A Multidialectal Parallel Corpus of Arabic

Houda Bouamor[1], Nizar Habash[2] and Kemal Oflazer[1]

[1]**Carnegie Mellon University in Qatar**
hbouamor@qatar.cmu.edu,ko@cs.cmu.edu
[2]**Center for Computational Learning Systems, Columbia University**
habash@ccls.columbia.edu

Abstract

The daily spoken variety of Arabic is often termed the colloquial or dialect form of Arabic. There are many Arabic dialects across the Arab World and within other Arabic speaking communities. These dialects vary widely from region to region and to a lesser extent from city to city in each region. The dialects are not standardized, they are not taught, and they do not have official status. However they are the primary vehicles of communication (face-to-face and recently, online) and have a large presence in the arts as well. In this paper, we present the first multidialectal Arabic parallel corpus, a collection of 2,000 sentences in Standard Arabic, Egyptian, Tunisian, Jordanian, Palestinian and Syrian Arabic, in addition to English. Such parallel data does not exist naturally, which makes this corpus a very valuable resource that has many potential applications such as Arabic dialect identification and machine translation.

Keywords: Arabic, Dialects, Parallel Corpus

1. Introduction

The Arabic language today is a collection of variants (i.e., dialects) that are historically related to Classical Arabic, but are the product of intense interaction among the various historical dialects of Classical Arabic, the pre-Islamic local languages in the Arab World (such as Coptic, Berber and Syriac), neighboring languages (such as Persian, Turkish and Spanish) and colonial era languages (such as Italian, French and English). Dialects of Arabic (DA) vary among themselves and the Modern Standard Arabic (MSA) variety, which is the language of media and education across the Arab World. The differences are not only lexical, but also phonological, morphological and to lesser degree syntactic. There are numerous linguistic studies in Arabic dialects, with the comparative studies being limited to small scale and old laborious field method techniques of information collection (Brustad, 2000).

In the context of natural language processing (NLP), some Arabic dialects have started receiving increasing attention, particularly in the context of machine translation (Zbib et al., 2012; Salloum and Habash, 2013) and in terms of basic enabling technologies (Habash et al., 2012b; Habash et al., 2013; Pasha et al., 2014). However, the focus is on a small number of iconic dialects, (e.g., Cairene Arabic). In this paper, we present the first multidialectal Arabic parallel corpus, a collection of 2,000 sentences in Standard Arabic, Egyptian, Tunisian, Jordanian, Palestinian and Syrian Arabic, in addition to English. Since such parallel data does not exist naturally (unlike parallel news, e.g.) this is a very valuable resource that has many potential applications such as Arabic dialect identification and machine translation.

The remainder of this paper is organized as follows. Section 2 discusses the differences between MSA and DA and within DAs . In Section 3, we review the main previous efforts for building dialectal resources. Our approach for building the multidialectal parallel corpus is explained in Section 4. Section 5 presents some preliminary corpus analysis and statistics. Finally, we conclude and describe our future work in Section 6.

2. Arabic Dialect Variation

While MSA is the shared official language of culture, media and education from Morocco to the Gulf countries, it is not the native language of any speakers of Arabic. Most native speakers of Arabic are unable to produce sustained spontaneous discourse in MSA; in unscripted situations where spoken MSA would normally be required (such as talk shows on TV), speakers usually resort to repeated code-switching between their dialect and MSA (Abu-Melhim, 1991; Bassiouney, 2009). Arabic dialects are often classified regionally (as Egyptian, North African, Levantine,

Gulf, Yemeni) or sub-regionally (e.g., Tunisian, Algerian, Lebanese, Syrian, Jordanian, Kuwaiti, Qatari, etc.). There are a number of additional ways to classify the dialects (as per social class or city level).

Arabic is a morphologically complex language that combines a rich inflectional morphology with a highly ambiguous orthography,[1] which poses many challenges for NLP (Habash, 2010). These features are shared by DA and MSA. The differences between MSA and DAs have often been compared to Latin and the Romance languages (Habash, 2006). Arabic dialects differ phonologically, lexically, and morphologically from one another and from MSA (Watson, 2007).

