Toward Building a Bilingual Dictionary for Libyan Dialect-modern Standard Arabic Machine Translation

Husien Alhammi[1] and Kais Haddar[2]
[1]*Faculty of Economics and Management, University of Sfax, Tunisia*
[2]*Laboratory MIRACL, University of Sfax, Tunisia*

Keywords: Libyan Dialect, Bilingual Dictionary, Modern Standard Arabic, Machine Translation.

Abstract: In this paper, a method for building a bilingual dictionary that can be used to translate words and phrases from one dialect to the native language is described. Obviously, dialects and their main languages have many features in common in terms of linguistic structure, lexicon, morphology, and so on. As a result, the method of creating a bilingual dictionary to translate texts from one language into another differs significantly from a method that is used to build a bilingual dictionary for translating dialects into their native languages. To this end, a specific dictionary including some linguistic information must be built to translate dialects into their native languages. In this paper, we discuss the main idea of the method that is used to build a bilingual dictionary to translate Libyan Dialect (LD) into its original language, Modern Standard Arabic (MSA). The advantages of the method are discussed.

1 INTRODUCTION

The Arabic language is spoken by about 300 million people in the Arab world, which consists of 22 different countries. The Arabic language usually refers to Standard Arabic (SA), which is divided by linguists into Classical Arabic (CA), Modern Standard Arabic (MSA), and the Arabic Dialects (AD). Although, MSA is the official language that is used in literature, academics, media, law, legislation, and formal education, it is not a daily language. In contrast, Arabic dialects are spoken mostly at home and in everyday life by Arabic speakers. Despite the fact that the dialects are rarely written down in a formal writing style, they have become widely used on social media networks such as Facebook and Twitter by Arabic users as an informal style. In fact, most of the Arabic dialects are mainly different from each other, making it difficult for speakers to understand them. However, Arabic dialects can be classified into five categories by linguists (Zaidan et al, 2014) (Ghoul et al, 2019) which are: Maghreb Arabic (Morocco, Algeria, Tunisia, Libya, western Sahara, and Mauritania), Levantine (Lebanese, North Syria, Damascus, Palestine and Jordan), Gulf Arabic (Northern Iraq, Baghdad, Southern Iraq, Gulf, Saudi-Arabia, and Southern Arabic Peninsula) and Nile Region (Egypt and Sudan). However, around six million out of 300 million Arabic speakers are also LD speakers. It is a member of the Maghreb family which are spoken in north Africa, it can be generally classified by linguists into three similar dialects depending on their geographic areas: the eastern, the western and the southern dialect being centered in Benghazi, Tripoli and Sabha respectively.

During the colonization era, several words in the current LD came from colonialist languages like Italian and Turkish. Furthermore, indigenous languages like Berber or Amazigh are extensively spoken in the LD (Abdulaziz et al, 2014). For example, the LD word "سفرانية" is borrowed from the Amazigh language which means in English "carrot" as well as the words "كاشيك" and "كوجينا" are borrowed from Turkish and Italian which mean in English "spoon" and "kitchen" respectively. As we see in the above example, LD is distinguished by the usage of a multitude of languages, including Arabic, Italian, Turkish, Tamazight, as well as some vocabulary from French, English, and even Persian. Here is an example of LD text: " قطوس مقعمزة على اروشن وتشبح وطا" which means in English "a cat sitting on the window and looking down", it is difficult for other Arabic speakers, even those who speak Maghreb dialects, to understand it. Obviously, Arabic speakers have a hard time understanding LD

