



# Understanding spoken language through TalkBank

Brian MacWhinney<sup>1</sup>

© The Psychonomic Society, Inc. 2018

## Abstract

Ongoing advances in computer technology have opened up a deluge of new datasets for understanding human behavior (Goldstone & Lupyan, 2016). Many of these datasets provide information on the use of written language. However, data on naturally occurring spoken-language conversations are much more difficult to obtain. A major exception to this is the TalkBank system, which provides online multimedia data for 14 types of spoken-language data: language in aphasia, child language, stuttering, child phonology, autism spectrum disorder, bilingualism, Conversation Analysis, classroom discourse, dementia, right hemisphere damage, Danish conversation, second language learning, traumatic brain injury, and daylong recordings in the home. The present report reviews these resources and describes the ways they are being used to further our understanding of human language and communication.

**Keywords** Child language · Aphasia · Conversation analysis · Bilingualism · Second language acquisition · Phonology · Computational linguistics · Corpora

Ongoing advances in computer technology have opened up a deluge of new datasets for understanding human behavior (Goldstone & Lupyan, 2016). Many of these datasets provide information on the use of written language. In contrast, data on naturally occurring spoken-language conversations are much more difficult to obtain. A major exception to this is the TalkBank system, which provides online multimedia data for 14 types of spoken-language data: language in aphasia, child language, stuttering, child phonology, autism spectrum disorder (ASD), bilingualism, Conversation Analysis, classroom discourse, dementia, right hemisphere damage (RHD), Danish conversation, second language learning, traumatic brain injury (TBI), and daylong recordings in the home. Five of these areas have accumulated very large data collections that are being used extensively to study the cognitive, neurological, developmental, and social bases of language processing and structure. The present report reviews these resources and describe the ways that they are being used to further our understanding of human language and communication. Given space limitations, this review cannot begin to cover the thousands of empirical contributions that have made use of these

resources. Instead, the focus here is on explaining the shape of the underlying system that has facilitated this work.

This review is designed to simultaneously address three rather different audiences. One group of researchers will already be familiar with segments of TalkBank. For these readers, the goal is to inform them of new resources and developments. Other researchers will not have made use of these resources. For them, the goal is to describe the basic ways in which TalkBank can be used to study language behavior. Finally, this review is intended to send a message to the larger scientific regarding the importance of access to real-life data in systems such as TalkBank. This is a message about the importance of the principles of data sharing, open access, uniform annotation standards, replicability, and responsivity to the needs of specific research communities. This message will hopefully encourage others to embark on and support similar enterprises.

## Components of TalkBank

The TalkBank system (<http://talkbank.org>) is the world's largest open-access integrated repository for spoken-language data. It provides language corpora and resources to support researchers in psychology, linguistics, education, computer science, and the speech sciences. The National Institutes of Health (NIH) and the National Science

---

✉ Brian MacWhinney  
macw@cmu.edu

<sup>1</sup> Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, USA

Foundation (NSF) have provided support for the construction of five of the components of TalkBank:

1. AphasiaBank, at <https://aphasia.talkbank.org>, for the study of language in aphasia in six languages;
2. CHILDES, at <https://childes.talkbank.org>, for the study of child language development in 42 languages, from infancy to age 6;
3. FluencyBank, at <https://fluency.talkbank.org>, for the study of fluency and disfluency in stuttering, aphasia, second language learning, and normal processing;
4. HomeBank, at <https://homebank.talkbank.org>, for the application of automatic speech recognition technology to untranscribed daylong recordings in the home and elsewhere; and
5. PhonBank, at <https://phonbank.talkbank.org>, for the analysis of children's phonological development in 18 languages.

The data in each of these banks involve multiple corpora that were contributed by individual researchers. In addition to support for these five funded areas, TalkBank also promotes the growth of spoken-language corpora in nine other areas:

6. ASDBank, at <https://asd.talkbank.org>, for the study of language in autism spectrum disorder;
7. BilingBank, at <https://biling.talkbank.org>, for the study of bilingualism and multilingualism;
8. CABank, at <https://ca.talkbank.org>, for the study of conversation using the methods of Conversation Analysis;
9. ClassBank, at <https://class.talkbank.org>, for the study of language in the classroom;
10. DementiaBank, at <https://dementia.talkbank.org>, for the study of language in dementia;
11. RHDBank, at <https://rhd.talkbank.org>, for the study of language in right hemisphere damage;
12. SamtaleBank, at <https://samtalebank.talkbank.org>, for the study of conversations in Danish;
13. SLABank, at <https://slabank.talkbank.org>, for the study of second language learning; and
14. TBIBank, at <https://tbi.talkbank.org>, for the study of language in traumatic brain injury.

