Lahjawi: Arabic Cross-Dialect Translator

Mohamed Motasim Hamed, Muhammad Hreden, Khalil Hennara
Zeina Aldallal, Sara Chrouf, Safwan AlModhayan
Misraj AI
Khobar, Saudi Arabia
{hamed,hreden,hennara,aldallal,sara.chrouf,safwan}@misraj.ai

Abstract

In this paper, we explore the rich diversity of Arabic dialects by introducing a suite of pioneering models called Lahjawi. The primary model, Lahjawi-D2D, is the first designed for cross-dialect translation among 15 Arabic dialects. Furthermore, we introduce Lahjawi-D2MSA, a model designed to convert any Arabic dialect into Modern Standard Arabic (MSA). Both models are fine-tuned versions of ***Kuwain-1.5B***[1] an in-house built small language model, tailored for Arabic linguistic characteristics. We provide a detailed overview of Lahjawi's architecture and training methods, along with a comprehensive evaluation of its performance. The results demonstrate Lahjawi's success in preserving meaning and style, with BLEU scores of 9.62 for dialect-to-MSA and 9.88 for dialect-to-dialect tasks. Additionally, human evaluation reveals an accuracy score of 58% and a fluency score of 78%, underscoring Lahjawi's robust handling of diverse dialectal nuances. This research sets a foundation for future advancements in Arabic NLP and cross-dialect communication technologies.

1 Introduction

Arabic is the official language of 22 countries, with an estimated 400 million speakers globally (Mohammed Ameen and Abdulrahman Kadhim, 2023), It stands out as one of the world's most linguistically rich. With more than 120 morphological patterns (Shaalan et al., 2019), Arabic offers a multitude of word formations that significantly amplify its expressive capacity. In everyday communication, Arabs primarily use dialects, which vary significantly across countries and regions, posing challenges for cross-dialect communication, particularly in informal contexts.

The importance of Arabic dialect translation has grown significantly over the last decade, driven by increasing demand for digital communication and cultural exchanges. While early research focused on Modern Standard Arabic (MSA) translation, the need for comprehensive cross-dialectal translation has recently gained attention due to the language's rich diversity. This diversity presents substantial challenges, including significant vocabulary disparities (see Table 1), varying sentence structures, and region-specific idiomatic expressions like folk proverbs. Additionally, grammatical differences in verb conjugations and plural forms further increase complexity. Despite advancements in Arabic Natural Language Processing (NLP), several challenges persist:

- *Lack of Cross-Dialect Translation Models*: lack of models addressing the dialect-to-dialect translation.
- *Absence of Comprehensive Solutions*: Current models fail to provide a holistic approach that addresses the full spectrum of Arabic dialectal diversity and translation needs.

To address these challenges, we present Lahjawi, a set of dialect translation models designed to address the challenges of cross-dialect communication in Arabic. Our key contribution, Lahjawi-D2D, is the first model developed for cross-dialect translation, covering 15 distinct Arabic dialects. Additionally, we introduce Lahjawi-D2MSA, which translates any Arabic dialect into Modern Standard Arabic (MSA). This work advances Arabic dialect translation and contributes to the broader goal of enhancing inclusivity and linguistic diversity in NLP.

The remainder of this paper is organized as follows: Section 2 reviews related works, Section 3 details our dataset creation steps, Section 4 presents our model and the proposed method, Sec-

[1] ***Kuwain-1.5B* (كُوَيْن)**: *an in-house built small language model designed to address the unique linguistic characteristics of Arabic.*

Proceedings of The 4th Workshop on Arabic Corpus Linguistics (WACL-4), pages 12–24
January 20th, 2025. ©2024 Association for Computational Linguistics

MSA	Levantine Arabic	Egyptian Arabic	Translation
كيف حالك؟	كيفك؟	إزيك؟	How are you?
أريد الذهاب للمنزل	بدي أروح عالبيت	عايز أروح البيت	I want to go home
ماذا يحدث؟	شو عم بصير؟	إيه اللي بيحصل؟	What's happening?

Table 1: Examples of dialectal variations in Arabic

tion 5 outlines our experimental setup. Section 6 discusses the findings, interprets them to existing research, and explores their broader implications. Section 7 acknowledges the approach limitations and suggests directions for future research. Through this structured approach, we deliver an in-depth analysis of Lahjawi's capabilities, highlighting its potential impact on Arabic NLP and cross-dialectal communication.

2 Related Work

Recent advancements in dialect translation research have been explored in various dimensions, with notable efforts focusing on translation from individual dialects to Modern Standard Arabic (MSA) and translation involving multiple dialects into MSA (AlMusallam and Ahmad, 2024). The former involves converting a specific dialect into MSA, aiming for precise linguistic alignment between regional speech and formal Arabic. The latter examines the translation of multiple dialects into MSA, offering broader applicability across diverse dialectal variations and enhancing mutual intelligibility. Beyond these, cross-dialect translation involves translating texts from one specific Arabic dialect directly into another, bypassing the need for Modern Standard Arabic (MSA) as an intermediary. This approach is particularly relevant for improving communication between speakers of different dialects. However, despite its practical importance, research in this area remains limited. This may be because most other languages do not exhibit the same level of dialectal variation as Arabic. As a result, cross-dialect translation is a challenge unique to Arabic and a few other languages, which might explain the relatively limited attention it has received from researchers. Consequently, only one foundational work (Meftouh et al., 2015) has addressed this underexplored domain.

2.1 Single Dialect Translation To MSA

Recent research in Arabic dialect translation has primarily focused on converting specific dialects to Modern Standard Arabic (MSA). Studies on Jordanian (Al-Ibrahim and Duwairi, 2020), Tunisian (Sghaier and Zrigui, 2020; Kchaou et al., 2020), and Egyptian (Faheem et al., 2024) dialects have highlighted various challenges and approaches. For instance, Jordanian-to-MSA translation has achieved high accuracy at both word and sentence levels (Al-Ibrahim and Duwairi, 2020), while Tunisian dialect translation has faced difficulties with longer, idiomatic phrases (Sghaier and Zrigui, 2020; Kchaou et al., 2020). Egyptian dialect research has emphasized the importance of both monolingual and parallel data in low-resource settings (Faheem et al., 2024). Multidialectal approaches, such as (Khered et al., 2023), have shown success in translating Egyptian, Emirati, Jordanian, and Palestinian dialects to MSA using separate models for each dialect.