Alhammi, H. and Haddar, K.
Toward Building a Bilingual Dictionary for Libyan Dialect-modern Standard Arabic Machine Translation.
DOI: 10.5220/0011560700003335
In *Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2022) - Volume 2: KEOD*, pages 75-81
ISBN: 978-989-758-614-9; ISSN: 2184-3228

because it employs a wide range of vocabulary from several languages, to make LD more understandable. To begin, LD text is translated into the MSA language to be understandable by Arabic speakers. Second, translation software solutions such as Google translate could be used to easily translate MSA text that expresses LD into another national language. Actually, many resources, such as a bilingual dictionary, can be used to translate dialects into their original languages. Building such dictionaries plays a crucial role in Natural Language Processing (NLP) applications not only in machine translation but also in named entity recognition and cross-lingual information retrieval. The final purpose of this work, regardless of language, is to explain the general method for creating a bilingual dictionary for translating dialects into their original languages. The method also takes into account the availability of preexisting monolingual dictionaries for original languages. In this research, a method for creating a bilingual dictionary for LD-MSA translation is given as a case study.

2 RELATED WORK

Many researchers have recently become interested in translating AD into MSA. To carry out their researches, they employed various approaches to build their parallel, bilingual, and monolingual dictionaries that are crucial for building machine translation systems. This section will discuss some of the most significant researches on creating dictionaries or corpora for translation between AD and MSA. (Kchaou et al., 2020) published a TD-MSA parallel corpus in 2020, which was collected using a variety of resources. The Parallel Arabic DIalectal Corpus (PADIC) is the first resource, which is a parallel corpus that combines Maghreb dialects (Algerian, Tunisian, and Moroccan), Levant dialects (Palestinian and Syrian), and the MSA (Meftouh et al., 2015). Multi Arabic Dialect Applications and Resources (MADAR), a TD-MSA parallel corpus (Bouamor et al., 2018), is the second resource. They then gathered text from the Tunisian corpus CONSTitution (TD-CONST), which contains the Tunisian constitution written in MSA and translated into the dialect of Tunisian. The Tunisian social media corpus COMments (TD-COM) is another resource that includes 900 Facebook comments that were then translated into MSA by a native speaker. Finally, they created a TD-MSA bilingual dictionary by aligning the collected parallel corpora. Starting with the two monolingual morphological dictionaries for TD and MSA, (Sghaier et al., 2020) made a great effort to generate the necessary resources from scratch. To map TD words to their MSA equivalents, a bilingual lexicon dictionary was built.

In 2012, (Salloum et al., 2012) presented their Elissa Rule-Based Machine Translation (RBMT) system, which allowed for the translation of a set of Arabic dialects into MSA utilizing AD-MSA dictionaries such as the Tharwa dictionary and other dictionaries they built. (Diab et al., 2013) Presented Tharwa in 2013, a three-way, large-scale lexicon that encompasses Egyptian Arabic, Modern Standard Arabic, and English. The Tharwa is the first three-way electronic resource for DA that includes rich and deep linguistic information for each entry. Egyptian Arabic is the resource's first pilot dialect, with intentions to expand to other Arabic dialects. The Tharwa were gathered from a variety of sources, both manually and automatically. (Tachicart et al., 2014), introduced their machine translation, which combines a rule-based approach and a statistical approach, using tools designed for Arabic standard and adapting them to the Moroccan dialect. To collect their bilingual dictionary corpus, they used the writings of some television production scenarios and some MSA dictionaries. The extension of the bilingual dictionary was done by collecting additional online resources to ensure maximum coverage of the vocabulary of the Moroccan dialect.

In 2018, (Mubarak et al., 2018) presented a parallel corpus called Dial2MSA, which contains dialectal Arabic tweets in four main Arabic dialects (Egyptian, Maghrebi, Levantine, and Gulf) and their corresponding MSA translations. The tweets were collected from Twitter, and then a set of distinctive words for each dialect were filtered. The crowdsourcing platform (CrowdFlower) was then utilized to hire native speakers of each dialect to translate each tweet into its MSA. The final corpus contains 16,000 Egyptian-MSA pairs, 8,000 Maghrebi-MSA pairs, and 18,000 of Gulf-MSA and Levantine-MSA pairs. In 2022, Torjmen, Roua, and Kais Haddar created a bilingual dictionary from various TD-MSA corpora. The TD-MSA bilingual dictionary has 4417 entries and generates approximately 174, 000 forms using derivational and inflectional grammars (Mubarak et al., 2019).

