Corpus of Tunisian Arabic: Design and Progress

Karen McNeil
Georgetown University
klh58@georgetown.edu

Miled Faiza
University of Virginia
mbf2@virginia.edu

After initially lagging behind other world languages, Arabic has recently made some great strides in the field of corpus linguistics. Yet in spite of the proliferation of many new Arabic corpora, most of them are limited in that they're excessively focused on formal sources. Many are especially over-dependent on media: several are composed entirely of collections of newspaper or newswire sources. Even those not focused exclusively on news sources are restricted to formal topics. Although these are important resources for certain kinds of linguistic analysis, they are obviously not balanced representations of the language as a whole.

This limitation is not a design flaw on the part of the corpora authors, but rather is a result of the diglossic nature of the Arabic language; the above-mentioned corpora are all for formal, written Arabic ("Modern Standard Arabic" or MSA). Although the compilers of some of these corpora, such as the International Corpus of Arabic, explicitly state that one of their goals was "collecting the texts to convey all regional variations of the Arabic Language," the "Arabic Language" that they are referring to is only Modern Standard Arabic which, in fact, varies very little between Arabic countries. An example of the variations studied using corpora like this is in the variation of the phrase "at the same time" between different areas of the Arab world. The two variants studied were *fī nafs al-waqt* and *fī al-waqt nafsihi* (Al-Sulaiti and Atwell 2006, 6). Most of the variation observable in MSA is similarly trivial.

To illustrate the difference between the spoken varieties and MSA, however, consider a 2005 study by Katrin Kirchhoff and Dimitra Vergyri (p. 41). Kirchhoff and Vergyri were working on developing speech recognition tools for Egyptian Arabic, and were testing whether adding a large MSA corpus to the small Egyptian corpus would improve the error rate of their program. They discovered that the vocabulary (unigram) overlap between the two corpora was only 10.3% (compared with 44.5% for a similar pair of American formal and British informal corpora) and found that, in the end, extending the Egyptian corpus with MSA did not improve their program's error rate.

In addition to the large difference between the written and spoken varieties, the differences *among* the spoken varieties are also significant. Although often segmented by national boundaries (Egyptian, Iraqi, Lebanese, etc.), the varieties of spoken Arabic actually encompass a graduated dialect continuum extending from Morocco to Iraq, with the dialects on either extreme being mutually incomprehensible.

Because of the heterogeneity of the language, there can never be "one" authoritative corpus of Arabic. To be of any practical use for either language-learning resources or natural language processing, MSA corpora need to be distinct from the spoken, and the spoken corpora will need to be segmented into at least the major dialectal groups. Some efforts in this direction have been undertaken. However, many of the available spoken corpora are quite short, and most of them are based on telephone conversations, resulting in a limited range of vocabulary. In addition, Egyptian and Levantine are represented in several corpora, whereas no North African dialect is represented at all.

To augment this list and address some of these shortcomings, we are currently working to create a corpus of the spoken Arabic of Tunisia. There are several challenges to developing a corpus of an Arabic dialect, the most significant of which is sources. Corpora are written resources, and the spoken varieties of Arabic are, by definition, not written. It is therefore very difficult to find written sources to include in the corpus. In the development of the Tunisian Arabic Corpus, we have utilized three categories of sources: 1) traditional written sources; 2) new written sources; and 3) transcription of audio sources.

Traditional written sources are the few exceptional types of writing in which dialect has always been used. These include folklore, folk poetry, popular songs, proverb collections, and screenplays for movies, television shows, and plays. New written sources are those that were not traditionally written in dialect (or did not exist), but in which the dialect has become a viable language choice, such as blogs and forum comments. The third source strategy that I am utilizing for the TAC is transcription. As time consuming and expensive as it is, there's just no avoiding the fact that many of the sources needed to make a balanced corpus of spoken Arabic will not be available in written form.

Another major problem in creating a corpus of dialectical Arabic is that of balance. It is difficult to imagine how a "balanced" corpus, in the conventional sense, can be possible. The spoken varieties of Arabic are, to a great extent, in complementary distribution with MSA. Just as media sources are vastly

overrepresented in all the MSA corpora, they will be vastly underrepresented in any variety of spoken Arabic.

