Aymara on the internet: a step toward interoperability and user access

Howard Beck, Sue Legg,
Elizabeth Lowe & M. J. Hardman

Proceedings of Conference on

Language Documentation & Linguistic Theory

Edited by Peter K. Austin, Oliver Bond & David Nathan

7-8 December 2007 School of Oriental and African Studies, University of London

School of Oriental and African Studies
Thornhaugh Street, Russell Square
London WC1H 0XG
United Kingdom

Department of Linguistics:
http://www.soas.ac.uk/academics/departments/linguistics

Hans Rausing Endangered Languages Project:
http://www.hrelp.org
elap@soas.ac.uk

ISBN: 978-0-7286-0382-0

Aymara on the internet: a step toward interoperability and user access

HOWARD BECK, SUE LEGG ELIZABETH LOWE & M. J. HARDMAN
University of Florida, USA

1. THE ETHNOGRAPHIC SETTING

Our work is with the Jaqi family of languages (Aymara, Jaqaru, Kawki) which is spoken in the Yauyos valley in the south central Andes mountains. 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. Torero 1974). 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 Aymara were the members of a great but little known culture centered in the ancient city of Tiahuanaco. From about 400 AD to 1000 AD their empire spanned the south central Andes mountains, and they established both a complex system of irrigation by canals and a complex system of mercantile exchange, involving the construction of roads. Today, 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. Jaqaru and Kawki are endangered sister languages to Aymara. Jaqaru is spoken in the Andes Mountains of Peru by a few thousand people resident in Tupe, Yauyos and Lima 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. Kawki is now almost extinct; there exist only a few native speakers.

2. THE AYMARA ON THE INTERNET PROJECT

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 draws upon Professor M.J. Hardman's field research in Peru in the 1950s as well as the classroom teaching materials she developed through a series of funded grants dating from the 1970s. A grant from the U.S. Department of Education (2004-2007) funded the project to convert the Aymara materials into a web-based delivery format for language learning. The vision expanded, however, when the possibilities of an ontology-based database backend for the instructional materials were introduced to the project planning team. The realization that the students as well as other linguists would have easy access to complex linguistic structures led the project to a methodology that offers not only language learning, but also an XML metadata driven archive of the materials, a web services architecture for sharing a generic database tools for linguistic data collection and analysis, and a new way of thinking about how data and tools to manipulate data will become shareable in a diverse community of linguists, educators and the general public.

3. AN ONTOLOGY MANAGEMENT SYSTEM FOR LANGUAGE ARCHIVING

We have constructed an open source ontology management system (OMS) called Lyra which provides a number of basic data management services. The OMS provides a framework for representing and integrating everything from raw data elements including sound recordings, transcripts, images, and video to more abstract linguistic elements including morphemes, words, phrases, phrase patterns (grammars), and dialogues. Lyra uses an ontology language for data modeling and representation based on OWL, the Web Ontology Language (W3 Consortium 2004), and it supports physical storage management optimized for storage and retrieval of large numbers of objects. While ontologies are traditionally used to represent semantics of words, we expand the use of ontologies to the level of a full-blown database management system. That is, an OMS is a database system that uses a formal ontology language such as OWL as the data definition language. There are several advantages of using an OMS rather than a traditional relational database or no database at all, e.g. an XML file system. First would be the more natural way in which an OMS models linguistic data. The ontology language provides a way to model taxonomic and other relationships among objects, with direct pointers between abstract generalizations and concrete data such as field observations (Beck et al. 2005). Though this can be done with relational databases, it is much more difficult as these complex structures must be mapped to normalized relational tables in which the object structure and relationships are no longer explicit. This can also lead to less efficient data retrieval (Lee 2004, Liu and Hu 2005).

Another advantage of using an OMS 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 the 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 (OWL-DL, the description logic version of OWL, is designed to balance this tradeoff). Finally, the OMS offers a variety of services to support security, data integrity, transaction management, and query processing that are necessary for publishing data as a web service with sharable XML data formats. XML files are better suited as best practice for exchange of data among different systems, but a database management system is needed to provide operational functionality.