3 LEXICAL RELATIONS AND VARIATIONS

In linguistics, words consist of three components which are spelling, pronunciation, and meaning. See figure 1, where spelling is a set of letters used to represent the basic sounds of a word, while pronunciation is a sound that refers to how a word is pronounced, whereas meaning is what a word means. In semantics and pragmatics, meaning is the message conveyed by words and text in a context. Although, DL and MSA vowels are often omitted in spelling, they are correctly pronounced by readers who are able to recognize them through the context. However, the following terms which deal with how words are spelled, pronounced, and meant, show how MAS and LD words are related. Understanding them better aided us in coming up with an idea for constructing our bilingual dictionary.

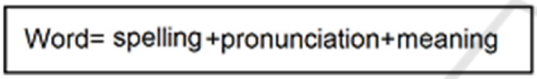

Figure 1: Three components of words.

- **Homographs** are words that share the same spelling but have different meanings, regardless of how they are pronounced. For example, the MSA word "قنينة", means "a bottle" and the LD word "قنينة", means "beautiful" for single feminine. They have the same spelling but are different in pronunciation and meaning.

- **Homophones** are usually defined as words that share the same pronunciation but have a different meaning, regardless of how they are spelled. Here is an examples, the MSA word "زبون" means "a customer" and the LD word "زبون" means "a folk costume". The words are pronounced and spelled the same but have a different meaning. They are also both homophones and homonyms.

- **Heteronyms** are words that are spelled identically but have different meanings when pronounced differently. Two words can be homographs, but not homophones. For example, the MSA word "بياض", pronounced "byAad", means "whiteness". However, the LD word "بياض", pronounced "byAd", means "charcoal".

- **A polysemy** refers to a single word or phrase that can be used in different contexts to express two or more different meanings. The word "حقها" is a case of polysemy, because its meaning is changed by context. For example, the LD expression "قداش حقها" means "how much is it" and the other MSA expression " من حقها" means "she has a right", the word "حقها" has different meanings in both expressions which are in English "price" and "rights" respectively. In addition, the word "حقها" has another meaning in LD which is in English "see her".

- **Homonyms** are words that are pronounced or spelled the same way as other words but have a different meaning. When homonyms have the same sound, they are called homophones, and if they have the same spelling, they are called homographs. Therefore, it is possible for a homonym to be both a homophone and a homograph. Words that have the same sound and different spellings do not exist in both LD and MSA because they are phonetically consistent.

- **A synonym** is a word or phrase having the same or nearly the same meaning as another word or phrase in certain contexts. For example, the MSA word "كأس", means "a cup" and the LD word "طاسة" which also means "a cup".

- **An antonym** is a word or phrase that means the opposite or nearly the opposite of another word or phrase. For example, the MSA word "طويل" means "tall" is the opposite of the LD word "دربة" means "short".

4 DICTIONARY CONSTRUCTION METHOD

For corpus collection, we used the concepts of words axioms introduced in the previous section to build our bilingual dictionary. Linguistically, there is no more than one word in a particular language having the same three components: spelling, pronunciation, and meaning. Each word in a given language has its own unique components. The main idea behind our method is to make use of the lexical

Table 1: The relationships between words in both sentences, LD (Sen1) and MSA (Sen2).

Spelling	Pronunciation	Meaning	English Meaning	Sen2	Sen1	No
different	different	same	the eggs	البيض	الدحي	1
same	same	same	and	و	و	2
different	different	same	oranges	البرتقال	الليم	3
same	same	same	in	في	في	4
same	same	same	the fridge	الثلاجة	الثلاجة	5

