Printed by: fromm@andrew.cmu.edu. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.

11
Clinical Corpus Linguistics

DAVIDA FROMM AND BRIAN MACWHINNEY

11.1 Preamble

Major advances in computer power and new technologies in machine learning have made it possible to study large corpora of language with increased efficiency and reliability. Manual transcription, coding, and analyses of large datasets require time and training that exceed the capacity of most research programs. Even the collection of the datasets themselves can be challenging, as they need to be large enough and representative enough across a variety of domains (e.g., type and severity of a particular disorder, demographic diversity) to conduct robust, powerful studies. The TalkBank project seeks to address these issues and to take advantage of these new opportunities.

TalkBank (https://talkbank.org) arose as an extension of the already existing child language acquisition data system called CHILDES, established by Catherine Snow and Brian MacWhinney in 1984. In 2000, we began the construction of a more general, distributed, web-based, data archiving system for transcribed video and audio data on communicative interactions in multiple forms. Since 2000, TalkBank has expanded to include 15 language banks, 7 of which focus on clinical data: AphasiaBank, ASDBank, DementiaBank, FluencyBank, PsychosisBank, RHDBank, and TBIBank. Additionally, the CHILDES and PhonBank databases have data from clinical populations that augment their corpora for the study of normal language and phonological development. This chapter will describe these clinical corpora and their impact on the study of language in a range of fields. First, we review the TalkBank principles which have been essential to its widespread adoption for advancing the study of spoken language.

DO NOT COPY fromm@andrew.cmu.edu

11.2 TalkBank Principles

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

1. *Maximally open data-sharing*: The social sciences have been slower to adopt data-sharing process than the physical sciences, mainly out of concerns regarding participant privacy. This has limited the scientific advancement facilitated by large datasets. To address this, TalkBank requires that members (licensed clinicians and researchers) agree to abide by data sharing ground rules and a code of ethics (https://talkbank.org/share). The system also provides options that preserve participant anonymity (e.g., de-identification, audio bleeping, password protection, controlled viewing).
2. *CHAT transcription format*: TalkBank uses a uniform transcription standard, called CHAT, that allows one to encode any of the features that play a role in spoken language. These features and codes are documented in the CHAT manual which can be downloaded from https://talkbank.org/manuals/chat.pdf. Although the system is quite extensive, individual projects usually only need to use specific subsections of the full format. CHAT allows for words and utterances to be linked to corresponding segments in the media files, thereby facilitating transcription, coding, and temporal analyses. Because it uses UTF-8 Unicode, all languages can be transcribed directly in CHAT files.
3. *CHAT-compatible software*: CHAT files are compatible with the CLAN program (http://dali.talkbank.org/clan) which allows for automatic parsing of the transcript and many automatic analyses of language and discourse for syntax, morphology, phonology, lexicon, timing, and pragmatics.
4. *Interoperability*: To archive data coming from non-CHAT formats and to provide users with options for using other programs with CHAT data, CLAN includes 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.
5. *Responsivity to research community needs*: TalkBank seeks to be maximally responsive to the needs of individual researchers and their research communities. We attempt to implement all features that are suggested by users in terms of software features, data coverage, documentation, and user support. This is accomplished in multiple ways, such as through Google Groups mailing lists, YouTube screencast tutorials, construction of web pages for each corpus, index pages for databanks, manuals for CHAT and CLAN, article publications, conference presentations and workshops, adding new computational resources, and regular updating of programs

