

Searching Interoperability Between Linguistic Coding and Ontologies for Language Description: Language Acquisition Data

Barbara Lust, Suzanne Flynn, Maria Blume, Jon Corson-Rikert, and Brian Lowe
Cornell University, MIT, Albert R. Mann Library

E-MELD (Electronic Metastructure for Endangered Languages Data)
2005 Workshop on Digital Language Documentation
July 1-3
Cambridge, Mass.

Abstract

In this paper we will present an overview of a project which has been developing over the last several years, centered at Cornell University, but integrating several national and international institutions (MIT, Rutgers/New Brunswick, Rutgers/Newark, California State University, Southern Illinois University, City University of New York, Columbia University and several sites in India, Taiwan and Peru for example). Funded by planning grants from programs at NSF, and working with collaboration of Cornell's Albert R. Mann Library, this project is now building an infrastructure for the shared collection, representation, preservation, access and dissemination of large amounts of cross-linguistic data in the field of language acquisition. This project involves the creation of materials at several levels: (i) best practice manuals for scientific research in the field; (ii) software for the mark up of metadata and data and their seamless integration in a coherent relational database; (iii) a multi-level Web-based ontology tool for managing diverse inter-disciplinary resources within and beyond the university library, as a platform for disseminating metadata about, and ultimately access to, detailed linguistic resources. We will begin to investigate the possibility for integration with current developments in GOLD (General Ontology for Linguistic Description). At first glance, realizing this potential appears to require developing interoperability with upper level ontologies (ULO) as well as with lower level ontologies (LLO). For example, with regard to LLO, our current coding of language attempts to verify and calibrate metadata and data ranging from subject to session to utterance transcription. Below that it begins to code linguistic elements in a manner which allows comparability across widely varying language data (English, Romance Languages, Hebrew, and several East Asian and South Asian languages) and across widely varying language acquisition stages from initial to adult state. In particular, we are now attempting to develop morphological coding in this system. For this challenging process, we will appreciate the independent developments in GOLD for morphological markup, and seek formats which can link to universal and standardized annotation systems at this critical morphological level. At each level (i-iii above) we will articulate both promises and problems involved in current work, in the possible integration with GOLD, and in linking cross-linguistic language acquisition data to the general purpose of making "our combined knowledge of the world's languages fully accessible and interoperable."

I. Introduction

Working in the field of language acquisition, our fundamental charge is to better understand the relation between nature and nurture. In certain paradigms, this leads to a deeper understanding of the relation between Universal Grammar (that which is biologically determined and common to all languages) and Specific Language Grammar (that which must involve experience and learning). Working in this field on these issues requires amassing and assessing an immense amount of language data (both experimental and naturalistic), necessarily including cross-linguistic data, and rendering these diverse data accessible to empirical study. Immediately this charge leads to the necessity of database development and interoperability of multiple databases at a cross-linguistic level. All the issues which GOLD addresses are confronted in this domain of language acquisition—to an exponential level. These data must be rendered comparable; thus they require theory-neutral standards for description. At the same time, they must be recognized as inherently variable. At early periods, the child language or the adult second language learner's language is expected to vary from the adult model, within each language. Potential variability across languages must be assumed, at the same time that this early language is continuous with the adult language. The question is how, and whether there are general cross-linguistic principles which explain these variations. Systems for language description and analysis must allow scientific evaluation of this question. At the same time, firm distinctions between syntax, semantics, phonology and morphology cannot simply be assumed. Questions regarding language acquisition involve how much the subject/learner may be depending on one or another of these in their early language. Scientific inquiry must find a way of representing these independently, but evaluating their role in every utterance. Ontologies¹ developed to guide annotation in linguistic description, such as GOLD, can critically aid the representation of language acquisition data and the study of language acquisition. However, they may assume consistency of data, or firm distinctions between grammatical and non-grammatical knowledge which cannot be assumed in study of language acquisition, but which must themselves be studied. They may assume that upper level and lower level ontologies for language representation can be independently developed and maintained.² In this paper, we pursue merging ontologies as a means of approaching these issues. We do so through the collaboration of a research lab (the Cornell Language Acquisition Lab) and a Virtual Center for the study of Language Acquisition (VCLA) which links language acquisition labs internationally, with a research library (Albert R. Mann Library at Cornell). Through the VCLA (www.clal.cornell.edu/vcla) a Virtual Linguistics Lab is being constructed (Figure 1) which includes a language data representation system (a Data Transcription and Analysis (DTA) tool). This tool and this general structure can by design link language data to both GOLD and OLAC (Open Language Archives Community) with assistance from a university library system (Figure 2), thus addressing the need for both data accessibility and comparability.

