Translation* (pp. 345–348). Association for Computational Linguistics. https://aclanthology.org/W12-3143/

Mondal, S. K., Zhang, H., Kabir, H. D., Ni, K., & Dai, H. N. (2023). Machine translation and its evaluation: A study. *Artificial Intelligence Review*, *56*(9), 10137-10226. https://doi.org/10.1007/s10462-023-10423-5

Morsly, D. (1986). Multilingualism in Algeria. In J.A. Fishman (Ed.), *The Fergusonian impact: In honor of Charles A. Ferguson on the occasion of his 65th birthday* (pp. 817-828). De Gruyter Mouton. https://doi.org/10.1515/9783110873641-060

Moudjari, L., Akli-Astouati, K., & Benamara, F. (2020). An Algerian corpus and an annotation platform for opinion and emotion analysis. In N. Calzolari et al. (Eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 1202–1210). European Language Resources Association. https://aclanthology.org/2020.lrec-1.151/

Moukafih, Y., Sbihi, N., Ghogho, M., & Smaïli, K. (2022). Improving machine translation of Arabic dialects through multi-task learning. In S. Bandini, F. Gasparini, V. Mascardi, M. Palmonari, & G. Vizzari (Eds.), *AIxIA 2021 – Advances in Artificial Intelligence. Lecture Notes in Computer Science, Vol 13196* (pp. 580-590). Springer. https://doi.org/10.1007/978-3-031-08421-8_40

Moulay-Lahssan, B. E., & Kerras, N. (2016). A sociolinguistic study of the Algerian language. *Arab World English Journal (AWEJ) Special Issue on CALL*, (3). 141-154. https://doi.org/10.2139/ssrn.2822984

Mueller, J. P., & Massaron, L. (2018). *Artificial intelligence for dummies.* John Wiley & Sons.

Munkova, D., Hajek, P., Munk, M., & Skalka, J. (2020). Evaluation of machine translation quality through the metrics of error rate and accuracy. *Procedia Computer Science*, *171*, 1327-1336. https://doi.org/10.1016/j.procs.2020.04.142

Nagao, M. (1984). A framework of a mechanical translation between Japanese and English by analogy principle. *Artificial and human intelligence*, 351-354.

Nagao, M. (2003). A framework of a mechanical translation between Japanese and English by analogy principle. In S. Nirenburg, H. L. Somers, & Y. A. Wilks (Eds.), *Readings in machine translation* (pp. 351-354). The MIT Press. https://doi.org/10.7551/mitpress/5779.003.0038

Nagoudi, E. M. B., Elmadany, A., & Abdul-Mageed, M. (2021). Investigating code-mixed modern standard Arabic–Egyptian to English machine translation. In T. Solorio, S. Chen, A. W. Black, M. Diab, S. Sitaram, V. Soto, E. Yilmaz, & A. Srinivasan (Eds.), *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching* (pp. 56–64). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.calcs-1.8

Neapolitan, R. E., & Jiang, X. (2018). *Artificial intelligence: With an introduction to machine learning* (2nd ed.). CRC Press, Taylor & Francis Group. https://doi.org/10.1201/b22400

Negadi, M. N. (2015). Investigating language contact situation in Algeria. *Traduction et Langues*, *14*(1), 145-150. https://doi.org/10.52919/translang.v14i1.797

Neubig, G. (2013). Travatar: A forest-to-string machine translation engine based on tree transducers. In M. Butt & S. Hussain (Eds.), *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 91–96). Association for Computational Linguistics. https://aclanthology.org/P13-4016/

Neubig, G. (2017). Neural machine translation and sequence-to-sequence models: A tutorial. *arXiv preprint arXiv:1703.01619*. https://doi.org/10.48550/arXiv.1703.01619

Nirenburg, S. (1989). Knowledge-based machine translation. *Machine Translation*, *4*, 5-24. https://doi.org/10.1007/BF00367750

Nirenburg, S., & Wilks, Y. (2000). Machine translation. *Advances in Computers*, *52*, 159-188. https://doi.org/10.1016/S0065-2458(00)80018-2

Nirenburg, S., Beale, S., & Domashnev, C. (1994). A full-text experiment in example-based machine translation. In *International Conference on New Methods in Language Processing (NeMLaP)* (Vol. 94, pp. 78-87).

