Creating Linked Data for the Interdisciplinary International Collaborative Study of Language Acquisition and Use: Achievements and Challenges of a New Virtual Linguistics Lab

María Blume, Suzanne Flynn, and Barbara Lust

Abstract In this paper, we describe and exemplify our development of a cyber-tool, the Data Transcription and Analysis tool (DTA tool) that is currently being implemented in the Virtual Center for Language Acquisition through a Virtual Linguistic Lab (VLL). We review this cyber-tool's design and accomplishments to date, assessing its ability to address "the challenge of our time to store, interlink and exploit this wealth of data" (Chiarcos et al, this vol.). We explicate the architecture and usability of the DTA tool, we summarize its current status, possibilities for expansion, and related challenges we currently confront. We focus on the conceptual and functional structure of this tool here, and not on technical aspects of its programming.

1 Introduction

Data collected from the fields of language acquisition and use are multi-lingual, multi-modal, 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 an immensely complex set of databases often appearing in diverse formats as different labs generally practice distinct forms of data management. Data from

María Blume
Department of Languages and Linguistics, University of Texas at El Paso, Liberal Arts Bld. Room 119, El Paso, TX, USA 79968, e-mail: mblume@utep.edu

Suzanne Flynn
Department of Linguistics and Philosophy, MIT, 77 Massachusetts Avenue, 32-DB808, Cambridge, MA 02139, USA, e-mail: sflynn@mit.edu

Barbara Lust
Department of Human Development, Cornell University, G57 Martha Van Rensselaer Hall, Ithaca, NY 14853, USA, e-mail: bcl4@cornell.edu

more than 20 languages and cultures and thousands of child and adult subjects exist in the Cornell Language Acquisition Lab and Virtual Center for Language Acquisition alone. 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). Language data collections are infinitely expandable and should be used, reused and, when possible, repurposed.

This scientific enterprise requires that these data be stored, accessible and exploited in a manner where relationships can be discovered within and across data sets. In keeping with fundamental insights of the "Linked Data" program (Berners-Lee, 2009), the more each data singleton can be significantly connected or "interlinked", the more powerful and useful it becomes. For example, in the study of language acquisition and use, such interlinks can be cross-disciplinary (e.g., connecting brain images with behavioral experimental results testing language comprehension or production), or linguistically specific (e.g., comparisons of certain properties of sentence structure or verb morphology in a Spanish- versus an English-, French- or Sinhala-speaking child's speech at a particular stage of language development). 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.

2 Approach

In the Virtual Center for Language Acquisition (VCLA),[1] faculty from eight US universities and one international university[2] (Peru) converged through US National Science Foundation, Cornell University and University of Texas at El Paso support to "Transform the Primary Research Process" in the area of language acquisition.[3] A Virtual Linguistics Lab (VLL)[4] was constructed to provide an infrastructure of principles, best practices, materials and cyber-tools including the Data Transcription and Analysis tool.

The Data Transcription and Analysis tool (DTA tool)[5] provides a web-based interface to guide the user – either researcher or student learning scientific methods of research – in primary data creation, data management and collaborative data use.

---

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

[2] Founding members: Suzanne Flynn, MIT; ClaireFoley, 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 and James Gair, Cornell University; María Blume, University of Texas at El Paso; and Affiliate member Jorge Iván Pérez Silva, Pontificia Universidad Católica del Perú.

[3] Lust, B. 2003 (NSF BCS-0126546); McCue and Lust 2004 (NSF 0437603). Blume and Lust 2007 (NSF OCI-0753415). Seed grant support from the following was also essential to this project: American Institute for Sri Lankan Studies, Cornell University Einaudi Center, Cornell University Faculty Innovation in Teaching Awards, Cornell Institute for Social and Economic Research (CISER), New York State Hatch grant.

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

[5] http://webdta.clal.cornell.edu

At the same time, data entry through this tool automatically feeds a structured, calibrated, and infinitely expandable cross-linguistic relational database. It provides means for structuring, storing and linking scientifically sound, diverse, but calibrated, language data, either naturalistic or experimentally derived, ranging from raw to structured forms, which can be accessed, connected and queried in linked fashion (Lust et al, 2010).

