TalkBank and CLARIN

Brian MacWhinney
Department of Psychology
Carnegie Mellon University, Pittsburgh USA
macw@cmu.edu

Abstract

TalkBank promotes the use of corpora, web-based access, multimedia linkage, and human language technology (HLT) for the study of spoken language interactions in a variety of discourse types across many languages, involving children, second language learners, bilinguals, people with language disorders, and classroom learners. Integration of these materials within CLARIN provides open access to access a large amount of research data, as well as a test bed for the development of new computational methods.

1 Introduction

The TalkBank system (http://talkbank.org) is the world's largest open access repository for spoken language data. It provides language corpora and resources for a variety of research topics in Psychology, Linguistics, Education, Computer Science, and Speech Pathology. There are currently seven funded TalkBank components. The National Institutes of Health (NIH) funds the development of the CHILDES database (http:childes.talkbank.org) for the study of child language development (MacWhinney, 2000), PhonBank (phonbank.talkbank.org) for the study of phonological development (Rose & MacWhinney, 2014), AphasiaBank (aphasia.talkbank.org) for the study of language in aphasia (MacWhinney & Fromm, 2015), and FluencyBank (fluency.talkbank.org) for the study of the development of fluency and disfluency in children and language learners (Bernstein Ratner & MacWhinney, 2016). The National Science Foundation (NSF) provides additional funding for FluencyBank, as well as funding for HomeBank (http://homebank.talkbank.org) with daylong audio recordings in the home (VanDam et al., 2016). The National Endowment for the Humanities (NEH) and the Deutsche Forschungs Gesellschaft (DFG) have provided funding for web-based access to materials from Classical Latin and Historical German (http://sla.talkbank.org).

In addition to these seven funded projects, TalkBank has developed resources for TBIBank (traumatic brain injury), RHDBank (right hemisphere damage), ASDBank (autism), DementiaBank (dementia), CABank (Conversation Analysis) (MacWhinney & Wagner, 2010), SamtaleBank (Danish), GestureBank (gesture), SLABank (Second Language Acquisition) (MacWhinney, 2015b), BilingBank (bilingualism), ClassBank (classroom interactions), and TutorBank (human tutors). All these resources use a common transcript format called CHAT which is used by the CLAN analysis programs and other open-access resources. Except for some of the corpora from clinical areas and the daylong recordings from the home, these resources are available without passwords.

TalkBank includes 348 corpora contributed by researchers across all these fields. After corpora have been contributed, they undergo additional reformatting, curation, indexing, annotation, and linkage to media. The result is a unified open-access database with a fully consistent system of transcription and annotation across all corpora. We believe that this type of data integration with open access is important for maximizing the value of the corpora contributed to TalkBank, and that this method can serve as a model for other CLARIN data sites.

In 2014, the TalkBank center at Carnegie Mellon University in Pittsburgh became a CLARIN-B site, and in 2016 it became a CLARIN-K site. TalkBank is the first CLARIN site outside of Europe.

This paper will summarize the principles underlying the design of TalkBank, the ways in which TalkBank has implemented CLARIN standards, and how it can provide resources for the CLARIN community.

2 The Motivation for TalkBank

Most language resources derive from written sources, such as books, newspapers, and the web. It is relatively easy to enter such written data directly into computer files for further linguistic (Baroni & Kilgarriff, 2006) and behavioral (Pennebaker, 2012) analysis. On the other hand, preparation of spoken language data for computational analysis is much more difficult. Despite ongoing advances in speech technology (Hinton et al., 2012), collection of spoken language corpora still depends on a time-consuming process of hand transcription. Because of this, the total quantity of spoken language data available for analysis is much less than that available for written language, although face-to-face conversation is the original and primary root of human language. Furthermore, unplanned spoken language (Givon, 2005; Redeker, 1984) includes many prosodic features, gestural components, reductions, and hesitation phenomena that further complicate transcription and analysis.

Because of its conceptual centrality, there are several major disciplines that examine aspects of face-to-face communication. These include Psycholinguistics, Development Psychology, Applied Linguistics, Phonology, Theoretical Linguistics, Conversation Analysis, Gestural Studies, Human-Computer Interaction, Social Psychology, Speech and Hearing, Neuroscience, Evolutionary Biology, and Political Science. To understand language acquisition, second language learning, language attrition, language change, language disorders, sociolinguistic variation, persuasion, and group communication, we will need to combine methods and insights from each of these disciplines. Through such comparisons, and by examining language usage across a range of timescales (MacWhinney, 2015a), we can address core issues such as: how language is learned, how it is processed, how it changes, and how it can be restored after damage.

Like written language (Biber, 1991), the forms of spoken language vary enormously from situation to situation (Hymes, 1962). However, individual speakers can operate smoothly within each of these varied contexts. This means that, to fully understand human language and its role in human culture, we need to compare language use across many situations, forms, and participants. Because of this diversity, contrasts between practices in individual laboratories and disciplines have made the forms of transcription and coding for spoken language corpora remarkably unstandardized, making it difficult to construct comparisons across corpora. The first goal of TalkBank is to provide a system that bridges across these differences by providing an inclusive standard that recognizes all the features required for these specific disciplinary analyses. To achieve this goal, TalkBank has elaborated the CHAT transcription standard.

The development and extension of the CHAT transcription standard represents a necessary precondition to the central goal of TalkBank, which is to encourage and support data-sharing across all the language sciences. In the physical sciences, the process of data-sharing is taken as a given. However, until recently, data-sharing has not been adopted as the norm in the social sciences. This failure to share research results – much of it supported by public funds – represents a huge loss to science. Researchers often cite privacy concerns as reasons for not sharing data on spoken interactions. However, as illustrated at http://talkbank.org/share/irb/options.html, TalkBank provides many ways in which data can be made available to other researchers, while still preserving participant anonymity.

3 Many Banks in One

TalkBank is composed of 17 component banks, each using the same CHAT transcription format and database organization standards. This section describes the contents and each of these component language banks. The homepage at http://talkbank.org provides links to each of these 17 banks, as well as related resources.