similarities between dialects and their original languages, as well as the availability of monolingual dictionaries for original languages which can be used to deal with shared words. In our method, LD words are mainly divided into two categories: shared words and other words. Shared words are LD words that are also used in MSA, while the other words are LD words that are not used in MSA. Usually, many common words are shared between dialects and their original languages. For example, the LD phrase "الدحي والليم في الثلاجة" which is in English "the eggs and oranges are in the fridge" and its translation in MSA "البيض والبرتقال في الثلاجة" share some words. Shared words are words that appear in both the text and its translation in terms of meaning and spelling. In the previous example, the words "الثلاجة", "في" and "و" that mean respectively in English "the fridge", "in", and "and" are found in both the LD and its MSA translation. Therefore, no translation is required to translate shared words. The rest of the LD words are not shared words because they have different corresponding words in MSA. Clearly, the LD words "الدحي" and "الليم" that have equivalent words in MSA are translated by mapping them respectively into their correspondent MSA words "البيض" and "البرتقال". Whereas LD words "الثلاجة", "في" and "و" that mean respectively in English "the fridge", "in" and "and are not mapped to MSA because they do not have equivalent words in MSA. The word relationships in the preceding example are shown in Table 1.

Words in LD that have MSA synonyms but differ in spelling can be represented in equation 1: (LD - MSA) and would be added to a bilingual dictionary as headwords or lemmas to be translated. In contrast, words with the same meaning and spelling in both LD and MSA that can be represented in equation 2: (LD ∩ MSA) would not be translated because they are shared words. Therefore, they do not need to be included in a bilingual dictionary. In summary, LD words are divided into two categories. The first category includes words that must be inserted into a bilingual dictionary. These words share words in MSA with meanings but differ in spelling and pronunciation. While the second category includes words that do not need to be added to a bilingual dictionary, these words share words in MSA with spelling and meaning, regardless of how they are pronounced. Equation (1) represents the first categories, while equation (2) represents the second. Figures 2 and 3 show the results of applying equations (1) and (2) to the previous example, respectively. Where A are MSA words and B are LD words.

Equation (1):

$$\{X: x \notin A \text{ and } x \in B \} = B - A \quad (1)$$

Where:
X: Libyan dialect (LD) words that have MSA synonyms with different spellings.

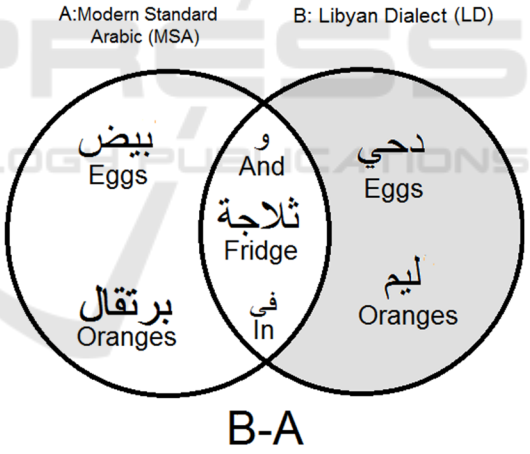

Figure 2: Shows the words that represent bilingual dictionary headwords.

Equation 2:

$$\{x: x \in A \text{ and } x \in B \} = A \cap B \quad (2)$$

Where:
X: Shred words.

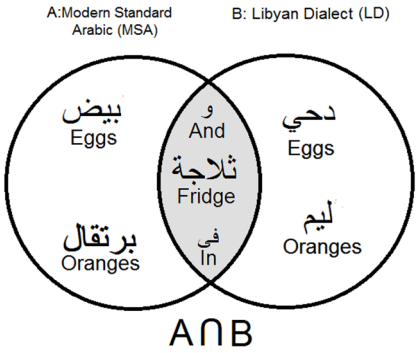

Figure 3: Shows the words that are not included to a bilingual dictionary.

