ArabLEX: A Comprehensive Full-Form Lexicon for Arabic NLP and Speech Technology

Jack Halpern
The CJK Dictionary Institute

Yannis Haralambous
IMT Atlantique

LREC 2026
Mallorca, Spain
May 13, 2026

السَّلَامُ عَلَيْكُمْ جَمِيعًا. ٱسْمِي جَاك هَالْبِرْن، وَأَنَا الرَّئِيسُ التَّنْفِيذِيُّ لِلْمُؤَسَّسَةِ الْمُعْجَمِيَّةِ لِلُّغَاتِ الشَّرْقِيَّةِ CJK فِي الْيَابَانْ.

Good afternoon. My name is Jack Halpern, CEO of The CJK Dictionary Institute based in Japan.

Overview

1. **Background: What is a Full-Form Lexicon?**
   Definition and importance for Arabic NLP
2. **Introducing ArabLEX**
   Scope, scale, and distinctive features
3. **Resource Architecture**
   Core modules and example data
4. **Compilation Methodology**
   Sources, generation, validation
5. **Comparison with Existing Resources**
   ArabLEX vs. Camel Morph
6. **Practical Applications**
   MT, speech, NER, LLMs, pedagogy
7. **Availability & Closing Remarks**

What is a full-form lexicon?

- a computational lexicon that provides comprehensive coverage of all wordforms
- in Arabic this includes all inflected, conjugated, declined, and cliticized forms
- ordinary dictionaries only contain canonical forms like *eat*
- full-form dictionaries include inflected forms like *ate*, *eating*, *eaten* and plurals (*boys*) as well as clitics (*boy's*)

Arabic wordforms

inflected forms	بُيُوتٌ	*buyūtun* ‘houses‘ (plural)
declined forms	كَاتِبٍ	*kātibin* 'writer' (genitive)
procliticized forms	وَلِكَاتِبٍ	*walikātibin* 'and to (a) writer'
encliticized forms	كَاتِبُكَ	*kātíbuka* 'your writer'
conjugated forms	كَتَبْتُ	*katábtu* 'I wrote'

What is *ArabLEX*

- a large-scale full-form Arabic lexicon with 570 million entries
- provides comprehensive coverage for inflected, conjugated and cliticized forms
- includes a rich set of attributes for NLP and language technology

Distinctive Features

- created by linguists, lexicographers and professional editors
- covers 570 million full-form entries
- includes general vocabulary and proper nouns with full inflections such as:
  - plurals
  - duals
  - feminine
  - case endings
  - stems
  - conjugated forms
  - proclitics
  - enclitics
- unvocalized and carefully curated fully vocalized Arabic

Distinctive Features (continued)

- carefully curated phonemic and IPA transcriptions for all entries
- rich grammatical tags include broken plurals, person, transitivity, case, and more
- comprehensive coverage of orthographic variants
- all wordforms are cross-referenced to their lemma (canonical form)

ArabLEX Modules

The full set of *ArabLEX* consists of the following major four modules:

Module	Description	Quantities
DAG	Database of Arabic General Vocabulary	120 million
DAN	Database of Arabic Names	218 million
DAF	Database of Arabic Foreign Names	226 million
DAP	Database of Arabic Place Names	6 million

Cliticized General Vocabulary - Morphological

ARAB_V	ENC_BW	STEM_V	STEM_BW	PROC_V	PROC_BW
وَكَاتِبٌ	-N	كَاتِب	kaAtib	وَ	wa-
وَكَاتِبُ	-u	كَاتِب	kaAtib	وَ	wa-
وَكَاتِبِي	-iy	كَاتِب	kaAtib	وَ	wa-
وَكَاتِبُكَ	-uka	كَاتِب	kaAtib	وَ	wa-
وَكَاتِبُكِ	-uki	كَاتِب	kaAtib	وَ	wa-
وَكَاتِبُهُ	-uhu	كَاتِب	kaAtib	وَ	wa-
وَكَاتِبُهَا	-uhaA	كَاتِب	kaAtib	وَ	wa-
وَكَاتِبُنَا	-unaA	كَاتِب	kaAtib	وَ	wa-
وَكَاتِبُكُمْ	-ukumo	كَاتِب	kaAtib	وَ	wa-
وَكَاتِبُكُنَّ	-ukun~a	كَاتِب	kaAtib	وَ	wa-
وَكَاتِبُكُمَا	-ukumaA	كَاتِب	kaAtib	وَ	wa-
وَكَاتِبُهُمْ	-uhumo	كَاتِب	kaAtib	وَ	wa-