Table 1 summarizes the size, use, and age of the five funded TalkBank databases. In that table, the contents of the other nine areas are included within the overall column labeled "TalkBank." The "Established" row indicates the number of years that the database has been in existence. The "Words" row gives the number of words, in millions. The "Media" row gives the size, in terabytes, of the linked audio or video media. The "Hits" row gives the number of web hits, in millions. The audio recordings in HomeBank are linked to transcripts that give vocal characteristics, but there is typically no full transcription of what is being said.

**Table 1** TalkBank descriptive statistics

	CHILDES	Aphasia Bank	Phon Bank	Fluency Bank	Home Bank	Talk Bank
Established	1984	2006	2010	2017	2016	2000
Words (mil)	59	1.8	0.8	0.5	audio	47
Media (TB)	1.8	0.6	1.6	0.3	2.8	1.1
Languages	41	6	18	4	3	24
Speakers	3,056	964	2,088	212	315	5,077
Publications	7,000+	256	480	5	18	320
Users	2,950	934	212	84	44	930
Hits (millions)	5.0	0.5	0.1	0.1	0.4	1.7

## 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 & Kilgariff, 2006) and behavioral (Pennebaker, 2012) analysis. On the other hand, the preparation of spoken-language data for computational analysis is much more difficult. Despite ongoing advances in speech technology (Hinton et al., 2012), the collection of spoken-language corpora still depends on the 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. This is unfortunate, because 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 express important aspects of processing and meaning, but that also complicate transcription and analysis.

Because of face-to-face communication's conceptual centrality, several major disciplines are concerned with it. These include psycholinguistics, developmental psychology, applied linguistics, clinical linguistics, phonology, theoretical linguistics, Conversation Analysis, gestural studies, human-computer interaction, social psychology, speech and hearing, neuroscience, evolutionary biology, and sociology. To understand and model the complex interactions and competitions (MacWhinney, 2014) involved in spoken language, researchers need to combine methods and insights from these various disciplines. Through such comparisons, and by examining language usage across a range of time scales (MacWhinney, 2015), 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. In this article, we will see how TalkBank has supported this process, leading to thousands of published articles, new methods for clinical practice, accessible support for education and professional development, and widespread adoption as a standard in many of these fields.

## TalkBank principles

The TalkBank system is grounded on six basic principles: maximally open data sharing, use of the CHAT transcription format, CHAT-compatible software, interoperability, responsiveness to research group needs, and the adoption of international standards.

1. **Maximally open data sharing** In the physical sciences, the process of data sharing is taken as a given. However, data sharing has not yet been adopted as the norm in the social sciences. This failure to share the results of research—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 is 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. Additionally, in this age of open social media, participants are often willing and eager to grant informed consent for open data sharing. Such consent for open access of identifiable material can override institutional review board concerns about the need to preserve anonymity and destroy data.
2. **CHAT Transcription format** Because individual researchers sample from the great diversity of language contexts, they tend to develop idiosyncratic methods for language transcription and analysis. Some subfields have developed transcription standards, but these are often not compatible with those used in related fields. To provide maximum harmonization across these formats, TalkBank has created an inclusive transcription standard, called *CHAT*, that recognizes all the features required by these different disciplinary analyses. The many possible features and codes available in this system are documented in the CHAT manual, which can be downloaded from <https://talkbank.org/manuals/chat.pdf>. CHAT can also be automatically converted to XML format through the use of the Chatter program (<https://talkbank.org/software/chatter.html>), in accord with the schema available at <https://talkbank.org/software/xsddoc/index.html>. Although the overall system is complex, much of this complexity is only relevant for special purposes, and the core methods of basic CHAT transcription are straightforward.
3. **CHAT-compatible software** TalkBank provides five systems for data analysis. Each of these systems provides a unique functionality that addresses a separate need for language analysis and research productivity.

