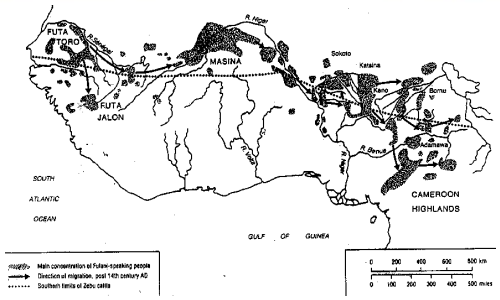
Digital Representation of Pulaar

David Robinson, Michigan State University
Cheikh Babou, Michigan State University
Bartek Plichta, Michigan State University

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
Source: H.P White and M. Gleave, *An Economic Geography of West Africa (1971)*

7. Jakalel sumi haa gasi, Sutukullenaaɓe nanngaa, baame njahi haa to Kuulum, daa-laabi njokki laawol. [23]

8. Yaa 'Alla ! Yaa Rabbi ! Yoo 'Alla hoynu jaggal Kuñakaari, koydo beelol. [24]

7. Jakkalel fut incendié et réduit à néant ; les habitants de Sutukulle furent capturés ; des détachements marchèrent sur Kuulum, d'où des filles de prisonniers se mirent en route.

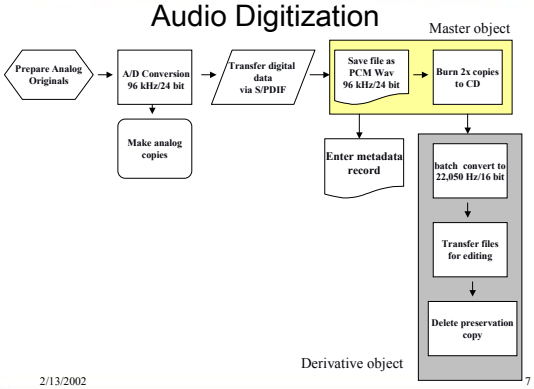
8. Oh Dieu ! Oh Seigneur ! Que Dieu répande la honte sur le seigneur de Kuñakaari, ce vaut-rien[1].

David Robinson, Moustapha Kane, Sonja Fagerberg-Diallo, "Une vision iconoclaste de la guerre sainte d'al-Hajj Umar Taal, " *Cahiers d'Etudes Africaines*, 1994.

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*

Audio Digitization

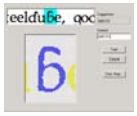


Text Digitization - OCR

- 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

Individual characters of the scripts that humans use to record and transmit their languages are encoded in the form of binary numerical codes.

Character on the screen	Binary value used to process it	Character on the screen	Binary value used to process it
1	0110001	A	1000001
2	0110010	B	1000010
3	0110011	C	1000011
4	0110100	D	1000100
5	0110101	E	1000101

Common character encoding schemes:
ASCII, ISO 646, ISO 8859 parts 1-14m (a wide variety of languages), JIS X 0201-1976 (Japanese), GB 2312-80 (Chinese), KS C 5601-1992 (Korean)

Text Digitization – Character Encoding

- There is exists 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.

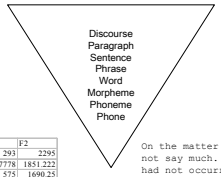
Unicode codes for Pulaar characters

Ɓ	&#385
ɓ	&#595
Ɗ	&#394
ɗ	&#599
ŋ	&#331
Ŋ	&#330
ƴ	&#435
Ƴ	&#436
Ň	&#209
ň	&#241

	labial	alveolar	palatal	velar	glottal
fricative					
voiceless	f	s			h
plosive					
voiceless	p	t		k	
voiced	b	d		g	
affricate					
voiceless			c		
voiced			j		
semi-vowel			y	w	
rolled fricative		r			
lateral		l			
nasal	m	n	ñ(ny)		
nasal compounds	mb	nd	nj	ng,ŋ	
glottalized	ɓ	ɗ	y		'

Text Digitization – Mark-up

Hierarchical and Sequential Nature of Linguistic data



vowel	F1	F2
1	293	2295
2	403.7778	1851.222
3	575	1690.25
4	384.125	2179.125
5	588.5	1804.75
6	701.1111	1238.889
7	632.5	1044.5
8	495	994.2857
9	478.5714	1207.571
10	335.3333	1456
11	592.5	1246.5

On the matter of the shipwreck he did not say much. He only told me that it had not occurred in the Mediterranean, but on the other side of Southern France---in the Bay of Biscay. ``But this is hardly the place to enter on a story of that kind,'' he observed, looking round at the room with a faint smile as attractive as the rest of his rustic but well-bred personality.

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

Relational database	Structured text
Offers powerful analysis tools	Offers good analysis tools
Fails to capture the hierarchical and sequential nature of linguistic data	Promises to capture the hierarchical and sequential nature of linguistic data
Oracle MySQL MS Access	SGML XML TEI

Linguistic Mark-up – example 1

Sentence

John works in the factory.

DTD – Document Type Definition

```
<!DOCTYPE Sentence [
        <!ELEMENT S (DP, VP|PP)>
        <!ELEMENT NP (N)>
        <!ELEMENT DP (D|NP)>
        <!ELEMENT VP (V|PP)>
        <!ELEMENT PP (P|DP)>
        <!ELEMENT N (#PCDATA)>
        <!ELEMENT V (#PCDATA)>
        <!ELEMENT P (#PCDATA)>
        <!ELEMENT D (#PCDATA)>
]>
```

XML mark-up

```
<?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>
```


Linguistic Mark-up – example 2

Conversation

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

DTD – Document Type Definition

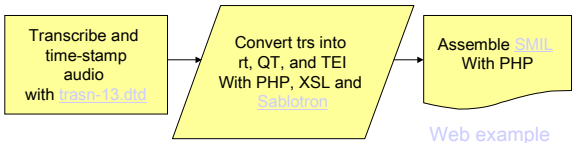
```
<!DOCTYPE list [

        <!ELEMENT conversation (male, female)>
        <!ELEMENT male (#PCDATA|hedge)>
        <!ELEMENT female (#PCDATA|hedge)>
        <!ELEMENT hedge (#PCDATA)>
]>
```

XML mark-up

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



Web example

Thank you!