1. Introduction

The current digital age has opened unprecedented opportunities for storing, disseminating, and manipulating complex and limitless data. New digitally based technology has provided sciences in the digital age with power for new discoveries (Borgman 2007) enabled by new methods and new possibilities for collaborative data sharing. Many of the sciences, including the social sciences, are being transformed by these developments. The president’s new Executive Order on Open Data and federal funding agency requirements for data management and data sharing plans appear to assume these developments are in place. In general, there has been new recognition that scientific use of data requires access to both metadata and raw and structured data. In keeping with fundamental insights of the “Linked Data” program (Berners-Lee 2009, Chiarcos et al. 2012), the more each data singleton can be significantly “interlinked”, the more powerful and useful it becomes.

At the same time, our digital age and the technical power it can provide have led to relatively limited advance in the field of language acquisition. New students and researchers are not systematically trained in a way which would provide the foundations of knowledge and skills which such technical power requires. This is in spite of the fact that the major theoretical motivations in this field, e.g. to uncover universal and language-specific characteristics of language development, require comparisons of various forms of data collected in different languages from participants at different ages using different research methods, both observational and experimental, including multiple forms of media, and crossing disciplines, e.g. linguistics, psychology, and neuroscience. Scientific methodology requires that these comparative data analyses be replicable (cf. Johnson 2014), and thus continually accessible and calibrated in data descriptions and analyses. Data must therefore be preserved in a comparable manner across languages and methods and researchers. Collaboration is essential across researchers spanning languages and disciplines.

In our recent work, following on pioneering efforts such as the Open Language Archives (OLAC) (Bird and Simons 2003, Simons et al. 2004) and the CHILDES database (MacWhinney 2000), and supported by the National Science Foundation Cyberinfrastructure CI-TEAM program, we have introduced a community-based approach to enabling such empowerment. As a supplement to efforts like CHILDES, it designs and offers a primary research tool which can contribute to original data collection, management and collaborative analyses. (See below for discussion of differences from CHILDES.) As a supplement to efforts like OLAC, it targets the language acquisition field’s data and its special needs and thus contributes to general metadata development for this project. As a supplement to endeavors in the field which target preservation of endangered languages (e.g., Electronic Metastructure for Endangered Languages Data, emeld.org) and the Language Archive, tla.mpi.nl/resources/archiving-service, it offers principles, methods and tools for data preservation, management and collaborative research.

---

1 University of Texas at El Paso, Boston College, Harvard Medical School and Massachusetts Institute of Technology Program in Speech and Hearing Bioscience and Technology, Massachusetts Institute of Technology, Cornell University
2. A case study for cyberinfrastructure development in the language acquisition field: Building community and technology

2.1. Building community

In recent publications we have described our work to develop a Virtual Center for Language Acquisition (VCLA/www.clal.cornell.edu/vcla) dedicated to the study of both first and second/multiple language acquisition, and consisting of founding faculty and students from eight United States universities and one Peruvian university (see list in acknowledgements). These founding members have converged in a commitment to shared research and collaboration, and to integrating research with teaching of a new generation of scientists in the language acquisition field—scientists trained in data management principles and methods, empowered by new cybertools, and inculcated with a commitment to collaboration.

This community of founding members has created a cyber-enabled international Virtual Learning Environment for the language sciences generally and the study of language development specifically. It trains students and new researchers from different disciplinary, geographical, linguistic and cultural backgrounds to use new cybertools as part of a Virtual Linguistics Laboratory, which prepares them to apply basic scientific methods to the study of language and language acquisition. It teaches methods of data management and a culture of collaboration.

Creating this digital research and learning environment has allowed us to more precisely articulate problems of working scientifically in the field of language acquisition, particularly challenges of data complexity and quality (Blume, Foley, Whitlock, Flynn and Lust, 2012, and in prep). We have laid out the problems of data creation in this complex field, and the challenge of appropriate metadata creation, and in this we have worked with University Librarians at both Cornell University and MIT to establish a metadata structure for management of shared language acquisition data and analyses (Lust, Flynn, Bloom, Westbrooks and Tobin, 2010).