The core system for TalkBank transcript analysis relies on a series of 30 commands compiled into a single program called

*CLAN*, at <http://dali.talkbank.org/clan/>. Written by Leonid Spektor and running on the Windows, Unix, and OSX platforms, CLAN allows users to conduct the basic operations of corpus analysis, such as frequency profiling, concordances of keywords in context, co-occurrence computation, word and utterance length computation, and many other functions. Currently, CLAN includes 30 analysis commands and 25 utility commands, each documented in the CLAN manual, which is freely downloadable from <https://talkbank.org/manuals/clan.pdf>. Much of the power of CLAN analyses derives from the fact that CLAN includes automatic taggers for part of speech (Parisse & Le Normand, 2000) and grammatical dependency structures (Le Franc et al., 2018; Lubetich & Sagae, 2014; Sagae, Davis, Lavie, MacWhinney, & Wintner, 2010) in 11 languages, including Cantonese, Danish, Dutch, English, French, German, Hebrew, Japanese, Italian, Mandarin Chinese, and Spanish.

The second set of TalkBank tools builds upon core CLAN programs to create a system for turnkey analysis of dozens of measures across a set of transcripts. Some of these measures, such as mean length of utterance, are computed by core CLAN commands. Others involve the computation of detailed grammatical profiles that previously had to be constructed tediously by hand. These include the Developmental Sentence Score (DSS, Lee, 1974) and the Index of Productive Syntax (IPSyn, Scarborough, 1990) for child language, as well as the Northwestern Narrative Language Analysis (NNLA, Thompson et al., 1995), Quantitative Production Analysis (QPA, Rochon, Saffran, Berndt, & Schwartz, 2000), and Computerized Propositional Idea Density Rater (CPIDR, Brown, Snodgrass, Kemper, Herman, & Covington, 2008) for aphasia. Both the basic measures and the more complex profile measures are then packaged together into either the KIDEVAL system for child language, the EVAL system for adult language analysis, or the FLUCALC system for stuttering (Bernstein Ratner & MacWhinney, 2018). Using these systems, a researcher or clinician can summarize the data for a group or a single participant and can compare a single participant with reference groups from the TalkBank database. These reference groups can be normal controls or participants with similar language disorders.

The third TalkBank support for language analysis is provided by the CLAN editor. This editor supports the creation of transcripts that use the CHAT transcription format. Utterances and words in these transcripts can be linked directly to corresponding segments of audio or video media, using either a waveform editor or the SoundWalker simulation of the function of the traditional foot pedal. The transcripts created by the editor are in a readable UTF-8 text format and can also be converted automatically into the TalkBank XML format, using the Chatter converter and validator written by Franklin Chen. Recently, Christophe Parisse of INSERM/CNRS and the Ortolang project built a powerful new editor called *TRJS* (<http://ct3.ortolang.fr/trjs/doku.php>) that works with the

CHAT, ELAN, and TEI formats. Because it uses more modern technology, we are hoping that the TRJS editor can soon replace the editor built into the CLAN program.

The fourth TalkBank support for language analysis is the TalkBank browser, which allows for direct playback from transcripts linked to media through a standard browser using HTML5 technology. This form of access to the database is particularly appropriate for exploratory analyses, because it allows for ready access to an enormous amount of data for the quick testing and generation of hypotheses. Because the transcripts and media are served from high-speed connections at the Carnegie Mellon Campus Cloud facility, playback is fairly smooth. This facility is also widely used for instructional purposes. For example, classes in neurolinguistics and clinical aphasiology can make use of the Grand Rounds instructional videos for aphasia, stuttering, TBI, and RHD, to familiarize students with the linguistic and communication patterns of different clinical types.

