BIRZEIT UNIVERSITY

Birzeit University releases corpora for six Arabic dialects

(/#facebook) (/#twitter)

(https://www.addtoany.com/share#url=https%3A%2F%2Fwww.birzeit.edu%2Fen%2Fnews%2Fbirzeit-university-releases-corpora-six-arabic-dialects&title=Birzeit%20University%20releases%20corpora%20for%20six%20Arabic%20dialects)

17 Dec 2022

Birzeit University released corpora for Libyan, Palestinian, Lebanese, Iraqi, Sudanese and Yemeni dialects that have 1.3 million words. Titled Currasat, the corpora aim to enrich artificial intelligence technologies and enable them to understand texts written in dialectal Arabic. The university has worked on some dialects in partnership with the American University of Beirut and the United Nations. Currasat was launched on December 15, 2022 at the United Nations Headquarter in New York.

The corpora consist of a collection of dialectal texts collected from social media platforms, such as Facebook, Twitter and YouTube. Each token in the corpora was segmented into prefixes, suffixes, stems, parts of speech, lemmas and English glosses.

The corpora can be used as a trilingual lexicon (Dialectal Arabic-Standard Arabic-English), especially by foreigners and researchers. It can also be used to construct computational applications capable of understanding written content on social media platforms, so that computers can understand texts written in dialectal Arabic and automatically convert them into Standard Arabic.

The Curras Palestinian corpus was previously launched in 2013 with the support of the Ministry of Higher Education. Later, it was revised and combined with Baladi, a Lebanese corpus that consists of 10k words. Both Curras and Baladi represent Levantine dialects.

The four-dialect corpus (Libyan, Sudanese, Iraqi and Yemeni) was constructed based on the methodology used to construct the Palestinian corpus. The Yemeni corpus was collected from Twitter; it includes 1.2 Million words. The Libyan, Sudanese and Iraqi corpora were collected from Facebook and YouTube; each includes 50k words. The corpora are the result of a collaboration project between Birzeit University, the American University of Beirut and the United Nations.

Researchers can use and download the corpora via the following link:

http://portal.sina.birzeit.edu/curras (http://portal.sina.birzeit.edu/curras)

Recent ★ Popular (/en/news/birzeit-university-releases-corpora-six-arabic-dialects?qt-news_tal

Birzeit University celebrates scientists with top 2% citation impact (/en/news/birzeit-university-celebrates-scientists-top-2-citation-impact)
19 Dec 2022


(/en/news/birzeit-university-celebrates-scientists-top-2-citation-impact)

Birzeit University Alumni win Gender and Intersectional Justice Award (/en/news/birzeit-university-alumni-win-gender-and-intersectional-justice-award)
19 Dec 2022


(/en/news/birzeit-university-alumni-win-gender-and-intersectional-justice-award)

Birzeit University releases corpora for six Arabic dialects (/en/news/birzeit-university-releases-corpora-six-arabic-dialects)
17 Dec 2022

(/en/news/birzeit-university-releases-corpora-six-arabic-dialects)

Meta launches My Digital World to secure online safety of students in Palestine (/en/news/meta-launches-my-digital-world-secure-online-safety-students-palestine)
15 Dec 2022


(/en/news/meta-launches-my-digital-world-secure-online-safety-students-palestine)

Birzeit University Museum hosts book launch for Returning by Vera Tamari (/en/news/birzeit-university-museum-hosts-book-launch-returning-vera-tamari)
15 Dec 2022


(/en/news/birzeit-university-museum-hosts-book-launch-returning-vera-tamari)

BIRZEIT UNIVERSITY
PO Box 14, Birzeit
West Bank, Palestine
Contact details