Highlights of program developments are listed below.

2.2. Creation of a Virtual Linguistics Lab

To establish a virtual learning environment, a Virtual Linguistics Lab has been developed and disseminated in a new web portal (www.clal.cornell.edu/vll). The Virtual Lab fosters international interdisciplinary collaboration and provides for courses in scientific methods for the study of language acquisition which address the challenges of the current digital environment. Eight US universities and one Peruvian university have provided the foundation for development and expansion of this Virtual Lab both nationally and internationally, by contributing to and participating in a series of inter-university courses based on its resources.

The Virtual Linguistics Lab includes:

- a series of web-based courses, and accompanying materials in English and Spanish that can be used to teach the same course in several institutions at the same time, or as a resource which professors and students can access at their convenience. They include learning modules focused on research methods with audio-visual samples of actual research being conducted in first, second and bilingual language acquisition. Methodologies for gathering data and for analyzing collected data using various assessments for eliciting language production, comprehension and judgment are collected, and exemplified, with assignments. (See Appendix A for illustrations.)
- Web conferences that have been developed throughout these courses to cultivate real time collaboration.
• a set of materials and tools for the scientific study of language acquisition, including, for example, a set of multilingualism questionnaires (adult and child, Spanish and English versions) and a research methods manual focused on establishing best practices for the study of language development (English and Spanish versions) (Lust and Blume in prep.).

Over time, across the various fields collecting language data, separate databanks have typically been created by individual researchers using different and often irregular procedures for collecting, labeling and storing language data. Crucial metadata is often missing and irretrievable; methods now must be developed post-hoc so these diverse data sources can at least partially be preserved, linked, calibrated and subjected to reliability standards. Only after this do data use and reuse and sound collaborative research become effectively possible. The VLL we have created begins to ameliorate this problem.

2.3. Creation of cybertool

Scientific advance in any field requires development of new tools as well as new theories. The creation of a functional Virtual Linguistics Lab requires development of cybertools to exploit its resources. Thus, a new cybertool has been created by the VCLA to assist the researcher and train students in data management in primary research and in collaborative data analyses: a Data Transcription and Analysis Tool (DTA tool). This cybertool provides a web-based experiment bank to store information about previous and active research (experimental or naturalistic) on language, thus establishing the basis for replicability. It also provides a hierarchical series of fields for entering metadata and transcribing and analyzing language data in a calibrated fashion thus allowing shared data and collaboration. It is distinct from CHILDES in part because of its highly structured interface, which guides the researcher or student through a sequence of metadata and data entry fields. At the same time, it is flexible and allows the researcher to adopt codings and analyses to their specific research questions and programs. We describe this tool in Blume, Flynn and Lust 2012.

We show how its power can be exploited in an educational setting in Blume and Lust 2012. The power of the DTA as a research tool can be illustrated through an example from a study investigating developmental patterns of acquisition in relative clauses across languages based on data from cross-linguistic experiments rendered comparable through the DTA tool. Research questions include:

(1) Do some relative clause types emerge sooner than others in first language acquisition across languages?

(2) What explains cross-linguistic similarities and differences in developmental patterns?

For example, in an elicited imitation study using an experimental design which varied relative clause head types, overall production of free relatives like (3) and (4) did not differ significantly across English and French (Foley 1996).