3.1 CHILDES

The CHILDES (Child Language Data Exchange System) database at http://childes.talkbank.org is the oldest of TalkBank's component banks. Brian MacWhinney (CMU) and Catherine Snow (Harvard School of Education) began the CHILDES system in 1984 with funding from the MacArthur Founda-

tion. Snow organized a meeting in the (appropriately named) town of Concord, Massachusetts at which many of the major figures in the child language field agreed on the basic principles for sharing child language data. In the early 1980s, researchers were just beginning to use personal computers and transcribed data was still stored in 9-track tapes, punch cards, and floppy disks. The Internet was not generally available for data transmission, so data was shared by mailing CD-ROM copies to members. At that time, there was no thought that the transcripts might eventually be linked to audio or video. As a result, researchers often destroyed or recycled their audio recordings. Since that early beginning, CHILDES has grown in coverage, membership, and output. Since 1987, the project has been funded by NIH with some additional support from NSF. The table at the end of this section shows that there now are over 7000 published articles based on the use of data or programs from CHILDES. This work extends across the areas of phonology, morphology, syntax, lexicon, narrative, literacy, and discourse.

Using CHILDES data and methods, researchers have evaluated alternative theoretical approaches to comparable data. For example, the debate between connectionist models of learning and dual-route models focused first on data regarding the learning of the English past tense (MacWhinney & Leinbach, 1991; Marcus et al., 1992; Pinker & Prince, 1988) and later on data from German plural formation (Clahsen & Rothweiler, 1992). In syntax, emergentists (Pine & Lieven, 1997) have used CHILDES data to elaborate an item-based theory of learning of the determiner category, whereas generativists (Valian, Solt, & Stewart, 2009) have used the same data to argue for innate categories. Similarly, CHILDES data in support of the Optional Infinitive Hypothesis (Wexler, 1998) have been analyzed in contrasting ways using the MOSAIC system (Freudenthal, Pine, & Gobet, 2010) to demonstrate constraint-based inductive learning. In these debates, and many others, the availability of a shared open database has been crucial in the development of analysis and theory. Based on these contributions, CHILDES serves as a model and inspiration for next-generation data-sharing projects in child development such as Databrary (http://databrary.org) and Wordbank (http://wordbank.stanford.edu).

3.2 PhonBank

During the first two decades of work on the CHILDES system, it was frustratingly difficult to adapt computer transcripts for the study of children's phonological development. Researchers used ASCII-based system such as ARPANET, SAMPA, PHONASCII, and UNIBET. However, application of these systems across languages was difficult and error-prone. The LIPP system (Nathani & Oller, 2001) solved some of these problems, but the proprietary nature of its font encoding made it difficult to integrate into transcripts, and it provided no linkage to media. With the introduction of Unicode in the 1990s and the promulgation of fonts supporting data entry for IPA such as Arial Unicode and the SIL Unicode IPA fonts (http://fonts.sil.org), it became increasingly easier to represent children's phonological productions in a standardized way. Building on this opportunity, Yvan Rose (Memorial University, Newfoundland) and Brian MacWhinney (CMU) initiated the PhonBank project. Working with a consortium of researchers in child phonology, and supported now for over 10 years by grants from NICHD, the PhonBank project has accumulated 40 corpora of early child phonological productions across 12 languages, all transcribed in IPA along with the target language forms and linked directly to the audio record. These new corpora are available in two formats: CHAT and Phon, and these two formats subscribe to the single underlying CHAT XML Schema that guarantees complete interoperability. Files in CHAT transcript format can be analyzed using the CLAN programs which we will describe later. Files in Phon format can be analyzed using the Phon program. Phon provides all the basic analyses required in the study of child phonology for tracking the growth of segments, features, prosodic patterns, and phonological processes. In addition, Phon incorporates the full source code of Praat (http://praat.org), making it possible to run Praat's acoustic analysis directly inside Phon and storing the results in the Phon transcript.

3.3 HomeBank

HomeBank, which began in 2015, is one of the newest components of TalkBank. It is supported by a grant from the National Science Foundation to Anne Warlaumont (UC Merced), Mark VanDam (Washington State University), and Brian MacWhinney (CMU). The primary data in HomeBank are daylong (i.e. 16-hour) audio recordings collected from children in the home through use of the LENA recording system (http://www.lena.org). This system uses a small digital recording device sewn into a

child's vest. The LENA software processes the captured audio to identify who is speaking when, but it does not attempt to recognize words. The output of this processing includes a text file in LENA's ITS format and the associated WAV file. To include these data in HomeBank, we use the LENA2CHAT conversion program in CLAN (http://childes.talkbank.org/clan) to output CHAT format. Researchers then select segments of these huge CHAT files for detailed language transcription. HomeBank currently includes 3.5 TB of these audio recordings and this number will soon grow well beyond this.

Because these data have no transcripts, we cannot provide public access to segments that may include potentially embarrassing material. Researchers interested in working with the non-public versions of these data must undergo careful debriefing regarding this issue before they are given access. To make at least some of this huge quantity of material publicly available, our students and research assistants listen through complete recordings to spot any questionable material, which they then tag in the CHAT transcript with a code for later silencing. Determining what should count as embarrassing material in these natural contexts is itself an interesting research topic.

Even without transcripts, these recordings can address many issues regarding the language environment of the young child. How much input is the child receiving and when? Do children who receive more input acquire language more quickly and does that help them in later years? How much responsivity do different adults show to child vocalizations? How do a child's intonational patterns change over time? These and many other questions can be addressed even without additional coding. However, when these recordings are accompanied by video or when various new methods for automatic analysis are used, the data can address an even broader range of research questions. For example, we are currently working with Florian Metze (Metze, Riebling, Warlaumont, & Bergelson, 2016) to apply the Speech Recognition Virtual Kitchen (SRVK) methodology (http://speechkitchen.org) to the CHAT and audio files derived from LENA. InterSpeech 2017 includes a challenge to see how well the SRVK methodology can diarize these recordings and identify the various speakers. If this methodology proves to be as good as that provided by the LENA system, we will work to make it available through open source, and we will work to create inexpensive recording devices that can be used with this non-proprietary software.

