frontiers | Frontiers in Communication

METHODS
published: 27 April 2022
doi: 10.3389/fcomm.2022.865498

Language Sample Analysis With TalkBank: An Update and Review

Brian MacWhinney and Davida Fromm*

Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, United States

This chapter examines state-of-the-art methods for coding, analyzing, and interpreting discourse-level language data from children and adults with language disorders using the data, tools, and methods provided by the TalkBank system (https://www.talkbank.org). These open and free methods have been used for language sample analysis (LSA) with several clinical populations (e.g., child language disorders, stuttering, aphasia, dementia, traumatic brain injury, right hemisphere brain damage), as well as with control participants without communication impairments. We review the six core principles guiding TalkBank, the current shape of the 15 TalkBank databanks, and the different analytic tools provided by TalkBank. We examine automatic TalkBank methods that use ASR (automatic speech recognition), NLP (natural language processing), database technology, statistics in R and Python, and ML (machine learning). The specific tools include corpus analysis methods, LSA profiling systems, online database searches through TalkBank, online browsing through transcripts linked to media, and a new system for online collaborative commentary. These systems provide multimedia access to transcripts from a wide variety of participants with and without language disorders.

Keywords: aphasia, child language, language sample analysis, TalkBank, automation, discourse assessment, automatic speech recognition

OPEN ACCESS

Edited by:
Lisa Milman,
Utah State University, United States

Reviewed by:
Chaleece Wyatt Sandberg,
The Pennsylvania State University (PSU), United States
Elizabeth Rochon,
University of Toronto, Canada

****Correspondence:***
Brian MacWhinney
macw@cmu.edu

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *29 January 2022*
Accepted: *04 April 2022*
Published: *27 April 2022*

Citation:
MacWhinney B and Fromm D (2022) Language Sample Analysis With TalkBank: An Update and Review. Front. Commun. 7:865498. doi: 10.3389/fcomm.2022.865498

INTRODUCTION

Language sample analysis (LSA) is the most complete and ecologically valid way to understand and assess language disorders. However, when done by hand, LSA can be tedious, incomplete, unreplicable, and inaccurate. The great hope is that technology can "come to the rescue" by automating language sample analysis for communication disorders. As we launch this rescue effort, we need to think about *who is being rescued* and *what are they being rescued from*? Perhaps a researcher needs to be rescued from the time and effort involved in finding and training research assistants to perform linguistic analyses on large amounts of discourse data collected from persons with aphasia (PWA) and controls. Perhaps, a University professor needs help finding good case examples to use in her graduate course on language disorders, especially if the University is not in a large, urban setting or if a pandemic makes it impossible to have the usual variety of clinical training experiences. Perhaps, a clinician wants to determine which aspects of discourse are most impaired in a child with language delay to quantify the child's impairment relative to a larger group matched on age and gender. However, that clinician can only dedicate perhaps 40 min to data collection and analysis. Perhaps, a graduate student at a small University in a rural area hopes to write a thesis about patterns in stuttering, but does not have the time, money, or resources to access the number or type of participants needed for the study. Perhaps, a working group of clinical researchers wants to determine

a core set of discourse tasks and measures with proven psychometric properties for use in research on treatments for improving expressive language in aphasia. For these and other applications that will be highlighted throughout this article, technology can make a big difference and positively impact teaching, diagnosis, and the study of language disorders.

The focus here is on state-of-the-art methods for coding, analyzing, and interpreting discourse-level language data from children and adults with language disorders using the data, tools, and methods provided by the TalkBank system (https://talkbank.org). These open and free methods can be applied to language samples from any clinical population (e.g., child language disorders, stuttering, aphasia, dementia, traumatic brain injury, right hemisphere brain damage), as well as to control participants without communication impairments (MacWhinney et al., 2018). Here, we will review the six core principles guiding TalkBank, the current shape of the 15 TalkBank databanks, and the different analytic tools provided by TalkBank. We will examine automatic TalkBank methods that use ASR (automatic speech recognition), NLP (natural language processing), database technology, statistics in R and Python, and ML (machine learning). These are designed to function within custom web browsers that provide multimedia access to transcripts from a wide variety of participants with and without language disorders.

THE TALKBANK SYSTEM

Component Databanks

TalkBank is a shared, multimedia database for the study of spoken language (MacWhinney, 2019). It includes separate databanks for these 15 population types:

1. AphasiaBank for language in aphasia,
2. ASDBank for language in autism spectrum disorder (ASD),
3. BilingBank for language in bilingualism,
4. CABank for Conversation Analysis (CA) data,
5. CHILDES for child language data,
6. ClassBank for classroom discourse data,
7. DementiaBank for the study of language in dementia,
8. FluencyBank for the study of disfluency,
9. HomeBank for daylong recordings in the home,
10. PhonBank for child phonology and phonological disorders,
11. PsychosisBank for language in psychosis,
12. RHDBank for language in right hemisphere disorder (RHD),
13. SamtaleBank for the study of conversation in Danish,
14. SLABank for the study of second language acquisition, and
15. TBIBank for the study of language in traumatic brain impairment (TBI).

Of these 15 databanks, the eight that focus on language disorders are AphasiaBank, ASDBank, DementiaBank, FluencyBank, PsychosisBank, RHDBank, and TBIBank, along with the clinical components of CHILDES and PhonBank. Much of the data in these clinical banks is password protected. However, access is given readily and quickly to researchers and clinicians who send an email request with their contact information and affiliation to macw@cmu.edu. In this review, we will focus on the use of TalkBank tools for aphasia and child language. However, most of these methods apply equally well to all eight databases for language disorders.

TalkBank Principles

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

Maximally Open Data-Sharing

In the physical biological 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 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 explained 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 (e.g., de-identification, audio bleeping, password protection, controlled viewing).

CHAT Transcription Format

Individual researchers and research groups tend to develop idiosyncratic methods for language transcription and analysis, thereby greatly complicating cross-corpus analysis. Some subfields have developed transcription standards, but these are seldom compatible with those used in related fields. To provide maximum harmonization across these formats, TalkBank has created an inclusive transcription standard, called CHAT, to recognize the many features of spoken language. These features and codes are documented in the CHAT manual which can be downloaded from https://talkbank.org/manuals/chat.pdf. CHAT can be converted automatically to XML and JSON format through 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 quite extensive, individual projects usually only need to use specific subsections of the full format.

CHAT-Compatible Software

Because all the data in TalkBank use the same transcription format, it is possible to create analysis programs and facilities that make use of this format. TalkBank provides ten analytic tools based on the CHAT format. These includes the CLAN analysis commands, a system for automatic morphosyntactic tagging, eight programs to produce clinical profiles, methods for ASR processing, the Phon program for phonological analysis, a system for doing CA (Conversation Analysis) transcription, the TalkBank browser for study of transcripts in the web browser, the TalkBankDB database search system, a new system for Collaborative Commentary, and web pages with teaching tools.

Interoperability

TalkBank emphasizes the use of CHAT format. However, there are other important transcript formats that are well-adapted to uses in specific communities. To unify the data coming from

these other formats, we have created a series of 14 programs for translating to and from these formats to CHAT. These other formats include Anvil, CA, CONLL, DataVyu, ELAN, LAB, LENA, LIPP, Praat, RTF, SALT, SRT, Text, and XMARaLDA. We are also developing interoperability with other language database systems through CLARIN's (https://clarin.eu) FCS (Federated Content Search) system.

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 through construction of web pages for each corpus, index pages for databanks, manuals for CHAT and CLAN, YouTube screencast tutorials, Google Groups mailing lists, article publications, conference presentations, and conference workshops. We receive overall guidance for the project from the TalkBank Governing Board.

International Standards

The sixth basic TalkBank principle is our adherence to international standards for database and language technology. In particular, we strive to adhere to the FAIR standards (Wilkinson et al., 2016) for open access to data. These standards hold that data should be *Findable, Accessible, Interoperable,* and *Reusable*. TalkBank promotes *Findability* by (1) registering websites for Google Search, (2) using OAI-PMH (https://openarchives.org) standards for harvesting of metadata in the CMDI format by the Virtual Linguistic Observatory (VLO at https://vlo.clarin.eu) and the Online Language Archives Community (OLAC), (3) including DOIs (digital object identifiers) and PIDs (permanent identifiers) for each TalkBank corpus, and (4) providing index pages in each databank with descriptions of datasets and cross-listings to releated datasets.

TalkBank promotes *Accessibility* by (1) providing fully open access to all TalkBank tools and programs, (2) providing open source to computer code, (3) documenting all aspects of the tools in the full CHAT, CLAN, and MOR manuals and online manuals for TalkBankDB and Collaborative Commentary, (4) providing YouTube tutorial screencast for the use of the tools, and (5) making transcript and media data open access whenever possible and accessible with a readily given password for other data. In no cases are any special DUAs (data use agreements) with special IRB (institutional research board) sign offs required.

TalkBank promotes *Interoperability* through the 14 data conversion programs mentioned earlier. It supports *Reusability* through methods for analysis replication. In accord with recent emphases on reproducibility of experimental (Munafò et al., 2017) and computational analyses (Donoho, 2010), TalkBankDB is configured to allow researchers to download data for accurate replication from any given time in the past back to 2018. To replicate studies based on use of the database before 2018, corpora can be pulled from the TalkBank git repositories. We are also developing methods for the exact replication of analyses by storing commands issued in TalkBankDB and in the TalkBank API for the R programming language.

TalkBank also adheres to the TRUST standards (Lin et al., 2020) for maintenance of reliable digital databases. These standards require *Transparency, Responsibility, User Focus, Sustainability,* and *Technology*. To comply with these standards, TalkBank has fulfilled the 16 requirements for the peer-reviewed CTS (Core Trust Seal at https://www.coretrustseal.org) certification. These requirements stipulate that all policies and procedures of the database be made publicly available through documentation at the website. Toward this end, we provide documentation regarding all aspects of governance, mission, licensing, continuity of access, confidentiality, organizational structure, expert guidance, data integrity, data intake appraisal, data storage, data preservation, data quality, workflows, data discovery, data reuse, technical infrastructure, and security. For example, there are TalkBank web pages that describe in detail how TalkBank data is managed and backed up by the CMU Campus Cloud facility. The fact that TalkBank has been given CTS certification for these 16 dimensions is based upon its adherence to the FAIR and TRUST standards.

