**An Accessible Research Database for the Endangered Jaqaru and Kawki Languages**

M.J. Hardman (P.I.), Elizabeth Lowe McCoy(Co-P.I.)
Howard Beck (Co-P.I.) with Sue Legg, Techical Liaison and Evaluation Consultant
The University of Florida, Center for Latin American Studies

**Overview**

This project will document the endangered Jaqaru and Kawki languages, which are members of the Jaqi family of languages (Jaqaru, Kawki and Aymara). Jaqaru is spoken in the Andes Mountains of Peru by a few thousand people resident in Tupe, Yauyos, Lima, Peru and in the cities of Lima, Huancayo, Chincha and Cañete. Diaspora Jaqaru speakers are located in small numbers throughout Latin America, the United States and Europe. The District of Tupe consists of three interrelated communities: Txupi (Tupe), Aysha (Ayza) and Qullqa (Colca). The population of the district is very low, having declined in the last twenty years due to the migration of young people to the cities and the increase in terrorism. The highland plaza is 9,000 feet above sea level and access to the area is by foot and pack animal. Kawki is a dying language spoken by only a few people around Cachuy, Chavin, and Lima, Peru. Access to high quality research materials for Jaqaru and Kawki is viewed as a priority by the people of Tupe. In July 2007 a town meeting was held with the Regional government and the Peruvian Ministry of Education to protest the scarcity and poor quality of educational materials in the languages. The meeting resulted in a resolution of support to improve bilingual education. The earthquake of August 2007 destroyed Tupe and caused damage to roads and adjacent villages, placing the languages in further danger of disappearance. This project will build on the life work of Principal Investigator, M.J. Hardman, who first began her work on the Jaqaru, Kawki and Aymara languages in 1958. Her first Jaqaru Grammar was published in 1966 by Mouton; the Spanish version was published in 1983 by the Instituto de Estudios Peruanos; the updated Hardman Jaqaru grammar was published in 2000 by Lincom Europa. She is the author of extensive publications on the Jaqi languages, including Aymara: Compendio de Estructura Fonológica y Gramatical (ILCA, 1988). The project will build on Dr. Hardman's past fieldwork and well as new fieldwork in Peru, to transform a corpus of 50 notebooks of texts, corresponding audiotapes, and related linguistic data into an accessible linguistic research database of the Jaqaru and Kawki languages. A unique feature of the database will be a dictionary that will be produced with interactive collaboration from the Jaqaru and Kawki speaking community, enabled by a Wiki function. The work responds to the call for interoperability of tools and common data identification standards made by Gary Simons at the Summer 2007 TILR workshop at Stanford University. It builds on existing computational and linguistic work, and conforms to the recommendations and standards proposed by the E-MELD consortium. The Jaqaru project continues the work currently being concluded on a linguistic database developed with a grant (M.J. Hardman, P.I.) from the U.S. Department of Education International Education and Graduate Programs (Title VI) titled: "The Aymara E-Learning Project: Using the Internet to Preserve and Promote an Indigenous Language." database envisioned for this project will have many of the features of the Aymara database, however unlike the Aymara project, which has instructional objectives, the Jaqaru-Kawki database will be designed for the use of researchers and native speakers to document and preserve these endangered languages. It will be multilingual (Jaqaru, Kawki, English and Spanish). It will contain dictionaries of the two languages, audio, digital photographs, and the grammatical and linguistic elements of the languages. The work to index and digitize the original materials, to develop the database and to translate the linguistic work into English and Spanish will be conducted primarily at the University of Florida, in Gainesville, Florida. A coordinator in Peru will be responsible for the collection of additional materials, the transcription of the audio tapes, as well as the deployment and training of local counterparts in the project implementation, the development of the interactive dictionary and the use of the database. The broader impact of this project will be to preserve and make available in four languages the grammar and dictionaries of two endangered Andean languages for linguistic research and for the use of heritage speakers in a strategically important world region. The intellectual merit of the project resides in the multifunctional nature of the database, which has features that are distinct from the SIL LinguaLinks project and the FLEx project, or work done by Charniak and Collins. The project will contribute to the multi-university community work on the standards to ensure the robustness, accessibility and interoperability of electronic archives for endangered languages. The Jaqaru-Kawki resources for the project will be catalogued on appropriate metadata forms and deposited in the Archive of Indigenous

Languages of Latin America at Texas, Austin (AILLA) and the accessible linguistic database will be mapped to the University of Florida Digital Collections with appropriate metadata and ontology standards for digital archiving and will become part of a forthcoming Aymara Collection within a larger South American collections library at the University of Florida.

The project description will be organized as follows. After introducing the project investigators and key personnel in section 1, we present some background information on Jaqaru and Kawki is presented. Section 2.1 describes the ethnographic setting; 2.2 discusses the endangered status of the languages; 2.3 surveys existing documentation for the Jaqi and Kawki languages. Section 3 describes the linguistic resources that will be digitized and analyzed for the primary product of this project: the accessible linguistic database for the Jaqaru and kawki languages. Section 4 describes the archiving plan (how the material maps to AILLA and to the University of Florida Digital Collections) and samples of the metadata. Section 5 describes the accessible linguistic database, tools and access standards and issues and compares and contrasts the proposed work to work done elsewhere. Section 6 discusses the Web Service Architecture, User Interface and Interoperability. The responsibilities of the project collaborators and the project timeline are specified in section 7. Section 8 provides a summary of broader impact, indigenous use and intellectual merit of the project. Section 9 briefly outlines planned future projects, including a plan for testing in other contexts.

**1. The Project Investigators and other key personnel**

This project is an interdisciplinary, international collaboration that brings together several areas of research interest and expertise at the University of Florida and with local institutions in Peru. It also draws upon the life work of the leading scholar on Jaqi languages working in the field of anthropological linguistics.

**M.J. Hardman (P.I.)** is a professor of Linguistic Anthropology at the University of Florida and is an affiliate faculty member of the Center for Latin American Studies. She is one of the pioneers of research with the Jaqi languages from an anthropological linguistic viewpoint, a lifelong work that she began in the 1950's as a field researcher in Peru. She has extensive publications on Aymara, Jaqaru and Kawki, most recently her Jaqaru Grammar (Lincom Europa, 2000). She is credited with being the first outside researcher to discover the existence of Kawki and the relationship with Jaqaru and doing the first linguistic research on them, in collaboration with her husband, Dr. Dimas Bautista Iturrizaga, D.V.M., who in the 1940's sought a way to write his language (Bautista, 2000). In 1965, with Dr. Julia Elena Fortún, she founded the Instituto Nacional de Estudios Lingüísticos (INEL) in La Paz, Bolivia, as a dependency of the Directorate of Anthropology of the Bolivian Ministry of Education and Culture. As a Fulbright-Hays Professor, she taught courses to prepare human resources in linguistics in Bolivia. She was the Director of the Aymara Language Program at the University of Florida from 1969-1990, which was funded by a grant from the U.S. Department of Education Title VI program. Dr. Hardman continues to do research on the Jaqi languages and spends several months each year in the region. Dr. Hardman is P.I. of the U.S. Department of Education Title VI Grant, "The Aymara E-Learning Project (2004-2007)." Her current research also involves language and gender and the patterning of worldview in language.