Phonology An example of phonological differences is in the pronunciation of dialectal words whose MSA cognate has the letter Qaf (ق *q*).[2] It is often observed that in Tunisian Arabic, this consonant appears as /q/ (similar to MSA), while in Egyptian and Levantine Arabic it is /'/ (glottal stop) and in Gulf Arabic it is /g/ (Haeri, 1991; Habash, 2010).

Orthography While MSA has an established standard orthography, the dialects do not. Often people write words reflecting the phonology or the history (etymology) of these words. DAs are sometimes written in Roman script (Darwish, 2013). In the context of NLP, a conventional orthography for DA (CODA) has been proposed and instantiated for Egyptian Arabic by Habash et al. (2012a) and was later extended to Tunisian Arabic (Zribi et al., 2014).

Morphology Morphological differences are quite common. One example is the future marker particle which appears as +س *sa+* or سوف *sawfa* in MSA, +ح*Ha+* or رح *raH* in Levantine dialects, +ه *ha+* in Egyptian and باش *bAš* in Tunisian. This together with variation in the templatic morphology make the forms of some verbs rather different: e.g. 'I will write' is سأكتب *saÂaktubu* (MSA), حأكتب *HaÂaktub* (Palestinian), هكتب *haktib* (Egyptian) and باش نكتب *bAš niktib* (Tunisian).

Lexicon The number of lexical differences is quite significant. The following are a few examples (Habash et al., 2012a): Egyptian بس *bas* 'only', طربيزة *tarabayzaħ* 'table', مرات *mirAt* 'wife [of]' and دول *dawl* 'these', correspond to MSA فقط *faqaT*, طاولة *TAwilaħ*, زوجة *zawjaħ* and هؤلاء *hawˆlA'*, respectively. For comparison, the LEV forms of the above words are بس *bas* (like EGY), طاولة *TAwliħ* (closer to MSA), مرة *mart* and هدول *hadawl*.

Syntax Comparative studies of several Arabic dialects suggest that the syntactic differences between the dialects are minor. For example, negation may be realized differently (ما *mA*, مش *mish*, مو *muw*, لا *lA*, لم *lam*, etc.) but its syntactic distribution is to a large extent uniform across varieties (Benmamoun, 2012).

3. Related work

Much work has been done in the context of standard Arabic NLP (Habash, 2010). There are lots of parallel and monolingual data collections, richly annotated collections (e.g., treebanks), sophisticated tools for morphological analysis and disambiguation, syntactic parsing, etc. (Habash, 2010). Efforts to create resources for Dialectal Arabic (DA) have been limited to a small number of major dialects (Diab and Habash, 2007; Habash et al., 2013; Pasha et al., 2014).

Several researchers have explored the idea of exploiting existing MSA rich resources to build tools for DA NLP. Al-Sabbagh and Girju (2010) described an approach of mining the web to build a DA-to-MSA lexicon. Chiang et al. (2006) built syntactic parsers for DA trained on MSA treebanks. Similarly Sawaf (2010), Sajjad et al. (2013) and Salloum and Habash (2013) translated dialectal Arabic to MSA as a bridge to translate to English. Boujelbane et al. (2013) built a bilingual dictionary using explicit knowledge about the relation between Tunisian Arabic and MSA.

Crowdsourcing to build specific resources (e.g., parallel data for translation) for a specific dialect has also been successful (Zbib et al., 2012). Some efforts on dialect identification at the regional level have been done (Habash et al., 2008; Elfardy and Diab, 2013; Zaidan and Callison-Burch, 2013).

In the context of DA-to-English SMT, Riesa and Yarowsky (2006) presented a supervised algorithm for online morpheme segmentation on DA that cut the OOVs by half. Zaidan and Callison-Burch (2011) crawled the websites of three Arabic Newspapers and extracted reader commentary on their articles to build the Arabic Online Commentary dataset. They also collected crowd-driven dialectal annotations on Arabic sentences using Mechanical Turk. More recently,

[1] Short vowels and consonantal doubling are represented with optionally written diacritics in Arabic orthography. This leads to a high degree of ambiguity and also hides some of the differences among dialects and MSA.

