First Steps in Transforming the Primary Research Process through a Virtual Linguistic Lab for the Study of Language Acquisition and Use: Challenges and Accomplishments

María Blume†
University of Texas at El Paso
Department of Languages and Linguistics
Liberal Arts Bldg., Room 119
El Paso, TX 79968
1-915-747-6320
mblume@utep.edu

Barbara Lust†
Cornell University
Department of Human Development
G57 Martha Van Rensselaer Hall
Ithaca, NY 14853
1-607-255-0829
bcl4@cornell.edu

ABSTRACT

This project involves both the development of a community of scholars committed to cross-institution, interdisciplinary and cross-linguistic collaboration (a Virtual Center for Language Acquisition, VCLA) and the creation of a web-based infrastructure through which a new generation of scholars can learn concepts and technologies empowered through this CI environment. These technologies, constituting a Virtual Linguistic Lab (VLL), provide the student with the structure for data creation, data management and data analysis as well as the tools for collaborative data sharing. This infrastructure, informed and executed through computational science, involves the coherent integration of an open web-based gateway (The VCLA website), linked to a specialized web-based VLL portal which includes not only real world examples and visualizations of data creation and analyses, but several cybertools by which these data can be managed and analyzed. This infrastructure subserves both the beginning student and the researcher pursuing calibrated methods and structured data sharing for collaborative purposes. Students continually engage in the development of the cybertools involved and in the scientific method involved in primary research. In this paper we summarize our objectives, the challenges we face and the solutions we have developed to these challenges. At this point, the project is just completing an implementation stage, and the first steps in creating a virtual community of practice, and is being readied to move to a diffusion stage.

Categories and Subject Descriptors

J.5 ARTS AND HUMANITIES. *Linguistics*

General Terms

Management, Documentation, Design, Experimentation, Security, Human Factors, Standardization, Languages, Theory, Legal Aspects.

Keywords

Community of practice, language acquisition, language use, language documentation, research, education, database, data management, standardization.

1. INTRODUCTION

Recent developments in cyber-infrastructure offer new possibilities to scientists for advancing research questions and methods [2, 4, 5, 11, 12, 13, 25, 29, 30, 33], opening possibilities for interdisciplinary collaborative research and empowering cross-linguistic and cross-cultural research in a global perspective. These developments can empower the study of the language sciences as they have empowered other areas of science.

However, these new possibilities challenge the field of linguistics and the language sciences to develop (1) an infrastructure of collaboration that will allow us to create a virtual community of practice [2, 3, 4, 12, 34, & 38]; (2) standardized tools and best practices which can be shared while at the same time allowing unique methods by individual researchers; (3) infrastructure for data storage, management, dissemination and access, including means for interfacing databases that differ in both type and format [5, 13, 15, 25, 31, and 32]; (4) preservation and 'portability' of data and related materials [6 & 37]; and (5) changes to the ways in which we educate our students and train new researchers in scientific methods.

As noted by King [24] for the social sciences:

> "The potential of the new data is considerable, and the excitement in the field is palpable. The fundamental question is whether researchers can find ways of accessing, analyzing, citing, preserving, and protecting this information." (p. 719)

Our purpose in this project is to train students in new methods of primary research which exploit new cyberinfrastructure-enabled possibilities to collect and manage complex data in a collaborative scientific environment and to develop cyber-infrastructure for documentation and accessibility of an ever-growing set of shared complex data, allowing data use, reuse and repurposing.

To this end, we created a new Virtual Learning Environment for the language sciences, through development of a Virtual Linguistics Lab (VLL).[1]

[1] http://clal.cornell.edu/vll

† With the collaboration of the founding members of the Virtual Center for Language Acquisition

In section 1 we introduce our project, audience and objectives. In section 2 we describe our multiple interrelated challenges. In section 3 we present the components of the VLL and then in section 4 we explain how they approach solution to the challenges we face. In section 5 we summarize our educational achievements to date. Section 6 describes the broad impact of the project. Section 7 presents future challenges and lessons learned. A description of the VLL especially with regard to its role in language documentation can be found in Lust et al. 2010 [27]. A description of cybertool development in the VLL can be found in Blume et al. 2012 [8]. Here we focus on the educational mission of our project.

1.1 Audience

Our program mobilized faculty from a new community of practice across eight diverse US Universities and one initial international extension (Peru).[2] Project members were interdisciplinary VCLA founding and contributing members who are linguists, developmental and cognitive psychologists, and neuroscientists.[3] Members come from different fields, institutions and countries.

Courses involved in this project were addressed to undergraduate and graduate students from linguistics, psychology, human development, and computer sciences and to researchers across the world wishing to collaborate on shared data and/or learn best practices for scientific methods in the study of language acquisition and use.

1.2 Objective

We seek to "transform the primary research process" by providing a systematic infrastructure and cybertools to foster and support scientific collaborative data collection and management starting from the initial stages of a research project and throughout to its report of results.

Thus, this project seeks to educate a new generation of interdisciplinary undergraduate and graduate students and future researchers –with diverse geographical and cultural backgrounds– who will gain a solid formation on language documentation, standardization and cybertool use through our courses[4]. It also seeks to support interdisciplinary researchers[5] interested in international and cross-institution collaboration that need to create and share data but do not have the means or training to do so. As they and their students use our virtual learning environment they can both give us feedback and later train other researchers or students at their institutions.

2. CHALLENGES

Our project faced several challenges related to education (2.1.), data complexity (2.2.), and cultivation of researcher collaboration (2.3).

2.1 Educational Challenges

2.1.1 Interdisciplinarity of the language sciences.

The major questions in Cognitive Science — *is the brain programmed for language knowledge and acquisition? What are the universals of language structure? What is innate and what is learned with regards to language? How is new linguistic knowledge developed over time?* — require us to be able to study all languages (of which 6,000-7,000 have been estimated) and all developmental stages of language acquisition. Language acquisition is therefore, by its very nature, a multidisciplinary area, which must be studied by linguists, developmental psychologists, educators, language pathologists, human development researchers, and computer scientists who have means for collaboration. Both researchers and students need to be able to collect, analyze, and compare large amounts of cross-linguistic data in interdisciplinary forms (e.g., brain images and language utterances)[6]. Our scientific enterprise thus requires cross-institution and international collaboration, in addition to a well-designed computational platform for its development.

2.1.2 Language documentation and data management

Students need training to manage language data in a scientifically-sound way. They must also be taught methods of data sharing. For example:

> "Finally, universities and individual disciplines need to undertake a vigorous programme of education and outreach about data. Consider, for example, that most university science students get a reasonably good grounding in statistics. But their studies rarely include anything about information management —a discipline that encompasses the entire cycle of data, from how they are acquired and stored to how they are organized, retrieved, and maintained over time. That needs to change: data management should be woven into every course in science, as one of the foundations of knowledge." [15][7]

This needs to be accompanied by education in a culture of collaboration and data sharing, as highlighted by King [24]:

> "[…] More importantly, when we teach we should explain that data sharing and replication is an integral

[2] MIT, Boston College, Rutgers University at New Brunswick, Rutgers University at Newark, California State University at San Bernardino, Southern Illinois University at Carbondale, Cornell University, University of Texas at El Paso, Pontificia Universidad Católica del Perú.

[3] FOUNDING MEMBERS: Suzanne Flynn (MIT), Claire Foley (Boston College), Liliana Sánchez (Rutgers University, New Brunswick), Jennifer Austin (Rutgers University, Newark), YuChin Chien (California State University at San Bernardino), Usha Lakshmanan (S. Illinois University at Carbondale), Barbara Lust, Claire Cardie, James Gair, Marianella Casasola, and Qi Wang (Cornell University), María Blume (University of Texas at El Paso), and Elise Temple (NeuroFocus). CONTRIBUTING MEMBERS relevant to this project: Jorge Iván Pérez Silva (Pontificia Universidad Católica del Perú), Gita Martohardjono (CUNY Graduate Center/Queens College), Cristina Dye (Newcastle University), Yarden Kedar (Ben Gurion University at the Negev).

[4] The courses had different learning objectives since they were taught at different institutions on different semesters; some focused on first language acquisition, others on bilingual acquisition, and one was focused on the acquisition of Spanish. However, they all incorporated this learning objective as a main objective.

[5] At this point, faculty at the nine institutions that helped us develop this project.

[6] See section 2.2 below.

[7] See also [1].

> part of the scientific process. Students need to understand that one of the biggest contributions they or anyone is likely to be able to make is through data sharing".

