LDC: PROJECTS

The Iraqi-English Lexicon

(Iraqi-Arabic ⇄ English)

The Dialectal Arabic Dictionary Project

In 2008 LDC and Georgetown University Press began a collaboration to enhance and update three dialectal Arabic dictionaries – Iraqi, Syrian and Moroccan – originally published by GUP in the 1960s.

There were notable differences among the original print dictionaries that impeded comparisons across the dialects and with Modern Standard Arabic (MSA). An important goal of the collaborative effort was to provide a common basis for word recognition across dialects.

Primary Goals of the Iraqi-English Lexicon

- Supplementing the original inventory by adding new Arabic headword entries from recent corpora of Iraqi conversational speech and research on contemporary usage.
- Developing reasonable orthographic conventions for applying the Arabic alphabet to the dialect, to promote ease of use by both language learners and researchers.
- Providing Arabic script spellings and International Phonetic Alphabet (IPA) pronunciations for Iraqi words and phrases.
- Creating both a print dictionary and lexical database.

The Challenge of Colloquial Arabic

While uses of colloquial Arabic in writing are increasing through literature and Internet channels (email, chat, blogs), there are no orthographic standards for regional dialects. This dictionary project presented an opportunity to develop and apply one general approach for normalizing the orthography of multiple Arabic dialects.

We set aside, as a separate problem, the issue of covering the ranges of variability observed in actual written usage so far. Our focus was to provide a consistent baseline to describe the given dialect, and also expose its similarities and differences relative to other forms of the Arabic language.

Method and Rationale for Spelling with MSA

Within each regional dialect community, literacy is based on learning MSA. Also, when native speakers of English learn Arabic, they typically learn MSA before becoming conversant in a given dialect.

The lexicon of each dialect contains a substantial core in common with MSA, and for much of this common core, the system of Semitic roots and root-based derivations remains a regular, rule-based feature of the grammar. The inventory of MSA roots thus provides a stable foundation for describing the roots and words that are shared by the dialects in this way.

However, the use of MSA roots as the basis for spelling cognates in the dialects means that the spellings cannot be interpreted to have the same phonetic representation as they do in MSA. For that reason, we deemed it essential to include pronunciations rendered in IPA.

Among the most important sources of dialect differences are how patterns of sound change affect certain classes of consonants. Each dialect exhibits different changes. Examples that appear in Iraqi are shown below.

MSA	Iraqi	Examples
ك	k, č (g)	كلمة kilma/čilma “a word” كلب čalib /kalib “dog” كلب galub “heart” (ك [g] is rare)
ق	q, g (k, ǰ)	قبلة qubla “kiss” قمر gumar “moon” قتل kital “to kill” قدر ǰidir “a pot”
ث	θ (f, t)	ثمن θaman “price” ثلث θiliθ “one third” ثالولة falu:la “a wart” ثلث tlaθ “three”

Prominent sound change relations between MSA and Iraqi for selected consonants

The Iraqi-English Lexicon

The Iraqi-English Lexicon contains entries drawn from the original print dictionary (digitized and normalized), from Iraqi-Arabic material held by LDC (transcripts of conversational Iraqi speech) and from research on contemporary usage.

Iraqi-English Lexicon Statistics

- 4965 roots (consonant bases) for Iraqi words
- 17,389 Iraqi headwords (lemmas)
- 24,193 English definitions of Iraqi words
- 10,731 English headwords
- 15,399 Iraqi definitions of English words
- 10,609 example phrases (for Iraqi and English entries combined)

Key Linguistic Features

- Reflects current vocabulary and usage
- Provides conventional Arabic script for main entries and example phrases with diacritics
- Provides a standardized phonetic transcription to aid pronunciation
- Organizes Iraqi headwords according to their root and part-of-speech, in accordance with established practice in Arabic lexicography

Technical Characteristics

Dialectal lexicon development at LDC is mediated through a relational database, with an isomorphic set of tables to store the Arabic roots, Arabic and English headwords, definitions and example phrases for each dialect. Software developed at LDC provides web-based tools for searching and editing the lexicon, as well as a procedure for exporting it to a portable XML format.

