Arabic Dialects in 10 Minutes

Image source: http://en.wikipedia.org/wiki/Basmala

Demographics

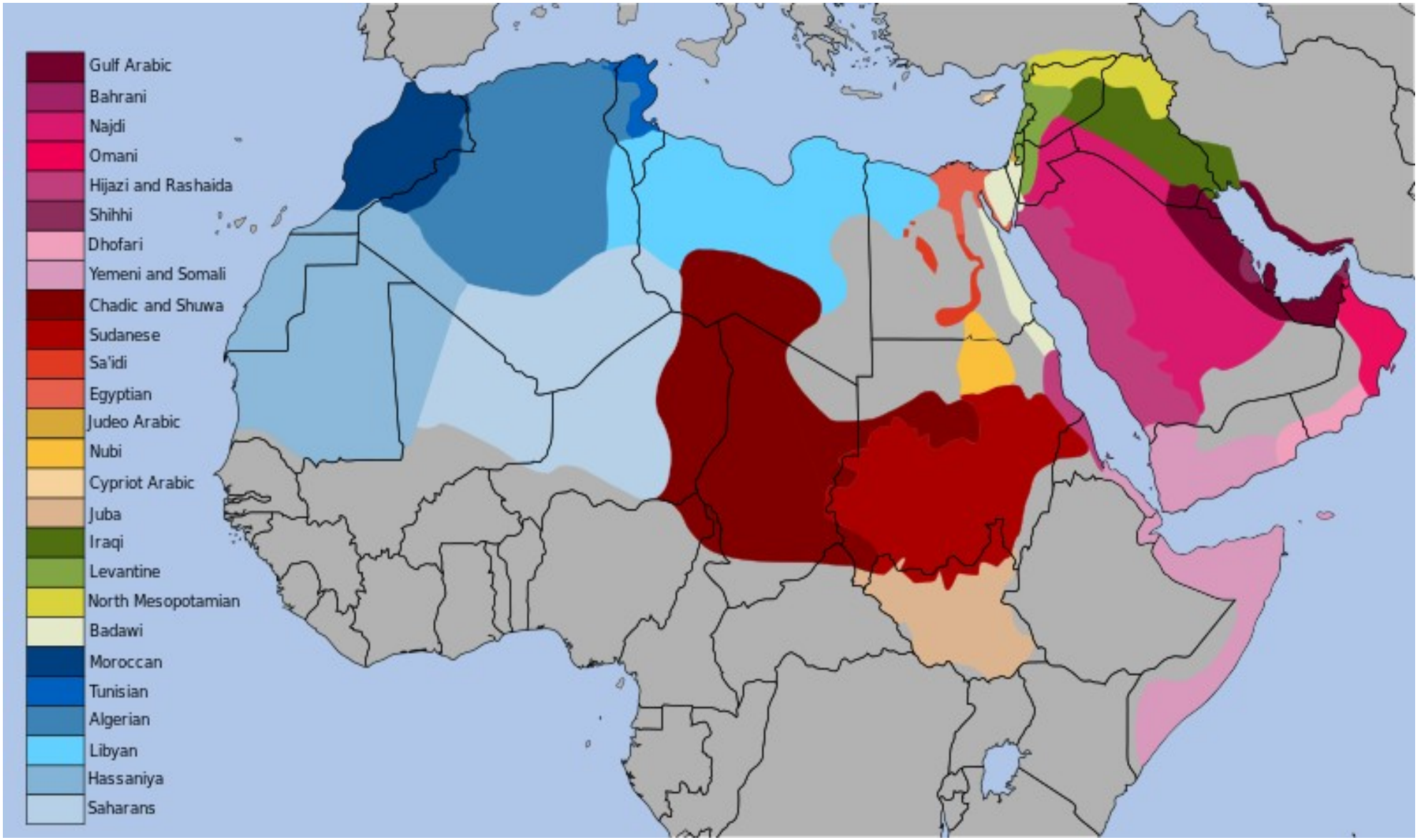


Image source: http://en.wikipedia.org/wiki/Varieties_of_Arabic

Statistics

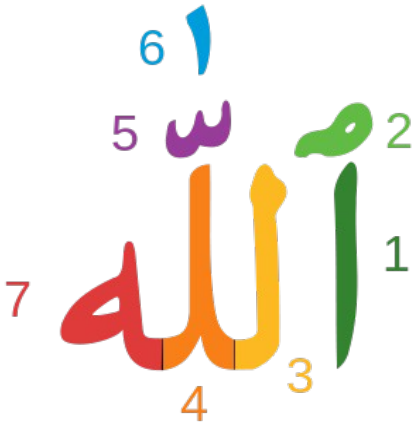
- ~300 million speakers (223 million L1)
- Diglossia: Modern Standard Arabic (MSA) and dialects, spoken in 59 countries
- Ethnologue calls Arabic a Macro-language
  - Egyptian (54m)
  - Algerian (28m)
  - Moroccan (21m)

MSA has no native speakers!