The ability to manage and share complex data, in its turn, depends on students being trained in basic computational skills needed for data management and analysis.

2.1.3 *Student background*

Students interested in language acquisition come from different fields, and may come to our courses without much of the necessary background. In particular, students from psychology, human development, and computer science need to learn linguistic theory and terminology; students from linguistics, human development and computer science need to learn about research design, and experimental methods in developmental psychology; students from psychology, human development, and linguistics need to receive basic training in computer science in terms of using and understanding complex databases, as well as in basic computational thinking to be able to create their own searches in the database; computer science students need to learn to apply their computing skills to language data. All students need to be trained in transcription of language data and in conducting basic linguistic analyses of natural language. Computer science will be necessary both for modeling and analyzing large data sets, but also for the students' contribution to the development of cybertools themselves.

2.1.4 *Research with human subjects*

To conduct research on natural language students require extensive training to work with human subjects and access to human subjects is tightly controlled. Students must be taught procedures to ensure confidentiality and informed consent that are set by individual Institutional Review Boards in conjunction with new mandates by federal funding agencies (e.g., National Institutes of Health (NIH).[8] Students must be taught that all records regarding human subjects must become part of the complete language documentation process.[9]

2.2 Data challenges

2.2.1 *Data Complexity*

In the fields of language acquisition and use, data are multi-linguistic, multi-modal (i.e., audio, video, transcripts in different formats, etc.), multi-formatted, and derive from multiple methods of data collection (i.e., observational and experimental, cross-sectional or longitudinal). In addition, they involve multiple aspects of data provenance (e.g., age and/or developmental or cognitive stage of speaker, social and pragmatic contexts, culture). These features result in a complex set of databases. The scientific use of any single record requires access to many levels of data, ranging from raw (establishing provenance) to structured and analyzed data (establishing intellectual worth).[10] The computational science necessary to accomplish analyses and interoperability over representations of such large, diverse and expanding data sets is challenging to students and researchers not trained in computer science.[11]

Various linguistic theories are invoked for data description and analysis, creating a need to interface theoretical vocabularies. The variety of languages needs to be described in language typology, while we search for language universals by the creation of uniform formats for cross-linguistic comparisons.[12] Audio or audiovisual samples provide the authoritative archival form of language data creating technical challenges [23]. Generating transcriptions of language requires a time consuming, cognitive and analytic process with variation expected across individual transcribers [16 & 17]. At every moment, different points of data creation must be linked and sound methods of data documentation must be applied. Language data collections are infinitely expandable and should be merged, used, reused and, when possible, repurposed. Continual data-driven computation and statistical analysis is required as is theoretical modeling through computational methods.

2.2.2 *Data Documentation and Standardization*

These features of language data result in a complex set of databases often appearing in diverse formats as different labs generally practice distinct forms of data collection and management. Therefore, there is a need in the field for standardization of data collection, labeling and storage methods that will allow for preservation and portability of such data.

Once the data are collected, researchers must develop ways to link diverse data sources, calibrate them and make sure they are subjected to the same reliability standards so that data can "speak to" data" [25; see also 5, 13, 30, & 32].

There is also a need for databases that provide access to all the information related to a project, from PI information, to project design, batteries, results, as well as the actual data from each subject. These background data are fundamental both as a teaching tool and as a prerequisite for data reliability and researcher collaboration. Data must be described and preserved with systematic and significant metadata, which include general concepts recognized across fields and linguistic concepts for specific inquiries (see [27] for further description of issues related to language documentation).[13]

2.2.3 *Data Management and Dataset Design*

Data must be stored so that relationships can be discovered within and across data sets. The more each data singleton can be significantly connected or "interlinked", the more powerful and useful it becomes [5 & 13]. Such links can be cross-disciplinary (e.g., connecting brain images with behavioral experimental results testing language comprehension or production), or specific

[8] http://grants2.nih.gov/grants/policy/data_sharing/

[9] In addition to collecting data and comparing data on multilingual populations, students and researchers need to be able to determine whether the multilingual populations of any two different projects are homogeneous [22], since numerous variables affect a speaker's language knowledge, dominance, and patterns of use.

[10] For example, data from more than 20 languages and cultures and thousands of child subjects[10] and adults exist in the Cornell Language Acquisition Lab and Virtual Center for Language Acquisition alone.

[11] See the introduction and papers collected in [13], for other efforts regarding open data in linguistics aided by a strong computational component.

[12] This challenge is being confronted by the General Ontology for Linguistic Description (GOLD) project [21] in the Electronic Metadata for Endangered Languages Data (EMELD) enterprise [19].

[13] See also [6, 31, & 32].

to linguistics. Data from any one language must be comparable to that in another if one pursues a hypothesis concerning linguistic universals or variation linked to language typology.

An effective data management infrastructure must not only provide a powerful database that can handle both experimental and naturalistic data, but, at the same time, it must structure the primary data creation process from its initial stages, providing a way to represent new data so that it can be analyzed subsequently in a standardized and theory-neutral way which ensures data comparability. At the same time, this representation must allow researchers to create theory-specific coding screens allowing multiple types of analyses in their own data or linking data across projects.

2.3 Cultivating Researcher Collaboration

2.3.1 Researcher training

The scientific study of language acquisition and use not only requires researchers to conduct field work and collect data according to sound scientific methods but also to manage international collaboration and to be trained in shared principles of data documentation, database use, and collaboration through cybertools. Researchers need a resource that allows them to compare data across datasets and projects, and to reuse previously collected data. As noted in *Nature Biotechnology* [14]

> "[…] More often, though, a failure to share simply reflects the considerable time and effort associated with formatting, documenting, annotating and releasing data. In this regard, the availability of new tools, […] should prove helpful"[14]

2.3.2 Intellectual property

Finally, intellectual property rights must be addressed in the case of language data as for research data in general. Language data painstakingly collected and created by individual scientists belongs primarily to the researcher and to the institution in which they work. Principles for sharing data or scientific materials must be developed in a manner that respects this premise [1, 3, 7, 8, 11, 12, 14, 15, 18, 27, 33, 34, 35, 38, & 39]. Such agreements must also become part of comprehensive language documentation where language is to constitute scientific data.

2.3.3 Cross-institutional and international collaboration

For active researcher collaboration to expand, our academic institutions need to standardize the ways in which IRB permissions are complied with across institutions. IRBs ordinarily do not have common rules and common standards for cross-institution research. In some countries, comparable IRBs do not exist.

3. COMPONENTS

Our project pursued solution to these challenges by building an infrastructure that includes two main components integrating a VCLA[15] with a Virtual Linguistics Lab whose elements are provided through a structured VLL portal which can be used in both synchronous and asynchronous courses collaboratively across institutions.

3.1 The Virtual Center for Language Acquisition (VCLA)

The VCLA unites a series of research labs across the country and the world. A set of founding members collaborated to build an infrastructure for its mission: to foster collaborative research among scientists working in the area of language acquisition, collaborations which are potentially interdisciplinary, which may be at a distance geographically and which may involve the comparative study of multiple languages, interactions on shared data, and a variety of lab methods.

3.2 The Virtual Linguistics Lab (VLL)

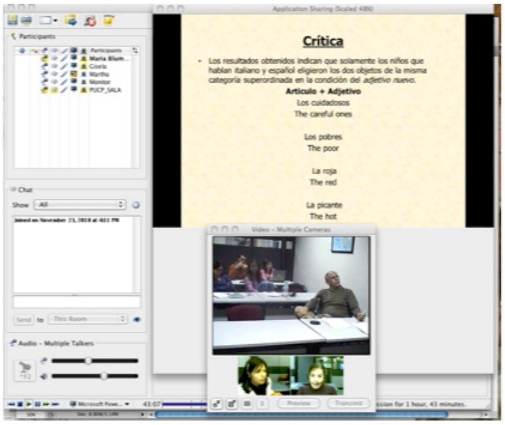
The VLL portal,[16] which is now accessible in English and Spanish, provides structured access to the components of a virtual linguistic lab, which are:

- Materials constituting a series of web-based courses, integrating synchronous and asynchronous forms of interactive information distribution that teach them the specific procedures for investigating language knowledge. Each topic contains:
    - PowerPoint presentations that can be used in class or for review.
    - Audio/visual samples that may be used as part of the lessons to demonstrate a particular method or issue in language acquisition.
    - Published or unpublished papers.
    - Specific exercises/homework, linked to the audio/visual materials so that students can practice data transcription and analysis with real research examples.
- A Research Methods Manual [10] explicating best practices for the scientific study of language acquisition. It provides students and researchers with standard methods for data collection as well as with the background knowledge the DTA tool presupposes.
- A glossary [26] as part of the general Research Manual further helps students from different fields learn the terminology and concepts of the other fields.
- A set of materials to assist in data collection, data management and data analyses, e.g., a multilingualism questionnaire for assessment of degree and nature of multilingualism.
- A series of web conferences (as exemplified in figure 1) through which students can participate in discussions with students and researchers at other institutions synchronically during courses, or later review recordings of these conferences asynchronically.
- A discussion board, with a blog and a wiki, to share ideas and post assignments, presentations and research.