The fifth, and newest, TalkBank support for researchers is the TalkBankDB database facility, at <https://talkbank.org/DB>. This system was inspired by the CHILDES-DB system at <https://childes-db.stanford.edu>, created by Michael Frank and colleagues (see the article in this issue), and by the LuCiD Toolkit at <http://gandalf.talkbank.org:8080/>, created by Franklin Chang of the ESRC International Centre for Language and Communicative Development. The basic goal of each of these systems is to provide browser-based access to all of the contents of TalkBank transcript data. In this way, they differ from processing and analysis by means of CLAN, which runs instead in data-stream mode across a series of transcripts, rather than through access to a database. For advanced users, access to the database allows for the extraction of large quantities of data into spreadsheets for further analysis in R. For less advanced users, the web interface itself directly provides basic graphical and statistical analysis. For users on both levels, two types of filters are available for data selection. First, there is a system for choosing transcripts on the basis of criteria such as language, corpus name, participant age, socioeconomic status, and so forth. Second, there is a system for selecting data on the basis of linguistic queries that look at specific lexical items, parts of speech, and other co-occurrences. This second set of filters is not yet fully implemented, but it will be closely based on the Corpus Query Language (CQL), which is the standard form of RegEx (regular expression) multitier searching in corpus linguistics.

4. **Interoperability** The PhonBank component of TalkBank (<https://phonbank.talkbank.org>) has developed a separate program called *Phon* (Rose & MacWhinney, 2014), which is available from <https://github.com/phon-ca/phon/releases>. This program provides extensive support for the analysis of phonological data. The entire code and

functionality of the popular Praat program, at <http://www.fon.hum.uva.nl/praat>, are now included inside the Phon program. Because Phon stores data in CHAT XML format, transcripts in CHAT and Phon are fully compatible after conversion through Chatter. Compatibility with other common formats, including Anvil, CONNL, ELAN, EXMARaLDA, LENA, Praat, SRT, SALT, and Transcriber is achieved through translation programs inside CLAN.

5. **Responsivity to research community needs** TalkBank seeks to be maximally responsive to the needs of individual researchers and their research communities. Our most basic principle is that we attempt to implement all features that are suggested by users, in terms of software features, data coverage, documentation, and user support. We provide this support in six ways:

*Corpus pages:* We have configured separate web servers for each of the 14 TalkBank communities, each within the [talkbank.org](http://talkbank.org) domain. Each website provides an index to the available corpora. For example, the index at <https://ca.talkbank.org> lists the 34 available corpora for Conversation Analysis, along with links to four related TalkBank adult conversation databases. Clicking on any one of these links, such as the one for “Bergmann,” brings up a page with a description of the corpus, as well as photos and contact information for the contributors, articles and a DOI number for citation, a link for downloading the media, and a link for downloading the transcripts. The corpora vary widely in terms of the amount of description provided. For example, the Bergmann corpus page only tells us that these are emergency calls to the fire department. Other corpora, such as the one described at <https://aphasia.talkbank.org/access/English/Aphasia.html>, provide more extensive documentation.

*TalkBank browser:* Each corpus page also includes a link to the TalkBank browser, which allows users to play back linked multimedia corpora directly in their web browser. Users can choose to have either continuous playback or playback of specific sections or utterances.

*Tutorials:* To facilitate the process of learning how to use the TalkBank data and tools, we have created screencast tutorials at <https://talkbank.org/screencasts/>. These are hosted both at our own servers and through YouTube, for better distribution in certain parts of the globe.

*Grand rounds:* For AphasiaBank, RHDBank, and FluencyBank, we have constructed a series of instructional pages that allow students to learn about language disorders through direct playback of videos

linked to transcripts and commentary from experts in language analysis.

*Mailing lists:* For each TalkBank area, we maintain a user-oriented mailing list at <https://groups.google.com>.

*Presentations and workshops:* We also conduct presentations and workshops each year at international conferences.

The guiding principle underlying all these methods is that we seek to be maximally responsive to the needs of researchers and research groups, as well as to instructors and clinicians. We try to fulfill all requests for new corpora, new methods, new protocols, and new computational resources. In this way, we are able to maximize the participation of individual research groups in TalkBank.

