ArabLEX:
Arabic Full-form Lexicon
معجم اللغة العربية الكامل

Presented by Jack Halpern, CEO
The CJK Dictionary Institute
日中韓辭典研究所

LREC 2022 Industry Day
June 22, 2022

What is a full-form lexicon?

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

Arabic wordforms

inflected forms	بَيْتٌ	*báytun* ('house') -> بُيُوتٌ *buyūtun* ('houses')
declined forms	كَاتِبٍ	*kā́tibin* 'writer' (genetive)
procliticized forms	وَلِكَاتِبٍ	*walikā́tibin* 'and to writer'
encliticized forms	كَاتِبُكَ	*kātíbuka* 'your writer'
conjugated forms	كَتَبْتُ	*katábtu* 'I wrote'

What is *ArabLEX*

- a large-scale full form Arabic lexicon with 530 million entries
- provides comprehensive coverage for inflected, conjugated and cliticized forms
- includes a rich set of attributes for natural language processing

Distinctive Features

- created by specialists in Arabic morphology and computational lexicography
- first release covers 530 million full form entries
- includes general vocabulary and proper nouns with full inflections such as:

plurals	conjugated forms
duals	proclitics
feminine	enclitics
case endings	roots
stems	

- unvocalized and precisely fully vocalized Arabic

Distinctive Features (continued)

- accurate phonemic transcriptions and IPA for all entries
- millions of orthographic variants
- rich grammatical codes include POS, person, gender, and case
- 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	83 million
DAN	Database of Arabic Names	218 million
DAF	Database of Arabic Foreign Names	226 million
DAP	Database of Arabic Place Names	6 million

Arabic General Vocabulary - Grammatical (DAG)

ARAB_V	ARAB_BW	LEMMA_V	POS	GEN	NUM	CASE	PER
وَكَاتِبٌ	wakaAtibN	كَاتِبٌ	N	M	S	NOM	000
وَكَاتِبُ	wakaAtibu	كَاتِبٌ	N	M	S	NOM	000
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

Arabic General Vocabulary - Phonological (DAG)

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
وَكَاتِبِي	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

Arabic General Vocabulary - Morphological (DAG)

ARAB_V	ENC_V	ENC_BW	STEM_V	STEM_BW	PROC_V	PROC_BW	ROOT
وَكَاتِبٌ	ٌ	N	كَاتِب	kaAtib	وَ	wa	ك-ت-ب
وَكَاتِبُ	ُ	u	كَاتِب	kaAtib	وَ	wa	ك-ت-ب
وَكَاتِبِي	ِي	iy	كَاتِب	kaAtib	وَ	wa	ك-ت-ب
وَكَاتِبُكَ	ُكَ	uka	كَاتِب	kaAtib	وَ	wa	ك-ت-ب
وَكَاتِبُكِ	ُكِ	uki	كَاتِب	kaAtib	وَ	wa	ك-ت-ب
وَكَاتِبُهُ	ُهُ	uhu	كَاتِب	kaAtib	وَ	wa	ك-ت-ب
وَكَاتِبُهَا	ُهَا	uhaA	كَاتِب	kaAtib	وَ	wa	ك-ت-ب
وَكَاتِبُنَا	ُنَا	unaA	كَاتِب	kaAtib	وَ	wa	ك-ت-ب
وَكَاتِبُكُمْ	ُكُمْ	ukumo	كَاتِب	kaAtib	وَ	wa	ك-ت-ب
وَكَاتِبُكُنَّ	ُكُنَّ	ukun~a	كَاتِب	kaAtib	وَ	wa	ك-ت-ب
وَكَاتِبُكُمَا	ُكُمَا	ukumaA	كَاتِب	kaAtib	وَ	wa	ك-ت-ب
وَكَاتِبُهُمْ	ُهُمْ	uhumo	كَاتِب	kaAtib	وَ	wa	ك-ت-ب

Arabic General Vocabulary - Orthographical (DAG)

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
ضَحِكِي	01	ضَحِكِي	ضحكي
ضَحِكِي	02	ضَحْكِي	ضحكي
ضَحِكِي	03	ضِحْكِي	ضحكي

Arabic Place Names - Romanized (DAP-ROM)

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 اَلْأَحْمَرُ	AalobaHoru {lo>aHomaru
Saint Lucia	سَانْتْ لُوتْشِيَا	saAnoto luwto$iyaA

Arabic Place Names - Grammatical (DAP-MOR)

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

Arabic Place Names - Phonological (DAP-MOR)

