Computerization of African languages-French dictionaries

DiLAF project: From published dictionaries to XML/LMF lexical resources

http://dilaf.org/

Chantal Enguehard
LINA Laboratory
Nantes, FRANCE
Chantal.Enguehard@univ-nantes.fr

Mathieu Mangeot
GETALP-LIG Laboratory
Grenoble, FRANCE
Mathieu.Mangeot@imag.fr

Dictionaries used

- Niger: Soutéba project 2004-2009

  support of basic education, german cooperation!

  - Hausa-French: 7,823 entries
  - Kanuri-French: 5,994 entries
  - Soŋay zarma-French: 6,916 entries
  - Tamajaq-French: 5,205 entries
- Mali: Bambara-French
  - Father Charles Bailleul (1996 edition)
  - > 10,000 entries

Number of speakers

- Bambara: 4 millions native; 10 millions total
- Fulfulde: 10 to 20 millions
- Kanuri: 9 millions
- Hausa: 40 to 50 millions
- Tamajaq: 5 millions
- Zarma: 2.1 millions

Source: wikipedia

National Languages in Niger

- Spoken languages:

  Hausa 50%, Zarma 21%, Tamajaq 8%, Fulfulde 8%, Kanuri 5%, Arabic 1%, +

- Writing systems
  - 1978: referenced african alphabet
    - Latin alphabet + IPA
    - Special characters: ɓ ɗ ə ɛ ɣ ƙ ɲ ŋ ɔ ƴ
    - Diacritics: âêîôûăãẽĩõũḍḷṣṭẓğǰšɾ
  - 1999: standardization of the alphabets
    - Hausa, Zarma, Kanuri, Tamajaq
  - 2001: law of the 10 national languages

Linguistic situation

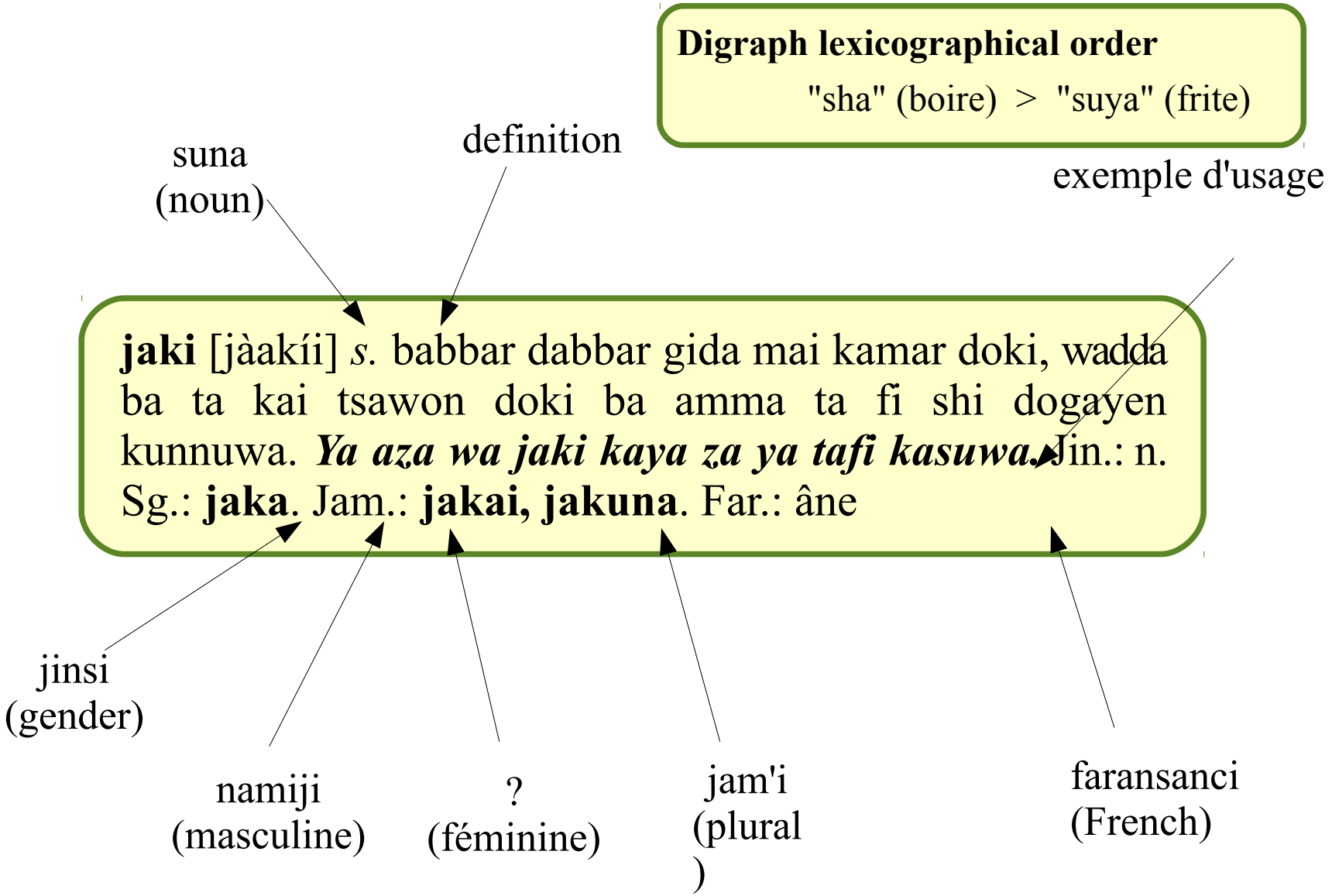
- Recently written languages
    - Problems with hacked fonts (Microsoft, SIL)
    - Problems with unicode and existing resources
        - Lack of some characters in Unicode
- Under-resourced languages
    - No written corpora, no spell-checkers, no MT, etc.
- Specialists of a language (linguists) are rare and busy

