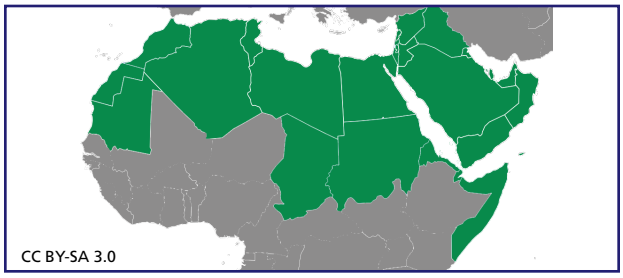
LDC: LANGUAGES
Arabic

An Important World Language

Arabic is the most widely spoken language of the Afro-Asiatic language family with over 300 million speakers concentrated principally in North Africa and the Middle East. It is classified as a Semitic language within the Afro-Asiatic family, and includes a standard form, Modern Standard Arabic (MSA), as well as several colloquial dialects. Arabic is one of the six official languages of the United Nations, reflecting its global importance.

LDC collects and develops digital Arabic language resources of all types, spanning text (newswire, web text, SMS, chat), speech (telephone, broadcast), video (broadcast, web), lexicons, morphological analyzers and language learning software. It also applies a range of annotations to source data, among them transcription, translation, alignment, co-reference and tagging of morphology, syntax and semantics. This work, represented in language resources distributed through LDC's catalog, supports ongoing research and human language technology development including automatic speech recognition, machine translation and content extraction.



Distribution of Arabic in the Middle East and North Africa

Challenges for Language Resource Development

Arabic is a highly inflected language with a rich grammar history. Researchers must grapple with issues like the following:

- a complex morphology that marks case and mood through vowel differences
- texts lack short vowels and other diacritics that distinguish words and those grammatical functions
- coexistence of MSA and multiple dialect forms
- orthographic standards for regional dialects are wanting

Addressing the Challenges

LDC applies creative and flexible solutions to the challenges of working in Arabic that solve present needs and prepare the ground for ongoing work. Examples include:

Morphological analysis. LDC's Standard Arabic Morphological Analyzer (SAMA) considers each Arabic word token in all possible prefix-stem-suffix segmentations and lists all known solutions with diacritic marks, morpheme boundaries, part-of-speech labels and glosses.

Syntactic annotation. Relying on traditional grammar, modern grammatical theories of MSA and computational approaches, including the Penn Treebank, LDC developed the Penn Arabic Treebank series of corpora annotated for morphological information, part-of-speech, English gloss at the token level and syntactic structure.

Dialect orthography and normalization. For *Egyptian* Arabic SMS and chat data containing a prevalence of romanized script, LDC worked in partnership with Columbia University to normalize spelling to facilitate morphological analysis and annotation. In its *Iraqi* lexical database, LDC used MSA roots as the basis for dialect spelling cognates with pronunciations rendered in the International Phonetic Alphabet (IPA).

Reading comprehension. LDC developed tools for language learners that provide multiple views of an annotated Arabic text with the ability to display or hide diacritic marks, to listen to readings of the text using an Arabic text-to-speech synthesizer, to view lexical or morphological information for highlighted words and to search for all occurrences of a word in a selected reading.

Projects, Evaluations and Collaborations

TDT, TIDES, EARS, ACE, GALE, MADCAT, and BOLT support Arabic resource development.

LRE, SRE, OpenMT, OpenHaRT, TRECVID, CoNLL and **SemEval** use LDC's Arabic resources.

Collaborators include **Al Akhawayn University, Columbia University, Georgetown University Press** and data collection and annotation teams in Egypt, Morocco and Tunisia.

