

Toward the Interoperability of Language Resources

July 13-15, in conjunction with the 2007 LSA Summer Institute



Contextual Archiving with Linguistic Analysis: An Ontology-Based Approach to Developing a Linguistic Database

Howard Beck and M.J. Hardman
University of Florida, Gainesville

We are utilizing an ontology management system (OMS) for representing and archiving knowledge associated with the language and culture of Aymara, the native language of Bolivia, Peru and Chile. Our project is based on the work of linguist M.J. Hardman, whose life work has been the study of the Jaqi languages (Aymara, Jaqaru and Kawki). Funded originally by a grant from the U.S. Department of Education Title VI program (2004-2007), we now wish to expand the work to create an ontology-based linguistic database for the Jaqi language family with the development of tools for archiving and sharing linguistic information on these endangered languages. 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, and dialogs), including grammatical information about each element. Logical relationships between any two elements are expressed using ontology property relationships.

While ontologies are traditionally used to represent semantics of words, we expand the use of ontologies to the level of a 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 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.

Ontologies support reasoning useful for organizing and searching the language database. Basic operations include automatic subsumption and classification, useful for categorizing elements. There are also conceptual clustering techniques that can compare and contrast the structure of two elements to see how they are the same or different and which can be used to induce new categories. Any new element added to the database is automatically compared with



the existing elements, and the database structure is modified as necessary to accommodate a new element. Thus the database accommodates, archives and makes accessible the actual texts and grammar of the language, in a manner useable by school teachers, native speakers and scholars in a fully contextualized environment.

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.