[14] See also [24 & 28]

[15] http://vcla.clal.cornell.edu

[16] Programming for the VLL portal was created by Tommy Cusick, then a Cornell undergraduate student in computer science, now at Google.



Crítica

- Los resultados obtenidos indican que solamente los niños que hablan italiano y español eligieron los dos objetos de la misma categoría superordinada en la condición del *adjetivo nuevo.*

Artículo + Adjetivo

Los cuidadosos
The careful ones

Los pobres
The poor

La roja
The red

La picante
The hot

Figure 1. Webconferences join institutions for discussions.

In addition, the VLL portal provides access to cybertools developed to structure the primary research process in the area of language acquisition and use.[17] These cybertools are:

- The Data Transcription and Analysis Tool that provides a structured interface for metadata and data collection, [8]. It not only guides researchers and students in the primary research process but also results in a web-based calibrated database of continually expanding cross-linguistic data, and an Experiment Bank.
- Virtual Workshops, and a technical user's manual [9] provide training for students on the use of the DTA tool and the Experiment Bank.

These materials are integrated into a university-supported cyberinfrastructure to underwrite potential courses and to ensure the high availability needs of a distance-learning program.

4. MEETING THE CHALLENGES

4.1 Educational Challenges

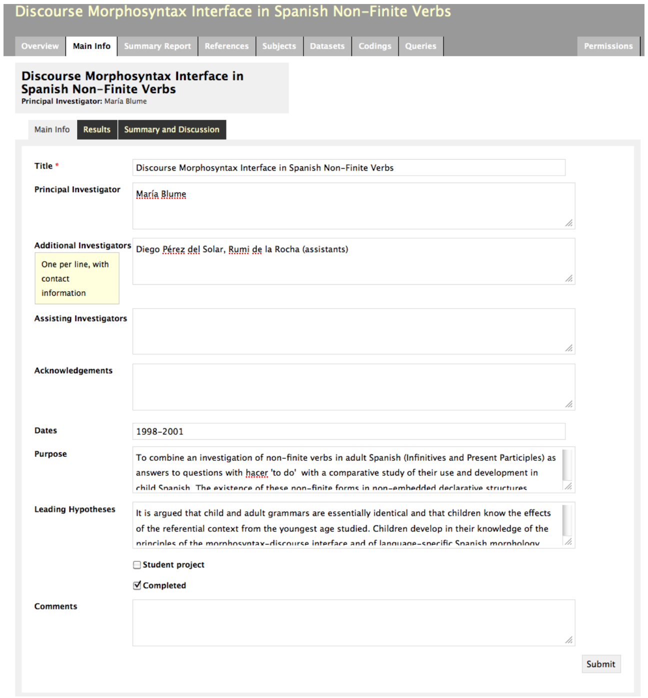
4.1.1 Interdisciplinarity

All VLL materials are developed to serve interdisciplinary access. Students across disciplines are encouraged to collaborate in original research projects, once they complete training, and they read about and discuss the challenges of collaboration [1, 2, 11, 12, 14, 18, 27, 28, 33, 34, 35, & 39]. Through the VCLA website, which lists projects by interdisciplinary VCLA members in order to give undergraduate and graduate students and other researchers ideas for future research and collaboration, our students have access to researchers and projects from outside their institution, country, and discipline, who they can contact for advice, or collaboration.

The students meet with interdisciplinary students and professors at other institutions through Elluminate/Blackboard Collaborate.

4.1.2 Language documentation and data management

Our VLL materials provide students with lessons on scientific methods and best practices for data collection and management in primary research and provide web available tools for learning and practicing these. In particular, the DTA tool guides students and/or researchers through the steps of data creation and management, including metadata representation (cf. 4.2 below). Through its Experiment Bank component, it collects all information related to a study (experimental or observational) in the same location, and makes it available to researchers seeking to replicate it, criticize it or consult it when reading a particular scientific research paper reporting an experiment's results. The first sets of screens in the DTA tool guide students and researchers to save/access the metadata information for a research study, thus providing an entry in an Experiment Bank. Main areas include project investigators, purpose and leading hypotheses, subjects, and results and discussion. The DTA tool also tracks publications, presentations, related studies, and bibliography related to a research project. At several different points, documents can be attached. Figure 2 shows the first screen a researcher completes when starting a new project.



Discourse Morphosyntax Interface in Spanish Non-Finite Verbs

Discourse Morphosyntax Interface in Spanish Non-Finite Verbs
Principal Investigator: María Blume

Title *	Discourse Morphosyntax Interface in Spanish Non-Finite Verbs
Principal Investigator	María Blume
Additional Investigators One per line, with contact information	Diego Pérez del Solar, Rumi de la Rocha (assistants)
Assisting Investigators	
Acknowledgements	
Dates	1998-2001
Purpose	To combine an investigation of non-finite verbs in adult Spanish (Infinitives and Present Participles) as answers to questions with hacer 'to do' with a comparative study of their use and development in child Spanish. The existence of these non-finite forms in non-embedded declarative structures
Leading Hypotheses	It is argued that child and adult grammars are essentially identical and that children know the effects of the referential context from the youngest age studied. Children develop in their knowledge of the principles of the morphosyntax-discourse interface and of language-specific Spanish morphology
	☐ Student project
	☑ Completed
Comments	

Figure 2. DTA Project Information

4.1.3 Student background

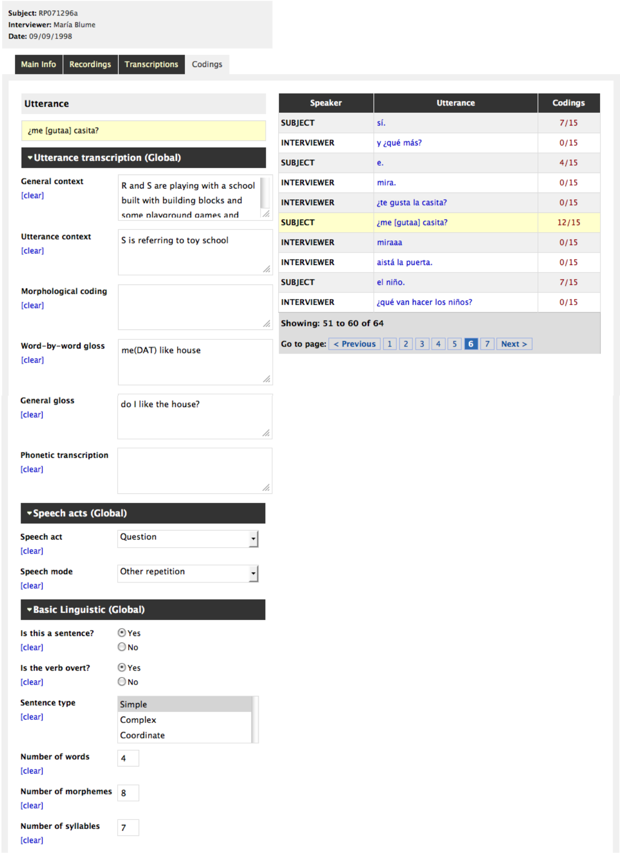
Students in need of linguistic background, get it through our course presentations, readings, manual, and glossary. Students without background in psychology or experimental methods receive training in research design and particular methods for data collection through our learning topics dedicated to basic concepts on scientific research, and through several others dedicated to particular methods. Students also read relevant chapters of our VCLA Research Methods Manual [10] and assigned papers, use the Experiment Bank component of the DTA tool to see detailed examples of previous research, do assignments that have them extract the research design of a published paper and enter it in the DTA tool, and complete a final project in which they are required to design their own study. Finally, interactive assignments teach

[17] These tools will be explained in more detail in section 4 when we describe their educational and research features.

all students to transcribe (cf. 4.2.2.1), reliability check[18], and analyze previously collected language data, using the DTA or the original project's format. For example, in the Elicited Imitation learning module in the VLL, students get access to information on this research task and to a set of research articles using elicited imitation as their primary method. In the assignments they have the option to focus on different projects related to the articles they read. They can then see samples of the original session recordings for the projects which they can score using systematic scoring guidelines and materials of the project. Then they can compare their own results to those reported in the article. They can also look at all the details of the relevant research project by looking at the project information in the DTA Tool. This hands-on experience with real research material is fundamental to help students understand all steps of research development, from creating a researchable question to designing a research project, to collecting data, testing hypotheses, and relating results to previous research.