4. THE ACCESSIBLE LINGUISTIC DATABASE

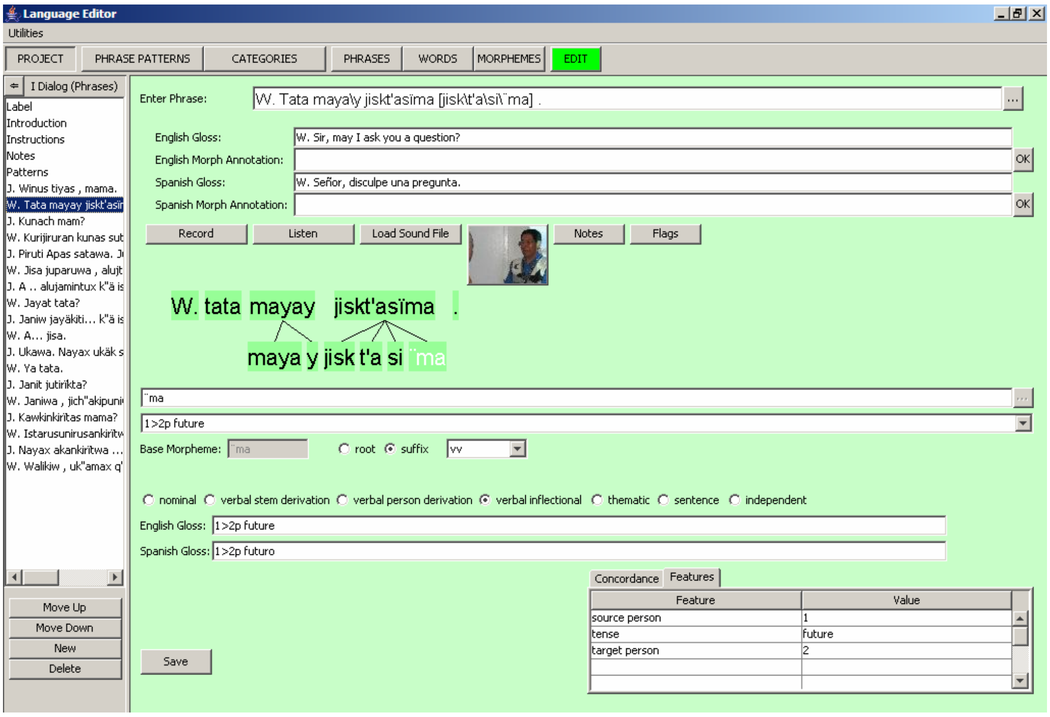
We emphasized the importance of total analysis using a linguistic database built on a collaborative work environment in order to explicitly capture the existing detailed analysis at the level of grammar for phrase and word structure and to interrelate all resources, connecting abstract theories with raw field data. The detailed analysis in the form of grammars describing phrase and word structure is used as a framework for organizing raw materials and cultural resources, and can also assist in error correction (detecting spelling and tagging errors). We have tried, with the database, to present a more complete description of the languages beyond what already exists in the grammatical statements, e.g., regarding a refinement of the rules for vowel dropping/keeping.

5. COLLABORATIVE AUTHORING TOOLS

Tools for creating and managing a linguistic database must be on-line, web-based, multi-user and cross-platform in order to provide a wiki-style collaborative work environment where many different people can contribute their expertise to building a shared database (Melby et al. 2006, Simons et al. 2007). Lyra contains several authoring tools, including LanguageEditor and ObjectEditor that provide these functions within 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 from distributed geographic locations (in the Aymara project this included Florida, Bolivia and Peru). The tools are web-based (run inside web browsers using plugins) and form the basis for a collaborative environment for creating data objects stored in and retrieved from a common database.