1 As in *E-MELD 2005 Ontology FAQ*, we assume the working definition of an ontology as “essentially a machine-readable formal statement of a set of terms and a working model of the relationships holding among the concepts referred to by those terms in some particular domain of knowledge.”

2 By upper level ontologies (ULO) we refer not only to formal defining structures such as SUMO but also administrative ontologies or general ontologies of scientific resources which can help situate language data in a general universe of knowledge. By lower level ontologies (LLO) we refer to domain specific ontologies such as those that capture detailed linguistic analysis of specific language data.

Virtual Linguistics Laboratory (VLL) - The Components

Laboratory Methods	Lab Methods Manuals
Libraries of Comparable Data	Archiving of Words of World's Children A Relational Database Library of Audio-visual samples
Linguistic Tools	Data Collection Data Transcription Data Analysis Experiment Bank
Distance Learning Course Materials	Lab Methods Text Book and CD Rom Virtual Linguistic Library Methods of analysis Survey Course

Figure 1 – Virtual Linguistics Laboratory (VLL)

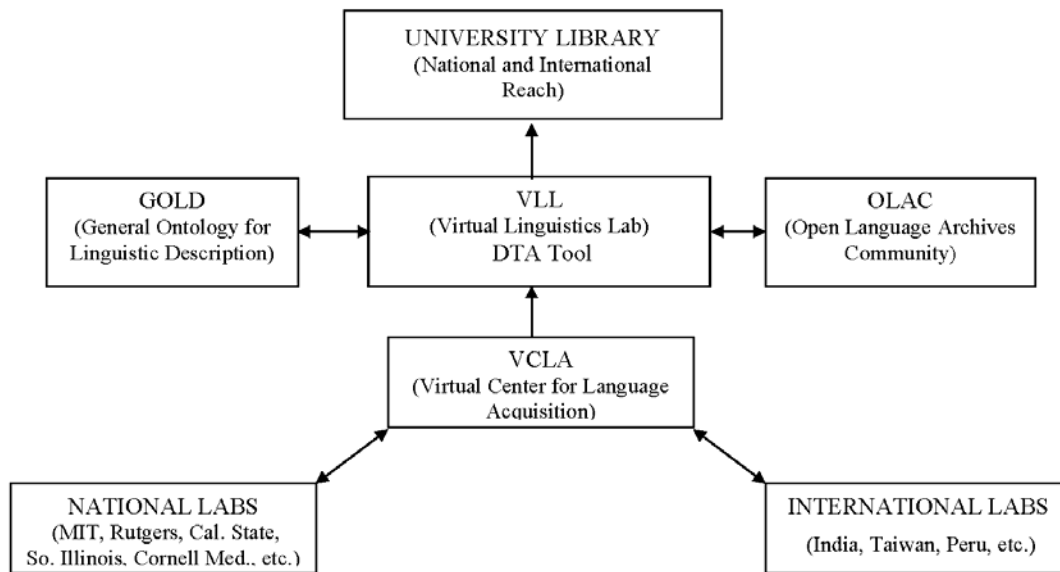


Figure 2 - Model of DTA tool cyberinfrastructure linkages

Although this model raises several issues regarding specific links to GOLD of a Data Transcription and Analysis system applicable to specific aspects of linguistic coding for language acquisition data, we will concentrate in this paper on the foundations for the linkages of this system to upper level ontologies and the associated role for the a university library.³

³ For example, with regard to LLO linkages, the DTA tool, after coding metadata regarding *subject* (speaker under study) and *session* of language data collection in a standardized format, leads the user to *initial linguistic codings* of the language data, which can calibrate subsequent linguistic analyses on these data, and then through an open-ended series of specific linguistic codings. Initial codings are designed to be theory neutral and to generalize across languages, while subsequent codings involve more theory specific hypotheses and assumptions. Initial codings require linguistically well motivated and shared concepts: e.g., what are basic possible ‘speech acts’ and how are they defined, what constitutes an ‘utterance,’ what constitutes a ‘sentence,’ what constitutes a ‘clause’ or a ‘word.’ Morphological glossing must be standardized to allow cross-subject and cross-linguistic comparability.

