CHAPTER 3

Methodological considerations for the development and use of sign language acquisition corpora*

Ronice Müller de Quadros, Diane Lillo-Martin
and Deborah Chen-Pichler
Federal University of Santa Catarina / University of Connecticut /
Gallaudet University

This chapter discusses the building of sign language acquisition corpora. We have developed methodology to collect, transcribe and store data from different contexts of acquisition. The corpora include: deaf children from deaf parents, deaf children from hearing parents, hearing children from deaf parents (Kodas) and deaf children with cochlear implants – all in the contexts of two sign languages – Brazilian Sign Language and American Sign Language, and two spoken languages in the bilingual bimodal cases: Brazilian Portuguese and American English. In this paper we also present the notion of Sign ID, software to indicate identities for each sign that is part of the database. It helps us make the annotations more consistent across transcribers. This kind of work is making it possible to compare data from the acquisition of these four languages.

1. Introduction

In order to address numerous linguistic research questions, we have been building several corpora of sign language acquisition data (Quadros e Pizzio, 2007; Lillo-Martin & Chen Pichler, 2008). Until recently, our focus had been on sign language acquisition for deaf children ages 1 to 4, from deaf parents, acquiring a sign language as a native language. For this research, we built corpora of longitudinal data collected over

* Research reported in this publication was supported by the National Institute on Deafness and other Communication Disorders of the U.S. National Institutes of Health under award number R01DC00183 and R01DC009263. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Support was also provided by a Gallaudet University Priority Grant; and by the Brazilian National Council for Research, CNPq Grant #CNPQ # 304102/2010-5 and # 471478/2010-5.

a period of several years: these corpora included spontaneous production data, with interactions between the child and an adult (usually the Deaf mother or a Deaf experimenter). To a lesser extent, we have also studied deaf children with hearing parents, to see possible effects of differences in the linguistic environment. On the Brazilian side, these children received fairly early input and were observed in the same age range (Quadros, 1997, 2010; Quadros & Lillo-Martin, 2009). In the U.S., children whose input was significantly delayed were studied between the ages of 6–10 (Lillo-Martin & Berk, 2003; Berk & Lillo-Martin, 2012).

Most recently, we have been studying two groups of bimodal bilinguals – children who are bilingual, with each language using a different modality (hence, bimodal). The first group comprises hearing children with deaf parents. These children receive input in sign language and spoken language, and grow up as native bimodal bilinguals. The second group are deaf children who use sign language, but they also have a cochlear implant (CI) and develop spoken language. This project is described in Chen Pichler et al. (2010), Chen Pichler et al. (submitted), Davidson et al. (2014), Lillo-Martin et al. (2010), Lillo-Martin et al. (2012), Quadros et al. (2010), Quadros et al. (2013), among others.

The analyses done so far in our own work – like work by others – indicates that in the specific context of deaf children with deaf parents, sign language acquisition is parallel to spoken language acquisition (see Lillo-Martin, 1999, 2009, and Newport and Meier, 1985 for reviews of some of this research). In these specific contexts, Deaf and hearing children acquire language in steps that reflect their growing understanding of the language used around them. The specific areas analyzed in these studies are related to various grammatical structures, as well as interactional aspects. These studies support the view that sign languages are full languages, on a par with spoken languages. Using similar data, Petitto (2000) draws the strong conclusion that "Deaf children exposed to signed languages from birth acquire these languages on an identical maturational time course as hearing children acquire spoken languages."

However, there are also findings showing that certain aspects of language acquisition in this context show modality effects (e.g. Meier & Newport, 1990; Marentette & Mayberry, 2000; Meier, 2006). These researchers found some specific aspects of sign languages can affect acquisition, because of factors such as simultaneity, iconicity, grammatical use of space and the requirement for visual accessibility (see a summary of current studies in Chen Pichler, 2012, and Chen Pichler et al., in press).

On the other hand, in the context in which a deaf child has limited contact with sign language, there is a lot of variability in the language development reported by different researchers. Children who do not receive accessible linguistic input may develop their own homesign systems, which has some properties of language, though not all (Goldin-Meadow, 2003; Goldin-Meadow & Mylander, 1984, 1990, 1998). If a child receives input but it is delayed (e.g., until past the age of five years), notable persistent differences between their use of sign language and that of native signers can be observed (Newport, 1990; Berk, 2003; Berk & Lillo-Martin, 2012).

Yet another situation is that in which children receive input from birth or an early age, but that input is not fully target-like, because the child has parents who

themselves learned sign language late (and the child has no or highly restricted access to sign language from others). In this context, the child develops his/her signing skills better than his/her parents, showing that the child is able to make better use of the mental language system in this situation (e.g. Singleton & Newport, 2004; Quadros & Cruz, 2011).

Our recent research includes bimodal bilingual children, that is, individuals who have received early exposure to languages in two different modalities: signed and spoken. This group includes both hearing children of deaf parents (Kodas) and deaf cochlear implanted children who are learning both signed and spoken language. We started building comparable corpora across two sign/spoken language pairs: Brazilian Sign Language and Brazilian Portuguese on the one hand, and American Sign Language and American English on the other. We are again collecting longitudinal data with babies from 1 to 4 years old, and we have added experimental data with children from 4 to 7 years old. We use different sets of researchers (deaf and hearing) to emphasize appropriate target language use, assuming the child's interlocutor sensitivity (Petitto et al., 2001). While the children do show such sensitivity, they also produce code-blending, that is, aspects of structure during which both signing and speaking occur productions (cf. van den Bogaerde & Baker, 2005, 2009; Emmorey et al., 2008). This is a real part of the language system being acquired, and is one of the foci of our investigations.

Recent research on childhood bilingualism has indicated that although children have two separate developing grammatical systems from very early on, there are instances of cross-linguistic influence, where grammatical structures from one language seem to exert a temporary influence on the child's grammar of the other language (e.g. Hulk & Müller, 2000). An important question is to identify the loci of such influences based on linguistic criteria. In order for us to address such issues, we are developing corpora from individual children acquiring both a sign language and a spoken language. Many of the same data collection issues arise as those for projects investigating only sign language (see Baker, van den Bogaerde & Woll, 2005 and Baker & Woll, 2009 for some best practices in this domain). However, in our current project, it turns out that there are specific concerns for which additional particular practices are needed; for instance, the frequent shift between code-blended and unimodal utterances. Language- or modality-specific properties as well as universals are found to be very interesting in these contexts (see Chen Pichler et al., 2010; Lillo-Martin et al., 2010, 2012; Quadros et al., in press).