Nirenburg, S., Carbonell, J., & Tomita, M. (1994). *Machine translation: A knowledge-based approach*. Morgan Kaufmann Publishers Inc.

Obeid, O., Inoue, G., & Habash, N. (2022). Camelira: An Arabic multi-dialect morphological disambiguator. In W. Che & E. Shutova (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 319–326). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-demos.32

Och, F. J., & Ney, H. (2000). Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics* (pp. 440–447). Association for Computational Linguistics. https://doi.org/10.3115/1075218.1075274

Och, F. J., & Ney, H. (2002). Discriminative training and maximum entropy models for statistical machine translation. In P. Isabelle, E. Charniak, & D. Lin (Eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 295–302). Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073133

Panov, D. Y., Lyapunov, A. A., & Mukhin, I. S. (1956). *Automation of translation from one language to another*. CCCP, Moscow.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2002). BLEU: A method for automatic evaluation of machine translation. In P. Isabelle, E. Charniak, & D. Lin (Eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

Pasha, A., Al-Badrashiny, M., Diab, M. T., El Kholy, A., Eskander, R., Habash, N., Pooleery, M., Rambow, O., & Roth, R. (2014). MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 1094–1101). European Language Resources Association (ELRA). https://aclanthology.org/L14-1479/

Pierce, J. R. (Ed.). (1966). *Languages and machines: Computers in translation and linguistics*. National Academy of Sciences. https://doi.org/10.17226/9547

Poibeau, T. (2017). *Machine translation*. MIT Press. https://doi.org/10.7551/mitpress/11043.001.0001

Pollack, J. B. (1990). Recursive distributed representations. *Artificial Intelligence*, *46*(1-2), 77-105. https://doi.org/10.1016/0004-3702(90)90005-K

Poole, D., Mackworth, A. K., & Goebel, R. (1998). *Computational intelligence: A logical approach*. Oxford University Press.

Post, M., Cao, Y., & Kumar, G. (2015). Joshua 6: A phrase-based and hierarchical statistical machine translation system. *Prague Bulletin of Mathematical Linguistics*, *104*, 5-16. https://doi.org/10.1515/pralin-2015-0009

Pugh, J. (1992). The story so far: An evaluation of machine translation in the world today. In J. Newton (Ed.), *Computers in translation: A practical appraisal* (pp. 34-52). Routledge.

Puspitaningrum, D. (2021). A study of English-Indonesian neural machine translation with attention (Seq2Seq, ConvSeq2Seq, RNN, and MHA): A comparative study of NMT on English-Indonesian. In *Proceedings of the 6th International Conference on Sustainable Information Engineering and Technology* (pp. 271–280). Association for Computing Machinery. https://doi.org/10.1145/3479645.3479703

Queffélec, A., Derradji, Y., Debov, V., Smaali-Dekdouk, D., & Cherrad-Benchefra, Y. (2002). *Le français en Algérie: Lexique et dynamique des langues*. De Boeck Supérieur. https://doi.org/10.3917/dbu.cherr.2002.01

Rahab, H., Zitouni, A., & Djoudi, M. (2021). SANA: Sentiment analysis on newspapers comments in Algeria. *Journal of King Saud University-Computer and Information Sciences*, *33*(7), 899-907. https://doi.org/10.1016/j.jksuci.2019.04.012

Ramírez-Sánchez, G., Sánchez-Martínez, F., Ortiz-Rojas, S., Pérez-Ortiz, J. A., & Forcada, M. L. (2006). Opentrad Apertium open-source machine translation system: An opportunity for business and research. In *Proceedings of the 28th Conference on Translating and the Computer*, London, UK. Aslib. https://aclanthology.org/2006.tc-1.12/

Ranta, A. (2011). *Grammatical framework: Programming with multilingual grammars* (Vol. 173). Center for the Study of Language and Information. https://doi.org/10.5555/2049998

Richens, R. H., & Booth, A. D. (1955). Some methods of mechanized translation. In W. N. Locke & A. D. Booth (Eds.), *Machine translation of languages: Fourteen essays* (pp. 24-46). MIT Press.

