Journal of Languages & Translation
P-ISSN: 2716-9359 E-ISSN: 2773-3505
Volume 04 Issue 02 July 2024 pp.161-174

Building the Oranian-English Parallel Corpus: Methodology and Compilation Process

Abdelbasset Dou[1]
DSPM Research Laboratory,
Abdelhamid Ibn Badis University of Mostaganem, Algeria
Email: abdelbasset.dou.etu@univ-mosta.dz

Khalida KISSI
Higher Teacher Training School of Oran (ENSO), Algeria
Email: khalida.kissi@ens-oran.dz

Received *20/04/2024* ***Accepted*** *24/05/2024* ***Published*** *01/07/2024*

Abstract

The scarcity of linguistic resources poses a major challenge for automated translation and processing of dialects. These resources are crucial for natural language processing experts conducting research on dialect recognition, processing, and machine translation. This paper describes the compilation of a dataset for an Algerian low-resource language as it emphasizes the importance of developing resources for Algerian dialects. It examines existing relevant corpora and details the creation process and unique features of the pioneering Oranian-English Parallel Corpus (OEPC). OEPC is the first parallel corpus built from scratch that pairs an Algerian dialect with its English counterparts. The paper outlines the criteria and steps involved in compiling a monolingual corpus for the Oranian dialect (ORN), including data sources and formats. ORN comprises 8500 sentences, which were then translated into English to form OEPC. This valuable linguistic resource is a product of the ERAD project, an initiative aimed at providing NLP professionals with diverse Algerian mono-, multi-, and cross-dialectal corpora. The paper also explains the data compilation and augmentation techniques used to expand the project's outputs.

Keywords: ERAD; machine translation; Oranian dialect; OEPC; parallel corpus.

[1] Corresponding author: Abdelbasset DOU, khalida.kissi@ens-oran.dz.

Introduction

The computer revolution has not only transformed research but has also opened up new avenues for exploring the human mind. Computational skills and tools, particularly in language processing, have revolutionized lexicography, language identification, automated translation, and related techniques. This revolution has led to advanced processing methods, including rule-based, statistical, and neural network methodologies, which greatly contribute to research goals. However, meaningful results and sound judgments depend on the development of machine-readable data, known as language resources (Haddow et al., 2022). Within computational linguistics, lexical resources play a crucial role. These resources enhance Natural Language Processing (NLP) research and applications by providing lexical and semantic knowledge. Language resources are essential for extracting and utilizing relevant information to improve lexical and semantic understanding in NLP.

Today, daily communication largely involves dialects rather than standard languages. Informal spoken and written linguistic variations are prevalent in television and social media. The computational processing of Modern Standard Arabic (MSA) has been studied since the 1970s (Farghaly & Shaalan, 2009). However, the prevalence of dialectal Arabic has shifted the focus of many researchers toward non-standard varieties (Zaidan & Callison-Burch, 2014; Harrat et al., 2015; Habash, 2021). In the past two decades, there was a surge in studies dedicated to processing and translating dialects. Arabic dialects, considered low-resource languages, have become a focal point for NLP tasks and advancements. However, the lack of robust linguistic coding hinders their performance in dialect identification, machine translation, speech recognition, and sentiment analysis (Alnassan, 2023). The scarcity of data for training intelligent systems on dialects has sparked diverse research endeavours in computational linguistics and NLP.

Several research projects have begun to experiment with NLP using a limited number of existing corpora for Algerian dialects. This highlights the ongoing need for additional data to further improve the processing of Algerian dialects. A dedicated initiative has been launched to facilitate machine translation experiments on the primary dialects in Algeria. This initiative involves creating relevant datasets by recording and transcribing naturally occurring linguistic practices. The resulting outputs include language corpora for selected Algerian dialects and parallel corpora with English as the target language. This paper aims to address the challenges of processing Algerian dialects, provide an overview of relevant corpora representing Algerian dialects, and discuss machine translation experiments conducted on them. Finally, the paper outlines the objectives of the ERAD project and explains the compilation process of ORN, a monolingual corpus of the Oranian dialect, and OEPC, the first Oranian-English parallel corpus.