In each one of these contexts, there are specific design concerns that must be considered in order to proceed with data collection, as well as with the annotation process, organization of the data and analyses. For example, in a bilingual bimodal context, in which a deaf child has hearing parents, a deaf experimenter interacts with the child in sessions alternating with a hearing interlocutor; the annotation process takes place with deaf transcribers and hearing transcribers, because of each language involved; and the organization of the data follows specific goals for each research sub-project to be considered in the analysis. Another example is that for deaf children from deaf

parents, native signer experimenters or the deaf parents by themselves interact with the child during the period of data collection. All the data is collected through videos that are stored in servers and have compressed versions for the work of annotation and analyses to take place.

The videos consist of samples collected longitudinally in places familiar to the children, usually, their homes or schools. In some cases, the children come to the university for filming. The environment is intended to be informal and based on ordinary activities that the children are used to. One of the parents or an experimenter interacts with the child during each session. The sessions are 30 to 60 minutes long each, and filming takes place two to four times a month for a period that varies across children from 1 to 5 years old. The videos also include experimental data that are conducted according to the requirements of each experiment. Children are invited to play in individual sessions with an experimenter who plays with them different language 'games' to target different aspects of language (see Quadros et al., in press for details).

For the bimodal bilingual project, we reorganized the form of the database previously used with our longitudinal data, and we built a new database for the experimental studies. The experimental studies include a set of 24 tests, evaluating different language aspects, such as, morphology, phonology, syntax, discourse and pragmatics. The goal of the tests is to provide a comprehensive profile of each bilingual child's developing competency in Libras (Brazilian Sign Language) and Brazilian Portuguese, or ASL (American Sign Language) and American English.

The data in sign and in speech adds considerable complexity to the already challenging prospect of corpus building. In this chapter, we present the organization of the sign language acquisition corpora developed on both sides of the project: Brazil and the United States of America.

There are few comparable projects for us to build on. In the area of adult sign language corpora, several signing communities have fairly recently embarked on corpus collection and annotation. We have learned a lot about sign language corpora construction and annotation from projects involving Australian Sign Language – Auslan <www.auslan.org.au/about/corpus/>, German Sign Language <www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/welcome.html>, Sign Language of the Netherlands <www.ru.nl/corpusngten/>, and British Sign Language <www.bslcorpusproject.org/>. In fact, our sign language acquisition data collection began before these projects were started (or before they were widely known), and we have learned how to improve on our early work, in part based on reports from these projects. Other projects have also collected sign language acquisition data, but no general guidelines or research reports focused on the data collection process were available when we started; Baker et al. (2005) provides a recent exception. The work reported in Casey (2003), van den Bogaerde & Baker (2005), Schick (2002), and Tang et al. (2007), among others, helped by showing ways that longitudinal spontaneous production data can be collected and utilized. Still, the focus in these reports has been on the acquisitional questions addressed by using corpus data. Here we focus on our own process of corpus construction.

2. Metadata

The metadata of the children involved in our study is organized through documents that are shared with researchers involved in the different steps of the investigation: data collection involving filming, transcribers, people that organize the data for specific purposes and people that analyze the data and write about the findings. These shared documents are posted in cloud-based resources (for example, Google docs and Dropbox), that is, they are password-protected online accessed documents. To maintain confidentiality, pseudonyms are used in all documents, for all participants. The main topics of the documents are the following:

Longitudinal study

a. Pseudonym of the child (for example, EDU)
b. Number of the session (from 000 up to the number of the sessions collected, for example, EDU_001, EDU_002, EDU_003)
c. Date of the filming
d. Age of the child (years; months.days)
e. Target language
f. Duration of the session
g. Adults involved in the session
h. Other participants involved in the session
i. Comments (for example, notes on topics of discussion or particularly fruitful sessions)
j. Transcribers – for speech and for sign
k. Checker/reviser of the transcription
l. Version of the manual used for annotation
m. Coding / analysis of the data for each purpose (for example, for WH analysis, for Modality analysis, etc.)

Experimental study

a. Name of the test
b. Pseudonym of the child
c. Condition (Coda, Deaf, CI, Coda adult)
d. Date
e. Age
f. Language
g. Duration
h. Comments
i. Transcriber
j. Reviser

The whole database is organized in a computer server. See Figure 1 for an illustrative sample of this organization. There are two main folders: the original archive (the original videos and backup materials) and the production archive (the compressed videos

Figure 1. Example of the organization of the database

and annotation files). The first one has the original videos that can be used eventually when we need to redo the compressed video or even when we may need a better quality video to access for reasons that we even do not know at the present time. The second one has the compressed videos for manipulation by the people that access the videos, as well as transcription and analysis files. The production folder includes the experimental data and longitudinal data in separate sections.

For the longitudinal data, the basic organization is to list the children in separate folders. Each child's folder includes the folders for each session containing the video and the transcript files (the basic one and the ones with the specific organization for specific purposes). The transcription is done using ELAN software made available by the Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands <http://tla.mpi.nl/tools/tla-tools/elan/>; Crasborn & Sloetjes, 2008). This program produces .eaf (Elan Annotation Format) files, with separate tiers of annotation capturing different types of information (see also below). The child's folder includes sub-folders for each session, starting with 001, since we usually have more than 100 videos for each child: IGOR_001, IGOR_002, IGOR_003, etc. In the IGOR_001 folder, we have the compressed video, the basic eaf file (the one that has the basic tiers for transcribers), the eaf files for analyses (for WH, for Modality, for NP, etc.), and Excel files with the organization of the data and charts.