(3) Big Bird pushes what bumps Ernie

(4) Aladdin choisit ce que Fifi achète
   Aladdin choose-3S ce that Fifi buy-3S
   ‘Aladdin chooses what Fifi buys.’ (Foley 1996, age 4;2)

---

2 Linguists are now urgently confronting the wide set of issues concerning portability of language data in the language sciences (Bird and Simons 2003, Simons 2004, Constable and Simons 2000, Lehmann 2004, Mereu 2004), as are psychologists (Johnson 2001, Johnson and Sabourin 2001.) Therefore, a culture of collaborative materials and data sharing must be cultivated, both socially and technically (e.g., ‘Fair Share’ 2006, Knorr Cetina 1999).
However, target-like lexically headed relative clauses emerged more slowly in English than in French. The DTA tool permitted data to be entered using the same codes across languages, thus allowing queries that compared and contrasted children’s utterances in the two languages. Comparisons within a subset of children between the ages of 4;6 and 4;11 revealed that, consistent with the overall finding, in English only 19% of responses were “correct”, in contrast to 53% for French. At the same time, examining the “incorrect” responses called up by the query for this subset illustrates the power of the DTA tool to call up data for fine-grained qualitative analysis to supplement quantitative experimental results:

- In both languages, there are conversions to free relatives.
- In both languages, “errors” include manipulation of clause-initial elements (e.g., that/which, qui/que) which are more similar to clause-initial elements of free relatives in French than in English—contributing to the smaller number of errors.

These findings are illustrated in Figures 1 and 2.
Figure 1. Sample DTA query: all “incorrect” response for English, ages 4;6-4;11 (Flynn and Lust 1981)