Additionally, to reduce the size of a dictionary and human effort, only the basic forms of words, known as lemmas or headwords would be added to a dictionary. In detail, the inflected forms that are produced by attaching possession affixes and definite articles to a noun lemma are rejected. For example, the LD noun lemma "بيرو" which means in English "a pen" would be added to a dictionary instead of its all inflected forms which are: "بيروي", "بيروك","بيروها","بيروه", "بيروهم", "بيرونا", "ابيرو" and "بابيرو" that respectively mean in English "my pen" , "your pen" ,"her pen" ,"his pen", " their pen", "our pen", "the pen", and "by the pen". The reason why new inflected forms of lemmas are excluded is that their translations can be easily generated. For instance, the translation of the LD word "بيروهم", means "their pen" which is an inflected form of the noun lemma "بيرو", means "a pen" could be obtained by adding the suffix "هم" means "their" to the end of the MSA word "قلم" means "a pen" which is the translation of the LD noun lemma "بيرو" means "a pen". Then the new MSA word "قلمهم" means "their pen" is obtained, which is the correct translation of the LD word "بيروهم" means "their pen". Likewise, all the inflected verb forms: "هدرزت", "هدرز", "هدرزت", "هدرزو", "هدرزنا" and "هدرز" that respectively mean in English "I talk", "he talks", "she talks", "they talk", "we talk", and "you talk" would be excluded, and their LD verb lemma "هدرز" which means in English "talk" would be inserted into a dictionary as a headword instead. It currently contains more than 2000 entries. The size of the dictionary is continuously increasing by adding new words daily.

4.1 Data Collection

The prior work on the creation of bilingual dictionaries can be mainly classified into two approaches: manually and automatically (Dubey et al., 2013). Due to a lack of LD resources, a manual approach is appropriate for building a bilingual dictionary for LD-MSA translation. In order to manually construct a bilingual dictionary, a variety of available resources have been being used to collect LD words, including social media networks, books, and people. Additionally, a web site was implemented to collect LD words from volunteers.

4.2 Data Pre-processing

A small process is done to add LD lemmas to the dictionary. If a given word is not a lemma, then it is converted to its lemma or inflected surface form by humans to be included. The word "قطوسنا " for example, which means "our cat", is the LD word but not a lemma. In this case, the LD "قطوسنا" is converted to its lemma which is "قطوسة" means "a cat", in order to be added to the bilingual dictionary. Furthermore, in some cases, looking up given lemmas in a MSA dictionary is required to confirm whether or not they are shared lemmas, because shared lemmas are ignored. If a given lemma is found in a MSA dictionary, that means it is a shared lemma and is excluded; otherwise, it is added as a headword to a dictionary. A LD-MSA bilingual dictionary contains LD lemmas with their MSA equivalents, as well as some additional linguistic information such as part of speech (POS), gender, number, and verb tense for each dictionary entry. We used the tagsets that were introduced by (Alqrainy et al., 2021) to tag our bilingual dictionary. See Table 2.

Table 2: An example of bilingual dictionary entries.

Bilingual Dictionary Entry	English Meaning	MSA Equivalent	LD Word
$قنينة.<Aj,Sn,Fe>.جميلة	Beautiful	جميلة	قنينة
$هدرز.<Pe,Ma,Th>.تحدث	Talk	تحدث	هدرز
$تريس.<Nu, Pl,Ma>.رجال	Men	رجال	تريس
$غادي.<Ad>.هناك	There	هناك	غادي

5 DISCUSSION

In this section, we discuss the advantages that result from using our method. However, in manual methods, a great human effort has to be made in order to manually create a bilingual dictionary as well as build a such dictionary that needs to collect such a massive quantity of data which is very time-consuming task. Furthermore, big data typically necessitates a significant amount of CPU time. To explain how our method mitigates the drawbacks of

conventional method, for this purpose, a set of LD candidate sentences were randomly picked up from the LD Twitter corpus (Alhammi et al., 2018), and ten of them were selected to serve as a sample to evaluate our method.