For the experimental studies, the basic organization is to have the folders with the places and years in which the fairs happened. Within each place, the folders are separated by test. These folders are further divided into two sets of data by child: one

for those whose data is without restriction *("without restriction"*), and another for restricted data ("*with restriction*"). The restrictions apply to the kind of access people have to the videos. Some of the parents do not want students to have access to the videos of their child or for the researchers to use frames of the videos in conferences, for example. Within these two folders based on restriction, the children, then, are listed with the video and the eaf or the form of the test scanned with the results, depending on each test.

In the case of the experimental studies, the database is organized as well using FileMakerPro – FMP (Figure 2). Our FMP database has separate layouts for each test, and information for each participant. Participant responses are entered in the layouts where appropriate; for some tests, we use ELAN to transcribe the responses and then put links to the ELAN files in FMP bins. This database includes all four languages. It is accessed through a server by all the researchers involved in the project. Each group completes the data entry, and then reports are generated to allow us to compare the data from all four languages involved in each test. This approach then facilitates the comparison among the experimental results over the four languages.

On the experimental side, separate folders are made for each testing location (e.g., Porto Alegre), year, language (Libras), test (Carl), and level of restriction (com/sem restricao), then by participant pseudonym. The longitudinal studies are divided by participant pseudonym.

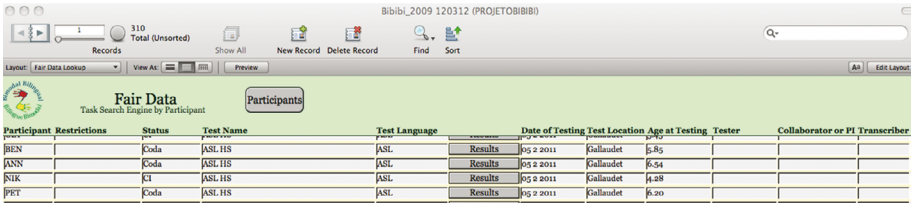

Figure 2. FileMakerPro

The figure illustrates only a small portion of our database, to indicate aspects of the information that is available. Clicking on the 'Results' tab leads to the results for that participant on that test.

3. Designing annotation patterns

Following video collection, we invest considerable energy in the production of transcripts, to be used in conjunction with the videos for linguistic analyses. Following our earlier sign-only research, we use ELAN for time-locked videos with transcription. The basic minimal level of annotation we use includes tiers for the utterance level and for free translation. Signs are glossed using Sign IDs – words that we have identified as the label for each sign. This is described in more detail below.

For the bilingual research, we designed a different template so that both languages are parent tiers, to optimize the study of (sequential or simultaneous) bimodal productions. As presented by Chen Pichler et al. (2010), our template has different tiers that are essentially the same for both countries and for all languages, except that Brazilian Portuguese is used for BP and Libras annotations, while English is used for spoken English and ASL annotations. The ELAN updated tier structure is the following (illustrated as well in Figure 3):

1. Child ASL/Libras utterance – utterance is a signed propositional unit or fragment of a propositional unit identified by prosodic hints annotated with glosses
2. Child ASL/Libras individual – tier in which individual signs are each in separate annotations; this tier is produced by tokenization of the utterance tier
3. Child ASL/Libras Pho – phonological transcriptions of signs (as needed)
4. Child ASL/Libras right hand – when two signs are produced simultaneously with both hands they are recorded in the utterance tier and also specified (redundantly) in the right and left hand tiers, with transcription of the right hand sign in this tier
5. Child ASL/Libras left hand – (see right hand for explanation) transcription of the left hand sign in this tier
6. Child ASL/Libras syntactic unit – a unit that is driven by syntactic information during analysis phases (e.g., breaking up an utterance with phrasal repetition so that each repetition can be separately analyzed)
7. Child English/Portuguese utterance – utterance is a spoken propositional unit or fragment of a propositional unit identified by prosodic and syntactic hints
8. Child English/Portuguese individual – tier in which individual spoken words are each in separate annotations; this tier is produced by tokenization of the utterance tier
9. Child English/Portuguese pho – phonological transcriptions of words (as needed)
10. Child free translation – this is a free translation of the intended proposition in English/ Portuguese (it can combine parts from ASL/ Libras and English/ Portuguese as well as contextual information to form a complete proposition)
11. Child comments – specific comments regarding the child's utterance
12. Comments – general comments (e.g., people in the background)

The same tier structure applies to the adults that interact with the child. The tiers for each adult that interact with the child are added as Adult1, Adult2, Adult3, etc.

Besides the basic tiers, we add specific tiers for each analysis applied to the data. This is always driven by the specific goals of the research. We have developed up to now, specific tiers for modality, for transfer analysis, for NP structure, for cyclicity, and for IX. For instance, we added specific tiers to the basic eaf file regarding modality as presented in Figure 4.

Transcription conventions are agreed on between the two countries with specific adaptations imposed by each language. These conventions are continuously being improved and once a year, we update them, considering findings that we have. This is why we added a column to the metadata information, as a way to find out about the version of the conventions applied to a specific annotation and analysis. In Chen

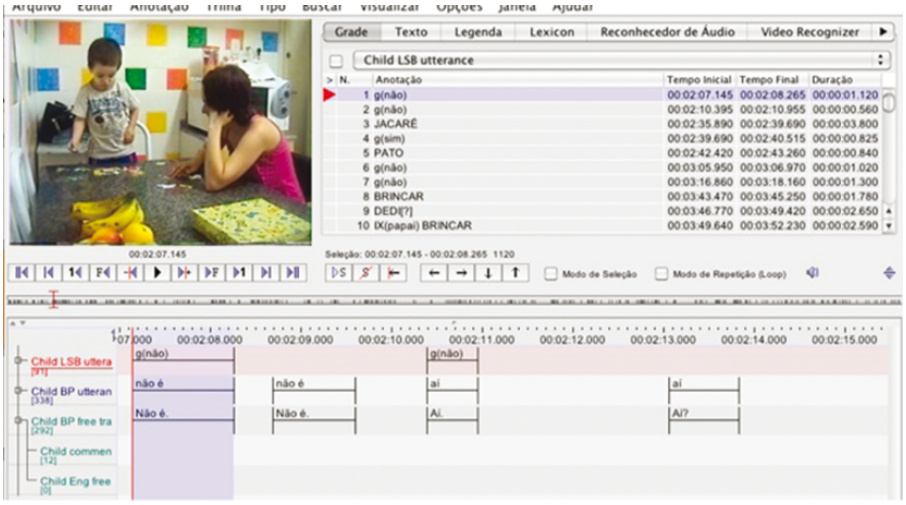

Figure 3. ELAN screenshot in the context of the Bibibi Project with the basic tiers for the child illustrated