6. **International standards** The sixth basic TalkBank principle is our commitment to defining international standards for database and language technology. Toward this end, TalkBank has joined the European CLARIN Federation (<https://clarin.eu>). CLARIN is an association of the computational linguistic communities in 21 European countries, supported by the European Union and the governments of the individual countries. CMU TalkBank is currently the only member of CLARIN outside of Europe. Much like TalkBank, CLARIN seeks to provide uniform computational methods for accessing and processing language data. Toward this end, CLARIN centers have implemented standards for publishing corpus metadata using the CMDI format with the Handle server and OAI-PMH software. On the basis of these metadata about corpora, CLARIN has constructed a Virtual Linguistic Observatory (<https://vlo.clarin.eu>) for locating linguistic resources, and nearly a third of the corpora in that system derive from TalkBank. CLARIN also promotes participation in the Core Trust Seal program for accreditation of data centers, and TalkBank has received this approval, as is noted in the extensive documentation at <https://www.coretrustseal.org/wp-content/uploads/2017/10/TalkBank.pdf>. The Core Trust Seal program emphasizes the adoption of international standards in areas such as ease of data access, protection of confidentiality, organizational infrastructure, data integrity, data storage, data curation, and data preservation. In accord with recent emphases on the reproducibility of experimental (Munafò et al., 2017) and computational (Donoho, 2010) analyses, TalkBank maintains incremental GIT repositories at <https://git.talkbank.org> for all of its datasets. Using this resource, researchers interested in replicating earlier analyses can obtain copies of segments of the database from any particular date.

## Many banks in one

TalkBank is composed of 14 component banks, each using the same CHAT transcription format and database organization standards. This section describes the contents and these component language banks, beginning with the five banks currently receiving federal support: CHILDES, PhonBank, HomeBank, FluencyBank, and AphasiaBank. For each of these, we will also consider some of the ways in which the data have been used, although a full review of the many thousands of published studies based on TalkBank data would be an overwhelming task.

**CHILDES** The Child Language Data Exchange System (CHILDES) database, at <https://childes.talkbank.org>, is the oldest of TalkBank's component banks. Brian MacWhinney (Carnegie Mellon University [CMU]) and Catherine Snow (Harvard School of Education) began the CHILDES system in 1984 with funding from the MacArthur Foundation. In the early 1980s, researchers were just beginning to use personal computers, and transcribed data were still stored in 9-track tapes, punch cards, and floppy disks. The Internet was not generally available for data transmission, so data were shared by mailing CD-ROM copies to members. Not imagining that transcripts might eventually be linked to audio and video, researchers often destroyed or recycled their audio recordings.

Since that early beginning, CHILDES has grown in coverage, membership, and output, thanks to continual support from NIH and NSF. 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 learning of the English past tense (MacWhinney & Leinbach, 1991; Marcus et al., 1992), and later on data from German plural formation (Clahsen, Rothweiler, Woest, & Marcus, 1992; Pine & Lieven, 1997). In syntax, emergentists (Pine & Lieven, 1997) have used CHILDES data to elaborate an item-based theory of how the determiner category is learned, 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 by using the MOSAIC system (Freudenthal, Pine, & Gobet, 2010) to demonstrate constraint-based inductive learning. CHILDES data are also frequently used to study how statistical learning and segmentation operate on parental input data (McCaughey, Monaghan, & Christiansen, 2015; Ngon et al., 2013). In these debates, and many others, the availability of a shared open database has been crucial to the development of analysis and theory. On the basis of these experiences and contributions, CHILDES has served as a model for other data-sharing projects in child development, such as Databrary (<http://databrary.org>) and Wordbank (<http://wordbank.stanford.edu>).

**PhonBank** During the first two decades of work on the CHILDES system, it was difficult to adapt computer transcripts for the study of children's phonological development. Researchers used ASCII-based systems, such as ARPANET, SAMPA, PHONASCII, and UNIBET, to encode phonological contrasts. However, the application of these systems across languages was difficult and error-prone. With the introduction of Unicode in the 1990s and the promulgation of fonts supporting data entry for the International Phonetic Alphabet (IPA), such as Arial Unicode and the SIL Unicode IPA fonts (<http://fonts.sil.org>), it became easy to represent children's phonological productions in a standardized way. Building on this opportunity, Yvan Rose at Memorial University, Newfoundland, and Brian MacWhinney at Carnegie Mellon University initiated the PhonBank project. Working with a consortium of researchers in child phonology, and supported by ongoing grants from the National Institute of Child Health and Human Development, the PhonBank project has accumulated 50 corpora of early child phonological productions across 18 languages, all transcribed in Unicode 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, both of which subscribe to the single underlying CHAT XML schema that guarantees complete interoperability. Files in CHAT transcript format can be analyzed using the CLAN programs. 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. These data have been used to evaluate the role in phonological development of factors such as phonological processes (dos Santos, 2007; Leonard & McGregor, 1991), individual variability (Costa, 2010), syllabic template structure (Vihman & Croft, 2007), and phonological disorders (McAllister Byun, 2012).