The Databanks

To understand the ways in which the TalkBank tools automate LSA, it helps to understand the current contents of the databanks. We will focus here on AphasiaBank and CHILDES, although most of these features we describe for AphasiaBank apply equally well to the other clinical databanks, and many of the features we describe for CHILDES apply to the other child language banks.

AphasiaBank

AphasiaBank (MacWhinney et al., 2011) is the only openly available data source for spoken language and communication in aphasia. It has served as a model for the development of several other adult language databases: TBIBank, RHDBank, and DementiaBank. Currently, AphasiaBank has over 1,250 members from more than 55 countries. Hundreds of published research articles have utilized AphasiaBank data and methods (e.g., see https://aphasia.talkbank.org/publications/). Additionally, many conference presentations (https://aphasia.talkbank.org/posters/) and graduate theses/dissertations have relied on the use of the AphasiaBank database and methods.

AphasiaBank contains corpora that use a standard discourse protocol and test battery with large numbers of participants, allowing for the development of new discourse assessment tools and norms. Briefly, the discourse protocol includes personal narratives, picture descriptions, storytelling, and a procedural task. Detailed administration instructions and a script for the investigator were developed to ensure consistent implementation across sites. Most of the data collected since AphasiaBank's initial funding in 2007 is in English and includes over 450 videos and transcripts of PWAs and more than 250 videos and transcripts for controls. The participants come from 26 different sites in the United States and 1 site in Canada. The standard discourse protocol has been translated into Cantonese, Croatian, French, German, Italian, Japanese, Mandarin, Romanian, and Spanish. These corpora are smaller but also available at the website.

Originally, the standard discourse protocol was administered in person and with materials downloaded from the website. It has recently been adapted for computer-based administration, making it easier and more efficient for clinicians and researchers to collect data using these tasks. A webpage (https://aphasia.talkbank.org/protocol/english/) provides various scenarios and hyperlinks for administering the protocol to PWAs and controls using web-based or PowerPoint instructions and materials. Recording can be done directly from the program (e.g., Zoom) or the computer, avoiding the need to acquire and manage recording equipment and transfer media files. Currently, this is available for English only.

In addition to the large corpus of data using the standard discourse protocol, AphasiaBank contains over 20 corpora contributed by researchers who collected language data specific to their research goals. Examples include: (1) the QAB corpus, which contains video files for 19 PWAs doing the Quick Aphasia Battery with transcripts for the 5 min conversation segment (Wilson et al., 2018); (2) the Olness corpus, which contains transcripts and audio files from 50 PWAs and 30 controls, half of whom are Caucasian and half African American, doing a wide variety of discourse tasks and an ethnographic semi-structured interview; and (3) the SouthAL corpus, which contains transcripts and media files for 9 PWAs and 8 controls reading passages from the Gray Oral Reading Test (Wiederholt and Bryant, 2012).

CHILDES

The CHILDES database includes over 50,000 transcripts. About 40,000 of these come from cross-sectional studies in which each child contributes only one or two samples. These cross-sectional samples focus either on children's narratives or their language during freeplay. There are also 88 longitudinal case studies in which a given child may be followed for as many as 5 or even 6 years from the beginning of speech up to school age. Major languages such as English, Dutch, French, German, Spanish, Japanese, Indonesian, Mandarin, and Cantonese are heavily represented. However, there are also corpora from languages such as Basque, Cree, Quechua, Nungon, Hungarian, Estonian, and Sesotho. There are also 30 corpora from children learning two or more languages simultaneously. There are several large corpora from children with language disorders, although these are primarily in English. Older datasets were often contributed without audio. However, recent additions almost always have audio or video linked to the transcripts.

Two databanks closely related to CHILDES are PhonBank and HomeBank. PhonBank provides transcripts linked to audio for the study of the development of child phonology. Like other TalkBank data, PhonBank data can be processed through CLAN, TalkBankDB, or the TalkBank Browser. However, they can also be processed through the *Phon* program (https://phon.ca) which provides dozens of standard phonological indices and profiles, and which includes the full capabilities of the *Praat* system (Boersma, 2001). The other major child language database is HomeBank which provides access to daylong (16 h) recordings taken in the home. Most of these recordings were collected using the LENA system (Gilkerson et al., 2017). The CHAT files derived from this system are linked to speech turns in the audio, but there is no actual transcription of the speech.

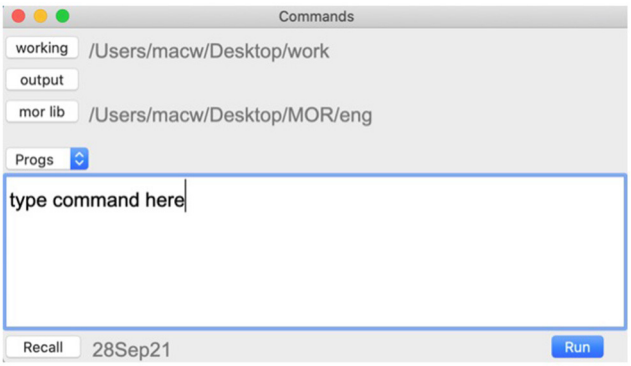

FIGURE 1 | CLAN commands window.

TALKBANK TOOLS

Having reviewed the shape of these two databanks, we turn next to a fuller description of the ways in which the nine TalkBank tools mentioned earlier can facilitate LSA for language disorders.

CLAN Commands

CLAN is the core desktop program for analysis of TalkBank data. It can be freely downloaded from https://dali.talkbank.org. CLAN includes 30 analysis commands and 25 utility commands, each documented in the CLAN manual that is freely downloadable from https://talkbank.org/manuals/clan.pdf. Commands are entered into a Commands window, as shown in **Figure 1**. In this window, the *working* directory identifies the location of the file(s) to be analyzed. The user sets this by clicking on the *working* button and navigating to the folder that contains the relevant CHAT files. The *mor lib* button should be set to the grammar for the language being analyzed (in this case *eng* for English). These grammars are accessed in the CLAN program from the *File* option in the Menu bar and the *Get MOR Grammar* selection.

For quick information about each CLAN command and its various options, users can type a command (e.g., EVAL) into the command window and press "Run" (or the return key). The CLAN Output page then provides a short description of the command and a list of "switches" (or options) that can be added to modify the command. For example, to use raw values instead of percentages in the EVAL command you can add +o4 to the command; or to select word mode analysis instead of syllables in FluCalc you can add +b. The Progs button in the Commands window can be used to pull up a menu-driven system for selecting commands, adding input files, and choosing program options. **Figure 2** illustrates the segment of this dialog system that selects input files. In this example, you see the main folder name, Aphasia, and the various corpora within the collection. Double clicking on any corpus opens a list of CHAT files in that corpus,

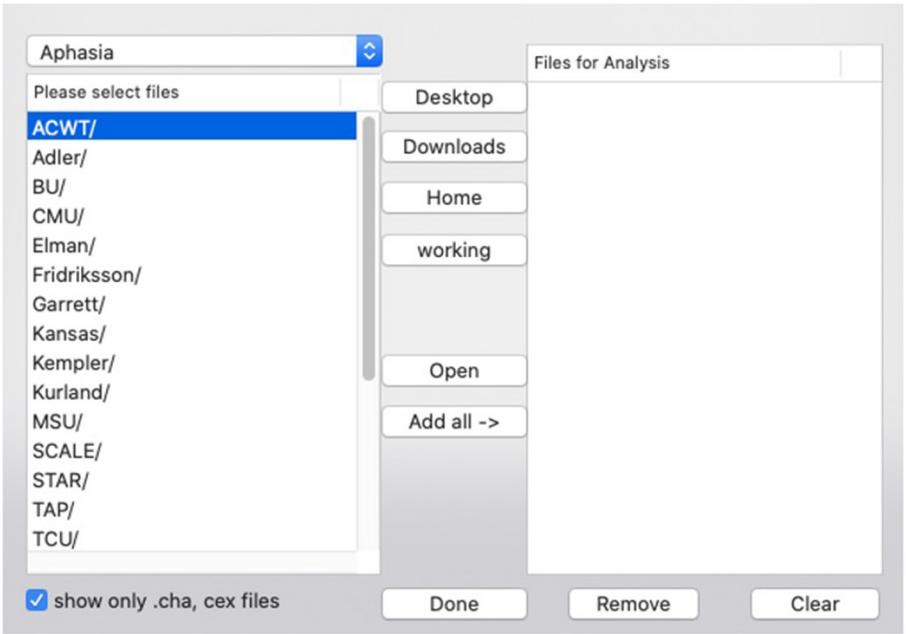

FIGURE 2 | Dialog for file selection.

which allows you to select the files you want to analyze. Then those files will appear in the "Files for Analysis" column.

CLAN commands can be divided into five groups:

1. Analysis commands. These provide basic corpus linguistic analysis functions, such as frequency lists, pattern searches, n-gram analysis, keyword and line (KWAL), mean length of utterance (MLU), lexical diversity, and others for a total of 27 commands.
2. Profiling commands. These include EVAL, KIDEVAL, FluCalc, C-NNLA, C-QPA, IPSyn, and DSS. Each of these will be discussed in detail below.
3. Morphosyntactic commands. These include the MOR, PREPOST, POST, POSTMORTEM, and MEGRASP commands which will be discussed in detail below.
4. Interoperability commands. These include the 14 commands for format conversion that were mentioned earlier.
5. Utility commands. These include 19 commands used to check, adjust, and improve the format of CHAT files.

CLAN Editor and CHAT

In addition to providing access to these various commands, CLAN can also serve as an editor. As much as possible, the functions of the CLAN editor mirror those that are familiar to users from MS-Word. However, unlike MS-Word, the files created by the CLAN editor are pure text files encoded in UTF-8 that can be read directly by other text editors. **Figure 3** displays a CLAN editor window with a transcript from AphasiaBank.