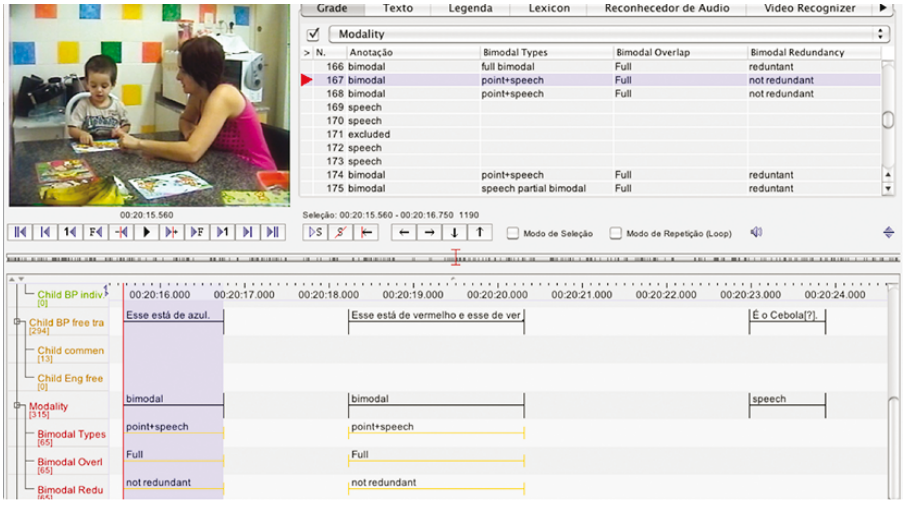

Figure 4. ELAN screenshot in the context of the Bibibi Project with specific tiers for modality analysis as well as basic tiers illustrated

Pichler et al. (2010) we presented the 2010 version of our conventions; it has already had two updates since (see further discussion below).

The general principles that guide the annotation of the data are to create a machine-readable record of language samples, not necessarily sufficient for the reader to reproduce the utterances in exactly the same way, but so that the records can be

searched to find all occurrences of phenomena of interest (in the way described by Johnston, 2001, 2010; Pizzuto & Pietrandrea, 2001). In addition to having a basic annotation of the utterance in each language, we use multiple annotation parses focusing on different phenomena. This documentation of the data is the foundation for our analysis decisions.

Where it is possible, we follow the CHILDES conventions established for child language data (MacWhinney, 2000; <http://childes.psy.cmu.edu/manuals/chat.pdf>) in transcribing both speech and sign (though we do not use the BTS system designed for writing sign language data morphologically). When the CHILDES conventions conflict with our sign-specific goals, we create new conventions to be followed for transcribing both sign and speech. It is important to keep the sign and speech transcriptions comparable.

As do many sign researchers, we use upper-case glosses from the spoken language to annotate signs. In order to ensure consistency in gloss-sign mappings, we have developed particular glosses, called Sign IDs (or ID glosses), for each sign language. The development and use of Sign IDs is discussed in detail in the following section. An example of the use of glosses is given below.

Sign:	CAR BIG
Free Translation:	*The big car.*

In addition to the Sign IDs, we use specific conventions to annotate sign and speech. These conventions help ensure consistency in our transcripts – an important asset both for readability and for machine-searching for analysis purposes. Our conventions are described in detail in Chen Pichler et al. (2010). The conventions are also available on our website (bibibi.uconn.edu). Some of the most commonly used include the following:

1. Interruptions: [/] retracing without correction; [//] retracing with correction; [///] retracing with reformulation
2. Pauses are marked with #
3. Unclear words: [?] indicates the transcriber's best guess, when the word is not completely clear but plausibly as written; yyy (speech)/ YYY (sign) is used when the word is not recognized, but it is possible to provide some phonetic information (written on the pho tier); xxx (speech)/ XXX (sign) is used when something is produced which is thought to be linguistic but not pronounced clearly enough to be recognizable.
4. Interjections are transcribed with i(xxx). We have developed a list of common interjections in sign and speech.
5. Actions are transcribed using &= (this applies to sound imitation, onomatopoeia, general actions such as movements imitating something, or communicative actions such as reaching).

Some of the specific conventions for signs are the following:

1. Pointing to people, objects, or locations is indicated using IX(referent).
2. Possessive or reflexive points are similarly indicated using POSS(referent) or SELF(referent).
3. Indicating verbs (traditionally known as agreeing verbs) are transcribed simply with the sign ID for the verb without including information about the referents. Such information can be added in a further pass according to a particular analysis.
4. Signs of the type known as classifiers or depicting verbs are annotated using DV(description).
5. Fingerspelling is identified by FS(name).
6. Name signs are transcribed using NS(name).
7. Timing: [_] is used for static signs held longer than usual or [+] for reduplication (or repetitive sequences of movement).
8. Emblems are signs which are also conventionalized gestures shared with the hearing community; they are transcribed with E(xxxx). Emblems are counted as signs in the computation of the sign units.
9. Mouthing (when the signer mouthes a word) is indicated with m(word-mouthed).
10. Non-manual signals are not transcribed; they may be added in subsequent passes for specific research.

The last update in our conventions changed the way that we use to transcribe gestures, as well as including more detailed conventions for speech. Chen Pichler et al. (2010) presented a list of gestures transcribed as g(description of the gesture). When we were analyzing the gestures, we found that many of them were interjections. Then, we decided to create a category called interjections to mark these differently (e.g., i(ow)). We also realized that many of what we would still be calling gestures would be better classified as actions (e.g., &=reaches). Finally, we have a category of emblems – signs that are also used as conventional gestures in the hearing community (e.g., E(come-here)). With these adjustments, we decided to exclude the gesture category in our annotations. What we categorize as emblems, interjections and actions would all be possible realizations of what general gesture researchers would consider as different categories of gestures.