Data source: http://www.ethnologue.com/statistics/size

Orthography

- Written from right to left
- Diacritics and vocalization marks for pronunciation
- No capitalization
- Use of ligature
- Dialects do not have a standard



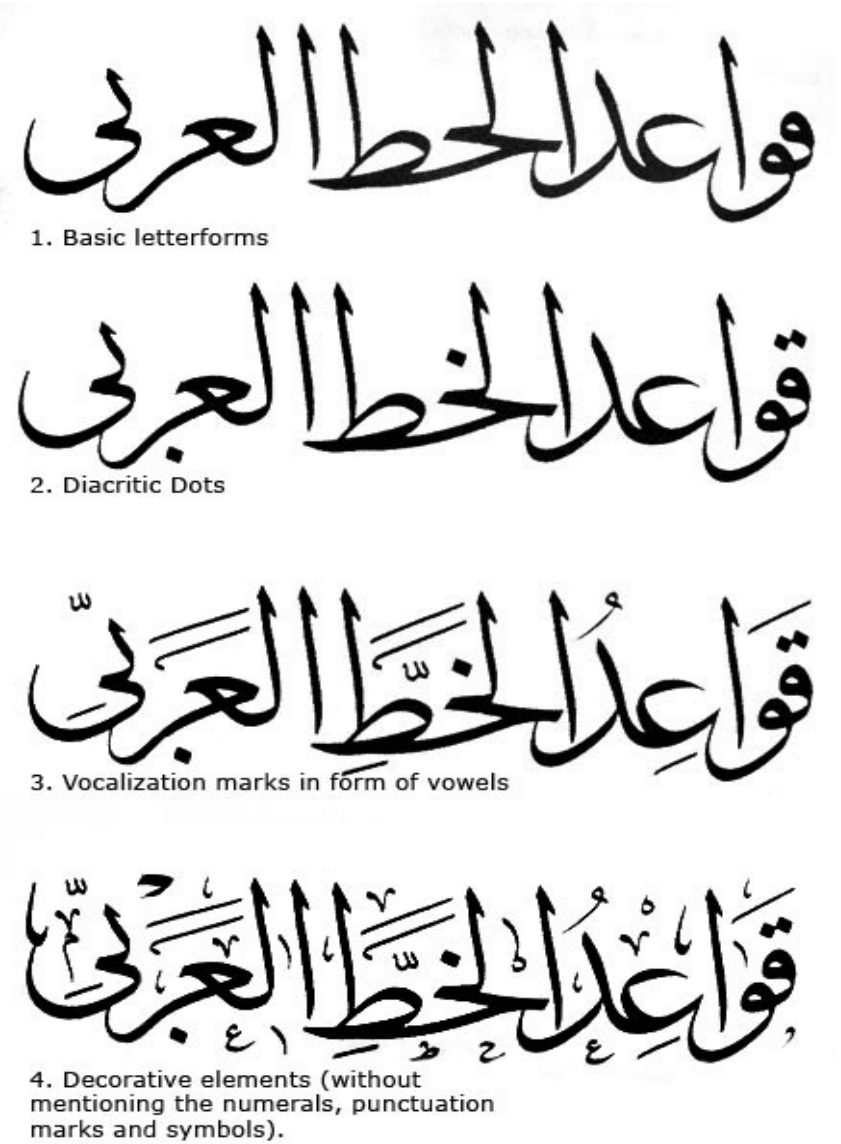


Image source: http://29letters.wordpress.com/2007/05/28/arabic-type-history/

Morphology

- Non-concatenative, root pattern or templatic
  - *k-t-b* (write)
    - *katabtu* (I wrote)
    - *aktabtu* (I dictated)
    - *kā**ā**tabtu* (I corresponded with someone)
    - *kit**ā**b* (book)
    - *k**ā**tib* (writer)
    - *maktabah* (library)
- Dialects have simpler inflectional morphology
  - Ex: MSA has dual markers and nominative case markers, dialects do not

Syntax

- Flexible word order
  - MSA prefers VSO
  - Egyptian prefers SVO
- Adjectives follow the noun they are modifying, and agree with the noun in case, gender, number, and state

Why study dialects?

- Most user generated content is dialectal
- MSA is only a formal standard - news, medium of education
- Dialects are very different from MSA at every level
  - Ex.: State of the art MSA tokenizer with 99.2% accuracy on MSA has 88% accuracy on DA

Source: Mona Diab, Kadri Hacioglu and Daniel Jurafsky. 2007. Automated Methods for Processing Arabic Text: From Tokenization to Base Phrase Chunking.

Computational Efforts

- Arabic NLP: Introduction to Arabic natural language processing, Nizar Habash, Synthesis Lectures on Human Language Technologies
- Processing MSA
  - A maximum entropy word aligner for Arabic-English machine translation, Ittycheriah and Roukos, ACL 2005
  - MADA+TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization, Habash et al., MEDAR 2009
  - Arabic preprocessing schemes for statistical machine translation, Habash and Sadat, 2006
- Processing dialects
  - Parsing Arabic Dialects, Chiang et al. EACL 2006
  - Morphological Analysis and Disambiguation for Dialectal Arabic, Habash et al. NAACL HLT 2013
  - CODACT: Towards Identifying Orthographic Variants in Dialectal Arabic, Dasigi and Diab , IJCNLP 2011
  - Dialectal Arabic to English Machine Translation: Pivoting through Modern Standard Arabic, Salloum and Habash, Proceedings of NAACL-HLT 2013

Resources

- Monolingual Corpora
  - MSA: Arabic Gigaword
  - Dialects: LDC Call Home Project
- Treebanks
  - MSA: Arabic Penn Tree Bank, Prague Dependency Treebank, Columbia Arabic Treebank
  - Dialects: Levantine Arabic Treebank
- Parallel Corpora
  - MSA: UN Corpus
  - Dialects: Egyptian - English parallel corpus at LDC