An Integrated Framework for Archiving, Processing and Developing Learning Materials for an Endangered Aboriginal Language in Taiwan

Meng-Chien Yang
Department of Computer and Communication Engineering
Providence University, Taiwan
mcyang2@pu.edu.tw

D. Victoria Rau
Department of English Language, Literature and Linguistics
Providence University, Taiwan
dhrau@pu.edu.tw

Abstract

Preservation of an endangered language is an important and difficult task. The preservation project should include documentation, archiving and development of shared resources for the endangered language. In addition, the project will consider how to revitalize this endangered language among the younger generation. In this paper, we propose an integrated framework that will connect the three different tasks: language archiving, language processing and creating learning materials. We are using this framework to document one Taiwanese aboriginal language: Yami. The proposed framework should be an effective tool for documenting other endangered languages in Asia.

1 Introduction

The impact of globalization and urbanization has caused many aboriginal languages on our planet to go extinct. This language death process not only reduces the number of native languages but also wipes out the cultural heritage connected with those languages (Xu 2001). Therefore, preservation and archiving of these endangered native languages is vital and critical. Many projects around the world are seeking to preserve these endangered native languages (e.g., Lublinskaya 2002; Psutka 2002).

The attempt to preserve an endangered language includes several steps: documenting and recording the oral and written literature, compiling the grammar and a dictionary of the language, and annotating the documentation related to this language. It is also important to find an effective approach to teach the endangered language to the ethnic group using the language, particularly to members of the younger generation, who often live in urban areas without any connection to their place of origin.

According to a study by Whaley (2003), the factors required to help an endangered language survive include:

1. a well developed preservation and maintenance program for the language;
2. use of information technology in the preserving project;
3. a new world order, especially economic and political shifts;
4. an environment for learning and exploring the language.

Based on the above discussion, it is important that an endangered language preservation and documentation project should be comprehensive and carefully planned. This project needs to take advantage of state-of-the-art technologies and establish an environment for learning.

In order to successfully document and preserve a Batanic language, Yami, we propose an approach of archiving and development of an environment that fosters learning of the language. The Yami language, used by the Yami tribe on Orchid Island, is an oral language in which most of the content is closely connected to the traditional life style and cultural heritage. However, many Yami people have moved to cities in Taiwan and have lost their connection to the Yami society on Orchid Island. The death

of the older generation has hastened the decline of the Yami language. According to Rau's (1995) sociolinguistic survey on Orchid Island, Iraralay is the only community of the six villages on the island where children still use some Yami for daily interaction. Although Yami has been offered as an elective in elementary school since 1998, Yami is gradually being replaced by Mandarin Chinese. Among the junior high school students on the island, 60% either believed Yami would die eventually or were uncertain about the fate of the language.

The approach proposes a comprehensive series of steps to collect and record the Yami language. In addition, the work includes development of a learning method that will be effective with Yami youngsters who live in urban areas. Although the complete work of documentation will take many years, the Yami language is in danger of being lost due to rapid urbanization. Therefore, we have developed a strategy to make language items available in learning materials as soon as they have been collected, taking advantage of information technology and computer networking. Using these technologies we have developed an integrated platform for documenting, processing and learning that will help both Yami youngsters and other students taking Yami as a second language.

The integrated platform is built on a main web server with several supporting servers. The main server is designed as the server for resource management and the supporting servers are designed for different purposes. The purpose of this design is to effectively edit the oral recording of the Yami language and to make the language learning materials. The proposed platform includes three subsystems:

1. a subsystem to manage and edit the digital archiving of the Yami language,
2. a subsystem to handle the workflow of collecting oral recordings of the Yami language,
3. a subsystem to create and manage the Yami language learning materials.

Each subsystem is installed on one or two servers. All these subsystems will be described in detail in Section 3.

Although most ideas in the proposed integrated framework has been used for other language documentation and learning, the proposed framework is an initiative for archiving and teaching an endangered language. The attempt of our study is not only to use technologies to preserve an endangered language but also to develop a well-accepted platform for this language. Hence, people can learn and appreciate this language and its cultural heritage.

The proposed framework is used in an ongoing grant-supported project for archiving and documenting the Yami language (ELDP, MDP0114). The collection of Yami language materials began in 1994. Currently, we are implementing the computer systems and database in this integrated framework. In the later section, we will report on our current progress.

The remainder of this paper is organized as follows. Section 2 is a description of the process of collecting the material for archiving. Section 3 shows the proposed integrated framework and a brief description of related methodologies. Section 4 illustrates the current development of the system, followed by conclusion and future directions in Section 5.