[2] Arabic transliteration is presented in the Habash-Soudi-Buckwalter scheme (Habash et al., 2007): (in alphabetical order) *AbtθjHxdðrzsšSDTDˇςγfqklmnhwy* and the additional symbols: ' ء, Â أ, Ǎ إ, Ā آ, ŵ ؤ , ŷ ئ , ħ ة, ý ى.

Dialect/Language	Example
English	*Because you are a personality that I can not describe.*
Modern Standard Arabic	لأنك شخصية لا أستطيع وصفها. *lÂnk šxSyħ lA ÂstTyς wSfhA.*
Egyptian Arabic	لأنك شخصية **وبجد** مش هعرف أوصفها. *lÂnk šxSyħ wbjd mš hς rf ÂwSfhA.*
Syrian Arabic	لأنك شخصية **وعنجد** ما رح أعرف أوصفها. *lÂnk šxSyħ wςnjd mA rH Âς rf ÂwSfhA.*
Jordanian Arabic	انت **جد** شخصية مستحيل اقدر اوصفه *Ant jd šxSyħ mstHyl Aqdr AwSfhA.*
Palestinian Arabic	**عن جد** ماشاء اللّه عليك شخصيتك ما بتنوصف. *ςn jd mA šA' Allh ς lyk šxSytk mA btnwSf.*
Tunisian Arabic	على خاطرك شخصية **بلحق** منجمش نوصفها. *ς lý xATrk šxSyħ blHq mnjmš nwSfhA.*

Table 1: A table comparing the translations for one sentence in the Multidialectal Arabic Corpus. Egyptian Arabic is the original sentence which was translated to MSA, English and the other dialects.

Zbib et al. (2012) demonstrated a crowd-sourcing solution to translating sentences from Egyptian and Levantine into English, and thus built two bilingual corpora. The dialectal sentences were selected from a large corpus of Arabic web text. They argued that differences in genre between MSA and DA make bridging through MSA of limited value.

4. Approach

In addition to Standard Arabic (MSA) and English (EN), our corpus covers five dialects: Egyptian (EG), Tunisian (TN), Syrian (SY), Jordanian(JO) and Palestinian (PA) . The last three dialects represent the Levantine group of Arabic dialects. In the future, we plan to expand this effort to other dialects.

In order to build our corpus, four translators (native speakers of Palestinian, Syrian, Jordanian and Tunisian) were asked to translate 2,000 sentences written in Egyptian into their dialects. Egyptian was chosen as a starting point because it is the most widely understood and used dialect throughout the Arab world. The Egyptian media industry has traditionally played a dominant role in the Arab world. A large number of cinema productions, television dramas and comedies have since long familiarized Arab audiences with the Egyptian dialect. A fifth translators (who happened to be Egyptian) was asked to translate the same text to MSA.

The sentences are selected from the Egyptian part of the Egyptian-English corpus built by Zbib et al. (2012). This corpus was translated to English by non-professional translators hired on MTurk. Since our translators saw the sentences out of context, we provided them with the equivalent ones in English to help disambiguate some readings if necessary.

Every translator was asked to: (a) read the sentences carefully and simply translate them without adding any new information; (b) avoid word by word translation; and (c) be consistent in their orthographic choices and avoid Roman script writing. We asked the translators to be internally consistent in spelling words since there is no standard orthography available for Arabic dialects at this time and we wanted to minimize unnecessary sparsity. We did not provide them with any orthographic guidelines (other than the request for internal consistency). A different approach would have been to collect the dialectal sentences in Arabic script following a general conventional orthography for DA such as CODA. However CODA guidelines at this time only cover Egyptian and Tunisian (Habash et al., 2012a; Zribi et al., 2014).

5. Preliminary Data Analysis

Table 1 illustrates the translations for one sentence in the multidialectal Arabic corpus. This example highlights the many lexical and morphological differences among the different dialects. For example, the Egyptian expression وبجد *wbjd* 'and seriously' was translated into وعنجد *wςnjd* in Syrian, and بلحق *blHq* in Tunisian. The example shows, as well, that there are many shared words that, on their own, cannot disambiguate among the different dialects.

