日中韓辭典研究所
The CJK Dictionary Institute

SPEECH TECHNOLOGY FOR ARABIC DIALECTS

Introducing: *DiaLEX* and *DARS*

معجم اللهجات العربية

by Jack Halpern

1. Background

The CJK Dictionary Institute (CJKI) has launched a project to develop lexical resources for Arabic dialects, especially a *full-form lexicon* for Egyptian Arabic (EA_LEX) and Palestinian Arabic (PA_LEX), called ***DiaLEX.*** Our institute is in a unique position to assist in the development of speech technology and machine translation systems for Arabic dialects (dialectal Arabic or DA) by providing large-scale lexical resources and consulting services based on our in-depth knowhow of Arabic phonology and morphology.

2. Differences between Arabic Dialects and MSA

MSA is used as the official language of 22 Arab League nations and is widely used in education and in the media. It is taught in schools and spoken formally as the language of the cultured and the elite. The Qur'an is written in Classical Arabic (similar to MSA), and all religious works, prayers, lectures etc. are either in Classical or Modern Standard Arabic.

However, in everyday life, Arabs normally use one of the 30 or so modern dialects, such as Egyptian Arabic, Moroccan Arabic, Gulf Arabic, and Palestinian Arabic. The dialects are used between family and friends, and in society in general for practical communication. The mother tongue of every Arab is his/her dialect, whereas MSA is the mother tongue of no one. Though the dialects are increasingly used in social media, they do not have a formal written language nor a standard orthography. Books, magazines, newspapers and websites are overwhelmingly written in MSA.

Though most educated Arabs can understand spoken MSA, rarely do they speak it well. In fact, it is hard to find an Arab that can utter several MSA sentences in a row without making grammatical mistakes, unless they are highly educated or work in broadcast media, education, the government, or are religious leaders. Ordinary Arabs will usually omit case endings, often use numbers incorrectly, elide rather than fully pronounce final vowels, not use correct inflections and declensions, and more.

Basically, for Arabs MSA is like an *acquired* foreign language. Those who do speak it tend to do so slowly, unnaturally, and in a formal manner, rather than in a relaxed natural manner as when talking to friends, joking, or having a drink with friends.

3. Why Arabs tend to dislike spoken MSA

On the whole, the use of MSA is on the decline. According to Dr. Hussam Abu Zahr, "Greece translates five times as many books into Greek as all 22 Arab nations combined[1]. Against this background, there is much evidence to indicate that many if not most Arabs not only do not have a good command of spoken MSA, but also have a tendency to dislike it. Although Arabs have great respect for MSA as the language of prestige, culture and religion, it makes them feel unnatural to use it in conversation, which most of them cannot do well anyway. According to several Arabic teachers surveyed, even native speakers who speak MSA very well feel uncomfortable using it in daily life or even at work.

I myself am an intermediate speaker of MSA and until recently have never studied the dialects, so allow me to share an anecdote based on personal experience. I am having dinner with two educated Arabs, one a school teacher of MSA. I asked them to speak in MSA, the only variety of Arabic that I knew, and they agreed to cooperate. But something strange happened: whenever one of them speaks to me directly, he speaks in informal MSA (without case endings and final vowels). But as soon as they speak *to each other,* they immediately switch to dialect. I pointed this out a few times but they literally could not help themselves: they could not bring themselves to speak MSA to each other even though both can speak it well. Additionally, I have been questioned by people like taxi drivers and even by an Arabic school official: "Why do you insist on speaking Fusha (MSA)? It sounds so strange and unnatural. We don't speak that way!"

4. Full-Form Lexicons

A *full-form lexicon* contains all inflected, conjugated, declined, and cliticized forms. To give an English example, the full set of *wordforms* for the canonical form *eat* includes *eat, eating, eaten, ate*, while for *boy* it includes *boy, boys, boy's, boys'*. Arabic morphology is far more complicated. For example, adding the proclitics و and ل to and the enclitic اتهما to the stem كاتب *kātibun* 'writer' yields the complex form as ولكاتباتهما *walikātibātihimā* 'and to the two female writers'. Full-form lexicons can significantly contribute to the training of language models for natural language processing applications, especially for speech technology.