Hausa-French dictionary

a b ɓ c d ɗ e f fy g gw gy h i j k kw ky ƙ ƙw ƙy l m n o p r s sh t ts u w y y̆ z

Digraph lexicographical order

"sha" (boire) > "suya" (frite)



jaki [jàakíi] *s.* babbar dabbar gida mai kamar doki, wadda ba ta kai tsawon doki ba amma ta fi shi dogayen kunnuwa. ***Ya aza wa jaki kaya za ya tafi kasuwa.*** Jin.: n. Sg.: **jaka**. Jam.: **jakai, jakuna**. Far.: âne

Bambara-French dictionary

a b c d e ɛ f g h i j k l m n ɲ ŋ o ɔ p r s t u w y z

example

kamalen.ya n. ① condition de jeune homme, jeunesse

màá tɛ se kà à ka kamalenya kɛ n'a ma filìkò kelen kɛ : on ne peut passer sa jeunesse sans faire au moins une fredaine

② en pleine force (pour un homme), bravoure

③ impudicité

(jeune homme.suf abs)

Dictionaries conversion into XML

- Starting point
  - styled text files (.doc)
- Steps
  1) Converting the characters into Unicode
  2) Converting the data into XML
  3) Adding/Replacing tags with regexps
  4) XML validation
  5) Correction of the data

1 Characters conversion into Unicode

- Problem: hacked fonts (mainly from SIL)
- Solution:
  - Convert the chars to Unicode
  - before conversion to XML

origine	Unicode
§	ã
é	ẽ
$	ɲ
ù	ŋ
£	Ɲ

noøe [nóoøèe] *aik.* **ðoye don kar a ga mutum.** ***Da ðarawo ya ga ÷an sanda sai ya noøe.*** *Far.:* disparaître en cachette

noƙe [nóoƙèe] *aik.* **ɓoye don kar a ga mutum**. ***Da ɓarawo ya ga yân sanda sai ya noƙe.*** *Far.:* disparaître en cachette

2 Conversion into XML

- Example entry of the zarma dictionary

abirillu [àbìríllù] *m.* • *avril* • annasaara handu taacanta kaŋ go marsu nda me game ra • *Abirillu, 15, 1974 no Sayni Kunce na hino sambu* • *f/b.* abirillo, abirilley

abiyanso [àbìyànsôo] *m.* • *aéroport* • batama kaŋ ra abiyey ga zumbu • *Tilbeeri nda Dooso sinda abiyanso kaŋ ra abiyo beeri ga zumbu* • *f/b.* abiyansa, abiyansey

abiyo [àbíyò] *m.* • *avion* • naarumay hari no kaŋ ra i ga boro nda jinay daŋ a ma deesi nd'ey • *Jidda no abiyey ga alfujaajey zumandi* • *him.* beene-hi • *f/b.* abiya, abiyey

abunaadam [àbúnăadàm] *m.* • *être humain, personne* • *f/b.* abunaadamo, abunaadamey • *di.* adamayze