3 Overall Architecture of the DTA Tool

The DTA tool[6] offers the user a structured annotation scheme for the representation of layers of metadata related to language data (i.e., the actual utterances) as well as for representation of reliability-checked transcriptions and analyses of the utterances themselves. Figure 1 provides an overview of the tool's structure showing the major areas from a user's perspective.

The DTA tool is based on 10 tables with the following basic markup categories: Project, dataset, subject, session, recording,[7] transcription, utterance, coding set, coding, and utterance coding.[8] Metadata codings involve the project and subject levels (Fig. 2) and the datasets themselves (Fig. 3) leading to transcribed utterances and related linguistic codings.

The DTA tool also provides some non-project-specific linguistic coding sets: an utterance level coding set (including a literal gloss and a general gloss as well as pragmatic context specification), a set (including speech act and speech mode), and a basic linguistic coding set (including sentence codings and syllable, morpheme and word counts). Users working on natural speech data are expected to use these basic codings, so that the data are calibrated across projects. A researcher working on an experimental project may select if s/he wants to use all, none or some of the established codings. Regardless of the project type, researchers can create new project-specific coding sets or global coding sets.[9]

---

[6] We concentrate on the latest 2011 version of the DTA tool, which has been reprogrammed by GORGES, a web and mobile technology firm in Ithaca, New York. (http://www.gorges.us). The DTA tool has been under construction for more than 20 years through several generations of students who contributed to its development and through several changing technical formats. For a history of its development see the DTA User's Manual (Blume and Lust, 2011; Blume et al, in prep).
The current version of the WebDTA tool is built on Yii, a PHP web development framework that uses the 'Model-View-Controller' pattern to structure the application and the 'ActiveRecord' pattern to manage records from the database. MySQL is used for the database platform. All are open source technologies.

[7] 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.

[8] An utterance coding records specified linguistic values (which the DTA tool refers to as 'codings') for a given utterance. 'Subject' refers to the participant providing language to be studied.

[9] Selected permission level is necessary for such new coding creation.

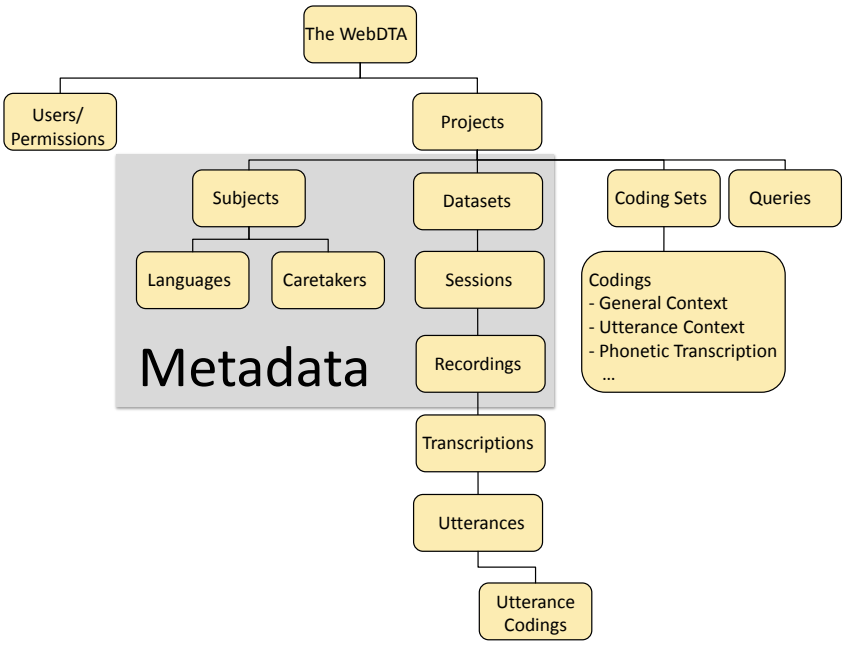

Fig. 1 WebDTA tool basic structure

Figure 4 exemplifies basic coding of an utterance of natural speech data of a Peruvian monolingual Spanish-speaking child from the 'Spanish Natural Speech-Blume' corpus. Such codings render the data ready for further analyses in connection with specific research questions.