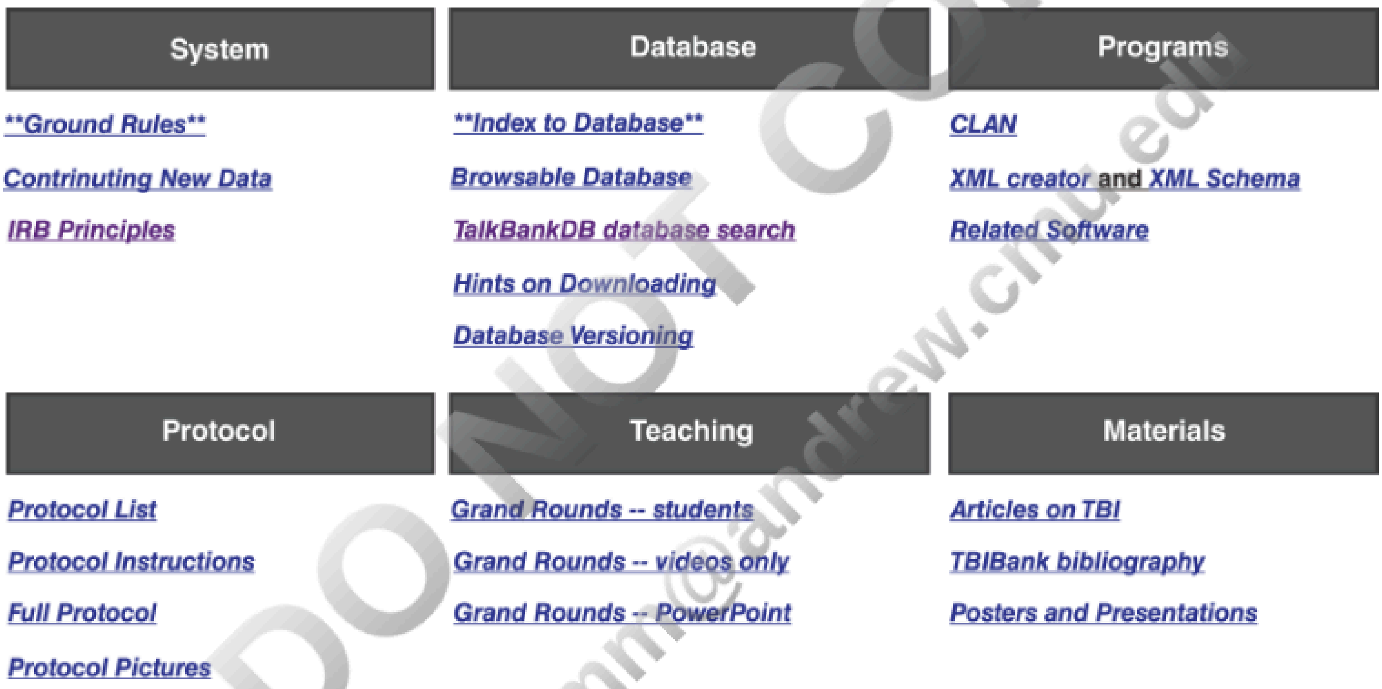
Printed by: fromm@andrew.cmu.edu. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.

conference presentations and workshops, adding new computational resources, and regular updating of programs and materials. The TalkBank Governing Board provides overall guidance for the project.

6. *International standards*: TalkBank adheres to international standards for database and language technology. In particular, the system adheres to (1) the FAIR standards (Wilkinson et al., 2016) for open access to data, which hold that data should be Findable, Accessible, Interoperable, and Reusable, and (2) the TRUST standards (Lin et al., 2020) for maintenance of reliable digital databases. The icons seen at the bottom of the TalkBank home page (CLARIN, Core Trust Seal, etc.) demonstrate our commitment to meeting state-of-the-art standards set by the international community of scientists interested in computational linguistics and corpora.

11.3 Clinical Corpora

All seven clinical databanks are listed and accessible from the main TalkBank webpage at https://talkbank.org. Clicking on a clinical bank's hyperlink takes one to the webpage for that clinical corpus, as in Figure 11.1 for the TBIBank homepage. These corpus webpages are organized with section headings such as *System*, *Database*, *Programs*, *Protocols*, *Teaching*, and *Manuals*. In the *Database* section, there is an 'Index to Corpora' link that provides a directory of what is contained in the collection. Clicking on that link opens a page with a table describing features of the individual corpora contained in the database. From that table, clicking on the hyperlink for any individual corpus opens a page with information about the contributors, publications, and the corpus. The corpus page also has links to view the Browsable database (explained more later) and to download the transcripts and media for that corpus. Most of the clinical corpora are password protected. However, access is given readily and quickly to verified professionals who send an email request with their contact information and affiliation to macw@cmu.edu.