General Vocabulary – Grammatical Attributes

ARAB_V	ARAB_BW	LEMMA_V	POS	GEN	NUM	CASE	PER
وَكَاتِبٌ	wakaAtibN	كَاتِبٌ	N	M	S	NOM	---
وَكَاتِبُ	wakaAtibu	كَاتِبٌ	N	M	S	NOM	---
وَكَاتِبِي	wakaAtibiy	كَاتِبٌ	N	M	S	NOM	1SC
وَكَاتِبُكَ	wakaAtibuka	كَاتِبٌ	N	M	S	NOM	2SM
وَكَاتِبُكِ	wakaAtibuki	كَاتِبٌ	N	M	S	NOM	2SF
وَكَاتِبُهُ	wakaAtibuhu	كَاتِبٌ	N	M	S	NOM	3SM
وَكَاتِبُهَا	wakaAtibuhaA	كَاتِبٌ	N	M	S	NOM	3SF
وَكَاتِبُنَا	wakaAtibunaA	كَاتِبٌ	N	M	S	NOM	1PC
وَكَاتِبُكُمْ	wakaAtibukumo	كَاتِبٌ	N	M	S	NOM	2PM
وَكَاتِبُكُنَّ	wakaAtibukun~a	كَاتِبٌ	N	M	S	NOM	2PF
وَكَاتِبُكُمَا	wakaAtibukumaA	كَاتِبٌ	N	M	S	NOM	2DC
وَكَاتِبُهُمْ	wakaAtibuhumo	كَاتِبٌ	N	M	S	NOM	3PM

General Vocabulary - Phonological

ARAB_V	CARS	IPA	XSAMPA
وَكَاتِبٌ	wakā́tibun	wa.ˈka:.ti.bun	wa."ka:.ti.bun
وَكَاتِبُ	wakā́tibu	wa.ˈka:.ti.bu	wa."ka:.ti.bu
وَكَاتِبِى	wakā́tibi_	wa.ˈka:.ti.bi	wa."ka:.ti.bi
وَكَاتِبُكَ	wakātíbuka	wa.ka:.ˈti.bu.ka	wa.ka:."ti.bu.ka
وَكَاتِبُكِ	wakātíbuki	wa.ka:.ˈti.bu.ki	wa.ka:."ti.bu.ki
وَكَاتِبُهُ	wakātíbuhu	wa.ka:.ˈti.bu.hu	wa.ka:."ti.bu.hu
وَكَاتِبُهَا	wakātíbuha_	wa.ka:.ˈti.bu.ha	wa.ka:."ti.bu.ha
وَكَاتِبُنَا	wakātíbuna_	wa.ka:.ˈti.bu.na	wa.ka:."ti.bu.na
وَكَاتِبُكُمْ	wakātíbukum	wa.ka:.ˈti.bu.kum	wa.ka:."ti.bu.kum
وَكَاتِبُكُنَّ	wakātibukúnna	wa.ka:.ti.bu.ˈku.n:a	wa.ka:.ti.bu."ku.n:a
وَكَاتِبُكُمَا	wakātibúkuma_	wa.ka:.ti.ˈbu.ku.ma	wa.ka:.ti."bu.ku.ma
وَكَاتِبُهُمْ	wakātíbuhum	wa.ka:.ˈti.bu.hum	wa.ka:."ti.bu.hum
وَكَاتِبُهُنَّ	wakātibuhúnna	wa.ka:.ti.bu.ˈhu.n:a	wa.ka:.ti.bu."hu.n:a