4. Sign IDs

Because sign languages do not have established conventional orthographies, sign researchers have typically relied on glossing: choosing a printed word from the local language whose meaning overlaps with the meaning of the sign, and using this word as a label for the sign. In some places or laboratories, phonetically-based systems for writing signs may be used, such as SignWriting, which is relatively popular in Brazil. Nevertheless, the common practice in sign linguistics is to use glosses. However, annotators might use different glosses for what is actually the same sign, possibly because the meaning of the sign might differ according to the context. For example, the ASL sign MOTHER might be transcribed as MOM or MAMA. Similarly, a single English

word might be used to annotate two different ASL signs, as when LIGHT is used for both illumination and light-weight.

Inconsistencies such as these are problematic for corpora, particularly when searching across files for all occurrences of a particular sign. Thus, it is important to ensure that signs are written using the same consistent gloss in all contexts (Johnston, 2010). By using the same gloss for a sign, the researcher can search the corpus for all occurrences of this sign. If the sign has more than one type of use, the researcher will be able to determine this through examining all occurrences.

In order to follow this practice, we are creating a specific identification for each sign to be used in our transcripts (in the same spirit of Johnston, 2010, for Australian Sign Language). These choices are called "Sign IDs" in the Brazilian group, and ID glosses in the US group.

In order to facilitate and expand the analysis of data collected in our project, we developed a sign ID lexicon containing the vocabulary items used most frequently by the children we are studying. Sign IDs are word labels chosen to represent each sign root systematically, so that every use of the sign has the same label, despite contextual or morphological differences which affect how the sign is interpreted. By using sign IDs in our transcripts, we are able to conduct our analyses more efficiently, using a wider range of data. The sign ID lexicon addresses the problem of transcript searchability and greatly facilitates the analysis of data collected for sign language corpora. This helps to standardize annotations and it can be more freely accessed by other researchers.

On the Brazilian side, we have been developing a sign IDs database by feeding it with the signs over which transcribers had doubts regarding transcription. We have periodic meetings to discuss these signs, then we christen each and add it to the ID list <www.idsinais.libras.ufsc.br> (see Figure 5 for the Sign ID screen). The search system has filters based on sign language parameters (132 handshapes divided in 13 groups and 8 locations). The signs included are the ones that are considered stable in the lexicon, that is, conventionalized lexical items including what we call emblems. Depicting

Figure 5. ID screen for Libras – the entry page to search for a particular sign

signs (known as classifiers for some researchers) are not included in the Sign ID, since they are highly productive and cannot be individually listed.

The signs can have geographic variation. When this happens, we have more than one entry, with each of the different realizations of the same sign indicating the geographic area that the sign is currently used. The sign ID for these variations have different forms, indicating the existence of multiple signs. For example, there is an entry for MOTHER that is the most common one, but there is also another sign for mother used in other regions of the country: MOTHER-RS.

An example with a group of handshapes chosen as a parameter to search for a specific sign is given in Figure 6 and the results of this search are shown in Figure 7.

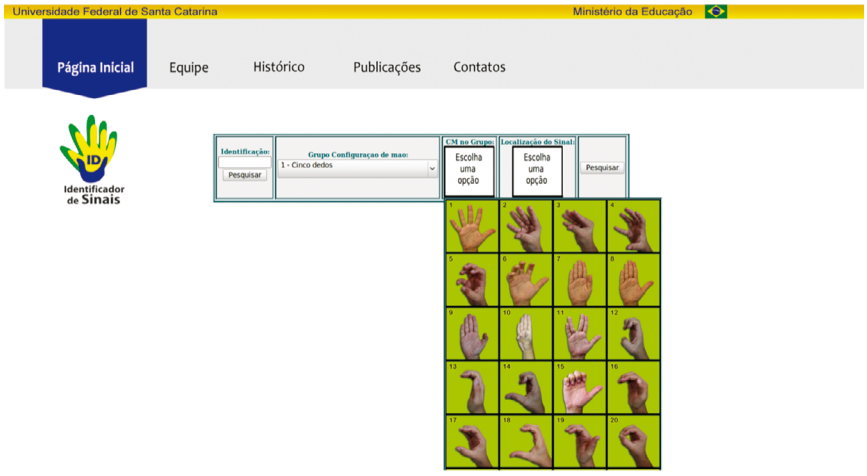

Figure 6. ID searching system: Handshape selection

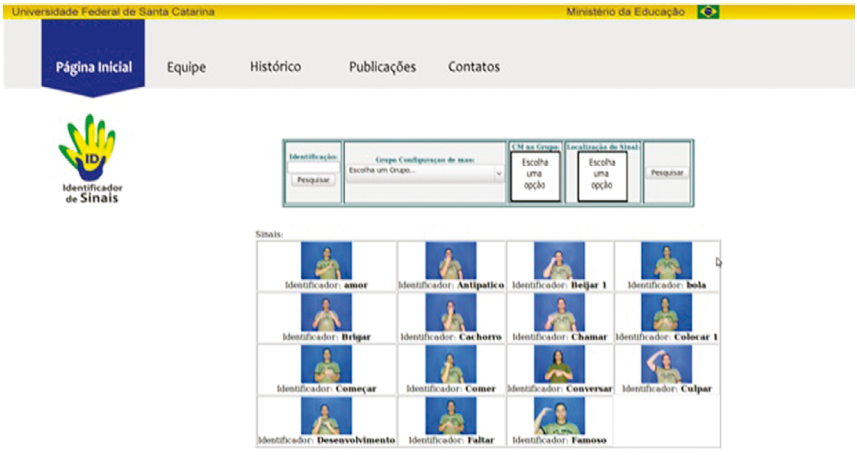

Figure 7. Sign IDs for the signs resulting from a search using a particular set of search criteria

The sign ID specifications include identification of the sign, Portuguese translation, English translation, written sign, handshape groups, handshapes, location and sign video (as illustrated in Figure 8). The searching may be done through handshapes, locations, handshape groups, location groups, the sign ID or the first letter of the sign ID. When a sign is located, the user may evaluate the appropriateness of the ID that has been assigned. In this way, the research team can rank each ID sign and replace the ones that have consistently low scores.