- Save .doc to .odt and change .odt to .zip
  - Note: MS XML (.docx) is too verbose and $$
- Unzip
- Open the file content.xml with Notepad++ (or equiv)

3 Replacing tags: choice of language

- It is time to decide which language to use for tags
- English
  - Not understood by the local linguists
  - Not present in the dictionaries
- French
  - OK but still post-colonial spirit
- National languages
  - Dicts are in fact monolingual with translations so OK
  - Useful for setting up a terminology in the language
  - Gives importance to the language

Result of the tagging process

<sanniize>abiyanso</sanniize>

<ciiyaŋ>àbìyànsôo</ciiyaŋ>

<kanandi>m.</kanandi>

<bareyaŋ>aéroport</bareyaŋ>

<feeriji>batama kaŋ ra abiyey ga zumbu</feeriji>

<silmaŋ>Tilbeeri nda Dooso sinda abiyanso kaŋ ra abiyo beeri ga zumbu</silmaŋ>

<f>abiyansa</f>abiyansey

5 Simple corrections

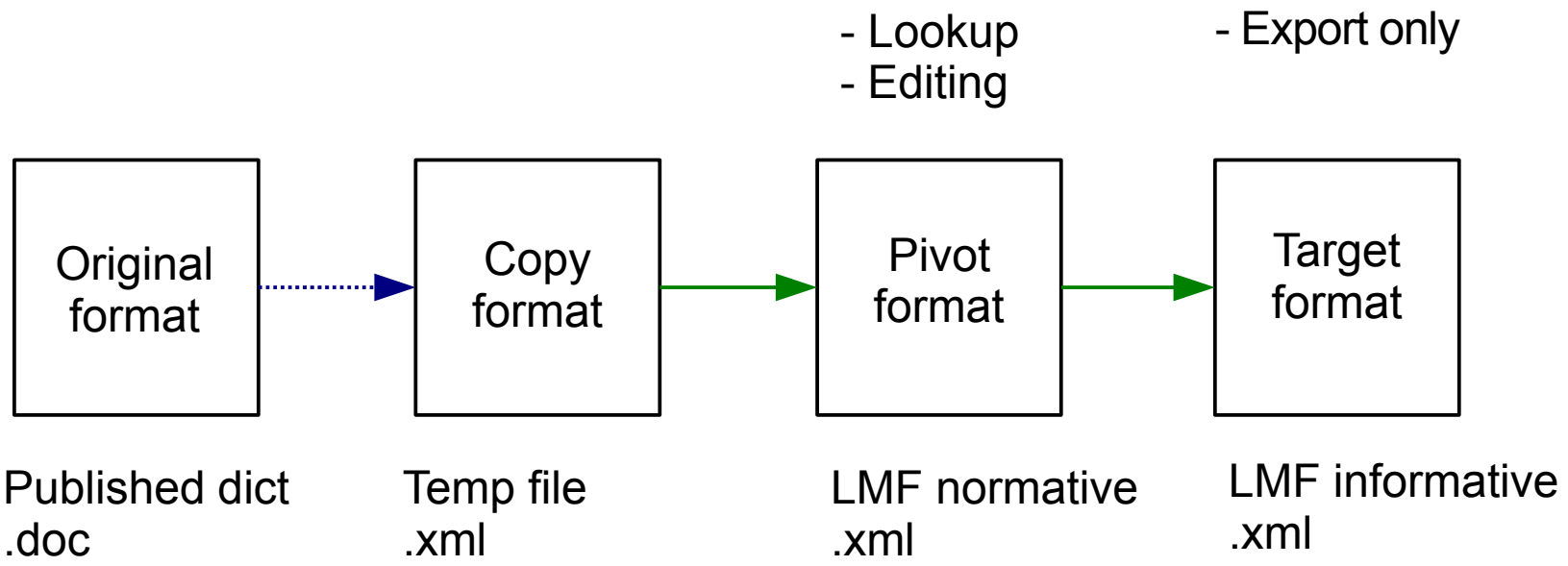
- CSS stylesheet for display
- XSLT stylesheet for cross ref links