The LanguageEditor is a web-based tool for browsing and editing the language database (Beck 2007). LanguageEditor is a cross-platform Java applet that runs in any web browser that has the Java plugin (Sun Microsystems Inc. 2007). It connects remotely to access the database over the Internet. It includes modules for dialogues, phrase patterns (grammar), phrases, words, and morphemes as well as cultural resources (images, sounds), and text narratives and other documentation. Shown in Figure 1 is an analysis of one of the phrases appearing in a dialogue (Aymara). Multilingual glosses are supported (in this application English and Spanish). Facilities for creating and storing sound recordings of the phrases are shown, and images can be associated with phrases and individual words. The analysis of the phrase includes a morphological breakdown, and each word and morpheme couples directly to its dictionary entry. Note the markers and features associated with a particular morpheme. Each object shown in the figure is internally cross-referenced with related data objects, for example, word senses are enforced so that it is known when the same sense of a word is used in two different phrases.

Figure 1
Language Editor illustrating the analysis of an Aymara phrase



Label
Introduction
Instructions
Notes
Patterns
J. Winus tiyas , mama.
W. Tata mayay jiskt'asi
J. Kunach mam?
W. Kurijiruran kunas sut
J. Piruti Apas satawa. J
W. Jisa juparuwa , alujt
J. A .. alujamintux k"ä is
W. Jayat tata?
J. Janiw jayäkiti... k"ä is
W. A... jisa.
J. Ukawa. Nayax ukäk s
W. Ya tata.
J. Janit jutirïkta?
W. Janiwa , jich"akipuni
J. Kawkinkirïtas mama?
W. Istarusunirusankirïtw
J. Nayax akankirïtwa ...
W. Walikiw , uk"amax q'

Enter Phrase: W. Tata maya\y jiskt'asïma [jisk\t'a\si\¨ma] .

English Gloss: W. Sir, may I ask you a question?

English Morph Annotation:

Spanish Gloss: W. Señor, disculpe una pregunta.

Spanish Morph Annotation:

¨ma

1>2p future

Base Morpheme: ¨ma ○ root ● suffix vv

○ nominal ○ verbal stem derivation ○ verbal person derivation ● verbal inflectional ○ thematic ○ sentence ○ independent