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: A modern approach.* (3rd ed.). Pearson Education Limited.

Saadane, H., & Habash, N. (2015). *A conventional orthography for Algerian Arabic*. In N. Habash, S. Vogel, & K. Darwish (Eds.), *Proceedings of the Second Workshop on Arabic Natural Language Processing* (pp. 69–79). Association for Computational Linguistics. https://doi.org/10.18653/v1/W15-3208

Sadat, F. (2015). Multi-dialect machine translation (MuDMaT). In I. D. El-Kahlout, M. Özkan, F. Sánchez-Martínez, G. Ramírez-Sánchez, F. Hollowood, & A. Way (Eds.), *Proceedings of the 18th Annual Conference of the European Association for Machine Translation* (p. 226). https://aclanthology.org/W15-4943/

Sadat, F., Kazemi, F., & Farzindar, A. (2014). Automatic identification of Arabic dialects in social media. In *Proceedings of the First International Workshop on Social Media Retrieval and*

Analysis (pp. 35–40). Association for Computing Machinery. https://doi.org/10.1145/2632188.2632207

Saeed, M., & Habash, N. (2025). Lemmatizing dialectal Arabic with sequence-to-sequence models. In K. Darwish et al. (Eds.), *Proceedings of The Third Arabic Natural Language Processing Conference* (pp. 117–129). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.arabicnlp-main.10

Sager, J. C. (1994). *Language engineering and translation: Consequences of automation*. John Benjamins Publishing. https://doi.org/10.1075/btl.1

Sajjad, H., Darwish, K., & Belinkov, Y. (2013). Translating dialectal Arabic to English. In H. Schuetze, P. Fung, & M. Poesio (Eds.), *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 1–6). Association for Computational Linguistics. https://aclanthology.org/P13-2001/

Sajjad, H., Durrani, N., Guzman, F., Nakov, P., Abdelali, A., Vogel, S., Salloum, W., El Kholy. A., & Habash, N. (2016). Egyptian Arabic to English statistical machine translation system for NIST OpenMT'2015. *arXiv preprint arXiv:1606.05759*. 1-10. https://doi.org/10.48550/arXiv.1606.05759

Salloum, W., & Habash, N. (2011). Dialectal to standard Arabic paraphrasing to improve Arabic-English statistical machine translation. In J. Jancsary, F. Neubarth, & H. Trost (Eds.), *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties* (pp. 10–21). Association for Computational Linguistics. https://aclanthology.org/W11-2602/

Salloum, W., & Habash, N. (2012). Elissa: A dialectal to Standard Arabic machine translation system. In M. Kay & C. Boitet (Eds.), *Proceedings of COLING 2012: Demonstration Papers* (pp. 385–392). The COLING 2012 Organizing Committee. https://aclanthology.org/C12-3048/

Salloum, W., & Habash, N. (2013). Dialectal Arabic to English machine translation: Pivoting through modern standard Arabic. In L. Vanderwende, H. Daumé III, & K. Kirchhoff (Eds.), *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 348–358). Association for Computational Linguistics. https://aclanthology.org/N13-1036/

Salloum, W., & Habash, N. (2014). ADAM: Analyzer for dialectal Arabic morphology. *Journal of King Saud University-Computer and Information Sciences*, *26*(4), 372-378. https://doi.org/10.1016/j.jksuci.2014.06.010