The DTA tool provides coding sets that train students in first steps of linguistic analysis in a theory-neutral design.[19] Students and researchers coding natural speech data are expected to use all these basic codings, so that the data are calibrated across projects. Figure 3 exemplifies basic coding of an utterance of natural speech data of a Peruvian monolingual Spanish-speaking child from the "Spanish Natural Speech Corpus-Blume", such as the ones that are coded by our students in their assignments. Such codings render the data ready for further analyses in connection with specific research questions.

In this way, the student or new researcher can create their own dataset and begin asking questions regarding how the child acquires the knowledge of question formation.



Subject: RP071296a
Interviewer: María Blume
Date: 09/09/1998

Utterance

¿me [gutaa] casita?

▾Utterance transcription (Global)

General context [clear]: R and S are playing with a school built with building blocks and some playground games and

Utterance context [clear]: S is referring to toy school

Morphological coding [clear]:

Word-by-word gloss [clear]: me(DAT) like house

General gloss [clear]: do I like the house?

Phonetic transcription [clear]:

▾Speech acts (Global)

Speech act [clear]: Question

Speech mode [clear]: Other repetition

▾Basic Linguistic (Global)

Is this a sentence? [clear]: ● Yes ○ No

Is the verb overt? [clear]: ● Yes ○ No

Sentence type [clear]: Simple / Complex / Coordinate

Number of words [clear]: 4

Number of morphemes [clear]: 8

Number of syllables [clear]: 7

Speaker	Utterance	Codings
SUBJECT	sí.	7/15
INTERVIEWER	y ¿qué más?	0/15
SUBJECT	e.	4/15
INTERVIEWER	mira.	0/15
INTERVIEWER	¿te gusta la casita?	0/15
SUBJECT	¿me [gutaa] casita?	12/15
INTERVIEWER	miraaa	0/15
INTERVIEWER	aistá la puerta.	0/15
SUBJECT	el niño.	7/15
INTERVIEWER	¿qué van hacer los niños?	0/15

Showing: 51 to 60 of 64

Go to page: < Previous 1 2 3 4 5 6 7 Next >

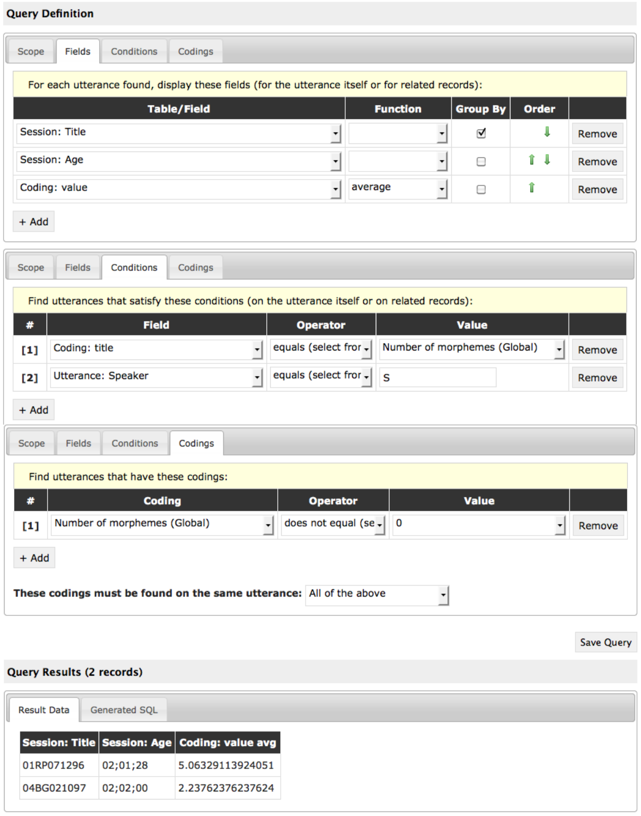
Figure 3. Basic linguistic coding screen.

A set of basic queries, which is essential to calibrating language data, is available in the DTA tool. Queries can be run on all sessions that have been coded for the relevant features in all projects in the DTA tool, thus linking across sessions, subjects, and projects. These basic queries are common queries in the field that are ready-made for the researcher, and also serve as teaching examples for students who can use them to complete assignments on selected samples of language data. The query screens are designed to guide the student or researcher through the different necessary steps to do the computation that would answer their search question. For example, students are asked to compute the subject's Mean Length of Utterance or MLU, a common measure of a child's language development, and compare the MLU they find with the MLUs for the different developmental stages reported for the subject's language. To do this they are required to code manually the number of morphemes in each of the utterances produced by the subject. Then they run a simple query that adds the utterance counts (looking only at the utterances produced by the subject) by the total number of utterances produced by the subject. Figure 4 shows one such query run on two children of comparable age from two different corpora[20]

[18] Reliability checking is the process by which a researcher's transcription or coding is cross-checked with that of another researcher to establish its reliability.

[19] These basic linguistic codings (analyses) include: an utterance-level coding set (i.e., literal gloss, general gloss, and pragmatic context), a speech act coding set (e.g., speech act and speech mode), and a basic linguistic coding set (e.g., sentence codings and syllable, morpheme and word counts).

[20] Spanish Natural Speech Corpus-Blume and English Natural Speech Corpus-Lust.



Query Definition

For each utterance found, display these fields (for the utterance itself or for related records):

Table/Field	Function	Group By	Order	
Session: Title		☑	⬇	Remove
Session: Age		☐	⬆ ⬇	Remove
Coding: value	average	☐	⬆	Remove

Find utterances that satisfy these conditions (on the utterance itself or on related records):

#	Field	Operator	Value	
[1]	Coding: title	equals (select fror	Number of morphemes (Global)	Remove
[2]	Utterance: Speaker	equals (select fror	S	Remove

Find utterances that have these codings:

#	Coding	Operator	Value	
[1]	Number of morphemes (Global)	does not equal (se	0	Remove

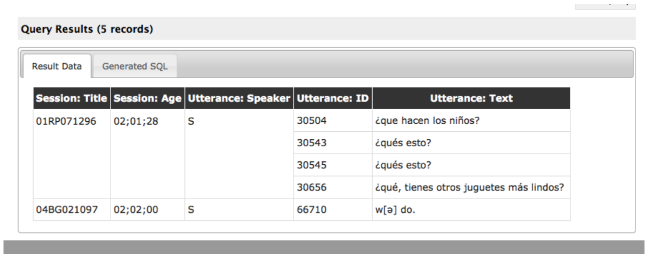
These codings must be found on the same utterance: All of the above

Query Results (2 records)

Session: Title	Session: Age	Coding: value avg
01RP071296	02;01;28	5.06329113924051
04BG021097	02;02;00	2.23762376237624

Figure 4. MLU query example.

Students are also asked to do queries searching for particular speech acts and sentence types and subtypes; for example, all utterances produced by the subject that have a question speech act, that are at the same time *wh*-questions[21] and simple sentences. Figure 5 shows the results of one such query run on the same two subjects of the query on figure 4.



Query Results (5 records)

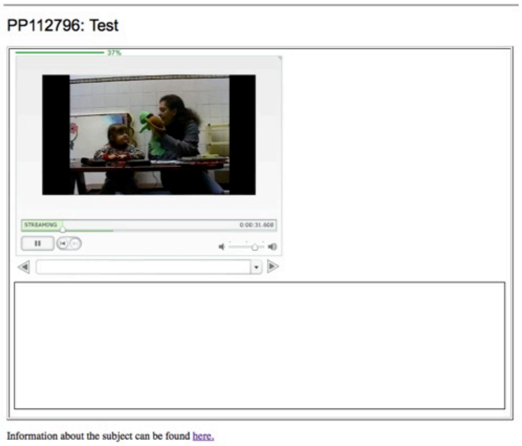
Session: Title	Session: Age	Utterance: Speaker	Utterance: ID	Utterance: Text
01RP071296	02;01;28	S	30504	¿que hacen los niños?
			30543	¿qués esto?
			30545	¿qués esto?
			30656	¿qué, tienes otros juguetes más lindos?
04BG021097	02;02;00	S	66710	w[ə] do.

