DiaLEX:
Arabic Dialects Full-Form Lexicon

Jack Halpern
The CJK Dictionary Institute

LREC 2026 Industry Day
Mallorca, Spain
May 14, 2026

What is a full-form lexicon?

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

Palestinian Arabic Wordforms

DiaLEx is a computational lexicon that provides comprehensive coverage of all wordforms.

inflected forms	بَيْتْ بْيُوتْ	*bēt* *byūt*	*house* *houses*
declined forms	كَاتِبْ	*kātib*	writer (genitive)
procliticized forms	وِلْكَاتِبْ	*wilkātib*	and to (a) writer
encliticized forms	كَاتِبَكْ	*kātibak*	your writer
conjugated forms	كَتَبْتْ	*katabt*	I wrote

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

- comprehensive full form entries
- rich morphology: all inflected, cliticized and negated forms
- numerous orthographic variants
- high frequency proper nouns (personal/place names)
- fully vocalized and unvocalized Arabic
- accurate phonemic phonetic transcriptions
- all wordforms cross-referenced to lemma

DiaLEX 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

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

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
- **Pedagogical applications**
  - complete 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
- supports morphological analysis, including 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

Thank You

Muchas gracias

شكرا جزيلا