  => compact view in the browser

abarba [ábàrbà] *m.* type de banane **banaana dumi no kaŋ i ga haagu ga ŋwa** *Abarba gani ŋwaayaŋ ga hin ga te boro se gunde-kuubi* budde abarbaa abar
abirillu [àbìrílù] *m.* avril **annasaara handu taacanta kaŋ go marsu nda me game ra** *Abirillu, 15, 1974 no Sayni Kunce na hino sambu* abirillo abirill
abiyanso [àbìyànsôo] *m.* aéroport **batama kaŋ ra abiyey ga zumbu** *Tilbeeri nda Dooso sinda abiyanso kaŋ ra abiyo beeri ga zumbu* abiyansa abiyansey
abiyo [àbíyò] *m.* avion **naarumay hari no kaŋ ra i ga boro nda jinay daŋ a ma deesi nd'ey** *Jidda no abiyey ga alfujaajey zumandi* beene-hi abiya abiyey
abunaadam [àbúnăadàm] *m.* être humain, personne abunaadamo abunaadamey adamayze

- Check tags and closed lists of values
  - A small number may be an error

Conversion process towards LMF



Pivot format: respects LMF normative part

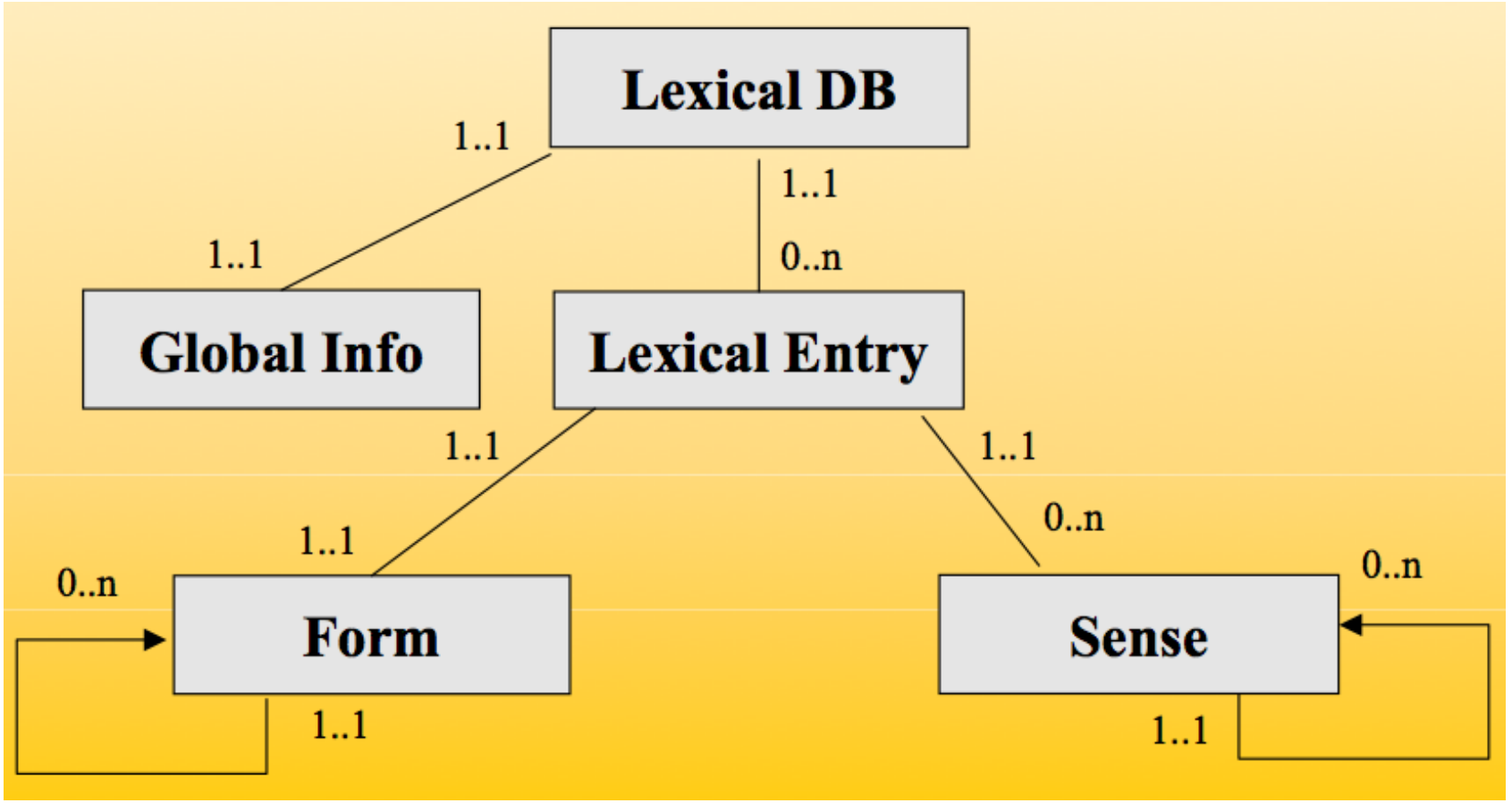
- Structuring the entries
- The structure must respect the LMF core