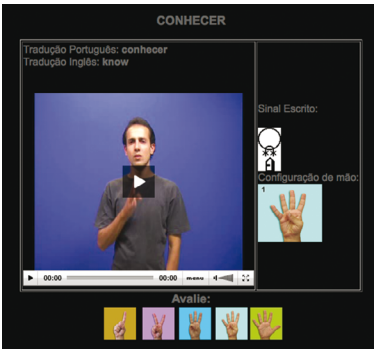

Figure 8. ID sign screen, showing the ID, Portuguese translation, English translation, video of the sign, SignWriting, and handshape. At the bottom, alternative handshapes are shown

On the American side, the development of an ID gloss database has taken into consideration the needs of different research groups across the country, each of which uses a different system for writing signs. Different groups of researchers use different glosses, and we found it desirable to create a database with a structure which allows it to be used by multiple groups, representing the various glosses in such a way that cross-group comparisons can be made. The database was set up so that different local groups can enter their own information about each sign, and each group can also view the information entered by the others. This approach facilitates the comparison of transcriptions used across different groups, and may eventually lead to greater convergence in the glossing systems used.

At this point in development, the database has been programmed and information has been entered on 1000 signs from three research groups (see Hochgesang et al., 2010; and Fanghella et al., 2012 for details). Each research group enters the gloss used for the sign, alternative glosses, and phonological information following the phonological system of choice (one group enters information using the Berkeley Transcription System, Hoiting & Slobin, 2002; others use Stokoe et al., 1965 and/or Johnson & Liddell, 2011). The database contains fields for various other types of information about each sign, including lexical category and sociolinguistic information (such as regional variation).

In order to assess the need for additional glosses, we compared the signs in the database to the signs used by two of the children in our project, one Deaf and one

Coda (Fanghella et al., 2012). We found that approximately 2/3 of the children's lexical types were included in the database. This indicates that our first selection of signs for inclusion in the database was good, but that the database needs to be expanded for optimal usage. This expansion is on-going.

The American group has begun to integrate the ID glosses and the transcription process. The database was originally developed as an independent project which would eventually tie in with the transcription. We are using the 'external controlled vocabulary' (ECV) function of ELAN to list the glosses in the database. Transcribers are able to consult the ECV to see whether their chosen gloss is included. Glosses that are not included may need to be changed in the transcript, or added to the database. Streamlining of this process is currently in progress.

5. Conclusion

One of our major goals has been cross-site comparability, that is, establishing the same criteria, approach to data collection, ELAN template, and general transcription principles to be used across our three universities. The metadata and data are shared through the use of a common server, as well as online services including Google docs and Dropbox. The analyses of the results are being conducted through regular meetings and we are on the right track to answer our research questions (e.g., Lillo-Martin et al., 2010; Chen Pichler et al., 2010; Quadros et al., 2013).

We have not yet resolved the following linguistic issues, but we hope that our project will contribute to their discussion in the field as a whole. Does bimodal bilingualism lead to cross-language influence different from that found in mono-modal bilingualism (e.g., due to code-blending, or use of non-manuals)? What is the best theoretical mechanism to account for this apparent cross-language influence? When bimodal bilinguals code-blend, are they choosing grammatical structures which are permitted in both languages for maximum accommodation? What kinds of syntactic representations can account for code-blends? These are the types of research questions our project can address through the use of the corpora we are now building.

Our template and corpus-building decisions can be applicable to the development of adult only bimodal bilingual corpora. In addition, many similar issues arise in the study of co-speech gesture, and researchers in this area may take advantage of aspects of our procedures. And, we hope that our collaboration across continents may contribute to and promote cross-linguistic research on sign languages as well.

Acknowledgements

We sincerely thank the Deaf consultants, research assistants, children, and their families who work with us in our research.

References

Baker, Anne, van den Bogaerde, Beppie & Woll, Bencie. 2005. Methods and procedures in sign language acquisition studies. *Sign Language & Linguistics* 8(1–2): 7–58. DOI: 10.1075/sll.8.1-2.03bak

Baker, Anne & Woll, Bencie (eds). 2009. *Sign Language Acquisition* [Benjamins Current Topics 4]. Amsterdam: John Benjamins. DOI: 10.1075/bct.14

Berk, Stephanie. 2003. Sensitive Period Effects on the Acquisition of Language: A Study of Language Development. PhD dissertation, University of Connecticut, Storrs.

Berk, Stephanie & Lillo-Martin, Diane. 2012. The two-word stage: Motivated by linguistic or cognitive constraints? *Cognitive Psychology* 65: 118–140. DOI: 10.1016/j.cogpsych.2012.02.002

van den Bogaerde, Beppie & Baker, Anne. 2005. Code-mixing in mother-child interaction in deaf families. *Sign Language & Linguistics* 8(1–2): 151–174. DOI: 10.1075/sll.8.1.08bog

van den Bogaerde, Beppie & Baker, Anne. 2009. Bimodal language acquisition in Kodas (kids of deaf adults). In *Hearing, Mother-father Deaf: Hearing People in Deaf Families*, Michele Bishop & Sherry L. Hicks (eds), 99–131. Washington DC: Gallaudet University Press.

Casey, Shannon. 2003. 'Agreement' in Gestures and Signed Languages: The Use of Directionality to Indicate Referents Involved in Actions. PhD dissertation, University of California, San Diego.

Chen Pichler, Deborah. 2012. Acquisition. In *Sign Language. An International Handbook*, Roland Pfau, Markus Steinbach & Bencie Woll (eds), 647–686. Berlin: Walter de Gruyter.