3.4 AphasiaBank

Aphasia involves the loss of language abilities, often arising from a stroke or embolism. This condition affects nearly 2 million people in the United States alone, making it the most common adult communication disorder. To improve our understanding of language usage and recovery in aphasia, NIH has been funding the AphasiaBank project for 10 years. Unlike the other language banks, AphasiaBank emphasizes the collection of data based on a tightly specified elicitation protocol. This protocol requires that the investigator follow a script in terms of asking questions and eliciting narratives. The detailed components of the protocol can be found at http://aphasia.talkbank.org/protocol. Using this standardized protocol, we have collected, transcribed and analyzed 402 hour-long interviews from persons with aphasia (PWAs) and 220 age-matched control participants. All transcripts are linked to the video at the utterance level and can be played back using the TalkBank browser over the web. Analysis of these materials have generated 256 publications, and the videos are used as teaching materials in universities and clinics throughout the English-speaking world. AphasiaBank also has smaller numbers of recordings for French, Cantonese, Spanish, and German, collected through translations of the protocol and the protocol materials into these languages.

We plan several extensions of AphasiaBank. First, we will record and transcribe increasingly naturalistic interactions in both group therapy sessions and conversations in the home. Second, we will test out the effects on language recovery of the use of tablet-based teletherapy lessons. Finally, we will use the Speech Kitchen methodology noted above to analyze the productions of people with aphasia and people with apraxia of speech (AoS) when reciting a scripted passage. The advantage of this method for speech recognition is that the words that must be recognized are restricted to those in the scripted passage.

3.5 Other Clinical Banks

Following the lead of AphasiaBank, we have developed protocols for data collection from four other varieties of language disorder. DementiaBank already includes a fairly large sample from earlier projects on language in dementia. We will formulate a data collection protocol for this area. RHDBank

examines the language and problem-solving abilities of people who have suffered from right hemisphere damage. TBIBank examines language from people suffering from traumatic brain lesions. Both RHDBank and TBIBank use a protocol close to that of AphasiaBank. Finally, ASDBank includes data from both children and adults with autism spectrum disorder.

3.6 FluencyBank

The other most recently funded TalkBank component is FluencyBank, based on a collaboration between Nan Bernstein Ratner (University of Maryland) and Brian MacWhinney (CMU). The development of FluencyBank is supported by two separate federal grants. The grant from NIDCD seeks to characterize the development pathway of fluency and disfluency in children between the ages of 3 and 7. During this period, many of the children that show signs of early disfluency end up as normally fluent, with only a fraction of this population developing stuttering. How and why this occurs developmentally remains a mystery, largely because data from this period are incomplete. To address this, we are using TalkBank methods to conduct a longitudinal study across this period. To supplement this work, NSF has provided support for incorporating data from earlier studies of disfluency from a variety of laboratories, much of it coded in SALT format.

Work in speech technology is centrally important for the development of FluencyBank. We need to not only analyze transcripts for lexicon, morphology, and syntax, but also carefully track word and segment repetitions, retraces, drawls, and overall durations. Ideally, these data should be linked to the audio records through a process of automatic diarization. Our initial work with this method indicates that this is feasible.

3.7 SLABank and BilingBank

SLABank currently includes 31 corpora from second language learners, and BilingBank includes 10 corpora from bilinguals. Nearly all of these corpora are accompanied by audio, although only a few have been linked to the audio at the utterance level. In addition to these corpora from adult learners and bilinguals, the CHILDES database has 32 corpora tracing the development of childhood bilingualism. To facilitate the analysis of grammatical development, we have developed a method for tagging multilingual corpora using a combination of unilingual taggers. This system is based on the taggers and parsers we have developed for Cantonese, Danish, Dutch, English, French, German, Hebrew, Japanese, Italian, Mandarin, and Spanish (MacWhinney, 2008). For bilingual corpora that use any combination of these languages, we use marks to encode the language source of each word. To minimize the actual marks being used, we establish the notion of a matrix (Myers-Scotton, 2005) language, so that only intrusions into the matrix are marked. This form of coding not only allows efficient tagging, but also provides a good profile of code-switching behavior.

We hope to be able to link this growing corpus collection with data from experimental and tutorial approaches to second language learning as characterized in a recent proposal for establishment of an SLAWeb (MacWhinney, in press).

3.8 CABank and SCOTUS

Conversation Analysis (CA) is a methodological and intellectual tradition stimulated by the ethnographic work of Garfinkel (1967) and systematized by Sacks, Schegloff, and Jefferson (1974) among others. With support from the Danish BG Bank Foundation, Johannes Wagner (Southern Denmark University) and Brian MacWhinney (CMU) developed methods for producing Jeffersonian CA transcription within CHILDES. We then collected and formatted a database of CA materials, including such classics as Jefferson's Newport Beach transcripts and the Watergate Tapes. There are currently 20 other corpora in CABank. One particularly large corpus that is not yet in CA format is the SCOTUS corpus developed in collaboration with Jerry Goldman (University of Illinois). This corpus – the largest in TalkBank – includes 50 years of oral arguments from the US Supreme Court linked on the utterance level to the audio. We also have a CHAT-encoded versions of the Santa Barbara Corpus of Spoken American English (SBCSAE) and the Michigan Corpus of Academic Spoken English (MICASE). CHAT/CA is being used in a variety of labs internationally that are planning to contribute additional data.

3.9 ClassBank

ClassBank includes 15 corpora of transcripts linked to video from classroom interactions. The largest of these are the Curtis corpus from a year-long study of instruction in Geometry in fourth grade (Lehrer & Curtis, 2000) and the seven-nation TIMMS study of teaching in Math and Science (Stigler, Gallimore, & Hiebert, 2000).