Sawaf, H. (2010). Arabic dialect handling in hybrid machine translation. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers* (pp. 222-227). Association for Machine Translation in the Americas. https://aclanthology.org/2010.amta-papers.5/

Scott, B., & Barreiro, A. (2009). OpenLogos MT and the SAL representation language. In J. A. Pérez-Ortiz, F. Sánchez-Martínez, & F. M. Tyers (Eds.), *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation* (pp. 19–26). Association for Computational Linguistics. https://aclanthology.org/2009.freeopmt-1.5/

Senez, D. (1994). Developments in SYSTRAN. In *Proceedings of Translating and the Computer 16* (pp. 65–77). Aslib. https://aclanthology.org/1994.tc-1.7/

Sennrich, R. (2012). Perplexity minimization for translation model domain adaptation in statistical machine translation. In W. Daelemans (Ed.), *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 539–549). Association for Computational Linguistics. https://aclanthology.org/E12-1055/

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. In K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics ACL 2016, Volume 1: Long Papers*, (pp. 1715–1725). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1162

Sghaier, M. A., & Zrigui, M. (2020). Rule-based machine translation from Tunisian dialect to modern standard Arabic. *Procedia Computer Science*, *176*, 310-319. https://doi.org/10.1016/j.procs.2020.08.033

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of information*. University of Illinois Press.

Sharma, S., Diwakar, M., Singh, P., Singh, V., Kadry, S., & Kim, J. (2023). Machine Translation Systems Based on Classical-Statistical-Deep-Learning Approaches. *Electronics*, *12*(7), 1716. https://doi.org/10.3390/electronics12071716

Slim, A., & Melouah, A. (2024). Low resource Arabic dialects transformer neural machine translation improvement through incremental transfer of shared linguistic

features. *Arabian Journal for Science and Engineering, 49,* 1-17. https://doi.org/10.1007/s13369-023-08543-9

Slim, A., Melouah, A., Faghihi, U., & Sahib, K. (2022). Improving neural machine translation for low resource Algerian dialect by transductive transfer learning strategy. *Arabian Journal for Science and Engineering*, *47*(8), 10411-10418. https://doi.org/10.1007/s13369-022-06588-w

Slim, A., Melouah, A., Faghihi, Y., & Sahib, K. (2021). Algerian dialect translation applied on COVID-19 social media comments. In M. Hatti (Ed.), *Artificial Intelligence and Renewables Towards an Energy Transition. ICAIRES 2020. Lecture Notes in Networks and Systems, Vol 174.* (pp. 716-726). Springer. https://doi.org/10.1007/978-3-030-63846-7_68

Slocum, J. (1980). An experiment in machine translation. In *18th Annual Meeting of the Association for Computational Linguistics* (pp. 163–164). Association for Computational Linguistics. https://doi.org/10.3115/981436.981487

Slocum, J. (1985). A survey of machine translation: Its history, current status and future prospects. *Computational linguistics*, *11*(1), 1-17.

Somers, H. (1999). Review article: Example-based machine translation. *Machine Translation*, *14*(2), 113–157. https://doi.org/10.1023/A:1008109312730

Somers, H. (Ed.). (2003). *Computers and translation: A translator's guide*. John Benjamins Publishing. https://doi.org/10.1075/btl.35

Somers, H. L. (1992). Current research in machine translation. *Machine Translation*, *7*, 231-246. https://doi.org/10.1007/BF00398467

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

Soumeur, A., Mokdadi, M., Guessoum, A., & Daoud, A. (2018). Sentiment analysis of users on social networks: Overcoming the challenge of the loose usages of the Algerian dialect. *Procedia Computer Science*, *142*, 26-37. https://doi.org/10.1016/j.procs.2018.10.458

Stahlberg, F. (2020). Neural machine translation: A review. *Journal of Artificial Intelligence Research*, *69*, 343-418. https://doi.org/10.1613/jair.1.12007

Suleiman, Y. (2013). *Arabic sociolinguistics: Issues and perspectives*. Routledge. https://doi.org/10.4324/9781315026190