English Gloss: 1>2p future

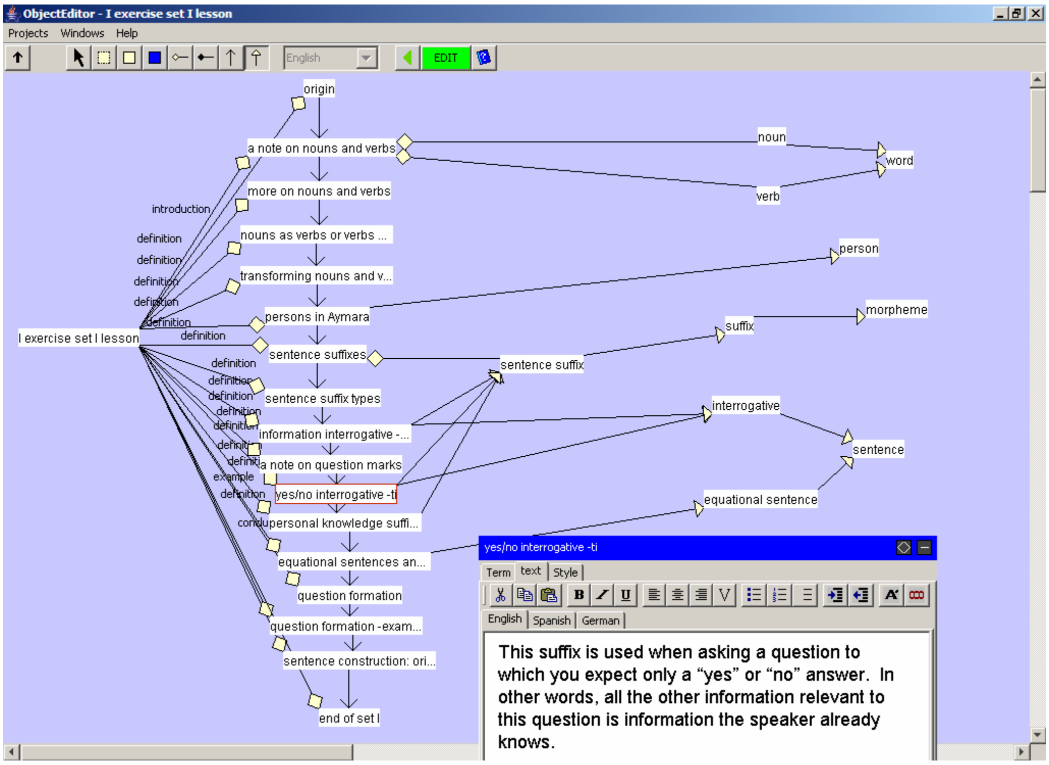
Spanish Gloss: 1>2p futuro

Feature	Value
source person	1
tense	future
target person	2

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. For presentation to students, 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 2 shows 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. This sequence is the order in which concepts are presented to a student. Labeled associations from the lesson ('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 the database groups objects across all lessons, and integrates them with the complete Aymara grammar, as well as to the original data (all words and phrases) captured in the database.

Figure 2
The Object Editor showing introductory Aymara grammar from the standpoint of students (left side) as well as a taxonomy of grammatical elements (far side)



This suffix is used when asking a question to which you expect only a “yes” or “no” answer. In other words, all the other information relevant to this question is information the speaker already knows.

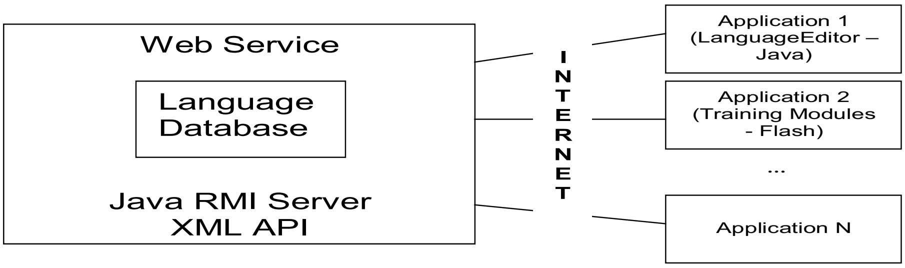
6. WEB SERVICE ARCHITECTURE

Lyra supports an architecture for publishing the language database on the Internet in multiple forms. The database is wrapped inside web services that enable the data objects to be accessed from remote applications. The web service 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 as Java objects. The RMI server bypasses the XML API and can result in faster performance for certain types of applications. However, only Java applications can access the RMI Server. The LanguageEditor and ObjectEditors are examples of applications that attach to the RMI Server, and the Aymara training module Flash environment (Beck et al. 2007, Adobe Systems Inc. 2007) is an example of an application that uses the XML API. Figure 3 shows the web service architecture with the linguistic database wrapped inside. Remote applications can attach to these services to both access and enter data. Here are a few of the methods supported by the XML API for accessing data. Note that 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:

- getDialog (dialogID, language). Gets a particular dialog.
- getPhraseAnalysis (phraseID, language). Gets syntactic analysis of a particular phrase.
- getPhrasePatternAnalysis (phrasePatternID, language). Gets a particular phrase pattern (grammar rule).
- get Dictionary(letter, language). Gets dictionary entries beginning with letter.
- getDefinition (wordID,language). Gets dictionary entry for a particular word.

Figure 3

Web service architecture for publishing the database over the Internet


Web Service
Language Database
Java RMI Server
XML API
INTERNET
Application 1 (LanguageEditor – Java)
Application 2 (Training Modules - Flash)
...
Application N

7. STANDARDS AND INTEROPERABILITY

While we have developed an OMS comprised of many software systems (authoring tools, visualization tools, physical storage managers, e-Learning), it is not necessary to think of the ontology as physically residing within a particular centralized database or proprietary software system. Rather the ontology is a knowledge network distributed worldwide, using XML as an exchange format, with different parts of the ontology managed by different people and organizations. It is very 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. We also need to distinguish between archiving standards such as OLAC (Simons and Bird, 2003) which provide coarse-level metadata descriptions of archived resources, and standards that allow sharing of fine-grained data structures. While we are confident in our ability to archive our raw field data using OLAC, we can only outline an approach for sharing the detailed database objects.