System

Ground Rules

Contrinuting New Data

IRB Principles

Database

Index to Database

Browsable Database

TalkBankDB database search

Hints on Downloading

Database Versioning

Programs

CLAN

XML creator and XML Schema

Related Software

Protocol

Protocol List

Protocol Instructions

Full Protocol

Protocol Pictures

Teaching

Grand Rounds -- students

Grand Rounds -- videos only

Grand Rounds -- PowerPoint

Materials

Articles on TBI

TBIBank bibliography

Posters and Presentations

Figure 11.1 **TBIBank homepage** © Davida Fromm & Brian MacWhinney 2023.

11.3.1 AphasiaBank

Aphasia results from damage to the language areas of the brain (usually the left hemisphere) and may impair expression, comprehension, reading, and writing. AphasiaBank (https://aphasia.talkbank.org) started in 2005 with a planning meeting of 20 experienced aphasia researchers who agreed on the need for a shared protocol, a shared database, and increased availability of computational tools for the aphasia research community (MacWhinney et al., 2011). With NIH funding beginning in 2007, the project team developed a standard discourse protocol and test battery and began collecting data from university clinics and aphasia centers around the United States and Canada.

For English, there are currently 467 transcripts from persons with aphasia (PWAs) and 286 from age-matched controls. Croatian has 53 PWAs; French has 13 PWAs and 14 controls; German has 4 PWAs; Italian has 10 PWAs; Mandarin has 11 PWAs and 31 controls; Romanian has one PWA; and Spanish has 4 PWAs. In addition to these transcripts that use the standard one-hour discourse protocol, there are four collections of non-protocol data including 658 with various types of discourse data, 377 files with script training samples before and after treatment, 207 recordings of aphasia group therapy sessions, and 99 recordings using the Famous People Protocol (Holland et al.,

DO NOT COPY fromm@andrew.cmu.edu

Printed by: fromm@andrew.cmu.edu. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.

recordings of aphasia group therapy sessions, and 33 recordings using the Famous People Protocol (Holland et al., 2019) for PWAs with severe aphasia. Non-protocol PWA data in other languages include 26 for German, 53 for Greek, 64 for Hungarian, 8 for Italian, 15 for Mandarin, and 32 for Spanish. All the transcripts are in CHAT, and most have been linked to either audio or video on the utterance level and coded for discourse features.

The standard discourse protocol includes personal narratives, picture descriptions, storytelling (Cinderella), and a procedural task. Detailed administration instructions and a script for the investigator were developed to ensure consistent implementation across sites. The protocol is video-recorded and can be administered in person or via the internet. Demographic data are collected, and a variety of formal and informal tests are administered as well. These materials (e.g., protocol, instructions, script) are all accessible from the main webpage. The English-speaking protocol database includes participants from 37 different sites, providing good geographic diversity.

The non-protocol aphasia discourse collection contains over 20 corpora contributed by researchers who collected language samples that were specific to their purposes. Some examples that illustrate the range of materials in this collection are: (1) the Fridriksson corpus, which contains WAB-R picture descriptions from 19 PWAs before and after speech entrainment treatment (Fridriksson et al., 2012); (2) the Pawleys corpus, which is a longitudinal set of 48 short (5–10 minute) conversations with a 68-year old man with fluent, Wernicke's type aphasia, during the first 2–6 months following an ischemic stroke; (3) the SouthAL corpus (Smith & Clark, 2019), which contains transcripts and media files for 9 PWAs and 8 controls doing an oral reading test; and (4) the Olness corpus (Olness et al., 2002), 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. Along with the large numbers of group treatment videos from six different sites, script training samples from two sites, and administrations of the Famous People Protocol from 13 sites, these materials greatly increase the ways in which this clinical database can impact clinical research and teaching opportunities.

