Development of Sign Language Acquisition Corpora

Diane Lillo-Martin[*], Deborah Chen Pichler[#]
[*]University of Connecticut and Haskins Laboratories; [#]Gallaudet University
Storrs, CT 06269-1145; Washington, DC 20002
E-mail: lillo.martin@uconn.edu, deborah.chen.pichler@gallaudet.edu

Abstract

Longitudinal, spontaneous production data have long been a cornerstone of language acquisition studies, but building corpora of sign language acquisition data poses considerable challenges. Our experience began with the development of a sign language acquisition corpus more than 15 years ago and has recently included a small-scale experiment in corpus sharing between our two research groups. Our combined database includes regular samples of deaf and hearing children between the ages of 1;06 to 3;06 years acquiring ASL as their native language. The process through which we generate and share transcripts has undergone dramatic changes, always with the triple goal of creating transcripts with sufficient information for the reader to locate regions of interest, while keeping the video fully accessible and minimizing the time required to generate transcripts. In this paper we summarize the various incarnations of our transcription system, from simple Word documents with minimal integration of video, to a combination of FileMaker Pro software integrated with Autolog, to a fully integrated transcript+video package in ELAN. Along the way, we discuss the potential of ELAN to surmount several obstacles that have traditionally stood in the way of large-scale corpus sharing in the sign language acquisition community.

1. Longitudinal Spontaneous Production Corpora in Language Acquisition Research

Longitudinal, spontaneous production data have long been a cornerstone of acquisition studies, offering a wealth of information on the processes by which children develop language. Research based on longitudinal spontaneous production data has already led to significant discoveries about the acquisition of a number of languages.

Spontaneous production data provide several advantages to the researcher. (a) A particular child participant is observed in a natural environment, interacting with people she is very familiar with. (b) The child's development over a period of time is carefully documented. (c) Researchers working within a wide variety of theoretical frameworks are able to use such data to address a great range of theoretical issues. For example: (d) The researcher can use the data to address research hypotheses concerning acquisition sequence (the ordering of constructions hypothesized to have particular pre-requisites) and hypotheses concerning simultaneous acquisition (constructions related by an underlying common principle). (e) The researcher can investigate hypotheses about non-target structures used by the child. (f) The input provided to the child can also be sampled and studied, when the child is recorded interacting with a parent. Useful resources on the use of spontaneous production data are found in Stromswold (1996) and Snyder (2007), among others.

The development and distribution of multiple corpora of child language data through the CHILDES project (MacWhinney 2000) has resulted in thousands of publications (see http://childes.psy.cmu.edu/bibs/). Corpus sharing through CHILDES allows multiple researchers to independently examine the same data, making it possible to test each other's analyses for reliability, or propose alternative approaches to interpreting the data. Such activity greatly increases the scientific rigor of the research community.

For the sign language acquisition community, corpus sharing on the scale of CHILDES is still far in the future, due to a number of challenges that we discuss below. However, experiments in smaller-scale corpus sharing can allow us to begin addressing these challenges now. In this paper, we will discuss our experiences in small-scale sharing of sign language longitudinal corpora between our respective research groups at the University of Connecticut and Gallaudet University. We will point out some of the difficulties we have encountered over the course of our collaboration, and the modifications we have adopted in response to them. Although many challenges remain, we are encouraged by the success of our experiment so far and by the enormous potential benefits of corpus sharing for the field of sign language acquisition.

2. Challenges of Creating and Sharing Sign Language Acquisition Corpora

Research on sign language acquisition has expanded significantly over the last thirty years, propelled in large part by a growing number of video corpora of signing children. Corpora have now been developed for an ever-increasing number of natural sign languages, creating the potential for fruitful cross-subject and cross-linguistic comparison. Yet much of what we know about sign language development remains limited to a small number of reports on a very small number of children. For example, our collective understanding of early word order acquisition in American Sign Language

(ASL) was for decades determined by a single study based on longitudinal data from three deaf children. Furthermore, details on how data are coded and analyzed are often unavailable, making it impossible for other researchers to test the reliability of analyses presented in the literature. In short, the sign language acquisition community has yet to enjoy the important benefits of corpus sharing that databases such as CHILDES have brought to spoken language acquisition researchers.