General Vocabulary – Extensive Coverage of Variants

ARAB_V	VARID	VAR_V	VAR_U
ضَحِكٌ	01	ضَحِكٌ	ضحك
ضَحِكٌ	02	ضَحْكٌ	ضحك
ضَحِكٌ	03	ضِحْكٌ	ضحك
ضَحِكُ	01	ضَحِكُ	ضحك
ضَحِكُ	02	ضَحْكُ	ضحك
ضَحِكُ	03	ضِحْكُ	ضحك
ضَحِكِى	01	ضَحِكِى	ضحكى
ضَحِكِى	02	ضَحْكِى	ضحكى
ضَحِكِي	03	ضِحْكِي	ضحكي

Place Names - Romanized

ENGLISH	LEMMA_V	LEMMA_BW
Burkina Faso	بُورْكِينَا فَاسُو	buwrokiynaA faAsuw
Egypt	مِصْرُ	miSoru
Guinea-Bissau	غِينِيَا بِيسَاوْ	giyniyaA biysaAwo
Hong Kong	هُونْغْ كُونْغْ	huwnogo kuwnogo
Japan	اَلْيَابَانُ	AaloyaAbaAnu
Jefferson City	جِيفِرْسُونُ سِيتِي	jiyfirosuwnu siytiy
Libya	لِيبْيَا	liyboyaA
New Jersey	نْيُو جِرْسِي	noyuw jirosiy
Palau	بَالَاوْ	baAlaAwo
Porto-Novo	بُورْتُو نُوفُو	buwrotuw nuwfuw
Red Sea	اَلْبَحْرُ ٱلْأَحْمَرُ	AalobaHoru {lo>aHomaru
Saint Lucia	سَانْتْ لُوتْشِيَا	saAnoto luwto$iyaA

Place Names - Grammatical

ARAB_V	ARAB_BW	LEMMA_V	GEN	NUM	CASE	PER
وَمِصْرُ	wamiSoru	مِصْرُ	F	S	NOM	000
وَمِصْرِي	wamiSoriy	مِصْرُ	F	S	NOM	1SC
وَمِصْرُكَ	wamiSoruka	مِصْرُ	F	S	NOM	2SM
وَمِصْرُكِ	wamiSoruki	مِصْرُ	F	S	NOM	2SF
وَمِصْرُهُ	wamiSoruhu	مِصْرُ	F	S	NOM	3SM
وَمِصْرُهَا	wamiSoruhaA	مِصْرُ	F	S	NOM	3SF
وَمِصْرُنَا	wamiSorunaA	مِصْرُ	F	S	NOM	1PC
وَمِصْرُكُمْ	wamiSorukumo	مِصْرُ	F	S	NOM	2PM
وَمِصْرُكُنَّ	wamiSorukun~a	مِصْرُ	F	S	NOM	2PF
وَمِصْرُكُمَا	wamiSorukumaA	مِصْرُ	F	S	NOM	2DC
وَمِصْرُهُمْ	wamiSoruhumo	مِصْرُ	F	S	NOM	3PM
وَمِصْرُهُنَّ	wamiSoruhun~a	مِصْرُ	F	S	NOM	3PF
وَمِصْرُهُمَا	wamiSoruhumaA	مِصْرُ	F	S	NOM	3DM

Foreign Names - Romanized

ENGLISH	LEMMA_V	TYPE	GEN	RS_FREQ	RG_FREQ
Izabella	إِزَابِيلَا	G	F	-	0025717
Jack	جَاك	GS	MF	0015256	0696625
Janet	جَانِيت	GS	MF	0000437	0557605
Juliet	جُولِيت	G	F	-	0030202
Peterson	بِيتِرْسُون	GS	M	0278297	0000756
Schmidt	شْمِيت	S	-	0147034	-
Smith	سْمِيث	GS	MF	2442977	0004733
William	وِيلِيَام	GS	MF	0013373	4133327