Methodologies in this field have evolved from traditional rule-based systems to advanced Deep Learning techniques. Early rule-based machine translation (RBMT) systems (Sghaier and Zrigui, 2020) struggled with complex phrases, while statistical machine translation (SMT) (Kchaou et al., 2020) offered moderate improvements through data augmentation. Deep learning methods, particularly recurrent neural networks (RNNs) and transformer models, have shown superior accuracy. For example, RNN-based approaches (Al-Ibrahim and Duwairi, 2020) demonstrated high accuracy for Jordanian dialect translation, while transformer models (Torjmen and Haddar, 2024; Khered et al., 2023) significantly outperformed rule-based approaches. Semi-supervised approaches (Faheem et al., 2024) have effectively combined parallel and monolingual data, outperforming both supervised and unsupervised models in low-resource contexts.

2.2 Multiple Dialects Translation To MSA

Recent advancements in multi-dialect translation to MSA have centered on model fine-tuning, data augmentation, and applying large language models (LLMs). Fine-tuning pre-trained transformer models, particularly AraT5, has shown significant improvements in translation quality for various dialects including Palestinian, Jordanian, and Egyptian (AlMusallam and Ahmad, 2024; Alahmari, 2024; Derouich et al., 2023). Joint models trained on multiple dialects (Khered et al., 2023) have leveraged cross-dialectal information to achieve high performance.

Data augmentation and dataset expansion have been crucial strategies. Studies like (Nacar et al., 2024) and (Fares, 2024) have employed back-translation and incorporated multiple corpora to expand training data, leading to substantial improvements in translation performance. The introduction of novel datasets, such as SADA (Abdelaziz et al., 2024), created using automated translation methods with ChatGPT 3.5, has further enhanced model training.

The application of LLMs has shown great potential, especially in low-resource settings. Research utilizing models like GPT-3.5, AraT5, and No Language Left Behind (NLLB) (Atwany et al., 2024) has achieved high BLEU scores across multiple dialects. Notably, the Arabic Train Team demonstrated the superior performance of the Jais (Sengupta et al., 2023), an Arabic-focused model, which outperformed GPT-3.5, GPT-4, and NLLB in translating dialects into MSA (Demidova et al., 2024). Additionally, the fine-tuning of models like LLaMA-3 using Parameter-Efficient Fine-Tuning (PEFT) methods (Ibrahim, 2024) has demonstrated the effectiveness of resource-efficient approaches for complex dialect translations. These advancements underscore the growing impact of LLMs and the importance of dialect-specific datasets and efficient fine-tuning techniques in improving translation quality across Arabic dialects.

2.3 Cross-Dialect Translation

While most research in Arabic dialect translation has focused on converting dialects into Modern Standard Arabic (MSA), cross-dialect translation has received comparatively less attention. A notable exception is the work by (Meftouh et al., 2015), who introduced PADIC, a parallel corpus of five Arabic dialects from the Maghreb and the Middle East (Algerian, Tunisian, Syrian, and Palestinian). PADIC represents an early attempt at facilitating machine translation between dialects themselves. The study found that dialects from the same region, such as Algerian and Tunisian, achieved better translation accuracy due to their linguistic similarities. In contrast, dialects from different areas, like Syrian and Algerian, posed greater challenges due to their divergence. This groundbreaking work underscores both the potential and the current limitations of machine translation systems when applied to under-resourced Arabic dialects.

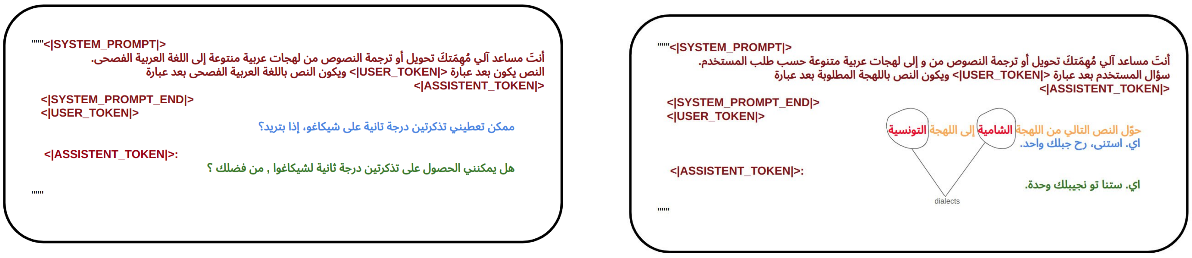
3 Dataset Preparation and Preprocessing

Our multi-dialect Arabic translation model was developed using a combination of open-source datasets: MADAR (Bouamor et al., 2018), PADIC (Meftouh et al., 2015), NADI (2023) (Abdul-Mageed et al., 2023), Dial2MSA (Mubarak, 2018), Arabic STS (Al Sulaiman et al., 2022), UFAL Parallel Corpus of North Levantine 1.0 (Sellat et al., 2023) and Multidialectal Parallel Corpus of Arabic(MDPCA) (Bouamor et al., 2014). These datasets were processed uniformly using two distinctive templates, with system prompts employed throughout, one for translating any dialect-to-MSA, and another for translating between specific dialects (see Figure 1).

Applying these templates results in two types of datasets: Dialect-to-MSA, and Dialect-to-Dialect datasets. The *Dialect-to-MSA (D2MSA)* dataset consists of 197,042 samples, which are used to train the Lahjawi-D2MSA models. Figure 2 shows the distribution of dialects within this dataset. As shown in the figure, the dataset exhibits significant dialect imbalance, with Syrian Arabic dominating at 66%, while other dialects have minimal representation ranging from 0.8% to 5.7%.

The *Dialect-to-Dialect (D2D)* dataset contains 266,871 samples and is used to train the Lahjawi-D2D model. This dataset was created by generating every possible combination of dialect pairs from all previously mentioned datasets, encompassing a wide range of dialect variations. The dataset includes 210 possible dialect translation pairs (see Figure 3). The dataset shows significant skewness in the number of samples for each pair across the 15 dialects, with an over-representation of Levantine dialects, specifically Syrian, Palestinian, and Jordanian, and Maghrebi dialects, particularly Tunisian, Moroccan, and Algerian.