Linguistic Data Consortium 3600 Market St., Suite 810, University of Pennsylvania, Philadelphia, PA 19104-2653 USA
Telephone: +1.215.898.0464 • Fax: +1.215.573.2175 • ldc@ldc.upenn.edu • www.ldc.upenn.edu

Selected Resources

LDC's catalog contains over 125 Arabic resources in multiple genres, including:

- Arabic broadcast speech and transcripts and parallel, word-aligned and tagged text: broadcast news and conversation, newswire, web text
- Arabic Gigaword Fifth Edition, LDC2011T11 (newswire from 1994-2010)
- Arabic Treebanks: over one million words of newswire, broadcast and web data with morphological and syntactic annotation
- Conversational telephone speech and transcripts: Egyptian, Gulf, Iraqi and Levantine Arabic
- LDC Standard Arabic Morphological Analyzer (SAMA) Version 3.1, LDC2010L01
- Arabic handwriting in diverse styles and genres by writers from many backgrounds

```
▼<token_Arabic>
   لبعض
  ▼<variant>
     lbED
    ▼<solution>
       <lemmaID>baEoD_1</lemmaID>
       <voc>libaEoD</voc>
       <pos>li/PREP+baEoD/NOUN_QUANT</pos>
       <gloss>for/to + some, a few, little, part</gloss>
     </solution>
    ▼<solution>
       <lemmaID>baEoD_1</lemmaID>
       <voc>libaEoDi</voc>
       <pos>li/PREP+baEoD/NOUN_QUANT+i/CASE_DEF_GEN</pos>
       <gloss>for/to + some, a few, little, part + [def.gen.]</gloss>
     </solution>
    ▼<solution>
       <lemmaID>baEoD_1</lemmaID>
       <voc>libaEoDK</voc>
       <pos>li/PREP+baEoD/NOUN_QUANT+K/CASE_INDEF_GEN</pos>
       <gloss>for/to + some, a few, little, part + [indef.gen.]</gloss>
     </solution>
    ▼<solution>
       <lemmaID>baEoD_1</lemmaID>
       <voc>labaEoD</voc>
       <pos>la/PREP+baEoD/NOUN_QUANT</pos>
       <gloss>for/to + some, a few, little, part</gloss>
     </solution>
    ▼<solution>
       <lemmaID>baEoD_1</lemmaID>
       <voc>labaEoDi</voc>
       <pos>la/PREP+baEoD/NOUN_QUANT+i/CASE_DEF_GEN</pos>
       <gloss>for/to + some, a few, little, part + [def.gen.]</gloss>
     </solution>
```

Sample output from LDC Standard Arabic Morphological Analyzer (SAMA) Version 3.1

LDC Arabic Resources by Dialect and Type

	Text	Speech	Video	Lexicon
MSA	√	√	√	√
Gulf	√	√		
Levantine	√	√		
Iraqi	√	√		√
Egyptian	√	√		√
Maghrebi	√			

Papers and Publications

LDC has presented and published numerous papers about its work in Arabic which are found on the LDC Papers page, https://www.ldc.upenn.edu/language-resources/papers/LDC-papers. A sample follows:

The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus (Maamouri, et al., 2004)

Linguistic Resources for Arabic Machine Translation (Bies, et al., 2012)

Developing LFM-XML Bilingual Dictionaries for Colloquial Arabic Dialects (Graff, et al., 2012)

Expanding Arabic Treebank to Speech: Results from Broadcast News (Maamouri, et al., 2012)

Developing ARET: An NLP-based Educational Tool Set for Arabic Reading Enhancement (Maamouri, et al., 2012)

Developing an Egyptian-Arabic Treebank: Impact of Dialectal Morphology on Annotation (Maamouri, et al., 2014)

Transliteration of Arabizi into Arabic Orthography: Developing a Parallel Annotated Arabizi-Arabic Script SMS/Chat Corpus (Bies, et. al, 2014)

Conclusion

LDC's expertise in Arabic language resource development is unmatched. Its resources are widely used, and many of its pioneering approaches have been adopted and built upon, ensuring that progress continues. Among ongoing projects at LDC are the development of additional dialectal Arabic dictionaries and an Arabic heritage corpus tracing the extensive history of the language.