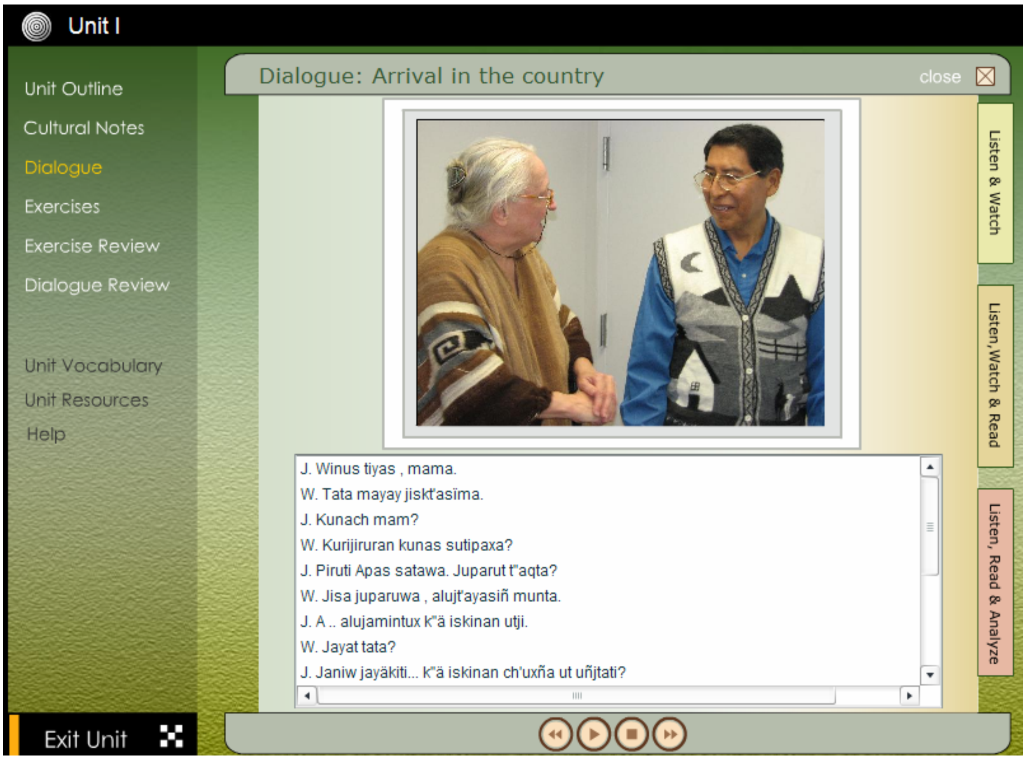
Existing standards are piecemeal at best (a particular standard addresses only a subset of the database). We know of no standard that addresses all content in a linguistic database such as Lyra in a uniform way. OWL addresses the generic data model used in Lyra, but not the particular linguistic elements which are created on top of the ontology language. Standards exist for subcategories of elements, and can be characterized as formal, informal, implicit, or proprietary (Rumble et al. 2005). Standards of all four types have been identified (Thieberger et al., 2007) in lexicography such as TEI (2007), LMF (2007) and WordNet (Fellbaum 1998), terminology including TMF (2007) and GOLD (Farrar and Langendoen 2003), metadata including OLAC (Simons and Bird 2003) and DDI (2007), word lists including IDS (Key 2007), and semantic fields including DDP (2007). No notable standards exist for grammars, interlinear glossed text (IGT), or corpus annotation. Until more and better defined standards emerge, we can create generators and parsers that will export and import data in the format of these piecemeal standards.

