1

日中韓辭典研究所
The CJK Dictionary Institute

ArabLEX
Comprehensive Arabic Full Form Lexicon
معجم اللغة العربية الكامل

by Jack Halpern

The CJK Dictionary Institute (CJKI) is pleased to announce the release of the **Arabic Full Form Lexicon**, or *ArabLEX*, the most comprehensive Arabic computational lexicon ever created.

Covering approximately 530 million entries. *Full form* means that it includes *all* inflected forms. It covers not only general vocabulary, but also, for the first time, fully inflected proper nouns (personal and place names).

ArabLEX is, quite literally, *the ultimate resource* for Arabic NLP and AI, ideally suited for such applications as morphological analysis, machine translation, speech technology, and deep learning. No other Arabic lexicon comes close to it the scope, coverage and comprehensiveness.

ArabLEX is rich in morphological, grammatical, phonological, and orthographical 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. These features can significantly contribute to the training of language models for NLP and AI applications.

The **Burj Khalifa** (*Alburj*), soaring at 832 meters above the skies of Dubai, is the tallest structure ever created. Similarly, *ArabLEX* is the largest Arabic lexicon ever created. We thus chose *Alburj* as an emblem to epitomizes the essence of *ArabLEX.*

This unparalleled computational lexicon, the fruit of nearly a decade of intense development and validation, is now available to the NLP and AI communities for research and product development.

What is a Full Form Lexicon?

A *full form lexicon* is a computational lexicon that contains all inflected, conjugated, declined, and cliticized forms that occur in a language (referred to as *wordforms*). Unlike ordinary dictionaries, which include only the canonical forms (base lexemes), a full form lexicon includes all wordforms. For example, the full set of wordforms for *eat* includes *eating*, *eaten* and *ate*, while for *boy* it includes *boys*, *boy's* and *boys'*. Arabic morphology is more complicated. Adding the *proclitics* وَ *wa* 'and' لِ li 'to' and the *enclitic* اتهِمَا *tíhimạ* to the stem كَاتِبٌ kātibun 'writer' yields such a complex form as وَلِكَاتِبَاتِهِمَا *walikạtibātíhimạ* 'and to the two female writers'.

Distinctive Features

Various features of *ArabLEX* offer special benefits to developers of Arabic NLP and AI applications, especially speech technology and machine translation.

- Created by a team of specialists in Arabic morphology and computational lexicography.
- The first release in early 2021 will probably cover some 530 million full form entries, including proper nouns.
- Includes all inflected, conjugated, declined, and cliticized wordforms, including plurals, dual, feminine, case endings, conjugated forms, as well as proclitics, enclitics, stems and roots.
- Unvocalized and precisely fully vocalized Arabic.
- Accurate phonemic transcriptions and IPA for all entries.
- Millions of orthographic variants for both vocalized and unvocalized Arabic.
- A large variety of grammatical codes include part-of-speech, person, gender, and case codes -- currently about 30 attributes for each entry.
- All wordforms are cross-referenced to their lemma (canonical form).
- Constantly maintained and expanded.

Enhancing Speech Technology

The quality of Arabic TTS lags considerably behind that of the major languages. One reason is that the Arabic script is highly ambiguous. For example, كاتب can represent as many as *seven* pronunciations. In addition, the complexity of such cliticized forms as ولكاتباتهما *walikạtibātíhimạ*, and the absence of vowels, makes Arabic TTS especially challenging.

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. The time has come to dramatically improve the intolerably low quality of Arabic TTS. *ArabLEX* helps developers

significantly enhance the quality of Arabic TTS by training ASR systems to achieve higher recognition rates.

To summarize, ArabLEX can bring 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 orthographical disambiguation.
- Future versions will provide 'importance flags' to help indicate the most likely alternative.
- Highly accurate phonemic transcriptions for all wordforms, including precise stress and vowel neutralization.

Enhancing Machine Translation

Some issues in Arabic MT are (1) the high orthographic ambiguity, (2) the morphological complexity (forms like ولكاتباتهما are difficult to analyze), (3) the recognition of named entities (which are often cliticized), and (4) the large number of wordforms for Arabic nouns and verbs.

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 be used as a special kind of corpus to train the language model and enable more accurate morphological, syntactic, and semantic analysis.

Conclusions

ArabLEX provides a rich set of morphological and phonological features. It brings the following benefits to speech technology and MT:

- Enhance the quality of MT, NLP and AI applications.
- Full support for accurate morphological analysis, including stemming, lemmatization, segmentation and tokenization.
- Supplements corpora for training speech technology models.
- Improved accuracy of word and entity recognition and extraction.
- Support for query processing in information retrieval applications.
- Support for automatic conjugators for pedagogical and NLP applications.
- Part-of-speech analysis and POS tagging.
- Accurate determination of the root for each wordform.

In summary, *ArabLEX* aims to serve as the ultimate resource for Arabic natural language processing. (see white paper *ArabEX.pdf* for full details).