日中韓辭典研究所
The CJK Dictionary Institute

DiaLEX
ARABIC DIALECTS FULL-FORM LEXICON

with 878 million entries

Overview

While Modern Standard Arabic is used as the official language of 22 Arab nations, Arabs normally use one of the 30 or so modern dialects for communicating with family and friends in their daily life. ***DiaLEX*** is the most comprehensive Arabic computational lexicon ever created for Arabic dialects. Designed for NLP applications like MT, NER and morphological analysis, it is ideally suited for training speech technology models.

Dialect	Lemmata	Entries
Egyptian	33,000	280 million
Hejazi	31,000	112 million
Emirati	30,000	166 million
Syrian	25,000	77 million
Lebanese	20,000	109 million
Palestinian	27,000	134 million

Coverage

DiaLEX covers the following major Arabic dialects: Egyptian, Emirati, Saudi Arabian Hejazi, Syrian, Lebanese, and Palestinian. It is rich in morphological, grammatical, phonological, and orthographic attributes, and currently covers approximately 878 million for six dialects. In addition, it maps all unvocalized forms to their vocalized forms and the lemma, and provides phonemic transcriptions and graphemic transliterations.

Speech Technology

DiaLEX is especially valuable for speech technology, since Arabic dialects are the natural medium of everyday spoken interaction yet remain significantly under-resourced compared with MSA. Their non-standard orthography, abundant spelling variation, and large number of cliticized forms create major challenges for both TTS and ASR. Built on the same principles as ***ArabLEX***, our full-form lexicon for MSA, ***DiaLEX*** addresses these challenges with comprehensive full-form coverage, extensive orthographic variants, and highly accurate phonemic transcriptions, making it an ideal resource for training and improving dialect speech systems.

Distinctive Features

- Extremely comprehensive full form entries
- Rich in morphological attributes: all inflected, cliticized and negated forms
- Numerous orthographic variants
- Includes high frequency proper nouns (personal names and place names)
- Fully vocalized and unvocalized Arabic
- Accurate phonemic and phonetic transcriptions
- All wordforms are cross-referenced to their lemma

Grammatical and phonetic attributes (8 out of 25 shown)

ARABIC	PHONEMIC	PHONETIC	LEMMA	POS	GEN	NUM	NPG
بَيْتْ	bēt	be:t	بَيْتْ	N	M	S	000
اِلْبَيْتْ	ilbēt	ɪlbe:t	بَيْتْ	N	M	S	000
بَيْتِي	bēṭi	be:ti	بَيْتْ	N	M	S	S1C
بَيْتَكْ	bētak	be:tak	بَيْتْ	N	M	S	S2M
بَيْتِكْ	bētek	be:tek	بَيْتْ	N	M	S	S2F
بَيْتُو	bēto	be:tʊ	بَيْتْ	N	M	S	S3M
بَيْتْهَا	bēt˘hạ	be:tha	بَيْتْ	N	M	S	S3F
بَيْتْنَا	bētnạ	be:tna	بَيْتْ	N	M	S	P1C
بَيْتْكُمْ	bētkom	be:tkʊm	بَيْتْ	N	M	S	P2C
بَيْتْهُمْ	bēt˘hom	be:thʊm	بَيْتْ	N	M	S	P3C

The CJK Dictionary Institute

The CJK Dictionary Institute (CJKI), founded in 1993 and based in Saitama, Japan, compiles large-scale dictionary databases of proper nouns and technical terms for CJK and Arabic, currently totaling over 50 million entries. It is a leading provider of lexical resources, educational tools, and consulting services for the IT industry.

Jack Halpern (春遍雀來), CEO of CJKI, is a lexicographer by profession, specializing in Japanese and Chinese. His work as an editor in chief of learner's dictionaries resulted in various renowned standard reference works. An avid polyglot who speaks 12 languages, he has lived in five countries and has been a permanent resident of Japan for decades.