Arabic Names – Romanized

(Subset of approx. 200 variants)

R_NAME	LEMMA_V	LEMMA_BW	GEN	TYPE	R_TYPE	FREQ
Muhammad	مُحَمَّد	muHam~ad	M	GS	I	0005300000
Mohammad	مُحَمَّد	muHam~ad	M	GS	V	0003410000
Mohd	مُحَمَّد	muHam~ad	M	GS	V	0002870000
Mohamad	مُحَمَّد	muHam~ad	M	GS	V	0001140395
Muhamad	مُحَمَّد	muHam~ad	M	GS	V	0000258000
Muhd	مُحَمَّد	muHam~ad	M	GS	V	0000191000
Mohamud	مُحَمَّد	muHam~ad	M	GS	V	0000063400
Mukhammad	مُحَمَّد	muHam~ad	M	GS	V	0000059021
Mouhammad	مُحَمَّد	muHam~ad	M	GS	V	0000042000
Mochamad	مُحَمَّد	muHam~ad	M	GS	V	0000036305
Mahamad	مُحَمَّد	muHam~ad	M	GS	V	0000028900

Compilation Methodology

Sources & Reference Materials

- compiled by a multidisciplinary team (lexicographers, linguists, engineers)
- major dictionaries and corpora (CJKI, Oxford, Wiktionary, etc.) used for reference and validation
- includes proprietary CJKI resources and large MSA corpora
- proper nouns extracted from CJKI's proprietary large-scale databases

Quality Control Process

- vocalization and morphology fully validated
- automated error detection and correction
- manual review for unresolved cases
- continuous refinement of rules and exception lists
- iterative cycle repeated over many years → high reliability

Compilation Methodology

Inflection & Conjugation

- nouns/adjectives: full paradigms (gender, number, case)
- verbs generated from CAVE (180 categories, fully vetted)
- covers all tense/person combinations

Cliticization

- cliticized forms (enclitics and proclitics) are generated and fully validated
- cliticization rules are selected according to morphosyntactic context

Constraints & Validation

- over 2,000 rule-based valid combinations
- human-vetted constraint tables
- not simple concatenation → linguistically controlled generation

Enclitics-Proclitic Combinatorics

Clitic templates indicate valid enclitic-proclitic combinations for accurate generation.

Proclitic	**Enclitic**	**Gen Num**
0, >a, wa, fa, >awa, >afa, Aalo, ...	a, u	M → S
0, >a, wa, fa, >awa, >afa	N, FA, FY	M → S
0, >a, wa, fa, >awa, >afa	uhaA, uhu, uhumaA, uhumo, uhun~a, uka, uki, ukumaA, ...	M → S

Comparing ArabLEX with Camel Morph

Dataset	Version	Entries (approx.)	Lemmata (approx.)
ArabLEX	v1.1	530,000,000	—
Camel Morph	—	535,000,000	110,000
ArabLEX	v1.2	570,000,000	390,000

Note: Entry counts are not strictly comparable due to structural differences between ArabLEX and Camel Morph.

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
- **Named-entity recognition**
  - dramatically improved due to comprehensive coverage of inflected names compared to algorithmic methods
- **LLM model training**
  - improve tokenization and morphological features to support RAG

Benefits to NLP

- enhances quality of MT, NLP and AI applications
- supports morphological analysis, including lemmatization and tokenization
- supplements corpora for training speech technology models
- improves accuracy of entity recognition and extraction
- support for query processing in information retrieval applications
- supports automatic verb conjugation and verb lemmatization
- part-of-speech analysis and POS tagging
- accurate determination of the root for each wordform

Availability

- subsets available free for academic research
- commercial licensing available through CJKI
- also available through ELRA

Thank You

Muchas gracias

شكرا جزيلا

www.cjk.org • jack@cjki.org