4 Examples of DTA Tool Content

Users complete a series of tables through a graphical user interface intended to facilitate and structure a user's data and metadata entry. The tool first leads the user to enter detailed project-level information and metadata. Main areas include project investigators, purpose and leading hypotheses, references (publications, presentations, related studies, bibliography), subjects, and results and discussion. At several different points documents can be attached. The tool also provides summary reports showing subsets of metadata in addition to the datasets of the project.

The user enters information on the datasets that form a project. 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. When a research project is completed, results and conclusions can be linked. The DTA tool provides a report at

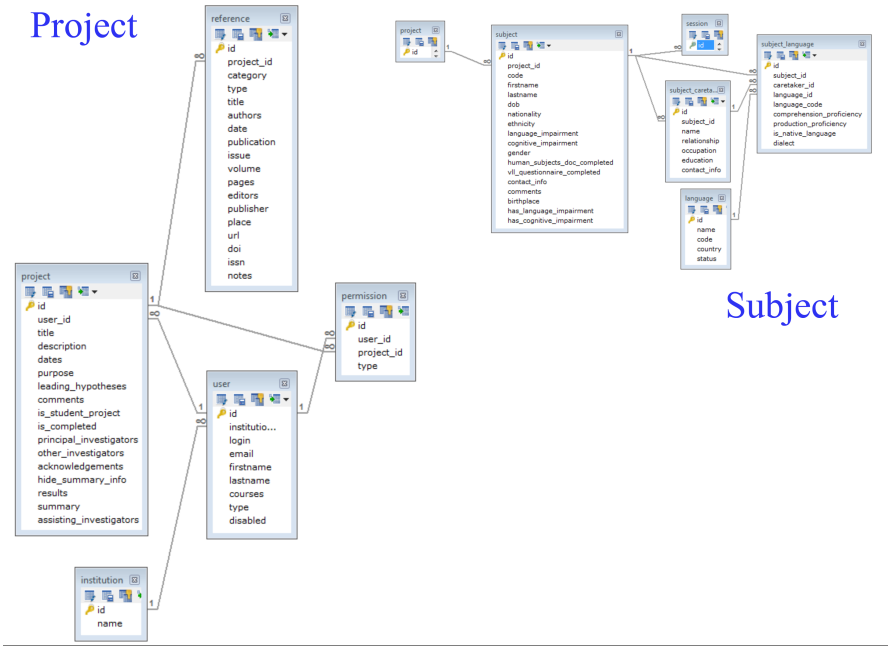

Fig. 2 Project and subject metadata

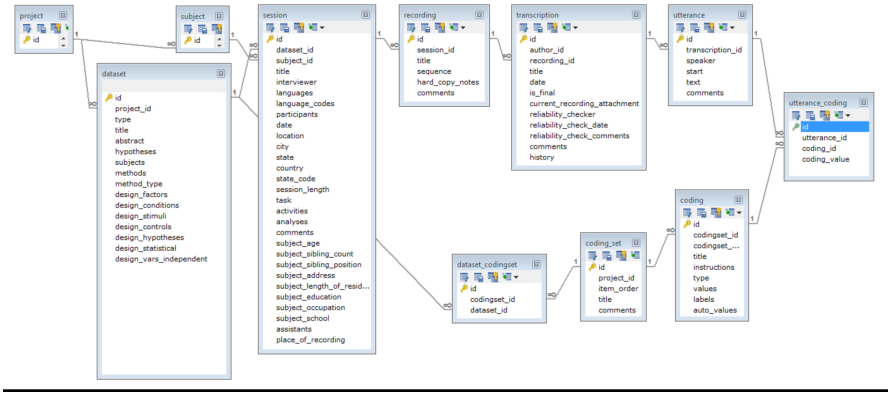

Fig. 3 Dataset metadata

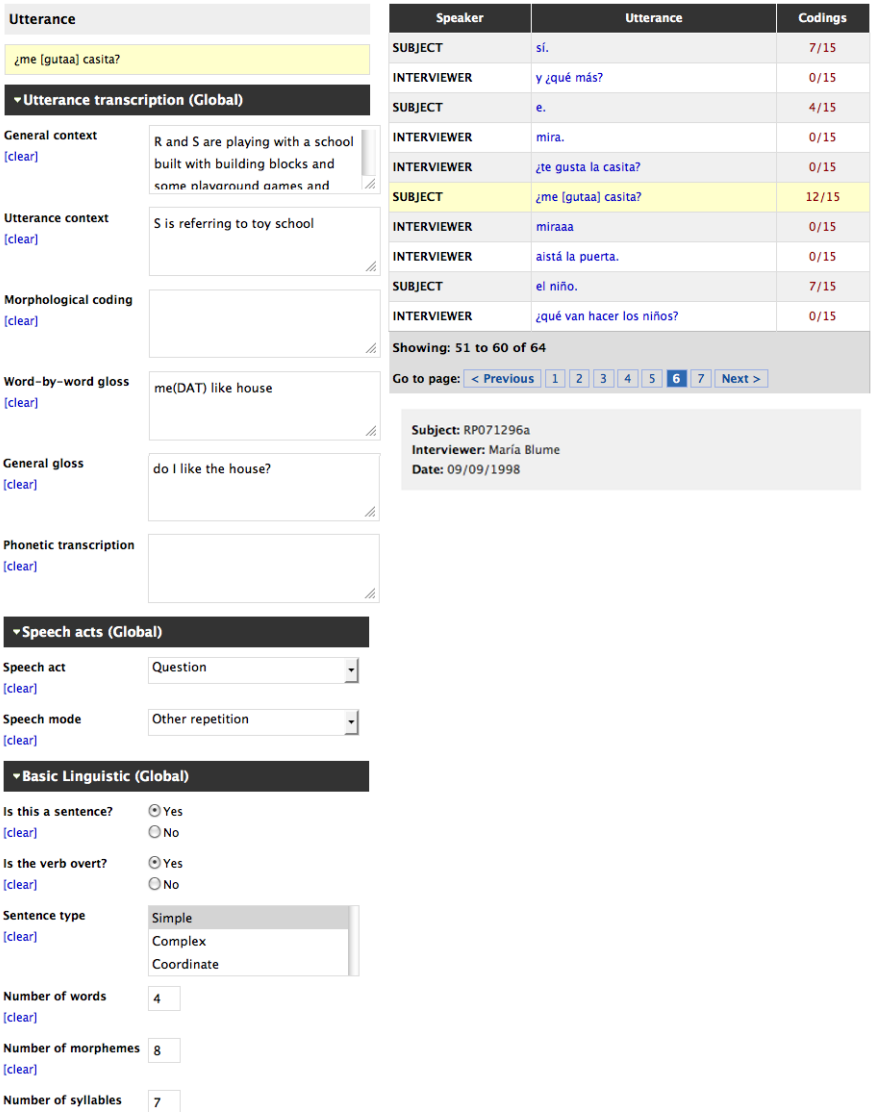

Fig. 4 Coding screen: Example of Spanish child language utterance entry

the dataset level, the Experiment Bank report, including all the information for the project and each of its datasets.

After the user provides specified metadata for all the subjects in the project, the user enters information for the sessions pertaining to each dataset. Each session[10] has associated to it a recordings screen, a transcription screen, and a coding screen.

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

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. This screen supports all files supported by the JW Player,[11] *QuickTime* player, PDF, HTML, and image files, and, with additional software, other file formats such as *Microsoft Office* files. 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. Finally, the user moves to the coding screen where he/she can code for any of the global coding sets or for codings created specifically for the particular project. Figure 5 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".