Sutskever, I., Martens, J., & Hinton, G. (2011). Generating text with recurrent neural networks. In L. Getoor, & T. Scheffer (Eds.), *Proceedings of the 28th International Conference on International Conference on Machine Learning* (pp. 1017–1024). Omnipress. https://doi.org/10.5555/3104482.3104610

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Proceedings of the 28th International Conference on Neural Information Processing Systems NIPS'14* (Vol. 2, pp. 3104–3112). MIT Press. https://doi.org/10.48550/arXiv.1409.3215

Tachicart, R., & Bouzoubaa, K. (2014). A hybrid approach to translate Moroccan Arabic dialect. In *2014 9th International Conference on Intelligent Systems: Theories and Applications (SITA-14)* (pp. 1-5). IEEE. https://doi.org/10.1109/SITA.2014.6847293

Tachicart, R., Bouzoubaa, K., & Jaafar, H. (2014). Building a Moroccan dialect electronic dictionary (MDED). In *Proceedings of the 5th International Conference on Arabic Language Processing*, Morocco. 216-221.

Tachicart, R., Bouzoubaa, K., Harrat, S., Smaïli, K. (2022). Arabic dialects morphological analyzers: A survey. In M. Al-Emran & K. Shaalan (Eds.), *Recent Innovations in Artificial Intelligence and Smart Applications. Studies in Computational Intelligence, Vol 1061* (pp. 189–203). Springer, Cham. https://doi.org/10.1007/978-3-031-14748-7_11

Tan, Z., Wang, S., Yang, Z., Chen, G., Huang, X., Sun, M., & Liu, Y. (2020). Neural machine translation: A review of methods, resources, and tools. *AI Open*, *1*, 5-21. https://doi.org/10.1016/j.aiopen.2020.11.001

Tawfik, A., Emam, M., Essam, K., Nabil, R., & Awadalla, H. H. (2019). Morphology-aware word-segmentation in dialectal Arabic adaptation of neural machine translation. In W. El-Hajj, L. H. Belguith, F. Bougares, W. Magdy, I. Zitouni, N. Tomeh, M. El-Haj, & W. Zaghouani (Eds.), *Proceedings of the Fourth Arabic Natural Language Processing Workshop* (pp. 11–17). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4602

Toma, P. (1977). SYSTRAN as a multilingual machine translation system. In European Congress on Information Systems and Networks (Ed.), *Proceedings of the Third European Congress on Information Systems and Networks, Overcoming the Language Barrier* (pp. 569-581). Verlag Dokumentation.

Torjmen, R., & Haddar, K. (2021). The automatic recognition and translation of Tunisian dialect named entities into modern standard Arabic. In B. Bekavac, K. Kocijan, M. Silberztein, & K. Šojat (Eds.), *Formalising Natural Languages: Applications to Natural Language*

Processing and Digital Humanities: NooJ 2020. Communications in Computer and Information Science, Vol 1389 (pp. 206-217). Springer. https://doi.org/10.1007/978-3-030-70629-6_18

Torjmen, R., & Haddar, K. (2022). Translation system from Tunisian Dialect to modern standard Arabic. *Concurrency and Computation: Practice and Experience*, *34*(6), e6788. 1-16. https://doi.org/10.1002/cpe.6788

Yousuf, H., Lahzi, M., Salloum, S. A., & Shaalan, K. (2021). A systematic review on sequence-to-sequence learning with neural network and its models. *International Journal of Electrical & Computer Engineering*, *11*(3), 2315-2326. http://doi.org/10.11591/ijece.v11i3.pp2315-2326

van der Wees, M., Bisazza, A., & Monz, C. (2016). A simple but effective approach to improve Arabizi-to-English statistical machine translation. In B. Han, A. Ritter, L. Derczynski, W. Xu, & T. Baldwin (Eds.), *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)* (pp. 43–50). The COLING 2016 Organizing Committee. https://aclanthology.org/W16-3908/