Table 3 shows the ten LD sentences that were used to evaluate our method. The words that have an underline are LD words and are not used in MSA. The remaining words are MSA words that are also used in LD, which are called shared words. In fact, LD words are either shared words or LD words that are not used in MSA. In Table 3, the number of LD words among the LD sentences shows considerable diversity. However, in sentences 2, 4, and 9, LD words make up the majority of words, while they are the minority of words in sentences 1 and 3. The number of LD words is almost equal to the number of the shared words in sentences 5, 6, 8, and 10. In contrast, sentence 7 has only shared words. Additionally, some LD words, such as "شنو" which means "what" are repeated more than once in the LD sample. To evaluate our method, two evaluation assumptions are made. In the first assumption, a conventional method is used to build a dictionary. To this end, we suppose that the total dictionary headwords are made up of all the words in the LD sentences in Table 3. In this case, each word represents a dictionary headword. Consequently, the entire dictionary would contain 54 unique headwords or entries, which is the same as the total number of words in Table 3. In the second assumption, the dictionary would be rebuilt by applying our method to the same set of given words as in the first assumption. In detail, 23 out of 54 words are LD words that are not used in MSA, representing a smaller proportion with 42.60%. And 31 out of 54 words are LD words that are also used in MSA, accounting for 57.40% of all words. In our method, the dictionary would only contain LD words that are not used in MSA. Therefore, only 42.60% of words would be included in the dictionary as headwords, while approximately 57.40% of words would be excluded. See Table 4. Clearly, using our method to build the dictionary leads to a significant reduction in dictionary size, human effort and construction time, paving the way for developing machine translation systems to translate dialects into their native languages.

Table 3: LD sentences from a LD Twitter corpus.

English Meaning	LD Sentence	No
For the first time, I know, they look alike.	اول مرة نفطن انهم يشبهوا بعضهم	1
Will I visit you tomorrow?	شن غدوة نديرو دورة عندكم	2
It's OK, what you say does not mean that.	ياهي تمام□كن كلامك ما يوحي بشي هدا	3
It is very hot and full of flies, is there anything worse than this?	نو وصرهادي و ذبان شن فيه اسواء من هكي؟	4
You saw him playing, did not you?	حاقيته هو يلعب ولا شنو؟	5
Why is the country's situation becoming so bad?	كنها□بلاد قعدت شينة هكي؟	6
Honestly, T.V program needs some spice.	بصراحة□برنامج يحتاج ملح	7
men, how are you?	شنو حا□كم يا تريس	8
Do you stay for me, to talk to me?	قاعد على خاطري باش يتهدرز معاي؟	9
The issue is that the short boy does not understand anything.	□مشكلة ان□درية مش زابط	10

Table 4: Applying the evaluation assumptions to the LD sample.

Assumption 1	Assumption 2		
LD Sample	LD Sample		
	Equation 2 (LD ∩ MSA)	Equation 1 (LD-MSA)	
اول –مرة- نفطن-انهم -يشبهوا –بعضهم -شن –غدوة- نديرو -دورة –عندكم – باهي -تمام□كن -كلامك -ما - يوحي -بشي –هدا –نو-صرهادي-و -ذبان -فية -اسواء -من -هكي -حاقيته -هو -يلعب -ولا –شنو- كنها □بلاد -قعدت -شينة -بصراحة □برنامج -يحتاج –ملح- حا□كم -يا -تربس - قاعد -على -خاطري -باش -بتهدرز –معاي-□مشكلة - ان -□درية -مش -زابط	اول –مرة- انهم -يشبهوا –بعضهم - عندكم -تمام□كن -كلامك -ما -يوحي - بشي –هدا -و-فية -اسواء -من -هو - يلعب □بلاد -شينة -بصراحة □برنامج -يحتاج –ملح- حا□كم -يا -قاعد -على - □مشكلة -ان	نفطن -شن –غدوة- نديرو -دورة - باهي –نو-صرهادي -ذبان - هكي -حاقيته -ولا –شنو -كنها - قعدت -خاطري -باش -بتهدرز – تريس –معاي-□درية -مش -زابط	**Candidate Headwords**
54	31	23	**Words number**
included	excluded	included	**Case**
100%	57.40%	42.60%	**Percentage**