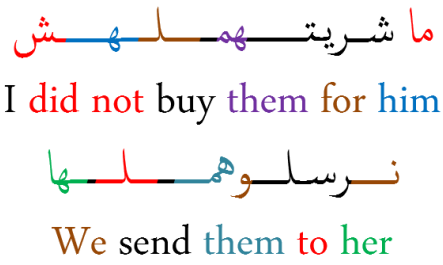
1. Challenges of Processing the Algerian Dialect(s)

Algerian colloquial language is part of the Maghrebi dialects and is spoken by over forty million people in North Africa. The Maghrebi dialects are distinct from eastern varieties such as Gulf, Levantine, and Egyptian. The Algerian dialects, in particular, are more complex due to Algeria's sociolinguistic situation (Morsly, 1986). Additionally, the influence of Arabic, Latin, and Turkish on Algeria can be seen through various historical factors (Chami, 2009). As a result, the Algerian dialects differ from MSA in terms of vocabulary, phonology, morphology, and syntax. However, computational linguists have rarely addressed this level of complexity (Souag, 2006; Saadane & Habash, 2015). Sociolinguists and linguists have classified Algerian dialects based on geographical areas, provinces, and ethnicity (Derradji et al., 2022).

The lack of language resources, orthography, and morphosyntactic aspects greatly impact the accuracy of dialect identification and machine translation. The size of the language resources used during the machine learning phase plays a significant role in these processes. When it comes to processing Arabic dialects, a major challenge is the scarcity of preserved examples available in blogs, websites, and social media. However, existing written forms, despite their inability to accurately represent pronunciation, do contain distinct terms specific to each dialect. Initial research on processing Arabic dialects was conducted by Habash (2010), with subsequent efforts focusing on building resources for Middle Eastern dialects, particularly the Egyptian dialect (Shoufan & Al-Ameri, 2015). While Algerian dialects have been the subject of numerous NLP studies, there is still a pressing need for resources and improved machine translation for these dialects.

In most Arab countries, diglossia, which refers to the coexistence of multiple varieties or languages, is prevalent alongside MSA. The latter is used in formal education, academia, religion, the media, and constitutional and regulatory documents. Conversely, the written form of Algerian dialects is informal and lacks regulation. Algerians commonly use either Arabic characters or transliterated Latin characters, known as Arabizi, when writing the dialect on social media (Bies et al., 2014). Arabizi is predominantly used by Algerians in online communication. Since there are no established rules for writing the dialect, it is challenging to determine the correctness of a written form. Additionally, as vernacular Algerian dialects are not taught or used in formal contexts, the choice of written form depends on the perspective of the language user. Nevertheless, language resource builders and corpus compilers may adopt specific rules to accurately represent the dialect in written texts.

Due to the scarcity of written forms and theoretical linguistic studies on Algerian dialects, computational linguists cannot rely on a rule-based approach. Additionally, such an approach would fail to capture all the features of the various sub-dialects. From the lexical level to the morphological level, the complexity of processing increases as more inflected forms are added to a lemma. Arabic dialects, with their rich morphologies, present several challenges in automatic processing. There are several shared inflectional morphology features among Maghrebi dialects. Below, we present two examples of the morphological complexity and diversity found in Algerian dialects: "ماشريتهملهش" (*mā shrithmlhsh*) and "نرسلوهملها" (*nerslwhmlhā*).



MSA, unlike most Arabic dialects, does not have negation suffixes. Additionally, when written, prepositional phrases are attached to the verb. Numerous studies have addressed this challenge and have achieved high-quality lemmatization (Habash et al., 2013; Zalmout & Habash, 2019; Hashem, 2021). Moreover, various morphological analysers have been developed for processing Arabic dialects, such as MAGEAD (Habash & Rambow, 2006), ADAM (Salloum & Habash, 2014), and YAMAMA (Khalifa et al., 2016).

As for Arabic dialects' machine translation, there is a limited number of relevant works that focus on the dialect as a source. This is primarily due to the complex nature of processing all the dialect's syntactic features. Therefore, the establishment of reliable and robust language models for rule-based or statistical machine translation systems necessitates the availability of large-scale, clean datasets. Having tools that generate syntax will enhance dialect processing and reduce its challenges as a source. However, Harrat et al. (2019a) note that most Arabic dialects currently lack such tools. Conversely, Habash et al. (2022) have recently proposed a new approach to modelling Arabic morphology with a specific emphasis on multi-dialectness.

