Arabic Dialect Processing Tutorial

Mona Diab and **Nizar Habash**
Center for Computational Learning Systems
Columbia University
{mdiab,habash}@cs.columbia.edu

Language exists in a natural continuum, both historically and geographically. The term *language* as opposed to *dialect* is only an expression of power and dominance of one group/ideology over another. In the Arab world, politics (Arab nationalism) and religion (Islam) are what shape the perception of the distinction between *the* Arabic language and an Arabic dialect. This power relationship is similar to others that exist between languages and their dialects. However, the high degree of difference between standard Arabic and its dialects and the fact that standard Arabic is not any Arab's native language sets the Arabic linguistic situation apart.

As such, the Arabic language can be conceived of as a collection of multiple variants among which Modern Standard Arabic (MSA) has a special status as the formal written standard language of the media, culture and education across the Arab world. The other variants are informal spoken dialects that are the mediums of communication for daily life. Arabic dialects substantially differ from MSA and each other in terms of phonology, morphology, lexical choice and syntax.

MSA is the official language of the Arab world. It is the primary language of the media and culture. MSA is syntactically, morphologically and phonologically based on Classical Arabic, the language of the Qur'an (Islam's Holy Book). Lexically, however, it is much more modern. It is not a native language of any Arabs but is the language of education across the Arab world. MSA is primarily written not spoken.

The Arabic dialects, in contrast, are the true native language forms. They are generally restricted in use for informal daily communication. They are not taught in schools or even standardized although there is a rich popular dialect culture of folktales, songs, movies, and TV shows. Dialects are primarily spoken not written. However this is quite changing since more Arabs are gaining access to electronic media of communication such as emails and newsgroups. Arabic dialects are loosely related to Classical Arabic. They are the result of the interaction between different ancient dialects of Classical Arabic and the indigenous languages that existed in today's Arab world together with influences from colonization and interaction with neighboring countries. For example, Algerian Arabic has a lot of influences from it's ancient indigenous language Berber as well as French due to the French occupation.

Arabic dialects vary on many dimensions – primarily, geography and social class. Geolinguistically, the Arab world can be divided in many different ways. The following is only one of many: **Levantine Arabic** includes the dialects of Lebanon, Syria, Jordan, Palestine and Israel. **Gulf Arabic** includes the dialects of Kuwait, Saudi Arabia, United Arab Emirates, Bahrain, and Qatar. Iraqi and Omani Arabic are included some times. **Egyptian Arabic** covers the dialects of the Nile valley: Egypt and Sudan. **North African Arabic** covers the dialects of Morocco, Algeria, Tunisia and Mauritania. Libya is sometimes included. **Yemenite Arabic** is often considered its own class. **Maltese Arabic** is not always considered an Arabic dialect. It is the only Arabic variant that is considered a separate language and is written with Latin script.

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL pages 5–6,
Rochester, April 2007. ©2007 Association for Computational Linguistics

Socially, it is common to distinguish three sub-dialects within each dialect region: city dwellers, peasants/farmers and Bedouins. The three degrees are often associated with a class hierarchy in which rich settled city dwellers are on top and Bedouins are on bottom. Different social associations exist as common in many other languages around the world. For example, the city dialect is considered less marked, better and smarter; whereas the Bedouin dialect is considered lower class, rough, yet pure to the origin of the language.

The relationship between MSA and the dialect in a *specific region* is rather complex. Arabs do not think of these two as separate languages. This particular perception leads to a special kind of coexistence between two forms of language that serve different purposes. This kind of situation is what linguists term *diglossia*. Although the two variants have clear domains of prevalence: formal written (MSA) versus informal spoken (dialect), there is a large gray area in between and it is often filled with mixing of the two forms.

For Natural Language Processing (NLP), the existence of dialects for any language constitutes a challenge in general since it adds another set of variation dimensions from a known standard. The problem is particularly interesting and challenging in Arabic and its different dialects, where the diversion from the standard could, in some linguistic theories, warrant a classification as a different language. This problem would not be as pronounced if standard Arabic were to be a living language, however it is not. Any realistic and practical approach to processing Arabic will have to account for dialectal usage since it is so pervasive.

In this tutorial, we highlight different dialectal phenomena. We discuss how dialects migrate from the standard and why they pose challenges to NLP. Our tutorial will have four different parts: First, we describe a background layout of issues for standard Arabic NLP. Then we discuss a high level generic view of dialects and their different aspects that are of interest for the NLP community. We address both text and speech issues in addition to standardization issues. We focus in depth on two aspects of dialect processing in the third and fourth parts of the tutorial, namely, dialectal morphology and dialectal syntactic parsing. Throughout the presentation we will make references to the different resources available and draw contrastive links with standard Arabic and English. Moreover, we will discuss annotation standards as exemplified in the Linguistic Data Consortium Arabic Treebank. We will provide links to recent publications and available toolkits/resources for all four sections.

This tutorial is designed for computer scientists and linguists alike. No knowledge of Arabic is required. However, we recommend taking a look at Nizar Habash's Arabic NLP tutorial[1] which will be reviewed in the first quarter of the tutorial.

[1] http://www.ccls.columbia.edu/cadim/presentations.html