The first 6 lines in this example display header tiers that describe the participants and media. Line 7 indicates the beginning of the segment of the AphasiaBank protocol that asks the participant to describe "how your speech is these days." The "@G:" is a gem marker that facilitates later retrieval and analysis of specific segments from one or multiple transcripts. After that, the lines marked as *INV for the Investigator and *PAR for the Participant give the spoken words. Speaker IDs like *INV and *PAR can be quickly inserted through keystroke shortcuts. Each utterance ends with a little bullet mark that encodes the beginning and end time of the utterance in milliseconds for direct playback from the audio or video. If you expand the bullets, you can see the time stamp, and the bullet on the Investigator's first utterance would look like this: •0_2927•. Under each utterance are dependent tiers. In this transcript they include only the %mor and %gra lines which provide the automatically computed morphological and grammatical relations analysis, both of which are explained below.

Transcript files in TalkBank all have a ".cha" or CHAT extension which allows them to be opened directly in CLAN by double-clicking. The editor provides four methods to speed transcription through direct linkage to the audio, a system for checking correct use of CHAT, and a variety of other methods to speed transcription. The default font for CHAT files is Arial Unicode which allows for representation of the characters of all languages. Entry of characters from languages that write from right to left is possible. However, combining right-to-left script with the left-to-right features of CHAT can be tricky. For that reason, we recommend the use of romanization for languages with right to left orthographies.

CHAT has many other codes for special features of spoken language, some of which can be seen in **Figure 3**. Commonly used codes in AphasiaBank transcripts include: &-for fillers, &+ for sound fragments, &= for gestures, [/] repetitions, [//] for

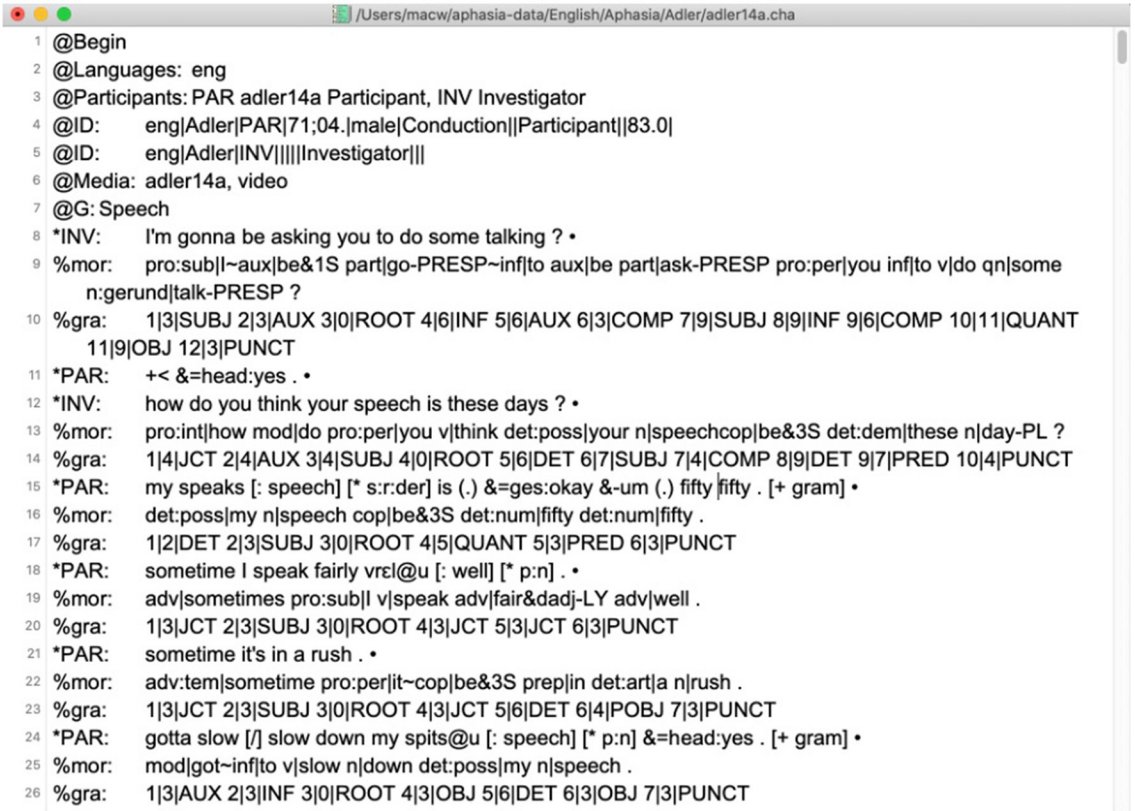

FIGURE 3 | An AphasiaBank transcript opened in the CLAN Editor.

revisions, +… for a trailed-off utterance, xxx for unintelligible content, @u placed at the end of phonetic transcriptions, + < to indicate overlapping speech, [*] to indicate an error production, [: target] for the target word following an error production, [+] for optional utterance level coding, and (.) for a short pause. In addition to 32 special characters with keystroke entry methods (https://ca.talkbank.org/codes.html) for Conversation Analysis (CA) coding, there is also a comprehensive system for coding errors using the [* code] format shown in **Figure 3**.

Compared with other computer-based programs, word/text files, or spreadsheets, transcription in CHAT has several advantages. First, because morphology and syntax can be automatically analyzed through the MOR program discussed below, there is no need for special marking of these features, and transcription can just use standard orthography. Second, because the CLAN editor allows direct linkage to the media, transcription can be faster and more accurate. Third, the use of consistent systems for marking of behaviors such as revisions, repetitions, fillers, and sound fragments allows for tabulations and searches of these features. Most importantly, transcription in CHAT makes it possible to include the results of the work in the shared Open Science TalkBank database with all its additional methods for analysis and profiling.

ASR and Transcription

For many years, practical use of automatic speech recognition (ASR) seemed like a promise that was always disappearing over the horizon. However, in the last 8 years, there have been steady improvements in the word-level accuracy of ASR, driven by new computational methods implemented on increasingly powerful hardware that uses huge collections of spoken language derived from web platforms. These advances now make it possible to use web-based ASR systems to create the initial version of a transcript for further human-driven checking and formatting. There are many available commercial systems for this. We have tested only 10 of them and most perform reasonably well, but we have found that Amazon Web Services (AWS) ASR was able to provide the best accuracy for our purposes. Our methods for going from AWS output to CHAT transcripts can be found at https://talkbank.org/info/ASR/. However, these methods only work for well-recorded audio from non-clinical adult participants speaking standard American English (SAE). When recording quality goes down, when other versions of English are involved, or when the participants are children or adults with language disorder, then ASR accuracy is no longer acceptable. Once large training sets for these other populations become available, we hope that this situation can continue to improve.

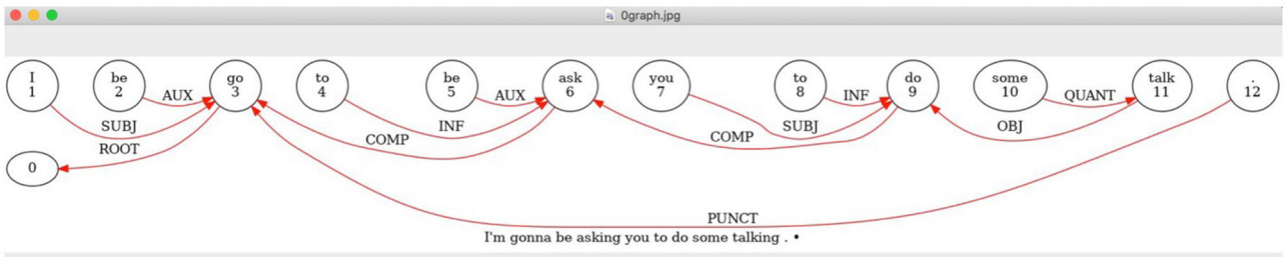

FIGURE 4 | Dependency graph for the first utterance in **Figure 3**.

MOR/POST/MEGRASP

After creating a transcript, the user can run CLAN's MOR command to automatically insert two lines under each utterance: the %mor tier, which has morphological and part-of-speech parsing; and the %gra tier, which shows pairwise grammatical relations between words, as illustrated in **Figure 3**. We can use the first word in line 9 in **Figure 3** to illustrate how to read items on the %mor line. This code analyzes the word *I'm* as *pro:sub|I~aux|be&1s*. The tilde sign (~) in the middle of this analysis indicates that this is a cliticization of the auxiliary onto the pronoun. The first person pronoun is coded as pro:sub|I where *pro:sub* stands for subject pronoun and the form after the bar is the lemma or stem. For the auxiliary, the form after the bar is the lemma *be* which is marked as being in the first person singular. In the %gra tier on line 10, the first word is tagged as *1|3|SUBJ*. Here, the "1" indicates that this is the first word. The "3" indicates that the word is grammatically related to the third word which is the verb "go." For the syntactic analysis, the cliticized auxiliary is treated as an item. One can double click the %gra line to fire up a web service that throws a graphic display to your screen of the utterance's dependency structure with arcs labeled for the relevant grammatical relations, as in **Figure 4**.

The MOR command runs in a matter of seconds, firing these five programs in linked succession on anything from a single file to a collection of folders of files:

1. MOR generates possible morphological analyses of the words on the main line (excluding repetitions). Using declarative rules for allomorph generation, it builds a runtime tree structure which is then processed in a left-associative manner (Hausser, 1999) through a set of continuation rules until the end of the word is reached. Each successful run of a chain of continuation rules activates a morpheme (stem or affix) to be added to the analysis. The result is a set of possible analyses which then need to be disambiguated to choose the correct analysis.
2. PREPOST filters the readings generated by MOR by applying simple context-sensitive rules based on alternative readings of part-of-speech (POS) sequences.
3. POST is a trainable system based on probabilities of sequences. It looks at a window of 2 words before and 2 words after the target word to find the single most probable candidate, based on the POS categories of the words. Because it relies on POS categories, it cannot disambiguate alternatives with a given part of speech.
4. POSTMORTEM uses a set of context-sensitive rules much like those of PREPOST to make a final corrective pass over the results produced by POST. Given careful formulation of PREPOST, POST, and POSTMORTEM, along with blocking rules in MOR, only a few ambiguities will remain, and those can be resolved through a command in the CLAN editor or a set of exclude rules for special cases. The result of this chain is a fully disambiguated morphological analysis written out to the %mor line. The information on the %mor is used for many of the automatic discourse analysis commands we will discuss later. Morphological tagging accuracy of CLAN for English has consistently been between 95 and 97% (MacWhinney et al., 2011). There are also versions of MOR for 10 other languages (https://talkbank.org/morgrams/).
5. MEGRASP then uses the output on the %mor line to generate a grammatical dependency analysis (Kübler et al., 2009) of the type illustrated graphically in **Figure 4**. The MEGRASP parser is trained using a disambiguated training set to create a support vector network (SVN) classifier that creates the dependency structure. If the %mor line is accurate, the accuracy of the %gra line is about 92%.