3.10 SamtaleBank

The creation of SamtaleBank was supported by a DK-CLARIN grant to Bente Maegaard (University of Copenhagen) and Johannes Wagner (Southern Denmark University). This bank includes the conversational component of the current DK-CLARIN corpus for Danish. All materials are carefully transcribed in CA format and linked to either the audio or video media. This collection serves as a model for the further construction of well-prepared materials for Conversation Analysis.

3.11 GestureBank

Creating a database of videotaped, transcribed, and coded interactions for the study of gestures during speaking has proven to be one of the most difficult challenges facing TalkBank. Coding and transcribing gestures is much more difficult than coding and transcribing spoken language. Unlike spoken language, there is no accepted method for gesture coding or transcription. Even if one tries to implement one of the dozens of proposed methods, it can take as long as a week to code one hour of gesture. Partly as a result of these problems, data-sharing has not taken hold as a norm in this community. Faced with these challenges, our work in this area has focused on the construction of a coding system that can be deployed more simply within the framework of the CLAN editor and programs. Our initial proposals along these lines are included along with other tutorial screencasts at http://talkbank.org/screencasts .

3.12 LangBank

With support from an NEH/DFG binational grant, Anke Lüdeling (Humboldt University, Berlin), Detmar Meurers (Tübingen), and Brian MacWhinney (CMU) are developing methods based on TalkBank, SLAWeb, and ANNIS (http://corpus-tools.org). In this project, we are creating systematized and aligned JSON versions of corpora for both Classical Latin and Historical German. This language bank represents an exception to the TalkBank focus on spoken language, because neither of these classical languages is actively spoken in a language community today. Instead, the focus here is on the development of these corpora in the SLAWeb framework (MacWhinney, in press) to support effective language learning. Moreover, this collaboration allows us to make contact with CLARIN-related groups studying issues such as corpus analysis (Berlin) (Lüdeling, Walter, Kroymann, & Adolphs, 2005) readability (Tübingen) (Meurers, 2005, 2012; Meurers et al., 2010), and the learning of classical languages (Leipzig).

3.13 Usage

To monitor the usage of the various components of TalkBank, we can track indices such as articles published and web hits. We are able to rely on http://scholar.google.com to track usage, because we have requested that people using these data include a reference to (MacWhinney, 2000) in their reference list. Table 1 summarizes these indices for the six major funded TalkBank components.

	CHILDES	Talk Bank	Aphasia Bank	Phon Bank	Fluency Bank	Home Bank
Age (years)	28	12	8	6	0.5	1
Words (millions)	59	47	1.8	0.8	0.5	audio
Linked Media (TB)	2.8	1.1	0.4	0.7	0.3	3.5
Languages	41	22	6	18	4	2

Publications	7000+	320	256	480	5	5
Users	2950	930	390	182	50	18
Web hits (millions)	4.3	1.3	0.3	0.1	-	0.2

Table 1: TalkBank Usage

4 Principles

TalkBank relies on a series of principles for data sharing, formatting, access, analysis, and user involvement. In this section, we will review these principles. Many of these principles adhere closely to CLARIN standards. Others seek to expand on these standards.

4.1 Data-sharing

The most fundamental TalkBank principle is the idea that the results of scientific investigations should be shared with the scientific community. This principle may seem like a platitude. We all know that scientists are supposed to open their ideas to further testing and development. However, as we noted earlier, data-sharing has not been adopted as the norm in many areas of the social sciences. The core goal of TalkBank is to correct this situation by building easy methods for data-sharing that will lead to important results for scientific investigation. CLARIN subscribes to similar principles.

Data-sharing can be encouraged through either the carrot or the stick. However, the only stick that carries much weight is one wielded by a funding agency. Agencies such as NIH and NSF now stipulate that, at the end of a project, the results of the project should be fully shared and archived. Funding agencies in Europe have also moved increasingly toward this standard. However, there remain large gaps in the enforcement of these standards. This is beginning to change, as granting agencies have begun to require that proposals must document the effective sharing of data from earlier funded research.

Researchers often claim that they cannot share data because of IRB (Human Subjects) restrictions. However, if there is proper planning and administration of informed consent at the beginning of a study, IRB problems can all be resolved. Similarly, investigators often complain that, if they contribute their data to a database like TalkBank, other researchers could publish analyses that might preempt or "scoop" their own plans for publication. This concern can be addressed by contributing data along with the specification of an embargo period, after which data will become publicly available.

The other approach to data-sharing involves carrots. In past decades, carrots have been more effective than sticks. Researchers have learned that contributed data will be cited, thereby increasing their citation index. To facilitate citation, we associate DOI (Digital Object Identifier) numbers with each corpus. Also, researchers find that by working with TalkBank they become members of a community of interest that furthers their communication with researchers having similar interests. In addition, by contributing data to TalkBank, researchers can use the increasingly powerful TalkBank tools to perform new analyses on their own corpora. This could be done without data contribution, because the programs are all open access. However, if we know that corpora are to be contributed to TalkBank, we will devote special attention to customizing analytic programs for the needs of particular projects.

4.2 Open access

Data-sharing implies open access. If a researcher contributes a corpus to a database, but refuses to permit open access, this is not real data-sharing. Corpora can be protected by passwords if necessary, but these passwords should be readily granted to qualified researchers. Provision of open access to corpora has been a problem for other database efforts, including some of those in CLARIN. Some archives only permit access to data through a search interface. This may work for certain types of queries, but it places restrictions on the types of questions that a researcher may pose regarding a dataset. In other cases, corpora are really not available at all. For example, many of the materials in The Language Archive (tla.mpi.nl) are not available for access. Limits on access also make it difficult for projects such as Linked Open Data or Federated Content Search. Hopefully many of these restrictions on access to corpora will be removed in the future.

4.3 Consistent format

A third important TalkBank principle is that all the data in TalkBank are transcribed in a single consistent format. This is the CHAT format which is compatible with the CLAN programs. This format has been developed over the years to accommodate the needs of a wide range of research communities and disciplinary perspectives. The format is described discursively in the CHAT manual, which is available from http://talkbank.org/manuals/CHAT.html. The full computational description is provided in the XML Schema viewable and browsable from http://talkbank.org/software/xsddoc/ . This XSD description includes links between the XML characterizations of CHAT elements and their description in the MS-Word manual.

