Representing Pulaar Digitally

Bartek Plichta

Matrix, The Center for Humane Arts, Letters, and Social Sciences Online 310 Auditorium
East Lansing, MI 48823 USA
+1517 355 6187
plichtab@msu.edu

David Robinson

History Department, Michigan State University, East Lansing, MI 48824; telephone 517-353-8898
+1517
robindav@mail.matrix.msu.edu

ABSTRACT

This paper considers the issues of digital representation and preservation of Pulaar language data.

Many African languages are seriously under-represented both on the World Wide Web and in digital libraries. To help preserve these languages, disseminate information to underserved populations, and aid language learning, we need to find ways to digitize and deliver linguistic content on the World Wide Web. In the course of our research, we have developed tools and methodologies that help to represent and preserve African language content (particularly, Pulaar) in a, robust, open-source, electronic environment.

Our methodology is built on (1) Optical Character Recognition (OCR), (2) Unicode, (3) Extensible Markup Language (XML), and (4) Synchronized Multimedia Integration Language (SMIL). We have built a complete text digitization application suite that allows accurate digitization and Unicode encoding of written Pulaar data by means of a custom OCR "training" method. We have also developed a Unicode keyboard mapping application for Pulaar and a SMIL-based methodology to create time-aligned multimedia corpora of Spoken Pulaar.

Keywords

Digitization, linguistics, mark-up, TEI, XML, SMIL, West Africa, Pulaar

INTRODUCTION

Pulaar (also Fulfulde) is the most widely spoken of the West Atlantic languages of Africa. At the beginning of the 21st century, after 200 years of writing, and with roughly 20 years of "mass-market" publishing, a rich written (printed text) and recorded (speech) heritage has been developed in Pulaar, a rich heritage that is in need of preservation. Presently, digital preservation offers many advantages; however, representing Pulaar also presents many technical challenges that make digitizing efforts especially.

Within this context, the National Science Foundation has funded a project at Michigan State University in conjunction with research teams in the Senegal based Institut Fondamental d'Afrique Noire (IFAN) and the West African Research Center (WARC). Working in collaboration with Associates in Research and Education for Development (ARED), a non-profit publisher which specializes in the Pulaar language (also based in Senegal). This initiative aims to help narrow the digital divide. As information technologies transform education and communications around the globe, the digital divide is enlarging the informational and educational gap between those countries with the resources to use Internet technologies and those without. Thanks to this four-year grant, research teams are working to narrow this divide by building an "African Online Digital Library" (AODL). This project is headquartered at Matrix, an Online Service in the Humanities and Social Sciences at MSU, which also serves as the center for other digital library projects, such as Historical Voices, South African National Heritage, and many others.

One of the primary goals of AODL is to digitize Pulaar language data to preserve existing analog materials, as well as both facilitate World Wide Web access and offer new, computer-aided research tools. The materials to be digitized include everything from taped oral interviews, to Pulaar publications, to Pulaar language teaching materials. With access to multiple forms of Pulaar-language material (handwritten historical texts, tapes, publications, etc.) AODL is seeking to combine techniques designed for native speakers, who are learning to read and write in their own languages, with those techniques designed for non-native speakers of Pulaar, who need a full range of materials to improve their oral and aural language skills.

DIGITIZING TEXT

The advantages of digitizing archival materials, such as text, image, audio, and video are well-known and need no further justification here. Most leading digital archives have established best practices and procedures for digitizing printed text and images. Audio and video materials are a little more problematic, but progress is to be expected fairly soon.

The Pulaar language materials constitute a particular subset of these types of archival materials. The range of available materials is vast in both form and content. ADL has access to taped interviews, handwritten text in *ajami*, published

translations of modern laws, original novels, transcriptions of oral traditions, dictionaries, language-teaching materials, literacy teaching manuals, etc. While most archival digitization efforts are motivated primarily by the issues of preservation and access, it is our goal to establish digitization practices that make these materials available, searchable, and useful for teaching and learning, as well as research. In short, our digitization efforts are geared toward converting printed and audio data into structured, yet uncluttered, text and multimedia corpora that can then be accessed and exploited by multiple disciplines for a wide array of uses.

Using OCR and UNICODE for Pulaar texts.

There are a number of challenges related to using OCR with printed Pulaar data. The most important one is that there exist no Pulaar-specific OCR or proofing tools. It is therefore necessary to develop new tools by adapting existing OCR technology to meet the particular character recognition and encoding needs of Pulaar.

We have developed a custom "OCR training" methodology whereby the software "learns" to recognize the particular bitmap shapes of the 12 unique Pulaar characters. Those characters, as well as the entire text, are then encoded in Unicode and stored as an XML document.

Mark up

In addition to character encoding, we have developed a methodology for marking up both the structural and linguistic detail of the original text.

Given the limitations of traditional text-encoding formats, it appears that only a mark-up schema derived from SGML (Standard Generalized Mark-up Language) will be able to offer adequate tools for this step in the digitization of Pulaar texts. The advantages of SGML mark-up are well-known. It ensures data independence and interchange, offers powerful structure-preserving mechanisms, and provides virtually unlimited possibilities for extension. While SGML has been known to be unnecessarily complex and cumbersome, its recent, slimmed-down derivative, XML (Extensible Mark-up Language), however, is quickly becoming a mark-up standard of choice in both business and scholarly communities.

KEYBOARD MAPPING

To date, several Pulaar keyboard lay-outs have been successfully used on a variety of platforms. However, in each case, the keyboard maps to a proprietary, platform-specific, character encoding.

Based on the Tavultesoft Keyman system, we have developed a Unicode-compliant keyboard application that lets the users map the 12 unique Pulaar characters to virtually any key or combination of keys on the standard keyboard. We are giving Pulaar speakers the flexibility and convenience of using the most suitable keyboard layout, while ascertaining proper Unicode encoding.

MULTIMEDIA LANGUAGE CORPUS

The emergence of electronic text has given researchers unconventional, yet powerful research tools. Structured, computer-readable text has become available for sophisticated quantitative analysis. Linguists, historians, literary scholars have produced a considerable body of innovative research, which was made possible by the powerful analysis tools of SMGL parsers and concordancers. However, most of language corpora to date have dealt primarily with text. Given the wealth of oral history materials in Pulaar, a traditional approach to language corpora seems inadequate. Ideally, we should be able to perform the same kind of quantitative analysis and with as much ease both on text and audio. We have achieved that by creating a time-synchronized corpus of a digital audio (or video) file and its transcript. The XML-encoded transcript can be searched and analyzed in traditional ways, yet the parser synchronizes the text with audio, adding an aural dimension to the analysis. Some possible uses of this approach include phonological and sociolinguistic analysis, language learning and teaching, translation, access for people with hearing disabilities, etc. Time-synchronized corpora present us with sophisticated tools that range from speech recognition to customized online search and delivery.

ACKNOWLEDGMENTS

We would like to thank Mark Korbluh, Dean Rehberger, Dennis Boone, Mike Fegan and Scott Pennington. Their research and work have added greatly to this paper.