Profiling

TalkBank provides eight tools for creating clinical profiles of individual participants or clients. Clinical profiling has a long history in the field of Speech-Language Pathology with systems such as DSS (Lee, 1966), IPSyn (Scarborough, 1990), and LARSP (Crystal, 1982; Ball et al., 2012) targeting child language and systems like NNLA (Thompson et al., 1995b) and QPA (Rochon et al., 2000) targeting language in aphasia. These systems were all based on hand analysis of specific lexical and structural items found in an LSA transcript. For most of these, the analysis can compare the target participant with a control reference group matched for age, gender, and other features. However, these reference groups generally included as few as 20–30 subjects.

More recently there have been at least three efforts to automate LSA-based profiling. The most extensive effort uses the SALT program and database to evaluate a target transcript on six measures (Tucci et al., 2021). However, these six measures focus primarily on the quantity of speech produced in a recording session without tracking the morphological, lexical, and syntactic

details of systems such as DSS, IPSyn, or LARSP. A second system called SUGAR (Pavelko and Owens, 2017) offers a Microsoft Word-based method for quickly computing a basic profile of a similar type. TalkBank's CLAN program provides a third approach to this issue. Profiling in CLAN combines automaticity of analysis with the linguistic and analytic detail of the original measures. The eight CLAN profiling commands are EVAL, C-QPA, C-NNLA, and CoreLex for aphasia and KIDEVAL, C-IPsyn, C-DSS, and FluCalc for child language. Each of these relies on a comparison set taken from the many hundreds of transcripts available in either AphasiaBank or CHILDES.

EVAL

EVAL produces a language profile for PWAs with 34 output measures such as total utterances, total words, mean length of utterance, type-token ratio, words per minute, percent or raw number of various parts or speech, noun-verb ratio, and open-class to closed-class word ratio (Forbes et al., 2012). An important aspect of this command is the option to compare an individual's performance to the full AphasiaBank database for any of the six tasks in the standard AphasiaBank discourse protocol. For example, one could compare a client's description of the Cat in Tree picture (Brookshire and Nicholas, 1994) to controls or to other PWAs with the same type of aphasia. The comparison group can also be specified by age and sex. Results, in spreadsheet format, show means and standard deviations for the client and the comparison group, with asterisks indicating where the target transcript differs from the group mean by one or two standard deviations. Another feature of this command is the option of comparing a given individual's performance pre-treatment and post-treatment to see where changes occurred. Researchers have used this command to generate large datasets and select the variables of interest for their studies. For example, Boucher et al. (2020) assessed the relationship between quantitative measures of connected speech and performance in confrontation naming in 20 individuals with early post-stroke aphasia and 20 controls. EVAL was used to extract 10 micro-linguistic variables such as duration, speech rate, total number of words, mean length of utterance, and lexical diversity from CHAT transcriptions of a picture description task. Stark (2019) used the EVAL command to extract six primary linguistic measures including propositional density, verbs per utterance, and type-token ratio in her large study comparing three discourse elicitation methods in 90 PWAs and 84 controls. Finally, the Teaching resource section of AphasiaBank includes a classroom activity using the EVAL program on a picture description task from three PWAs with different types of aphasia (anomic, Broca's, conduction) and comparing them with controls. The activity has multiple options as well as questions to guide students in using the information provided by the analysis results.

C-QPA and C-NNLA

C-NNLA and C-QPA commands automatically compute outcome measures from two well-established grammatical analysis systems, the Northwestern Narrative Language Analysis (Thompson et al., 1995a) and the Quantitative Production Analysis (Saffran et al., 1989; Rochon et al., 2000). These systems have been used in aphasia research for decades, providing highly detailed analyses of aspects of morphological content (number of regular and irregular plurals, possessives), general language measures (mean length of utterance, number of words and utterances), lexical variables (e.g., number of nouns, verbs, pronouns), and structural analysis (e.g., number of utterances, embeddings, verb phrases, subject noun phrases) that have advanced the science, specifically in our understanding of agrammatic speech. When scored by hand, both systems require considerable training, linguistic expertise, and time. The automated commands can be of huge benefit to researchers for efficient and reliable analyses of large numbers of discourse samples. These analyses require slightly more extensive CHAT transcription (e.g., with full error coding as explained at the AphasiaBank *Discourse Analysis* webpage) and may therefore be less practical for busy clinicians.

CoreLex

CoreLex computes the number of core lexicon words used based on normed core lexicon lists for the five AphasiaBank discourse protocol tasks (Dalton et al., 2020). This command produces a spreadsheet showing how many and specifically which core lexicon words were used in a language sample or set of language samples. These results can be used to assess typical language usage (Dalton and Richardson, 2015; Kim et al., 2019). A recent study compared automated and manual CoreLex scoring and found them to be highly correlated, with automated scoring again requiring a small fraction of the time that it takes to train scorers and score manually[1].

SCRIPT

SCRIPT compares a participant's transcript to a model transcript such as a therapy script or a reading passage. The spreadsheet output computes the number and percent of correct words, number and percent of omitted words, number of added words, number of recognizable errors, number of unrecognizable errors, number of utterances with unintelligible content, and number of missing utterances. This command was useful in a study examining the treatment effects of script training (Szabo et al., 2014) and increased the efficiency of clinically relevant efficacy analyses across participants. We were also able to use the SpeechKitchen software (https://srvk.github.io) to do phoneme-level diarization of script productions from persons with apraxia of speech (AoS) to determine which phoneme patterns were causing the greatest difficulty (MacWhinney et al., 2017).

KIDEVAL

The KIDEVAL program for child language evaluation is similar in concept to the EVAL program for evaluation of language in aphasia. Like EVAL, it allows the analyst to compare a target transcript with the larger CHILDES database in terms of matching age range, gender, and recording type (freeplay, narrative, interview). KIDEVAL includes many of the same measures as EVAL, along with automatic runs of the IPSyn and

[1]Dalton, S. G., Stark, B., Fromm, D., Apple, K., MacWhinney, B., Rensch, B., et al. (2022). Comparing automated and manual scoring modalities for core lexicon analysis. *Am. J. Speech Lang. Pathol.* doi: 10.31234/osf.io/ex7q5 (under review).

DSS profiling schemes. However, for IPSyn and DSS, it only outputs the overall score and not the detailed profile. It also outputs frequencies of usage for the 14 grammatical morphemes tracked in Brown (1973).

IPSYN

Beginning in 2005, CLAN included a computerized version of the popular Index of Productive Syntax (IPSyn) (Scarborough, 1990) method for profile analysis of children's productions up to age 6. IPSyn provides scores along 59 grammatical structures, including the noun phrase, the verb phrase, questions and negation, and overall sentence structure. The initial versions of the IPSyn command did not adequately match the results from hand coders. Fortunately, recent improvements in the rule set now allow it to compete successfully with hand coding (MacWhinney et al., 2020). Moreover, use of the automatic version greatly speeds analysis and permits proper replication (Munafò et al., 2017). Based on a new analysis of IPSyn from Yang et al. (2021), KIDEVAL and IPSyn now require an input corpus of 50 acceptable child utterances, rather than the earlier requirement for 100 utterances. Moreover, IPSyn now runs with the reduced rule set recommended by Yang et al. which removes items that were found to reduce the predictive power of the test. However, it is still possible to run the classic version of IPSyn that requires 100 utterances with the original rule set.

C-DSS

The Developmental Sentence Scoring (DSS) profile method (Lee, 1974) examines many of the same features as IPSyn. Because it focuses more on morphological structures and lexical aspects of syntax, it is somewhat easier to compute automatically.

FluCalc

FluCalc provides analysis of raw and proportioned counts of disfluencies (e.g., prolongations, silent pauses, filled pauses, phonological fragments) marked in the transcript. This command was originally developed for use in studies of childhood stuttering (Bernstein-Ratner and MacWhinney, 2018), but can be applied to aphasia as well, given that fluency is central to aphasia diagnosis and treatment. Transcripts need to have specific markings in them to capture the behaviors such as prolongations, blocks, filled pauses, and unfilled pauses. The FluCalc command then provides an analysis of raw and proportioned counts of individual types of dysfluencies, average repetition unit frequency for word and part-word repetitions, and overall counts and proportions of dysfluencies. In addition to providing data on fluency behaviors in aphasia, FLUCALC could be used on transcripts from individuals with apraxia of speech, where speech may be slow and halting, with effortful groping, lengthened and repeated sound segments, and disturbed prosody (Peach, 2004). Automated analyses of larger shared datasets may contribute useful information to the differential diagnosis of these and related disorders.

Advantages of Automated Analyses

The advantages of automated analysis of the types described above cannot be overstated. They allow for faster analysis (in seconds) on one or as many transcripts as desired, less demand for training and expertise of coders and scorers, excellent replicability, and comparisons to existing databases. For researchers, the combination of large data sets and automated analyses has allowed for the application of multivariate and machine learning approaches to aphasia classification (Fraser et al., 2014; Day et al., 2021; Fromm et al., 2021). In the DementiaBank database, the Pitt corpus (Becker et al., 1994) has been used in hundreds of projects to create tools that automate the detection of dementia directly from audio files using various computational speech processing and machine learning methods (de la Fuente Garcia et al., 2020; Luz et al., 2021).