2 Materials to be Documented

In addition to digitally archiving the 20 narratives, reference grammar, trilingual dictionary with 2000 entries (Rau & Dong, 2005), and multimedia pedagogical materials (Rau et al. 2005), we also collaborated with local consultants to document daily conversations, business transactions, festivals, and ceremonies.

The topics were selected based on consultation of previous research on Yami ethnography, and are designed to meet the standards stipulated by the R.O.C. Ministry of Education for developing Austronesian teaching materials in Taiwan. The topics are closely related to those selected for inclusion in four volumes of Yami multimedia teaching materials the second author is currently developing.

3 Integrated Framework

In this section, we will describe our design and the theoretical framework behind the design. The project is divided into four major steps:

(1) field recording: recording the oral sound data of the Yami language,
(2) archiving: editing the sound data and annotating the data using the metadata,
(3) multimedia transformation: analyzing the original data and creating a multimedia Yami dictionary and text description,
(4) e-Learning: creating online Yami language learning materials.

The framework is designed to meet two requirements of our Yami language archiving project:

(1) to build a complete and original archiving database for Yami language including speech of various genres, grammar, vocabulary and cultural artifacts.
(2) to create learning materials in an easy-to-learn environment via internet and computer.

3.1 Field Recording

First of all, the existing records collected by the research team since 1994 will be organized and digitalized, along with new field recordings. In our project, we will develop an oral speech archiving database to store these oral recordings. Each recording will be scanned to find the basic sound characteristics and transferred to digital data. The sound characteristics are used for comparing and tracking these recordings. Following a study by Chen (1996) about tone and stress patterns in Asian languages, we will extract information on intonation and stress from the field recording. This information will later be used to create the learning material. The field recordings are arranged by segments, ranging from words in isolation to "idea units" or "tone units" (Chafe 1979) in continuous speech.

Once a segment of the field recording has been completed, the original data is stored in the computer and two different types of digital data are created. These include MP3 data that will be used for creating the learning materials and the annotated digital data in which the recordings are separated into phrases with Chinese and English translations. All these data are stored in a relational database with the recording date used as the searching key.

The processing of field recordings is considered to be the preparation and preprocessing stage of the Yami language documentation project. The voice database is used to create the archived data and learning materials.

3.2 Archiving

The archiving step begins with editing the voice database and construction of the OLAC metadata for each entity in the voice database. The original sound tracks in the field recording database are edited to improve clarity of the sound by using sampling techniques (Kientzle 1998). The edited sounds are stored as the new sound records in the voice database.

The metadata used for describing Yami language is the OLAC metadata, an extended Dublin Core set with basic elements of language resources. To meet the requirement of the linguistic community, certain new extension elements are put in the OLAC set following DCMI guidelines (DCMI 2000). To build a proper OLAC metadata for the Yami language, we have chosen to adopt the OLAC set proposed by Bird and Simons (Bird et al. 2001, Bird & Simons 2003) for this project. Because Yami is primarily an oral language, we use a subset of this OLAC set. The OLAC elements used in this project are: {Title, Creator, Subject, Subject language, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Rights}. The reason for selecting these elements is to create a common description of the Yami language. Furthermore, after reviewing the field study materials, we can show that the above OLAC subset can meet the basic requirement for describing the Yami language. The rules to apply these OLAC elements to each recording of the Yami language are:

(1) Each OLAC element can be optional and repeatable;
(2) Each OLAC element can describe only one single identification or one single range;
(3) Data format of each OLAC element follows the rules in DCMI (DCMI 2002).

Each OLAC element used in describing the Yami language is given following the OLAC and ELDP guidelines. Suppose there is a Yami language sound track to be described, the OLAC

element set of this sound track is shown as follows:

Title: the Chinese name of the Yami language sound track. A second Title element is used to store English translation.
Creator: the Yami speaker who uttered this speech. A second Creator element is used to store his/her Chinese name.
Subject: the keyword used to classify the content of the Yami language sound track. The keywords and controlled vocabularies are being collected.
Subject language: the Chinese linguistic description of the Yami language. A second element is the corresponding English description.
Description: the usage and the multimedia data related to this Yami language sound track. Some multimedia data are collected using the Multimedia Transformation step described in Section 3.3.
Publisher: the research teams and the sponsoring institutions.
Contributor: the research teams and the person who recorded this sound track.
Date: the date this sound track was recorded and the date the archiving process was completed.
Type: the genre of the content of the Yami language sound track. We are transferring many Yami language linguistic and anthropological terms into DC-type. These DC-type terms will be used as the Type element.
Format: the digital data type of the Yami language sound track.
Identifier: the ELDP identifier for this Yami language sound track. We will follow ELDP guidelines to create identifiers for the archived sound track.
Source: the location of the archiving database and the location for storing the field study draft.
Language: English and Chinese (traditional and simplified characters)
Relation: the related Yami language sound tracks.
Rights: copyright information of this sound track.