Chen Pichler, Deborah, Hochgesang, Julie, Lillo-Martin, Diane & de Quadros, Ronice Müller. 2010. Conventions for sign and speech transcription of child bimodal bilingual corpora in ELAN. *Language, Interaction and Acquisition* 1: 11–40. DOI: 10.1075/lia.1.1.03che

Chen Pichler, Deborah, de Quadros, Ronice Müller & Lillo-Martin, Diane. 2010. Effects of bimodal production on multi-cyclicity in early ASL and LSB. In *A Supplement to the Proceedings of the 34th Boston University Conference on Language Development*, Jane Chandlee, Katie Franich, Kate Iserman & Lauren Keil (eds). <http://www.bu.edu/linguistics/BUCLD/supp34.html>

Chen Pichler, Deborah, Kuntze, Marlon, Lillo-Martin, Diane, de Quadros, Ronice Müller & Stumpf, Marianne Rossi. In press. *Sign Language Acquisition by Deaf and Hearing Children: A Bilingual Introductory Digital Course*. Washington DC: Gallaudet University Press.

Chen Pichler, Deborah, Hochgesang, Julie, Lillo-Martin, Diane, de Quadros, Ronice Müller & Reynolds, Wanette. Submitted. Best practices for building a bi-modal bi-lingual bi-national child corpus.

Crasborn, Onno & Sloetjes, Han. 2008. Enhanced ELAN functionality for sign language corpora. In *Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, 39–43.

Davidson, Kathryn, Lillo-Martin, Diane & Chen Pichler, Deborah. (2014). Spoken English language measures of native signing children with cochlear implants. *Journal of Deaf Studies and Deaf Education* 19(1). DOI: 10.1093/deafed/ent045

Emmorey, Karen, Borinstein, Helsa B., Thompson, Robin & Golan, Tamar H. 2008. Bimodal bilingualism. *Bilingualism: Language and Cognition*. 11(1): 43–61. DOI: 10.1017/S1366728907003203

Fanghella, Julia, Geer, Leah, Henner, Jonathan, Hochgesang, Julie, Lillo-Martin, Diane, Mathur, Gaurav, Mirus, Gene & Pascual-Villanueva, Pedro. 2012. Linking an ID-Gloss database of ASL with child language corpora. In *Proceedings of LREC Workshop on the Representation and Processing of Sign Languages (Interactions between Corpus and Lexicon)*. Istanbul.

Goldin-Meadow, Susan. 2003. *The Resilience of Language: What Gesture Creation in Deaf Children Can Tell us about How All Children Learn Language*. New York NY: Psychology Press.

Goldin-Meadow, Susan & Mylander, Carolyn. 1984. Gestural communication in deaf children: The effects and noneffects of parental input on early language development. *Monographs of the Society for Research in Child Development* 49(3–4, Serial No. 207). DOI: 10.2307/1165838

Goldin-Meadow, Susan & Mylander, Carolyn. 1990. Beyond the input given: The childs role in the acquisition of language. *Language* 66: 323–355. DOI: 10.2307/414890

Goldin-Meadow, Susan & Mylander, Carolyn. 1998. Spontaneous sign systems created by deaf children in two cultures. *Nature* 391: 279–281. DOI: 10.1038/34646

Hochgesang, Julie A., Pascual Villanueva, Pedro, Mathur, Gaurav & Lillo-Martin, Diane. 2010. Building a database while considering research ethics in sign language communities. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta*, Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, & Daniel Tapias (eds). Paris: ELRA.

Hoiting, Nini & Slobin, Dan I. 2002. Transcription as a tool for understanding: The Berkeley Transcription System for sign language research (BTS). In *Directions in Sign Language Acquisition* [Trends in Language Acquisition Research 2], Gary Morgan & Bencie Woll (eds), 55–75. Amsterdam: John Benjamins.

Hulk, Aafke & Müller, Natasha. 2000. Bilingual first language acquisition at the interface between syntax and pragmatics. *Bilingualism: Language and Cognition* 3(3): 227–244. DOI: 10.1017/S1366728900000353

Johnson, Robert E. & Liddell, Scott K. 2011. A segmental framework for representing signs phonetically. *Sign Language Studies* 11(3): 408–463. DOI: 10.1353/sls.2011.0002

Johnston, Trevor. 2001. The lexical database of Auslan (Australian Sign Language). *Sign Language and Linguistics* 4(1–2): 145–169. DOI: 10.1075/sll.4.1-2.11joh

Johnston, Trevor. 2010. From archive to corpus: Transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics* 15: 104–129. DOI: 10.1075/ijcl.15.1.05joh

Lillo-Martin, Diane. 1999. Modality effects and modularity in language acquisition: The acquisition of American Sign Language. In *Handbook of Language Acquisition*, Tej Bhatia & William C. Ritchie (eds), 531–567. San Diego CA: Academic Press.

Lillo-Martin, Diane. 2009. Sign language acquisition studies. In *The Cambridge Handbook of Child Language*, Edith Bavin (ed.), 399–415. Cambridge: CUP. DOI: 10.1017/CBO9780511576164.022

Lillo-Martin, Diane & Berk, Stephanie. 2003. Acquisition of constituent order under delayed linguistic exposure. In *Proceedings of the 27th Annual Boston University Conference on Language Development*, Barbara Beachley, Amanda Brown & Frances Conlin (eds), 484–495. Somerville MA: Cascadilla Press.

Lillo-Martin, Diane & Chen Picher, Deborah. 2008. Development of sign language acquisition corpora. In *Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora; 6th Language Resources and Evaluation Conference,* Onno Crasborn, Eleni Efthimiou, Thomas Hanke, Ernst D. Thoutenhoofd & Inge Zwitserlood (eds),129–133.

Lillo-Martin, Diane, de Quadros, Ronice Müller, Koulidobrova, Helen & Chen Pichler, Deborah. 2010. Bimodal bilingual cross-language influence in unexpected domains. In *Language Acquisition and Development: Proceedings of GALA 2009*, João Costa, Ana Castro, Maria Lobo & Fernanda Pratas (eds), 264–275. Newcastle upon Tyne: Cambridge Scholars.

Lillo-Martin, Diane, Koulidobrova, Helen, Quadros, de Ronice Müller & Chen Pichler, Deborah. 2012. Bilingual language synthesis: Evidence from wh-questions in bimodal bilinguals. In *Proceedings of the 36th Annual Boston University Conference on Language Development*, Alia K. Biller, Esther Y. Chung & Amelia E. Kimball (eds), 302–314. Somerville MA: Cascadilla Press.