The combination of large, shared databases and automated analyses has allowed researchers to develop new tools and norms, examine psychometric properties of discourse measures, and answer some basic questions with robust, powerful statistics. For example, Richardson and Dalton (2016) created checklists of main concepts (MCs) from the five discourse tasks in the AphasiaBank discourse protocol by using the large set of control data. The checklists show the MCs used by 33, 50, and 60% of the respondents. Clinicians can use these checklists to get an objective measure of a PWA's ability to provide "essential content" on these tasks. Fergadiotis et al. (2013) were able to use Cinderella storytelling transcripts from 101 PWAs in the AphasiaBank database to examine the validity of four measures of lexical diversity and determine which ones yielded the strongest evidence for producing unbiased scores. Their findings led to a strong recommendation for using either the Moving-Average Type-Token Ratio (MATTR) or the Measure of Textual Lexical Diversity (MTLD) as the best measures of lexical diversity in aphasia. Stark (2019) explored differences in language produced in three different AphasiaBank standard discourse tasks in 90 PWAs and 84 controls. Results demonstrated that each discourse type tapped different aspects of language output in both groups. For example, propositional density was highest and speech rate was reduced in narrative discourse (Cinderella storytelling) compared with the expository (picture description) and procedural discourse tasks. These are just a few examples of the ways researchers have advanced the science of discourse in aphasia by taking advantage of these rich resources.

Finally, an overarching advantage of shared databases is also the greater transparency it affords for clinical and scientific endeavors. The media files, transcripts, and analyses are available for purposes of replication or testing alternative theories and analysis methods.

Phon

The PhonBank Project at https://phonbank.talkbank.org has developed the *Phon* program (Rose and MacWhinney, 2014) at https://phon.ca. *Phon* provides extensive support for the analysis of data on phonological development. Because *Phon* stores data in CHAT XML format, data that are transcribed in either *Phon* or CHAT are fully compatible and interchangeable. Here are some of its major features.

1. **Time alignment:** As with transcripts in CHAT format, transcripts in *Phon* can be aligned to the media at the utterance or word level for playback and analysis.

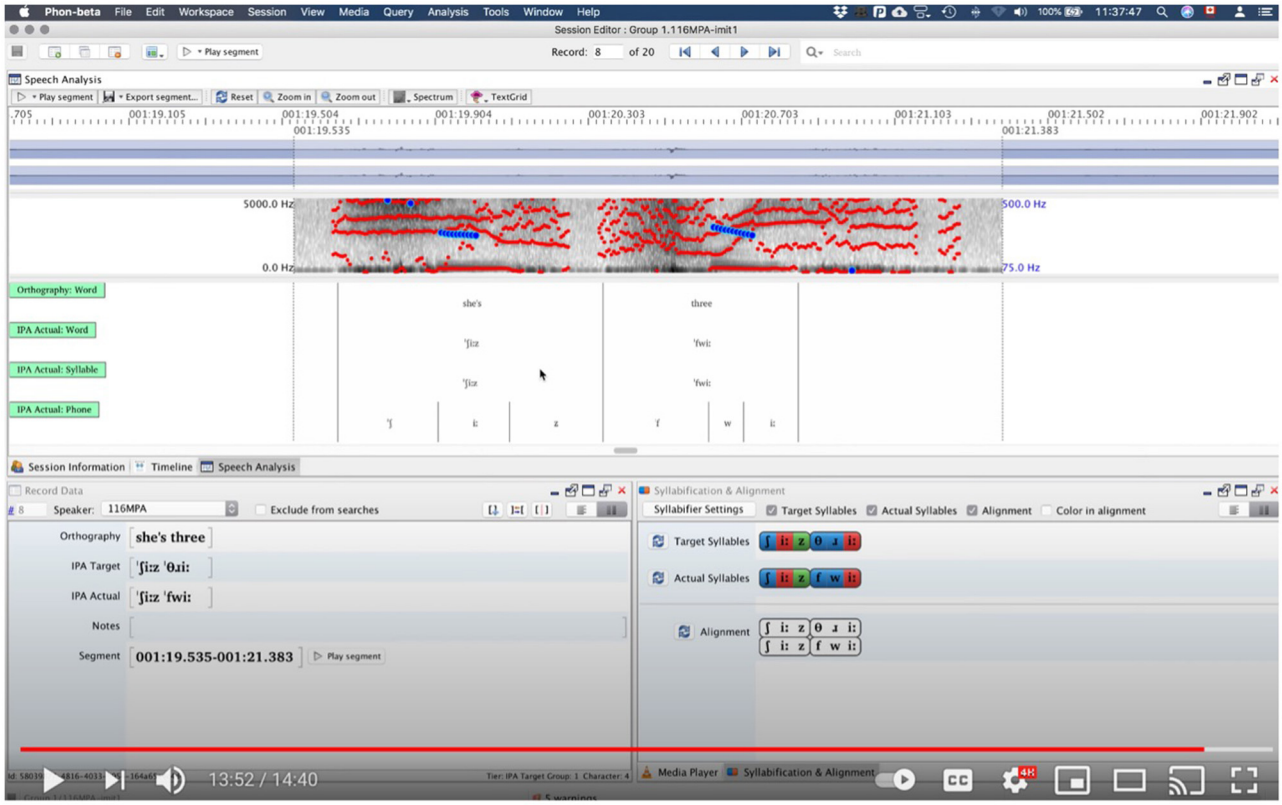

FIGURE 5 | *Phon* analysis of a child's production of "she's three".

2. ***Praat* integration:** *Praat* is the most widely used tool for acoustic and phonological analysis. All *Praat* functions are completely integrated into and available within *Phon*.
3. **Dictionaries and transliterators:** Analyses of phonological development often examine mismatches between a child's production and the form in the target language. To make this comparison, Phon provides tools for automatic insertion of the target phonology for 16 languages. In **Figure 5**, the IPA Target line is automatically inserted in this way.
4. **Syllabification:** *Phon* has syllabification algorithms for 23 languages, as well as a general algorithm for syllabifying babbling. The bottom right window in **Figure 5** displays a sample automatic syllabification with color coding for the onset, nucleus, and coda of syllables.
5. **Data query and reporting:** The query system relies on an easy-to-use mix of textual, regular, and phonological expressions for searching and reporting. The results of each query can be reported using different report formats customizable by the user.
6. **"Canned" analyses:** We provide packaged versions of all the common analyses used by clinical phonologists. We have also configured new analyses such as the Percentage of Tones Correct for tonal languages. We have also built into *Phon* a system to calculate inter-transcriber reliability, which we can assess for consonants and/or vowels using different settings of the PPC analysis in terms of Levenshtein distance.
7. **Analysis Composer:** This system enables users to combine their own sets of queries, reports and/or canned analyses into single custom packages. Besides the convenience it offers, this facility provides support for replication of published analyses.
8. **CHAT Interoperability:** *Phon* can directly open any file in the CHAT format. Once open, the file is in PhonXML format. Automatic IPA lookup then adds the target phonology (%mod) and temporary actual phonology (%pho) lines and the user can adjust the temporary actual phonology line to capture the correct phonological forms. The result can then be re-exported in CHAT format for inclusion in the other databanks and analysis by CLAN and TalkBanksDB.

Conversation Analysis (CA)

When it was introduced in the 1960s (Schegloff, 2007), Conversation Analysis (CA) relied on transcription through either pen and paper or typewriter. To mark special features such as overlaps, the typewritten transcript was marked up afterwards by hand. The introduction of Unicode in 1991 (https://www.unicode.org/versions/Unicode1.0.0/) made it possible to create symbols to represent all the features of CA, along with IPA and

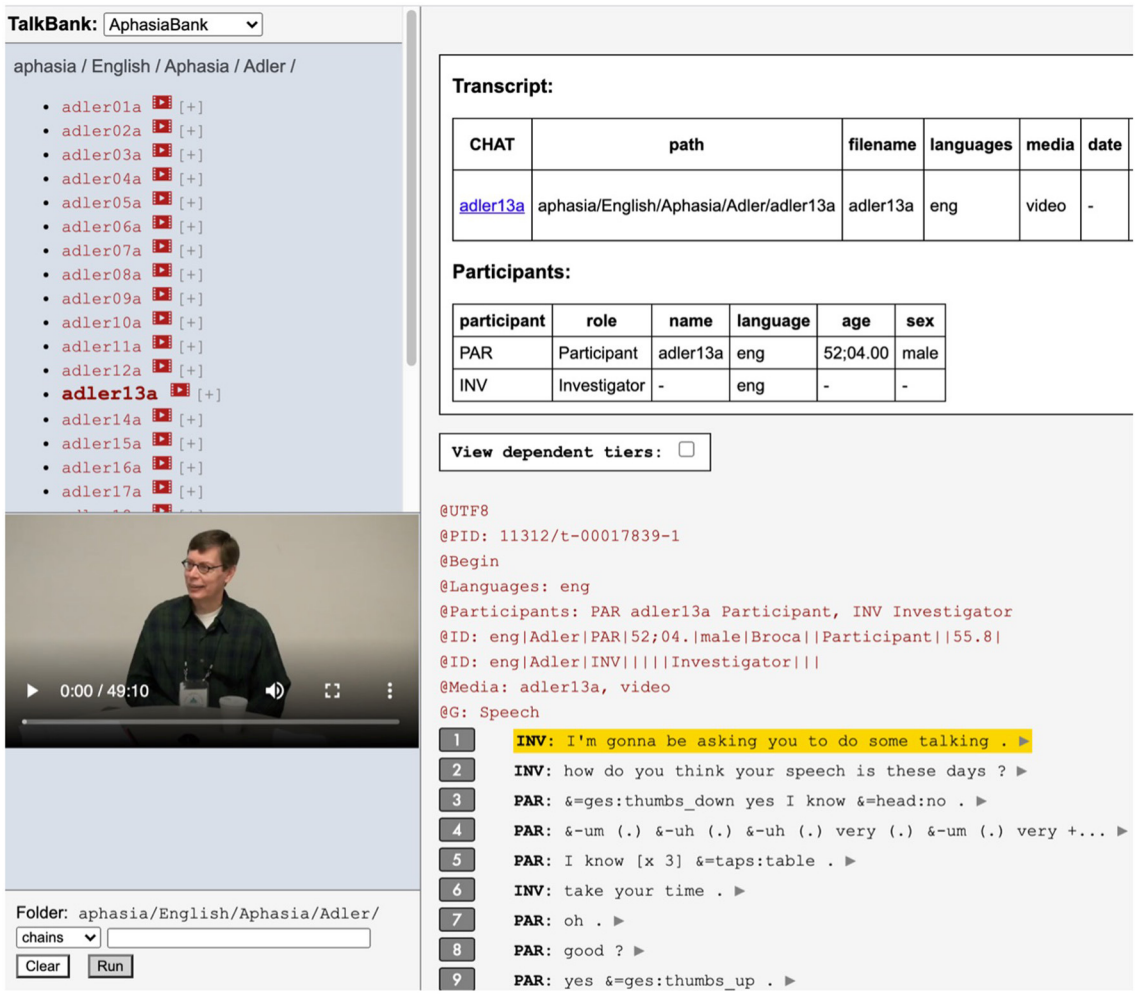

FIGURE 6 | TalkBank Browser screenshot.

the orthographies of all languages in the world. To further adapt the CLAN editor for CA, we introduced an option that allowed for utterances to be terminated through a series of prosodic shifts, rather than punctuation such as the period and question mark. We also introduced special forms of overlap marking brackets for the beginning and end of the initial segment and the following segment. The web page at https://ca.talkbank.org/codes.html illustrates 32 symbols for marking changes in volume, tempo, and pitch, along with markings for whispering, creaky voice, laughter, yawning, and other vocal characteristics. With all these features in place, CA transcriptions can be analyzed by all TalkBank facilities.