**Elizabeth Lowe McCoy (Co-P.I.)** is Associate Director for Program Development and Distance Learning at the U.F. Center for Latin American Studies and founding director of the Partnership in Global Learning, an international distance learning initiative funded originally by the Lucent Foundation and Bell Labs. She is establishing a Research and Training Program and Academic Specialization in Indigenous Languages and Language Policies of the Americas within the Center. Dr. Lowe is the Co-P.I. of the U.S. Department of Education Grant "The Aymara E-Learning Project," and is responsible for overall fiscal and administrative management of the project, serving as leader of the interdisciplinary team. Former Executive Director of the U.F. International Center (1991-1998), and founder of the U.F. certificate program in translation studies, she has been a project manager for large and complex international programs and is an expert on the use of technology for foreign language and translation instruction. Her publications are in the field of Latin American literature, language and culture.

**Howard Beck (co-P.I.)** is a professor in the Agricultural and Biological Engineering department at University of Florida, where he has been on the faculty since 1990. He has been working at the UF College of Agriculture since 1977 and has involved in a broad range of information technology-related projects. He completed his Ph.D. in 1990 in the UF Computer and Information Sciences department in 1990, within the Database Systems Research Center. Dr. Beck's research combines artificial intelligence techniques with database management in order to create knowledge management systems that can

organize vast collections of knowledge within a particular domain. He pioneered the use of object database management systems to develop digital libraries in agriculture and natural resources. More recently the work has evolved to incorporate formal ontologies as the core of a database management system, resulting in his development of the Lyra ontology management system. A unique feature of Lyra is its support for natural language processing, and current work includes development of a database on the Aymara language using Lyra. Other successfully database systems developed and deployed by Dr. Beck include EDIS (Extension Digitial Information System, which is a digital library of Extension publications for the College of Agriculture, FAWN (Florida Automated Weather Network), DISC (Decision Information Systems for Citrus), SPDN (Southern Plant Diagnostics Network), and CBC (Crop Bioscurity Curriculum). These applications incorporate a variety of techniques including expert systems, computer simulation, eLearning, and information retrieval**.**

**Sue Legg (Technical Liaison and Evaluation Consultant)** was the Associate Director of the Office of Instructional Resources (OIR), now the Office of Academic Technology at the University of Florida from 1980-1995 and Director from 1995-2001. In 2002, she also became Director of the Center for Instructional and Research Computing Activities (CIRCA) and U.F. Coordinator for Distance Education. In 2002, Dr. Legg retired from her administrative responsibilities and joined the Center for Latin American Studies, where she was Director of the Partnership in Global Learning (PGL) until 2004, when she assumed the role of Research Director. PGL is an international university and K-12 consortium whose mission is to develop and implement online learning technology and training. Also in 2004, she assumed responsibility for technical coordination of database development for the “Aymara e-learning project." Her area of specialization is in research methods, measurement, and evaluation. She was principal investigator for a number of contracts and grants from the Florida Department of Education and national foundations. Her publications are in the area of assessment and evaluation. She also continues to serve as the assessment consultant for the Florida Bar Board of Legal Certification and Education.

**Yolanda Nieves Payano Iturrizaga (Peru Coordinator)** is a linguist who has been appointed by the Regional Government of Lima to supervise bilingual education in Tupe, Peru at the Instituto Superior Pedagógico Yauyos in the Department of Lima. Ms. Payano studied linguistics and languages at the Universidad Mayor de San Andrés (USMA) in Bolivia and studied English in the United States for one year, while assisting with a Field Methods course at the University of Florida. **Dr. Dimas Bautista Iturrizaga, D.V.M (Jaqaru Language Resource Consultant**) was mayor of Tupe, Province of Yauyos, Peru from 1984-1986 and has been involved with efforts to preserve the Jaqaru language since the 1950’s. He met the famous linguist Kenneth Pike when he approached the Peruvian Ministry of Education, in search of someone who could help him write his language. Trained as a doctor of veterinary medicine, Dr. Bautista began to publish on the Jaqaru and Aymara languages as a result of his work in Puno with the alpaca, for which he discovered and developed a vaccine that made possible herds of alpacas and thus today’s alpaca wool industry. While inoculating the animals, he noticed someone had used an Aymara term to refer to pasturing (*awata* in Jaqaru, *awatina* in Aymara) which he understood and replied in Jaqaru . This resulted in his publication “Tupe y el Jaqaro o Kawke” (1959). Currently in preparation is his book on the history of Tupe, relating what he was told by his elders ( *Mark Qillqa: Tupe, Reseña histórico-cultural del pueblo de Marka año 750-2007*,” Instituto de Estudios Peruanos, 2007). Dr. Bautista has been a principal consultant to Dr. Hardman throughout the years and frequently the source of her publications.

## 2. The Languages

### 2.1 The Ethnographic Setting

The Yauyos valley is a very steep, rugged valley that has long impeded the intrusions of outsiders because it is so difficult to negotiate. It is also the most linguistically varied area of the Andes (cf., Alfredo Torero). It is thought that the Incas never entered the valley, and even the Spanish were slow to do so. Eventually the Spanish penetrated the area, and in 1761 the Virrey Amat issued an establishment document for Tupe. The Jaqi people are primarily farmers, herders, weavers, musicians, dancers and marketers. The Jaqi civilization (400-1000 C.E.) established both a complex system of irrigation by canals and a complex system of mercantile exchange, involving the construction of roads. Some of the market relationships still function today and the road system is clearly visible in ruins. One cultural invention was the archipelago land holding system where an individual would attempt to hold plots of land, however small, in as many different ecological niches as possible. The land is steep so that very different crops can be cultivated not far apart by going up or down the mountain. Cultivating crops at

different elevations ensures diversity in what they produce and protects against weather and pests. The archipelago system remains the desired form of land holding. This system, together with extensive marketing, today specializing in cheese, means a great deal of movement as people go about ordinary tasks. Travel to Watxuqu (Catahuasi), some 15 miles distant and 5000 feet lower, and returning the same day, is routine. School has always been attractive to Jaqi people and there are today many professionals, including teachers, doctors, veterinarians, engineers and others who have come from Tupe. The modern city has become one more ecological niche within the Jaqi archipelago system of land holding, with foodstuffs and weavings coming from Tupe and manufactured goods, especially electronics such as radios and batteries returning from the city.