We implemented a straightforward preprocessing pipeline to standardize the training data. This process includes the normalization of Arabic characters and numerals, as well as the standardization of punctuation and spacing. These preprocessing



```
"""<|SYSTEM_PROMPT|>
أنتَ مساعد آلي مُهِمَتكَ تحويل أو ترجمة النصوص من لهجات عربية منتوعة إلى اللغة العربية الفصحى.
النص يكون بعد عبارة <|USER_TOKEN|> ويكون النص باللغة العربية الفصحى بعد عبارة
<|ASSISTENT_TOKEN|>
<|SYSTEM_PROMPT_END|>
<|USER_TOKEN|>
ممكن تعطيني تذكرتين درجة تانية على شيكاغو، إذا بتريد؟
<|ASSISTENT_TOKEN|>:
هل يمكنني الحصول على تذكرتين درجة ثانية لشيكاغوا , من فضلك ؟
"""
```

```
"""<|SYSTEM_PROMPT|>
أنتَ مساعد آلي مُهِمَتكَ تحويل أو ترجمة النصوص من و إلى لهجات عربية متنوعة حسب طلب المستخدم.
سؤال المستخدم بعد عبارة <|USER_TOKEN|> ويكون النص باللهجة المطلوبة بعد عبارة
<|ASSISTENT_TOKEN|>
<|SYSTEM_PROMPT_END|>
<|USER_TOKEN|>
حوّل النص التالي من اللهجة الشامية إلى اللهجة التونسية
اي. استنى، رح جبلك واحد.
<|ASSISTENT_TOKEN|>:
اي. ستنا تو نجيبلك وحدة.
"""
```

Figure 1: Illustration of the two system prompt templates used in Lahjawi. (Left) Template for translating any dialect-to-MSA, with **system prompt** in dark red, **input** in light blue, and **output** in green. (Right) Template for translating between specific dialects, with **system prompt** in dark red, **template question** in orange, **dialect** name in red, **input** in light blue, and **output** in green.

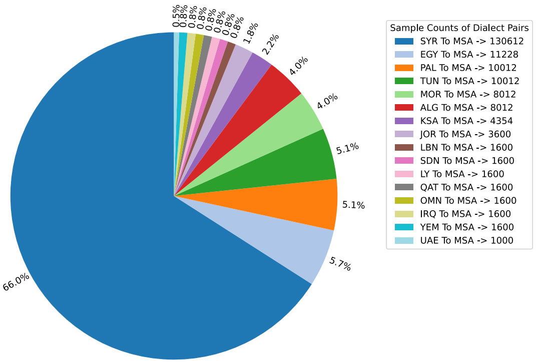


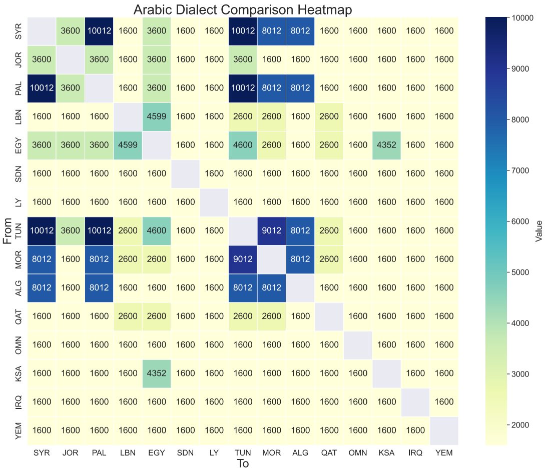
Figure 2: The distribution of dialect-to-MSA samples in D2MSA dataset

steps ensure consistency across the datasets, enabling more accurate and reliable model training. To evaluate the performance of our models, we utilized two datasets.

The first dataset is the NADI-2024 DA-MSA test and development data (Abdul-Mageed et al., 2024), which is available in four dialects: Egyptian, Emirati, Jordanian, and Palestinian. This benchmark facilitates the comparison of our results with others. Additionally, we selected the MADAR parallel corpus test set (Bouamor et al., 2018) to assess our model's performance on additional dialects, considering the absence of a standardized benchmark for testing the translation of other dialects into MSA. We applied the same benchmark to evaluate Lahjawi-D2D for cross-dialectical translation, leveraging the fact that MADAR offers parallel translations between our targeted dialects.

4 Model

Lahjawi models are a fine-tuned adaptation of an in-house small language model *Kuwain 1.5B*, specifically optimized for the challenging task of Arabic dialect translation. In our approach, we reformulated the translation problem into a Question-Answering (QA) framework, which enabled more precise and focused training. This reframing allowed us to capture the nuances of dialect-specific translations better.



	SYR	JOR	PAL	LBN	EGY	SDN	LY	TUN	MOR	ALG	QAT	OMN	KSA	IRQ	YEM
SYR		3600	10012	1600	3600	1600	1600	10012	8012	8012	1600	1600	1600	1600	1600
JOR	3600		3600	1600	3600	1600	1600	3600	1600	1600	1600	1600	1600	1600	1600
PAL	10012	3600		1600	3600	1600	1600	10012	8012	8012	1600	1600	1600	1600	1600
LBN	1600	1600	1600		4599	1600	1600	2600	2600	1600	2600	1600	1600	1600	1600
EGY	3600	3600	3600	4599		1600	1600	4600	2600	1600	2600	1600	4352	1600	1600
SDN	1600	1600	1600	1600	1600		1600	1600	1600	1600	1600	1600	1600	1600	1600
LY	1600	1600	1600	1600	1600	1600		1600	1600	1600	1600	1600	1600	1600	1600
TUN	10012	3600	10012	2600	4600	1600	1600		9012	8012	2600	1600	1600	1600	1600
MOR	8012	1600	8012	2600	2600	1600	1600	9012		8012	2600	1600	1600	1600	1600
ALG	8012	1600	8012	1600	1600	1600	1600	8012	8012		1600	1600	1600	1600	1600
QAT	1600	1600	1600	2600	2600	1600	1600	2600	2600	1600		1600	1600	1600	1600
OMN	1600	1600	1600	1600	1600	1600	1600	1600	1600	1600	1600		1600	1600	1600
KSA	1600	1600	1600	1600	4352	1600	1600	1600	1600	1600	1600	1600		1600	1600
IRQ	1600	1600	1600	1600	1600	1600	1600	1600	1600	1600	1600	1600	1600		1600
YEM	1600	1600	1600	1600	1600	1600	1600	1600	1600	1600	1600	1600	1600	1600	

Figure 3: Heatmap of Arabic Dialect Comparison

As outlined in the previous section, we implemented a consistent template transformation across the entire training dataset, tailoring the input-output structures to align with dialect-specific translation tasks, as illustrated in Figure 1. This step was crucial in adapting the general-purpose *Kuwain* model to specialize in translating input text from one dialect to another, based on the prompt provided.

The fine-tuning process followed the next-token prediction paradigm, with system prompts and embedding tokens carefully masked to ensure the model focused on relevant dialectal context. The training was conducted over three epochs using a *cosine learning rate schedule*, with meticulously adjusted hyperparameters to maximize performance. These optimizations ensured the model's ability to capture both subtle and overt linguistic distinctions across the dialects, delivering robust translation quality across diverse sentence structures. See Appendix A for configuration details.