	EG	SY	JO	PA	TN	MSA
# tokens	11,131	11,586	9,866	11,131	10,896	11,048
# unique tokens	4,588	4,167	4,055	3,675	4,483	4,436
# tokens per sentence	9.22	9.61	8.17	8.52	8.98	9.70

Table 2: Statistics on our corpus Arabic dialect corpus(a sample of 1,000 sentences for each dialect and MSA)

	Raw Sentences				
	MSA	**TN**	**PA**	**JO**	**SY**
EG	39.12	37.30	43.42	42.33	44.66
SY	28.23	31.61	38.87	47.42	
JO	26.97	32.74	45.29		
PA	26.97	30.20			
TN	26.32				

Table 3: Sentence pair average similarities using the Overlap Coefficient

	Orthographically Normalized Sentences				
	MSA	**TN**	**PA**	**JO**	**SY**
EG	44.64	41.32	50.29	49.33	53.81
SY	34.47	35.12	47.56	54.05	
JO	33.30	34.26	49.81		
PA	33.40	34.22			
TN	31.41				

Table 4: Orthographically normalized sentence pair average similarities using the Overlap Coefficient

Table 2 provides various statistics for a sample of 1,000 sentences extracted from our multidialectal corpus.

To compare the similarity of the sentence pairs, we compute the Overlap Coefficient (OC), representing the percentage of lexical overlap between the vocabularies for each dialect pair D_1 and D_2. The OC is computed as follows:

$$OC = \frac{|D_1 \cap D_2|}{min(|D_1|, |D_2|)}$$

We conducted a preliminary lexical analysis restricted to simple matches. The results are given in Table 3. It is important to note the high lexical overlap of Egyptian Arabic with the rest of dialects studied. This could be explained by the fact that the translations were originally obtained from Egyptian. We first observe that the MSA and Egyptian closeness is particularly high. The fact that the MSA translator is Egyptian possibly introduced a bias in the translation process which explains this higher degree of similarity (39.12). The other dialects are all less similar to MSA than Egyptian. Their overlap degree with MSA ranges from 26 to 28. If we focus on different dialects (without MSA), we notice that Tunisian has the least overlap with all other dialects. This is not surprising since Tunisian is a Western dialect, whereas Levantine (PA, JO, SY) and Egyptian are all Eastern dialects. The highest degree of similarity across these dialects seems to be within the Levantine family (Syrian and Jordanian). When we compare Levantine against Egyptian, we observe that the highest degree of similarity is between Syrian and Egyptian. This could be explained by the fact that the Syrian translator is currently living in Egypt, which might introduce a bias in the translation process. In the future, we plan to consider carefully such biasing factors when creating translations.

In order to study the impact of an orthographic normalization on the similarity degree between dialects, we compute the overlap coefficient on pairs of sentences in which all the Hamzated Alif forms (آ *Ā*, أ *Â*, إ *Ǎ*) are replaced by a bare Alif ا *A*, the Alif-Maqsura ى *ý* by Ya ي *y* and the Ta-Marbuta ة *ħ* by Ha ه *h*. Similarity results are reported in Table 4. It is important to notice that the normalization does not change anything in the similarity ranking, which suggests that the kind of orthography errors made by the translators are naturally distributed across the different corpora.

Figure 1 presents some additional examples of sentences from this corpus.