Jaqaru is a member of the Jaqi family of languages which also includes Kawki and Aymara. Jaqaru is spoken by a few thousand people around Tupe, Yauyos, Lima, Peru and in the cities of Lima, Huyancayo, Chincha and Cañete, with a few diaspora speakers scattered throughout Latin America and in the United States and Europe. The district of Tupe consists of three interrelated communities; Txupi (Tupe), Aysha (Ayza) and Qullqa (Colca). The current population in the District is very low, consisting of only around 600 people. It has dropped from 5000 in the past 100 years, due to the outward migration of young people and the advent of terrorism. Kawki is a dying language spoken by only a very few people in and around Cachuy, Yauyos, Lima, Peru. Aymara is spoken by two to three million people, the first language of a third of the population of Bolivia and the major native language in Southern Peru and northern Chile.

**2.2 Endangered Status**

When Hardman first began her study of the Jaqaru language in 1958 there were still monolingual speakers of Jaqaru and quite a few people who had learned Spanish only late in life, whose knowledge of Spanish was limited. Today all the young people of Tupe are bilingual and a number of children now do not speak (but still understand) Jaqaru. There are no living monolingual speakers; even the oldest living bilinguals are fully fluent in both languages. Jaqaru, therefore, is an endangered language. In the early 1900's, the elders of Tupe sought to establish a high quality elementary school. Over a period of about 30 years, they succeeded. They brought books for the school from Boston, materials for a physics and chemistry lab, a globe and recruited teachers from the outside. Tupe became the educational center of all of south Yauyos; students came as boarders to study in Tupe. In an era when the eradication of native languages was the goal, children in the school were punished for speaking Jaqaru. Within two generations, by the middle of the century, virtually all speakers were at least bilingual, though there remained some for whom Jaqaru was indeed dominant, who spoke a rudimentary Spanish. That is no longer the case. During the 1980's, Tupe suffered from the intrusion of Shining Path terrorists. A massacre led to migration, which accelerated the shift from Jaqaru to Spanish. In the late 1990's, when Shining Path was no longer a threat, children no longer spoke Jaqaru as a primary language. Those who were children at the beginning of the terrorist assault, including some current school teachers, now wish to recover the language, even though their own control of the language is sometimes shaky. Some people, including the high school students in Tupe itself, are hoping that bilingual education will preserve the language as part of their endangered cultural heritage. Recent regional government initiatives seek to strengthen bilingual education. In July 2007 a town meeting was held with the Regional Government and the Peruvian Ministry of Education to protest the scarcity and poor quality of educational materials in the language. The meeting resulted in a resolution of support ("Acta") to improve bilingual education and encourage the language training of heritage speakers. The endangered status is even more grave due to the August 2007 earthquake that destroyed Tupe and caused damage to roads and adjacent villages.

The extinction of the Kawki language was provoked by the discovery of a miraculous image, el Señor de Cachuy. The religious icon attracted hordes of pilgrims, and unleashed a number of changes that threaten their language and culture In May of every year, the town of 200 swells to over 20,000 pilgrims, including evangelical missionaries whose actions hasten the shift from Kawki to Spanish. By the middle of the last century, the language was spoken only by the elderly, with the exception of one individual who had been raised by his grandmother, Valerio Luciani Ascencio, on whose shoulders the future of the language now rests. He is almost 60, but he has been, for the last 30 years, attempted to teach Kawki to school children, laboring in the absence of official recognition and financial support. .

**2.3 Previous Documentation**

The first person to be interested in a study of the language was Dr. Dimas Bautista Iturrizaga, D.V.M., who in the 1940's sought a way to write his language. The story is told in his forthcoming book on the history of Tupe (Bautista, 2007). In recent years two young women have begun to work with the language, the linguist Yolanda

Nieves Payano Iturrizaga (Payano, 1988) and Neli Belleza Castro (Belleza, 1995). There is also a periodical that publishes some material in Jaqaru (Ramirez, 1989-to present). Also, Dr. Jose Matos Mar collected some materials in the 1950's. The meager results of early studies, are referenced in Hardman's 1966 and 1983 Jaqaru grammars. She has written numerous articles on Jaqaru and Kawki, referenced in the bibliography. Over the years Hardman has continued to work with Jaqaru as well as with the sister languages, Aymara and Kawki. Included in the Kawki documentation are bilingual Kawki-Spanish primers (Hardman, 1982,1983), Kawki texts for school use (Hardman, 1981, 1980), a Kawki-Castellano-English trilingual dictionary (Hardman, 1986) and an alphabet of the Jaqaru, Kawki and Aymara languages (Hardman, 1991).

Hardman's recent Jaqaru Grammar (2000) is the most systematic attempt to document the language. The grammar illustrates some of the language's unique features and provides the organizational framework for the proposed accessible research database, to be adapted from the current Aymara database project (U.S. Department of Education Title VI, 2004-2007). The phonemic system distinguishes 36 consonants but only 3 vowels. Vowel dropping is significant, complex and pervasive, marking case and phrase structure as well as style. The language makes extensive use of morphology, with all verbs carrying several suffixes. Syntax is morphologically marked; verbal person suffixes mark simultaneously object/subject; data source is marked at all levels of grammar. Within the nominal system inclusive/exclusive and humanness are marked.

## 3. Source Data

The primary source of the data for the project will be the Jaqaru field materials collected by Hardman during her fifty years of research in the region, along with new text, and audio and photographic materials that will be collected and digitized by the coordinator in Peru. The materials include:

1. 50 hand written field notebooks containing transcriptions of existing audiotapes, notes about the content of the tapes, grammatical analysis and illustrative hand drawings.
2. corresponding audiotapes of native speakers of Jaqaru and Kawki. Many of these recordings are of now deceased village elders.
3. Over 1000 slides taken between 1959-1975 of Jaqaru and Kawki speakers, villages, farms, market scenes, homes, and schools.
4. Interactive dictionaries of the two languages, allowing for user collaboration with a Wiki function.

Figure 1 shows a screen shot of one of the field notebooks.

## 4. The Archiving Plan

### 4.1 Archive of Indigenous Languages of Latin America at the University of Texas, Austin (AILLA)

**A sample of the metadata form is in the Appendix**

### 4.2 University of Florida Digital Collections

Source Data will be digitized by the University of Florida's Digital Library Center. Text page images will be scanned at 300 dpi, while graphical images will be scanned at 600 dpi. Resulting images will be mastered in the TIFF v.6 uncompressed format, with derivatives for Internet use in JPEG and *zoomable* JPEG 2000 formats. Searchable text, generated by double-key method to 100% accuracy, will be procured from the Center's vendor. Audiotapes will be sampled at 24-bit 96.0 kHz and retained as WAV uncompressed format, with derivatives for Internet use in down-sampled 16-bit 44.1 kHz MP3 format. The high-resolution WAV will also be available for download. Internet accessible versions of all digitized source data will be available in the University of Florida Digital Collections' *Jaqi Collection* (http://www.uflib.ufl.edu/ufdc/?s=aymara). The Digital Collections (UFDC) utilize both Greenstone digital library and Aware JPEG 2000 data stores, as well as a locally designed audiovisual store. All files and metadata will be fully searchable and available freely through the Internet. All digitized resources will be supplied to the Project PIs for use in the Language Editor and other project applications.