There are a number of reasons why sharing of sign acquisition corpora has been slow to catch on. One is that the filming of signing children and their families raises extra concerns about confidentiality. The faces of children and their families must be clearly visible for linguistic analysis to be possible (i.e. their identities can not be concealed by masking or distorting their faces). This increases the already high probability that subjects will be recognized by members of the research community, which draws heavily from comparatively small Deaf communities. Recently, we have noted that video is quickly replacing audio and written records as the standard for acquisition studies, for spoken as well as signed languages, and with this trend has come a general increased tolerance for the inevitable accompanying loss of anonymity. Still, acquisition researchers using video corpora have the responsibility of ensuring as high a degree of confidentiality for their subjects as possible, and this poses a challenge for which we cannot offer any solution at this time.

Instead, we will focus our discussion on two other major obstacles to sharing sign corpora. The first is an absence of standards for annotation or transcription of signed data. Although CHILDES supports the Berkeley Transcription System (Slobin et al. 2001) as a new standard for sign transcription, this system has not been universally adopted, and researchers continue to employ a wide variety of largely idiosyncratic notational conventions. This has made cross-corpus comparisons difficult, if not impossible. Second, our field has until recently lacked a standardized system for efficiently linking annotation or transcription files to large amounts of video data. In the next sections of this paper, we summarize the various ways in which we have addressed these two concerns for our sign acquisition corpus over the past decade.

3. Our Corpus

Although we refer to our corpus as a single entity in this paper for ease of exposition, it is actually composed of two distinct sets of naturalistic, longitudinal corpora, one focused on Deaf signers and the other on hearing, bimodal bilingual (coda) signers. The former was collected as part of the *Cross-Linguistic Early Syntax Study* (CLESS) at the University of Connecticut, Department of Linguistics. This project, funded by the U.S. National Institutes of Health (NIH), and the National Science Foundation (NSF), supported data collection of early child English, Spanish, Japanese, and ASL (Lillo-Martin & Snyder 2002). Over the years, data from the acquisition of Russian, Brazilian Portuguese, and Brazilian Sign Language (LSB) have also been included in the project. For this paper, we will focus on the ASL corpus, which includes data from Deaf children (ages 1;6-3;6) acquiring ASL from Deaf, signing parents, and Deaf children of hearing parents (ages 5;9-10;0) whose exposure to ASL began only after the age of five years. Data from hearing, bimodal bilingual children (ages 1;6-4;0) acquiring both ASL from Deaf, signing parents and spoken English are currently being collected as part of a separate project, *Effects of bilingualism on word order and information packaging in ASL*, at Gallaudet University, Department of Linguistics.

As is clear from the names of our projects, our initial focus has been on early syntactic development, beginning at or before the point when children first combine words into 2-word phrases. A great deal of syntactic development occurs within two years from this point, so we extend data collection until the children are about 3-and-a-half to 4-years old. The children are/were filmed regularly (in most cases, on a weekly basis) for about 30-60 minutes at a time.

Information about the age range and amount of data collected for each child in our combined corpora is shown in Table 1. (NB: as coda data collection is still in progress, the information for those children is projected.)

	Child	Age Range	#Sessions	#Hours (approx)
ASL				
D/D	Abby	1;05 – 3;04	79	75
	Jill	1;07 – 3;07	77	79
	Ned	1;05 – 4;02	44	40
	Sal	1;07 – 2;10	18	16
D/H	Cal	6;10 – 10;01	115	50
	Mei	6;07 – 10;0	111	50
H/D	Ben	1;04 – [4;04]	[100]	[80]
	Tom	1;04 – [4;04]	[100]	[80]
	Pete	1;07 – [4;07]	[100]	[80]

Table 1: Data collection – ASL participants.