The machine translation of Arabic dialects has emerged as an area of interest for NLP researchers and computational linguists. Harrat et al. (2019a) and Babaali and Salem (2023) provide comprehensive reviews of experiments conducted in this field. In particular, Harrat et al. (2018) contextualize their NLP research review within the framework of Maghrebi dialects. In the context of this survey, it is crucial to specifically evaluate machine translation of Algerian dialects based on existing corpora.

2. Machine Translation of the Algerian Dialects

Machine translation of Arabic dialects has seen various experiments recently. Initially, rule-based approaches faced challenges due to the lack of linguistic tools and time-consuming nature. To make progress, researchers have emphasized the importance of parallel corpora availability. However, data-driven approaches have not been successful due to the lack of data representing pairs of dialects and languages (Harrat et al., 2019a). This shortage of language resources is a common issue for many global dialects. Despite this, using a resourced language as a pivot code to translate low-resourced languages has been attempted in several machine translation experiments, showcasing the connection and similarity between a language and its dialects. Notable works in Arabic dialects machine translation using MSA as a pivot language have been conducted by Sawaf (2010), Salloum and Habash (2013), and Jeblee et al. (2014).

Algerian (sub-)dialects continue to face the challenge of limited language resources. Some dialects have low resources, while others have none at all. However, some researchers have managed to create quality machine-readable datasets for certain Algerian dialects, enabling NLP tasks such as machine translation and transliteration. Harrat et al. (2017) have reviewed how textual resources for Arabic dialects are developed, while Babaali and Salem (2023) have surveyed the machine translation of existing dialectal Arabic corpora. In the case of Algerian dialects, the available corpora are limited and described in Table 1.

Amazouz et al. (2018) have created the most notable speech corpus that includes the Algerian dialect. As for textual corpora, we present only a few corpora with a small number of sentences. Two notable corpora contain code-switched data (Cotterell et al., 2014; Abainia, 2020), and one comparable corpus with the largest size (CALYOU). Additionally, Meftouh et al. (2012) have introduced the first Algerian dialect-MSA parallel corpus by constructing a dictionary of words in Annaba's dialect translated into MSA. Their dataset has since been incorporated into a larger parallel corpus.

Table 1. *Existing Textual Corpora of Algerian Dialects*

Research works	(Corpus) #tokens	Sorce	Script
Cotterell et al. (2014)	19k	News Comments	*Arabizi*, *FR*.
Abainia (2020)	(DZDC12) 44.4k	Social media	*Arabizi*, *FR*.
Abidi et al. (2017)	(CALYOU) 12.7M	YouTube comments	*Arabizi*, *MSA*, *FR*.
Meftouh et al. (2015)	(PADIC) 290k	Recordings	*Arabizi*, *MSA*.
Lichouri and Abbas (2021)	(PADIC 2) 331k	PADIC	*Arabizi*,*MSA*, *Arabized Kabyle*.

Among the highlighted datasets representing the Algerian dialect, PADIC is the most prominent and relevant. In addition to MSA, PADIC includes two Algerian dialects: one spoken in Annaba (ANB) and one in Algiers (ALG). The other dialects represented are Tunisian, Syrian, and Palestinian. PADIC has been expanded to include a Moroccan dialect (Meftouh et al., 2018) and has been used by researchers for dialect identification (Lichouri et al., 2018), machine transliteration (Guellil et al., 2017), and neural machine translation (Slim et al., 2022). Moukafih et al. (2022) recently utilized PADIC as a dataset, applying a multi-task learning approach to improve cross-dialectal neural machine translation. Additionally, a zero-resourced dialect, Kabyle (KAB), written in Arabic, has been added to PADIC (Lichouri & Abbas, 2021). When a corpus is constructed, it receives attention from experts who use it to improve NLP applications for Algerian and Arabic dialects. For example, Harrat et al. (2019b) used both PADIC and CALYOU to enhance an application for dialect morphological segmentation. Other researchers relied on collected documents for processing the Algerian dialect (Adouane & Dobnik, 2017) or used a collection of comments and posts in conjunction with existing corpora (Slim et al., 2020).

