日中韓辭典研究所
The CJK Dictionary Institute

ArabLEX:
ARABIC FULL-FORM LEXICON

with 530 million entries

Overview

ArabLEX is the most comprehensive Arabic computational full form lexicon ever created, covering over 530 million inflected, conjugated, declined, and cliticized wordforms. It is ideally suited for NLP applications like MT, NER and morphological analysis and especially for speech technology, including training ASR and TTS models. No other Arabic lexicon comes close to it in scope, coverage and comprehensiveness.

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.

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

Enhancing Arabic NLP Technology

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

- Hundreds of millions of full-form entries, including millions of proper nouns.
- Covers all combinations of proclitics and enclitics for inflected wordforms.
- Tens of millions of orthographic variants.
- Exhaustively lists alternative pronunciations for orthographic disambiguation.
- Highly accurate phonemic transcriptions, including stress and vowel neutralization.

ArabLEX can 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.

日中韓辭典研究所
The CJK Dictionary Institute

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
وَكَاتِبُهُنَّ	wakaAtibuhun~a	كَاتِبٌ	N	M	S	NOM	3PF
وَكَاتِبُهُمَا	wakaAtibuhumaA	كَاتِبٌ	N	M	S	NOM	3DM
وَكَاتِبُهُمَا	wakaAtibuhumaA	كَاتِبٌ	N	M	S	NOM	3DF

DAG: Database of Arabic General Vocabulary – some grammatical Information

The CJK Dictionary Institute

The CJK Dictionary Institute (CJKI) was founded in 1993. Its principal activity is the compilation of large-scale dictionary databases of proper nouns and technical terms for CJK (Chinese, Japanese, Korean) and Arabic, currently with over 50 million entries. CJKI has become the world's prime source for CJK lexical resources for the IT industry and software developers, providing high-quality comprehensive dictionary data, educational tools, and consulting services. Based in Saitama, Japan, CJKI is headed by Jack Halpern, editor in chief of *The Kodansha Kanji Learner's Dictionary* and several other dictionaries that have become standard works for learning Japanese.

Jack Halpern (春遍雀來), CEO of The CJK Dictionary Institute, 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. He has been a resident of Japan for over 40 years, but was born in Germany and has lived in France, Brazil, Japan, and the United States. He is an avid polyglot who has studied 18 languages.

The CJK Dictionary Institute, Inc.
34-14, 2-chome, Tohoku, Niiza-shi, Saitama 352-0001 JAPAN
Phone: 048-473-3508 Fax: 048-486-5032
Email: jack@cjki.org Web: www.cjk.org