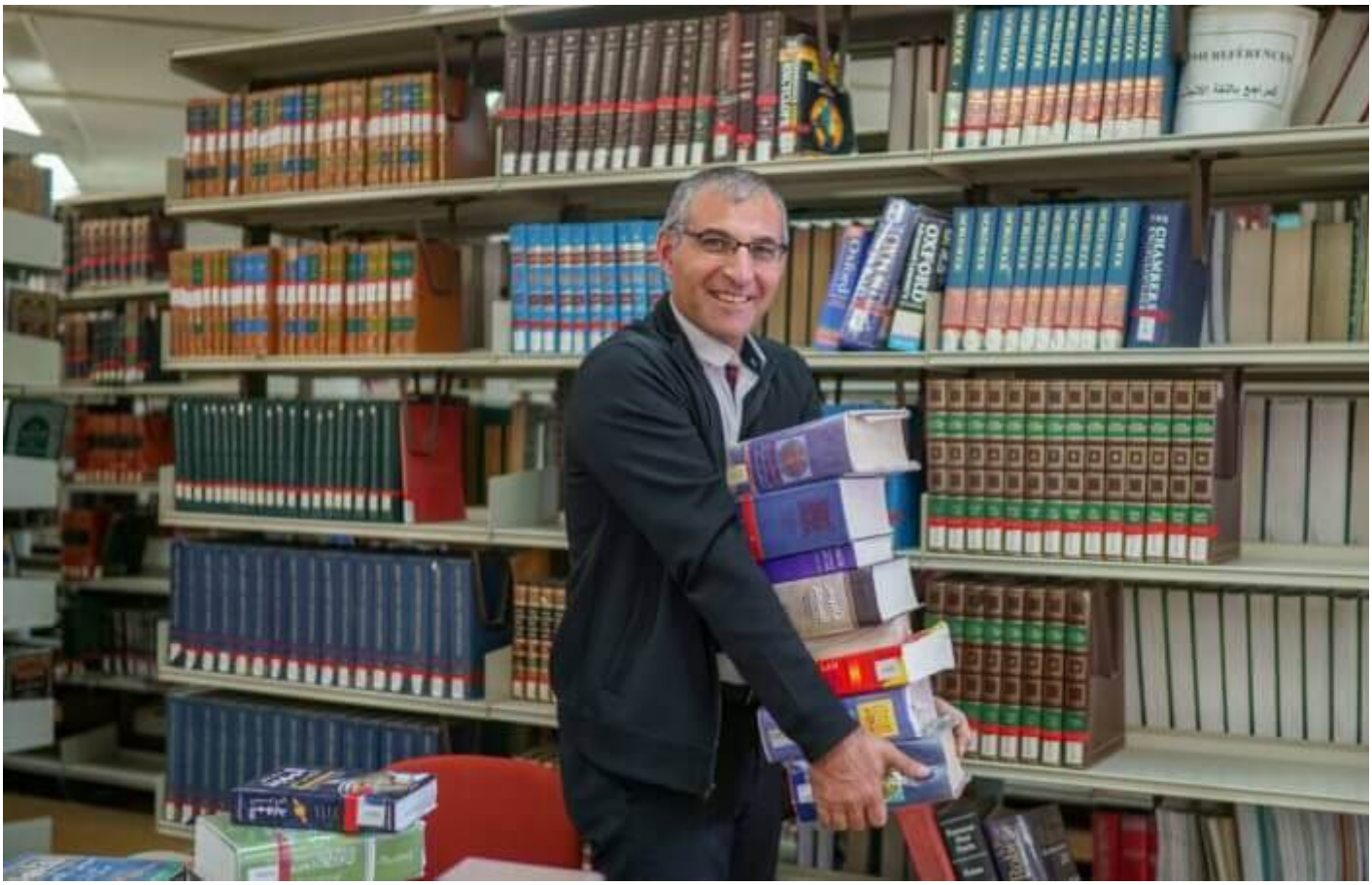
LANGUAGE , LITERATURE & TRANSLATION , RESEARCH

An Online Arabic Dictionary Makes Its Debut

Edward Fox / 28 Nov 2018



Mustafa Jarrar with some of the 150 Arabic dictionaries that were scanned into the Arabic Ontology online database (Photo: Eyad Jadallah/ Birzeit University).

A computer science professor at Birzeit University has built an online tool called Arabic Ontology that is both a comprehensive dictionary of Arabic and a system that will enable the creation of new Arabic-language software, such as better machine translation.

Mustafa Jarrar, an associate professor in the department of computer science at Birzeit, in the West Bank, has worked for eight years to create the new tool. It functions not only as a searchable lexicon of Arabic that can be used as both a dictionary and a thesaurus, but also as a logical system—an ontology—that recognizes the unique characteristics of the Arabic language. It can find relationships between the meanings of Arabic words in a natively Arabic way for the first time.