In the archiving step we will also consider how to build a database of the controlled vocabularies for the Yami language. We will use three sources for the controlled vocabulary in this project: lexicon, primary text and language description.

The table of OLAC metadata is created in two forms, one XML text table format and one relational table format. The voice database from the first step is edited and connected to the metadata table.

Another goal of this step is to build a Yami language online phrase dictionary. The OLAC metadata are used for parsing and editing with the voice database to create a Yami language online phrase dictionary. We will develop an auto dictionary-generating program that can process the OLAC metadata and find suitable terms. In addition, we use the grammar and course materials of Yami language multimedia courseware created by Rau et al. (2005) to build our on-line multimedia Yami language phrase dictionary.

When the metadata of a set of the Yami language sound tracks are completed, the results will be published online on our web site. This year, our focus is aligning the OLAC metadata of the Yami language sound tracks with the multimedia courseware by Rau et al. (2005). Later, we will try to use ontology to determine rules for creating metadata automatically and to develop an automatic metadata generator for the Yami language.

3.3 Multimedia Transformation

The Yami language is basically a spoken language, although an orthography is being developed and standardized as texts are collected. To preserve the Yami language, we will use an image database to annotate the language. In addition, each word in Yami is annotated with its orthography stored in a sound database. The purpose of this transformation is to build an image for each Yami word. Therefore, the meaning of the word can be related directly to a picture. The reasons why we have chosen to use this approach to annotate the Yami language are as follows:

(1) The Yami language, like all other languages, has culture-specific words and expressions, of which pictures are direct representations.
(2) The annotated pictures help learners understand the traditional lifestyle on Orchid Island and give them more incentive to learn the language.
(3) The pictures include many Yami cultural artifacts. The annotated pictures can thus preserve descriptions of their cultural heritage.

The steps for multimedia transformation of the Yami language are as follows:

(1) Collect suitable images for building the annotated image database. We will consult many other research teams to borrow Yami images and video recordings.
(2) Design criteria to choose the images. We will select appropriate images and develop possible connections between Yami expressions and a set of pictures.
(3) Build a special annotated database and use the Yami language to annotate the image data. The annotated algorithms are based on the fuzzy logic style (Kecman 2001) or the Coherent Language model (Jin 2004).
(4) Build a corresponding mapping relation between a Yami expression and a set of annotated images. The mapping relations are a set of contexts and symbolic tables similar to a set of induction rules.
(5) Build a sound connection between each Yami word and its phonetic symbols by using the fuzzy logic learning algorithm.

The results of multimedia transformation can be used as a foundation for creating online learning material. The results are stored in a relational multimedia database as well as the XML pages.

3.4 e-Learning

The final task of our project is to find an effective way to teach the Yami language to urban Yami youngsters and other learners of Yami as a second language. To build an open and self-learning environment, the computer-based learning or the webs for learning is our choice. There have been various discussions about how to use information technologies and the web to learn a different language. Gerbault (2002) showed that it is viable to set up a proper multimedia environment for leaning a language without a teacher's participation. Fujii et al. (2000) demonstrated a project using the Internet as a tool for the teacher to post course materials and create an online learning environment. In addition, Lamb (2005) suggested rethinking pedagogical models for e-learning from the what, the why and the how. e-Learning consists of self-access, reference sources, discussion forum, and virtual learning classrooms. The main motives for introducing e-learning include improving student multimedia learning experience, enhancing learner autonomy and widening participation. Finally, e-learning can be controlled primarily by tutors or students, depending on objectives, contents, learning tasks, length/time/place of study, or choice of assessment activities.

As mentioned in a study by Leung (2003), the computer-based learning environment is very important as a way to help students learn effectively. In order to provide an effective learning environment, Leung (2003) suggested that four contextual issues should be considered in design and implementation of computer-based learning. These issues are topic selection, authenticity, complexity, and multiple perspectives. The design of the web-based computer-assisted learning program for the Yami language takes these four issues into consideration. We outline our design as follows.