TalkBank Browser

The custom TalkBank Browser (https://sla.talkbank.org/TBB/aphasia) provides direct access to the entire collection of media files (video and audio) and transcripts. The directory that appears in the upper left corner of the screen allows users to select the language, the corpus, and the file of interest. **Figure 6** shows the Browsable Database screen. In the left corner, the directory shows this example is from: AphasiaBank, English language, aphasia group (as opposed to controls), Adler corpus, participant adler13a. From here, the video can be played by pressing the play arrow on the video screen or by pressing the play arrow at the end of any speaker tier in the transcript. As the video plays, a yellow highlighting line shows the transcript line that corresponds to what the speaker is saying.

The ability to browse these collections is beneficial for researchers, professors, and clinicians in a variety of ways. A professor teaching a course on diagnostics in aphasia may have students watch a selection of the hundreds of video administrations of confrontation naming tests (Boston Naming Test and Verb Naming Test) and practice scoring them according to the test scoring rules. Some corpora in the AphasiaBank Non-Protocol collection could be relevant for purposes such as: how to transcribe using conversation analysis markings, as done in the Goodwin corpus; or how to have informal conversations with PWAs, as done expertly by Dr. Audrey

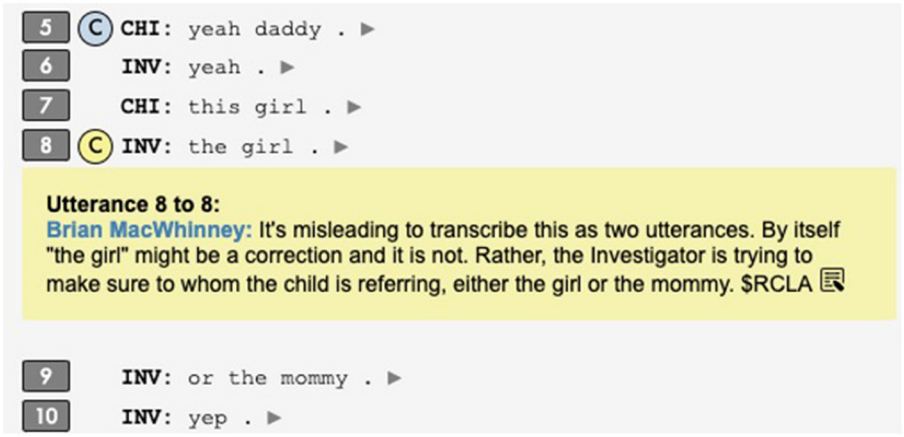

FIGURE 7 | A comment in the Collaborative Comment system.

Holland in the Tucson corpus. Researchers may use this tool to identify participants who meet specific selection criteria for their study, such as Broca's aphasia with and without apraxia of speech. Clinicians may use this tool to identify behaviors that facilitate a participant's successful self-corrections or communication. The Famous corpus has over 100 videos showing administration of the Famous People Protocol, which was designed specifically to identify any useful strategies people with severe aphasia can benefit from to communicate (Holland et al., 2019). This is a rich source of material to mine for students, clinicians, and researchers alike. Finally, clinic instructors may use this tool for student clinician training by finding examples of clinical styles to emulate in the administration of language tests and language sample collection.

Collaborative Commentary

Collaborative Commentary (CC) allows researchers, instructors, and clinicians to form commentary groups directed by a single supervisor but composed of multiple group members. Members can insert comments or codes directly into online transcript display with each comment or code being tagged to a specific utterance. **Figure 7** illustrates one comment that has been added in a child language transcript to note that the Investigator's utterance was incorrectly broken up and to explain why this was done. The utterance is tagged as $RCLA for "request for clarification" in accord with the INCA-R speech act coding system (Ninio and Wheeler, 1986). **Figure 8** illustrates the result of a search for all comments entered in a particular commentary group by a single user. This links to one comment in AphasiaBank and several in CHILDES. Clicking on these links takes the screen to the linked transcripts. It is also possible to search for comments with a given tag, such as $RCLA.

This is a new technology with many potential applications. For example, a clinic director may ask her clinical staff to watch the videos of aphasia group therapy sessions and identify (by marking directly in the transcript) behaviors that contribute to effective group management (the AphasiaBank shared database contains a large collection of aphasia group treatment videos from six different sites). A professor teaching a course on aphasia may give students a set of videos and transcripts representing different types and severities of aphasia and ask students to identify specific examples of behaviors such as word-finding difficulties, agrammatism, paragrammatism, phonemic paraphasias, semantic paraphasias, jargon, neologisms, perseverations, circumlocutions, empty speech, self-correction, conduite d'approche, and comprehension difficulties. A research team may use this to establish reliability for identifying and scoring measures of interest such as correct information units, main concepts, local coherence, global coherence, story grammar components, and gestures. All of these and many other applications of this technology will directly and positively impact the field and ultimately the quality of care provided to PWAs and their families.

TalkBankDB

To provide fuller and more direct access to the entire TalkBank database, we have developed a web-based postgreSQL system called TalkBankDB at https://talkbank.org/DB. TalkBankDB permits downloading of large segments of the database in seconds. The manual for this tool can be accessed by clicking on the manual icon in the upper right next to the Login button. **Figure 9** displays the results of a search for all the tokens (words) by English-speaking PWAs AphasiaBank. The result includes 926,626 words. Clicking on the Save button downloads this in 4 s in spreadsheet form to the desktop and it then takes another 5 s to open in Excel or 12 s to fully open in R.

TalkBankDB provides an intuitive on-line interface for researchers to explore TalkBank's media and transcripts, specify data to be extracted, and pass these data on to statistical programs for further analysis. It supports n-gram and CQL (Corpus Query Language) searches across all tiers in CHAT and allows for a variety of visualizations and analyses of data. Users can download data sets directly from Python or R. **Figure 10** illustrates a CQL

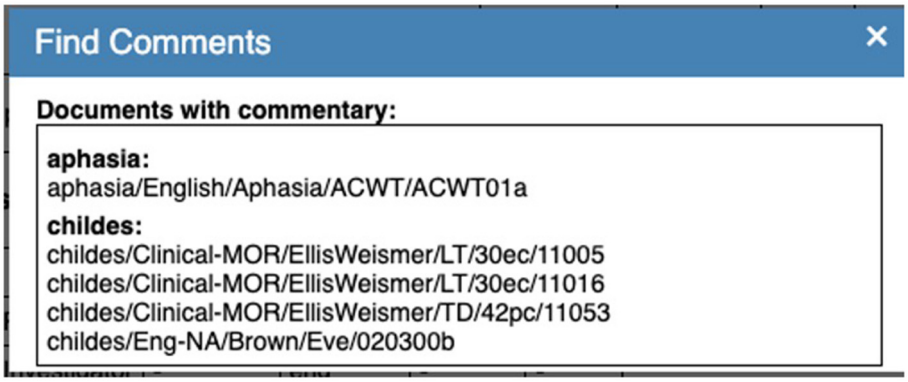

FIGURE 8 | Finding collaborative comments.

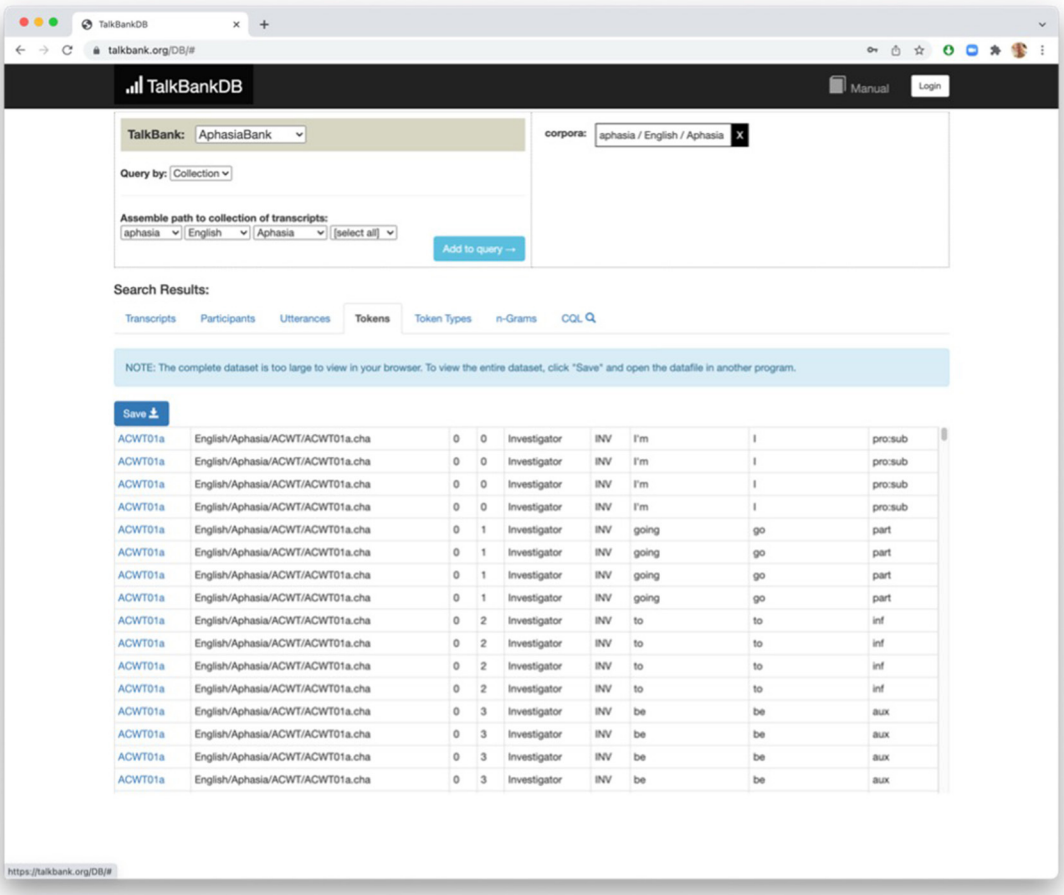

FIGURE 9 | A TalkBankDB search for tokens from English-speaking PWAs.

search for the word "my" followed by a noun from English-speaking PWAs.