**HomeBank** HomeBank, which began in 2015, is one of the newest components of TalkBank. It is supported by a grant from the NSF to Anne Warlaumont (UC Merced), Mark VanDam (Washington State University), and Brian MacWhinney (CMU). The primary data in HomeBank are daylong (i.e., 16-h) audio recordings collected from children in their 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 in order 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 nonpublic 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 the complete recordings to spot any questionable material, which they then tag in the CHAT transcript with a code for later silencing. Even without transcripts, these recordings can address many issues regarding the language environment of the young child (VanDam et al., 2016). How much input does the child receive, 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 to apply the Speech Recognition Virtual Kitchen (SRVK) methodology (<https://github.com/srvk>) to the CHAT and audio files derived from LENA (Metze, Riebling, Warlaumont, & Bergelson, 2016), to obtain further details regarding diarization and speech recognition. InterSpeech 2018 included 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 make it available through open source, along with inexpensive recording devices that can be used with this nonproprietary software.

**AphasiaBank** Aphasia involves the loss of language abilities, often arising from an ischemic stroke that blocks blood supply to an area of the brain. This condition affects nearly two million people in the United States alone, making it the most common adult communication disorder. Unlike many of the other language banks, AphasiaBank emphasizes the collection of data based on a tightly specified elicitation protocol that requires the investigator to follow a script for 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 has generated 256 publications across the areas of discourse, grammar, lexicon, gesture, fluency, syndrome classification, social factors, and treatment effects, as summarized and reviewed in MacWhinney and Fromm (2016). AphasiaBank videos are used as teaching materials, as well, in universities and clinics globally. AphasiaBank also has smaller numbers of recordings for French, Cantonese, Mandarin, Spanish, and German, collected through translations of the protocol and the protocol materials into these languages. We are working on several extensions of AphasiaBank. First, we are recording and transcribing 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 SRVK methodology mentioned above to analyze the productions of people with aphasia and of people with apraxia of speech 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.

**FluencyBank** The most recently funded TalkBank component is FluencyBank, based on a collaboration between Nan Bernstein Ratner (University of Maryland) and Brian MacWhinney (CMU). FluencyBank seeks to characterize the development pathways of fluency and disfluency in children between the ages of 3 and 7 years. 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 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. In addition to this data collection work, we are incorporating data from earlier studies of disfluency from a variety of laboratories, many of them coded in SALT format (Miller & Chapman, 1983).

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 (Le Franc et al., 2018), in work that is also relevant to HomeBank and AphasiaBank. We have now packaged analyses of this type, along with specific codings for patterns of disfluency, into a new program called *FLUCALC*. The same methods that we are using in *FLUCALC* to characterize patterns of disfluency in children are also relevant to the study of second language fluency. In fact, researchers in the task-based language-teaching framework (Skehan, Foster, & Shum, 2016) have already begun using many of the TalkBank methods, such as

the computation of vocabulary diversity (Malvern, Richards, Chipere, & Purán, 2004) and syntactic complexity, that are embedded in *FLUCALC*.

**Other clinical banks** Following the lead of AphasiaBank, we have developed protocols for data collection from four other varieties of language disorder. DementiaBank includes a large set of audio-linked transcripts from earlier projects on language in dementia. These data are being widely used internationally to develop automatic methods for diagnosis of the onset and types of dementia through speech technology. RHDBank focuses on the language and problem-solving abilities of people who have suffered from RHD. TBIBank contains 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. Unlike the other components of TalkBank, data in the clinical banks are password protected. However, password access is readily granted to responsible researchers and clinicians.

**SLABank and BilingBank** Two other components of TalkBank focus on adult multilingualism. SLABank currently includes 31 corpora from second language learners, and BilingBank includes 12 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 that encode the language source of each word. To minimize the actual marks being used, we rely on 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 the establishment of an SLAWeb (MacWhinney, 2017).

**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). A priority for future TalkBank work is to extend our work in this important area.

