Rafat Abushaban April 12, 2016

Palestinian tech bridges Arabic language online gaps

Read In

Arabic

The development of the Arabic dialect search engine was a team effort. (Images via Curras).

Every day millions of Arab speakers use social media platforms to discuss, debate, and express their opinions, but different dialects can sometimes make communicating the finer points challenging.

Within the sea of Arabic content that's generated daily, Birzeit University associate professor in linguistics Mustafa Jarrar recognized that a vast majority of it was in the Common Arabic dialect, (which is the common dialect tailored to each region in the Arab World).

As a result, Jarrar his team of researchers launched Curras ('notebook' in Arabic), a search engine tool to process, recognize, classify and translate Common Arabic Dialect, with a focus on the Palestinian dialect, into Modern Standard Arabic (MSA) and English.

A sea of dialects

Modern Standard Arabic is the formal form of the language, but every country and even regions within countries have their own dialects. There are the widely understood eastern-

MENA dialects that include those from the Levant, Egypt, and Gulf. Western-MENA dialects from Algeria, Morocco and Tunisia are not as well understood in eastern areas.

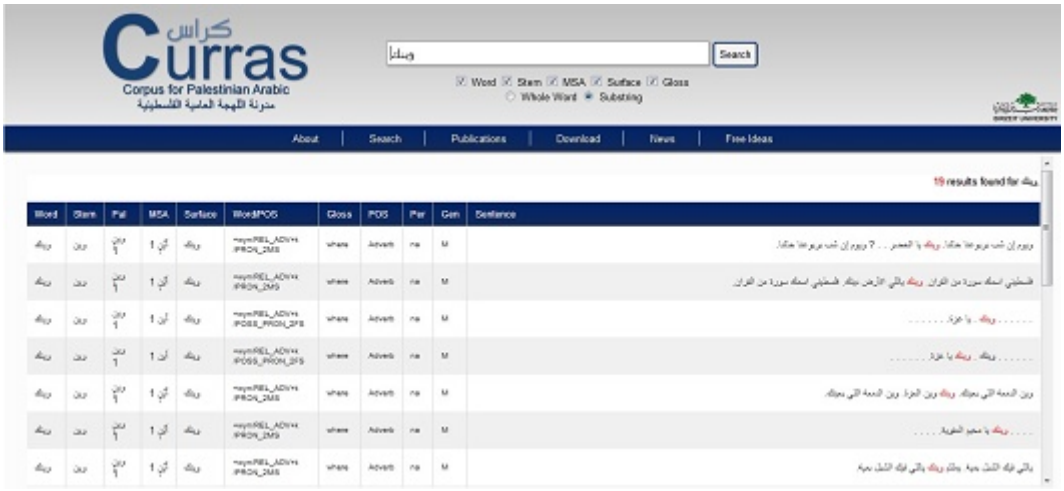
Each dialect tended to be for spoken use only and wasn't used for writing, until the internet and social media provided a launch pad for informal, written communication. As a result the written use of Arabic dialects has boomed online, and with it the need for some intra-Arabic translation services, as words as simple as 'now' differ between dialects.

Curras's Launch

Dialect is used not only on social media, but is also available all around the internet in text, audio and video; the Curras team wanted to contribute to closing the gaps in the digital Arabic content sector by categorizing and classifying this content that was otherwise unknown to computer software, such as Google Translate.

The project was originally meant to help researchers classify different words and grammatical structures, as well as helping developers make sense of online Arabic content. But once it went live the team noted that many Arabic language students started using Curras as a dictionary to link dialect vocabulary with MSA.

Jarrar said the project had created a knowledge base which could have further applications for problems such as sentiment analysis, a system that analyzes the text found in comments and forums to determine the user's opinion on a given issue.



The current format of the search engine.

The making of an online dialect search engine/dictionary

Curras has a simple Google-like interface where a user enters a dialect word, which is then broken down into elements such as its stem, prefixes, suffixes and gender forms, translated into MSA and English, and finally displays dictionary-like metadata to classify the words.

The main component in the system is the Corpus, a Database which contains the data of all collected words with their annotations and properties. When a user searches for a certain word the system determines if it is a verb or noun, prefix or suffix, masculine or feminine, among other categories.

“We had to manually enter 16 properties for over 55,000 words, half of them from a local popular show (Watan Ala Watar) that includes a variety of Palestinian local dialects, put them in the Corpus and develop the search engine,” Jarrar said.

The system was built gradually by Jarrar and four research assistants over two years, working with regional and international teams from Columbia University and New York University in Abu Dhabi, and the work itself was funded by the Scientific Research Council of the Palestinian Ministry of Higher Education.

Curras and the Arabic Language Digitization

Interestingly, the team conducted analyses to cross compare different Arabic dialects and found a 75 percent rate of similarity between the phonetics in Palestinian and Egyptian dialects, and even higher similarities between dialects from Levantine countries Jordan, Lebanon, and Syria are expected.

Curras isn’t Jarrar’s only venture into the digitization of the Arabic language. He recently won a Google Research Award of $50,000 in recognition for his work on natural language processing and language digitization, and has worked on other linguistic databases for humans and computers alike. His main and long-term project is the development of the Arabic Ontology, which is expected to launch in a few months.

The Curras team is working on a prototype that anyone can use, and hopes to include all Levant dialects, to integrate audio input into the system, and to release the Corpus for public use.

arabic search engine
Curras
Birzeit University
Arabic dialect
Arabic language
palestine

Rafat Abushaban