The Center digitally archives *all* content – including the projects executable files and data-tables, as well as digitized resources – with the Florida Digital Archive (FDA). FDA (http://www.fcla.edu/digitalArchive/) is a service of the Florida Center for Library Automation (FCLA), based on the framework for Open Archival Information Systems (OAIS). For more information on FDA, see http://www.fcla.edu/digitalArchive/pdfs/IJDL_article.pdf. Procedures for data validation, format migration and redundancy are key; archived resources are validated against DTDs and checksums and

retained by FCLA's FDA, backed up on the Northwest Regional Computing Center in Tallahassee, Florida, and copied to Data Center at Stanford University. FDA is funded by the State of Florida through continuing allocations to FCLA. The service is free to the University of Florida.

## 5. The Accessible Linguistic Database

**5.1 An Ontology Management System for Language Archiving.** Central to the goal of building a collaborative environment enabling team members to work together is an environment consisting of on-line authoring tools that team members can use to create and edit data objects, and a core database facility for storing these objects.

We are utilizing an ontology management system (OMS) to provide this environment, including facilities for representing and archiving knowledge associated with the language and culture of Aymara. The OMS provides a framework for integrating everything from raw data elements (sound recordings, transcripts, images, video), to more abstract linguistic elements (morphemes, words, phrases, phrase patterns (grammars), and dialogues). Logical relationships between any two elements are expressed using ontology property relationships (i.e. how the data elements are related). While we have built a database for Aymara, the system can be applied to any language.

The Lyra Ontology Management System. While ontologies are traditionally used to represent semantics of words, we expand the use of ontologies to the level of full-blown database management system, that is a database system that uses a formal ontology language (such as OWL, the Web Ontology Language []) as the data definition language, rather than tables as used in conventional relational databases, or general purpose persistent objects as used in object database management systems. The ontology language provides a more natural and logical way of describing concepts and relationships than previous database languages. Concepts are arranged taxonomically, through part/whole relationships, and through other relationships such as needed to capture the connection, for example, between a phrase and the original transcripts. Every abstract element is thus not only tied directly to original data, but such abstractions can be generated based on generalizations obtained over raw and intermediary structures. Thus the OMS provides a complete environment for modeling and storing all data objects and relationships needed to build a linguistic database, including raw field data and cultural artifacts.

We have constructed an ontology management system called Lyra, that provides a number of basic data management services. Lyra uses an ontology language for data modeling and representation. Lyra is similar to OWL, and we have developed mappings between Lyra and OWL. Lyra also supports physical storage management optimized to storage and retrieval of large numbers of objects. Lyra includes a number of authoring tools for viewing, designing, and editing objects. The authoring tools provide a data visualization environment, and tools can be customized to a particular application in order to provide experts with tools that they can be most comfortable with. We are currently incorporating reasoning capabilities for Lyra that support traditional ontology reasoning (subsumption and classification), as well as parsers and inductive learning algorithms (for use in grammar induction). The ontology reasoners will thus provide a toolkit for searching the database, and discovering new categories and relationships by analysis of new and existing data objects. Finally, the OMS publishes data using web services and XML interfaces provide interoperability between internal Lyra data object structures sharable and sharable XML data formats.

There are several advantages of using an OMS rather than a traditional relational database, or no database at all (XML file system). First would be the more natural way in which an OMS models linguistic (and other) data. The ontology language provides an way to model taxonomic relationships, and other relationships among objects, with direct pointers among abstract generalizations and concrete data (such as field observations). The object structure can model complexities of language elements. Though this can be done with relational databases, it is much more difficult as these complex structures must then go through an additional mapping to normalized relational tables in which the object structure and relationships is no longer explicit. This can also lead to less efficient data retrieval []. Another advantage would be the focus on reasoning, and in particular the computational complexity of reasoning processes that operate on the data. It is known that the details of the data modeling language have a direct impact

on computational complexity of reasoners. The ontology language can be designed to capture just enough detail, but not too much to lead to computational complexity problems (for example, OWL-DL, the description logic version of OWL, is designed to balance this tradeoff). Finally, an OMS offers a variety of services necessary to support security, data integrity, transaction management, and query processing. XML files are ideal for exchange of data among different systems, but the database is needed to provide operational functionality. One could also write custom tools that read and write XML files directly, but they would need to duplicate the functions provided by an OMS.

Lyra was not developed solely for the Aymara project, but has evolved over nearly two decades of research and development on combining databases with knowledge representation techniques from artificial intelligence to build a system capable of organizing and storing knowledge in broad domains. Lyra has been used successfully to model applications in agriculture [], environmental sciences [], mathematics [], decision support systems [], and linguistics []. The ability to support natural language in an integral fashion is one of the core advantages of this approach.

**5.2 Tools.** Lyra includes several authoring tools that provide data visualization environments for browsing, creating, and modifying data objects. The tools are designed for use directly by subject matter experts, and are used by members of our team to construct the database. Such tools can be developed by anyone to access the database using web services (See next section). The tools are graphic (providing visualization) and Web-based (run inside Web browsers using plugins), thus they form the basis for a collaborative environment for creating data objects stored in and retrieved from a common database. The tools can also be customized to meet the needs of experts working in a particular domain.

5.2.1 Language Editor

Figure (3) is a view of the LanguageEditor, a tool for browsing and editing the language database. LanguageEditor is a Java applet that runs in any web browser that has the Java plugin. It connects remotely to access the database over the Internet. Shown here is a complete phrase analysis including a morphological analysis. The gloss for the dictionary entry for one of the suffixes is displayed at the bottom of the screen.

The LanguageEditor is the primary tool used for constructing the Aymara database. It includes a section for entering each unit in the training database (as shown in Figure (1). It also includes sections for phrase patterns (grammar), categories (maintenance of grammatical categories), access to all phrases in the database, and dictionaries for individual words and morphemes. Shown in the figure is an analysis of one of the phrases appearing in a dialogue. Multilingual glosses are supported (in this application English and Spanish). Facilities for recording and storing sound recordings of the phrases are shown, as well as an image associated with this phrase. The phrase analysis includes a morphological breakdown, and each word and morpheme couples directly to the dictionary. Note the markers and features associated with a particular morpheme as shown in figure.