the normative part of LMF is a meta-model

  - Specifies how the information must be structured
  - Does not specifies how it is represented
    - Any tag name can be used
    - Elements or attributes can be used for same info

Structure of the LMF standard

- Lexical Markup Framework meta model



Result of the structuring

```
<article>
 <bloc-vedette>
  <sanniize>abiyanso</sanniize>
  <ciiyaŋ>[àbìyànsôo]</ciiyaŋ>
 </bloc-vedette>
 <bloc-grammatical>
  <kanandi>m.</kanandi>
  <f>abiyansa</f><b>abiyansey</b>
  <bloc-sémantique>
   <bareyaŋ>aéroport<bareyaŋ>
   <feeriji>batama kaŋ ra abiyey ga zumbu.</feeriji>
   <silmaŋ>Tilbeeri nda Dooso sinda abiyanso kaŋ ra abiyo beeri
ga zumbu</silmaŋ>
  </bloc-sémantique>
 </bloc-grammatical>
</article>
```

Target format: LMF informative view

```
<LexicalEntry id="abiyanso">
  <Lemma>
    <feat att="writtenForm" val="abiyanso"/>
    <feat att="phoneticForm" val="àbìyànsôo"/>
  </Lemma>
    <feat att="partOfSpeech" val="m."/>
  <Sense id="1">
    <Equivalent><feat att="writtenForm" val="aéroport"/></Equivalent>
    <Definition><feat att="writtenForm" val="batama kaŋ ra abiyey ga
zumbu"/></Definition>
    <Context>
          <TextRepresentation><feat att="language" val="dje"/><feat att="writtenForm"
       val="Tilbeeri nda  Dooso sinda abiyanso kaŋ ra abiyo beeri ga
       zumbu."/></TextRepresentation>
          <TextRepresentation><feat att="language" val="fra"/><feat att="writtenForm"
       val="Tillabery et Dosso ne possèdent pas d'aéroports."/></TextRepresentation>
    </Context>
  </Sense>
</LexicalEntry>
```

Managing the data with the Jibiki platform

- Java Web server platform
  - user and groups management
  - heterogeneous entry lookup (with CDM pointers)
  - generic entry editing (with XML schema)
  - remote programming (with REST API)
- CMS for lexical data
- Open Source
  - available at http://jibiki.ligforge.imag.fr/

How to handle different microstructures?

- Pb: each dictionary has its own microstructure
- How to handle them easily?
- Solution 1: convert them into one unique format (LMF)

  => loss of information
- Solution 2: write code for each structure

  => not generic
- Solution 3: use common pointers for each part of information into each microstructure

Common Dictionary Markup

CDM Pointer	FeM	OHD	JMdict
Volume	/volume	/volume	/JMdict
Entry	/volume/entry	/volume/se	/JMdict/entry
Entry ID	/volume/entry/@id		/JMdict/entry/ent_seq
Headword	/volume/entry/headword	/volume/se/hw	/JMdict/entry/k_ele/keb
Pron	/volume/entry/prnc	/volume/se/pr/ph	
PoS	//sense-list/sense/pos-list	/volume/se/hg/ps	/JMdict/entry/sense/pos
Domain		//u	
Example	//sense1/expl-list/expl/fra	//le	/JMdict/entry/sense/gloss