Before data are entered into one of the TalkBank databases, they must first pass through two levels of format checking. The first level relies on the CHECK program built into CLAN. Because this checker is built right into the CLAN Editor, it is easy for users to check their work frequently to make sure they are following the requirements of CHAT. This checker is able to catch most potential errors in the use of CHAT format. However, the fullest checking is done through the Chatter XML formatter and validator that can be downloaded from http://talkbank.org/software/chatter.html. Chatter is able to convert files in CHAT format into XML and vice versa. It can also output PHON format.

4.4 Interoperability

Using conversion programs available inside CLAN, transcripts in CHAT format can be automatically converted both to and from the formats required for Praat (praat.org), PHON (childes.talkbank.org/phon), ELAN (tla.mpi.nl/tools/elan), CoNLL, ANVIL (anvil-software.org), EXMARaLDA (exmaralda.org), LIPP (ihsys.com), SALT (saltsoftware.com), LENA (lenafoundation.org), and Transcriber (trans.sourceforge.net). To provide fuller database and corpus facilities, we created a Pepper importer (Zipser & Romary, 2010) from CHAT data to ANNIS (http://corpus-tools.org) as well as a local ANNIS server (http://gandalf.talkbank.org:8080/annis-gui-3.4.4/).

Because CHAT recognizes such a wide variety of information types (dates, speaker roles, intonational patterns, retrace markings, etc.), when data are converted into the other formats, there must be methods for protecting CHAT data types not recognized by these other programs against loss. This is done in two ways. First, we can often "hide" CHAT data in special comment fields that are not processed by the program, but which will be available later for export. Second, when using the other programs, users are warned to be careful not to alter codes in CHAT format that mark aspects not recognized by the other programs. There are no cases in which information created in the other programs cannot be represented in CHAT, because CHAT is a superset of the information represented in these other programs.

In some cases, these conversions between CHAT and other formats involve the minimalist level of interoperability characterized by annotation graphs (Bird & Liberman, 2001). This level simply marks the beginning and end of some annotation in terms of its time from the beginning of the media. This is the type of remapping achieved for imports and exports to ELAN, ANVIL, Transcriber, and EXMARaLDA. However, other forms of conversion, such as those involving LIPP, LENA, SALT, ANNIS, and PHON include a full remapping of the semantics of the codes used in each format in their corresponding values in CHAT. The two screenshots in Figure 1 give example of the results of these types of transfer. The screenshot on the left shows data from a CHAT transcript that has been exported to and opened in PHON; the one on the right shows CHAT data has been exported to and opened in ANNIS.

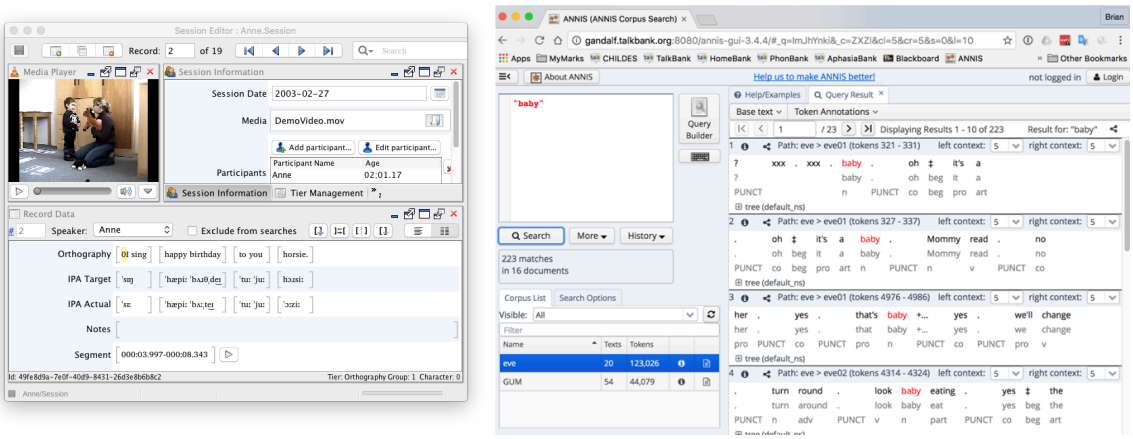

Figure 1: CHAT data that have been exported to PHON (left) and ANNIS (right)

During the process of building the database, we often needed to reformat files from still other, less documented formats. However, now most of the material we receive is already in CHAT format.

4.5 Media Linkage, Diarization

The majority of corpora in TalkBank have transcripts linked to either audio or video at the utterance level. By linking transcripts to the original recordings, we have lifted a burden off of the shoulders of the transcriber. Without linkage, transcription is forced to fully represent all of the important details of the original interaction. With linkage, transcription serves as a key into the original recording that allows each researcher to add or modify codes as needed. If a phonetician does not agree with the transcription of a segment of babbling, then it is easy to provide an alternative transcription.

The linkage of transcripts to recordings opens up a new way of thinking about corpora and the process of data sharing. In the previous model, we could only share the computerized transcripts themselves. For some important child language corpora, such as the Brown corpus, the original recordings have been lost. For others, however, we have been able to locate the original reel-to-reel recordings and convert them to digital files that we then link to the transcripts. Now, when corpora are contributed to TalkBank, we make sure that contributors provide both the transcripts in CHAT and the media.

Linkage to media on the utterance level is valuable for many aspects of language analysis from CA to child language. However, diarization through automatic speech recognition (ASR) methods can provide a more precise characterization of the temporal profiles of words and utterances. Diarization marks the time values for each word, allowing us to also find the values of intra-sentential and inter-sentential pauses. This type of analysis is important for work on language disorders and studies of turn-taking. One of our goals for the future is to increase the diarization of TalkBank corpora.