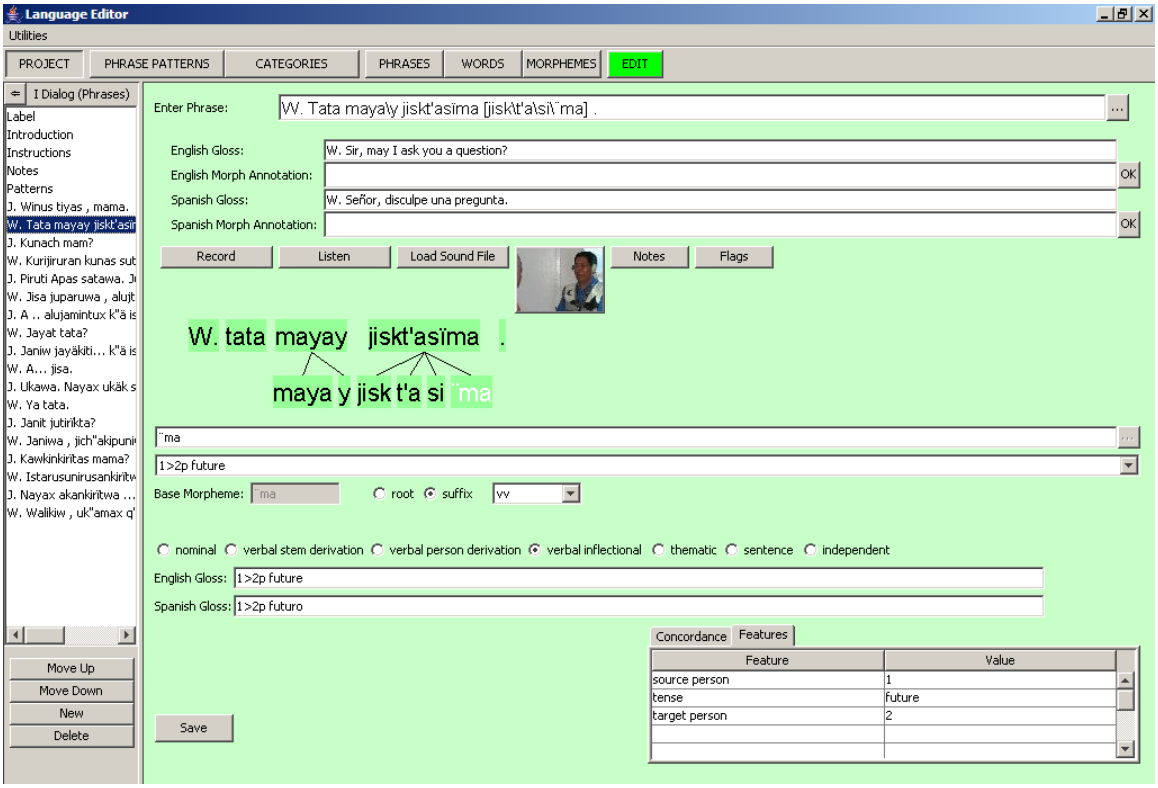


Figure 3. The LanguageEditor is a Web-based tool for creating the Aymara database. It is used on-line by field workers.

### 5.2.2 ObjectEditor

The ObjectEditor is another authoring tool in Lyra, and illustrates how data are entered into the database as conceptual units and examples. The ObjectEditor is a more general purpose object creation tool, in contrast to the LanguageEditor which is customized for creating linguistic objects. We use ObjectEditor to develop documentation for the Aymara grammar. Authors can sequence the same content in different ways as well as add to or eliminate content in order to adapt it for different audiences. In our application, the objects are organized by the issues being discussed. Alternative organizations of the same objects lead to different perspectives. For example, it is also possible to categorize the content by grammatical elements. From this approach, it would be possible to generate an index or several indices to highlight research and/or learning issues.

Figure (4) shows both a presentational view and a conceptual taxonomic view of a set of objects describing the introductory grammar. On the left is a column of topics (origin, nouns and verbs, etc). arranged sequentially (arrows with plain heads represent sequence) within a grammar lesson for Unit I Exercise Set I. This sequence is the order in which concepts are presented to a student in a linear fashion. Labeled associations from the lesson (represented by the node, “I exercise set I lesson”) to each topic indicate the role of the object within the lesson (introduction, definition, conclusion). An expansion of one of the concepts (“yes/no interrogative –ti”) shows a text description of the concept written using a multilingual text editor. The right side of the diagram shows an alternative view in which objects are categorized within a taxonomy of grammatical terms (word, morpheme, sentence, etc.). Only the portion of this taxonomy relevant to this particular set of objects is shown, but in fact the complete database is much larger than this, grouping objects across all lessons, and integrating them with the complete Aymara grammar, as well as to the original data (all words and phrases) captured in the database.