II. The role for the research library

The partnership between the VCLA and Cornell's Albert R. Mann Library has provided an important opportunity to explore the ways in which libraries can adapt their traditional roles in the management of information resources to aid in the exchange of online, electronic research data. The VCLA's intrinsic interest in wide-ranging collaboration between scholars of different disciplines coupled with its concern for best practices in information management make library collaboration natural. That the VCLA's data differ significantly from materials traditionally handled by the library creates unique challenges and opportunities for research. Funding agencies have expressed the need for ways to ensure that the results of research projects be shared and allowed to have influence beyond an immediate community of disciplinary colleagues.¹ Not only can the development of domain-specific ontologies like GOLD inform and assist these efforts in important ways, but the involvement of theory- and discipline- neutral institutions like libraries may also help speed the deployment and adoption of such ontologies.

Discovery, access, and preservation

Research libraries have long had responsibility for selecting important scholarly resources and facilitating their long-term preservation and accessibility. It is relatively easy to visualize the processes involved for traditional physical resources. Catalogers assign controlled subject headings to facilitate exhaustive searches, use of standardized classification systems ensures that related works are shelved together for easy browsing and serendipitous discovery, loose journal issues are bound in order to increase their longevity, brittle books are microfilmed, and thoughtful collection development policies ensure that results of research that may be unpopular today is available to inform the scholars of tomorrow. These same principles are no less important in the digital domain, but it is less clear how to implement them. Libraries have embraced the need for metadata specialists to help organize and describe the vast array of emerging electronic material. The digitization of analog resources presents many challenges, including the definition of technical best practices, the provision for long-term storage, and the navigation of a myriad of copyright and privacy issues. In addition, the popularity and ease of use of internet search engines such as Google challenges the ways libraries deliver services and makes it less obvious to the public why libraries are even necessary.

Example of ontology use: a virtual life sciences library

Cornell University, being an administratively complex institution with statutory and endowed colleges and several campuses worldwide, has a similarly complex distribution of life sciences research activities. To support the effort to integrate these activities under the umbrella of a new life sciences initiative, a Life Sciences Working Group in Cornell University Library was formed in 2003 to develop, among other goals, an integrated Web presence for library resources and services relevant to life sciences.² The committee recognized the need to transcend individual services and staff expertise in ten unit libraries to create a sense of "our library" within the life sciences community, and set out to craft an online information service that offers the simplicity of use of Google, and, more importantly, highlights the interconnections among all of the stakeholders in a vibrant academic research community.

The resulting service, named "VIVO" (vivo.library.cornell.edu), is implemented using metadata in locally-developed MySQL database persistence layer served by Java servlets and Java Server Pages. A single table stores all of the entities, each of which is an instantiation of a class type, while additional tables store the properties and relationship types possible in the VIVO ontology. The ontology involved is an outgrowth³ of the Harmony Project's ABC Ontology³ and

the AKT ontology developed by Advanced Knowledge Technologies,⁴ with additional classes and properties to model the activities of the life science community. VIVO functions as an “administrative ontology.” Rather than attempt to define and model the specific intellectual concepts involved in researchers’ activities, VIVO models more tangible relationships such as authorship and affiliation that associate related records without applying intellectual classifications that might be more arbitrary and controversial. For example, relationships can associate a person as faculty member in a graduate field or discipline, or as author with a recent journal article, or as principal investigator with a grant. The relationships represent “isness” more than “aboutness,” and by simple inferencing enable aggregating across multiple individual relationships to report the article titles published by a department. A user makes her own judgments about how the individual or collective entities and relationships relates to her own activities or interests. At a more basic level, the ontology also allows search results to be grouped into categories such as semester courses, journal article citations, and academic department names. Users can then browse those categories and traverse the links therein, offering a visible structure and context for information often not present in results from search engines.

Such a use of ontology is quite different from the type of complex formal modeling involved in systems designed for heavy inferencing from and comparison between detailed datasets; libraries are unlikely to have the personnel or mandate to get heavily involved at such a level. It seems apparent, however, that the interdisciplinary collaboration activities of organizations such as the VCLA would benefit from a multiple-level approach that allows for the interaction of ontologies.

Cascading triangles: multiple levels of ontology and access

In 2004 the Cornell Language Acquisition Lab (CLAL) and Albert R. Mann Library, in collaboration with VCLA members at MIT, began to explore the issues of organizing, formatting, preserving, and sharing the VCLA’s data and how the concepts might be extended to other communities of practice, under a Small Grant for Exploratory Research (SGER) from the National Science Foundation. We took the VIVO model as a starting point for applying ontology to the problem. The VIVO principle of using an ontology structure to organize the presentation of search results while remaining as neutral as possible about scientific theory seemed like a logical way to address the problems of relating the data content (recordings) to data analysis (marked-up transcriptions) and to begin examining issues of data structure display. We also realized that the VCLA project would involve more “levels” of information than VIVO. While VIVO tends to relate research groups and initiatives that are physically or politically separate but similar in function or scope, and with clear administrative divisions, the VCLA needs to deal with more forms of collaboration between different participating “co-laboratories,” each of which contributes to and is aided by the Virtual Center. In turn, the interdisciplinary VCLA should be able to make certain kinds of information about its activities known to the broader disciplines across which it operates: linguistics, developmental psychology, and brain science. An online search tool would need to serve the interests of diverse groups of users, each of which would desire to be presented initially with different levels of abstraction, and with the ability subsequently to “drill down” further toward primary data in topics of interest. We formulated a visual diagram of successive vertical tiers of triangles (Figure 3), signifying the fact that each tier desires to see a certain select set of metadata from the tier below, while masking the main body of primary data from appearing in initial search result screens.