The overarching goal of AphasiaBank is the construction of methods for improving clinical management in aphasia. The database has facilitated that primarily through its extensive research and teaching resources. In terms of research, hundreds of publications, presentations, and theses have made use of the materials. (A full bibliography of relevant articles and conference presentations is available at the AphasiaBank webpage.) Examples of the research projects include: creating clinician-friendly discourse assessment tools with norms (Dalton et al., 2020; Richardson & Dalton, 2020; Richardson et al., 2021); developing automated analysis programs for profiling language output specific to aphasia and creating benchmarks for comparison (Forbes et al., 2012); automating established grammatical analysis systems (Fromm, Katta et al., 2021; Fromm et al., 2020); examining spontaneous speech predictors of fluency measures (Clough & Gordon, 2020; Gordon & Clough, 2020); comparing measures for lexical diversity (Fergadiotis et al., 2013); evaluating psychometric properties of language outcome measures (Boyle, 2015; Kim & Wright, 2020) and automated error analysis (Smith et al., 2019); formalizing an auditory-perceptual rating scale for connected speech (Casilio et al., 2019); examining microlinguistic aspects of language across discourse genres (Stark, 2019); investigating how gesture conveys information that is essential to understanding communication (Sekine et al., 2013; van Nispen et al., 2017); comparing manual and automated analysis of connected speech (Hsu & Thompson, 2018); training, testing, and evaluating technologies for automatic speech recognition (ASR) systems (Le et al., 2018; Perez et al., 2020); and using machine learning approaches to enhance classifications of aphasia based on spontaneous speech output (Fromm, Greenhouse et al., 2021).

Hundreds of university programs around the world use the AphasiaBank *Grand Rounds* guided tutorial to teach about aphasia and to expose students to a wider range of people with aphasia than they might otherwise encounter. The *Grand Rounds* pages include case histories for 16 individuals, 40 captioned video clips of their discourse and performance on different tasks (e.g., confrontation naming, repetition), and clinically oriented questions to stimulate thought and discussion. The cases were carefully curated from the larger collection in the database to make it easier for instructors to find illustrative examples of individuals with different types and severities of aphasia to augment their course material. Another resource, the *Examples* page, zeroes in on the connected speech of PWAs with definitions and short video clips of common features such as paraphasias, anomia, agrammatism, and circumlocution. Finally, a *Classroom Activities* page includes suggestions for assignments that make use of the AphasiaBank corpus as well as others (e.g., RHDBank) for cross-disorder comparisons. With good resources for academic and clinical instructors, students will be better prepared to provide effective diagnostic and treatment services to patients. Currently, AphasiaBank has over 1,250 members from more than 55 countries making use of the educational, clinical, and research materials.

11.3.2 TBIBank

TBIBank is a repository for multimedia interactions for the study of communication in people with traumatic brain injury (TBI). TBI can result in cognitive-communication disorders that may affect all aspects of language (e.g., speaking, listening, reading, writing, pragmatics) as well as attention, reasoning, memory, and executive function.

DO NOT COPY
fromm@andrew.cmu.edu

Printed by: fromm@andrew.cmu.edu. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.

speaking, listening, reading, writing, pragmatics) as well as attention, reasoning, memory, and executive function. Discourse in TBI has been described as disorganized, inappropriate, tangential, unclear, redundant, and self-focused. Like AphasiaBank, TBIBank includes media files and transcripts from a standard discourse protocol and test battery. The discourse protocol has several tasks that overlap with the AphasiaBank protocol (e.g., the Cinderella story narrative), making it possible to conduct cross-disorder comparisons. This Togher-Protocol corpus is from Australia and has the added advantage of containing longitudinal discourse samples from 58 participants that allow for the study of recovery during the first two years, post-onset. The corpus contains a total of 237 transcript and media files. Several other sizeable and valuable non-protocol corpora have been contributed to this databank, yielding a total of over 800 additional files. One corpus has 55 participants with closed head injuries and 52 controls doing a variety of discourse tasks, such as story retell, story generation, and informal conversation (Coelho et al., 2003). No other languages besides English are included in this databank yet.

