

### Language Families

Few regions of the world boast as diverse and dense a collection of language families as West Africa. It is home to the three major language families on the African continent: Niger-Congo, Afro-Asiatic and Nilo-Saharan.

Except for Arabic and Hausa, the major languages are from the Niger-Congo family: Yoruba, Igbo, Fulfulde/Pulaar, Akan, and Wolof. However, the remaining language groups are well-represented as the continental sub-region is home to native speakers of over 500 languages. West African languages have the largest concentration of first and second-language speakers on the continent, after Amharic and Swahili in East Africa.

Language	# of Native Speakers
Akan (AKA)	8,300,000
Fulah (FUL)	12,269,800
Hausa (HAU)	24,988,000
Igbo (IBO)	18,000,000
Mandingo (MAN)	4,537,000
Mòoré (MOS)	5,061,700
Ngomba (JGO)	63,000
Yemba (YBB)	300,000
Yoruba (YOR)	19,380,800
Wolof (WOL)	3,976,500

*Number of native speakers of select languages, ISO 639-3 codes in parentheses. Source data: Ethnologue.com*

### Challenges for Language Resource Development

Human language technology (HLT) development relies on digital language resources, including: lexicons, grammars, monolingual and parallel corpora, morphological analyzers, taggers and segmenters.

West African languages present numerous challenges to the researcher, both technological and not, some language specific and some common to the region.

Among these challenges are:

- Complex phonology and morphology (Bantu)
- Verb serialization (Kwa)
- Complex pronoun systems (Yoruba)
- Diaspora forms in addition to multiple dialects (Yoruba)
- Absence of established writing systems (many)

#### Fieldwork in West African Languages

For languages with limited text resources or no writing tradition, speech data and corresponding transcripts provide fundamental information about the language.

In addition to the usual technologies (digital recorders and laptop computers), such field work requires customized transcription software and novel approaches to subject recruitment and elicitation.

Compounding these challenges are dynamic morphophonological processes, such as segment deletion, segment assimilation, vowel harmony, nasalization and word contraction. Also, resource developers must address the lack of standardized orthographies, fonts and encodings.

The interactions of these phenomena make West African languages challenging for human analysts as well as HLT systems.



*The West African sub-region, an area stretching from Mauritania and Liberia inland to Chad and Cameroon*

## Addressing these Challenges

LDC has funded, developed and/or published electronic resources in Mawukakan (Mandingo), Yoruba, and Dschang and Ngomba (Bantu). Each corpus represents a creative and flexible solution to a particular language challenge:

- Tonological and phonetic transcription for portions of Dschang and Ngomba tone paradigms
- Diaspora dialects included in Yoruba lexicon to capture language's global impact
- International Phonetic Alphabet (IPA) used for Mawukakan lexicon
- Tone marking in Mawukakan lexicon to support multiple research applications
- Bidirectional English and French glosses in Mawukakan lexicon to accommodate speakers in a francophone context

## West African Language Resources at LDC

The following data sets can be found in the LDC Catalog:

- Global Yoruba Lexical Database v. 1.0, *LDC2008L03*
- Grassfields Bantu Fieldwork: Dschang Lexicon, *LDC2003L01*
- Grassfields Bantu Fieldwork: Dschang Tone Paradigms, *LDC2003S02*
- Grassfields Bantu Fieldwork: Ngomba Tone Paradigms, *LDC2003S16*
- Mawukakan Lexicon, *LDC2005L01*

### Focus on Yoruba

LDC's Global Yoruba Lexical Database is a set of related dictionaries providing definitions and translations for over 450,000 words from Yoruba and its variants.

The Yoruba language diaspora stretches from Nigeria and Benin to the Caribbean and islands along the southeastern United States coast. The dictionary contains gloss pairs that reflect the language's multiple contexts:

- Yoruba → English (142,000+ words)
- English → Yoruba (226,000+ words)
- Lucumi → Spanish → English → Yoruba (8000+ words)
- Gullah → English → Yoruba (3500+ words)
- Trinidadian Yoruba → English → Yoruba (1000+ words)

The Yoruba dictionary has been tested and found to complement classroom teaching methods at the University of Pennsylvania. It is currently used by Yoruba students and provides relevant and valuable background information on the language and culture.

LDC's work on West African languages continues. Within the Less Commonly Taught Languages program, LDC created multiple resources for Yoruba, including bidirectional parallel texts in English. LDC has also collected Yoruba audio and written texts in preparation for treebank annotation, and a new version of the Yoruba lexicon is in progress.

LDC's ongoing work in Mandingo includes the development of lexicons for Maninkakan, Bambara and Jula. These data sets will complete LDC's resource development for Manding languages.

```
- <lxGroup>
  <lx>a</lx>
  <t>0</t>
- <cGroup>
  <c>Q</c>
- <dGroup>
  <d>yes/no question marker</d>
  <dfi>marque de la question oui/non</dfi>
- <eGroup>
  <e>Kú á?</e>
  <g>Yam?</g>
  <gfi>De l'igname?</gfi>
</eGroup>
- <eGroup>
  <e>Dí à?</e>
  <g>Honey?</g>
  <gfi>Du miel?</gfi>
```

Screenshot of Mawukakan Lexicon, LDC2005L01

## Language Threats and Future Outlook

Current preservation methods are not adequate to prevent many African languages from moving toward extinction. Insufficient funding, a lack of public resources and limited access to online materials hinder research and documentation tasks.

The enormity of the work necessary to develop the languages of this sub-region requires large-scale, sustained, collaborative effort. LDC contributes to this effort as both a creator and publisher of language resources.

LDC's work on West African languages is principally conducted by Dr. Yiwola Awoyale ([awoyale@ldc.upenn.edu](mailto:awoyale@ldc.upenn.edu)) and Dr. Moussa Bamba ([bamba2@ldc.upenn.edu](mailto:bamba2@ldc.upenn.edu)). Their research focuses on Yoruba and Manding languages, respectively.