By combining the strengths of the *Kuwain* model with our specialized fine-tuning approach, Lahjawi models are uniquely positioned to address the complexities of Arabic dialect translation. This tailored methodology enables Lahjawi to serve as a powerful tool for facilitating cross-dialectal communication, offering more accurate and context-aware translations between the various Arabic dialects.

5 Experiments and Results

This section presents the results from four experiments conducted on Arabic dialect translation. Each experiment was designed to evaluate different aspects of the translation process. The *first experiment*, Lahjawi-QuadD, follows the methodology of several papers participating in NADI 2024 competition (Abdul-Mageed et al., 2024), to translate from specific Arabic dialects to MSA, serving as a benchmark to compare results. The *second experiment*, Lahjawi-4Isolate, was inspired by (Khered et al., 2023), which suggested that training a model separately for each dialect improves performance. However, our results contradicted this hypothesis, leading us to the *third experiment*, Lahjawi-D2MSA, which investigated the impact of increasing the number of dialects on overall performance. The *fourth and final experiment*, Lahjawi-D2D, represents our primary contribution to developing the first-ever model for direct translation between Arabic dialects.

5.1 Lahjawi-QuadD: A Comprehensive Model on 4 Dialects

The experiment focused on fine-tuning a model to translate four Arabic dialectsJordanian, Palestinian, Emirati, and Egyptianinto Modern Standard Arabic (MSA), as part of the NADI-2024 (Abdul-Mageed et al., 2024) subtask DA-MSA machine translation. The model was trained on sample sizes of 3,600 for Jordanian, 10,012 for Palestinian, 14,227 for Egyptian, and 1,000 for Emirati. The data is a subset of D2MSA data for translating input text to MSA. Table 2 presents the model's evaluation measured by the BLEU metric, for the NADI-2024 DA-MSA test data across various Arabic translation systems.

5.2 Lahjawi-4Isolate: The Effect of Separates Models Training

In this experiment, four distinct models were trained, each specifically dedicated to translating one of the four target dialects in the NADI-2024 into Modern Standard Arabic (MSA). The main objective of this experiment was to explore the impact of training separate models for each dialect versus using a unified model, as done in the first experiment. The results in Table 2 illustrate the inefficiency of training separate models for each dialect, demonstrating that the previous experiment significantly enhances translation quality compared to this one. Results and findings will be discussed in the section 6

5.3 Lahjawi-D2MSA: A Unified Model for Translating All Arabic Dialects to MSA

This experiment focused on developing a robust model for translating various Arabic dialects into Modern Standard Arabic (MSA). The training utilized (D2MSA) dataset, enabling the model to handle the linguistic variations effectively across these diverse dialects. The dataset includes 197,042 samples, with detailed information on the dialects and their corresponding sample sizes provided in Figure 2. Tables 2 and 3 present the BLEU metrics of the unified model derived from the NADI-2024 DA-MSA and MADAR test data, respectively.

Table 4 in Appendix B demonstrates Lahjawi-D2MSA translation examples from different Arabic dialects to Modern Standard Arabic (MSA). The table specifically presents the original dialect sentences alongside their corresponding Lahjawi-D2MSA outputs, illustrating how the translations capture the essence of the original expressions while adapting them to the standardized form of Arabic.

5.4 Lahjawi-D2D: A Model for Cross-Dialect Translation

Lahjawi-D2D is an Arabic model for Arabic cross-dialect translation, capable of translating between 15 dialect pairs shown in Figure 3. The model was developed using a standardized format for conversion between any two dialects. The model's performance was evaluated using the BLEU metric on the MADAR test data, with results detailed

System	Overall	Egy.	Emi.	Jor.	Pal.
Arabic Train	20.44	16.57	23.38	21.37	20.62
Alson	17.46	16.76	17.53	20.94	18.43
ASOS	17.13	14.82	19.39	15.80	18.38
CUFE	16.09	14.86	17.35	15.98	16.82
Lahjawi-QuadD	13.55	12.64	12.51	14.96	14.20
Lahjawi-D2MSA	13.30	11.39	11.37	17.40	13.67
Lahjawi-4Isolate	12.13	10.54	15.27	7.87	14.41
MBZUAI BLEU	10.54	8.53	11.51	11.79	10.44
VBNN	9.24	8.62	6.30	11.79	10.54
AraT5v2	6.87	9.38	4.61	4.90	8.13
mT5	2.81	3.08	2.33	3.11	2.95
MBZUAI BADG	2.78	3.03	2.53	1.98	2.58
AraBART	0.87	0.77	0.81	1.11	0.88

Table 2: Performance Metrics: BLEU Scores Across Various Arabic Translation Systems Evaluated on NADI-2024 DA-MSA Test Data.

Dialect	**Test BLEU**	**Dialect**	**Test BLEU**
KSA	10.81	ALG	9.40
OMN	11.31	LY	7.89
QAT	8.77	MOR	8.58
IQR	8.37	TUN	6.47
JOR	11.52	EGY	10.55
LBN	11.29	SDN	8.97
PAL	11.24	YEM	7.80
SYR	11.39		
Overall: 9.62			

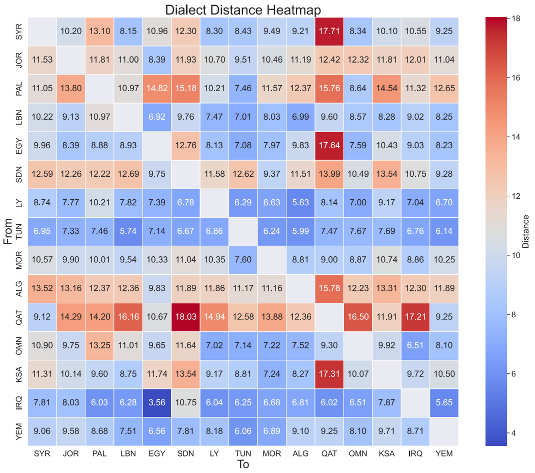
Table 3: *Lahjawi-D2MSA* BLEU Scores on MADAR Test Datasets for Arabic Dialects

in Figure 4. Additionally, human assessments were conducted on 50 sentences for the most commonly spoken dialects, including Syrian, Jordanian, Palestinian, Tunisian, Egyptian, Saudi, and Moroccan. These evaluations, which assess accuracy and fluency, were assigned scores ranging from 1 to 5. The combined outcomes of the human evaluations and the BLEU scores provide valuable insights into the model's effectiveness in cross-dialect translation. Table 5 and 6 in Appendix C demonstrate Lahjawi-D2D translations examples from the Egyptian, and Syrian dialects to various Arabic dialects respectively. It highlights how dialectal variations affect phrasing and vocabulary, showcasing similarities and unique features across all dialects.