<table>
<thead>
<tr>
<th>Session</th>
<th>Title</th>
<th>Session</th>
<th>Age</th>
<th>Utterance Text</th>
<th>Coding value</th>
</tr>
</thead>
<tbody>
<tr>
<td>01AC0713TH</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>...Would you say it again? (Ex: &quot;OK, now listen hard. Ready?&quot; [R] Herman touches the thing that- that- that- that- that Fozzie... (Ex: &quot;Um hmm... You through?&quot; etc.)</td>
<td>2</td>
</tr>
<tr>
<td>01CH0919TH</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>...Would you say that again? (Ex: &quot;OK, now listen hard. You’re getting all distracted. Just concentrate on the, OK? You ready?&quot; [R]) ... I forget. (Ex: &quot;OK, we can come back.&quot;)</td>
<td>2</td>
</tr>
<tr>
<td>01CH1111TH</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>Big Bird pushes the ballon that- that- Can you say that again? (Ex: &quot;Sure thing.&quot; [R] Big Bird pushes the ballon which- which- touches Ernie.</td>
<td>2</td>
</tr>
<tr>
<td>01CH1111TH</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>Ber[?] pushed the baloon that kick Fozzie. Ernie.</td>
<td>2</td>
</tr>
<tr>
<td>01CH1111TH</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>Ernie touched the baloon... what Big Bird touches.</td>
<td>2</td>
</tr>
<tr>
<td>01CH1111TH</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>Kernit Frog... touches the block... into Fozzie.</td>
<td>2</td>
</tr>
<tr>
<td>01CU1112TH</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>Scooter grabs it the candy what... what... Kernit Frog eats.</td>
<td>2</td>
</tr>
<tr>
<td>01CH1111TH</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>...Ernie throws the bell what Big Bird durnates (touchless!).</td>
<td>2</td>
</tr>
<tr>
<td>01CH1111TH</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>Big Bird... pushes a baloon... sn... went... went into Ernie.</td>
<td>2</td>
</tr>
<tr>
<td>01CH1111TH</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>Big Bird... pushes... gets a candy.</td>
<td>2</td>
</tr>
<tr>
<td>01CH1111TH</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>Fozzie... (Ex offers to do another one) (R) Fozzie bear... exss... (Ex: &quot;Try another one&quot;?...)</td>
<td>2</td>
</tr>
<tr>
<td>01CH1111TH</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>Der... Kernit... he touched the block... n’ hurt. (Ex reminds her to tell the exact same story she tells.)</td>
<td>2</td>
</tr>
<tr>
<td>01CH1111TH</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>Ernie... bump... um... bumps... bump bump bump... Bump, Bump. (Ex: &quot;Do you want me to say it again? At through? Or do you want me to say it again?&quot; Yes. (Ex: &quot;OK, here we go.&quot;&quot;) (R) Big Bird... pushes a big block... hurts Ernie.</td>
<td>2</td>
</tr>
<tr>
<td>01CH1111TH</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>The Big Bird... throws... the hurt... um... um... hurts... Bear... (Ex: &quot;OK, do you want me to say it again? Now listen hard. You ready?&quot;&quot;) (R) Ernie... throws... the... baloon... hurts... um... the Big Bird.</td>
<td>2</td>
</tr>
<tr>
<td>01CH1111TH</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>Um... Fozzie bear... eats... uh... uh... (Ex: &quot;Do you want me to say it again? So you can hear it again? OK, you got to listen hard now, you ready? OK?&quot;&quot;) (R) Fozzie bear... he did some candy... if eats.</td>
<td>2</td>
</tr>
<tr>
<td>01CH1111TH</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>Kernit Frog... bumps the block that - bumps Fozzie.</td>
<td>2</td>
</tr>
<tr>
<td>01CH1011TH</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>(Sigh internally after &quot;Kernit Frog&quot;. Starts over.) Kernit Frog touches the - block what is what? (unintelligible). (Ex: &quot;Do you want me to say it again?&quot;&quot;) Yeah. (Talks to him he can ask for a repeat.) (R) Kernit Frog touches the block what touches Fozzie.</td>
<td>2</td>
</tr>
<tr>
<td>01CH1011TH</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>Bert pushes the thing what bumps Ernie.</td>
<td>2</td>
</tr>
<tr>
<td>01CH1011TH</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>Scooter grabs the candy what Fozzie Bear eats.</td>
<td>2</td>
</tr>
<tr>
<td>01CH1112TH</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>Big Bird... uh... picks up the candy and give it to Ernie &quot;uh&quot; eat.</td>
<td>2</td>
</tr>
<tr>
<td>01CH1112TH</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>Ernie touches the baloon... but... Big Bird (H)nows.</td>
<td>2</td>
</tr>
<tr>
<td>01CH1112TH</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>Ernie... uh... (Ex asks if he wants him to say it again) knows... the... Ernie kisses the block... when? (R) Big Bird touches the baloon. Rob: Fozzie touches the block... an Ernie... uh... into the brick.</td>
<td>2</td>
</tr>
<tr>
<td>01CH1011TH</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>(Note: So shadows the ex as he reads the sentence.) Big Bird what every (Ex: &quot;You have to say it over, not when I say it. You want to hear that one again? Okay, you just remember everything and when I’m through saying it you say it, OK? And so let’s try again, OK?&quot;&quot;) (R) Big Bird... say... throws.</td>
<td>2</td>
</tr>
<tr>
<td>01CH1011TH</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>Ernie bump the baloon in- in... (Ex: &quot;Want me to say it again? OK, here we go.*”) (R) Ernie bumped into the baloon... but he eated... mmm... Ernie sat (Ex: &quot;Are you saying what I said, or any you making it up?&quot;&quot;)</td>
<td>2</td>
</tr>
<tr>
<td>01CH1011TH</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>Fozzie bump the block it touches the Fozzie.</td>
<td>2</td>
</tr>
<tr>
<td>01CH1011TH</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>00:00:00</td>
<td>Scooter... um... grats the candy... what he eats.</td>
<td>2</td>
</tr>
</tbody>
</table>
2.4. Linked Data

Any cybertool which accesses, preserves and exploits data for continual use must be capable of linkage to other internet based addresses, i.e., it must be interoperable. In a recent paper (Pareja Lora, Blume and Lust 2013) we begin to address the challenges of linking the DTA tool to ontologies that formalize its metadata conceptual structure and relate it to others.

3. Conclusions

This is the first project of its kind in the social sciences and humanities, having an international cross-linguistic framework, and thus can serve as a model for other social and cognitive sciences. It empowers a wide array of collaborative and interdisciplinary research and teaching agendas and 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 (in both child and adult).