Currently, TBIBank has over 200 members from around the world. This clinical corpus has been used in several dozen published reports, conference presentations, and graduate theses. Stubbs et al. (2018) showed that the simple procedural discourse task in the standard discourse protocol was sensitive to qualitative changes between 3- and 6-months post-injury, with significant increases in use of relevant information (macrostructure). The task also distinguished the TBI group from a control group at both time points based on speech rate and two macrostructural categories (essential and optional steps). Power et al. (2019) also reported differences on macrostructural and superstructural measures but not microstructural measures from a single picture description task done by participants with TBI controls. In a final example, Norman et al. (2022) recently selected adults with moderate to severe TBI from the TBIBank database to for their study comparing discourse-level language performance in adults with three other groups: mild TBI, healthy adults, and orthopedic controls. Studies like these require standard discourse protocols and large databases to help identify which measures are sensitive to group differences and, thereby, inform effective assessment and treatment planning for this clinical population.

Like AphasiaBank, TBIBank has a *Grand Rounds* tutorial that includes case histories and 25 captioned video clips. The tutorial addresses the range of spoken cognitive-communication disorders that can result from TBI, discourse analyses to complement assessment, treatment approaches that target real-life discourse level communication activities, comorbidities, and recovery. It is designed as an online learning module and begins with a pre-learning quiz that allows for measurement of newly acquired knowledge and skills.

11.3.3 RHDBank

RHDBank (https://rhd.talkbank.org) was created for the study of communication in adults with right hemisphere disorder (RHD) resulting from damage to the right hemisphere (Minga et al., 2021). Symptoms of RHD include cognitive-communication deficits that impair pragmatic skills, resulting in difficulties producing and comprehending discourse. Deficits commonly seen in people with RHD include difficulty with topic maintenance, discourse coherence and cohesion, inference generation, turn-taking, question use, and the integration of contextual nuance. Individuals with RHD have typically been underserved clinically because their symptoms tend to be more subtle than those with aphasia resulting from left hemisphere stroke. However, the consequences of these deficits can negatively impact quality of life in many ways, such as successful return to work and social relationships with family and friends. A paucity of research and clinical resources also contributes to gaps in service to this population.

As with the main AphasiaBank corpus, RHDBank contains corpora that use a standard discourse protocol, demographic data collection, and set of assessment procedures. The materials were chosen to have some overlap with those in the other clinical banks to allow for cross-disorder comparisons. In addition to the tasks in the AphasiaBank protocol, the RHD discourse protocol includes a first-encounter conversation (Kennedy et al., 1994). This task provides an opportunity to assess behaviors such as turn taking, question use, and pragmatics. The test battery includes assessments for visuospatial neglect and cognitive-linguistic functioning. To date, the protocol database has media and transcripts from 23 adults with RHD and 23 controls, but data collection is ongoing. A few non-protocol corpora are also available in this databank. The Hopkins corpus includes Cookie Theft picture descriptions from 42 participants who were seen acutely following right hemisphere strokes and then followed at various time intervals thereafter. A Spanish corpus contains discourse transcripts from 11 individuals with RHD. All materials are available from the webpage.

Currently, RHDBank has over 150 members who requested access for research projects, clinical training, and educational applications. A *Grand Rounds* tutorial contains 13 video clips and material that highlight language production behaviors and cognitive-linguistic deficits associated with RHD. It includes clinically oriented discussion questions as well as evidence-based literature on treatment of cognitive-linguistic deficits. Research studies have reported findings such as differences in the types of questions used by participants with RHD compared with controls (Minga et al., 2020; 2022), the utility of the procedural discourse task for clinical evaluation of individuals with RHD (Cummings, 2019), and differences in macrostructural measures (e.g., main concepts and global coherence) across

DO NOT COPY
fromm@andrew.cmu.edu

Printed by: fromm@andrew.cmu.edu. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.

(Cummings, 2019), and differences in macrostructural measures (e.g., main concepts and global coherence) across various tasks when compared with controls and PWA (Johnson et al., 2019). A bibliography of publications and links to conference presentations are available at the webpage. The goal is to fill the knowledge gaps and provide more exposure and training in this area to increase the likelihood that clinicians will have better tools and experience to assess and treat this population.