Figure 5. *Wh*-question, simple sentence query.

Students are also asked to create some queries of their own invention so that we know they understand the logic behind the search engine and also to check that they are able to generate new research questions for an existing dataset. Thus all students with all backgrounds are taught methods of data management and analysis as well as research inquiry.

4.1.4 Human subjects

Various topics provide the student and/or researcher with the virtual experience of working with human subjects through audio-visual examples of research sessions in each learning module. They allow students to learn a method to use in their own research before going into the field. Figure 5 demonstrates an experimental study using the Elicited Imitation task done with a 2-year-old in Peru (Discourse Morphosyntax Interface in Spanish Non-Finite Verbs-Blume).



PP112796: Test

Information about the subject can be found here.

Figure 5. A video showing a particular research study that exemplifies the Elicited Imitation method.

Initial VLL topics integrate the IRB training and tests to ascertain that all students comply with their institution's regulations in this respect and learn Human Subjects requirements in general.

4.2 Data Challenges

4.2.1 Data complexity

The DTA cybertool in the VLL provides a structured annotation scheme for the representation of layers of metadata related to language data (i.e., language utterances). It does so in addition to providing structure for representation of reliability-checked language transcriptions and analyses of the utterances in those transcriptions. It thus helps to make data complexity tractable. Figure 6 provides an overview of the tool's structure showing the major area of data and metadata entry from a user's perspective.

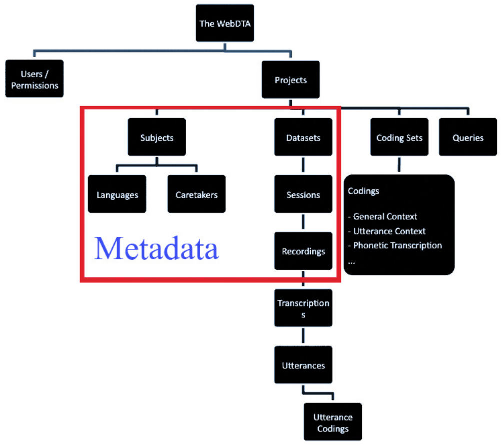


Figure 6. DTA structure diagram.

[21] Questions that in English start with a *wh*-word such as *what, why, how,* etc.

The DTA tool is based on 10 tables with the following basic markup categories: Project, dataset, subject, session[22], recording,[23] transcription, utterance, coding set, coding, and utterance coding.[24] Metadata codings involve the project and subject levels (figure 7) and the datasets themselves (figure 8) leading to transcribed utterances and related linguistic codings.

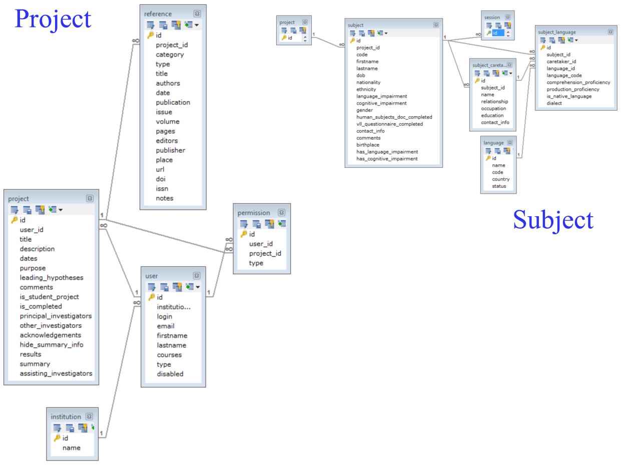


Figure 7. Project and subject metadata

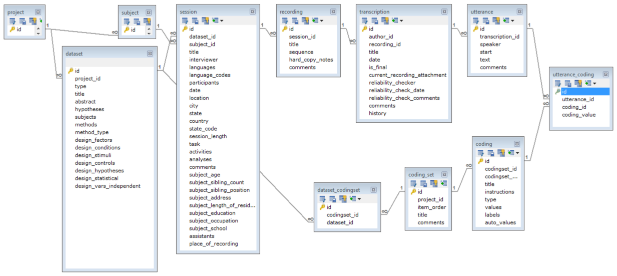
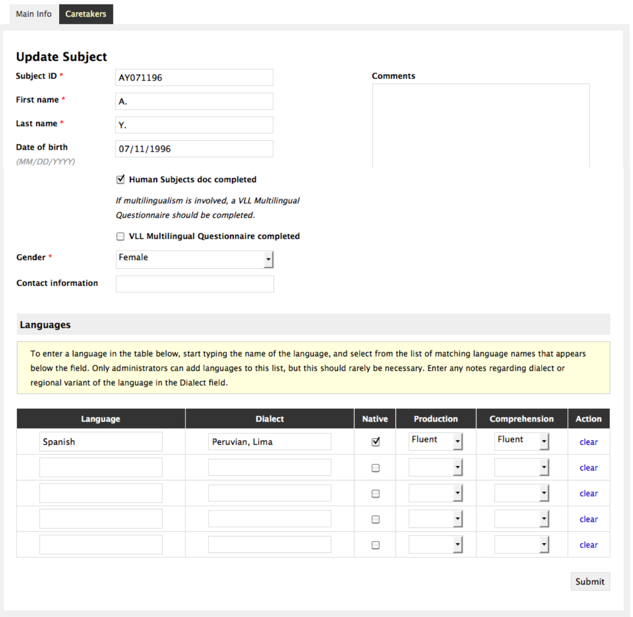


Figure 8. Dataset metadata.

4.2.2 Data documentation and standardization

The DTA tool provides the user with a web interface that guides him/her through steps for generating, storing and accessing both metadata and data. It trains researchers and students on how to organize research data. Users contribute data in a structured, uniform manner and they access calibrated data from a shared relational database. Therefore diverse data become comparable at many levels.

After entering the project's background information (cf. section 4.1.2), the student and/or researcher enters subject information, exemplified in Figure 9.[25]



Update Subject

Subject ID * AY071196

First name * A.

Last name * Y.

Date of birth (MM/DD/YYYY) 07/11/1996

Comments

☑ Human Subjects doc completed

If multilingualism is involved, a VLL Multilingual Questionnaire should be completed.

☐ VLL Multilingual Questionnaire completed

Gender * Female

Contact information

Languages

To enter a language in the table below, start typing the name of the language, and select from the list of matching language names that appears below the field. Only administrators can add languages to this list, but this should rarely be necessary. Enter any notes regarding dialect or regional variant of the language in the Dialect field.

Language	Dialect	Native	Production	Comprehension	Action
Spanish	Peruvian, Lima	☑	Fluent	Fluent	clear
		☐			clear
		☐			clear
		☐			clear
		☐			clear

Figure 9. The subject screen.

Then, a series of screens help students and researchers represent the research study's design. For each dataset, the user provides the main information (experiment/investigation, topic, abstract, related WebDTA projects/datasets), hypotheses, general subject description, methods, design, stimuli, procedures, and scoring procedures. These screens have an important educational purpose in teaching students how a particular experiment is designed. When a research project is completed, results, and conclusions can be linked.

Next, a session information screen guides the user to enter information for every time a subject was recorded for a given dataset. Each session has associated to it a recordings screen, a transcription screen, and a coding screen. The recordings screen houses information on all available primary data for a given session (audio or video files or previous transcripts in a number of formats) plus an inventory of the location of such files, and their backups.[26] The user then moves to a transcription screen where he/she can watch and listen to all available recordings (switching from one to the other as needed), transcribe and manually set timings to align the transcript with the recordings. The transcription screen is shown in figure 10.

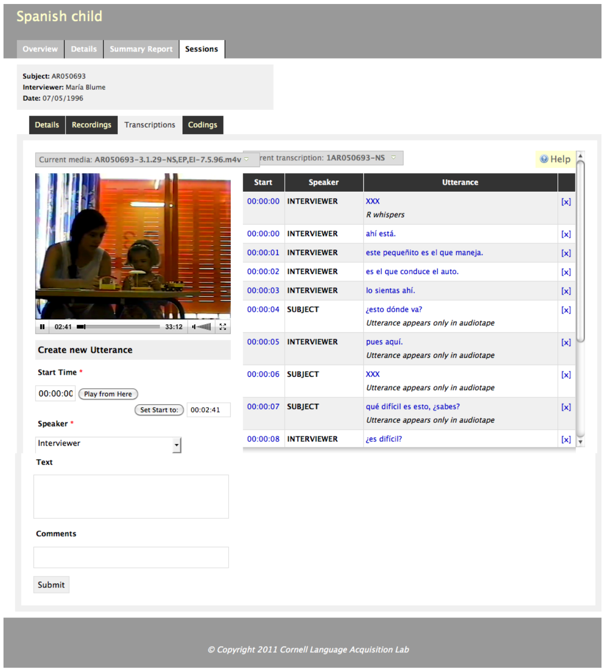
[22] A 'session' refers to a particular time in which a particular set of language data is recorded.