While there is wide agreement that interoperability requires standards for data structures, establishing them is difficult. Attempts at achieving interoperability include standardizing on system-specific data structures such as FLEx (SIL International 2006) and WordNet (Fellbaum 1998), conversion of proprietary standards into a common accepted standard (Dipper et al. 2006), and process-based models that use tool-to-database converters that access shared data structures through a common API (Cochran et al. 2007). We propose that the ontology, viewed as a distributed database accessed through web services, can provide a platform for solving the interoperability problem (similar to Simons et al. 2004). 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). The ontology would be physically distributed over many geographic locations, essential wherever linguistic databases are being built. Communities of Practice (COP) such as adopted for the GOLD standard (Farrar and Langendoen 2003) participate in working on low-level data structures needed for specific domains while paying attention to related work in the area. Standards building thus becomes a database building process on a global scale. This will require a shift in thinking by tool builders who need to create special data structures to handle the custom features of their tools. Such data structures can be created in the context of a global framework as long as they are registered within an appropriate community of practice.

8. AYMARA E-LEARNING STUDENT INTERFACE

The student interface was constructed using Flash (Adobe Systems Inc. 2007) that downloads from the Aymara web site and runs directly in the student’s web browser. The Flash program accesses the remote database to retrieve data objects in XML form which are then formatted by the Flash program for display to the student. The student interface is multilingual and contains 12 chapters that include cultural notes, dialogues, exercises, and review materials in Spanish or English. Figure 4 shows one of the three options, 'Listen, Watch & Read', for studying dialogues in which the dialogue is presented using audio, images and text. Another option allows the students to view a grammatical analysis of each phrase in the dialogue. The student interface also shows grammatical analysis of each phrase in the dialogue. The screen includes the gloss for the phrase as well as for each morpheme. A concordance table lists other phrases in which a particular word or morpheme is found. Finally, the student interface presents exercises that relate to each dialogue. In the exercises, students read, listen and respond to a question by writing an answer using the provided cues. The student checks his/her response by comparing it to the correct answer. The student can click on an expanded version of the question and answer to see the grammatical analysis of the phrase. Exercises present students with questions, cues, hints (pull down lists), and correct answers (sound and grammatical expansion are also available).

Figure 4
Aymara student interface showing the Dialogue option: 'Listen, Watch & Read'



J. Winus tiyas , mama.
W. Tata mayay jiskt'asïma.
J. Kunach mam?
W. Kurijiruran kunas sutipaxa?
J. Piruti Apas satawa. Juparut t"aqta?
W. Jisa juparuwa , alujt'ayasiñ munta.
J. A .. alujamintux k"ä iskinan utji.
W. Jayat tata?
J. Janiw jayäkiti... k"ä iskinan ch'uxña ut uñjtati?

9. CONCLUSION

We have successfully deployed and developed a web-based collaborative language documentation system based on the Lyra ontology management system. The resulting database is published on the Internet as a web service. Client programs, including authoring tools and educational programs, can attach to this database to create as well as to extract linguistic data structures. In this system, language preservation (creating these data structures) and language training (e-Learning software that utilizes these data structures) are closely integrated. We hope to expand the system by making the authoring tools available to more people within the Aymara-speaking regions of Peru and Bolivia. Moreover, the system can be applied to any language, and we plan to add Jaqaru and Kawki. So far the tools have been used only by individuals working directly with the Aymara project. As more users participate in documenting these languages, we will need to implement editorial review and quality control procedures. We will also need to develop proficiency testing tools to assess user outcomes. We hope to work closely with organizations developing standards and that our efforts will contribute towards global interoperability of language resources.

REFERENCES

Adobe Systems, Inc. 2007. Flash. http://www.adobe.com/products/flash.

Beck, H., M. J. Hardman, G. Lord and M. Pineros. 2005. A linguistics database supporting language education. III PGL Database Conference. Sao Paulo, Brazil. Invited Paper.