Vandeghinste, V., Martens, S., Kotzé, G., Tiedemann, J., Van den Bogaert, J., De Smet, K., Van Eynde, F., & van Noord, G. (2013). Parse and corpus-based machine translation. In P. Spyns & J. Odijk (Eds.), *Essential speech and language technology for Dutch: Results by the STEVIN programme* (pp. 305–319). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-30910-6_17

Vasconcellos, M. (2000). The Georgetown project and Leon Dostert: Recollections of a young assistant. In W. J. Hutchins (Ed.), *Early years in machine translation: Memoirs and*

biographies of pioneers (pp. 87-96). John Benjamins. https://doi.org/10.1075/sihols.97.09vas

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 6000–6010). Curran Associates Inc. https://doi.org/10.5555/3295222.3295349

Vauquois, B. (1968). A survey of formal grammars and algorithms for recognition and transformation in mechanical translation. In *IFIP Congress (2)* (Vol. 68, pp.1114-1122). https://hal.science/hal-04701802v1

Versteegh, K. (2014). *The Arabic language*. Edinburgh University Press. https://doi.org/10.1515/9780748645299

Wahlster, W. (1993). *Verbmobil: Translation of face-to-face dialogs* In O. Herzog, T. Christaller, & D. Schütt (Eds.), *Grundlagen und Anwendungen der Künstlichen Intelligenz* (pp. 393–402). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-78545-0_37

Watanabe, T., Imamura, K., & Sumita, E. (2002). Statistical machine translation based on hierarchical phrase alignment. In *Proceedings of the 9th Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages: Papers*. https://aclanthology.org/2002.tmi-papers.20/

Weaver, W. (1949). The mathematics of communication. *Scientific American*, *181*(1), 11-15. https://doi.org/10.1038/scientificamerican0749-11

Weaver, W. (1970). *Scene of change: A lifetime in American science*. Scribner.

Whitelock, P., & Kilby, K. (1995). *Linguistic and computational techniques in machine translation system design*. UCL Press.

Wilks, Y. (2008). *Machine translation: Its scope and limits*. Springer Science & Business Media.

Wilss, W. (1999). *Knowledge and skills in translator behavior*. John Benjamins Publishing.

Winston, P. H. (1992). *Artificial Intelligence.* Addison-Wesley Publishing.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., St. Amant, Y., Liu, Y., Johnson, M., Luong, T., Soricut, R., Mané, D., Schalkwyk, C., Palmer, J., ... Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation *arXiv preprint arXiv:1609.08144.* (2). 1-23. https://arxiv.org/abs/1609.08144

Xiong, D., & Zhang, M. (2016). *Linguistically motivated statistical machine translation*. Springer Verlag, Singapore. https://doi.org/10.1007/978-981-287-356-9

Yafooz, W. M. (2024). Enhancing Arabic dialect detection on social media: A hybrid model with an attention mechanism. *Information, 15*(6), 316. https://doi.org/10.3390/info15060316

Yamada, K., & Knight, K. (2001). A syntax-based statistical translation model. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics* (pp. 523-530). https://doi.org/10.3115/1073012.1073079

Yngve, V. H. (2000). Early research at M.I.T.: In search of adequate theory. In W. J. Hutchins (Ed.), *Early years in machine translation: Memoirs and biographies of pioneers* (pp. 39-72). John Benjamins. https://doi.org/10.1075/sihols.97.07yng

Younes, J., Souissi, E., Achour, H., & Ferchichi, A. (2020). Language resources for Maghrebi Arabic dialects' NLP: A survey. *Language Resources and Evaluation*, *54*, 1079-1142. https://doi.org/10.1007/s10579-020-09490-9

Zaidan, O. F., & Callison-Burch, C. (2014). Arabic dialect identification. *Computational Linguistics*, *40*(1), 171-202. https://doi.org/10.1162/COLI_a_00169