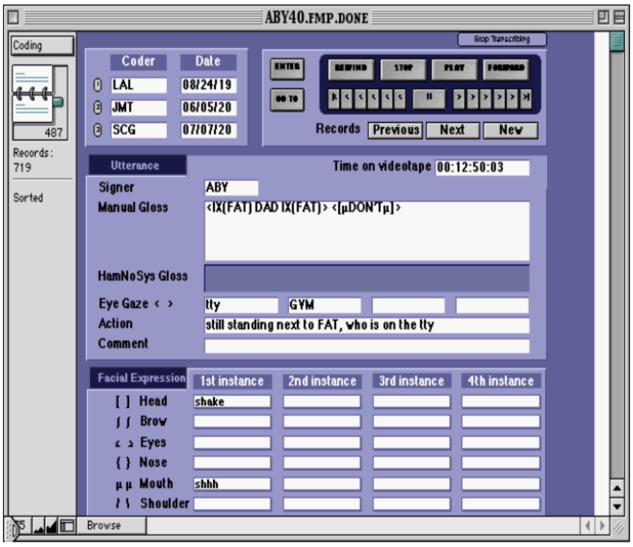
4. Early Transcription System

The very first incarnation of our sign transcripts took the form of Word documents, with a format patterned loosely after the CHAT format used in the CHILDES database.

Each child or adult utterance appeared on its own line, accompanied by information about context and phonological form. Time code was noted every ten lines or so to help users of the transcripts locate regions of interest. However, as may researchers found, entering time code was tedious and actually did very little to facilitate data mining, since video data was stored on analog VHS tapes that had to be manually rewound and fast forwarded to find specific utterances. Furthermore, the video integration of this early system was rather unwieldy, so we quickly sought a way to allow easier and more rapid access to video.

5. FMP+Autolog System

The next incarnation of our transcription system featured File Maker Pro (FMP) software integrated with Autolog, a program that allowed us to control a VCR via the computer and link each utterance to its corresponding SMPTE time code on the VHS tape. We designed different interfaces (coding screens) so that different transcribers could focus on the children's sign utterances, adult utterances, non-manual markers (for both child and adult utterances), and non-linguistic context (action and comments). A screen shot of one coding screen, with spaces for all of this information, showing one sample child utterance, is given in Figure 1.



	Coder	Date
1	LAL	08/24/19
2	JMT	06/05/20
3	SCG	07/07/20

Utterance

Time on videotape 00:12:50:03

Signer: ABY

Manual Gloss: <IX(FAT) DAD IX(FAT)> <[μDON'Tμ]>

HamNoSys Gloss:

Eye Gaze < >: tty | GYM

Action: still standing next to FAT, who is on the tty

Comment:

Facial Expression	1st instance	2nd instance	3rd instance	4th instance
[ ] Head	shake			
∫ ∫ Brow				
‹ › Eyes				
{ } Nose				
μ μ Mouth	shhh			
/ \ Shoulder				

Figure 1: FMP Screen Shot

Drop-down menus and semi-automated time code grab made the job of transcribing easier and greatly decreased the time required to generate transcripts. In addition, FMP included useful features for searching and sorting data, and could print out sections of transcripts for quick reference. Most importantly, this system dramatically increased the speed and ease with which we could locate video for specific utterances of interest, leading to more accurate data analysis.

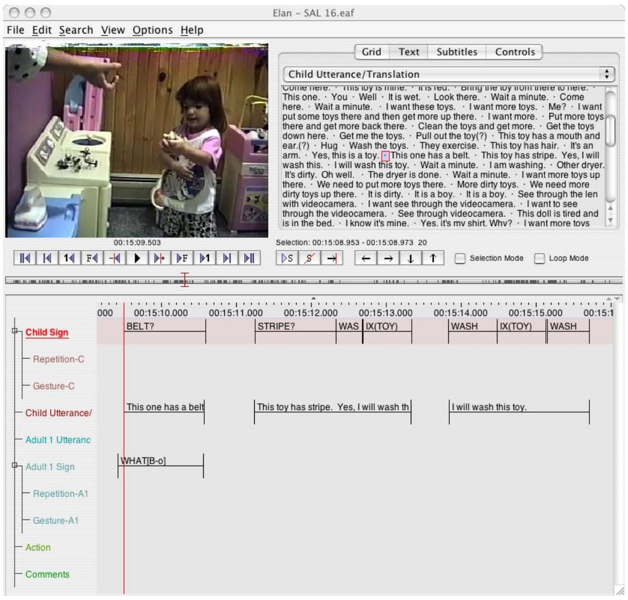
Unfortunately, this system required access to Autolog and specially-modified VCRs, and was not widely adopted by linguistics researchers. Furthermore, the VHS tapes on which data was stored and viewed deteriorated quickly from heavy use, making it necessary to dub copies. If subsequent copies started a few seconds later than the original, all our time code stamps on the corresponding transcript would then be off, a small annoyance that eventually caused considerable inconvenience for analysis.

