日中韓辭典研究所
The CJK Dictionary Institute

ArabLEX
ARABIC FULL-FORM LEXICON

with 570 million entries

Overview

ArabLEX is the most comprehensive Arabic computational full form lexicon ever created, covering over 570 million inflected, conjugated, declined, and cliticized wordforms. It is ideally suited for NLP applications like MT, NER, morphological analysis, and for speech technology, including training ASR, TTS and LLM models. No other Arabic lexicon comes close to it in scope, coverage and comprehensiveness.

DAG	General Vocabulary	120 million
DAN	Names	218 million
DAF	Foreign Names	226 million
DAP	Place Names	6 million

ArabLEX is rich in morphological, grammatical, phonological, and orthographic attributes (currently about 30). In addition, it maps all unvocalized forms to their vocalized counterparts and to the lemma, and provides precise phonemic and phonetic transcriptions.

Speech Technology

The quality of Arabic speech technology still lags considerably behind that of other major world languages. The extreme orthographic ambiguity of Arabic has led to unacceptably high error rates in both TTS and ASR. In a survey we found that sometimes over 50%, and even 80%, of the words in a sentence are mispronounced, especially cliticized forms. ***ArabLEX*** offers the following key benefits for speech technology:

- Hundreds of millions of full-form entries, including millions of proper nouns.
- Complete coverage of proclitic and enclitic combinations for inflected wordforms.
- Tens of millions of orthographic variants and alternative pronunciations.
- A comprehensive pronunciation dictionary for accurate orthographic disambiguation.
- Carefully curated phonemic and phonetic transcriptions, including stress and vowel neutralization.

ArabLEX can also significantly enhance the translation accuracy of Arabic MT. Not only can it be integrated into NMT systems to provide comprehensive coverage of cliticized forms, but it can also be used as a special kind of corpus to train the language model and enable more accurate morphological, syntactic, and semantic analysis. In summary, ***ArabLEX*** aims to serve as the ultimate resource for Arabic natural language processing. This unparalleled lexicon is now available to the NLP and AI communities for research and product development.

Grammatical and phonetic attributes (9 out of 28 shown)

ARAB	CARS	IPA	LEMMA	POS	GEN	NUM	CASE	PER
وَكَاتِبٌ	wakấtibun	waˈkaːtibun	كَاتِبٌ	N	M	S	NOM	000
وَكَاتِبُ	wakấtibu	waˈkaːtibu	كَاتِبٌ	N	M	S	NOM	000
وَكَاتِبِي	wakấtibi̱	waˈkaːtibi	كَاتِبٌ	N	M	S	NOM	1SC
وَكَاتِبُكَ	wakātíbuka	wakaːˈtibuka	كَاتِبٌ	N	M	S	NOM	2SM
وَكَاتِبُكِ	wakātíbuki	wakaːˈtibuki	كَاتِبٌ	N	M	S	NOM	2SF
وَكَاتِبُهُ	wakātíbuhu	wakaːˈtibuhu	كَاتِبٌ	N	M	S	NOM	3SM
وَكَاتِبُهَا	wakātíbuha̱	wakaːˈtibuha	كَاتِبٌ	N	M	S	NOM	3SF
وَكَاتِبُنَا	wakātíbuna̱	wakaːˈtibuna	كَاتِبٌ	N	M	S	NOM	1PC
وَكَاتِبُكُمْ	wakātíbukum	wakaːˈtibukum	كَاتِبٌ	N	M	S	NOM	2PM
وَكَاتِبُكُنَّ	wakātibukúnna	wakaːtibuˈkuːna	كَاتِبٌ	N	M	S	NOM	2PF
وَكَاتِبُكُمَا	wakātibúkuma̱	wakaːtiˈbukuma	كَاتِبٌ	N	M	S	NOM	2DC
وَكَاتِبُهُمْ	wakātíbuhum	wakaːˈtibuhum	كَاتِبٌ	N	M	S	NOM	3PM
وَكَاتِبُهُنَّ	wakātibuhúnna	wakaːtibuˈhuːna	كَاتِبٌ	N	M	S	NOM	3PF
وَكَاتِبُهُمَا	wakātibúhuma̱	wakaːtiˈbuhuma	كَاتِبٌ	N	M	S	NOM	3DM
وَكَاتِبُهُمَا	wakātibúhuma̱	wakaːtiˈbuhuma	كَاتِبٌ	N	M	S	NOM	3DF

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.