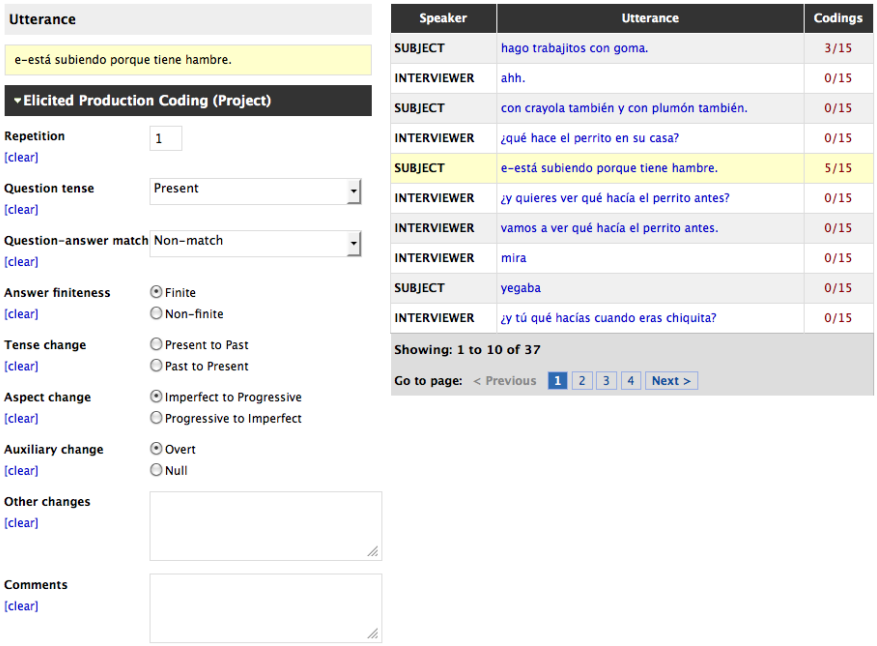

Fig. 5 Project-specific linguistic coding set

[11] JW Player: H.264 video (.mp4, .mov, .f4v), Flash video (.flv), YouTube video, 3GPP video (.3gp, .3g2), MP3 audio (.mp3), and AAC audio (.aac, .m4a). QuickTime: AIFF audio (.aif, .aiff). Image files: .jpg, .jpeg, .gif, .png

5 Linking Data Through Queries

A set of queries, which is essential to calibrating language data, is available initially in the DTA tool. Queries search properties of the data such as MLU (Mean Length of Utterance) in words, syllables, and morphemes; coded information on all subjects across projects who speak a particular language or have a certain age, utterances that are sentences, utterances that are NPs, sentences with overt verbs, simple vs. complex sentences, specific speech acts or speech modes. Figure 6 illustrates a basic query searching for all simple sentences with verbs for a particular session for a subset of one subject's data in the 'Spanish Natural Speech Corpus-Blume' corpus.[12] Figure 7 shows the results of the query. 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 and subjects.

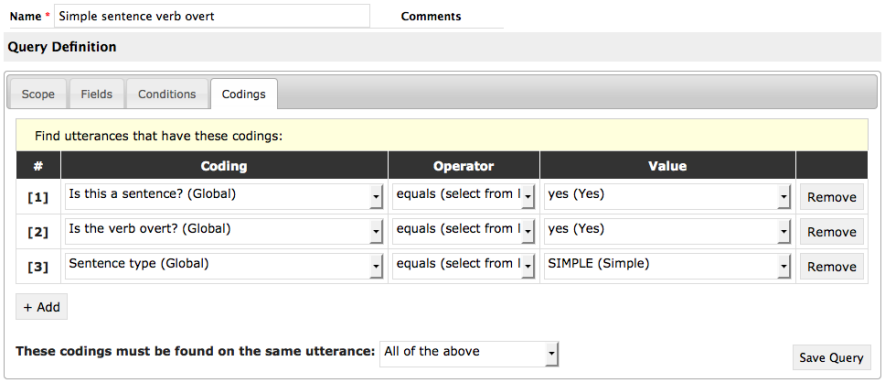

Fig. 6 Values of codings in query for all simple sentences with overt verbs.

In addition, the tool allows continual generation of new queries based on available codings that derive from a particular research question. For example, we can ask about a specific matching of tense and aspect forms in a verb in a question and in the answer in an Elicited Production task. We can also query about sentences that exemplify standard word orders across natural speech samples of children speaking different languages at the same MLU stage, etc.

Given the possibility for such content generation through the DTA tool, it can be seen that the DTA tool is a *primary* research tool, guiding the researcher or student in data collection and management; its potential usability extends from specific research projects to use in educational domains as well. At the same time, the tool automatically provides a rich, continually growing archive allowing present and future collaboration on shared data, potentially long distance, and potentially interdisci-

[12] To create a query the user would also need to define the scope, conditions for the query and the fields one would like to see when results are found, which we do not show here due to space limitations.