6 Discussion

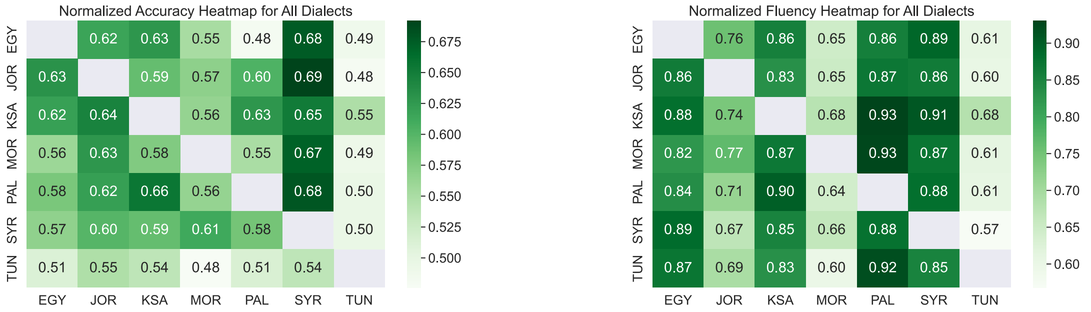
First, we will examine the impact of the first experiment involved training a comprehensive model on four dialects collectively (*Lahjawi-QuadD*), compared to the second experiment focused on training separate individual models for each of these dialects (*Lahjawi-4Isolate*). As shown in Table 2, training a comprehensive model demonstrated a relatively consistent performance across all dialects, showing slightly better results in Jordanian and Palestinian dialects. Despite the limited number of samples for the Jordanian dialect in the training data, this did not significantly impact the model's performance. This could potentially be attributed to the benefits of shared knowledge across dialects, leading to improved overall model performance.

Observing the results of training individual models for each dialect reveals that separate



	SYR	JOR	PAL	LBN	EGY	SDN	LY	TUN	MOR	ALG	QAT	OMN	KSA	IRQ	YEM
SYR		10.20	13.10	8.15	10.96	12.30	8.30	8.43	9.49	9.21	17.71	8.34	10.10	10.55	9.25
JOR	11.53		11.81	11.00	8.39	11.93	10.70	9.51	10.46	11.19	12.42	12.32	11.81	12.01	11.04
PAL	11.05	13.80		10.97	14.82	15.18	10.21	7.46	11.57	12.37	15.76	8.64	14.54	11.32	12.65
LBN	10.22	9.13	10.97		6.92	9.76	7.47	7.01	8.03	6.99	9.60	8.57	8.28	9.02	8.25
EGY	9.96	8.39	8.88	8.93		12.76	8.13	7.08	7.97	9.83	17.64	7.59	10.43	9.03	8.23
SDN	12.59	12.26	12.22	12.69	9.75		11.58	12.62	9.37	11.51	13.99	10.49	13.54	10.75	9.28
LY	8.74	7.77	10.21	7.82	7.39	6.78		6.29	6.63	5.63	8.14	7.00	9.17	7.04	6.70
TUN	6.95	7.33	7.46	5.74	7.14	6.67	6.86		6.24	5.99	7.47	7.67	7.69	6.76	6.14
MOR	10.57	9.90	10.01	9.54	10.33	11.04	10.35	7.60		8.81	9.00	8.87	10.74	8.86	10.25
ALG	13.52	13.16	12.37	12.36	9.83	11.89	11.86	11.17	11.16		15.78	12.23	13.31	12.30	11.89
QAT	9.12	14.29	14.20	16.16	10.67	18.03	14.94	12.58	13.88	12.36		16.50	11.91	17.21	9.25
OMN	10.90	9.75	13.25	11.01	9.65	11.64	7.02	7.14	7.22	7.52	9.30		9.92	6.51	8.10
KSA	11.31	10.14	9.60	8.75	11.74	13.54	9.17	8.81	7.24	8.27	17.31	10.07		9.72	10.50
IRQ	7.81	8.03	6.03	6.28	3.56	10.75	6.04	6.25	6.68	6.81	6.02	6.51	7.87		5.65
YEM	9.06	9.58	8.68	7.51	6.56	7.81	8.18	6.06	6.89	9.10	9.25	8.10	9.71	8.71	

Figure 4: Lahjawi-D2D's BLEU scores on MADAR test set.



(a)

(b)

Figure 5: Dialect-to-Dialect Human evaluation: (a) accuracy scores, (b) fluency score.

training does not consistently lead to better performance, particularly for the Jordanian model, which showed a notable drop in accuracy. This aligns with findings from (AlMusallam and Ahmad, 2024), who observed that the Jordanian and Palestinian dialects tend to achieve high accuracy with minimal differences between them when used together for training, likely due to their close similarity to each other and Modern Standard Arabic (MSA).To confirm this, we compared the Jordanian and Palestinian models on the NADI development set. The results were as follows: Jordanian achieved a BLEU score of 5.67, while Palestinian achieved 14.12, Interestingly, when we tested the Palestinian model on Jordanian data, it scored 19.34, while the Jordanian model scored 6.84 on Palestinian data. These results suggest that the Jordanian dataset is relatively small, and given the similarity between the two dialects, combining them leads to improved BLEU scores. Despite the limited number of training samples for the Emirati dialect, the model performed well. This success could be attributed to the fact that *Kuwain* was exposed to more Gulf dialects during the pre-training phase, leading to a better understanding and representation of the Emirati dialect within the model.

As for the third experiment, training*Lahjawi-D2MSA* as a unified model on 15 dialects yields slightly different scores, with similar overall averages. These small differences indicate that increasing the number of dialects adds translation challenges in some dialects due to the increase in complexity, while others may benefit from the existence of other dialects since similar words and contexts may be the same in different dialects. Nevertheless, the model demonstrated strong adaptability across the diverse linguistic variations.

Compared to other models, our model Lahjawi-D2MSA produced mediocre results. In contrast, teams like Arabic Train (Demidova et al., 2024) and CUFE(Ibrahim, 2024), with superior models such as the Jais-13B and LLaMA-8B multilingual model, leveraged much larger architectures. Additionally, teams like Alson (AlMusallam and Ahmad, 2024) and ASOS (Nacar et al., 2024) enhanced their performance by augmenting their datasets with higher quality and more extensive data. This suggests that using a larger model along with higher-quality data could significantly improve performance.