1 https://www.atlanticcouncil.org/blogs/menasource/standard-arabic-is-on-the-decline-here-s-what-s-worrying-about-that/

4.1 Full-Form Lexicon for Standard Arabic

For nearly a decade, CJKI has been in engaged in the development of an extremely comprehensive **Arabic Full-Form Lexicon** (ArabLEX) for Modern Standard Arabic (MSA) covering approximately 600 million entries including all inflected, conjugated and declined forms for both general vocabulary and proper nouns. Ideally suited for morphological analysis, machine translation and speech technology, it is currently being used to develop the world's most advanced speech technology and intelligent assistant systems for Arabic. For details see:

ArabLEX: Comprehensive Arabic Full Form Lexicon

□

4.2 *DiaLEX:* Full-Form Lexicons for Arabic Dialects

CJKI has launched a project to develop similar full-form lexicons for the major Arabic dialects and subdialects. In the initial stage, **EA_LEX** and **PA_LEX** (Full Form Lexicon for Egyptian and Palestinian Arabic) are being developed on the model of ArabLEX. PA_LEX covers the morphology and phonology of the urban variety of South Levantine Arabic (spoken in Palestine, most of urban Israel and western Jordan), commonly known as Palestinian Arabic (PA). This will be accompanied by a PA-to-English dictionary designed to support machine translation. In parallel we are launching projects to develop full-form lexicons for other dialects, especially EA_LEX for Egyptian Arabic and several dialects in the Gulf region, including Saudi Arabia and the UAE.

The data fields in PA_LEX and EA_LEX are similar to the ones in ArabLEX as shown in the data sample at https://www.cjk.org/wp-content/uploads/2020/11/ArabLEX.xls but fine tuned to those dialects.

4.3 *DARS:* Transcription System for Dialectal Arabic

We are also introducing ***DARS*** (Dialectal Arabic Romanization System), a transcription system fine tuned to speech technology and pedagogy of Arabic dialects in general, but as a first step focusing on Palestinian Arabic in particular. DARS represents PA phonemes unambiguously, explicitly indicating such features as allophonic variation, vowel neutralization, contextual vowel retraction ([a] --> [ɑ]), epenthesis, and word stress DARS pays special attention to epenthesis, the insertion of helping vowels between consonant clusters for ease of pronunciation. For example, قبل 'before' is rendered as *ʾábᵉl,* showing both word stress and epenthesis as well as the glottal stop realization of /q/.

□

5. Speech technology and MT for Dialectal Arabic

In light of the above, users of virtual assistants, like Alexa and Siri, would no doubt be much more satisfied to speak to their devices in their dialect (DA) rather than in MSA, but would not mind getting answers in MSA.

CJKI proposes that the development of speech technology for Arabic dialects be seriously considered, and that voice-activated smart platforms support the major dialects. The user would speak to the device in DA, which would be converted by ASR to a phonemic transcription and then translated by MT to MSA in the Arabic script. Next, the queries would be processed in MSA, and, most importantly, the answer would be generated by TTS output in MSA, ***not in DA.***

Translating MSA to PA or EA will not work well, and will often be impossible, since the dialects do not have the rich vocabulary set that MSA does, especially in such domains as science, technology, and the arts. Moreover, though native Arabs are not comfortable in *speaking* MSA, they have no problem with listening and understanding it.

6. Conclusion

In view of the central role that dialects play in Arab society, there is an urgent need to support speech technology for at least for the major dialects. With our extensive experience in developing lexical resources for Arabic NLP and our in-depth knowledge of Arabic phonology, CJKI can play a central role in achieving that goal, with the release of *DiaLEX* as the first step, followed by other dialects. A detaled document on *DiaLEX* will soon be released