SGER Conceptual Framework

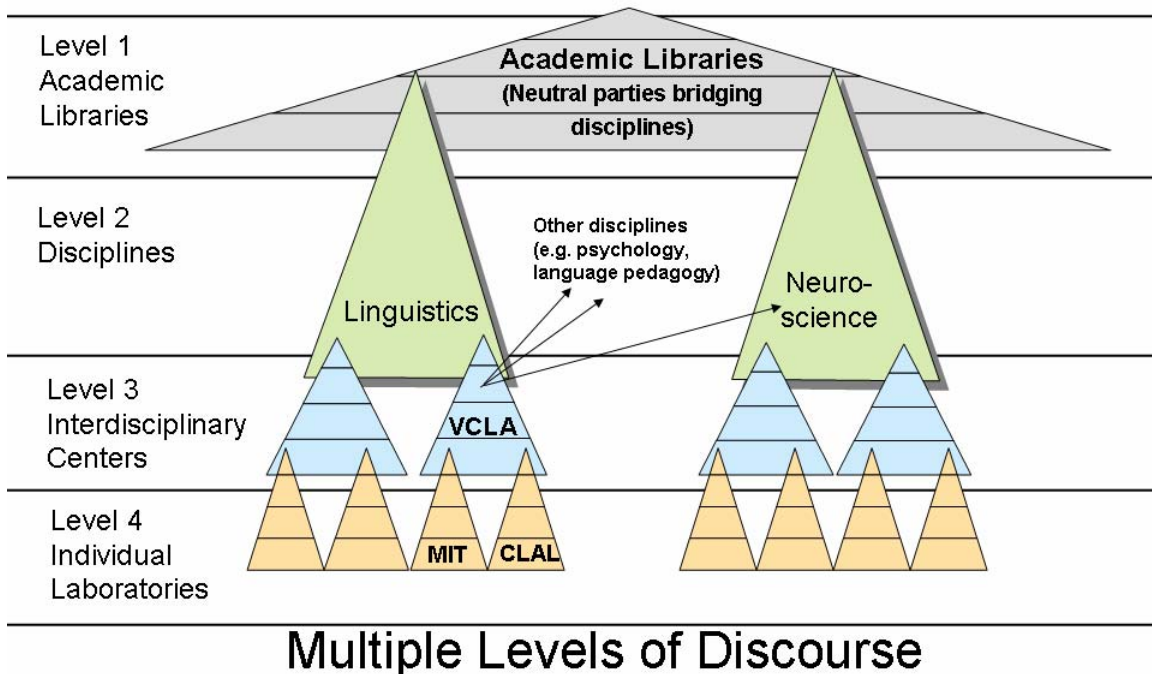


Figure 3 - NSF SGER Conceptual Framework

For example, at the lowest level, laboratories will generate very detailed GOLD-type markup of experiment session transcriptions. SQL databases or XML-serialized files of these data will reside on local laboratory or library servers, or in general repositories such as DSpace.⁵ In the short run, much data will exist only in paper transcriptions, and accompanying analog tape recordings may continue to await digitization. There may be privacy or copyright issues to be sorted out on a case-by-case basis before another researcher can get direct access to the primary "data content." There may also be issues about what, if any, data content is truly "primary" and what has been artificially influenced by the research process and how.