Previously, browsing TalkBank's databases required knowing the name of a corpus or area of research, finding its location within the talkbank.org domain (e.g., aphasia.talkbank.org), then browsing/downloading the media and annotations and installing the CLAN tools. To make these resources more accessible, TalkBankDB provides a single online interface to query across all

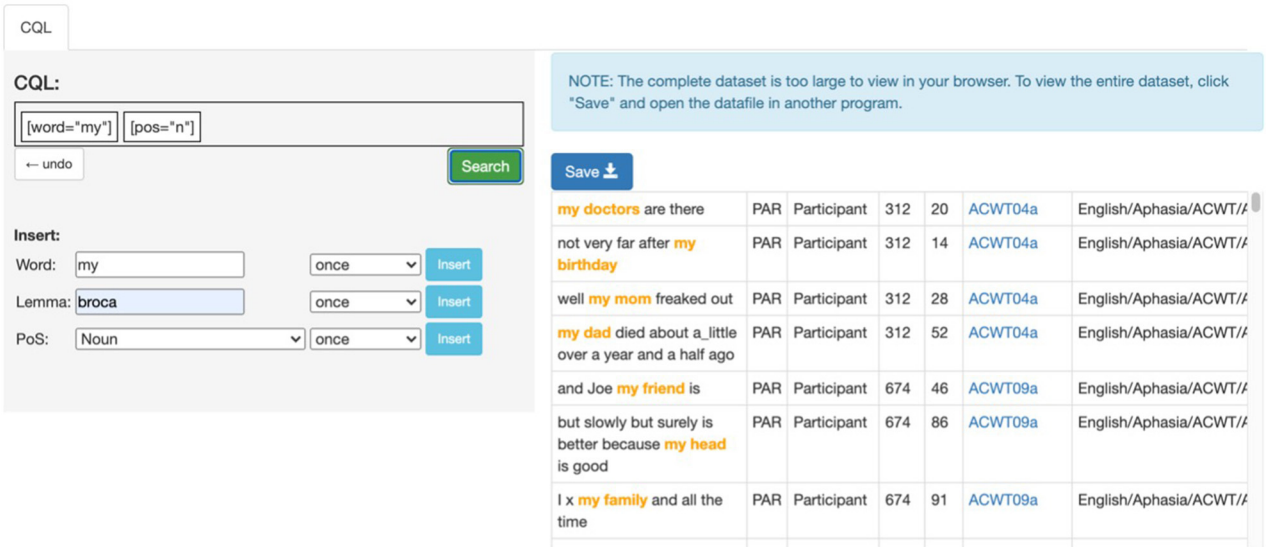

FIGURE 10 | A TalkBankDB CQL search.

the materials in TalkBank. Users can visually explore data directly in the browser, and if desired, download retrieved data sets for further analysis in a statistical software package.

With the entirety and richness of TalkBank freely accessible from a simple web interface, resources that were previously known only by advanced users are now open to a broader community. Features such as utterance length, lexical variables, morphological content, or error production by demographics or aphasia type can easily be selected, output, plotted, and analyzed through the web interface. By also providing a GitHub account link for users to upload scripts and analyses, the TalkBankDB site provides a single point where users can explore, share their research, and see what others are doing in the TalkBank community.

Learning Resources

Beginning users may find themselves overwhelmed by all the methods, data, and resources available in TalkBank. To help guide users toward the methods and data most relevant to their interests and to help them learn how to use the tools, we provide four types of learning resources.

1. **Grand Rounds.** For each of the clinical databases, we have carefully curated the collections to provide a set of Grand Rounds pages to familiarize students with the various presentations of the disorders. The traditional concept of Grand Rounds is to present a hands-on opportunity for medical professionals and students to improve their clinical knowledge of a disorder. The process involves hearing a clinical history and case presentation, doing an examination to assess relevant symptoms, and discussing ideas about diagnosis and treatment. The Grand Rounds for AphasiaBank at https://aphasia.talkbank.org/education/class/ is configured to echo this format. It includes case histories of individuals with different types and severities of aphasia, 40 captioned video clips of these individuals' discourse and performance on different tasks (e.g., confrontation naming, repetition), as well as clinically oriented questions to stimulate thought and discussion. Both TBIBank and RHDBank have Grand Rounds pages as well. The TBIBank Grand Round includes 25 video clips and provides material on characteristics of discourse impairments, discourse analyses to complement assessment, and treatment approaches that target "real-life" discourse level communication activities in adults with TBI. The modules begin with a pre-learning quiz that allows for measurement of new knowledge and skills. The RHDBank Grand Rounds (Minga et al., 2021) contains 13 video clips and material that highlight language production behaviors and cognitive-linguistic deficits associated with RHD. It, too, provides clinically oriented discussion questions as well as evidence-based literature on treatment of cognitive-linguistic deficits.
2. **Grand Rounds extensions.** Some instructors have shared their ideas for specific classroom activities that use the Grand Rounds materials (see "Classroom Activities" link in the AphasiaBank Teaching section). One assignment guides students in using the EVAL program to generate discourse data to compare Cat in Tree picture descriptions from three individuals with different types of aphasia: anomic, Broca's, conduction. Cross-disorder comparisons are the focus of another assignment, examining correct information units in language samples from RHD, aphasia, and control samples of the same picture description task. Several other assignments use specific case examples from the Grand Rounds, augmented by their test results (WAB, BNT, and VNT scores), and then poses questions about language

abilities, further assessment recommendations, and rationales for specific treatment approaches.

3. **Examples Page.** To further supplement the materials in Grand Rounds, AphasiaBank provides a page at https://aphasia.talkbank.org/education/examples/ linked to short video examples of common features from the connected speech of PWAs at the word-level (e.g., anomia, circumlocution, paraphasias) and at the sentence-level (agrammatism, empty speech). Two additional examples at the discourse level highlight how PWAs manage to communicate successfully despite having language filled with neologisms and jargon (the one with Wernicke's aphasia) and very limited language output (the one with Broca's aphasia). Further development of these types of examples of common behaviors can be useful for the other clinical language banks as well.
4. **Screencast Tutorials.** To guide learning about the database and tools themselves, we have constructed 48 screencast tutorials that usually last between 3 and 8 min. These are available both from our website and through YouTube. Topics covered include transcribing, linking transcripts to media files, running various commands, and more.
5. **Manuals.** We have produced detailed manuals for CHAT, CLAN, and MOR, along with a special manual for SLP practitioners and translations of the manuals into other languages. These materials are updated regularly as new tools are added to the program.
6. **Discussion Lists.** We maintain Google Groups mailing lists for aphasia, child language, bilingualism, and CA. These have proven very useful in a variety of ways, such as keeping users up to date on new features and new recording technologies, discussing IRB issues around new ASR technologies, answering questions from users about analysis command options, and receiving bug reports or requests for new features.

CONCLUSION

Construction of the TalkBank databases has benefitted from the commitment of participants and our colleagues to open data-sharing. Development of the programs and systems described here has benefited from advances in computer software and hardware, the hard work of our programmers, and support from NIH and NSF. These automated analyses provide many advantages that can improve the quality and quantity of information clinicians and researchers obtain from language samples. As a result, important strides are being made in understanding learning, recovery, disfluency, and problems in language disorders. All of the material covered here, though focused on aphasia and child language, can also be used with the other TalkBank clinical language banks to advance the work in those areas as well. We encourage a wide range of academic and clinical communities to contribute datasets to these shared databases and to make use of these tools to advance science.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: https://talkbank.org.

ETHICS STATEMENT

Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

This work was supported by NIDCD grant (R01-DC008524) for AphasiaBank and NICHD grant (R01-HD082736) for CHILDES.

ACKNOWLEDGMENTS

We are indebted to Audrey Holland for inspiration and direction in the construction of AphasiaBank and to Margie Forbes for her work on constructing the database and tools. We are similarly indebted to Catherine Snow and Nan Bernstein Ratner for their ongoing contributions and support for CHILDES. We also gratefully acknowledge the scores of other researchers, teachers, and clinicians who have contributed data to the shared database and the participants who have consented to sharing their data.

REFERENCES

Ball, M. J., Crystal, D., and Fletcher, P. (2012). *Assessing Grammar: The Languages of LARSP*. Clevedon: Multilingual Matters.