The only project that focused on machine translation of Algerian dialects is Torjman. This project, launched in 2011 by the Directorate General for Scientific Research and Technological Development (DGSRTD), was led by the Scientific and Technical Research Centre for the Development of Arabic Language (STRCDAL) and funded by the Algerian Ministry of Higher Education and Scientific Research. The Torjman team provided solutions, with PADIC being one of their improved products. However, the cross-dialectal dataset and machine translation experiments under this project did not include foreign languages, or at least the lingua franca. In the context of statistical machine translation, the existing corpora for Algerian dialects lack bilingual texts and are still relatively small. Therefore, building an efficient machine translation system requires larger corpora. In this regard, we advocate for a new dataset building project that includes English within a parallel or multilingual corpus.

4. ERAD Project

4.1. Objectives

This project, called Empowering and Resourcing Algerian Dialects (ERAD)[2], aims to create language resources that accurately represent naturally-occurring linguistic data from various Algerian dialects. The project focuses on Algerian dialects and English as the codes for resourcing, specifically in parallel corpora. Additionally, the project aims to use the datasets created to conduct automated translation experiments, including dialect-dialect and dialect-English machine

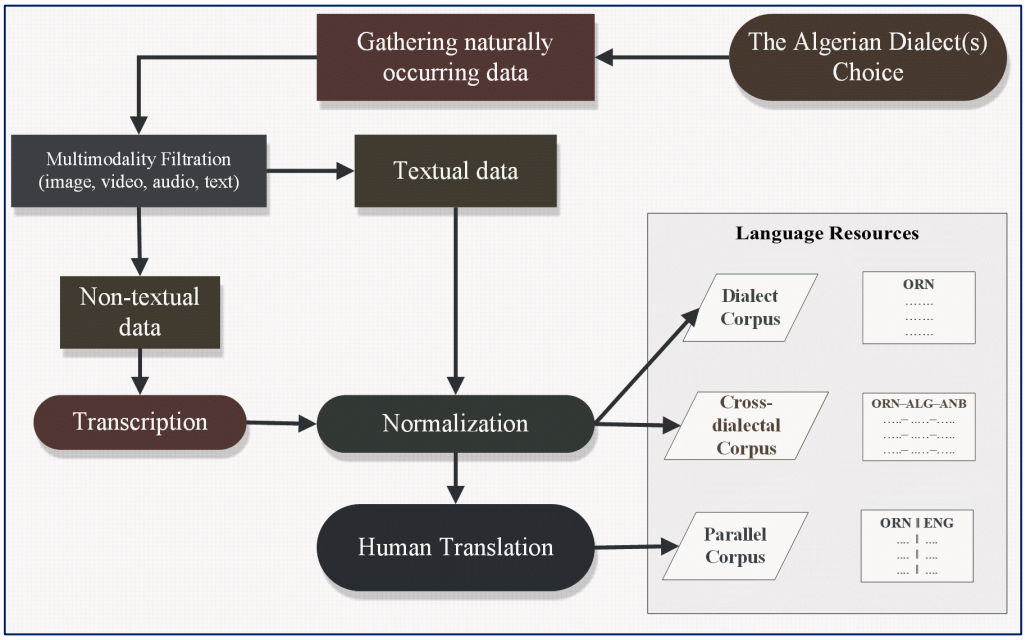
[2] The project homepage https://sites.google.com/view/eradproject

translation. The process and results of this project will be valuable for linguists, computational linguists, translation researchers, and NLP experts, as it will help identify the challenges and promising aspects of resourcing Algerian dialects, constructing corpora, and machine translation. By providing samples from these under-resourced varieties to the research community, the ERAD team hopes to evaluate the quality of inter-dialectal and dialect-into-English translations, which will in turn enhance the practicality of the project.

4.2. Procedure

In a study by Harrat et al. (2017), several pitfalls in creating Arabic dialect corpora were identified. Based on the recommendations from this study, ERAD has made certain disposals and considerations regarding scripting conventionality, normalization processes, and corpora compilation. ERAD's outcomes consist of three types of corpora: mono-dialectal, cross-dialectal, and parallel. The preliminary procedures for data collection and filtration are shared, and although the transcripts are built from scratch, they require revision and undergo the normalization process. Figure 1. illustrates the major steps involved in creating language resources under the ERAD project.

