Creating Parallel Arabic Dialect Corpus: Pitfalls to Avoid

Salima Harrat, Karima Meftouh, Kamel Smaïli

Creating Parallel Arabic Dialect Corpus: Pitfalls to Avoid

Salima Harrat[1], Karima Meftouh[2], and Kamel Smaili[3]

[1] Ecole Supérieure d'Informatique (ESI), Algiers, Algeria, ENSB *
[2] Badji Mokhtar University, Annaba, Algeria
[3] Campus Scientifique LORIA , Nancy, France

Abstract. Creating parallel corpora is a difficult issue that many researches try to deal with. In the context of under-resourced languages like Arabic dialects this issue is more complicated due to the nature of these spoken languages. In this paper, we share our experiment of creating a Parallel Corpus which contain several dialects and Modern Standard Arabic(MSA). We attempt to highlight the most important choices that we did and how good were these choices.

Keywords: Modern Standard Arabic, Arabic dialects, Dialectal corpora, Parallel corpora

1 Introduction

In overall Arab world, Modern Standard Arabic (MSA) is the official language. It is used in the formal speeches, newspapers, official documents, press releases, education etc... In parallel of this, in everyday life, Arab people use dialects. They are spoken languages highly influenced by MSA but also by other foreign languages like French, Spanish, English and Turkish. These languages, also called vernaculars, have emerged on forums, social media and TV shows; and even in the last years, an important number of movies and series are subtitled in these languages whereas before they were translated to MSA. The diglossia phenomenon [1] is more accentuated on the internet and mobile telephony , and has created new needs in the area of NLP related to Arabic language. In fact, all NLP applications dedicated to Arabic language like machine translation, speech recognition, speech synthesis, sentiment analysis... do not take into account the Arabic dialects. These dialects could be considered as under-resourced languages as they lack NLP resources. In the same vein, dealing with Arabic dialects is more complicated than MSA in NLP area. Indeed, MSA is a natural language with a full writing system including strict rules that cover all orthographic varieties, a rich morphology and a strong syntactic system. In contrast, Arabic dialects are spoken languages with no writing system. They differ from

* Ecole Normale Supérieure Bouzaréah.
[1] A situation in which two distinct varieties of a language are spoken within the same speech community

one Arab country to another and they simplify most Arabic language rules. Creating resources for Arabic dialects presents additional challenges to Arabic NLP. If MSA monolingual corpora for instance are available with very acceptable sizes and many parallel corpora including Arabic are downloadable for free, most of Arabic dialects are still under-resourced in terms of such data. Monolingual and parallel textual corpora for some dialects exist but they are few in number. Their creation is costly in time and effort due to their spoken nature.

In this paper, we describe an experiment of creating a dialectal parallel corpus that includes several Arabic dialects in addition to MSA. This corpus has been created for the purpose of machine translation. We try to evaluate our experiment and highlight the most important conclusions that we made. The rest of the paper is organized as follows: in Section 2, we give an overview of Arabic language and its dialects whereas Section 3 presents the most important works dedicated to create textual resources for Arabic dialect. Section 4 describes the parallel dialectal corpus that we created and in Section 5, we mention the most important choices that we made and how good they were. Section 6 concludes this work.

2 Arabic language and its dialects

Arabic is a generic term that covers: Classical Arabic, Modern Standard Arabic and Arabic dialects. The first term is related to the Arabic used in the Qur'an and in the earliest literature from the Arabian peninsula, and it is still used in the literature. Modern Standard Arabic is a modern variant of classical Arabic. In all Arab countries, it is taught in school and used in written correspondence, religious discourses,... but not in everyday conversations nor in households.

Arabic dialects or colloquial Arabic are spoken varieties of Arabic language. They are influenced by European languages such as French, Spanish, English, and Italian because most of the Arab countries were European colonies during the 19th century.

Arabic dialects are described in the literature according to east-west dichotomy [1]; Maghreb dialects regroup dialects spoken in north Africa: Algeria, Tunisia, Morocco, Libya and Mauritania while Middle-east dialects regroups the dialects spoken in Middle-east countries: Levantine dialect (Syria, Lebanese, Palestinian and Jordan), dialect of Arabian peninsula (Gulf countries and Yemen), Iraqi dialect Egyptian and Sudan dialect.

3 Textual resources for Arabic dialects

Availability of resources is a big issue for NLP applications dedicated to Arabic dialects. In [2], authors drew a survey of available Arabic corpora and lexicons and underlined the lack of this kind of resources for Arabic dialects. Several works are dedicated to build such resources.