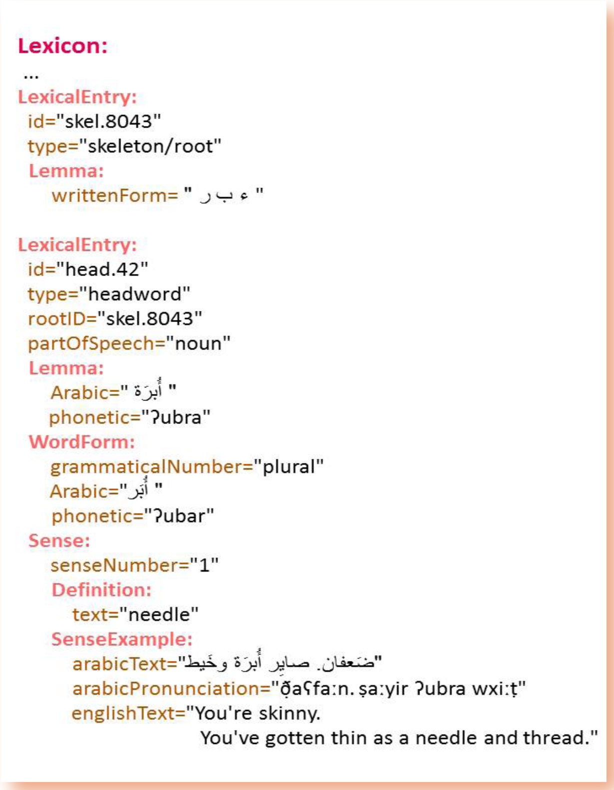
The Iraqi-English Lexicon employs the XML-based Lexical Markup Framework (LMF, ISO 24613), a common model for the creation and use of lexical resources. Using LMF ensures the lexicon's interoperability with a growing range of tools and systems.

Both the Iraqi-English and English-Iraqi halves are represented by a single XML structure, comprising LexicalEntry elements (Iraqi and English headwords). These in turn contain Sense and SenseExample elements (definitions and example phrases), with both Arabic script and IPA orthographies presented in parallel for Iraqi words and phrases.

The roots (or consonantal skeletons) of Iraqi headwords are presented as LexicalEntry elements as well, but without English definitions or IPA transcription. Numeric IDs and sequential order within the XML file serve to link headwords with their specific roots.

Benefits of Dialectal Lexicons

The lexicons developed by LDC in the dialectal dictionary project benefit Arabic learners and teachers as well as those engaged in language-based technology research. English speakers learning or teaching Arabic find that current, accessible lexical information increases proficiency in colloquial Arabic and ability to communicate with a variety of Arabic speakers. Natural language processing applications include speech synthesis and language recognition and identification, among other areas.



```
Lexicon:
 ...
LexicalEntry:
  id="skel.8043"
  type="skeleton/root"
  Lemma:
    writtenForm= " ء ب ر "

LexicalEntry:
  id="head.42"
  type="headword"
  rootID="skel.8043"
  partOfSpeech="noun"
  Lemma:
    Arabic=" أُبرَة "
    phonetic="ʔubra"
  WordForm:
    grammaticalNumber="plural"
    Arabic="أُبَر "
    phonetic="ʔubar"
  Sense:
    senseNumber="1"
    Definition:
      text="needle"
    SenseExample:
      arabicText="ضَعفان. صايِر أُبرَة وخَيط"
      arabicPronunciation="ðaʕfa:n. ṣa:yir ʔubra wxi:ṭ"
      englishText="You're skinny.
                   You've gotten thin as a needle and thread."
```

Example of LMF-based Lexicon Structure

Sponsorship

We are grateful for the generous support of our sponsors. Funding for the lexical research and software development described here was provided by the U.S. Department of Education: DOE-IRS P017A050040.

The print version of the Iraqi-English Lexicon is available from Georgetown University Press. The database will be available through LDC.