DiaLEX: Arabic Dialects Full-Form Lexicon

Jack Halpern, Director • The CJK Dictionary Institute • Niiza, Japan • jack@cjki.org

What is a full-form lexicon

- comprehensive coverage of all wordforms
- includes all inflected, conjugated, declined and cliticized forms
- ordinary dictionaries contain canonical forms like *eat*
- full form dictionaries include inflections like *ate*, *eating*, *eaten*, and plurals

What is DiaLEX

- a large-scale full form lexicon for Arabic dialects with approx. 878 million entries
- covers Egyptian, Emirati, Saudi Arabian Hejazi, Syrian, Lebanese, and Palestinian
- comprehensive coverage for inflected, conjugated and cliticized forms
- rich set of attributes for NLP

Distinctive Features

- extremely comprehensive full-form entries
- rich in morphological attributes: covers all inflected, cliticized and negated forms
- numerous orthographic variants
- includes high frequency proper nouns (personal names and place names)
- fully vocalized and unvocalized Arabic
- carefully curated phonemic phonetic transcriptions and transliteration
- all wordforms are cross-referenced to their lemma

Palestinian Arabic Variants

SUB_ID	VAR_ID	VAR_V	DARS	LEMMA	TENSE	NPG
01	01	بِكْتِبْ	bíkteb	كَتَبْ	bi-imperfect	S3M
01	02	بْيِكْتِبْ	bʸíkteb	كَتَبْ	bi-imperfect	S3M
01	03	بُكْتُبْ	búktob	كَتَبْ	bi-imperfect	S3M
01	04	بْيُكْتُبْ	bʸúktob	كَتَبْ	bi-imperfect	S3M
02	01	بْتِكْتِبْ	btíkteb	كَتَبْ	bi-imperfect	S3F
02	02	بْتُكْتُبْ	btúktob	كَتَبْ	bi-imperfect	S3F
03	01	بْنِكْتِبْ	bníkteb	كَتَبْ	bi-imperfect	P1C
03	02	مْنِكْتِبْ	mníkteb	كَتَبْ	bi-imperfect	P1C
03	03	بْنُكْتُبْ	bnúktob	كَتَبْ	bi-imperfect	P1C
03	04	مْنُكْتُبْ	mnúktob	كَتَبْ	bi-imperfect	P1C

Coverage

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

Grammatical and Phonetic Attributes

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
بَيْتُهَا	bēt˘ha	be:tha	بَيْتْ	N	M	S	S3F
بَيْتْنَا	bētna	be:tna	بَيْتْ	N	M	S	P1C
بَيْتْكُمْ	bētkom	be:tkʊm	بَيْتْ	N	M	S	P2C
بَيْتُهُمْ	bēt˘hom	be:thʊm	بَيْتْ	N	M	S	P3C

Practical Applications

- **Speech Technology**
  - training ASR and TTS models
- **Machine Translation**
  - enhanced MT quality due to full inflections
- **Morphological Analysis**
  - simplifies algorithms so decliticization unnecessary
- **Pedagocial Applications**
  - full verb conjugation paradigms
- **LLM Model Training**
  - integrate into LLMs to improve tokenization and support RAG

Benefits to NLP

- enhances quality of MT, NLP and AI applications
- morphological analysis, lemmatization and tokenization
- used for training speech technology models
- enhances entity recognition and extraction
- query processing in IR applications
- supports automatic verb conjugation
- part-of-speech analysis and POS tagging