Authors of [3] elaborated a Levantine lexicon using transductive learning. For Iraqi dialects, authors of [4] used an annotated recorded speech to build a lexicon,

whereas in [5] a spelling corrector was presented for this dialect. For Tunisian dialect, authors of [6] created a lexicon by converting MSA patterns to Tunisian dialect patterns and then extracting specific roots and patterns from a training corpus that they created for this purpose. This tunisian dialect lexicon was then used to adapt Al-Khalil [7] morphological analyzer. Another work dedicated for tunisian dialect is described in [8], where authors used the Penn Arabic Treebank (PATB)[9][10] to create a Tunisian dialect text corpus and a bilingual dictionary (MSA/Tunisian dialect) by using a transformation method based on the parts of speech of ATB words. In [11] a lexicon for MSA/Egyptian dialect (of Cairo the capital city of Egypt) was created. All dialect entries are mapped to their MSA synonyms and thus they get access to the quite available NLP tools and resources for MSA.

Regards to textual corpora, several works attempted to create this kind of resources. Authors of [12] developed a treebank for levantine dialect which consisted of morphological and syntactic annotation of approximately 26K words of Levantine Arabic conversational telephone speech. Also, a treebank for egyptien dialect was built in [13] in addition to a morphological analyzer. In [14] YADAC (a multi-genre Dialectal Arabic (DA) corpus) is created for Egyptian dialect using Web data from microblogs (Twitter), blogs/forums and on line knowledge market services.

Other important resources for Arabic dialects are multi-dialects corpora. Several researches were dedicated to create such kind of resources like [15] where the web was used as a source to build dialect corpus for Gulf, Levantine, Egyptian and North African dialects, or [16] where authors presented a multi-dialect, multi-genre, human annotated corpus of Levantine, Gulf, Egyptian, Iraqi and Maghrebi dialects. In [17], authors collected also a multi-dialect corpus from Twitter for a large set of Arabic dialects.

In terms of dialect parallel corpora, resources are more scarce because of the required effort to build them. In [18], a multidialectal Arabic parallel corpus is presented, it is a collection of 2K sentences in MSA, Egyptian, Tunisian, Jordanian, Palestinian and Syrian Arabic, in addition to English. The starting point of this work was the 2K egyptian sentences extracted from the egyptian part of the Egyptian-English corpus built in [19]. A parallel dialectal corpus including several dialects and MSA was created in [20] for the purpose of machine translation.

4 Overview of the parallel Arabic dialect corpus

In this section we give a brief description of our parallel corpus and the methodology of its creation. The corpus includes in addition to MSA, fives dialects, namely Algerian dialects (from Algiers the capital city of Algeria and Annaba a city in the east of the country), Tunisian, Palestinian and Syrian.

For creating the Algiers dialect corpus (ALG), we have transcribed the scenarios of films and sitcoms. This transcription allowed the collection of 2,5K sentences that we translated into standard Arabic. A corpus of Annaba di-

alect(ANB) was also created in the same way but by transcribing recordings of the daily life of some people of Annaba (university, doctor's waiting room, households). This transcription allowed to collect 3,9K sentences which were also translated into standard Arabic. In order to increase the size of the two corpora, they were translated each one to the dialect of the other. Thus, at the end of this operation, a first tri-lingual (MSA, ALG, ANB) corpus of 6,4K sentences was created.

In order to introduce the Tunisian, Syrian and Palestinian dialects, we used MSA as a pivot language. The standard Arabic part of the tri-lingual corpus constructed as explained above has been translated to Tunisian, Syrian and Palestinian (see Figure 1). The Tunisian corpus was produced by native speakers

