

A New Data Management Cybertool Supports

Cross-linguistic Collaborative Research and Student Training

María Blume, Claire Foley, Jordan Whitlock, Suzanne Flynn, Barbara Lust¹

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, Westbrook 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 possible². 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)

Query Results (26 records)

Session: Title	Session: Age	Utterance: Text	Coding: value
01AC071574	00;00;00	...Would you say it again? (Ex: "OK, now listen hard. Ready?") (R) Herman touches the thing that- that- that- that- that... that Fozzie... (Ex: "Um hm?...You through?" etc.)	2
		...Would you say that again? (Ex: "OK now listen hard. You're getting all distracted. Just concentrate on this, OK? You ready?") (R)I forget. (Ex: "OK, we can come back.")	2
		Big Bird pushes the balloon that- that... Can you say that again? (Ex: "Sure thing.") (R) Big Bird pushes the balloon which... which- touches Ernie.	2
01CH091974	00;00;00	Bert(?) pushed the balloon that kick Burnie, Ernie.	2
		Ernie touched the balloon... what Big Bird touches.	2
		Kermit Frog... pushes the block... hits Fozzie.	2
		Scotie... grabs it the candy what... what... Kermit Frog eats.	2
01CH111174	00;00;00	... Ernie throws the ball what Big Bird dunstes (touches?).	2
		Big Bird... pushes a balloon... an... went... went on Ernie.	2
		Big Bird... pushtes... gets a candy.	2
		Fozzsss... (Ex offers to do another one) (R) Fostie Bear... exsss... (Ex: "Try another one?")....	2
01CZ111574	00;00;00	Der... Kermit... he touched the block... n' hurt. (Ex reminds her to tell the exact same story she tells.)	2
		Ernie... bumps... um... bumps- bump (Ex: "Bump?") Bump. (Ex: "Do you want me to say it again? All through? Or do you want me to say it again?") Yes. (Ex: "OK, here we go.") (R) Big Bird... pushes... a big b'oon... hurts Ernie.	2
		The Big Bird... throws... the hurts... um... um... hurts... Bear. (Ex: "OK, do you want me to say it again? Now listen hard. You ready?") (R) Ernie... throws... the... balloon... hurts... um... the Big Bird.	2
		Um... Fozzie Bear... eats... uh... uh... (Ex: "Do you want me to say it again? So you can hear it again? OK, you got to listen hard now, you ready? OK.") (R) Fozzie Bear... he did some candy... n' eats.	2
01EH000000	00;00;00	Scotie bumps the thing that - (R) Kermit Frog bumps the block that - bumps Fozzie.	2
01GH111574	00;00;00	(Su interrupts after "Kermit Frog". Starts over.) Kermit Frog touches the- block what- what what (unintelligible). (Ex: "Do you want me to say it again?") Yeah. (Tells him he can ask for a repeat.) (R) Kermit Frog touches the block what touches Fozzie.	2
		Bert pushes the thing what bumps Ernie.	2
		Scotie grabs the candy what Fozzie Bear eats.	2
01JC112674	00;00;00	Big Bird... uh... pickups the candy and give its to Ernie "-duh" eat.	2
		Ernie touches the balloon..."bu"...Big Bird (th)rows.	2
		Ernie... uh...(Ex asks if he wants him to say it again) kisses...the b... Ernie kisses the block... when (?) Big Bird touches the balloon. Redo: Fozzie touches the block... an' Ernie... uh... hits the block.	2
01MK120874	00;00;00	(Note: Su shadows the Ex as she reads the sentence.) Big Bird what frow (Ex: "You have to say it after me, not when I say it. You want to hear that one again? OK you just remember everything and when I'm through saying it you say it, OK? And so let's try again, OK?") (R) Big Bird... ser... throws.	2
		Ernie bump the balloon in-to... (Ex: "Want me to say it again? OK, here we go.") (R) Ernie bumped into the balloon... but he eated it... mm... Ernie sat (Ex: "Are you saying what I said, or are you making it up?" etc.)	2
		Fozzie bumps the block n' touches the Fozzie.	2
		Shootie...ea...grabs the candy... what he eats.	2

© Copyright 2010–2012 Barbara Lust and Maria Blume

Figure 2. Sample DTA query: all “incorrect” response for English, ages 4;6-4;11 (Foley 1996)

Query Results (15 records)

Session: Title	Session: Age	Utterance: Text	Coding: value
1AL101589	04;10;24	Donald goute la soupe que Mickey aime	0
1CD090189	04;09;21	Aladdin il goute la chose...[op] que Mickey aime	0
1ES091389	04;11;09	Aladdin goute la soupe de Mickey	0
		Donald fait le dessin que Pierre aime bien	0
		Pierre cherche la balle que Pierre lance	0
1GP012189	04;10;01	Aladdin, goute la soupe, de Mickal aime	0
		Donald fait le dessin pour amuser Tintin	0
		Gargamel lance la balle	0
		Mickal, lit le livre, amuser, Fifi	0
1LD111889	04;09;21	Aladdin goute la soupe que Ladin	0
		Aladdin, choisit la chose quay Fifi achète	0
1LH110489	04;07;18	Donald fait le dessin c'qui intéresse, Tintin	0
		Gargamel lance la balle que Donald lance	0
		Mickey prend c'qui amuse Fifi	0
1PD080189	04;10;22	Donald fait le dessin qu'intéresse Tintin	0

© Copyright 2010–2012 Barbara Lust and María Blume

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 María 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 (Kyungsoong 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


- Berman, F & Cerf, V. 2013. Who will pay for public access to research data? *Science* 341, 616-617.
- Berners-Lee, T. 3/2009. TED Lecture. Tim Berners-Lee on the next Web.
http://en.wikipedia.org/wiki/Linked_data.
- Bird, S, & Simons, G. 2003. Seven dimensions of portability for language documentation and description. *Language* 79: 557-582.