The Virtual Lab materials and the new courses that they enhance created 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.

Our case study provides evidence that it is possible to create a new generation of scholars, crucially different from previous generations, educated through a Virtual Lab in the use of web-based technology, not typically part of the training in the language sciences. These scholars value large-scale, long-distance collaboration, are able to manage and reuse large amounts of scientifically validated data, and contribute to research projects that may be more ambitious because of superior collaborative and data management skills. This empowers the
language sciences and allows us to tackle crucial questions of the Cognitive Sciences in an interdisciplinary manner, and to resolve the tremendous gap between immense amounts of cross-linguistic language data and limited research resources, which we now face, as does science at large (Berman and Cerf 2013).

Acknowledgements

This project was supported by several funding sources: “Transforming the Primary Research Process Through Cybertool Dissemination: An Implementation of a Virtual Center for the Study of Language Acquisition.” National Science Foundation grant to Maria Blume and Barbara Lust., NSF OCI-0753415; “Planning Grant: A Virtual Center for Child Language Acquisition Research. National Science Foundation grant to Barbara Lust., NSF BCS-0126546; “Planning Information Infrastructure Through a New Library-Research Partnership.” National Science Foundation Small Grant for Exploratory Research to Janet McCue and Barbara Lust; American Institute for Sri Lankan Studies, Cornell University Einaudi Center; Cornell University Faculty Innovation in Teaching Awards, Cornell Institute for Social and Economic Research; New York State Hatch grant; Grant Number T32 DC00038 from the National Institute on Deafness and Other Communication Disorders (NIDCD).

We gratefully acknowledge the collaboration of the Virtual Center for Language Acquisition’s other founding members who have contributed over the years: James Gair, Claire Cardie, Qi Wang and Marianella Casasola (Cornell University); Elise Temple (NeuroFocus); Liliana Sánchez (Rutgers University at New Brunswick); Jennifer Austin (Rutgers University at Newark); YuChin Chien (California State University at San Bernardino); and Usha Lakshmanan (Southern Illinois University at Carbondale). We are grateful for the collaboration of scholars who are VCLA affiliates Gita Martohardjono, Valerie Shafer, and Isabelle Barrière (City University of New York); Cristina Dye (Newcastle University UK); Yarden Kedar (The Center for Academic Studies, Israel); Sujin Yang (Singapore Management University), Joy Hirsch (previously Columbia University); Ellen Courtney and Alfredo Urzúa (University of Texas at El Paso); Sarah Callahan (University of California at San Diego); Jorge Iván Pérez Silva (Pontificia Universidad Católica Del Perú), Kwee Ock Lee (Kyungsung University); K.V. Subbarao (University of Delhi India emeritus), R. Amritavalli (Central Institute of English and Foreign Languages, India); A. Usha Rani (Osmania University).

We thank application developers Ted Caldwell and Greg Kops (GORGES); consultants Cliff Crawford and Tommy Cusick; student research assistants Darlin Alberto, Gabriel Clandorf, Natalia Buitrago, Poornima Guna, Jennie Lin, Marina Kalashnikova, Martha Rayas Tanaka, Lizzeth Pattison, María Jiménez, and Mónica Martínez; research associate Alicia Ah Young Kim, and the students at all the participating institutions who helped us with comments and suggestions.

References


Blume, M. & Lust, B. 2012. First steps in transforming the primary research process through a Virtual Linguistic Lab for the study of language acquisition and use: Challenges and accomplishments. *Journal of Computational Science Education* 3, 1, 34-46.


Child Language Data Exchange System (CHILDES) [http://childes.psy.cmu.edu/](http://childes.psy.cmu.edu/)


Electronic Metastructure for Endangered Languages Data (EMELD) [emeld.org](http://www.emeld.org)


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


Appendix A. Illustrations of example Virtual Lab training modules

Above: UTEP and Cornell students interacting during first course. Right: A UTEP student presents her research proposal to students and faculty in Pontificia Universidad Católica del Peru.