[23] Digital audio or video file, an electronic document (e.g. a Word, Excel, or PDF file), or information in a non-digital format such as a tape recording or paper transcription.

[24] An utterance coding records specified linguistic values (which the DTA tool refers to as 'codings') for a given utterance.

[25] In this figure, the confidential information of the subject has been eliminated. Only this screen provides confidential identifying data from subjects. The rest of the screens refer to subjects by an anonymous ID (initials and date of birth). Confidential data are only open to project PIs and selected primary collaborators. As the project moves to a diffusion stage, permission levels will structure this access.

[26] This screen supports all files supported by the JW Player,[26] *QuickTime* player, PDF, HTML, and image files, and, with additional software, other file formats such as *Microsoft Office* files.



Subject: AR050693
Interviewer: María Blume
Date: 07/05/1996

Current media: AR050693-3.1.29-NS,EP,EI-7.5.96.m4v

rent transcription: 1AR050693-NS

Start	Speaker	Utterance	
00:00:00	INTERVIEWER	XXX *R whispers*	[x]
00:00:00	INTERVIEWER	ahí está.	[x]
00:00:01	INTERVIEWER	este pequeñito es el que maneja.	[x]
00:00:02	INTERVIEWER	es el que conduce el auto.	[x]
00:00:03	INTERVIEWER	lo sientas ahí.	[x]
00:00:04	SUBJECT	¿esto dónde va? *Utterance appears only in audiotape*	[x]
00:00:05	INTERVIEWER	pues aquí. *Utterance appears only in audiotape*	[x]
00:00:06	SUBJECT	XXX *Utterance appears only in audiotape*	[x]
00:00:07	SUBJECT	qué difícil es esto, ¿sabes? *Utterance appears only in audiotape*	[x]
00:00:08	INTERVIEWER	¿es difícil?	[x]

02:41 33:12

Create new Utterance

Start Time * 00:00:00 Set Start to: 00:02:41

Speaker * Interviewer

Text

Comments

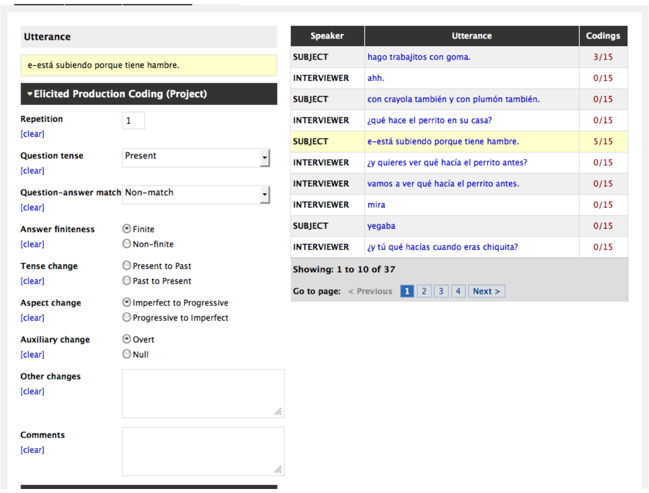
Figure 10. Transcription screen.

The utterances in these transcriptions, once they are reliability checked, constitute the basis for linguistic analyses.[27] Finally, researchers and students can code their data.

Through this tool we achieve systematization and calibration of metadata and data, thus allowing collaborative research programs and addressing the challenges of data documentation and standardization.

4.2.3 Data management and dataset design

As mentioned above, we wanted to design a structure that was open enough to accommodate both experimental and natural speech data from various types of populations[28] because such a resource did not exist in our field. The creation of this database, thus, provides previously unavailable resources for collaboration among researchers. The DTA tool structures data creation and analysis but allows the researcher to create project-specific coding sets and queries. Figure 11 illustrates a specific coding set created for an utterance from a Peruvian child participant in the experimental Project, "Discourse Morphosyntax Interface in Spanish Non-Finite Verbs-Blume" where language production is being systematically elicited from a child by the experimenter following an experimental design.



Utterance

e-está subiendo porque tiene hambre.

▾Elicited Production Coding (Project)

Repetition: 1

Question tense: Present

Question-answer match: Non-match

Answer finiteness: ● Finite ○ Non-finite

Tense change: ○ Present to Past ○ Past to Present

Aspect change: ● Imperfect to Progressive ○ Progressive to Imperfect

Auxiliary change: ● Overt ○ Null

Other changes

Comments

Speaker	Utterance	Codings
SUBJECT	hago trabajitos con goma.	3/15
INTERVIEWER	ahh.	0/15
SUBJECT	con crayola también y con plumón también.	0/15
INTERVIEWER	¿qué hace el perrito en su casa?	0/15
SUBJECT	e-está subiendo porque tiene hambre.	5/15
INTERVIEWER	¿y quieres ver qué hacía el perrito antes?	0/15
INTERVIEWER	vamos a ver qué hacía el perrito antes.	0/15
INTERVIEWER	mira	0/15
SUBJECT	yegaba	0/15
INTERVIEWER	¿y tú qué hacías cuando eras chiquita?	0/15

Showing: 1 to 10 of 37

Go to page: < Previous 1 2 3 4 Next >

Figure 11. Project specific linguistic coding set.

At the same time that the DTA tool provides a primary research tool, it automatically provides a rich, continually growing archive allowing present and future collaboration on shared data, potentially long distance and potentially interdisciplinary. In general, with external linkages, through Linked Data formats [5, 8, 13, it can be linked to a wide intellectual knowledge base, e.g., linking published forms of research to the actual data and data methods used to create the results reported.[29]

4.3 Researcher Collaboration

4.3.1 Researcher training

Membership in the VCLA provides researchers a new medium for accessing peers at other institutions and countries to engage in collaborative projects under considered principles for collaboration and data sharing, and our web-conferencing allows them to have online meetings. The VLL portal materials provide essential readings on issues related to distance collaboration. The DTA tool helps researchers find detailed information about other researchers and student projects and helps organize collaborative research materials for a specific project. Researchers, like students, can get trained in cybertool use through the virtual workshops and the DTA User manual.

In addition, the tool allows continual generation of new queries on data based on codings that derive from a particular research question, so that each researcher or student researcher can get the results they are looking for in their specific projects. Data analyses are cumulative, as they are stored in the resulting database. For more detail on this aspect of the DTA tool see [8].

[27] The DTA tool keeps a record of all recordings and transcriptions available for a particular session. The user can easily switch between recordings and transcriptions to view all versions of the raw and primary data. In addition, it keeps a record of who was the transcriber and when was the date of the first transcription, as well as of any subsequent reliability checked versions of a transcript, including reliability checker and date information.

[28] e.g. child vs. adult, first vs. second language acquisition, child vs. adult second language acquisition, normal vs. disabled populations.

[29] In order to maximize generalizability across fields of our tools, The DTA tool is designed to maximize the possibility for linked data by integrating with field standards. For example, the application uses the UTF-8 encoding to store text, which can represent any language. For this, the application adopts ISO 639-3 standard language codes [36], which lists over 7000 languages, developed by Ethnologue/SIL (http://www.ethnologue.com/codes/default.asp). It links with GeoNames (http://www.geonames.org/) [20] in geographic reference.

4.3.2 Intellectual property

Founding members of the VCLA have, through a series of video conference meetings, begun to design principles of agreement by which to assure the protection of each VLL member's intellectual property rights, while at the same time allowing for collaboration in new projects and the repurposing of previously collected data. Although this is still work in progress, a first summary of our vision and principles can be found on the VCLA website.[30]

4.3.3 Cross-institutional and international collaboration

To begin to address the issues we identified above which challenge cross-institutional and cross-country collaboration, VCLA founding members have begun meeting collectively with representatives of the IRB committees at their institutions and have begun collecting cross-institutional data to determine commonalities and differences across them.