4.6 Protocol Formulation

Projects such as AphasiaBank and FluencyBank rely heavily on the construction of a data elicitation protocol to maximize the comparability of results across participants. The composition of these protocols is determined by an Advisory Board composed of members of each research community. The goal here is to be systematically compare data from speakers at different ages, speaking different languages, in different tasks and situations, at different stages of learning, and with different clinical profiles. To facilitate these comparisons, we have developed a series of programs for each relevant database. For aphasia, the program is called EVAL. Using this program, we can extract group means for individual aphasia types (Broca's, Wernicke's, anomia, global, transcortical motor, and transcortical sensory) which we then use as comparisons for the results from individual participants with aphasia. The screenshot in Figure 2 shows some of the options which can be used when comparing a given participant with the larger database.

Figure2: Options for comparing a transcript with a database in EVAL

For child language data, the parallel program is called KIDEVAL and it uses mother-child play sessions in the full CHILDES database as its comparison sample. The comparison database for FluencyBank is under construction. Comparisons of this type are fundamental to the process of clinical assessment, as well as the study of basic developmental processes.

4.7 Analysis Tools

For ten of the languages in the database, we provide automatic morphosyntactic analysis using the MOR, POST, and MEGRASP programs built into CLAN. These languages are Cantonese, Chinese, Dutch, English, French, German, Hebrew, Japanese, Italian, and Spanish. Tagging is done by MOR, disambiguation by POST, and dependency analysis by MEGRASP. MOR was written by Mitzi Morris, based on specifications for a left-associative morphology (LA-MORPH) from Roland Hausser (Hausser, 1999). POST was developed by Christophe Parisse (Parisse & Le Normand, 2000) and MEGRASP was developed by Kenji Sagae (Sagae, Davis, Lavie, MacWhinney, & Wintner, 2007). Details regarding the operation of the taggers, disambiguators, and dependency analyzers for these languages can be found in MacWhinney (2008). In each of these languages processing involves unique computational challenges. The complexity and linguistic detail required for analysis of Hebrew forms is perhaps the most extensive. In German, special methods are used for achieving tight analysis of the elements of the noun phrase. In French, it is important to mark various patterns of suppletion. Japanese requires quite different codes for parts of speech and dependency relations. Eventually, the codes produced by these programs will be harmonized with the GOLD ontology (Farrar & Langendoen, 2010). In addition, we compute a dependency grammar analysis for each of these 10 languages, which we are harmonizing with the Universal Dependency tagset (http://universaldependencies.org). It is also possible to use other dependency taggers rather than MEGRASP by reformatting a CHAT into CONLL format using the CHAT2CONLL and CONLL2CHAT programs in CLAN. The results of the morphological analysis by MOR and POST are stored in the %mor lines of a CHAT files and the results of the grammatical dependency analysis produced by MEGRASP are stored in the %gra lines. Triple clicking on a %gra line in a CHAT files invokes the GraphViz web service that produces a graph of the utterances for display on the user's screen, such as the one in Figure 3 for the first sentence from Julius Caesar's *De Bello Gallico*.

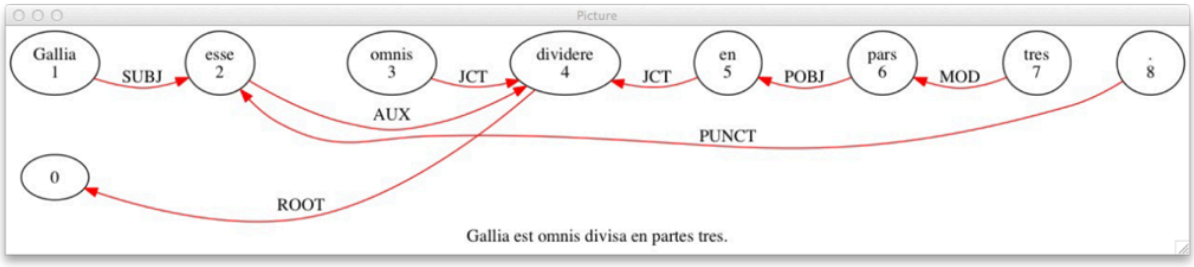

Figure 3: A dependency graph produced by GraphViz from a CHAT %syn line

Because these morphosyntactic analyzers use a parallel technology and output format, CLAN commands can be applied to each of these 10 languages for uniform computation of indices such as MLU (mean length of utterance), VOCD (vocabulary diversity) (Malvern, Richards, Chipere, & Purán, 2004), pause duration, and various measures of disfluency. In addition, we have automated language-specific measures such as DSS (Lee, 1974) (for English and Japanese) and IPSyn (Scarborough, 1990). Following the method of Lubetich and Sagae (2014), we are now developing language-general measures based on classifier analysis with SVN that can be applied to all 10 languages using the codes in the morphological and grammatical dependency analyses. However, there are many other languages in the database for which we do not yet have morphosyntactic taggers. This means that it is a priority to construct MOR systems for languages with large amounts of CHILDES and TalkBank data, such as Catalan, Indonesian, Polish, Portuguese, and Thai.

4.8 Metadata Publication

Metadata for the transcripts and media in the TalkBank databases are included in the two major systems for accessing linguistic data: OLAC, and CMDI/TLA. Each transcript and media file has been assigned a PID (permanent ID) using the Handle System (www.handle.net). In addition, each corpus has received an ISBN number and a DOI (digital object identifier) code. PID numbers are encoded in the header lines of each transcript file and the ISBN and DOI numbers are entered in 0metadata.cdc files included in each corpus as well as in HTML web pages that include extensive documentation for each corpus, photos and contact information for the contributors, and articles to be cited when using the data. All these resources are periodically checked and synchronized using the SCONS program that relies on the fact that there is a completely isomorphic hierarchical structure for the CHAT data, the XML versions of the CHAT data, the HTML web pages, and the media files. If information is missing for any item within this parallel set of structures, the updating program reports the error and it is fixed. All this information is then published using an OAI-PMH compatible method for harvesting through systems such as the Virtual Linguistic Observatory (VLO) developed through the CLARIN. Currently 13% of the records in the VLO come from TalkBank.

