Ahmad Yacout, Hoon Yoo

Computational Tools for Cultural Comparisons

Professor Dennis Shasha

11 January 2015

Project 1

Mapping the Linguistic Differences of Arabic Dialects

Arabic dialects do not differ just in the way that sounds are pronounced, but they also vary immensely in grammar structure and word formation. This project aims to capture these difference, map them in a way that will allow for the analysis of parameter like structures similar to those of Noam Chomsky and finally will be presented in a format that allows for a cross cultural comparison of these languages based on temporal and special factors at the root of each dialect.

The data for such research is not widely available due to the lack of the formalization of these dialects therefore, most of the data will be gathered from independent websites, sadly from Wikipedia, interviews conducted with various speakers of the dialects, and papers and books which have published independent research, both in Arabic and English about the structures of the various dialects. Definitely consult with Nizar.

Because of time constraints, the only qualitative factors that are going to be mapped are letter pronunciation, pronoun formation, colonizers, cultures in which it started, formation and uses of the definitive article, the formality of it (creole, official or pidgin), plural formation, possessive suffix formation, whether it exists above or below the equatorial line, and question structures

 Additionally there is a reddit page dedicated to deciphering dialects that we will use to identify our main parameters. This is where the python element becomes a part of the project.

 As background to what reddit is and what the dialects page accomplishes, a brief summary would be the following. Reddit is a large-scale internet community which is subdivided into many smaller communities by interest. One of these interest groups was created for the Arabic dialect. It is this sub-group that we want to use for our project.

 Within this group there is an ongoing project in which one excerpt (from a children’s story) is read by anyone from within the group in their own native dialect. After reading it they then transcribe it phonetically in order to show both the differences in pronunciations and the syntaxical structure. Using python, and the help of Juan Felipe, we will systematically pair the transcribed paragraphs to the native language in order to begin identifying and extracting our parameters.

I love it.