Becker, J. T., Boller, F., Lopez, O. L., Saxton, J., and McGonigle, K. L. (1994). The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Arch. Neurol.* 51, 585–594. doi: 10.1001/archneur.1994.00540180063015

Bernstein-Ratner, N., and MacWhinney, B. (2018). Fluency Bank: a new resource for fluency research and practice. *J. Fluency Disord.* 56, 69–80. doi: 10.1016/j.jfludis.2018.03.002

Boersma, P. (2001). *Praat: Doing Phonetics by Computer*. Available online at: https://www.praat.org

Boucher, J., Marcotte, K., Brisebois, A., Courson, M., Houzé, B., Desautels, A., et al. (2020). Word-finding in confrontation naming and picture descriptions produced by individuals with early post-stroke aphasia. *Clin. Neuropsychol.* 1–16. doi: 10.1080/13854046.2020.1817563

Brookshire, R. H., and Nicholas, L. E. (1994). Test-retest stability of measures of connected speech in aphasia. *Clin. Aphasiol.* 22, 119–133.

Brown, R. (1973). *A First Language: The Early Stages*. Cambridge, MA: Harvard.

Crystal, D. (1982). *Profiling Linguistic Disability*. London: Edward Arnold.

Dalton, S. G., Kim, H., Richardson, J., and Wright, H. H. (2020). A compendium of core lexicon checklists. *Semin. Speech Lang.* 41, 045–060. doi: 10.1055/s-0039-3400972

Dalton, S. G., and Richardson, J. (2015). Core-lexicon and main-concept production during picture sequence description in non-brain-damaged adults and adults with aphasia. *Am. J. Speech Lang. Pathol.* 24, S923–938. doi: 10.1044/2015_AJSLP-14-0161

Day, M., Dey, R. K., Baucum, M., Paek, E. J., Park, H., and Khojandi, A. (2021). "Predicting severity in people with aphasia: a natural language processing and

machine learning approach," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (IEEE)*.

de la Fuente Garcia, S., Ritchie, C., and Luz, S. (2020). Artificial intelligence, speech, and language processing approaches to monitoring Alzheimer's disease: a systematic review. *J. Alzheimer's Dis.* 1–27. doi: 10.3233/JAD-200888

Donoho, D. L. (2010). An invitation to reproducible computational research. *Biostatistics* 11, 385–388. doi: 10.1093/biostatistics/kxq028

Fergadiotis, G., Wright, H., and West, T. M. (2013). Measuring lexical diversity in narrative discourse of people with aphasia. *Am. J. Speech Lang. Pathol.* 22, 397–408. doi: 10.1044/1058-0360(2013/12-0083)

Forbes, M., Fromm, D., and MacWhinney, B. (2012). AphasiaBank: a resource for clinicians. *Semin. Speech Lang.* 33, 217–222. doi: 10.1055/s-0032-1320041

Fraser, K. C., Hirst, G., Meltzer, J. A., Mack, J. E., and Thompson, C. K. (2014). "Using statistical parsing to detect agrammatic aphasia," in *Proceedings of BioNLP 2014* (Baltimore, MD), 134–142.

Fromm, D., Greenhouse, J., Pudil, M., Shi, Y., and MacWhinney, B. (2021). Enhancing the classification of aphasia: a statistical analysis using connected speech. *Aphasiology*. doi: 10.1080/02687038.2021.1975636. [Epub ahead of print].

Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Kimbrough Oller, D., et al. (2017). Mapping the early language environment using all-day recordings and automated analysis. *Am. J. Speech Lang. Pathol.* 26, 248–265. doi: 10.1044/2016_AJSLP-15-0169

Hausser, R. (1999). *Foundations of Computational Linguistics: Man-Machine Communication in Natural Language*. Berlin: Springer.

Holland, A., Forbes, M., Fromm, D., and MacWhinney, B. (2019). Communicative strengths in severe aphasia: the famous people protocol and its value in planning treatment. *Am. J. Speech Lang. Pathol.* 28, 1010–1018. doi: 10.1044/2019_AJSLP-18-0283

Kim, H., Kintz, S., and Wright, H. H. (2019). Development of a measure of function word use in narrative discourse: core lexicon analysis in aphasia. *Int. J. Lang. Commun. Disord.* 56, 6–19. doi: 10.1111/1460-6984.12567

Kübler, S., McDonald, R., and Nivre, J. (2009). *Dependency Parsing*. San Rafael, CA: Morgan and Claypool.

Lee, L. (1966). Developmental sentence types: a method for comparing normal and deviant syntactic development. *J. Speech Hearing Disord.* 31, 331–330. doi: 10.1044/jshd.3104.311

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

Lin, D., Crabtree, J., Dillo, I., Downs, R. R., Edmunds, R., Giaretta, D., et al. (2020). The TRUST principles for digital repositories. *Sci. Data* 7, 1–5. doi: 10.1038/s41597-020-0486-7

Luz, S., Haider, F., de la Fuente Garcia, S., Fromm, D., and MacWhinney, B. (2021). Editorial: Alzheimer's dementia recognition through spontaneous speech. *Front. Aging Neurosci.* 3, 780169. doi: 10.3389/fcomp.2021.780169

MacWhinney, B. (2019). Understanding spoken language through TalkBank. *Behav. Res. Methods* 51, 1919–1927. doi: 10.3758/s13428-018-1174-9

MacWhinney, B., Fromm, D., Forbes, M., and Holland, A. (2011). AphasiaBank: methods for studying discourse. *Aphasiology* 25, 1286–1307. doi: 10.1080/02687038.2011.589893

MacWhinney, B., Fromm, D., Riebling, E., and Metze, F. (2017). *Automatic Speech Recognition of Scripted Productions*. PWAs Academy of Aphasia, Baltimore, MD. Available online at: https://psyling.talkbank.org/years/2017/MacW-Academy.pdf

MacWhinney, B., Fromm, D., Rose, Y., and Bernstein Ratner, N. (2018). Fostering human rights through TalkBank. *Int. J. Speech Lang. Pathol.* 20, 115–119. doi: 10.1080/17549507.2018.1392609

MacWhinney, B., Roberts, J., Altenberg, E., and Hunter, M. (2020). Improving automatic IPSyn coding. *Lang. Speech Hear. Serv. Sch.* 51, 1187–1189. doi: 10.1044/2020_LSHSS-20-00090

Minga, J., Johnson, M., Blake, M. L., Fromm, D., and MacWhinney, B. (2021). Making sense of right hemisphere discourse using RHDBank. *Top. Lang. Disord.* 41, 99–122. doi: 10.1097/TLD.0000000000000244

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., du Sert, N. P., et al. (2017). A manifesto for reproducible science. *Nat. Human Behav.* 1, 0021. doi: 10.1038/s41562-016-0021

Ninio, A., and Wheeler, P. (1986). A manual for classifying verbal communicative acts in mother-infant interaction. *Transcript Anal.* 3, 1–83.

Pavelko, S., and Owens, R. (2017). Sampling utterances abd grammatical analysis revised (SUGAR): New normative values for language sample analsysi measures. *Lang. Speech Hear. Serv. Sch.* 197–215. doi: 10.1044/2017_LSHSS-17-0022

Peach, R. K. (2004). Acquired apraxia of speech: features, accounts, and treatment. *Top. Stroke Rehabil.* 11, 49–58. doi: 10.1310/ATNK-DBE8-EHUQ-AA64

Richardson, J., and Dalton, S. G. (2016). Main concepts for three different discourse tasks in a large non-clinical sample. *Aphasiology* 30, 45–73. doi: 10.1080/02687038.2015.1057891

Rochon, E., Saffran, E., Berndt, R., and Schwartz, M. (2000). Quantitative analysis of aphasic sentence production: Further development and new data. *Brain Lang.* 72, 193–218. doi: 10.1006/brln.1999.2285

Rose, Y., and MacWhinney, B. (2014). "The PhonBank project: data and software-assisted methods for the study of phonology and phonological development," in *The Oxford Handbook of Corpus Phonology*, eds J. Durand, U. Gut, and G. Kristoffersen (Oxford: Oxford University Press), 380–401.

Saffran, E., Berndt, R., and Schwartz, M. (1989). The quantitative analysis of agrammatic production: procedure and data. *Brain Lang.* 37, 440–479. doi: 10.1016/0093-934X(89)90030-8

Scarborough, H. (1990). Index of productive syntax. *Appl. Psycholinguist.* 11, 1–22. doi: 10.1017/S0142716400008262

Schegloff, E. A. (2007). *Sequence Organization in Interaction: A Primer in Conversation Analysis I (Vol. 1)*. Cambridge: Cambridge University Press.

Stark, B. (2019). A comparison of three discourse elicitation methods in aphasia and age-matched adults: implications for language assessment and outcome. *Am. J. Speech Lang. Pathol.* 28, 1067–1083. doi: 10.1044/2019_AJSLP-18-0265

Szabo, G., Fromm, D., Heimlich, T., and Holland, A. (2014). *Script Training and Its Application to Everyday Life in an Aphasia Center Clinical Aphasiology Conference*. St. Simons Island, GA.

Thompson, C. K., Shapiro, L. P., Li, L., and Schendel, L. (1995a). Analysis of verbs and verb-argument structure: a method for quantification of aphasic language production. *Clin. Aphasiol.* 23, 121–140.

Thompson, C. K., Shapiro, L. P., Tait, M. E., Jacobs, B. J., Schneider, S. L., and Ballard, K. J. (1995b). A system for the linguistic analysis of agrammatic language production. *Brain Lang.* 51, 124–129.

Tucci, A., Plante, E., Heilmann, J. J., and Miller, J. F. (2021). Dynamic norming for systematic analysis of language transcripts. *J. Speech Lang. Hear. Res.* 1–14. doi: 10.1044/2021_JSLHR-21-00227

Wiederholt, J., and Bryant, B. (2012). *Gray Oral Reading Test-Fifth Edition (GORT-5): Examiner's Manual*. Austin, TX: Pro-Ed.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3, 1–9. doi: 10.1038/sdata.2016.18

Wilson, S. M., Eriksson, D. K., Schneck, S. M., and Lucanie, J. M. (2018). A quick aphasia battery for efficient, reliable, and multidimensional assessment of language function. *PLoS ONE* 13, e0192773. doi: 10.1371/journal.pone.0199469

Yang, J. S., MacWhinney, B., and Ratner, N. B. (2021). The index of productive syntax: psychometric properties and suggested modifications. *Am. J. Speech Lang. Pathol.* 1–18. doi: 10.1044/2021_AJSLP-21-00084

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.