The Pulaar Language

- Pulaar (also known as Fulfulde) is the most widely spoken of the West Atlantic languages (Niger-Congo) of Africa.
- In some states (Futa Jalon in Guinea and northern Nigeria, in particular), the elites devoted considerable attention to the development of Pulaar pedagogy.

This included an *ajami* system, that is, the writing (usually for recitation and instruction) of Pulaar texts in the Arabic script.
- Indigenous authorities, writers, and expatriate linguists have opted for the Roman script in recent decades.
- An important and largely untapped resource remains available in the Arabic-language texts written largely in the 18th and 19th centuries.

Fulbe Migration and Distribution

Representing Pulaar digitally

- **Digitization**
  - Digitization is a process of converting an analog source material into a computer-readable format
- **Language Digitization**
  - Language digitization is a process of representing language in a computer-readable format
  - Representing sound (phonology, phonetics) through audio digitization
  - Representing structure (syntax, semantics, discourse) through mark-up
  - Combining sound and structure with SMIL

David Robinson, Moustapha Kane, Sonja Fagerberg-Diallo,
"Une vision iconoclaste de la guerre sainte d’al-Haj Oumar Taal, “
Audio Digitization

1. Prepare Analog Originals
2. Make analog copies
3. Transfer digital data via S/PDIF
4. Save file as PCM Wav 96 kHz/24 bit
5. Burn 2x copies to CD
6. Batch convert to 22,050 Hz/16 bit
7. Transfer files for editing
8. Delete preservation copy
9. Enter metadata record
10. Record derivative object

Text Digitization - OCR

- Optical Character Recognition (OCR)
  - There exist no Pulaar-specific OCR or proofing tools.
  - We have developed a Pulaar-compliant OCR methodology

The matrix matching process maps Pulaar characters to Unicode codes
The feature analysis process is updated by adding new bitmap shapes to its inventory and assigning Unicode codes to them

Text Digitization - Character Encoding

- There is no standard character set for Pulaar.
- Unicode provides a unique number (code) for more characters and languages than any existing system.
- Unicode is platform-independent and has been adopted by such industry leaders as Apple, HP, IBM, Microsoft, Oracle, Sun, and many others.
- Unicode is required by modern standards such as XML and Java.
- Unicode is supported by many operating systems and all modern web browsers.

Tierarchical and Sequential Nature of Linguistic data

<table>
<thead>
<tr>
<th>Discourse</th>
<th>Paragraph</th>
<th>Sentence</th>
<th>Phrase</th>
<th>Word</th>
<th>Morpheme</th>
<th>Phoneme</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>⬃</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>⬃</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Unicode codes for Pulaar characters

<table>
<thead>
<tr>
<th>Character</th>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>&amp;#385</td>
<td>✁</td>
</tr>
<tr>
<td>b</td>
<td>&amp;#595</td>
<td>✁</td>
</tr>
<tr>
<td>D</td>
<td>&amp;#394</td>
<td>✁</td>
</tr>
<tr>
<td>d</td>
<td>&amp;#599</td>
<td>✁</td>
</tr>
<tr>
<td>L</td>
<td>&amp;#331</td>
<td>✁</td>
</tr>
<tr>
<td>l</td>
<td>&amp;#330</td>
<td>✁</td>
</tr>
<tr>
<td>Y</td>
<td>&amp;#435</td>
<td>✁</td>
</tr>
<tr>
<td>y</td>
<td>&amp;#436</td>
<td>✁</td>
</tr>
<tr>
<td>N</td>
<td>&amp;#209</td>
<td>✁</td>
</tr>
<tr>
<td>n</td>
<td>&amp;#241</td>
<td>✁</td>
</tr>
</tbody>
</table>
How do we represent linguistic data in a computer-readable format?

<table>
<thead>
<tr>
<th>Relational database</th>
<th>Structured text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Offers powerful analysis tools</td>
<td>Offers good analysis tools</td>
</tr>
<tr>
<td>Fails to capture the hierarchical and sequential nature of linguistic data</td>
<td>Promises to capture the hierarchical and sequential nature of linguistic data</td>
</tr>
</tbody>
</table>

Oracle
MySQL
MS Access

SGML
XML
TEI

Relational database
Structured text

How do we represent linguistic data in a computer-readable format?

Sentence XML mark-up

```xml
<?xml version="1.0" encoding = "UTF-8"?>
<S>
  <DP>
    <NP>
      <N>John</N>
    </NP>
  </DP>
  <VP>
    <V>works</V>
    <PP>
      <P>in</P>
      <DP>
        <D>the</D>
        <NP>
          <N>factory</N>
        </NP>
      </DP>
    </PP>
  </VP>
</S>
```

John works in the factory.

Linguistic Mark-up – example 1

Sentence
XML mark-up

```
<?xml version="1.0" encoding = "UTF-8"?>
<conversation>
  <male>Hi, Mary, how you doing?</male>
  <female>I am, like, OK, like I guess</female>
</conversation>
```

Mary, how do you like my new baseball hat?
It's, like, OK, I guess.

Linguistic Mark-up – example 2

Conversation
XML mark-up

```
<?xml version="1.0" encoding = "UTF-8"?>
<conversation>
  <male>Mary, how do you like my new baseball hat?</male>
  <female>It's, like, OK, I guess.</female>
</conversation>
```

Combining Text and Audio with SMIL

Synchronized Multimedia Integration Language (SMIL) is a simple but powerful markup language for assembling multimedia presentations.

We use SMIL to assemble time-synchronized multimedia language corpora, as well.

We have developed a methodology for making SMIL corpora:

Transcribe and time-stamp audio with Transcribe tool
Convert trs into rt, QT, and TEI
With PHP, XSL and Sablotron
Assemble SMIL
With PHP

Web example

Thank you!