**CABank, SCOTUS, and SamtaleBank** 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 30 corpora in CABank, although only 20 are in real CA format. One particularly large corpus that is not yet in CA format is the SCOTUS corpus, developed in collaboration with Jerry Goldman at the University of Illinois, Chicago. This corpus—the largest in TalkBank—includes 50 years of oral arguments from the US Supreme Court linked on the utterance level to audio. We also have CHAT-encoded versions of the Santa Barbara Corpus of Spoken American English (SBCSAE), the Michigan Corpus of Academic Spoken English (MICASE), and the spoken-language component of the British National Corpus. CHAT/CA is being used in a variety of labs internationally that are planning to contribute additional data. The SamtaleBank corpus (<https://samtalebank.talkbank.org>), developed by Johannes Wagner, includes an extremely well-transcribed set of CA materials linked to either video or audio for Danish.

## Conclusions

By providing maximally open access to data and analysis methods, TalkBank has stimulated many thousands of published research studies. It is also supporting practical applications for language therapy, clinical diagnosis, and second language teaching. Within each of the components of TalkBank, there is a continual need for the collection and analysis of additional languages and additional data types. Because language is such a central and complex aspect of our social, educational, and economic life, the need for TalkBank data will continue to grow, even as the methods for collecting and analyzing these data continue to improve. Given these forces, it seems safe to predict that the TalkBank system, or something evolving from it, will continue to be a major part of our scientific research infrastructure for many years to come.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## References

- Baroni, M., & Kilgarriff, A. (2006). Large linguistically-processed Web corpora for multiple languages. Paper presented at the Eleventh Conference of the European Chapter of the Association for Computational Linguistics, Trento.
- Bernstein Ratner, N., & MacWhinney, B. (2018). Fluency Bank: A new resource for fluency research and practice *Journal of Fluency Disorders*, 56, 69–80. <https://doi.org/10.1016/j.jfludis.2018.03.002>
- Brown, C., Snodgrass, T., Kemper, S. J., Herman, R., & Covington, M. A. (2008). Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior Research Methods, Instruments, & Computers*, 40, 540–545. <https://doi.org/10.3758/BRM.40.2.540>
- Clahsen, H., Rothweiler, M., Woest, A., & Marcus, G. (1992). Regular and irregular inflection in the acquisition of German noun plurals. *Cognition*, 45, 225–255.
- Costa, T. (2010). *The acquisition of the consonantal system in European Portuguese: Focus on place and manner features* (PhD dissertation), University of Lisbon, Lisbon, Portugal.
- Donoho, D. L. (2010). An invitation to reproducible computational research. *Biostatistics*, 11, 385–388.
- dos Santos, C. (2007). *Développement phonologique en Français langue maternelle: Une étude de cas*. (PhD dissertation), University Lumière Lyon 2, Lyon.
- 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.
- Garfinkel, H. (1967). *Studies in ethnomethodology*. Englewood Cliffs: Prentice-Hall.
- Givon, T. (2005). *Context as other minds: The pragmatics of sociality, cognition, and communication*. Philadelphia: John Benjamins.
- Goldstone, R., & Lupyán, G. (2016). Discovering psychological principles by mining naturally occurring data sets. *Topics in Cognitive Science*, 8, 548–568. <https://doi.org/10.1111/tops.12212>
- 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, 82–97.
- Le Franc, A., Riebling, E., Karadayi, J., Yun, W., Scaff, C., Metze, F., & Cristia, A. (2018). *The ACLEW DiViMe: An easy-to-use diarization tool*. Paper presented at Interspeech 2018, Mumbai, India.
- 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, 324.
- Leonard, L., & McGregor, K. (1991). Unusual phonological patterns and their underlying representations: A case study. *Journal of Child Language*, 18, 261–271.
- Lubetich, S., & Sagae, K. (2014). Data-driven measurement of child language development with simple syntactic templates. Paper presented at the 25th International Conference on Computational Linguistics (COLING 2014), Dublin.
- 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. (2014). Presentation. In L. Scliar-Cabral (Ed.), *O português na plataforma CHILDES* (pp. 9–20). Florianópolis, Portugal: Editora Insular.
- MacWhinney, B. (2015). Introduction: Language emergence. In B. MacWhinney & W. O'Grady (Eds.), *Handbook of language emergence* (pp. 1–32). New York: Wiley.
- MacWhinney, B. (2017). A shared platform for studying second language acquisition. *Language Learning*, 67, 254–275.