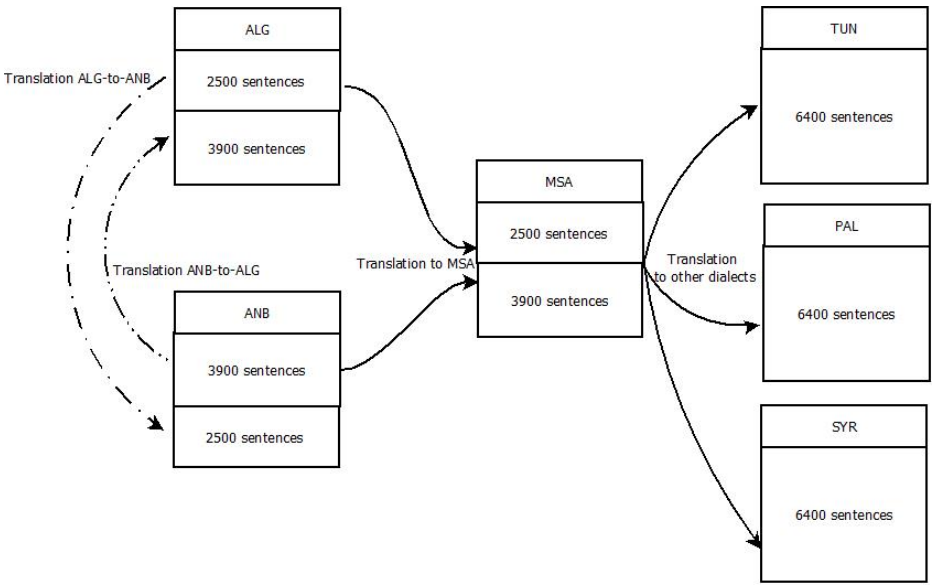

Fig. 1. Parallel Arabic dialect corpus creation

from the south of Tunisia. The Syrian and Palestinian corpora were created in the same way as the Tunisian, except that each one was produced by only two native speakers of these Dialects. The Palestinian dialect concerned by this study is mainly the dialect of Gaza Strip. The Syrian dialect is that of the city of Damascus.

Briefly, the parallel corpus consists of six texts of 6.4K sentences in Algeria dialect (two texts), Tunisian, Palestinian, Syrian in addition to MSA. The average size of each vocabulary corpus is about 9398, while the average number of distinct words is 38539. It is important to note that all operations described above were done manually. It was costly in terms of time and effort. A total of twenty five persons participated to create this corpus, of which twenty for ANB

and TUN, two for both PAL and SYR and one person for ALG. It should be noted that these people had done this work for free.

In language modelling, perplexity is frequently used as a quality measure for language models [21]. In order to have an idea about each monolingual corpus, we computed related perplexity. Each side of the parallel corpus was split into test and training set of 800 and 5600 sentences, respectively. We used tri-grams language models for which we calculated perplexity values as shown in Table 1.

Table 1. Perplexity values computed for each language model of the parallel corpus

Language/Dialect	MSA	ALG	ANB	TUN	SYR	PAL
Perplexity	51.79	57.79	61.49	62.96	59.88	56.09

It is clear that the lowest perplexity is related to MSA, this was expected sine MSA is a natural language with a strong orthographic system. The highest values are those for ANB and TUN. This could be explained by the number of persons who participated to create these two corpora (twenty persons for each corpus). Indeed, Every person used its own dialectal words and wrote them with little regard for orthography.

5 Pitfalls to avoid when creating dialectal corpora

In this section, we discuss the most important points that must be taken into account when creating parallel corpora for Arabic dialects.

5.1 Pivoting throw close languages

Pivoting throw MSA was an intuitive choice since Arabic dialects are variants of this language. Unfortunately, we noticed this pivoting has in some way influenced the translators. Indeed, we observed that they tend to apply the Arabic sentences structure to dialectal ones when translating. The influence of standard Arabic on dialect corpora is clear in all of them translated from MSA (Tunisian, Palestinian and Syrian). We give in Table 2 some examples collected from the parallel corpus where influence of MSA is noticeable.

In the same vein, it is not recommended to translate between dialects especially close ones. In fact, the translator could be easily influenced by the source dialect and unconsciously uses words or terms belonging to the source dialect even if they are rarely used in the target dialect.

We think that pivoting throw a foreign language such as french (for Maghreb dialects) or English (for middle-east dialects) could produce "more dialectal" sentences because the distance between languages imposes the use of purely dialectal words since borrowing words is not acceptable as the case in translating from MSA.

Table 2. Examples of MSA pivoting influence from the parallel corpus

Dialect/Language	Sentence	Meaning
MSA	هذا شيء غير معقول	It is impossible
ALG	ماكانش منها هادي	
ANB	ماكانش منها ادي	
TUN	مهوش معقول	
SYR	هالشي مو معقول بنوب	
PAL	هذا اشي مش معقول	
MSA	و هل في الَامر شك	Is there a doubt
ALG	و هَادي فيهَا هَدرة	
ANB	و ادي فَاهَا هَدرة	
TUN	عندك شك	
SYR	مَا في شك	
PAL	و هو الموضوع في شك	
MSA	شخصيًا لَا يمكن ان امر	Personally I can not pass
ALG	انَا مَا نقدرش نجوز	
ANB	انَا مَا نقدرش نعدي	
TUN	شخصيًا منجمش نمر	
SYR	شخصيًا مَا بقدرش اني امر	
PAL	شخصيًا انَا مَا بمر	