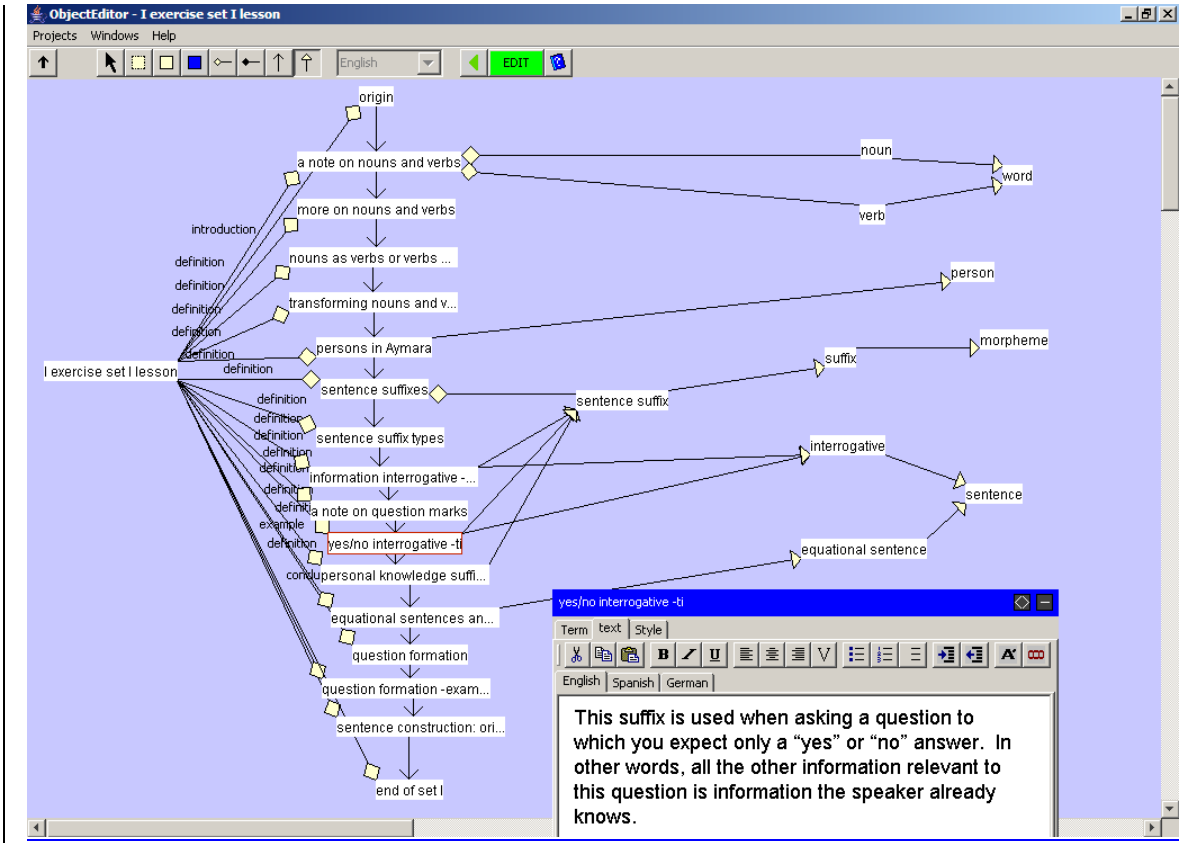


Figure 2. The ObjectEditor is a general purpose tool for creating and browsing database objects. Shown here is the text description of a portion of the Aymara grammar. There are two different classifications, a linear student presentation (left side) and a classification by grammatical category (right side).

**5.3 Database operations**

A significant advantage of ontologies is that they support reasoning facilities which are not available in a conventional relational database. These facilities can be exploited in linguistic research studies. Reasoning facilities are based on comparing the structure of two objects to see how they are similar or different. Reasoners automatically check the consistency of ontology to see that objects are properly classified. Furthermore, automatic classification can be used to determine where in the taxonomy a new class or a new individual should be placed. Classification supports many applications, including query processing. This enables a new way to query databases, based on object structure, that goes well beyond the industry standard SQL to exploit data semantics in query processing.

Another important reasoning facility is automatic clustering, also known as conceptual clustering, in which similarities among objects are automatically discovered to form new categories. We can apply this in the discovery of new phrase patterns by analyzing the similarities over sets of phrases, resulting in machine learning of grammar. Analysis of a new phrase is accomplished by attempting to apply existing phrase patterns. Parsing proceeds as a recognition+learning process. First an attempt is made to map phrase patterns to the new phrase by using a parser (we use a chart parsing algorithm). From another direction, an attempt is made to match similar phrases to the new phrase to try to discover new patterns. Since the database contains thousands of phrases (coming directly from the source data corpus) as well as the abstracted phrase patterns, it is possible to apply case-based reasoning to language learning by relating the new phrase to this case-base of existing phrases (REF Beck 1994). In our Aymara language teaching project, we do phrase generation by instantiating phrase patterns (filling in categories appearing in patterns with individuals that match the phrase pattern constraints).

The ability to reason supports automation of data analysis, leads to machine learning, and intelligent interaction. While these are useful facilities for studying a single language, they can also be

used to study similarities between languages, in particular, since Aymara, Jaqaru and Kawki are closely related, the database reasoner will be able to automatically identify cross-language similarities among these languages.

**5.4 Data Access, Standards and Issues**

We propose to follow standards at the level of data structures for data exchange (XML), rather than at the level of the particular local software systems being used for implementation. That is, the particular database management system used to implement our system is not as important as the ability to exchange the data contained in that database with international partners on other projects using data exchange standards. Furthermore, the data must also be published in a standardized way, preferably using a Web Service, so that data can be accessed not only manually by people browsing a Web site (Such as Figure 1), but by programs needed to access and analyze the data remotely.

The core XML standard used by our system is the Web Ontology Language (OWL), a W3 standard (OWL, 2004). This means that all the objects in our database, including source data, morphemes, words, phrases, phrase patterns, and dialogs can be accessed in OWL format. Furthermore we will incorporate the E-MELD GOLD standard as a particular ontology for linguistic categories. We are not aware of standard Web Service interfaces for the proposed work, but we will collaborate with E-MELD in the creation of such a standard.

In our analysis of the GOLD standard, we discovered that many of the linguistic categories needed in our Aymara project, and that will also be needed for the the Jaqaru and Kawki languages, are not included. Such categories as fourth person, morphological conditioning features, and the morpheme categories shown in Figure 1 must be added. We will work with the E-MELD community to extend the standard to support these additional categories. We see GOLD as an evolving standard, and the introduction of the languages we propose to study provide new insights into categories needed to support languages internationally.

**5.5 Comparison to Work Done Elsewhere (HB)**

**6. Web Service Architecture, User Interface and Interoperability**

There will be two ways to access the database: manually via a Web-based user interface, and automatically via a Web service that will allow remote programs to attach to and access the database.

6.1 Web Service Architecture

Lyra supports an architecture for publishing the Aymara database on the web in multiple forms. The database is wrapped inside services that enable the database to be accessed from remote applications. Primarily the database is wrapped in a web service that supports an application program interface (XML API) for making calls to the database in order to retrieve data in XML format. Alternatively we also support a Java RMI (Remote Method Invocation) server that allows remote Java applications to attach to the database to retrieve data objects directly. The RMI server bypasses the XML API, and that can result in faster performance for certain types of applications. However, only Java applications can access the RMI Server (thus it is restricted to that particular language). The LanguageEditor and ObjectEditors are examples of applications that attach to the RMI Server, and the Aymara training module Flash environment is an example of an application that uses the XML API.

Figure (6) shows the basic architecture with the linguistic database wrapped inside services, and remote applications accessing data in various formats. Here are a few of the methods for accessing data supported by the The XML API:

- getProject(projectID,language) Gets metadata for a particular project in specified language.
- getUnit(unitID,language) Gets lesson unit information in specified language.
- getUnitExercises(unitID,language) Gets exercises for a particular unit.
- getDialog(dialogID,language) Gets a particular dialog
- getPhraseAnalysis(phraseID,language) Gets syntactic analysis of a particular phrase

- getPhrasePatternAnalysis(phrasePatternID,language) Gets a particular pharse pattern (grammar rule)
- getLesson(lessonID,language)  Gets a particular grammar lesson.
- getDictionary(letter, language)  Gets dictionary entries beginning with letter.
- getDefinition(wordID,language)  Gets dictionary entry for a particular word.

Note while these methods are all for getting information from the database, there are also methods for submitting data, although those have restricted access for security reasons.