- Blume, M., Flynn, S. & Lust, B. 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 and connecting language data and language metadata*. Berlin/Heidelberg: Springer, 85-96.
- Blume, M., Foley, C., Whitlock, J., Flynn, S. & Lust, B. 2012. Principles and new cybertools for interlinking data in the study of language acquisition: Leveraging the advantages of a digital environment. Georgetown University Round Table on Languages and Linguistics. Georgetown University, DC, March 10, 2012.
- 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.
- Borgman, C. L. 2007. *Scholarship in the digital age*. Cambridge: MIT Press.
- Chiarcos, C., Nordhoff, S., & Hellmann, S. (Eds.) 2012. *Linked data in linguistics: Representing and connecting language data and language metadata*. Berlin/Heidelberg: Springer.
- Child Language Data Exchange System (CHILDES) <http://childes.psy.cmu.edu/>
- Constable, P. & Simons, G. 2000. Language identification and IT: Addressing problems of linguistic diversity on a global scale. *SIL Electronic Working Papers* 2000-2001. Dallas: SIL International (<http://www.sil.org/silewp/2000/001>).
- A Fair Share. 2006. *Nature* 444, 7120, 653-654.
- Knorr Cetina, K. 1999. *Epistemic cultures: How the sciences make knowledge*. Cambridge, MA: Harvard University Press.
- Electronic Metastructure for Endangered Languages Data (EMELD) emeld.org
- Flynn, S., and Lust, B. 1981. Acquisition of relative clauses: Developmental changes in their heads. In W. Harbert and J. Herschednsohn (Eds.), *Cornell Working Papers in Linguistics*. Ithaca, NY: Dept. Modern Languages and Linguistics, Cornell University.
- Foley, C. 1996. *Knowledge of the syntax of operators in the initial state: The acquisition of relative clauses in French and English*. Ph.D. dissertation, Cornell University.
- Johnson, D. 2001. Three Ways to Use Databases as Tools for Psychological Research. *APS Observer* 14, 10, 7-8.
- Johnson, D. & Sabourin, M. 2001. Universally accessible databases in the advancement of knowledge from psychological research. *International Journal of Psychology* 36, 3, 212-220.
- Johnson, G. 2014. New truths that only one can see. *New York Times*, January 20.
- Lehmann, C. 2004. Data in linguistics. *The Linguistic Review* 21, 3-4, 175-210.
- Lust, B. & Blume, M. (with collaboration of C. Dye, Y. Kedar and founding members of Virtual Center for Language Acquisition). In prep. *Virtual Linguistics Lab (VLL) Research Methods Manual: Scientific Methods for Study of Language Acquisition*. American Psychological Association.
- Lust, B., Flynn, S., Blume, M., Westbrooks, E., & Tobin, T. 2010. Constructing adequate documentation for multi-faceted cross-linguistic and language data: A case study from a Virtual Center for study of language acquisition. In Furbee, L, and Grenoble, L, eds., *Language documentation: Practice and values*. Philadelphia: John Benjamins, 127-152.

- MacWhinney, B. 2000. *The CHILDES Project: Tools for analyzing talk*. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates.
- Mereu, L. 2004. Linguistic data as complex items. *The Linguistic Review* 21, 3-4, 211-232.
- Open Language Archives Community (OLAC), <http://www.language-archives.org/>
- Pareja-Lora, A., Blume, M., & Lust, B. 2013. Transforming the DTA tool metadata and labels into a linguistic linked open data cloud resource. In Cimiano, P., McCrae, J., Chiarcos, C., and Declerck, T., eds., *Proceedings of Workshop on Linguistic Linked Data*, Pisa.
- Simons, G. F., Farrar, S. O., Fitzsimons, B., Lewis, W.D., Langendoen, D.T., & Gonzalez, H. 2004. The semantics of markup: Mapping legacy markup schemas to a common semantics. *Proceedings of the 4th Workshop on NLP and XML (NLPXML-2004): Held in cooperation with ACL-04*. Barcelona, Spain, 25-32.


Appendix A. Illustrations of example Virtual Lab training modules

Administration of Battery A
The procedures for administering Battery A are outlined.




After introducing the props, explain to the child that you will read the child a story, and that you'd like the child to act the story out, like a puppet show. Explain that you would like all the toys and props put back at the end of the story, so that you know the story is done. Praise the child after each action, even if the action is incorrect.

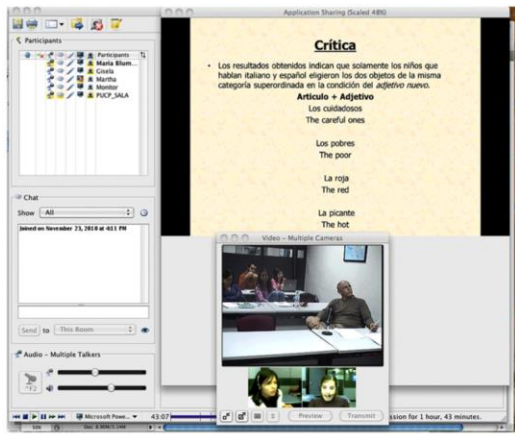
PP112796: Test



Information about the subject can be found [here](#).



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.

- Two samples of A/V module:
1. Teaching experimental procedures
 2. An actual experiment conducted in Peru.