4.9 Community Support and Sustainability

Corpus creators may believe that making a database easily available will lead to its general usage. The idea is that you "build it and they will come". However, a fuller version of this motto would be "build it, curate it, publicize it, make usage easy, construct clear documentation, and provide workshops and free snacks, and they will eventually come." In practice, all these things are necessary, and we have worked continually on all these fronts to incorporate the use of TalkBank data and methods into training and research practice.

Making the system easily available is closely linked to the goal of sustainability. TalkBank's approach to sustainability focuses on integrating our corpora and tools with the basic research agenda of each of our participating language research communities. To the degree that we achieve such integration, funding for our work is tied to ongoing funding for basic research. For example, when developing tools for the study of child language development, we focus on methods for automatic morphosyntactic coding, because of the importance of grammatical analysis in language acquisition theory. For aphasia, we focus on morphosyntax, lexical access, error analysis, and aspects of fluency. For the projects on disfluency and stuttering, we work on the application of tools for automatic speech recogni-

tion (ASR), including diarization and word-level alignment to characterize the linguistic environment and distribution of disfluencies. We also seek to achieve sustainability and survivability by using open-source software tools with full documentation and by linking to tool chains in the CLARIN infrastructure.

5 Integration with CLARIN

It is our goal to make TalkBank materials fully accessible and discoverable for CLARIN users, and to integrate CLARIN tools into the TalkBank analysis chains. The award of CLARIN-B Centre status indicates that much of this integration has already been achieved. We have implemented all the requirements for this status both for CLARIN and for Data Seal of Approval recognition. We achieved further integration in 2016, through the recognition of TalkBank as a CLARIN-K Centre for Knowledge distribution. In this role, TalkBank will provide information for researchers interested in working with spoken language corpora, using either CLAN or any of the other software analysis system with which CLAN and CHAT are compatible. We can offer support through email, mailing lists, and phone with extremely quick turnaround. We have been creating online screencasts demonstrating the use of TalkBank tools, and we welcome suggestions for the creation of additional methods. These resources can become particularly important if CLARIN seeks to provide a higher level of support for the study of spoken language interactions.

The major challenge currently facing TalkBank integration into CLARIN is a fiscal one. Because the United States is not a member of the European Union, it has no clear mechanism for providing financial support for CLARIN membership. In 2017, the CMU University Library agreed to provide modest support for integration with CLARIN. We hope to extend this first step by creating a CLARIN Infrastructure with multiple research sites in the United States, such as Brandeis, the University of Pennylvania, the University of Illinois, or Columbia. How we can secure long-term funding across these sites remains to be seen.

The process of integration of TalkBank with CLARIN can also be viewed from a slightly different perspective. In addition to making sure that TalkBank aligns with CLARIN standards, we can consider how CLARIN could benefit more fully from TalkBank as a model. First, if CLARIN could adopt the CHAT coding system as the default for data representation for spoken language, it would greatly enhance the value of its resources. Such a step would require buy-in from many parties and additional work in reformatting, but it would be a major step forward. Second, adoption of TalkBank methods for promoting open access and data-sharing would be of great value to CLARIN. Finally, CLARIN could benefit from developing ways of linking sustainability to the development of specific corpora and tools that are crucially relevant to individual research groups. By making its tools a fundamental part of the infrastructure of research communities, CLARIN could guarantee its long-term survival.

6 Conclusion

TalkBank seeks to provide data that can help us integrate insights about language from across all the human sciences. To achieve this goal, it has developed a series of component data banks focusing on specific aspects of human language, all made comparable through a uniform transcription standard and principles for data-sharing.

TalkBank plays an important role within the larger CLARIN infrastructure in terms of providing resources for the analysis of spoken language interactions. Unlike many other resources in this area, TalkBank resources are available through completely open access and rely on a consistent data format. The individual components of TalkBank are each responsive to the practices and agenda of individual research communities. These features of TalkBank may serve as a model for parallel developments in CLARIN.

Acknowledgements

The development of TalkBank is currently supported by NICHD grant HD082736 to Brian MacWhinney for CHILDES, NICHD grant HD051698 To Yvan Rose and Brian MacWhinney for PhonBank, NIDCD grant DC008524 to Brian MacWhinney for AphasiaBank, NSF SBE RIDIR Grants 1539129, 1539133, and 1539010 to Anne Warlaumont, Mark VanDam, and Brian MacWhinney for HomeBank,

NEH/DFG grant to Brian MacWhinney, Anke Lüdeling, and Detmar Meurers for LangBank, NIDCD Grant DC015494 to Brian MacWhinney and Nan Ratner for new FluencyBank data, and NSF SBE Grant to Brian MacWhinney and Nan Ratner for FluencyBank archiving.

References

Baroni, M., & Kilgarriff, A. (2006). *Large linguistically-processed web corpora for multiple languages.* Paper presented at the Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations.

Bernstein Ratner, N., & MacWhinney, B. (2016). Your laptop to the rescue: Using the Child Language Data Exchange System archive and CLAN utilities to improve child language sample analysis. *Seminars in Speech and Language, 37*, 74-84.

Biber, D. (1991). *Variation across speech and writing*: Cambridge University Press.

Bird, S., & Liberman, M. (2001). A formal framework for linguistic annotation. *Speech Communication, 33*, 23-60.

Clahsen, H., & Rothweiler, M. (1992). Inflectional rules in children's grammars: Evidence from German participles. In G. Booij & J. Van Marle (Eds.), *Yearbook of Morphology*. Dordrecht: Kluwer.

Farrar, S., & Langendoen, D. T. (2010). An owl-dl implementation of gold *Linguistic Modeling of Information and Markup Languages* (pp. 45-66): Springer.

Freudenthal, D., Pine, J., & Gobet, F. (2010). Explaining quantitative variation in the rate of Optional Infinitive errors across languages: A comparison of MOSAIC and the Variational Learning Model. *Journal of Child Language, 37*, 643-669.