Example 1	**EG**	واحد سافر لعند صديقه الى افريقيا لما وصل عزمو ينام عندو
	MSA	شخص سافر الى صديق له فى افريقيا و عندما وصل استضافه للمبيت عنده
	SY	واحد سافر لعند رفيقو إلى افريقيا لما وصل عزمو ينام عند
	JO	واحد سافر عند صاحبو بافريقيا ، لما وصل عزمو ينام عند
	PA	واحد سافر على صديق اله في افريقيا ويوم وصله عزموا ينام عنده
	TN	واحد سافر و مشا عند صاحبو في افريقيا كيف وصل استدعاه يبات عندو
	EN	*One left to his friend to Africa, when he got there he invited him to sleep over.*
Example 2	**EG**	المفروض ان اي اتنين بيحبوا بعض ميجرحوش بعض
	MSA	من المفترض ان اى شخصين فى علاقه حب لا يجرحوا بعضهما
	SY	المفروض أنو أي اتنين بيحبو بعض ما يجرحو بعض
	JO	المفروض الي بحبو بعض ما يجرحو بعض
	PA	لازم انو اي اثنين بحبو بعض ما يجرحوا بعض
	TN	المفروض اي زوز يحبو بعضهم ما يجرحوش بعض
	EN	*Its supposed to be that any two that loves each other not hurt each other*
Example 3	**EG**	دولابك مليان بفساتين سهرة مش محتاجاها ومش عارفة تعملى بيها ايه
	MSA	دولابك ممتلئ بفساتين سهرة لا تحتاجينها و لا تعرفى ماذا تفعلين بها
	SY	خزانتك مليانة فساتين سهرة و انتي مو بحاجتا ومو عرفانه تعملي فيها شي
	JO	خزانتك ملاينة فساتين سهرة شو رح تعملي فيه
	PA	خزانتك مليانه فساتين سهره انت مش بحاجتها ومش عارفه شو بدك تسوي فيها
	TN	خزانتك معبية بروب السهرات ماعادش حاشتك بيهم و ما تعرفش اش بش تعمل بيهم
	EN	*Your closet is full of dresses that you don't need or don't know what to do with*

Figure 1: Translations for three sentences in the Multidialectal Arabic Corpus

6. Conclusion and Future Work

We presented the first multidialectal Arabic parallel corpus, a collection of 2,000 sentences covering in addition to English and MSA, the Egyptian, Tunisian, Jordanian, Palestinian and Syrian Arabic dialects. The methodology we used relied on the familiarity of Egyptian Arabic to most other dialect speakers. A preliminary analysis of the data confirms known expectations about degrees of similarity between some of the dialects, but also points out to possible bias created by the choice of a specific dialect (in our case Egyptian) as the starting point.

We plan on extending the corpus in terms of size and number of dialects. We also plan to enrich it with additional annotations such as a CODA version, morphological tokenization, POS tagging and manual word alignments. Having all these rich annotations can be very helpful to supporting research in Arabic dialect NLP.

The corpus will be made freely available for research purposes.

Acknowledgements

The first and third authors were supported by grant NPRP-09-1140-1-177 from the Qatar National Research Fund (QNRF), a member of the Qatar Foundation. The second author was supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-12-C-0014. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of QNRF or DARPA.

7. References

Abu-Melhim, Abdel-Rahman. (1991). Code-switching and Linguistic Accommodation in Arabic. In *Perspectives on Arabic Linguistics III: Papers from the Third Annual Symposium on Arabic Linguistics*, volume 80, pages 231–250. John Benjamins Publishing.

Al-Sabbagh, Rania and Girju, Roxana. (2010). Mining the Web for the Induction of a Dialectical Arabic Lexicon. In *LREC*, Valetta, Malta.

Bassiouney, Reem. (2009). *Arabic Sociolinguistics*. Edinburgh University Press.

Benmamoun, Elabbas. (2012). Agreement and Cliticization in Arabic Varieties from Diachronic and Synchronic Perspectives. In Bassiouney, Reem, editor, *Al'Arabiyya: Journal of American Association of Teachers of Arabic*, volume 44-45, pages 137–150. Georgetown University Press.

Boujelbane, Rahma, Ellouze Khemekhem, Mariem, and Belguith, Lamia Hadrich. (2013). Mapping Rules for Building a Tunisian Dialect Lexicon and Generating Corpora. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 419–428, Nagoya, Japan.

Brustad, Kristen. (2000). *The Syntax of Spoken Arabic: A Comparative Study of Moroccan, Egyptian, Syrian, and Kuwaiti Dialects*. Georgetown University Press.

Chiang, David, Diab, Mona, Habash, Nizar, Rambow, Owen, and Shareef, Safiullah. (2006). Parsing Arabic Dialects. In *Proceedings of EACL*, Trento, Italy.

Darwish, Kareem. (2013). Arabizi Detection and Conversion to Arabic. *CoRR*.