Figure 1. ***Procedure of Textual Corpora Compilation in ERAD***



Given that the main difference among dialects lies in lexical choices, human translation tasks are primarily conducted for the creation of parallel corpora. However, even though aligned sentences exist in cross-dialectal data, they cannot be considered 'parallel' due to the structural closeness of the dialects (Harrat et al., 2015).

4.2.1. Scripting

Several preprocessing steps are necessary to handle the non-traditional volumes of the dialect. While it is common for other Arabic dialects to be written in Arabic script, the Maghrebi dialects are widely written in Arabizi (the Latin script), particularly on the internet. Despite the prevalence of Arabizi usage for Algerian dialects, language resources in Arabic script are still required for NLP experiments. The lack of data is acknowledged as a limitation in multiple studies that rely on corpora of Algerian dialect(s) in any script. Therefore, scripting the dialect in Arabic from scratch is a sensitive yet valuable task.

The spoken versions of these dialects, which are the native languages of the majority, have a non-standard orthography. Transcribing Arabic dialects from speech to text involves representing phonemes as they are pronounced. This means that some non-Arabic sounds are represented differently, for example, /G/ and /P/ are written as "ڤ" and "پ" respectively. However, one might question whether we are standardizing the dialect by using MSA alphabet or transliterating it from Arabizi to Arabic. The former is related to sociolinguistic policy, while the latter involves converting from one code to another (Kaur & Singh, 2014). It is important to note that the Algerian dialect is considered a sub-code of dialectal Arabic, regardless of its script when written. Therefore, the scripting process in ERAD follows an MSA-based convention, focusing on orthographic and lexical resources. Only Arabic letters are used in the transcription. Furthermore, idiosyncratic variations or individual usage of dialect orthography do not impact the automatic identification of the dialect.

4.2.2. Normalization

To reduce the occurrence of out-of-vocabulary items (OOVs) in a corpus and ensure consistency in written forms, a normalization task is applied to the raw data. This task is not necessary for datasets created from scratch, as transcription already takes various factors into account. However, if raw texts are not normalized, dialect processing systems may encounter issues (Darwish et al., 2012). The normalization process addresses various writing issues, as shown in Table 2.

Table 2. ***Normalized Written Forms***

Text issues	**Definition**	**Examples**
Elongations	repeating vowel / consonants forms	y‘āwduuuu يعاودووو
Spelling mistakes	dropped / transposed letters	rwāḥu رواحو / lwlbārḥ لولبارح
Abbreviations	(borrowed) short forms	bac باك / CV سيفي
Special tokens	Emotional / ejaculative expressions	Lol لول / hahaha هاهاها

Normalizing the spelling or writing of linguistic units is essential to facilitate accurate searching and usage of the corpora. Additionally, normalizing instances like "نيشاااان" and "نيييشان" to the word "نيشان" (in English: Right) does not affect the dialectic distinctiveness of the target variety. While there are no established writing guidelines for Arabic dialects, some researchers have utilized MSA-conformed normalization features (Dahou & Cheragui, 2022). The ERAD project aims to transcribe Algerian dialects, normalize them, and align their written forms with MSA spelling rules. To account for the usage of dialect orthography by non-experts, the corpus includes different forms of several linguistic items based on their frequency of occurrence in the dialect spoken by native speakers.

5. Corpora Compilation

Language resources are the product of gathering and utilizing linguistic forms that represent a specific variety. The ERAD project adopts a systematic approach to create datasets. The construction of language corpora involves three primary phases: selecting raw data, transcribing it, and normalizing it. The ERAD project follows the same formal process. Furthermore, mono-dialectal corpora are translated by humans and then aligned with parallel corpora, where chunks

or sentences in the source and target languages are placed side by side. Pre-processing is necessary before training the parallel texts for automated translation and evaluation.

In any cross-dialectal or parallel corpus, there must be a source language. In this case, the dialect of Oran, known as ORN, is used as the source. ORN comprises naturally occurring linguistic instances from the inhabitants of Oran city in western Algeria. The Oranian variety is encountered in various modes of everyday communication and media discourse, including text, audio, and audio-visual data (statistics are shown in Table 3.). These have been collected and transcribed to create a corpus consisting of 8.5k sentences.