Beck. 2007. Lyra Language Editor. http://orb2.at.ufl.edu/LyraEditor/languageeditor.html.

Beck, H. M.J. Hardman, G. Lord, S. Legg, E. Lowe, & J. Llanque Chana. 2007. Aymara On-Line. http://orb2.at.ufl.edu/Aymara.

Cochran, M., J. Good, D. Loehr, S. Miller, S. Stephens, B. Williams, & I. Udoh. 2007. Report from TILR working group 1: Tools interoperability and input/output formats. In E-MELD Workshop: Toward the Interoperability of Language Resources. http://emeld.mseag.org/wiki.

DDI. 2007. Data Documentation Initiative. http://www.icpsr.umich.edu/DDI.

DDP. 2007. Dictionary Development Process. SIL International. http://www.sil.org/computing/ddp.

Dipper, S., E. Hinrichs. T, Schmidt, A. Wagner, & A. Witt. 2006. Sustainability of Linguistic Resources. In: E. Hinrichs. N. Ide, M. Palmer; and J. Pustejovsky (eds.): *Proceedings of the LREC 2006 Satellite Workshop on Merging and Layering Linguistic Information*. Genoa 2006.

Farrar, S. & T. Langendoen. 2003. A Linguistic Ontology for the Semantic Web. GLOT International 7(3), pp. 97-100.

Fellbaum, C. ed. 1998. *WordNet: An electronic lexical database.* The MIT Press.

Key, M. (ed). 2007. Intercontinental Dictionary Series. http://www.eva.mpg.de/lingua/files/ids.html.

Lee, R. 2004. Scalability Report on Triple Store Applications. SIMILE. http://simile.mit.edu/reports/stores.

Liu, B. & B. Hu. 2005. An evaluation of RDF storage systems for large data applications. *First International Conference on Semantics, Knowledge and Grid (SKG'05)*. p. 59.

LMF. 2007. Lexical Markup Framework. ISO/DIS 24613

Melby, A., P. Fields and M. Carmen. 2006. Language Databases, Statistics and Social Networks. LACUS Forum XXXII. Dartmouth College, Hanover, New Hampshire

Rumble, John Jr., Bonnie Carroll, Gail Hodge, & Laura Bartolo. 2005. Developing and Using Standards for Data and Information in Science and Technology. PV 2005. Edinburgh, Scotland.

SIL International. 2006. Fieldworks Language Explorer. http://www.sil.org/computing/fieldworks/flex.

Simons, G. & S. Bird. 2003. OLAC metadata. Standard, Open Language Archives Community. http://www.language-archives.org/OLAC/metadata.html

Simons, G., B. Fitzsimons, D.T. Langendoen, W. Lewis, S. Farrar, A. Lanham, R. Basham, & G. Hector. 2004. A model for interoperability: XML documents as an RDF database. E-MELD Workshop on Linguistic Databases and Best Practice. Wayne State University. Detroit, Michigan.

Simons, G., A. Sevigny, J. Park, A. Kibort, S. Legg, E. Pyatt, E. Lowe, P. Cash Cash, D. Chang, & T. Kendall. 2007. Report from TILR working group 5: Web Services and Web 2.0. In E-MELD Workshop: Toward the Interoperability of Language Resources. http://emeld.mseag.org/wiki.

Sun Microsystems, Inc. 2007. Java. http://java.sun.com.

TEI. 2007. Text Encoding Initiative. http://www.tei-c.org.

Thieberger, N., E. Hinrichs, M. Cysouw, H. Sloetjes, H. Yi, L. Veselinova, D.T. Langendoen, H. Beck, & D. Anderson. 2007. Report from TILR working group 6: Standards and Data Models. In E-MELD Workshop: Toward the Interoperability of Language Resources. http://emeld.mseag.org/wiki

Torero, Alfredo. 1974. El Quéchua y la Historia Social Andina. Universidad Ricardo Palma. Lima, Peru.

W3 Consortium. 2004. Web Ontology Language. http://www.w3.org/2004/OWL