5.2 Choosing theme

Regards to themes, we did not impose any restrictions. Users charged to record conversations were free to record any kind of discussion. We aimed to diversify subjects in order to have as broad a vocabulary as possible. But we noticed that this diversification has generated some conversations in particular domains like religion and studies with specific vocabularies. For example, Arabic people when speaking about religion (what ever is their religion) tend to use standard Arabic. They switch from dialect to standard Arabic spontaneously, especially when they evoke religious texts. Due to that, some passages in dialectal corpora include standard Arabic expressions that translator left unchanged. We give in Table 3 some Arabic religious expressions present in all dialectal corpora.

Table 3. Example of religious expressions present in all sides of the parallel corpus

Expression	Meaning
علامات الساعة	Hereafter signs
سورة الكهف	Cave Surat
بسم الله الرحمن الرحيم	In the name of Allah most gracious most merciful
الرسول عليه الصلاة و السلام	The prophet peace be upon him

In the same context, we noticed also that some conversations were very specific and concerned the area of studies. This is due to the fact that some records related to Annaba dialect were done by students who tended to discuss about subjects related to courses and university in general. These discussions generated some expressions in French language (which is used in university studies in Algeria for technical specialties). These expressions have been translated to MSA and except for the two Algerian dialects texts, these expressions have been left in MSA since no equivalent dialectal words are available because of the poorness of these vernacular languages. We give in Table 4 some examples of these expressions.

5.3 Writing rules

One of the most challenging issue related to Arabic dialects is their writing system. Since they were at a near time only spoken they do not have any orthography to respect. This issue is more complicated then we thought at the beginning of this work. In fact, for transcribing speech to text we have agreed among ourselves to adopt a set of rules like writing words in Arabic if it is possible or writing them as they are uttered in the other case and transcribing non-Arabic phonemes like $/P/$ and $/G/$ as "پ" and "ڤ" respectively. We noticed that these rules have not been fully observed. For instance, we found some dialectal expressions like قلتلك (i said to you) or قَالتلو (she said to him) which had to be written like in Arabic قلت لك and قَالت له in that order. Likewise, foreign letters rule has been applied only in ALG and ANB corpora, for other dialects these letters have not been used. An example of this is the word بورتَابل (mobile phone) from the Tunisian corpus which is originally a french word [2], this word had to be written as it was uttered, that is پورتَابل.

An other important point regards to writing dialect is the numerals. We did not impose any rules for them, they have been transcribed with letters. With a little hindsight, we see that it would have been better if we transcribed them with numerals instead of letters. This will have a positive impact on machine translation scores.

[2] Also widely used in Algeria and Morocco

Table 4. Some examples of french words existing in source dialect corpora and their translation in MSA and target dialect corpora

Dialect/Language	Sentence	Meaning
MSA	الافراط المعرفي	Cognitive overload.
ALG	سورشارج كونيتيف	
ANB	سورشارج كونيتيف	
TUN	الافراط المعرفي	
SYR	الافراط المعرفي	
PAL	الافراط المعرفي	
MSA	من المفروض جميع الاختصاصات تدرس الخوارزميات	Normally in all specialties algorithms are taught
ALG	نورمالون ليزوپسيون كامل يقريوهم لالقوريتميك	
ANB	نورمالون ليزوپسيون كل يقريوهم لالقوريتميك	
TUN	من المفروض جميع الاختصاصات تقرا الحساب	
SYR	لازم كل الاختصاصات تدرس الخوارزميات	
PAL	من المفروض جميع الاختصاصات تدرس الخوارزميات	
MSA	المهم المسابقة	The contest is the most important
ALG	الصح فالكنكور	
ANB	الصح الكنكور	
TUN	المهم المسابقة	
SYR	المهم المسابقة	
PAL	المهم المسابقة	

6 Conclusion

We presented our experiment in the area of building textual resources for Arabic dialects. We attempted to point out the main pitfalls to avoid when dealing with such a problems. First of all, Since MSA is the common point between dialects of Arab countries, we intuitively used it as a bridge between them (Arabic dialect). The influence of MSA on translators was apparent. The facility to keep certain words unchanged was a good option since we noticed that translating between so close languages was a very hard task according to all persons who participated to this operation. Regards to themes, it is recommended to keep them far from specific domains like religion or scientific subjects. Indeed, Arab people when they address religious topics switch to MSA. This generates important MSA passages in dialectal corpora. In this respect and regards to the lexical poorness of Arabic dialects, domain-specific discussions where scientific and technical words are used have to be avoided. In general, Arabic dialect vocabulary does not cover this area and people switch to standard languages like MSA, French and English. For writing Arabic dialects, we noticed that it was difficult to impose strict rules.