Table 3. ***Sources of ORN data***

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

The table shows the formats and sizes of the sources. As distributed, texts, audios, and videos that represent the native Oranian dialect from online sources comprised for the corpus construction from scratch. That is, the data was not collected from a pre-existing database. The table shows that the corpus contains more sentences from video sources (4500) than from text (2500) and audio (3000). Videos were the most authentic sources where the identification of the dialect and its users is easier. Indeed, this suggests that the video data contains a large portion of Oranians' speech that was transcribed into text. The raw sentences will be pre-processed in terms of orthography and normalization in order to compile a final clean dataset.

Compared to Middle Eastern dialects, Algerian dialects are distinguished by their multilingualism and incorporation of French vocabulary. The Oranian dialect, in particular, incorporates words from Latin, Turkish, and Berber origins. Our methodology entails transcribing the Oranian corpus using MSA orthography, despite the presence of French words or expressions in the dialect. If the words have no equivalents in the dialect, the common alternatives in French are included in the ORN data transcription. However, certain French words without equivalents are included in the corpus and transcribed using Arabic script, particularly in online communication. This meticulous process is further elucidated with illustrative examples in Table 4.

Table 4. ***Inserted and Non-Inserted Lexis in ORN***

	Latin Lexis	**Latin Lexis in Arabic Script**
French	arbre/ chanter/ exactement/ fatigue.	سبيطار/ بورطابل/ طاكسي/ كارتي
(English)	tree/ sing/ exactly/ tired.	Neighborhood/ taxi/ mobile/ hospital
	Non-Inserted in ORN	***Inserted in ORN***
Reasons	Oranians have alternatives to these words. They may appear in Arabizi.	Such words are frequently used, have less or no alternative, and are commonly written in Arabizi.

In the ORN scripting process, conventional spelling utilizes Arabic letters to transcribe linguistic elements, irrespective of their variational pronunciation by Oranians. This approach will also be employed for all forthcoming datasets within this project. The ORN dataset is meticulously constructed, with native Algerians taking responsibility for translating it into English. They meticulously review the transcription in the Oranian dialect (ORN) and provide the corresponding English equivalents (ENG). Through the alignment of sentence pairs, the inaugural Oranian-English parallel corpus (OEPC) was established, as depicted in Figure 2.

Figure 2. ***A Sample of Aligned Sentences in OEPC***[3]

ORN	ENG
نزل بركا ما تنقز في السرير ولا نقول لمك	get down stop jumping on the bed or I tell your mom
مكانش منها هادي قاع ما صرانش	this is impossible it never happened
كانت تقارعلي خالتي برا	my aunt was waiting for me outside
علاه راكي توغي ؟	why are you shouting ?
مين ذاك تبحرلي نعاود من لول	sometimes I feel lost I start over
راكم في شحال ساكنين هنا	you are living here for a long time
حمبوكي جيبيلي الما راني ميت بالعطش	please bring water for me I am thirsty to death
سمعت الحس رقبت	I heard noise so I checked
حميت عليه الزونقا و ما خافش	I shouted at him and he didn't fear
قعدو غير الشيبانيات المقرقبات	only old senile ladies stayed
بركا ما ترطبلهم قولهالهم نيشان و صاي	do not suck up to them , tell them directly and that's it
كون غير جات ساجية كيما الشيرات	if only she was quick-witted like the girls

Based on statistics from OEPC, as shown in Table 5., the average number of words per sentence is 5.2 in ORN and 7.3 in ENG. Furthermore, the percentage of distinct items is 31.14 in ORN and 31.09 in ENG.

Table 5. ***Major Statistics of OEPC***

Corpus	Sentences (Avg. Length)	#Words	#Distinct
ORN	8.5K (5.2)	53,255	16,584
ENG	8.5K (7.3)	62,361	19,392

While the size of OEPC may be sufficient for initial machine translation experiments, it falls short of the large-scale datasets that yield high scores. To enhance Algerian dialects machine translation, an extension of OEPC will be created based on previous research works and the same procedural plan of ERAD project.