11.3.4 DementiaBank

DementiaBank (https://dementia.talkbank.org) includes transcripts and media from individuals with various types of dementia as well as individuals with primary progressive aphasia (PPA). Dementia has many potential causes and presentations, but usually involves gradually worsening impairments in memory, communication, reasoning, and orientation. Language symptoms depend largely on the type and severity of dementia. Language production deficits generally include word-finding problems, empty speech, paraphasias, circumlocution, perseveration, and reduced output. A corpus currently being collected by Alyssa Lanzi at the University of Delaware is using a standard discourse protocol to collect data from individuals with neurotypical and mild cognitive impairment and dementia. Again, this protocol has some overlaps with the other clinical databank protocols as well as some unique tasks and assessments designed for this clinical population.

Most of the corpora in this databank were contributed from other larger studies conducted years ago. One corpus in this repository, the Pitt Corpus (Becker et al., 1994) contains longitudinal data for four language tasks (Cookie Theft picture descriptions, a sentence construction task, word fluency tasks, and a story retell task) from hundreds of individuals with Alzheimer's disease (AD) and other types of dementia as well as elderly controls. Another large corpus, WLS (Herd et al., 2014), includes a subset of tasks from the Wisconsin Longitudinal Study, which is a long-term study of a random sample of over 10,000 high school graduates from 1957. The WLS corpus in the DementiaBank collection contains 1,369 audio and transcript files of Cookie Theft picture description from the 2011 test sessions, with additional data on demographics and a variety of related test scores (e.g., letter and category word fluency). Most of these participants would be considered healthy controls. Other corpora in DementiaBank include conversations and other language tasks from individuals with AD. Corpora have also been contributed from German, Mandarin, Spanish, and Taiwanese.

Currently, DementiaBank has over 600 members from all over the world. The large datasets in this clinical corpus have been of particular interest to researchers who are using a variety of ASR, machine learning, and language processing techniques to automatically identify AD from short narrative samples (de la Fuente Garcia et al., 2020). Several computational challenges have been hosted, for example at InterSpeech conferences (e.g., ADReSS Challenge), where teams from across the world test methods for detecting Alzheimer's disease and predicting cognitive decline based on spontaneous speech samples. Carefully curated sets of data are made available to the participating teams for comparison of results. The goal is to develop the most effective clinical applications of these techniques for early diagnosis and implementation of devices to promote patient health and safety. To date, the best classification accuracy (without transcripts or human intervention) is around 78% and the best cognitive score prognosis accuracy is around 66% (Luz et al., 2021). Accuracy is improved if text files are used with the audio signals. This important line of work relies on large, high-quality datasets, which are unfortunately limited in supply. A bibliography of all articles that use DementiaBank corpora is available from a link at the webpage.

11.3.5 FluencyBank

FluencyBank (https://fluency.talkbank.org), organized by Nan Bernstein Ratner at the University of Maryland, is one of the newer databases, established in 2016 to address the need for a shared open-access database that could allow for a broader and deeper understanding of the development of fluency and disfluency (stuttering) in normal and atypical speech (Bernstein Ratner & MacWhinney, 2018). Stuttering is a significantly disabling, lifelong communication disorder with severe psychological, educational, social, economic, and vocational impacts (Gerlach et al., 2018). Fluency of speech production is involved in many stages and processes of language encoding (Ferreira, 2007). When disfluency exceeds listener expectations in frequency and/or quality, it is often perceived as stuttering, a disorder producing substantial personal handicap, with academic, vocational, and social impacts spanning the lifetime (Tichenor & Yaruss, 2020). Disfluency is also involved in other expressive communication disorders.

The FluencyBank database currently contains 13 corpora in English, and one each in Dutch, French, and German. The English samples include almost 500 audio or video files of children and adults who stutter; some of the corpora include control participants as well. Teaching resources include a large selection of videos of adults and children who stutter as well as suggested classroom activities.