6. ELAN System

More recently, we have converted our transcription system to ELAN (http://www.lat-mpi.eu/tools/elan/), which enables our transcripts to be time-locked to corresponding digitized video data. This makes the relation between the transcription and the video image much tighter than in either of the previous systems and eliminates the problematic dependence on tape media.

We continue to use traditional upper-case English glosses for transcribing ASL signs, a convenient system, but one with well-known limitations. Nevertheless, we chose this system for its relative readability and ease of use. We keep a running list of glossing conventions, developed and modified through transcriber discussion, that ensures relatively consistent use of glosses across transcribers. However, we recognize the limits of this system, and are willing to accept them only because it is now so easy to consult the video for any given utterance in any transcript in ELAN.

Although ELAN also offers drop-down menus ("controlled vocabularies") and other time-saving features for transcription such as on-the-fly segmentation, generating transcripts is still a time-consuming endeavor, and we have elected to focus our attention on manual activity only, leaving our nonmanul tiers blank for the time being. Again, thanks to the tight integration of transcript and video, researchers will still have access to nonmanual information by watching the video. A screen shot from a sample ELAN transcript is provided in Figure 2.



Come here. · This toy is mine. · It is red. · Bring the toy from there to here. · This one. · You · Well · It is wet. · Look there. · Wait a minute. · Come here. · Wait a minute. · I want these toys. · I want more toys. · Me? · I want put some toys there and then get more up there. · I want more. · Put more toys there and get more back there. · Clean the toys and get more. · Get the toys down here. · Get me the toys. · Pull out the toy(?) · This toy has a mouth and ear.(?) · Hug · Wash the toys. · They exercise. · This toy has hair. · It's an arm. · Yes, this is a toy. · This one has a belt. · This toy has stripe. Yes, I will wash this. · I will wash this toy. · Wait a minute. · I am washing. · Other dryer. It's dirty. Oh well. · The dryer is done. · Wait a minute. · I want more toys up there. · We need to put more toys there. · More dirty toys. · We need more dirty toys up there. · It is dirty. · It is a boy. · It is a boy. · See through the len with videocamera. · I want see through the videocamera. · I want to see through the videocamera. · See through videocamera. · This doll is tired and is in the bed. · I know it's mine. · Yes. it's my shirt. Why? · I want more toys

This one has a belt

This toy has stripe. Yes, I will wash th

I will wash this toy.

Figure 2: ELAN Screen Shot

With ELAN, corpus sharing among sign language researchers is finally becoming an attainable reality. Because the video is integrated so completely with the transcript, idiosyncratic notation conventions no longer pose as great an obstacle as they once did; researchers using transcripts generated by another research team can readily see which signs are represented by which glosses, then modify them with convenient Find and Replace features. ELAN offers a host of powerful features to facilitate annotation and analysis, yet is free to all researchers. In addition, it is available in both Mac and PC versions and is compatible with a variety of commonly used video file formats. These attractive features have established ELAN as the new standard for sign annotation. The existence of a widely-used standard has already opened the door for corpus sharing of adult sign data (eg. the European Cultural Heritage Online/ECHO site for Case 4: Sign Languages at http://www.let.ru.nl/sign-lang/echo/), paving the way for similar sharing of child sign data in the near future.