- MacWhinney, B., & Fromm, D. (2016). AphasiaBank as big data. *Seminars in Speech and Language, 37*, 10–22. <https://doi.org/10.1055/s-0036-1571357>
- MacWhinney, B., & Leinbach, J. (1991). Implementations are not conceptualizations: Revising the verb learning model. *Cognition, 29*, 121–157.
- Malvern, D., Richards, B., Chipere, N., & Purán, P. (2004). Lexical diversity and language development. New York: 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, 57*(4). <https://doi.org/10.2307/1166115>
- McAllister Byun, T. (2012). Positional velar fronting: An updated articulatory account. *Journal of Child Language, 39*, 1043–1076.
- McCauley, S., Monaghan, P., & Christiansen, M. (2015). Usage-based language learning. In B. MacWhinney & W. O’Grady (Eds.), *The handbook of language emergence* (pp. 415–436). New York: Wiley.
- Metze, F., Riebling, E., Warlaumont, A. S., & Bergelson, E. (2016). *Virtual machines and containers as a platform for experimentation*. Paper presented at Interspeech 2016, San Francisco, CA. 10.21437/Interspeech.2016-997
- Miller, J., & Chapman, R. (1983). SALT: Systematic analysis of language transcripts, user’s manual. Madison: University of Wisconsin Press.
- Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., du Sert, N. P., ... Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*, 0021. <https://doi.org/10.1038/s41562-016-0021>
- 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: Oxford University Press.
- Ngon, C., Martin, A., Dupoux, E., Cabrol, D., Dutat, M., & Peperkamp, S. (2013). (Non) words, (non) words, (non)words: Evidence for a protolexicon during the first year of life. *Developmental Science, 16*, 24–34. <https://doi.org/10.1111/j.1467-7687.2012.01189>
- Parris, C., & Le Normand, M.-T. (2000). Automatic disambiguation of the morphosyntax in spoken language corpora. *Behavior Research Methods, Instruments, & Computers, 32*, 468–481. <https://doi.org/10.3758/BF03200818>
- Pennebaker, J. W. (2012). *Opening up: The healing power of expressing emotions*. New York: 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.
- Redeker, G. (1984). On differences between spoken and written language. *Discourse Processes, 7*, 43–55. <https://doi.org/10.1080/01638538409544580>
- Rochon, E., Saffran, E., Berndt, R., & Schwartz, M. (2000). Quantitative analysis of aphasic sentence production: Further development and new data. *Brain and Language, 72*, 193–218.
- 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.
- Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50*, 696–735. <https://doi.org/10.2307/412243>
- Sagae, K., Davis, E., Lavie, A., MacWhinney, B., & Wintner, S. (2010). Morphosyntactic annotation of CHILDES transcripts. *Journal of Child Language, 37*, 705–729. <https://doi.org/10.1017/S0305000909990407>
- Scarborough, H. S. (1990). Index of productive syntax. *Applied Psycholinguistics, 11*, 1–22. <https://doi.org/10.1017/S0142716400008262>
- Skehan, P., Foster, P., & Shum, S. (2016). Ladders and snakes in second language fluency. *International Review of Applied Linguistics, 54*, 97–112.
- 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*, 87–100.
- Thompson, C. K., Shapiro, L. P., Tait, M. E., Jacobs, B. J., Schneider, S. L., & Ballard, K. J. (1995). A system for the linguistic analysis of agrammatic language production. *Brain and Language, 51*, 124–129.
- Valian, V., Solt, S., & Stewart, J. (2009). Abstract categories or limited-scope formulae? The case of children’s determiners. *Journal of Child Language, 36*, 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*, 128–142. <https://doi.org/10.1055/s-0036-1580745>
- Vihman, M., & Croft, W. (2007). Phonological development: Toward a “radical” templatic phonology. *Linguistics, 45*, 683–725.
- Wexler, K. (1998). Very early parameter setting and the unique checking constraint: A new explanation of the optional infinitive stage. *Lingua, 106*, 23–79.