Query Results (15 records)

Result Data | Generated SQL

Transcription: Title	Utterance: Speaker	Utterance: Text	Coding: Title	Coding: value
RP071296-María	SUBJECT	¿dónde estás?	Is the verb overt?	yes
			Is this a sentence?	yes
			Sentence type	SIMPLE
		¿me [gutaa] casita?	Is the verb overt?	yes
			Is this a sentence?	yes
			Sentence type	SIMPLE
		eee trabajando.	Is the verb overt?	yes
			Is this a sentence?	yes
			Sentence type	SIMPLE
		no, así se pasea.	Is the verb overt?	yes
			Is this a sentence?	yes
			Sentence type	SIMPLE
		yo tengo [e] disco [deautoβu] mágico.	Is the verb overt?	yes
			Is this a sentence?	yes
			Sentence type	SIMPLE

Fig. 7 Previous query's results

plinary.[13] In general, with external linkages, through Linked Data formats, the DTA tool's database 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.

6 Further Linkages: External

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.[14] For this, the application adopts ISO 639-3 standard language codes, which lists over 7000 languages, developed by Ethnologue/SIL.[15] It links GeoNames.org in geographic reference.

[13] 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 and Cornell (Barbara Lust): "Alzheimer's language project".

[14] Given the availability of language specific fonts.

[15] http://www.ethnologue.com/codes/default.asp

In addition, we are collaborating with Cornell University's Albert Mann Library in their current pilot program, DataStaR (Data Staging Repository).[16] This library project is intended to help researchers to "create high quality metadata in the formats required by external repositories…" (Steinhart, 2010, 1) where metadata systems are cross-disciplinary[17] and flexible, and to provide a temporary repository for data sharing while research is in progress. DataStaR thus fosters an infrastructure for data sharing and preparation for publication to external repositories where available (Steinhart, 2010). The program adopts a semantic web approach to metadata "on the assumption that scientific communities will increasingly adopt semantic web technologies, and that Linked Data will become increasingly common…" (Steinhart (2010, 5); also see Khan et al (2011); Lowe (2009) for details). At present, one VCLA dataset (Sinhala language) from more than 400 children studied in Sri Lanka has been entered in DataStaR, linking the VCLA database to the Library staging repository, and is available for collaborative use through this repository.[18]

As our linguistic coding system develops further, we intend to pursue further integration with attempts at standardization of linguistic category descriptions such as GOLD[19] (Farrar and Langendoen, 2003), also see (Simons et al, 2004); as well as with conventions for glossing such as developed by Bickel et al (2008). We have initiated integration with OLAC (Open Language Archives Community) and are pursuing further development of this link.[20]

7 Current Status and Challenges

At this point we are completing beta-testing of the current newly programmed version of the DTA tool (cf. fn 5). This new version of the tool is currently being populated by data from a previous version of the tool that is now being ported to this new version, and by new data from several current collaborative research and educational projects.

Several cross institutional courses, including two with Peru, at the undergraduate and graduate levels have introduced the VLL and the DTA tool through our structured VLL web portal[21] and a series of web conferences over the last three years in order to train a new generation of scholars in scientific methods for data collection and management.

---

[16] Funded by the National Science Foundation (Grant No. 111-0712989)

[17] DataStaR uses RDF (Resource Description Framework (RDF)) statements and OWL (Web Ontology Language) classes in order to integrate different metadata frameworks across disciplines.

[18] http://datastar.mannlib.cornell.edu/display/n6291 and http://www.news.cornell.edu/stories/Oct11/SinhalaTools.html

[19] Generalized Ontology for Linguistic Description. 2010. Available online at http://www.linguistics-ontology.org/gold.htm

[20] Open Language Archives Community (OLAC), http://www.language-archives.org/ (24 Feb. 2011) (See also Open Archives Initiative (OAI), http://www.openarchives.org/ (15 Mar. 2005)).

[21] The VLL portal is accessible under http://www.clal.cornell.edu/vll. It was constructed by Tommy Cusick, previous Cornell student, now Google, Inc.