6 CONCLUSION AND FUTURE WORK

This paper proposes a method that can be used to build a bilingual dictionary at a word level to translate dialects into their original languages. However, this work also provides a bilingual dictionary that can be used as an essential tool for any LD-MSA machine translation system. In comparison to MSA, the texts available in electronic form for LD are extremely limited. LD also lacks basic NLP tools and resources. In general, creating a bilingual dictionary manually has many cons. It is a costly, complex, and time-consuming task that requires a lot of human effort. To collect dictionary words, crowdsourcing techniques are used where individuals are truly engaged in linguistics. To evaluate our method, an assumption was made to show the advantages of our method. It yields several benefits compared to conventional approaches. It is clear that it effectively reduces dictionary entries, human effort, building time, and CPU time. Broadly, our method might be applied to creating bilingual dictionaries for any machine translation systems that translate dialects into their main languages, particularly Arabic dialects.

In the future, we intend to develop an LD word segmenter which is essential for not only matching given words to dictionary entries but also a variety of NLP applications. For mid-term planning, we will create tools that will be critical in the development of advanced LD-MSA machine translation systems.

REFERENCES

Zaidan, Omar F., and Chris Callison-Burch. 2014. "Arabic dialect identification." Computational Linguistics 40.1: 171-202.

Ghoul, Dhaou, and Gaël Lejeune. 2019. "MICHAEL: Mining character-level patterns for Arabic dialect identification (MADAR Challenge)." Proceedings of the Fourth Arabic Natural Language Processing Workshop.

Abdulaziz, Ashour Saleh. 2014. Code switching between Tamazight and Arabic in the first Libyan Berber news broadcast: An application of Myers-Scotton's MLF and 4M models. Portland State University.

Kchaou, Saméh, Rahma Boujelbane, and Lamia Hadrich Belguith. 2020. "Parallel resources for Tunisian Arabic dialect translation." Proceedings of the Fifth Arabic Natural Language Processing Workshop.

Meftouh, Karima, et al. 2015. "Machine translation experiments on PADIC: A parallel Arabic dialect corpus." Proceedings of the 29th Pacific Asia conference on language, information and computation.

Bouamor, Houda, et al. 2018. "The MADAR Arabic Dialect Corpus and Lexicon." LREC.

Sghaier, Mohamed & Zrigui, Mounir. 2020. Rule-Based Machine Translation from Tunisian Dialect to Modern Standard Arabic. Procedia Computer Science. 176. 310-319. 10.1016/j.procs.2020.08.033.

W. Salloum and N. Habash. 2012. “Elissa: A Dialectal to Standard Arabic Machine Translation System,” 24th International Conference on Computational Linguistics (COLING), p. 385–392, 8-15 December.

Diab, M., et al. 2013. Tharwa: A multi-dialectal multi-lingual machine readable dictionary.

R. Tachicart and K. Bouzoubaa. 2014. “A hybrid approach to translate Moroccan Arabic dialect,” Proceedings of the 9th International Conference on Intelligent Systems: Theories and Applications (SITA-14), p. 1–5, 7-8.

Mubarak, Hamdy. 2018. "Dial2msa: a tweets corpus for converting dialectal arabic to modern standard arabic." Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), OSACT2018 Workshop.

Torjmen, Roua, Nadia Ghezaiel Hammouda, and Kais Haddar. 2019. "A NooJ Tunisian Dialect Translator." International Conference on Automatic Processing of Natural-Language Electronic Texts with NooJ. Springer, Cham.

Dubey, Ajay, and Vasudeva Varma. 2013. "Generation of bilingual dictionaries using structural properties." Computación y Sistemas 17.2: 161-168.

Alqrainy, Shihadeh and Alawairdhi, Muhammed. 2021. "Towards Developing a Comprehensive Tag Set for the Arabic Language" Journal of Intelligent Systems, vol. 30, no. 1, pp. 287-296. https://doi.org/10.1515/jisys-2019-0256

Alhammi, Husien Albashir, and Ramadan Alsayed Alfard. 2018. "Building a twitter social media network corpus for libyan dialect." International Journal of Computer Electrical Engineering 10.1.