The learning environment in this project is a virtual classroom without teacher participation. Students can select the Yami language learning materials prepared by the second author. If a student asks for clues or explanation of a specific Yami word or expression, a suitable image or video clip is retrieved from the multimedia database. If a student is not familiar with a specific Yami sound, a similar phonetic symbol is provided to him/her. The learning materials are arranged in three different settings, scenario setting, easy-to-difficult condition setting and learner's choice setting. The scenario setting uses related scenes in Yami society such as the flying fish festival as a main theme of the learning materials. The easy-to-difficult condition

setting allows the learner to select different levels of the Yami language materials. The levels are based on word frequencies and complexity of grammar. The learner can arrange his/her learning materials in the learner's chosen setting. The learning system will give detailed guidelines to explain how to choose the learning materials. If a student wants to learn the Yami language, he/she can choose different learning materials based on his/her interest. The learning materials are designed as theme units with exercises and rubrics for self-assessment. The design of these Yami language exercises is based on a study about the reactions of students to using a web-based system for learning Chinese in Taiwan (Yang, 2001).

We use the annotated image database as a tool to help the learners understand the meaning of Yami words or expressions. To make the pictorial explanation more understandable, an animation clip combined with several images is created to explain them.

A study by Aist (2002) showed that different designs of the oral-reading interactions can help students understand the language more. The learning system will provide several reading modes for students to listen and practice. These modes include: to read the entire sentence without interruption, to read the entire sentence by isolating each word, to read a word slowly syllable-by-syllable, recue the whole sentence and recue the selected words.

The interface of the proposed learning environment is built on a web server with a dynamic web page. To establish a more efficient learning environment, all the learning materials are edited into reusable learning objects. The user interface is developed as an adaptive style following Mich et al.'s (2004) PARLING system.

The proposed framework is illustrated in Figure 1.

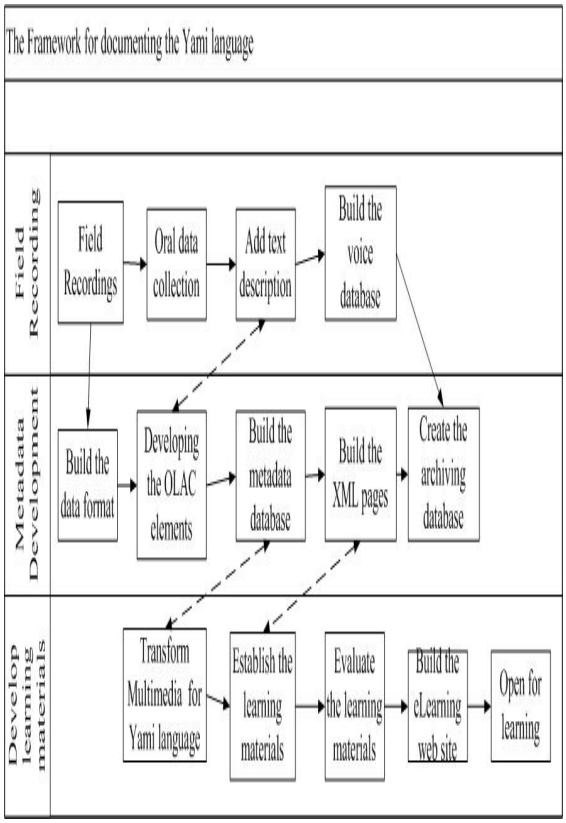

Figure 1: The integrated framework for the Yami Language preservation project

4 Implementation of the proposed Framework

We implement the proposed framework as a hybrid system with many different processes including:

(1) Data collection and formulation: to collect the original Yami language data and to build the metadata and the table for digital archiving.

(2) System design and analysis: to design and develop suitable computer systems and servers to accommodate the proposed framework.

(3) Research and construction of proposed framework: to develop each subsystem or database shown in Figure 1, such as the OLAC metadata database, the annotated image database and the Yami language learning materials.

(4) Assessment and evaluation: to test the effectiveness of the proposed learning ma-

terials and to evaluate whether the project goals were accomplished.

Currently, we are collecting the Yami language materials and building the system server for the proposed framework. We will use a SQL server as the main sever to manage the workflow and the documentation logs. A PHP web server with mySQL server is used as a server for multimedia transformation. Another SQL server is used as the archiving server. The system diagram of the proposed framework is shown in Figure 2.