ARAB_V	CARS	IPA	XSAMPA
وَمِصْرُ	wamíṣru	wa.ˈmis̱.ru	wa."mis_-.ru
وَمِصْرِي	wamíṣri̱	wa.ˈmis̱.ri	wa."mis_-.ri
وَمِصْرُكَ	wamíṣruka	wa.ˈmis̱.ru.ka	wa."mis_-.ru.ka
وَمِصْرُكِ	wamíṣruki	wa.ˈmis̱.ru.ki	wa."mis_-.ru.ki
وَمِصْرُهُ	wamíṣruhu	wa.ˈmis̱.ru.hu	wa."mis_-.ru.hu
وَمِصْرُهَا	wamíṣruha̱	wa.ˈmis̱.ru.ha	wa."mis_-.ru.ha
وَمِصْرُنَا	wamíṣruna̱	wa.ˈmis̱.ru.na	wa."mis_-.ru.na
وَمِصْرُكُمْ	wamíṣrukum	wa.ˈmis̱.ru.kum	wa."mis_-.ru.kum
وَمِصْرُكُنَّ	wamiṣrukúnna	wa.mis̱.ru.ˈku.nːa	wa.mis_-.ru."ku.nːa
وَمِصْرُكُمَا	wamiṣrúkuma̱	wa.mis̱.ˈru.ku.ma	wa.mis_-."ru.ku.ma
وَمِصْرُهُمْ	wamíṣruhum	wa.ˈmis̱.ru.hum	wa."mis_-.ru.hum
وَمِصْرُهُنَّ	wamiṣruhúnna	wa.mis̱.ru.ˈhu.nːa	wa.mis_-.ru."hu.nːa
وَمِصْرُهُمَا	wamiṣrúhuma̱	wa.mis̱.ˈru.hu.ma	wa.mis_-."ru.hu.ma

Arabic Place Names - Orthographical (DAP-MOR)

ARAB_V	ENGLISH	VARID	VAR_U
أَنْغُولَا	Angola	01	أنغولا
أَنْغُولَا	Angola	02	انغولا
أَنْغُولَا	Angola	03	أنجولا
أَنْغُولَا	Angola	04	انجولا
أَنْغُولَا	Angola	05	أنغوله
أَنْغُولَا	Angola	06	انغوله
أَنْغُولَا	Angola	07	أنجوله
أَنْغُولَا	Angola	08	انجوله

Arabic Foreign Names - Romanized (DAF-ROM)

ENGLISH	LEMMA_V	TYPE	GEN	RS_FREQ	RG_FREQ
Halpern	هَالْبِرْن	S	-	0004121	-
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 Foreign Names - Morphological (DAF-MOR)

ARAB_V	ARAB_BW	GEN	NUM	CASE	PER
وَلِجَاكَيْنِ	walijaAkayoni	C	D	GEN	000
وَلِجَاكَيْ	walijaAkayo	C	D	GEN	000
وَلِجَاكَيَّ	walijaAkay~a	C	D	GEN	1SC
وَلِجَاكَيْكَ	walijaAkayoka	C	D	GEN	2SM
وَلِجَاكَيْكِ	walijaAkayoki	C	D	GEN	2SF
وَلِجَاكَيْهِ	walijaAkayohi	C	D	GEN	3SM
وَلِجَاكَيْهَا	walijaAkayohaA	C	D	GEN	3SF
وَلِجَاكَيْنَا	walijaAkayonaA	C	D	GEN	1PC
وَلِجَاكَيْكُمْ	walijaAkayokumo	C	D	GEN	2PM
وَلِجَاكَيْكُنَّ	walijaAkayokun~a	C	D	GEN	2PF
وَلِجَاكَيْكُمَا	walijaAkayokumaA	C	D	GEN	2DC
وَلِجَاكَيْهِمْ	walijaAkayohimo	C	D	GEN	3PM
وَلِجَاكَيْهِنَّ	walijaAkayohin~a	C	D	GEN	3PF

Arabic Names - Romanized (DAN-ROM)

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

Database of Arabic Names - Morphological (DAN-MOR)

ARAB_V	ARAB_BW	LEMMA_V	GEN	NUM	CASE	PER
وَمُحَمَّدِينَ	wamuHam~adiyna	مُحَمَّد	M	P	GEN	000
وَمُحَمَّدِي	wamuHam~adiy	مُحَمَّد	M	P	GEN	000
وَمُحَمَّدِيَّ	wamuHam~adiy~a	مُحَمَّد	M	P	GEN	1SC
وَمُحَمَّدِيكَ	wamuHam~adiyka	مُحَمَّد	M	P	GEN	2SM
وَمُحَمَّدِيكِ	wamuHam~adiyki	مُحَمَّد	M	P	GEN	2SF
وَمُحَمَّدِيهِ	wamuHam~adiyhi	مُحَمَّد	M	P	GEN	3SM
وَمُحَمَّدِيهَا	wamuHam~adiyhaA	مُحَمَّد	M	P	GEN	3SF
وَمُحَمَّدِينَا	wamuHam~adiynaA	مُحَمَّد	M	P	GEN	1PC
وَمُحَمَّدِيكُمْ	wamuHam~adiykumo	مُحَمَّد	M	P	GEN	2PM
وَمُحَمَّدِيكُنَّ	wamuHam~adiykun~a	مُحَمَّد	M	P	GEN	2PF
وَمُحَمَّدِيكُمَا	wamuHam~adiykumaA	مُحَمَّد	M	P	GEN	2DC

Practical Applications

- Speech technology
  - training ASR and TTS models
- Machine translation
  - enhanced translation quality
- Morphological analysis
  - significantly simplified algorithms
- Pedagogical applications
  - automatic conjugation systems
- Named-entity recognition
  - dramatically improved

Benefits to NLP

- enhances quality of MT, NLP and AI applications
- supports morphological analysis, including stemming, 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

Thank You

شكرا جزيلا

Merci