[3] A sample of OEPC is available at https://sites.google.com/view/eradproject/products/parallel-corpora

Conclusion

This paper emphasizes the significance of language resources in enhancing NLP experiments on Algerian dialects. It provides a concise overview of machine translation of Algerian dialects and introduces the ERAD project. The ERAD project encompasses its initial mono-dialectal corpus (ORN) and parallel corpus (OEPC). The project's objective is to document and preserve a substantial number of instances of Algerian dialects in textual format for researchers to utilize. The datasets can also be converted into audio recordings for training dialect speech recognition. The funding for the ERAD project will be sourced from donations by researchers and their institutions when they incorporate it into a national research project. This support will enable ERAD products to be assigned identification schemas in the International Standard Language Resource Number (ISLRN) for proper usage and referencing.

The preliminary experiments underscore the necessity of creating dictionaries and converting textual corpora into speech corpora. To enhance the productivity of ERAD, researchers should utilize automatic data collection and pre-processing more specifically. ERAD's datasets will be beneficial for NLP experts in tasks such as dialect identification, dialect-dialect and dialect-English machine translation, testing tools for dialect (de-)romanization, and machine transliteration. Additionally, corpus linguists can rely on these datasets for discourse-related studies, sociolinguistic research, and analysis of language variation. The outcomes of this project will make a significant contribution to our research consortium and the NLP research community in the Arab world.

Bibliography

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

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

Adouane, W., & Dobnik, S. (2017). Identification of languages in Algerian Arabic multilingual documents. In *Proceedings of the third Arabic natural language processing workshop* (pp. 1-8). https://doi.org/10.18653/v1/W17-1301

Alnassan, A. (2023). Automatic Standardization of Arabic Dialects for Machine Translation. *arXiv preprint arXiv:2301.03447*.

Amazouz, D., Adda-Decker, M., & Lamel, L. (2018). The French-Algerian code-switching triggered audio corpus (FACST). In *LREC 2018 11th Edition of the Language Resources and Evaluation Conference*.

Babaali, B., & Salem, M. (2023). Survey of the Arabic Machine Translation Corpora. In *International Symposium on Modelling and Implementation of Complex Systems* (pp. 205-219). Springer, Cham. https://doi.org/10.1007/978-3-031-18516-8_15

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 *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, (pp. 93–103). https://doi.org/10.3115/v1/W14-3612

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

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 workshop programme* (p. 34).

Dahou, A. H., & Cheragui, M. A. (2022). Impact of Normalization and Data Augmentation in NER for Algerian Arabic Dialect. In *International Symposium on Modelling and Implementation of Complex Systems* (pp. 249-262). Springer. https://doi.org/10.1007/978-3-031-18516-8_18

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* (pp. 2427-2430). https://doi.org/10.1145/2396761.2398658

Derradji, Y., Debov, V., Queffélec, A., Dekdouk, D. S., & Benchefra, Y. C. (2002). Le français en Algérie: Lexique et dynamique des langues. *AUF*.

Farghaly, A., & Shaalan, K. (2009). Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing (TALIP)*, *8*(4), 1-22. https://doi.org/10.1145/1644879.1644881

Guellil, I., Azouaou, F., Abbas, M., & Fatiha, S. (2017). Arabizi transliteration of Algerian Arabic dialect into modern standard Arabic. *Social MT*, 1-8.

Habash, N. (2010). Introduction to Arabic natural language processing. *Synthesis lectures on human language technologies*, *3* (1), 1–187. https://doi.org/10.1007/978-3-031-02139-8

Habash, N. (2021). Arabic Dialect Processing. In M. Zampieri & P. Nakov (Eds), *Similar Languages, Varieties, and Dialects: A Computational Perspective*, (pp. 279-302). Cambridge University Press. https://doi.org/10.1017/9781108565080.017

Habash, N., & Rambow, O. C. (2006). MAGEAD: A morphological analyzer and generator for the Arabic dialects. https://doi.org/10.3115/1220175.1220261

Habash, N., Marzouk, R., Khairallah, C., & Khalifa, S. (2022). Morphotactic Modeling in an Open-source Multi-dialectal Arabic Morphological Analyzer and Generator. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology* (pp. 92-102). 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 *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 426-432).

Haddow, B., Bawden, R., Barone, A. V. M., Helcl, J., & Birch, A. (2022). Survey of low-resource machine translation. *Computational Linguistics*, *48*(3), 673-732. https://doi.org/10.1162/coli_a_00446

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

Harrat, S., Meftouh, K., Abbas, M., Jamoussi, S., Saad, M., & Smaïli, K. (2015). Cross-dialectal Arabic processing. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 620-632). Springer, Cham. 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 *18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING)*.