Zakraoui, J., Saleh, M., Al-Maadeed, S., & Alja'am, J. M. (2021). Arabic machine translation: A survey with challenges and future directions. *IEEE Access*, *9*, 161445-161468. https://doi.org/10.1109/ACCESS.2021.3132488

Zalmout, N., & Habash, N. (2019). Joint diacritization, lemmatization, normalization, and fine-grained morphological tagging. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8297–8307). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.736

Zbib, R., Malchiodi, E., Devlin, J., Stallard, D., Matsoukas, S., Schwartz, R., Makhoul, J., Zaidan, O.F., & Callison-Burch, C. (2012). Machine translation of Arabic dialects. In E. Fosler-Lussier, E. Riloff, & S. Bangalore (Eds.), *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 49–59). Association for Computational Linguistics. https://aclanthology.org/N12-1006/

Zhirkov, L. I. (1956). Granitsy primenimosti mašinnogo perevoda [The Limits of Applicability of Machine Translation]. *Voprosy Jazykoznanija*, *5*(5), 121-124.

Zribi, I., Ellouze Khemakhem, M., & Hadrich Belguith, L. (2013). Morphological analysis of Tunisian dialect. In R. Mitkov & J. C. Park (Eds.), *Proceedings of the Sixth International Joint Conference on Natural Language Processing* (pp. 992–996). Asian Federation of Natural Language Processing. https://aclanthology.org/I13-1133/

Abstract in Arabic

تُمثّل الترجمة الآلية العصبية للّهجات العربية تحديًا كبيرًا ضمن معالجة اللغات الطبيعية بسبب النقص الحاد في الموارد الرقمية .تظهر هذه المشكلة بحدة خاصة بالنسبة للسلسلة اللغوية المغاربية، حيث تعاني اللهجات الجزائرية مثل اللهجة الوهرانية من قصور تكنولوجي. تتصدى هذه الأطروحة لهذا الفراغ المطلق من خلال معالجة المشكلة المزدوجة المتمثلة في غياب الموارد والافتقار إلى معايير الترجمة الآلية لزوج اللغة الوهرانية-الإنجليزية. تُفصّل الدراسة عملية إنشاء مدونة وهرانية-إنجليزية متوازية جديدة، وهي مجموعة بيانات تحتوي على 8.5 ألف زوج من الجمل تم تجميعها لتصميم وتدريب نظام ترجمة آلية عصبية من نوع سلسلة-إلى-سلسلة .ويعتمد هيكل النموذج على شبكة ذاكرة طويلة-قصيرة المدى من طبقتين، معززة بشكل حاسم بآلية انتباه جمعي. حقق النموذج النهائي درجة BLEU بلغت 30.9 متجاوزًا بشكل كبير خط الأساس الذي لا يحتوي على آلية انتباه. والجدير بالذكر أن هذه النتيجة تنافسية للغاية، حيث تتفوق حتى على الأنظمة الأكبر حجمًا وشبه الموجهة للّهجات العربية الأخرى، مما يبرز تأثير المدونة عالية الجودة والمتسقة لغويًا. يؤكد تحليل الأخطاء النوعي كفاءة النموذج في التعامل مع التراكيب النحوية الأساسية، بينما يحدد التحديات الكلاسيكية المتعلقة بالموارد المنخفضة مثل الكلمات غير الموجودة في المفردات، والتعابير الاصطلاحية، والغموض الدلالي . تُثبِت النتائج صحة استراتيجية "العمق مقابل الاتساع" للمدونة كنهج فعال لبدء البحث في اللهجات عديمة الموارد . يقدم هذا العمل مخططًا منهجيًا للأعمال المستقبلية، مشيرًا إلى التوجه نحو هياكل المحولات والتعلم شبه الموجه.

الكلمات المفتاحية :الترجمة الآلية العصبية، اللغات منخفضة الموارد، العربية الجزائرية، اللهجة الوهرانية، ، سلسلة -إلى- سلسلة.