5. EDUCATIONAL ACHIEVEMENTS TO DATE

A structured series of synchronous cross-institutional courses, including two with Peru, at the undergraduate and graduate levels have introduced the components of the VLL through our structured VLL web portal. Students from Computer Science, Human Development, Linguistics, and Psychology participated in these courses and web conferences.[31] These several courses took students through initial introduction to scientific methods for data collection and management, followed by advanced cybertool learning through specialized research agendas. A series of cross-institutional web-conferences supplemented these courses in order to cultivate collaboration among students and faculty.

Our cross-institutional and international courses gave students new perspectives. For example, in a course on bilingualism three different professors and three groups of students (University of Texas at El Paso, Rutgers University, New Brunswick, and Pontificia Universidad Católica del Peru) participated and shared information on three different multilingual situations: New Jersey: dominant English, Spanish minority language, El Paso: border city with majority Hispanic population and strong ties to Mexico, and Peru: Spanish dominant language and various indigenous languages.

Students, through use of the VLL, engaged in several stages of original data analyses, culminating in original experimental research proposals. [32] The synchronous courses provided accumulated syllabi, materials and assignments that were used asynchronously as well.

Our main educational achievement has been to train students from the beginning on documenting data in such a way that will spare them on having to do what previous generations of researchers, us included, spend too much time having to do, i.e., find our old data, find our old records, find our old tapes, try to connect them all, hunt up metadata, experimental designs and stimulus sentences, etc. In this exploratory stage of our project, we have educated a small section of a new generation of students which can now pass such information to colleagues and future students, so that our efforts will, hopefully, payoff in the future.

We have also exposed these students to all the arguments in favor of large-scale data sharing and research collaboration, as well as to several of the problems surrounding such collaborative projects so that they can avoid pitfalls in the future. This is a topic that is not traditionally discussed in our field. We have, thus, planted a seed and achieved some beginning collaborative projects; whether students will embrace this new culture is yet to see, but we have at least given them the opportunity, the technology, and the computational skills to do so.

Student surveys conducted during synchronic cross-institutional courses to date have indicated high satisfaction with the course and in particular with its cybertools.[33] Critically, students asked for more interaction with students at other universities, indicating a positive inclination towards collaboration.[34] Figure 12 shows the overall results of the surveys.

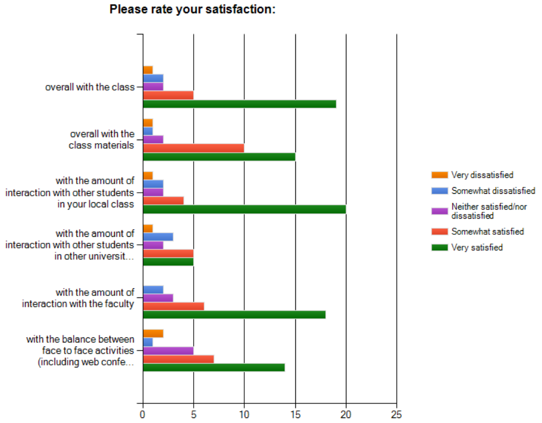


Figure 12. Student satisfaction survey.

[30] http://vcla.clal.cornell.edu/en/principles#collaboration

[31] At this point we have no information available for some semesters in which we taught the courses due to changes in server administrators and a lack of recognition on our part that we had to keep a record of all users before deleting their accounts. We do have information on some semesters that were not different from other semester in terms of access. Fall 2010, 67 new users accessed our materials and courses. For 2011, 35 new accounts were created, but it is important to bear in mind that users from previous semesters/years do not need to get new access each semester, so more students from previous semesters were actually still using our VLL.

[32] Examples of collaborative research projects including students and researchers are at Cornell (Barbara Lust): "SAQL Phase 1: Expert Evaluation and Validation of a New Child Multilingualism Questionnaire", Newcastle University (Cristina Dye) and Boston College (Claire Foley): "Acquisition of VP ellipsis in mono and bilingual children"; MIT (Suzanne Flynn), Massachusetts General Hospital (Janet Cohen Sherman) and Cornell (Barbara Lust): "Alzheimer's language project" with Jordan Whitlock.

[33] We provided all students the opportunity to answer these end-of-semester online surveys. Only a small proportion (n. 29) of the students who took the courses synchronously answered the survey. We are looking at alternative ways of distributing surveys in the future to improve the number of respondents.

[34] In general, this part of the project suffered from scheduling and language barrier problems. We are discussing how to improve this in future courses.

6. BROAD IMPACT

This is the first project of its kind in the social sciences and humanities and thus can serve as a model for other Social and Cognitive Sciences, as well as other STEM sciences.

It empowers a wide array of collaborative and interdisciplinary research and teaching agendas. It incorporates sound scientific principles and structured data management in a cybertool that provides a distributed infrastructure for collaborative learning and research in the study of language, bilingualism, and language development.

It creates new learning and research possibilities for Hispanic students, usually underrepresented in the sciences, at University of Texas at El Paso, Rutgers University, New Brunswick, and Peruvian students at Pontificia Universidad Católica del Perú.

7. FUTURE CHALLENGES AND LESSONS LEARNED

Our main challenges now, as we approach the diffusion stage of our project, concern the dissemination of our infrastructure and materials to a broader community of practice, one that is both interdisciplinary and cross-linguistic.

In this, we face issues of sustainability. In order to open the VLL materials to a wide audience, we must build a sustainability model that includes licensing and/or subscription options. For this we have now initiated correspondence with Cornell's E-Cornell program (eCornell.com). To ensure long-term sustainability, we must also negotiate and fund long-term storage and maintenance of the DTA tool and its database. We must develop an infrastructure for long-term management of the tool and its access and use. In our view this would ideally be some form of a distributed infrastructure rather than a localized one.[35]

To extend the DTA tool to new users we must establish a set of user principles and agreements involving shared materials and data. This must involve establishment of a leveled set of permissions, e.g., read only, etc. Founding Members of the VCLA are currently addressing this challenge.

Some practical issues also create challenges for the development of a project such as ours. Ironically, one of them is language. To teach our first international class with Peru, we had to translate most materials to Spanish, and since the class was taught in Spanish, not all US sites were able to participate in our class discussions. Coordinating schedules across US and international time zones for joint courses also proved to be a challenge for those who wanted to participate synchronously in our courses. These challenges will arise with each new language and country added to the project (e.g., India, Korea, Israel planned for extensions). One of our next and continuous challenges includes translating materials to other languages and possibly providing interpretation to allow better collaboration across countries. Technical and administrative challenges in cybertool development required additional costs and time beyond that first expected.

Among the lessons learned is, hence, the fact that cross-institutional collaboration demands precisely planned infrastructure. Another issue that confronted us was how hard it is to foster cross-institutional and international collaboration, even when tools for collaboration are in place and the desire of collaboration exists in all parties. While students were relatively easily encouraged to collaborate, time constraints and previous commitments on faculty,[36] plus a lack of real support for collaborative work by academic institutions, as observed in *Nature* [39], will require additional support for faculty time and commitment, if collaborative projects such as this are to flourish.

8. ACKNOWLEDGMENTS

8.1 Founding

National Science Foundation

CI-TEAM program. "Transforming the Primary Research Process Through Cybertool Dissemination: An Implementation of a Virtual Center for the study of language acquisition" (Award No. OCI-0753415, María Blume and Barbara Lust)

Planning Grant: A Virtual Center for Child Language Acquisition Research (BCS-0126546) (Barbara Lust and Founding Members) NSF Development and Learning Sciences Program

Small Grant for Exploratory Research. Planning Information Infrastructure through a New Library-Research Partnership. (NSF#0437603) (Janet McCue, director of Cornell Albert Mann Library and Barbara Lust) NSF Science and Engineering Information Integration and Informatics.

Cornell:

CISER Seed Grant: Building an Interdisciplinary Infrastructure to Foster Collaboration among Research Scientists in the Language Sciences: Proposal to Create Instrument-based Virtual Linguistics Laboratory for the International Study of Language Acquisition and Neural Plasticity.

FABIT (Faculty Innovation in Teaching) Award. "Integrating Digital Multimedia resources in two interdisciplinary language development courses" and "A VRE (Virtual Research Environment) for Language Acquisition and Cognitive Science. "

New York State Hatch grant. "Cornell Language Acquisition Lab: Building an Electronic Library of Words of the World's Children".

Seed Grant from Cornell Einaudi Center for International Studies: Exploring Activation of an International component in a Virtual Center for the Study of Language Acquisition.