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

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)* (pp. 222-228). IEEE. https://doi.org/10.1109/IALP54817.2021.9675225

Jeblee, S., Feely, W., Bouamor, H., Lavie, A., Habash, N., & Oflazer, K. (2014). Domain and dialect adaptation for machine translation into Agyptian Arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)* (pp. 196-206). https://doi.org/10.3115/v1/W14-3627

Kaur, K., & Singh, P. (2014). Review of machine transliteration techniques. *International Journal of Computer Applications*, *107*(20). https://doi.org/10.5120/18866-0061

Khalifa, S., Zalmout, N., & Habash, N. (2016). Yamama: Yet another multi-dialect Arabic morphological analyzer. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: system demonstrations* (pp. 223-227).

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 *Proceedings of The Fourth International Conference on Natural Language and Speech Processing (ICNLSP 2021)* (pp. 33-38).

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

Meftouh, K., Bouchemal, N., & Smaïli, K. (2012). A study of a non-resourced language: the case of one of the Algerian dialects. In *The third International Workshop on Spoken Languages Technologies for Under-resourced Languages-SLTU'12* (pp. 1-7).

Meftouh, K., Harrat, S., & Smaïli, K. (2018). PADIC: extension and new experiments. In *7th International Conference on Advanced Technologies ICAT*.

Meftouh, K., Harrat, S., Jamoussi, S., Abbas, M., & Smaïli, K. (2015). Machine translation experiments on PADIC: A parallel Arabic dialect corpus. In *Proceedings of the 29th Pacific Asia conference on language, information and computation* (pp. 26-34).

Morsly, D. (1986). Multilingualism in Algeria. *The Fergusonian Impact: In Honor of Charles A. Ferguson on the Occasion of His*, *65*. https://doi.org/10.1515/9783110873641-060

Moukafih, Y., Sbihi, N., Ghogho, M., & Smaïli, K. (2022). Improving Machine Translation of Arabic Dialects through Multi-Task Learning. In *International Conference of the Italian Association for Artificial Intelligence* (pp. 580-590). Springer, Cham. https://doi.org/10.1007/978-3-031-08421-8_40

Saadane, H., & Habash, N. (2015). A conventional orthography for Algerian Arabic. In *Proceedings of the 2nd workshop on Arabic natural language processing* (pp. 69–79). Beijing, China. https://doi.org/10.18653/v1/W15-3208

Salloum, W., & Habash, N. (2013). Dialectal Arabic to English machine translation: Pivoting through modern standard Arabic. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 348-358).

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*.

Shoufan, A., & Al-Ameri, S. (2015). Natural language processing for dialectical Arabic: A survey. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistic (ACL), the Arabic Natural Language Processing Workshop (ANLP)* (pp. 36–48). https://doi.org/10.18653/v1/W15-3205

Slim, A., Melouah, A., Faghihi, Y., & Sahib, K. (2020). Algerian Dialect translation applied on COVID-19 social media comments. In *International Conference in Artificial Intelligence in Renewable Energetic Systems* (pp. 716-726). Springer, Cham. https://doi.org/10.1007/978-3-030-63846-7_68

Slim, A., Melouah, A., Faghihi, Y., & Sahib, K. (2022). Improving Neural Machine Translation for Low Resource Algerian Dialect by Transductive Transfer Learning Strategy. *Arabian Journal for Science and Engineering*, 1-8. https://doi.org/10.1007/s13369-022-06588-w

Souag, M. L. (2006). Explorations in the syntactic cartography of Algerian Arabic. Master Thesis, School of Oriental and African Studies (University of London).

Zaidan, O. F., & Callison-Burch, C. (2014). Arabic dialect identification. *Computational Linguistics*, *40*(1), 171-202. https://doi.org/10.1162/COLI_a_00169

Zalmout, N., & Habash, N. (2019). Joint diacritization, lemmatization, normalization, and fine-grained morphological tagging. *arXiv preprint arXiv:1910.02267*. https://doi.org/10.18653/v1/2020.acl-main.736