Givon, T. (2005). *Context as other minds: The pragmatics of sociality, cognition, and communication*. Philadelphia, PA: John Benjamins.

Hausser, R. (1999). *Foundations of computational linguistics*. Berlin: Springer.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., . . . Sainath, T. N. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine, 29*(6), 82-97.

Hymes, D. (1962). The ethnography of speaking. *Anthropology and human behavior, 13*(53), 11-74.

Lee, L. (1974). *Developmental Sentence Analysis*. Evanston, IL: Northwestern University Press.

Lehrer, R., & Curtis, C. L. (2000). Why are some solids perfect? *Teaching Children Mathematics, 6*(5), 324.

Lüdeling, A., Walter, M., Kroymann, E., & Adolphs, P. (2005). Multi-level error annotation in learner corpora. *Proceedings of corpus linguistics 2005*, 15-17.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk. 3rd Edition*. Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. (2008). Enriching CHILDES for morphosyntactic analysis. In H. Behrens (Ed.), *Trends in corpus research: Finding structure in data* (pp. 165-198). Amsterdam: John Benjamins.

MacWhinney, B. (2015a). Introduction: Language Emergence. In B. MacWhinney & W. O'Grady (Eds.), *Handbook of Language Emergence* (pp. 1-32). New York, NY: Wiley.

MacWhinney, B. (2015b). Multidimensional SLA. In S. Eskilde & T. Cadierno (Eds.), *Usage-based perspectives on second language learning* (pp. 22-45). New York, NY: Oxford University Press.

MacWhinney, B. (in press). A shared platform for studying second language acquisition. *Language Learning*.

MacWhinney, B., & Fromm, D. (2015). AphasiaBank as Big Data. *Seminars in Speech and Language, 37*, 10-22.

MacWhinney, B., & Leinbach, J. (1991). Implementations are not conceptualizations: Revising the verb learning model. *Cognition, 29*, 121-157.

MacWhinney, B., & Wagner, J. (2010). Transcribing, searching and data sharing: The CLAN software and the TalkBank data repository. *Gesprächsforschung, 11*, 154-173.

Malvern, D., Richards, B., Chipere, N., & Purán, P. (2004). *Lexical diversity and language development*. New York, NY: Palgrave Macmillan.

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., Xu, F., & Clahsen, H. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, i-178.

Metze, F., Riebling, E., Warlaumont, A. S., & Bergelson, E. (2016). *Virtual machines and containers as a platform for experimentation*. Paper presented at the Cognitive Science, Philadelphia, PA.

Meurers, D. (2005). On the use of electronic corpora for theoretical linguistics: Case studies from the syntax of German. *Lingua, 115*(11), 1619-1639.

Meurers, D. (2012). Natural language processing and language learning. *The Encyclopedia of Applied Linguistics*. doi:10.1002/9781405198431.wbeal0858

Meurers, D., Ziai, R., Amaral, L., Boyd, A., Dimitrov, A., Metcalf, V., & Ott, N. (2010). *Enhancing authentic web pages for language learners.* Paper presented at the Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications.

Myers-Scotton, J. (2005). Supporting a differential access hypothesis: Code switching and other contact data. In J. F. Kroll & A. M. B. DeGroot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 326-348). New York, NY: Oxford University Press.

Nathani, S., & Oller, D. K. (2001). Beyond ba-ba and gu-gu: Challenges and strategies in coding infant vocalizations. *Behavior Research Methods, Instruments, & Computers, 33*(3), 321-330.

Parisse, C., & Le Normand, M.-T. (2000). Automatic disambiguation of the morphosyntax in spoken language corpora. *Behavior Research Methods, Instruments, and Computers, 32*, 468-481.

Pennebaker, J. W. (2012). *Opening up: The healing power of expressing emotions*: Guilford Press.

Pine, J. M., & Lieven, E. V. M. (1997). Slot and frame patterns and the development of the determiner category. *Applied Psycholinguistics, 18*, 123-138.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a Parallel Distributed Processing Model of language acquisition. *Cognition, 29*, 73-193.

Redeker, G. (1984). On differences between spoken and written language*. *Discourse Processes, 7*(1), 43-55.

Rose, Y., & MacWhinney, B. (2014). The PhonBank Project: Data and software-assisted methods for the study of phonology and phonological development. In J. Durand, U. Gut, & G. Kristoffersen (Eds.), *The Oxford handbook of corpus phonology* (pp. 380-401). Oxford: Oxford University Press.

Sagae, K., Davis, E., Lavie, A., MacWhinney, B., & Wintner, S. (2007). High-accuracy annotation and parsing of CHILDES transcripts *Proceedings of the 45th Meeting of the Association for Computational Linguistics* (pp. 1044-1050). Prague: ACL.

Scarborough, H. S. (1990). Index of productive syntax. *Applied Psycholinguistics, 11*, 1-22. doi:10.1017/S0142716400008262

Stigler, J., Gallimore, R., & Hiebert, J. (2000). Using video surveys to compare classrooms and teaching across cultures: Examples and lessons from the TIMSS video studies. *Educational Psychologist, 35*(2), 87-100.

Valian, V., Solt, S., & Stewart, J. (2009). Abstract categories or limited-scope formulae? The case of children's determiners. *Journal of Child Language, 36*(04), 743-778.

VanDam, M., Warlaumont, A. S., Bergelson, E., Cristia, A., Soderstrom, M., Palma, P. D., & MacWhinney, B. (2016). HomeBank: An online repository of daylong child-centered audio recordings. *Seminars in Speech and Language, 37*(2), 128-142. doi:10.1055/s-0036-1580745

Wexler, K. (1998). Very early parameter setting and the unique checking constraint: A new explanation of the optional infinitive stage. *Lingua, 106*, 23-79.

Zipser, F., & Romary, L. (2010). *A model oriented approach to the mapping of annotation formats using standards.* Paper presented at the Workshop on Language Resource and Language Technology Standards, LREC 2010.