People when writing dialect feel free to write how they want. In some way, this is natural. In dialects there are no spelling mistakes since there are no rules.

References

1. Hetzron, R.: The Semitic Languages. Routledge language family descriptions. Routledge (1997)
2. Zaghouani, W.: Critical survey of the freely available arabic corpora. In: Proceedings of the Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools, LREC. (2014) 1–8
3. Duh, K., Kirchhoff, K.: Lexicon acquisition for dialectal arabic using transductive learning. In: Proceedings of the 2006 conference on empirical methods in natural language processing, Association for Computational Linguistics (2006) 399–407
4. Graff, D., Buckwalter, T., Jin, H., Maamouri, M.: Lexicon development for varieties of spoken colloquial arabic. In: Proceedings of the fifth international conference on language resources and evaluation (lrec). (2006) 999–1004
5. Rytting, C.A., Zajic, D.M., Rodrigues, P., Wayland, S.C., Hettick, C., Buckwalter, T., Blake, C.C.: Spelling correction for dialectal arabic dictionary lookup. ACM Transactions on Asian Language Information Processing (TALIP) **10** (2011) 3
6. Zribi, I., Khemakhem, M.E., Belguith, L.H.: Morphological analysis of tunisian dialect. In: International Joint Conference on Natural Language Processing. (2013) 992–996
7. Boudlal, A., Lakhouaja, A., Mazroui, A., Meziane, A., Bebah, M.O.A.o., Shoul, M.: Alkhalil morpho sys1: A morphosyntactic analysis system for arabic texts. In: International Arab Conference on Information Technology, ACIT 2010. (2010)
8. Boujelbane, R., BenAyed, S., Belguith, L.H.: Building bilingual lexicon to create dialect tunisian corpora and adapt language model. ACL 2013 (2013) 88
9. Maamouri, M., Bies, A.: Developing an arabic treebank: Methods, guidelines, procedures, and tools. In: Proceedings of the Workshop on Computational Approaches to Arabic Script-based languages, Association for Computational Linguistics (2004) 2–9
10. Maamouri, M., Bies, A., Kulick, S., Zaghouani, W., Graff, D., Ciul, M.: From speech to trees: Applying treebank annotation to arabic broadcast news. In: LREC. (2010)
11. Al-Sabbagh, R., Girju, R.: Mining the web for the induction of a dialectical arabic lexicon. In: LREC. (2010)
12. Maamouri, M., Bies, A., Buckwalter, T., Diab, M., Habash, N., Rambow, O., Tabessi, D.: Developing and using a pilot dialectal arabic treebank. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC06. (2006)
13. Maamouri, M., Bies, A., Kulick, S., Ciul, M., Habash, N., Eskander, R.: Developing an egyptian arabic treebank: Impact of dialectal morphology on annotation and tool development. In: LREC. (2014) 2348–2354
14. Al-Sabbagh, R., Girju, R.: Yadac: Yet another dialectal arabic corpus. In: LREC. (2012) 2882–2889
15. Almeman, K., Lee, M.: Automatic building of arabic multi dialect text corpora by bootstrapping dialect words. In: 2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA). (2013) 1–6

16. Cotterell, R., Callison-Burch, C.: A multi-dialect, multi-genre corpus of informal written arabic. In: LREC. (2014) 241–245
17. Mubarak, H., Darwish, K.: Using twitter to collect a multi-dialectal corpus of arabic. In: Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP). (2014) 1–7
18. Bouamor, H., Habash, N., Oflazer, K.: A Multidialectal Parallel Corpus of Arabic. In: Proceedings of the Language Resources and Evaluation Conference, LREC-2014. (2014) 1240–1245
19. Zbib, R., Malchiodi, E., Devlin, J., Stallard, D., Matsoukas, S., Schwartz, R., Makhoul, J., Zaidan, O.F., Callison-Burch, C.: Machine translation of arabic dialects. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics (2012) 49–59
20. Meftouh, K., Harrat, S., Jamoussi, S., Abbas, M., Smaili, K.: Machine translation experiments on padic:
a parallel arabic dialect corpus. In: Proceedings PaCLiC 29th Asia Conference on Language, Information and Computation. (2015) 26–34
21. Chen, S.F., Goodman, J.: An empirical study of smoothing techniques for language modeling. Computer Speech and Language **13** (1999) 359–394