MacWhinney, Brian. 2000. *The CHILDES Project: Tools for Analyzing Talk*, 3rd edn. Mahwah NJ: Lawrence Erlbaum Associates.

Marentette, Paula & Mayberry, Rachel. 2000. Principles for an emerging phonological system: A case study of acquisition of American Sign Language. In *Language Acquisition by Eye*, Charlene D. Chamberlain, Jill P. Morford & Rachel Mayberry (eds), 51–69. Mahwah NJ: Lawrence Erlbaum Associates.

Meier, Richard. 2006. The form of early signs: Explaining signing children's articulatory development. In *Advances in Sign Language Development by Deaf Children*, Brenda Schick, Marc Marschark & Patricia E. Spencer (eds), 202–230. Oxford: OUP.

Meier, Richard P. & Newport, Elissa L. 1990. Out of the hands of babes: On a possible sign advantage in language acquisition. *Language* 66: 1–23.

Newport, Elissa L. 1990. Maturational constraints on language learning. *Cognitive Science* 14: 11–28. DOI: 10.1207/s15516709cog1401_2

Newport, Elissa L. & Meier, Richard P. 1985. The acquisition of American Sign Language. In *The Cross-Linguistic Study of Language Acquisition*, Vol. 1, Dan I. Slobin (ed.), 881–938. Hillsdale, NJ: Lawrence Erlbaum Associates.

Petitto, Laura Ann. 2000. On the biological foundations of human language. In *The Signs of Language Revisited: An Anthology in Honor of Ursula Bellugi and Edward Klima*, Karen Emmorey & Harlan Lane (eds), 447–471. Mahwah NJ: Lawrence Erlbaum Associates.

Petitto, Laura Ann, Katerelos, Marina, Levi, Bronna G., Gauna, Kristine, Tetrault, Karine & Ferraro, Vittoria. 2001. Bilingual signed and spoken language acquisition from birth: Implications for the mechanisms underlying early bilingual language acquisition. *Journal of Child Language* 28(2): 453–496.

Pizzuto, Elena & Pietrandrea, Paola. 2001. The notation of signed texts: Open questions and indications for further research. *Sign Language and Linguistics* 4(1–2): 29–45. DOI: 10.1075/ sll.4.12.05piz

de Quadros, Ronice Müller. 2010. Sign language acquisition. In *Les llengües de signes com a llengües minoritàries: Perspectives lingüístiques, socials i polítiques*, Vol. 1, Joan Martí Castell & Josep M. Mestres Serra (eds), 121–142. Barcelona: Limpergraf.

de Quadros, Ronice Müller. 1997. *Educação de Surdos: A Aquisição da Linguagem*. Porto Alegre: ArtMed.

de Quadros, Ronice Müller, Lillo-Martin, Diane & Chen Pichler, Deborah. 2010. Desenvolvimento bilíngue intermodal. In *Anais do IX Congresso Internacional de Educação de Surdos*, 146–150. Rio de Janeiro: Instituto Nacional de Educação de Surdos.

de Quadros, Ronice Müller & Cruz, Carina Rabello. 2011. *Língua de Sinais: Instrumentos de Avaliação*. Porto Alegre: ArtMed.

de Quadros, Ronice Müller, Lillo-Martin, Diane & Chen Pichler, Deborah. 2013a. O que bilíngues bimodais tem a nos dizer sobre o desenvolvimento bilíngue? *Letras de Hoje* 48(3): 380–388.

de Quadros, Ronice Müller, Lillo-Martin, Diane & Chen Pichler, Deborah. 2013b. Early effects of bilingualism on WH-question structures: Insight from sign-speech bilingualism. In *Proceedings of GALA 2011*, Stavroula Stavrakaki, Marina Lalioti & Polyxeni Konstantinopoulou (eds), 300–308. Newcastle upon Tyne: Cambridge Scholars.

de Quadros, Ronice Müller, Chen Pichler, Deborah, Lillo-Martin, Diane, Cruz, Carina Rebello, Kozak, Laura, Palmer, Jeffrey Levi, Lemos Pizzio, Aline & Reynolds, Wanette. In press. Methods in bimodal bilingualism research: Experimental studies. In *The Blackwell Guide to Research Methods in Sign Language Studies*, Elini Orfanidou, Bencie Woll & Gary Morgan (eds). Oxford: Blackwell.

de Quadros, Ronice Müller & Pizzio, Aline Lemos. 2007. Aquisição da língua de sinais brasileira: Constituição e transcrição dos corpora. In *Bilingüísmo dos surdos*, Vol. 1, Heloisa Maria Moreira Lima-Salles (ed.), 49–72. Giânia: Cânone Editorial.

de Quadros, Ronice Müller & Lillo-Martin, D. 2009. Sign language acquistion of verbal morphology in Brazilian and American Sign Language. In *Psycholinguistics: Scientific and Technological Challenges*, Vol. 1, Leonor Scliar-Cabral (ed.), 252–262. Porto Alegre: Edipucrs.

Schick, Brenda. 2002. The expression of grammatical relations by deaf toddlers learning ASL. In *Directions in Sign Language Acquisition* [Trends in Language Acquisition Research 2], Gary Morgan & Bencie Woll (eds), 143–158. Amsterdam: John Benjamins.

Singleton, Jenny L. & Newport, Elissa L. 2004. When learners surpass their models: The acquisition of American Sign Language from inconsistent input. *Cognitive Psychology* 49: 370–407. DOI: 10.1016/j.cogpsych.2004.05.001

Stokoe, William C., Casterline, Dorothy C. & Croneberg, Carl G. 1965. *A Dictionary of American Sign Language on Linguistic Principles*. Silver Spring MD: Linstok Press.

Tang, Gladys, Sze, Felix & Lam, Scholastica. 2007. Acquisition of simultaneous constructions by deaf children of Hong Kong Sign Language. In *Simultaneity in Signed Languages: Form and Function* [Current Issues in Linguistic Theory 281], Myriam Vermeerbergen, Lorraine Leeson & Onno Crasborn (eds), 283–316. Amsterdam: John Benjamins.