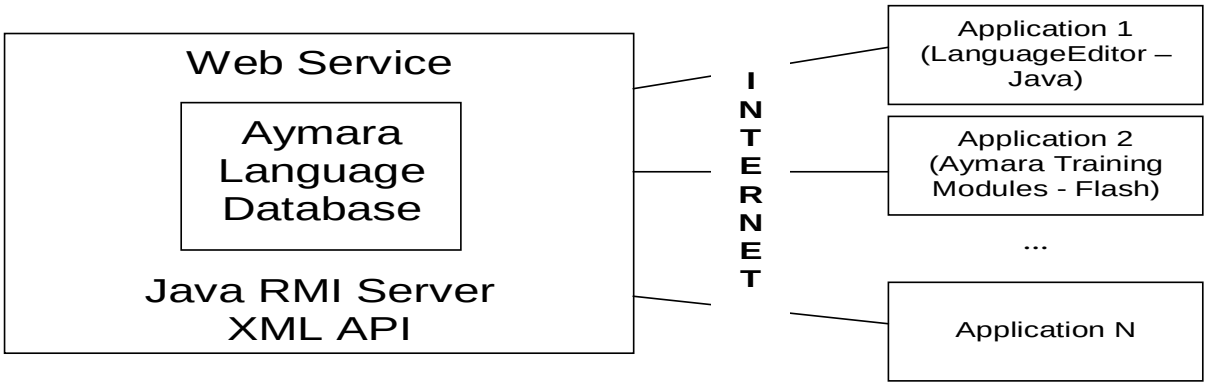


Figure 6. The Aymara language database is published on the Internet using various technologies including an XML API (provides data elements in XML format), and the Java RMI Server (provides data elements as Java objects). Remote applications can attach to these services to both access and enter data.

**6.2 Web-based User Interface**

The database can be accessed manually via a standard Web browser equipped with a Java plug-in (instructions for installing the plug-in are provided). The LanguageEditor in Figure 1 shows an example of such an interface. This Java applet runs in any browser, and enables users to browse all the database content (phrase patterns, phrases, word dictionary, morpheme dictionary, etc.). In addition, the applet provides utilities for data manipulation including automatic classification (which also provides query processing), conceptual clustering, parsing, generation, and case-based reasoning. The applet also supports data editing for users who are authorized to do so (usernames and passwords are provided). Facilities for manually extracting data in XML format will also be available through the Web-based user interface. Should installation of the Java-plugin be an obstacle for some users (though installation of the plug-in is not difficult), we will also provide a standard HTML-based interface to as many of the same functions as possible. Though not as convient as the graphical interface provided by the Java applet, the HTML-interface is more generic and conventional.

**6.3 Web-Service Automatic Remote Access**

Because it is also desirable for remote programs to automatically access the data, a Web-service interface to the database will also be provided. The Web service will enable programs located anywhere on the Internet and written in any language that can support the Web-service protocol (this includes Java, C++, C#, PHP and other popular languages) to remotely access and manipulate the data in real time. All the functions provided in the manual user interface will also be provided via the Web service. Web services operate by publishing a set of interface calls (methods) via a Web service registry that remote programs can use to access the service. For example, a method called “Word getWord(String word)”

could be called by a remote application that would retrieve the data object associated with the word identified in the argument. We will work with the E-MELD community to develop a standard description language for the Web-service interface.

### 6.4 **Interoperability**

While we have developed an OMS comprised of many software systems (authoring tools, visualization tools, physical storage managers, eLearning) it is not necessary to think of the ontology as physically residing within a particular database. Rather the ontology is a knowledge network distributed worldwide, using XML as an exchange format (OWL is XML based and already supports development of distributed ontologies), with different parts of the ontology managed by different people and organizations. We offer the OMS as a conceptual framework for achieving that goal. Furthermore, it is important to stress the significance of data structures over tools. It is no longer necessary to associate knowledge with the authoring tools used to create that knowledge. Rather the focus should be on the data structures created by tools. Different tools can operate on the same data structures. The first step towards achieving interoperability is to change focus from tools to data structures. We also need to distinguish between archiving standards such as OLAC, which provide course-level metadata descriptions of archived resources, and database standards that allow sharing of fine-grained data structures. Ultimately of course it is desirable to share data structures, but the standards are not yet available to achieve that goal.

Interoperability requires standards for data structures. While everyone seems to agree that such standards are needed, achieving them is difficult. Attempts at standard building include standardizing on tool-specific data structures [], an interlingua by which proprietary standards are converted to and from an intermediate accepted standard [], language-to-language converters that attempt to translate between different data structures without the use of interlingua []. We propose that the ontology, viewed as a comprehensive, distributed database, can solve the interoperability problem. We argue that other proposed solutions are basically special cases of this approach. The world-wide linguistics ontology would be an ever expanding set of data structures and formal definitions (abstractions in the form of ontology classes) for those structures. The ontology would be physically distributed over many geographic locations, essential wherever linguistic databases are being built. Communities of Practice (COP) such as currently envisioned for the GOLD standard [] participate in working on low-level data structures needed for specific domains while paying attention to related work in the area, and generalizing their work when appropriate to apply to other areas. Standards building thus becomes a database building process on a global scale. Most of all it requires a shift in thinking by tool builders who need to create special data structures to handle the custom features of their tools. Such structures can be created in the proposed global framework as long as they are registered within an appropriate community of practice.

### 6.5 Server Location and Support

The University of Florida's Office of Academic Technologies (OAT) is the central educational computer support facility at UF and has extensive experience with enterprise server support and secure long-term data archiving. OAT currently hosts the Lyra ontology management system and associated deployed applications (this includes EDIS, which receives 6 million visitors per year). The core servers consist of two dual-processor Linux systems containing 2GB+ main memory and ½ terabyte of disc storage. OAT is institutionally committed to maintaining the availability of the existing Aymara database and the proposed new Jaqaru and Kawki linguistic databases over the long term. We will use the servers that currently support the Aymara database for the proposed project. System administration for the resident servers (including regular backups, security and support) is provided by the U.F. Office of Academic Technologies.

## 7. Work Plan: Roles and Responsibilities

**M.J. Hardman**, P.I., is the chief researcher for the project and the investigator responsible for all linguistic matters. She will work closely with project personnel to provide guidance and to write specifications on the organization of the linguistic database, the selection and parsing of the linguistic and cultural

materials, the translation of the materials, and the involvement of appropriate colleagues in the assessment of the work. She will carry out the grammatical analysis for the database and act as general editor for all materials entered into the database. As P.I. , Hardman will assure that the project is subject to appropriate ethical oversight. She will write research papers for presentation at conferences and publication in journals.

**Elizabeth Lowe McCoy, Co-P.I**., will serve as the project manager and will oversee the budget, manage team communications and coordination, and write additional funding documents. She will be the primary liaison between the University of Florida team and Peruvian counterparts, as well as project liaison with outside technical bodies such as E-MELD and OLAC. She will work with the content experts, the team responsible for creating the electronic version of the materials, and the database expert to ensure that the materials are prepared appropriately for the database, the research function and the online environment. Dr. Lowe McCoy will oversee English and Spanish language translation of the database materials as well as supervise the work of the language consultant, Dr. Bautista, to ensure compliance with University of Florida conflict of interest policies.

**Howard Beck, Co-P.I.** will be the computer scientist for the project and will ensure the functionality, robustness and usefulness of the database for research and training purposes. Dr. Beck will also ensure that the database conforms to current E-MELD and GOLD standards. He will become involved in the work of E-MELD and write research papers for presentation at conferences and publication in journals.