The main challenges we must address now include the following: (i) In order to open the DTA tool productively to a wide audience, we must build a sustainability model that includes licensing options. For this we have now initiated correspondence with Cornell's E-Cornell program (eCornell.com); (ii) To widely extend the DTA tool to new users we must establish a set of 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; (iii) To ensure long-term sustainability, we must negotiate long-term storage of the database; (iv) 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.

In addition, several technical challenges remain as the new version of the DTA tool is being activated. For example, porting data from an old version of the DTA tool to the new directly instantiates the general issue of database linkage. We are currently working with Cornell Information Technologies to pursue the possibility for alternative video and audio streaming. A version of the tool allowing asynchronous uploading of data should be developed to facilitate work in field situations, e.g., cross-linguistic work, without Internet access.

Acknowledgements Several grants have provided funding for this project. The main support for our current development comes from Blume, María and Barbara Lust. 2007. Transforming the Primary Research Process Through Cybertool Dissemination: An Implementation of a Virtual Center for the Study of Language Acquisition. National Science Foundation. Award No. OCI-0753415. Other grants that have provided with fundamental resources are: Lust, Barbara. 2003. Planning Grant: A Virtual Center for Child Language Acquisition Research. National Science Foundation. NSF BCS-0126546, and Janet McCue and Barbara Lust 2004-2006. National Science Foundation Award: Planning Information Infrastructure Through a New Library-Research Partnership. (SGER=Small Grant for Exploratory Research). Seed grant support from the following was also essential to this project: American Institute for Sri Lankan Studies, Cornell University Einaudi Center, Cornell University Faculty Innovation in Teaching Awards, Cornell Institute for Social and Economic Research (CISER), New York State Hatch grant. 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); of our application developers Ted Caldwell and Greg Kops (GORGES); of our consultants Cliff Crawford and Tommy Cusick; of our 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. We would also liked to thank the students at all the participating institutions that helped us with comments and suggestions.

References

Berners-Lee T (2009) URL http://en.wikipedia.org/wiki/Linked_data, ted Lecture. Tim Berners-Lee on the next Web

Bickel B, Comrie B, Haspelmath M (2008) Glossing rules: Conventions for interlinear morpheme-by-morpheme glosses. URL http://www.eva.mpg.de/lingua/resources/glossing-rules.php

Blume M, Lust B (2011) Data Transcription and Analysis Tool User's Manual. With the collaboration of Shamitha Somashekar, and Tina Ogden

Blume M, Yang S, Lust B (in prep) Virtual Linguistics Lab (VLL) Research Methods Manual: Scientific Methods for Study of Language Acquisition. With the collaboration of Tina Ogden, Shamitha Somashekar, Yuchin Chien, Liliana Sánchez, Claire Foley, Marina Kalashnikova, Martha Rayas, Yarden Kedar and Natalia Buitrago

Chiarcos C, Hellmann S, Nordhoff S (this vol.) Introduction and overview. P. 1-12

Farrar S, Langendoen D (2003) A linguistic ontology for the semantic web. GLOT International 7:97–100

Khan H, Caruso B, Lowe B, Corson-Rikert J, Dietrich D, Steinhart G (2011) Datastar: Using the semantic web approach for data curation. International Journal of Digital Curation 2:209–221

Lowe B (2009) Datastar: Bridging xml and owl in science metadata management. Metadata and Semantics Research 46:141–150, URL http://www.springerlink.com/content/q0825vj78ul38712/

Lust B, Flynn S, Blume M, Westbrooks E, Tobin T (2010) Constructing adequate documentation for multi-faceted cross linguistic language data: A case study from a virtual center for study of language acquisition. In: Language Documentation: Theory, Practice and Values, John Benjamins, Amsterdam/Philadelphia, pp 127–152

Simons G, Lewis W, Farrar S, Langendoen D, Fitzsimons B, Gonzalez H (2004) The semantics of markup. In: Proc. 4th Workshop on NLP and XML (NLPXML-2004), Barcelona, Spain, pp 25–32

Steinhart G (2010) Datastar: A data staging repository to support the sharing and publication of research data. In: 31st Annual IATUL Conference - The Evolving World of e-Science: Impact and Implicationsfor Science and Technology Libraries, West Lafayette, URL http://docs.lib.purdue.edu/iatul2010/conf/day2/8/