In Table 3, Lahjawi-D2MSA demonstrates higher performance with Levantine dialects, which aligns with the significant representation of the Syrian Levantine dialect in the training dataset. Additionally, Gulf and Egyptian dialects exhibit decent translation performance, although not as robust as the Levantine dialects. However, the model encounters difficulties with Maghribi dialects, especially Tunisian. These challenges may stem from linguistic differences and the complexity inherent in those dialects, diverging from Modern Standard Arabic (MSA). This underscores the importance of additional training or refining the model to handle underrepresented dialects.

Analyzing the results presented in Figure 4 the Lahjawi-D2D model highlights that certain dialects consistently achieve higher scores (e.g., Qatari, Palestinian) compared to others, such as Iraqi and Libyan, which exhibit notably lower scores. Several factors could contribute to these disparities, including the quality and quantity of training data, as well as the presence of specific dialects during the pre-training phase of the *Kuwain* model . Moreover, it is observed that the model's translation capabilities are not always symmetrical. Some dialects may translate more effectively in one direction than the other. For instance, the translation score from Qatari to Iraqi is 17.21, whereas from Iraqi to Qatari, it is 6.02. This asymmetry in translation performance highlights the complexity and nuances involved in accurately capturing the linguistic variations between different dialects.

The results of the human evaluation accuracy in Figure 5a indicate that the Syrian dialect achieves the highest translation accuracy among the Arabic dialects, largely due to its large dataset and close similarity to Modern Standard Arabic and other Eastern dialects (Egyptian, Saudia, Palestinian, Jordanian, and Syrian) (see Figure 3). In contrast, although the Moroccan (Tunisia, Morocco) dialects have a large dataset, they achieve lower translation accuracy due to their divergence from Modern Standard Arabic and its most closely related dialects. Overall, the accuracy rating for this evaluation is 58%.

Figure 5b shows high fluency levels among most Arabic dialects, with the Eastern dialects showing high similarity and high fluency among them. While the Moroccan dialects show lower variation and percentages for the same reasons related to the nature of the dialect and its rarity in the basic training data in the original model. The overall fluency level, as assessed, reaches 78%.

7 Limitations

Our work faced significant challenges due to the complexity and diversity of Arabic dialects, which often deviate from Modern Standard Arabic in vocabulary and grammar. The lack of standardized sentence structures and written forms in many dialects complicated the training and evaluation of our models. A significant limitation is the quality and availability of Arabic dialect datasets, which are often small, unevenly distributed, and lack clear distinctions between dialects. Parallel training corpora are usually built separately for each dialect, without highlighting their similarities and differences, making it challenging to train models that accurately differentiate between them. Additionally, many translations in these datasets are rephrased rather than literal, adding complexity to both generating and evaluating precise translations. Finally, the model's tendency to generate inaccurate outputs (hallucinations), particularly in smaller models, highlighted the resource constraints in developing accurate cross-dialect translators.

8 Acknowledgement

This work was made possible through the support of Misraj Corporation, whose management and colleagues provided the necessary resources and technical infrastructure for this research. We express our sincere gratitude to Dr. Nada Ghneim from Arab International University for her thorough review and insightful feedback, which substantially improved this manuscript.

References

AhmedElmogtaba Abdelmoniem Ali Abdelaziz, Ashraf Hatim Elneima, and Kareem Darwish. 2024. LLM-based MT data creation: Dialectal to MSA translation shared task. In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation @ LREC-COLING 2024*, pages 112--116, Torino, Italia. ELRA and ICCL.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, Chiyu Zhang, El Moatez Billah Nagoudi, Houda Bouamor, and Nizar Habash. 2023. NADI 2023: The fourth nuanced Arabic dialect identification shared task. In *Proceedings of ArabicNLP 2023*, pages 600--613, Singapore (Hybrid). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Amr Keleg, AbdelRahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. 2024. NADI 2024: The fifth nuanced Arabic dialect identification shared task. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 709--728, Bangkok, Thailand. Association for Computational Linguistics.

Roqayah Al-Ibrahim and Rehab M. Duwairi. 2020. Neural machine translation from jordanian dialect to modern standard arabic. In *2020 11th International Conference on Information and Communication Systems (ICICS)*, pages 173--178.

Mansour Al Sulaiman, Abdullah M Moussa, Sherif Abdou, Hebah Elgibreen, Mohammed Faisal, and Mohsen Rashwan. 2022. Semantic textual similarity for modern standard and dialectal arabic using transfer learning. *PloS one*, 17(8):e0272991.

Salwa Saad Alahmari. 2024. Sirius_Translators at OSACT6 2024 shared task: Fin-tuning ara-t5 models for translating Arabic dialectal text to Modern Standard Arabic. In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation @ LREC-COLING 2024*, pages 117--123, Torino, Italia. ELRA and ICCL.

Manan AlMusallam and Samar Ahmad. 2024. Alson at NADI 2024 shared task: Alson - a fine-tuned model for Arabic dialect translation. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 764--768, Bangkok, Thailand. Association for Computational Linguistics.

Hanin Atwany, Nour Rabih, Ibrahim Mohammed, Abdul Waheed, and Bhiksha Raj. 2024. OSACT 2024 task 2: Arabic dialect to MSA translation. In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation @ LREC-COLING 2024*, pages 98--103, Torino, Italia. ELRA and ICCL.

Houda Bouamor, Nizar Habash, and Kemal Oflazer. 2014. A multidialectal parallel corpus of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1240--1245, Reykjavik, Iceland. European Language Resources Association (ELRA).

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 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).

Anastasiia Demidova, Hanin Atwany, Nour Rabih, and Sanad Sha'ban. 2024. Arabic train at NADI 2024 shared task: LLMs' ability to translate Arabic dialects into Modern Standard Arabic. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 729--734, Bangkok, Thailand. Association for Computational Linguistics.

Wiem Derouich, Sameh Kchaou, and Rahma Boujelbane. 2023. ANLP-RG at NADI 2023 shared task: Machine translation of Arabic dialects: A comparative study of transformer models. In *Proceedings of ArabicNLP 2023*, pages 683--689, Singapore (Hybrid). Association for Computational Linguistics.

Mohamed Faheem, Khaled Wassif, Hanaa Bayomi, and Sherif Abdou. 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.

Murhaf Fares. 2024. AraT5-MSAizer: Translating dialectal Arabic to MSA. In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation @ LREC-COLING 2024*, pages 124--129, Torino, Italia. ELRA and ICCL.