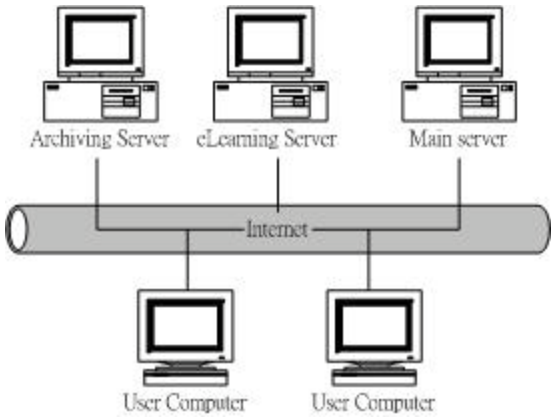

Figure 2 Diagram for the proposed framework

5 Conclusion and future studies

This paper describes an integrated framework for archiving and processing the Yami language. In addition, the framework includes the process for developing online learning materials for the endangered language. We use this framework for the Yami language preservation project. The project is continuously developing. We hope that this project can serve as a model for other endangered language preservation projects in Asia.

References

Aist, G. (2002) Helping Children Learn Vocabulary during Computer-Assisted Oral Reading, Educational Technology & Society 5(2).

Bird, S. Simons, G. Huang, C.-R. (2001) The Open Language Archives Community and Asian Language Resources, NLPRS 2001, pp. 31-38.

Bird, S. & Simons, G. (2003) Extending Dublin Core Metadata to support Description and Discovery of Language Resources, Computers and Humanities, No. 37, pp. 378-388.

Chafe, W. L. (1979) The flow of thought and the flow of language. In T. Givón (ed.), Syntax and Semantics 12: Discourse and Syntax. New York: Academic Press, pp. 159-181.

Chen, S. & Fu, M. (1996) Computer Assisted Language Learning in Teacher Education: Training of Tones and Stress Patterns in Asian Languages, IEEE International Conference on Multimedia Engineering Education, pp. 435-443.

DCMI (2000), Dublin Core Qualifiers,. [http://dublinecore.org/documents/2000/07/11/dememsqualifiers/]

DCMI (2002), DCMI Elements and Element Refinements – a current list.. [http://dublinecore.org/usage/terms/dc/current-elements/]

Fujii, S. Iwata, J. hattori, M., Iijima, M. & Mizuno, T. (2000) "Web-Call": a language learning support system using internet, Seventh International Conference on Parallel and Distributed systems, pp. 326-331.

Gerbault, J. (2002) Information technology and foreign language learning: what happens when no teacher is around?, International Conference on Computers in Education, pp. 394-398.

Jin, R. Chai, J. and Si, L. (2004) Effective Automatic Image Annotation via a coherent language model and active learning, MM 2004, pp. 892-899.

Kecman, V. (2001) Learning and soft computing: support vector machines, neural networks, and fuzzy logic model, MIT press.

Kientzle, T. (1998) A programmer's Guide to Sound, Addison-Wesley.

Lamb, T. (2005) Rethinking predagogical models for e-learning. Paper presented at AILA 2005, the 14th World Congress of Applied Linguistics. July 24-29, Madison, Wisconson.

Leung, A.C.K. (2003), Contextual Issues in the Construction of Computer-Based Learning Programs, J. Computer Assisted Learning, Vol. 19, 2003, pp. 501-516.

Lublinskaya, M. & Sherstinova, T. (2002) Audio Collections of Endangered Arctic Languages in the Russian Federation, TSD 2002, LNAI 2448, pp. 347-353.

Mich, O. Betta, E. & Giuliani, D. PARLING: e-Literature for Supporting Children Learning English as a Second Language, IUI 2004, pp. 283-285.

Psutka, J., et al. (2002) Automatic Transcription of Czech Language Oral History in the MALACH project: Resources and Initial Experiments, TSD 2002, LNAI 2448, pp. 253-260.

Rau, D. V., Dong, M.-N., Lin, M-Y, Chang, H.-H., & Hsu, Y-C, (2005) Multimedia Materials of Yami Language, Technical Report, Department of English Language, Literature and Linguistics, Providence University.

Rau, D. V. & Dong, M.-N. (2005). Yami Texts with Reference Grammar and Vocabulary, Language and Linguistics. A-10.

Rau, D. V. (1995) Yami Vitality. NSC report (NSC84-2411-H-126-001), presented at the Symposium on Language Use and Ethnic Identity, Institute of Ethnology, Academia Sinica (1995/5/16).

Whaley, L. (2003) The future of native languages, FUTURES 35, pp. 961-973.

Xu, S. X. (2001) Study on Language Endangerment. Beijing: Central Ethnic University.

Yang, S. C. (2001) Integrating computer-mediated tools into the language curriculum, J. Computer Assisted Learning, Vol. 17, 2001, pp. 85-93.