7. Continuing Challenges

Our experience of generating and sharing sign acquisition transcripts between our two research groups has been very promising so far, but of course, certain challenges remain. Some of these challenges are common to anyone in the business of generating corpora. Transcription of naturalistic video remains a long and tedious job, despite the welcome improvements that ELAN has brought us. We also struggle with decisions of how much of the video data to transcribe. We aim to generate transcripts that are as neutral as possible with respect to analysis, but invariably, the way we choose to gloss a sign or assign utterance breaks will reflect the analysis of the transcriber.

Others challenges are perhaps unique to naturalistic first language acquisition work. For instance, we film children as they play, which means that they are constantly in motion. To increase our chances of keeping the children on camera, we avoid tight shots. The result is that the children's hands and faces look very small on the ELAN video, even when viewed in detached mode. To maximize video resolution in ELAN, we generally rely on H.264 compression rather than the .mpg format that is the default standard for ELAN. Whereas .mpg versions of our video files would be relatively small and portable, our H.264 video files are over 1GB each, posing difficulties for storage and transfer of data. We are optimistic, however, that solutions to this and other challenges lie in technological advances that we have yet to exploit. For example, we are currently in the process of establishing a server to house our corpus data at Gallaudet, which would allow us to overcome the challenge of sharing large video files. For us, the benefits of corpus sharing between our two research groups have clearly outweighed the challenges, and we will continue to seek ways to streamline and refine the process.

8. Conclusion

In this short paper, we have traced the evolution of our sign acquisition corpora from a transcript-centered format with little video integration to a fully integrated transcript and video format. Our goal has always been to efficiently create transcripts that are rich enough for the reader to locate regions of interest by scanning text. At the same time, we require ready access to video, to mitigate the limitations of English-based glosses and guard against analyses based on the transcripts alone. ELAN has facilitated enormous progress towards these goals and made it possible for our two research groups to share our sign acquisition corpora with great success. In the long term, we are hopeful that experiments in small-scale corpus sharing such as ours will one day lead to sharing on a much broader scale, of the type currently available for the spoken language acquisition community through online databases such as CHILDES. As mentioned at the start of this paper, many of the seminal studies on early sign language development are based on tiny sample sizes of two or three children. Given the enormous resources of time and money required to collect and code longitudinal acquisition data, shared databases are absolutely crucial to achieving larger sample sizes, which will permit replication and expansion of basic studies as well as increased possibilities for statistical analyses.

9. Acknowledgements

Collection of monolingual ASL data at the University of Connecticut was supported by grants to Diane Lillo-Martin from the National Institutes of Health, National Institute on Deafness and Other Communication Disorders, Grant #DC-00183; the National Science Foundation, Division of Behavioral and Cognitive Sciences, Grant #BCS-0078788; and the University of Connecticut Research Foundation. The collection of bimodal bilingual data by Deborah Chen Pichler and her research assistants is supported by a priority grant from the Gallaudet Research Institute. Both groups wish to thank the signing children and their families who have graciously allowed us to film them over the last fifteen years, as well as the numerous research assistants who have been crucially involved in collecting and particularly in transcribing the data.

10. References

Lillo-Martin, D., and Snyder, W. (2002) Cross-linguistic study of early syntax. http://web.uconn.edu/acquisition/papers.html

MacWhinney, B. (2000). The CHILDES project: Tools for analyzing talk. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Stromswold, K. (1996). Analyzing children’s spontaneous speech. In. D. McDaniel, C. McKee, & H. Smith Cairns (Eds.), *Methods for Assessing Children’s Syntax*. Cambridge, MA: MIT Press, pp. 23-53.

Slobin, D. I, Hoiting, N., Anthony, M., Biederman, Y., Kuntze, M., Lindert, R., Pyers, J., Thumann, H., & Weinberg, A. (2001). Sign language transcription at the level of meaning components: The Berkeley Transcription System (BTS). *Sign Language & Linguistics*, 4(1-2), pp. 63-104.

Snyder, W. (2007). *Child Language: The Parametric Approach.* Oxford University Press.