To organize the corpus materials and workflow, we have created a web application using the programming language Python and the web framework Django. Although building this application required a significant amount of time, it is well worth it in that it allows us to manage and organize the corpus files and metadata, and perform basic linguistic processing (such as frequency lists, collocations, and concordancing). In addition, the application acts as a central portal and workflow manager for the people working on the project (including transcribers working remotely), allowing them to download yet-to-be-completed files, and add the completed transcripts to the corpus. Now that the tool is complete, we will be able to add large numbers of text through automated methods (for instance, using web crawlers to harvest internet forum comments), a task for which well-organized file and workflow management will be essential.

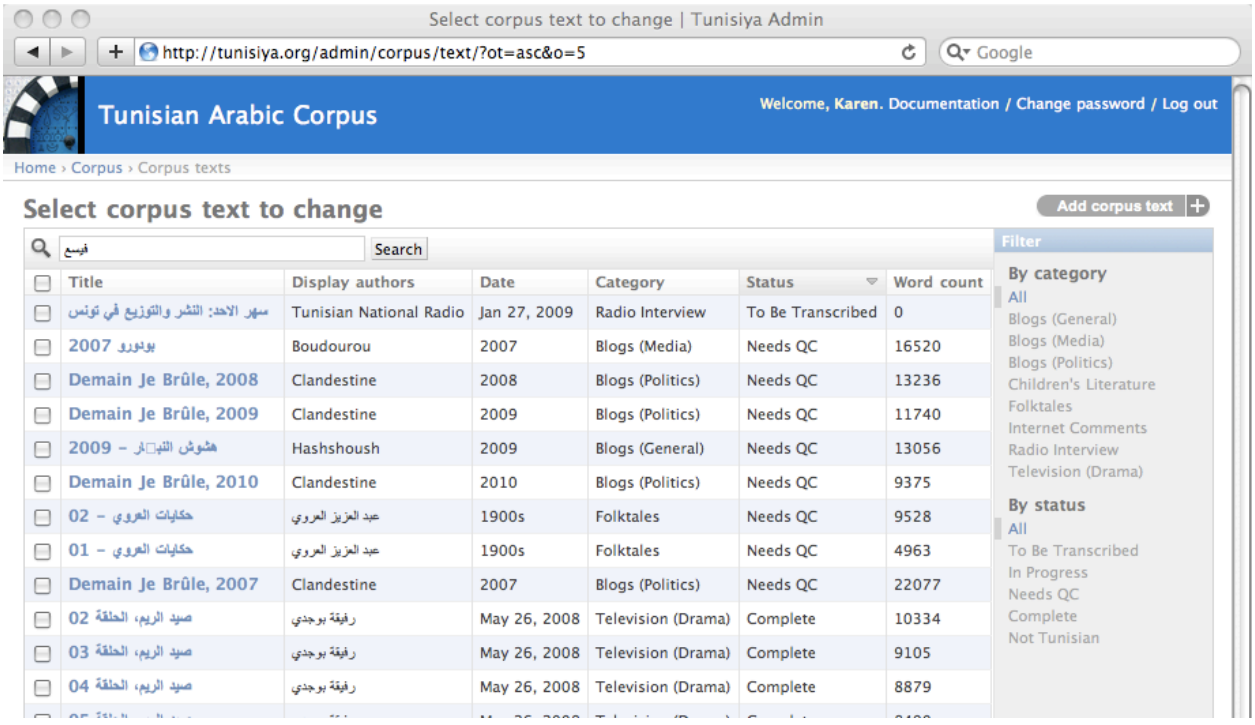

Figure 1: Tunisian Arabic Corpus Management Tool

Bibliography

Alansary, Sameh, Magdy Nagi, and Noha Adly. "Building an International Corpus of Arabic (ICA): Progress of Compilation Stage." Bibliotheca Alexandrina. 2007.

http://www.bibalex.org/isis/UploadedFiles/Publications/Building%20an%20Intl%20corpus%20of%20arabic.pdf.
—. "Towards analyzing the International Corpus of Arabic: Progress of Morphological Stage." Bibliotheca Alexandrina. 2008.
http://www.bibalex.org/isis/UploadedFiles/Publications/Morphological_analysis_o f_ICA_finalx.pdf.
Al-Sulaiti, Latifa, and Eric Atwell. "The design of a corpus of Contemporary Arabic." International Journal of Corpus Linguistics 11, no. 1 (2006): 1-36.
Kirchhoff, Katrin, and Dimitra Vergyri. "Cross-dialectal data sharing for acoustic modeling in Arabic speech recognition." Speech Communication 64, no. 1 (2005): 37-51.