8.2 Collaborators

We would also like to acknowledge the contributions to this project of:

VCLA members and affiliates: Elise Temple (NeuroLab) Gita Martohardjono (CUNY), Cristina Dye (Newcastle University), Isabelle Barrière (CUNY), Yarden Kedar (Ben Gurion University at the Negev), and Sujin Yang (Tyndale University College and Seminary)

Application developers: Ted Caldwell and Greg Kops (GORGES). Susan Mehringer (Cornell University Center for Advanced Computing) and Tommy Cusick (Google, cf. footnote 16 above), Cliff Crawford for early web DTA tool programming, Alex Immerman for portal and cybertool development.

[35] [8] lists further challenges specific to the DTA tool.

[36] "[…] As Gibbons and anthropologist Nancy Fried Foster observed in their 2005 postmortem, «The phrase 'if you build it, they will come' does not apply to IRs [institutional repositories].»" [28, p. 160].

Creators of previous versions of our DTA tool: Katharina Boser, David Parkinson, and Shamita Somashekar, then graduate students at Cornell.

Student RAs: Darlin Alberto, Gabriel Clandorf, Natalia Buitrago, Poornima Guna, Jennie Lin, and Jordan Whitlock at Cornell, and Marina Kalashnikova. Martha Rayas Tanaka, Lizzeth Pattison, María Jiménez, and Mónica Martínez at UTEP.

Undergraduate and graduate students who have participated in the courses and helped with their input at UTEP, Cornell, Rutgers New Brunswick, MIT, and Pontificia Universidad Católica del Perú.

9. REFERENCES

[1] "A fair share." *Nature* 444, 7120 (2006): 653-654.

[2] Atkins, D. 2005. CyberInfrastructure and the Next Wave of Collaboration, D. E. Atkins, Keynote for EDUCAUSE Australasia, Auckland, New Zealand, April 5-8, 2005.

[3] Bender, E. 2004. Rules of the Collaboratory Game. Science of Collaboratories papers: *MIT's Technology Review*. November 23, 2004

[4] Berman, F. and H. Brady. 2005. Workshop on Cyberinfrastructure for the Social and Behavioral Sciences: Final Report. http://*ucdata.berkeley.edu/pubs/****CyberInfrastructure_FINAL****.pdf*

[5] Berners-Lee, T. 7/26/2006. Linked data. http://www.w3.org/DesignIssues/LinkedData.html

[6] Bird, S. and G. Simons (2003). Seven dimensions of portability for language documentation and description. *Language*, vol. 79, 3. (557-582)

[7] Birney, E., T.J. Hudson, E.D. Green, C. Gunter, S. Eddy, J. Rogers, J.R. Harris, and S. Dusko Ehrlich. 2009. "Prepublication data sharing" *Nature* 461, 7261: 168-170.

[8] Blume, M., S, Flynn, and B. Lust. 2012. Creating Linked Data for the Interdisciplinary International Collaborative Study of Language Acquisition and Use: Achievements and Challenges of a new Virtual Linguistics Lab. In C. Chiarcos, S. Nordhoff, and S. Hellmann (Eds.) *Linked Data in Linguistics: Representing Language Data and Language Metadata*. Berlin/Heidelberg: Springer, pp. 85-96.

[9] Blume, M. and B. Lust, 2012. Data Transcription and Analysis Tool User's Manual. (with the collaboration of S. Somashekar and T. Ogden). http://webdta.clal.cornell.edu

[10] Blume, M., S. Yang, and B. Lust. (with the collaboration of T. Ogden, S. Somashekar, Y. Chien, L. Sánchez, C. Foley, M. Kalashnikova, M. Rayas, and N. Buitrago)(in prep) Cornell University Virtual Linguistics Lab (VLL) Research Methods Manual: Scientific Methods for Study of Language Acquisition.

[11] Borgman, C. 2007. *Scholarship in the Digital Age*. Cambridge: MIT Press.

[12] Bos, N., A. Zimmerman, G. Olson, J. Yew, J. Yerkie, E. Dahl, and G. Olson. 2007. From shared databases to communities of practice: A taxonomy of collaboratories. *Journal of Computer-Mediated Communication*, *12* (2), article 16. http://jcmc.indiana.edu/vol12/issue2/bos.html

[13] Chiarcos, C., S. Nordhoff, and S. Hellmann (Eds.) *Linked Data in Linguistics: Representing Language Data and Language Metadata*. Berlin/Heidelberg: Springer.

[14] "Credit where credit is overdue" *Nature Biotechnology* 27, 7 (2009): 579.

[15] "Data's shameful neglect" *Nature* 461, 7261 (2009): 145.

[16] Edwards, J. A. 1992a. Transcription of discourse. *International Encyclopedia of Linguistics*, ed. by William Bright, 367-370. Oxford: Oxford University Press.

[17] Edwards, J. A. 1992b. Computer methods in child language research: four principles for the use of archived data. *Journal of Child Language* 19.435-58.

[18] Elkins, K. 2012. Tiptoeing through Minefields: Launching Collaborations. *Association for Women in Science*, Winter 2012*,* vol. 43, no. 1, pp. 21-23.

[19] E-MELD: Electronic Metastructure for Endangered Languages Data http://www.emeld.org/index.cfm

[20] GeoNames: http://www.geonames.org/.

[21] GOLD Community: http://www.linguistics-ontology.org/

[22] Grosjean, F. 1998, 2004. Studying bilinguals: Methodological and conceptual issues. *Bilingualism: Language and Cognition*, 1, 131-149. And in T. K. Bhatia & W. C. Ritchie (Eds.). *The Handbook of Bilingualism* (pp. 32-63). Oxford, England: Blackwell Publishing.

[23] Grotke, R. W. 2004. Digitizing the world's largest collection of natural sounds: key factors to consider when transferring analog-based audio materials to digital formats. *RLG DigiNews*, Vol. 8, Number 1. Online: http://worldcat.org/arcviewer/2/OCC/2009/08/11/H1250010262952/viewer/file2.html

[24] King, G. 2011. "Ensuring the Data-Rich Future of the Social Sciences" *Science*, 331, 1: 719-721.

[25] "Let data speak to data." *Nature* 438, 7068 (2005): 531.

[26] Lust, B., M. Blume, Y., Kedar, S. Yang, and S. Callahan. A Glossary of Language Acquisition.

[27] Lust, B., S. Flynn, M. Blume, E. Westbrooks, and T. Tobin. 2010. Constructing Adequate Language Documentation for Multifaceted Cross-Linguistic Data: A Case Study from a Virtual Center for Study of Language Acquisition. In L. A. Grenoble and N. L. Furbee (eds.). *Language Documentation: Practice and Values*. pp. 89-107. Amsterdam and Philadelphia: John Benjamins.

[28] Nelson, B. (2009) "Empty archives" *Nature* 461, 7261: 160-163.

[29] NSF. 2003. *Revolutionizing Science and Engineering through Cyberinfrastructure*.

[30] NSF. 2007. *Cyberinfrastructure Vision for 21st Century Discovery*. March NSF 07-28.

[31] OLAC: Open Language Archives Community, http://www.language-archives.org/

[32] OLWG: Open Linguistics Working Group, http://linguistics.okfn.org.

[33] Olson, G., A. Zimmerman, and N. Bos. (eds.) 2008. *Scientific Collaboration on the Internet.* MIT Press.

[34] Pfirman, S., J. Collins, S. Lowes, and A. Michaels. 2005. February 11. Collaborative Efforts: Promoting Interdisciplinary Scholars. The Chronicle Review. *The Chronicle of Higher Education*. Vol. 51, issue 23, page B15. http://chronicle.com.

[35] Schofield, P.N., T. Bubela, T. Weaver, L. Portilla, S.D. Brown, J.M. Hancock, D. Einhorn, G. Tocchini-Valentini, M. Hrabe de Angelis, N. Rosenthal, and CASIMIR Rome Meeting participants. 2009. "Post-publication sharing of data and tools" *Nature* 461, 7261: 171-173.

[36] SIL. 2006. ISO/DIS 639-3. Dallas: SIL International. http://www.sil.org/iso639-3/.

[37] Simons, G. 2004. Ensuring that Digital Data Last. The priority of archival form over working form and presentation form. Paper presented at Symposium on Best Practice. Linguistic Society of America Annual Meeting, Boston, Mass.

[38] Wenger, E. and J. Lave. 1998. *Communities of practice: Learning, Meaning, and Identity*. N.Y.: Cambridge University Press.

[39] "Who'd want to work in a team?" *Nature* 424, 6944 (2003):1.