**Sue Legg, Technical Liaison and Evaluation Consultant,** will serve as the technical liaison for the project. In this capacity, she will assist in the planning for the data entry and data analysis functions. This planning will include the development of procedures and timelines. She will monitor the progress toward completing these functions and suggest alternatives for overcoming any obstacles. This role also includes coordinating with the database designer, Howard Beck, and the project directors to facilitate the resolution of any technical adjustments needed for improving the data entry process. Dr,. Legg will design evaluation instruments for the project and work with outside evaluators on assessing the use of the database materials in a variety of contexts.

**Dr. Dimas Bautista Iturrizaga**, **Language Consultant**, will be the general arbiter of all the Jaqaru data. He will be in charge of the writing of definitions for the dictionary and with the country coordinator will design and conduct training of native Jaqaru and Kawki speakers who will be learning to read and write their languages, to produce texts and to handle data for the database.

**Yolanda Nieves Payano Iturrizaga**, **Peru Coordinator**, will be responsible for recruiting and training persons in Peru and in Tupe for the entry of material into the database and for using the database for purposes within the Jaqaru and Kawki speaking communities. She will come to the University of Florida for short training periods during the life of the grant project.

**8. Project Timeline and Procedural Plan**

This project will take three years to complete. The project investigators will meet on a weekly basis during the academic year to assess completed work, monitor progress in relation to project goals, and to ensure that the project is progressing according to the work plan. The country coordinator will be brought to the U.F. campus for training and team consultations for eight weeks in the first year and for ongoing project work for eight weeks in the second year. Ongoing communications with the Peru Coordinator and her local assistants will occur regularly by phone, email and videoconferencing. The P.I. and the Language Consultant will travel once a year to Peru to work closely with the Peru Coordinator to provide support, as well as to monitor the progress of local training efforts and project related work. The work plan for each year is as follows:

**First Year (September 2008 – August 2009)**

| | |
|---|---|
| Put team in place; hire graduate and student assistants | October 2008 |
| Develop programming specifications for the database | |
| Plan the content of the public facing website | November 2008 |
| Digitize notebooks, audio tapes, slides and other materials | October 2008-March 2009 |
| Begin data entry into the linguistic database | January-August 2009 |
| Begin parsing of the linguistic material | January-August 2009 |
| Begin English translations of the linguistic material | February-August 2009 |
| Complete missing Spanish translations of the linguistic material | March-August 2009 |

| | |
|---|---|
| Conduct training for the Peru coordinator at the University of Florida | January-March 2009 |
| Catalog and prepare metadata of language resources for AILLA | September 2008-August 2009 |

**Second Year (September 2009-August 2010)**

| | |
|---|---|
| Recruit in-country linguists for work on the collaborative dictionary | September 2009 |
| Conduct training in Peru for Jaqaru and Kawki speakers | September 2009-August 2010 |
| Complete data entry of notebooks into the linguistic database | September 2009-August 2010 |
| Continue parsing of linguistic material | September 2009-August 2010 |
| Continue English translations of the linguistic material | |
| Disseminate the project in Peru and establish positive working ties | March 2010-July 2010 |
| With the Peruvian government | |
| Conduct further training and team consultations with Peru coordinator | January 2010-March 2010 |
| at the University of Florida | |
| Submit language resources to AILLA | October 2009 |

**Third Year (September 2010- August 2011)**

| | |
|---|---|
| Complete data entry of audio tapes into the linguistic database | September 2010- May 2011 |
| Complete parsing of all the linguistic material | September-December 2010 |
| Complete English translation of linguistic material | September -December 2010 |
| Edit text files | September-December 2010 |
| Edit digitized audio and photo files | September-December 2010 |
| Enter photo files into the database | September-December 2010 |
| Submit database to the UF Digital Collections | |
| Launch the linguistic database and public website | January 2011 |
| Present papers at conferences on standards | February-July 2011 |
| Conduct a dissemination event in Peru | April 2011 |
| Solicit funding for follow-on project to create Jaqaru and Kawki | September 2010 -January 2011 |
| Pedagogical materials | |

**9. Broader Impact and Intellectual Merit**

The broader impact of this project will be to preserve and make available in four languages the grammar and dictionary of two endangered Andean languages for linguistic research and for the use and collaboration of heritage speakers in a strategically important world region. The intellectual merit of the project resides in the multifunctional nature of the database, which has features that are distinct from the SIL LinguaLinks project and the FLEx project, or work done by Charniak and Collins For the Jaqaru and Kawki speakers, this database of rich text represents a repository of history as well as an opportunity to contribute to the archiving and preservation of their languages. Some of the recordings are of people who were born in the 19$^{th}$ century; they are the words of the grandparents of their grandparents. These ancestors are no longer with us; their voices constitute the major account of the history of both Tupe and Cachuy. By making available the language, as well as the notes and images contained in the database, the proposed project will retrieve and perpetuate the history and culture of the Tupe and Cachuy that will otherwise be lost. The database materials can be adapted for bilingual education, for literary materials as well as for historical and cultural purposes, including personal identity affirmation and continuing education. Training will be provided throughout the life of the grant to enable bilingual and heritage speakers, as well as those who have lost the language, to use the database for recording their own material, as well as for pedagogical and research purposes. It is anticipated that these trainees will become future trainers and teachers of the endangered languages. For the linguistic scholar, having this type of rich text with full grammatical analysis, in Jaqaru, Kawki, Spanish and English, provides the resources for comparative linguistics, for the reconstruction of proto-Jaqi, and for adding to the body of work on linguistic change. Also the existence of this database can make possible additional studies to further the grammatical analysis of the languages. As examples, the morphophonemics, with some phonological conditioning but primarily morphological conditioning, is extremely complex, including

noncontiguous conditioning. This database can provide data in a format that will allow further studies of this system. Also, it is hoped that the study of the distribution of the sentence suffixes in relation to syntactic structures can be furthered by the use of the database. The nature of the database, together with the nature of the original transcriptions provides data for future questions that we are not yet asking. Even after 50 years of research on the Jaqi languages, Hardman has found occasion to ask new questions, many of which have to do with discourse analysis, sociolinguistics, and dialogue structure. The existence of this database will allow future studies as new questions are discovered. It will open up as yet unforeseen possibilities to native speakers of the language as well as linguistic scholars, both in Peru and around the world.

## 10. Planned Future Projects

It is hoped that this project will lay the foundation for future work by speakers of Jaqaru and Kawki. In the future, we would like to bring young scholars to the University of Florida for training in linguistics as well as computational linguistics so that they can carry on this work in Peru at the university level. The project investigators would also like to give more direct help to the school teachers in the region who are doing the difficult work of preserving the language. We hope that the proposed database will help further research on linguistic databases and will become the foundation for future collaborative work on linguistic databases for rare and endangered languages.