Michael Ibrahim. 2024. CUFE at NADI 2024 shared task: Fine-tuning llama-3 to translate from Arabic dialects to Modern Standard Arabic. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 769--773, Bangkok, Thailand. Association for Computational Linguistics.

Saméh Kchaou, Rahma Boujelbane, and Lamia Hadrich-Belguith. 2020. Parallel resources for Tunisian Arabic dialect translation. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 200--206, Barcelona, Spain (Online). Association for Computational Linguistics.

Abdullah Khered, Ingy Abdelhalim, Nadine Abdelhalim, Ahmed Soliman, and Riza Batista-Navarro. 2023. UniManc at NADI 2023 shared task: A comparison of various t5-based models for translating Arabic dialectical text to Modern Standard Arabic. In *Proceedings of ArabicNLP*

2023, pages 658--664, Singapore (Hybrid). Association for Computational Linguistics.

Karima Meftouh, Salima Harrat, Salma Jamoussi, Mourad Abbas, and Kamel Smaili. 2015. Machine translation experiments on PADIC: A parallel Arabic DIalect corpus. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 26--34, Shanghai, China.

Zinah J Mohammed Ameen and Abdulkareem Abdulrahman Kadhim. 2023. Deep learning methods for arabic autoencoder speech recognition system for electro-larynx device. *Advances in Human-Computer Interaction*, 2023(1):7398538.

Hamdy Mubarak. 2018. Dial2msa: A tweets corpus for converting dialectal arabic to modern standard arabic. *OSACT*, 3:49.

Omer Nacar, Serry Sibaee, Abdullah Alharbi, Lahouari Ghouti, and Anis Koubaa. 2024. ASOS at NADI 2024 shared task: Bridging dialectness estimation and MSA machine translation for Arabic language enhancement. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 748--753, Bangkok, Thailand. Association for Computational Linguistics.

Hashem Sellat, Shadi Saleh, Mateusz Krubiński, Adam Pospíšil, Petr Zemánek, and Pavel Pecina. 2023. UFAL parallel corpus of north levantine 1.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Xudong Han, Sondos Mahmoud Bsharat, Alham Fikri Aji, Zhiqiang Shen, Zhengzhong Liu, Natalia Vassilieva, Joel Hestness, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Hector Xuguang Ren, Preslav Nakov, Timothy Baldwin, and Eric Xing. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *Preprint*, arXiv:2308.16149.

Mohamed Ali Sghaier and Mounir Zrigui. 2020. Rule-based machine translation from tunisian dialect to modern standard arabic. *Procedia Computer Science*, 176:310--319. Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 24th International Conference KES2020.

Khaled Shaalan, Sanjeera Siddiqui, Manar Alkhatib, and Azza Abdel Monem. 2019. Challenges in arabic natural language processing. In *Computational linguistics, speech and image processing for Arabic language*, pages 59--83. World Scientific.

Roua Torjmen and Kais Haddar. 2024. Translation from tunisian dialect to modern standard arabic: Exploring finite-state transducers and sequence-to-sequence transformer approaches. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*

A Configuration of Supervised Fine-Tuning

Lahjawi models are a fine-tuned adaptation of an in-house small language model *Kuwain 1.5B* with 8 Nvidia A100 80G GPUs. We employ the AdamW optimizer, with each batch consisting of 512 samples. We adopt 8×10^{-6} as the maximum learning rate based on the model architecture. A cosine scheduler is employed for learning rate adjustment, with a weight decay of 0.1. We train for one epoch using a variety of experiments discussed in section 5.

B Dialect-to-MSA Translation Examples (Lahjawi-D2MSA)

Dialect	Original Dialect Sentence	Lahjawi-D2D output
UAE	لا لو تسويلي عشاء، أنا مب طالع من هني إلين ما أعرف سالفة الذهب	لا لو تحضرين لي عشاء، أنا لن أخرج من هنا حتى أعرف قصة الذهب
SYR	ياخي مين سائل كلو مو فرقانة معو المهم انو تجي البضاعة لعندو ويا دار ما دخلك شر	من الذى يسأل و الذى يجيب و الذى يصله هذا من خلال المهم ان تاتى البضاعه اليه . و بالتالى فان من يسأل شر و من لا يسأل شر .
LY	ما نبي نمشي للشغل تواء الجو حامي وحاسس روحي محتاج قعدة قدام المكيف ونشرب شاهي بالنعناع.	لا أريد الذهاب إلى العمل الآن. الجو حار جدا. أتطلع إلى الحصول على استراحة من البرد.
IRQ	ممكن اعرف شنو هذا الهاشتاگ لان طبيت عليه ما افتهمت شي	هل من الممكن أن تخبرني ما هذا الهاشتاغ لأنني لم أفهمه
ALG	هادا كيتمان عفسه سهيله نفهموها غير حنا هه	هذا كيتمان على سهيلة فهموها نحن فقط
JOR	البطاريه فلسعت والكهربا في حد مزعلها منا اشبكلي كابل ياخي	البطارية ضعيفة و الكهرباء مقطوعة هل من الممكن ان تزودني بالكابل
EGY	ايه دا يعني ايه يعني انا كمان كام شهر لما اعدي التلاتين مش هابقي طفله	نعم ماذا يعني؟ يعني أنني عندما أصل إلى الثلاثين لن أكون طفلة
TUN	قداش كثروا ولاد و بنات الدار الكبيره الليله	كم عدد أطفال و بنات المنزل الكبير الليلة؟
PAL	بتعرف انك نكد؟ خذ امسك و مابديش احكي معك زيح هيك امشي	هل تعرف انك نكد؟ امسك و لا تريد ان تتحدث معي دعني اذهب
KSA	اعرف ان الأشياء اللي اسويها غلط بس كمان اسويها وما اهتم	أعلم أن ما أفعله خطأ ولكنى أفعله مرة أخرى ولا أهتم
OMN	واحد يسأل دلوع : وش اكثر نوع يعجبك من 'الابل'؟ رد الدلوع : تستعبط أكيد آيباد يقولون دفنوه والدلة ناشبة براسه هههههه	سأل شخص دلوع ما هو أكثر نوع تحبه من ابل؟ رد الدلوع: لا تستعجب، تستعبط، بالتأكيد آيباد، يقولون أنه تم دفنه و ربطه في شجرة و يقال أنه حي، ما هذا الهراء؟
MOR	الإنسان إبن بيئته، مافيها باس استنشقو الجو اللي فيه النكْير و شد لي نقطع ليك	الإنسان ابن بيئته، لا يمكن أن تغيره، استنشق فقط الهواء و أقطع لك مثلا
LBN	راكبه مواصلات مع شوية متخلفين حرفيا!!	راكب حافلة مع بعض المتخلفين حرفيا!
YEM	طالما وما فيش موعد مع حد نام على الجهة التي تريحك دون وجع راس وقلق من الموعد هههههههه	طالما ولا يوجد موعد مع احد نام على الجهة التي تريحك دون الم وقلق من الموعد
SDN	الفي معتقل موقف شندي بتطلع بكرا باذن الله فلو في زول عندو زول مفقود يجي بكرا ٢٦/١٢ الساعة ٧:٠٠ صباحاً	في معتقل موقف شندي تخرج غدا بإذن الله، فلو كان هناك شخص عنده شخص مفقود يأتي غدا 26/12/2012 الساعة 7:00 صباحا
QAT	شف عذر اخلق بأيّة طريقة كذبة بيضا او حقيقة .. المهم أنك تجي وغير هذا ما أبي	شف عذراً إخلق بأي طريقة، كذبة بيضاء أو حقيقة، المهم أنك تأتي، وغير هذا لا أريده.