The Arabic Ontology offers the prospect of more precise results from Google searches in Arabic, better online translation of Arabic text, and new insights into the language for students and scholars of Arabic literature.

The new tool, which is freely available for personal use at http://ontology.birzeit.edu/, was made public in a ceremony at Birzeit on September 25. The copyright is owned by Birzeit University.

"This is the first search engine of its type for a single language," Jarrar said. That is, the search engine offers results from 150 Arabic dictionaries. "Imagine you had the Oxford English Dictionary, the Merriam-Webster dictionary and all the others collected in one place, all integrated and unified in one database," he said.

The word "ontology" in the name of the project refers to the concept in the science of linguistics, meaning a way of classifying the meanings of and relationships between words in a language. It is originally a term in philosophy meaning the study of existence, a way of organizing things that have being.

"We took Arabic words and classified their meanings in the form of a tree," Jarrar explained. "Arabic Ontology is a tree of the meanings of Arabic words." Larger abstract concepts belong to the trunk of the tree, while the branches represent specific things and leaves represent a thing's characteristics.

Jarrar described the ontology as an Arabic-language successor to a project called WordNet, started in 1985 at Princeton University in the United States. WordNet is a searchable lexicon of English words, and functions as a dictionary and as a thesaurus.

The difference between Arabic Ontology and WordNet, Jarrar said, is that while in WordNet a word's meaning is defined according to a consensus of its usage by native speakers, the meanings of Arabic words in Jarrar's system are fixed according to their usage by specialists. "When I was building it, I consulted natural scientists, physicists, chemists, biologists and so on, because they determine the meanings of words in their disciplines," he said.

Christiane Fellbaum, a professor of computer science at Princeton University and the current director of WordNet, said that Jarrar's project is a step forward from the original idea of WordNet. "How he differs from WordNet is he wants to build an ontology," Fellbaum said. "This is an idea that goes back to the philosophy of Aristotle, who [in works such as *The Categories* and *The Metaphysics*] organized knowledge in terms of general and narrow categories, and also as part/whole relations. Mustafa's innovation is the ontology, which creates an understanding of the relationships between the meanings of words."

"He is an unusual and courageous person, and he deserves enormous credit," Fellbaum said.

Building the Database

To create the combined Arabic lexicon, the contents of 150 Arabic dictionaries had to be manually entered into a database. This was painstaking work. At first, Jarrar tried to capture information from books by using a scanner. But to extract the useful data from the books required optical character recognition software, known as OCR, that could read Arabic.

"I tried to use OCR, but it didn't work," Jarrar said. Arabic OCR software is still so poor, he said, that "the amount of corrections you have to do is more work than if you entered the text from scratch, manually."

Instead, Jarrar said, he crowdsourced the task to Birzeit University students.

Birzeit requires students to perform 120 hours of community service before they can graduate. The program's goal is to better connect the university with Palestinian society outside the campus. Typically, Birzeit students doing community service will pick olives on local farms, or help old people in their homes.

The Birzeit administration considered that working on Jarrar's project was a suitable activity for community service, because of its value to Arab culture and society as a whole, Jarrar said, so students could fulfil their community service duty by typing the contents of pages of Arabic dictionaries into his database. To improve accuracy, he gave the same page to more than one student to transcribe.

Eventually, he selected students who could do the work to a high standard, without mistakes. He engaged these students as paid workers. The work took eight years to complete.

Fadi Zaraket, an associate professor of computer engineering at the American University of Beirut who specializes in natural language processing in Arabic, emphasised the value of Jarrar's project in the way it follows the distinctive characteristics of the Arabic language, and the natural relationships between Arabic words.

Two Arabic words derived from the same three root consonants are likely to be related in meaning, directly or indirectly. The Arabic Ontology can identify these relationships in a way that WordNet could not, Zaraket explained.

"It will help us to discover new semantic relationships between Arabic words," he said.

The system also recognizes the morphology of Arabic words; that is, the conventional ways they are written. "In Arabic, people can read words even if the short vowels are not written, but automated tools that don't use a knowledge-based analysis, as the Arabic Ontology does, have a tough time understanding Arabic words without vowels," he said.

Incorporating Arabic Ontology into Google Translate, for example, would improve the quality of translations into and from Arabic, Zaraket said. Google Translate uses a statistical approach: It produces a translation by analyzing vast quantities of data, but sometimes the results lack accuracy. "Mustafa's thing," Zaraket said, "organizes the concepts of Arabic. If you used that in addition to the statistical techniques, you would boost accuracy. It is much needed."

Zaraket sees Jarrar's project as a contribution to the study of the Arabic language itself, as well as a useful research tool. "The Arabic Ontology is a documentation of the philosophical capacity of what the Arabic language can express," he said.