Diab, Mona and Habash, Nizar. (2007). Arabic Dialect Processing Tutorial. In *NAACL*.

Elfardy, Heba and Diab, Mona. (2013). Sentence Level Dialect Identification in Arabic. In *Proceedings of the Association for Computational Linguistics*, pages 456–461, Sofia, Bulgaria.

Habash, Nizar, Soudi, Abdelhadi, and Buckwalter, Timothy. (2007). On Arabic transliteration. In Soudi, Abdelhadi, Neumann, Guenter, and van den Bosch, Antal, editors, *Arabic Computational Morphology*, volume 38 of *Text, Speech and Language Technology*, chapter 2, pages 15–22. Springer.

Habash, N., Rambow, O., Diab, M., and Kanjawi-Faraj, R. (2008). Guidelines for Annotation of Arabic Dialectness. In *Proceedings of the LREC Workshop on HLT & NLP within the Arabic world*.

Habash, Nizar, Diab, Mona, and Rambow, Owen. (2012a). Conventional Orthography for Dialectal Arabic. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 711–718, Istanbul, Turkey.

Habash, Nizar, Eskander, Ramy, and Hawwari, Abdelati. (2012b). A Morphological Analyzer for Egyptian Arabic. In *Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology*, pages 1–9, Montréal, Canada.

Habash, Nizar, Roth, Ryan, Rambow, Owen, Eskander, Ramy, and Tomeh, Nadi. (2013). Morphological analysis and disambiguation for dialectal arabic. In *Proceedings of NAACL-HLT*, pages 426–432, Atlanta, Georgia, June.

Habash, Nizar. (2006). On Arabic and its Dialects. *Multilingual Magazine*, 17(81).

Habash, Nizar Y. (2010). *Introduction to Arabic natural language processing*, volume 3. Morgan & Claypool Publishers.

Haeri, Niloofar. (1991). Sociolinguistic Variation in Cairene Arabic: Palatalization and the qaf in the Speech of Men and Women.

Pasha, Arfath, Al-Badrashiny, Mohamed, Kholy, Ahmed El, Eskander, Ramy, Diab, Mona, Habash, Nizar, Pooleery, Manoj, Rambow, Owen, and Roth, Ryan. (2014). Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In *In Proceedings of LREC*, Reykjavik, Iceland.

Riesa, Jason and Yarowsky, David. (2006). Minimally Supervised Morphological Segmentation with Applications to Machine Translation. In *Proceedings of AMTA*, Cambridge, MA.

Sajjad, Hassan, Darwish, Kareem, and Belinkov, Yonatan. (2013). Translating dialectal arabic to english. In *Proceedings of ACL*, Sofia, Bulgaria.

Salloum, Wael and Habash, Nizar. (2013). Dialectal arabic to english machine translation: Pivoting through modern standard arabic. In *Proceedings of NAACL-HLT*, Atlanta, Georgia.

Sawaf, Hassan. (2010). Arabic dialect handling in hybrid machine translation. In *Proceedings of AMTA*, Denver, Colorado.

Watson, Janet CE. (2007). *The Phonology and Morphology of Arabic*. Oxford University Press.

Zaidan, Omar F. and Callison-Burch, Chris. (2011). Crowdsourcing Translation: Professional Quality from Non-Professionals. In *Proceedings of ACL*, Portland, Oregon, USA.

Zaidan, Omar and Callison-Burch, Chris. (2013). Arabic dialect identification. *Computational Linguistics*.

Zbib, Rabih, Malchiodi, Erika, Devlin, Jacob, Stallard, David, Matsoukas, Spyros, Schwartz, Richard, Makhoul, John, Zaidan, Omar F., and Callison-Burch, Chris. (2012). Machine translation of arabic dialects. In *Proceedings of NAACL-HLT*, Montréal, Canada.

Zribi, I., Boujelbane, R., Masmoudi, A., Ellouze Khmekhem, M., Hadrich Belguith, L., and Habash, N. (2014). A Conventional Orthography for Tunisian Arabic. In *In Proceedings of LREC*, Reykjavik, Iceland.