Table 4: Examples of Dialect to Modern Standard Arabic (Lahjawi-D2MSA) Translations.

C Dialect-to-Dialect Translation Examples (Lahjawi-D2D)

Dialect-to-Dialect	Lahjawi-D2D output
EGY to JOR	شو القصة؟ كل ما خلصت شغل لسا في مليون اشي تاني، بديت أتشتت.
EGY to PAL	شو؟ كل ما خلص شغل رح أكتشف شغلات جديدة، بس ما بقدر. بصراحة، ما بقدر أشتغل أكتر.
EGY to LBN	شو في؟ كل هلق لما بتخلص شي بتلاقي مليون شي تاني ناطرك، فما حدا عم بيقدر يركز هون.
EGY to SYR	شو القصة؟ لما بخلص شغل واحد بتكون عندي كم شغل تانية لازم أكملها، ما بعرف إذا كنت عم شوف شغلات تانية.
EGY to SDN	كيف هو؟ انا دائما بخسر حاجة ما حقتي، ما عارف السبب.
EGY to LY	شن القصة؟ كل ما خلصت حاجة أكتشفت حاجة أكبر منها. مش قادر ركز.
EGY to TUN	شنيا القصة؟ كل ما نحل حاجة نلقاها أخرى، ديما نطيح من عيني.
EGY to MOR	شنو كاي؟ كل ما تنهي شي حاجة تيحصل شي وحدا أخر ب شكل عام، و تتعرف بلي معندناش حل غير ندور.
EGY to ALG	واش راه يصرا، دوما وين نخلص حاجة نقددها نخلص حاجة اكبر. الانسان ما لقيتوش نفسه يسوي.
EGY to QAT	شنو؟ كل ما خلصت شي احصل في دوشتي غير مليون شي ثاني، ما اقدر اتم تركيزي.
EGY to OMN	شو؟ كلما خلصت شي لقيته زيادة، ما عرفت اذا كان هالشي مهم او لا.
EGY to KSA	أيش؟ كل ما خلصت حاجة كان يأتيني شيء أكبر منها، حتى وصلت لشيء مستحيل أعرف أتعامل معه.
EGY to IRQ	شنو رأيك؟ اذا سويت كل شي، راح اوصل الى مليون شيء آخر، ما اكدر اهتم بهي.
EGY to YEM	طيب، ما قصتك؟ انت دائما تبسر كل شي، والكل يقول لك ان تبسر غير هذا. انسان مثلك لازم يفلس.

Table 5: Translation from Egyptian Dialect to Other Dialects for this Sentence: "هوإيه الحكاية؟ كل ما أقوم أخلص حاجة ألاقي مليون حاجة تانية مستنياني، بجد الواحد مش عارف يركز."

Dialect-to-Dialect	Lahjawi-D2D output
SYR to JOR	امبارح رحت على سوق اخضرت كتير، وكلو خضرا و ما في حدا اشترى، كلو زعلان .
SYR to PAL	امبارح رحت على محل أبو جميل و اشتريت شوية خضروات زينين و غاليين، كلو ولا اشي، وبس في ناس كتير بيخافوا و يشتروا منهم.
SYR to LBN	مبارح رحت على أبو جميل جبت شوية خضرا، عنجد الأسعار كتير نار و كل شي غالي و ما حدا عم يشتري . مبلاة الناس و زعلانين .
SYR to EGY	امبارح رحت عند ابو جميل جبت شوية خضروات، الحقيقة ان الاسعار رهيبة، وكله محجوز.
SYR to SDN	البارح عند بائع الخضروات، عندي خضروات، كويس جدا، و خلاه غالي، ما في زول بيبيع زي دا.
SYR to LY	البارح نرفعو لعند سيده حسن و نجيبو شوية خضره، هدا هو الغلا و الفشار قاعدين غالينا، و كل واحد يسال الاخر.
SYR to TUN	البارح نجري عند بو خاطر باش اشتري شوية خضرة، أما الأمر باين باهي، كل شيء غالي شوية، باهي، الناس الكل توا لاهيها الفايدة متاع لا حديقة.
SYR to MOR	البارح مشيت عند خو جميل باش نجيب شوية ديال العشب، ولكن العشب غالي بزاف و ما كاين حتى واحد غادي يشتريه و كانو كلشي تيغضب.
SYR to ALG	البارح نروح عند بابا جميل نجيب شوية خضرة، كيما نقول بالأخص الكاجو والأسماك، ولكن الناس كل هابطة عندهم الغضب.
SYR to QAT	امبارح رحت ابو جميل عشان اشتري خضروات، والله الاسعار واجد غالية ومافيه احد يشتري، كلهم ساكتين.
SYR to OMN	انا رحت الى ابو ظبي عشان اشوف بعض الخضار، صدقني الاسعار عالية، وكل شي غالي في هاذي الأيام.
SYR to KSA	من امس رحت لشارع ابو جميل واشريت شوية خضروات، بس والله الاسعار تهيلت، ومافيه احد يشتري، وحنا.
SYR to IRQ	البارح جانت راحه عند ابو سامي بجيب شوية خضروات، بس شنو كلش غالي ما واحد يشتريه.
SYR to YEM	البارح اشلت ل السوق لاشلاح شوية خضروات، لكن الصعوبه ان كلها غاليه وماكانش ممكن اشتريها الان لانه هيه حاله غلاء الاسعار.

Table 6: Translation from Syrian Dialect to Other Dialects for this Sentence: مبارح رحت لعند أبو جميل نجيب شوية خضرة، لك والله الأسعار نار، وكل شي غالي وما في حدا عم يشتري.