The important thing is that some general set of metadata about the specific contents of a local collection "bubble up" to the level of the virtual center so that such negotiations may take place. In addition to aggregating metadata from their participating laboratories and research groups, virtual centers such as the VCLA will generate their own "primary data" (e.g. the VLL) in the form of best practices documents and software tools, as well as information about the theories tested by, and results of, different collaborative experiments, and forms in which such results are published. Metadata about the existence of these resources (the Open Language Archives Community (OLAC)'s "data," "tools" and "advice"⁶) will not only need to be harvestable using OAI-based systems⁷ such as OLAC, but also percolate up to the level of searches by the broader academic community, accompanied by a further-abstracted view of the primary data in order to aid in the forging of new collaborative partnerships and to offer the opportunity for someone who is not a specialist in the native domain of data to discover, reuse, or repurpose those data. Such an idea is quite different from many familiar Web searches, which rely on the spidering and caching of large amounts of document text in order to build keyword indexes of the actual content. This is not practical for our purposes, and in the case of primary data in the form of audio waveforms,

value between the dataset and its various component concepts. The script could also add entities to the search database for certain common lexical items or a human editor could add keywords and descriptions manually. Another critical component would be links to more standardized and less controversial concepts in the VIVO fashion, including concepts like “recording,” “session,” “experiment,” or “hypothesis,” and controlled language codes such as the SIL list. Links to different types of experimental “tasks” (such as elicited imitation, etc.) would also be made here, to aid in the future replication of experimental results. Such concepts would be standardized by a unit such as the VCLA through its DTA tool.

Such a process could occur any number of times as competing transcriptions, interpretations, and annotations are advanced about the same primary data; it is crucial that the successive refinements of data analysis not muddy the search results or obscure access to the primary data content. While traditional “full-text” style searches have a tendency to present somewhat unhelpful lists of results when confronted with multiple very similar documents, a set of competing transcriptions of a field recording could be quite useful when handled through an intermediating ontological framework. By deemphasizing the actual markup streams and retaining instead a network of relationships and relevance values to indicate the general ways in which a set of primary data relates to ontology-modeled concepts, multiple iterations of markup would add only minimal new information into the search tool's database. In the case that a researcher hears a segment previously untranscribed, the new metadata would help forge new upper-level links while leaving the existing relationships unchanged. This would ideally serve to enhance rather than dilute search results. Database entries allowing the free, ever-evolving mapping of GOLD concepts such as “clitic” or “simpleSyntacticWord” to general entries for subject areas such as “morphology” and “syntax,” or “morphosyntax” would begin to evolve a network of semantically-charged keywords. Collaborators focusing on one of these linguistic subcategories and on specific types of experiment tasks could then discover datasets or recording collections likely to be useful in comparative studies and be directed to laboratory websites or servers in order to gain access. Similarly, concepts like “orderingRelation” might acquire interesting links in the database, perhaps to aid programmers uninterested in the intricacies of morphosyntactic theory in finding language data that has been segmented into units of meaning. At even higher levels of abstraction, combinations of broad concepts like “language” or a specific language name, with “sound,” might be of interest to physiologists or language teachers who may wish to “drill down” to primary resource collections and put them to good use in ways never envisioned by their curators. As domain-specific ontologies continue to be linked with upper-level merged ontologies (such as GOLD with SUMO), dynamic taggings and mappings of abstract concepts might help researchers discover new opportunities for collaboration between disciplines. (See Figure 4.)

III. Conclusions

We have suggested that linking lower-level ontologies to administrative ontologies and multiple types of upper merged ontologies can have multiple benefits, including the exposure of research data to interested parties outside the immediate community of linguists, and have suggested some ways to approach such ontology linking. These ideas are of course still very theoretical, and as the project progresses we will test methods of implementation. There is, at least, a logical role for institutions like libraries, uninvolved with the specifics of data analysis, to develop tools to make use of domain-specific ontologies even while they are in formative, imperfect, or contested states. While codification, as much as possible, of shared linguistic concepts into a common ontology—such as pursued by GOLD—remains our quest, lack of

perfect interoperability or consensus on concepts need not be an impediment to the goals of data preservation and access.

Works Cited

1. National Science Board, “Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century” (draft), March 30, 2005.
2. Life Sciences Working Group, Cornell University Library, “Report of the Life Sciences Working Group Cornell University Library,” September 2004, <http://www.library.cornell.edu/staffweb/Life%20Sciences%20Report.pdf> (17 June 2005).
3. Carl Lagoze and Jane Hunter, “The ABC Ontology and Model,” <http://jodi.ecs.soton.ac.uk/Articles/v02/i02/Lagoze/> (15 Mar. 2005).
4. Advanced Knowledge Technologies, “AKT Reference Ontology,” <http://www.aktors.org/publications/ontology/> (15 Mar. 2005).
5. DSpace Federation [home page], <http://www.dspace.org/> (17 June 2005).
6. Steven Bird and Gary Simons, “OLAC Overview,” <http://www.language-archives.org/docs/overview.html> (17 June 2005).
7. Open Archives Initiative (OAI), <http://www.openarchives.org/> (15 Mar. 2005).