To study disfluency patterns in detail, we have introduced into CHAT a series of special codes and Unicode symbols. These mark features such as stalls, pauses, filled pauses, prolongations, broken words, blocking, repeated segments,

DO NOT COPY
fromm@andrew.cmu.edu

Printed by: fromm@andrew.cmu.edu. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.

These main features such as stalls, pauses, filled pauses, prolongations, broken words, blocking, repeated segments, lengthened repeated segments, phonological fragments, and various types of word and phrase repetition. These features can then be analyzed with CLAN’s FLUCALC program, which resembles the KIDEVAL program in many regards. One feature of FLUCALC is its ability to compute a disfluency index based on syllables, rather than words, as developed by the Illinois Stuttering Project (Yairi & Ambrose, 1999).

The first phase of construction of FluencyBank focused on the emergence of speech fluency over childhood in both typically developing preschoolers and children who stutter (CWS). This work has focused on ways of understanding what features can predict recovery from early childhood stuttering as opposed to persistent stuttering. Various factors have been proposed as potential predictors of recovery from childhood onset stuttering. Among these are negative family history of stuttering, female sex, earlier age of stuttering onset, better language/phonological skills, and genetic and brain indices. Interestingly, the initial speech fluency profile does not predict outcomes. We are now exploring ways in which neuroanatomical measures (Chang et al., 2019) collected in combination with language samples can further illuminate this issue.

11.3.6 ASDBank

ASDBank (https://asd.talkbank.org) includes data on language development from children and adolescents with autism spectrum disorder (ASD). It uses methods and analyses that are like those used in the larger CHILDES database. This is one of the smaller collections, and unfortunately it does not include media files for over half of the corpora. The most complete corpus in this collection contains Dutch language productions from 46 children with ASD. These data are from a project that investigates asymmetries between production and comprehension in unimpaired children, in young and elderly adults, and in autistic and ADHD children and adolescents (Kuijper et al., 2015). We hope to expand this databank in future work.

11.3.7 CHILDES

The oldest and most widely used component of TalkBank is CHILDES (Child Language Data Exchange System, https://childes.talkbank.org) which began in 1984 and has now been used in over 8000 publications. Although the bulk of work with CHILDES has focused on the analysis of normal language development, there are also 27 corpora from children with various developmental disorders, such as late talking, SLI, Downs Syndrome, and epilepsy. These are included in CHILDES, although similar data from children who stutter are in FluencyBank and data from children with ASD are in ASDBank.

11.3.8 PhonBank

The PhonBank project (https://phonbank.talkbank.org) is led by Yvan Rose at Memorial University, Newfoundland. PhonBank work has focused both on the creation of a database of phonologically transcribed productions from young children and the development of a program called Phon designed to analyze these data. In accord with the emphasis in TalkBank on interoperability, it is possible to open CHAT files directly in Phon, although the level of phonological coding and analysis in Phon is much deeper than that found in CHILDES corpora. Phon provides a wide range of analysis options, including automatic syllabification, automatic model phonology insertion, full analysis using Praat, dozens of standard measures, pre-configured reports, and user-configurable report formats.

The languages represented in PhonBank corpora include Arabic, Berber, Cree, Cantonese, Catalan, Dutch, English, German, Greek, Icelandic, Japanese, Mandarin, Norwegian, Polish, Portuguese, Quechua, Romanian, Spanish, Swedish, and Turkish. There are also phonologically transcribed corpora from bilingual children and second language learners. During the first phase of constructing PhonBank, the emphasis was on data from normally developing children. However, these data have now been supplemented with corpora from children with phonological disorders in English, French, Portuguese, and Spanish.

11.3.9 PsychosisBank

The most recent addition to TalkBank’s clinical banks is PsychosisBank. This bank focuses on language in psychosis in collaboration with the international Discourse in Psychosis consortium at https://discourseinpsychosis.org, led by Lena Palaniyappan at Western University. This group has formulated a standard spoken language elicitation protocol based on the AphasiaBank protocol with extensions to psychosis. Projects using this new protocol are now underway.

11.4 Other Tools

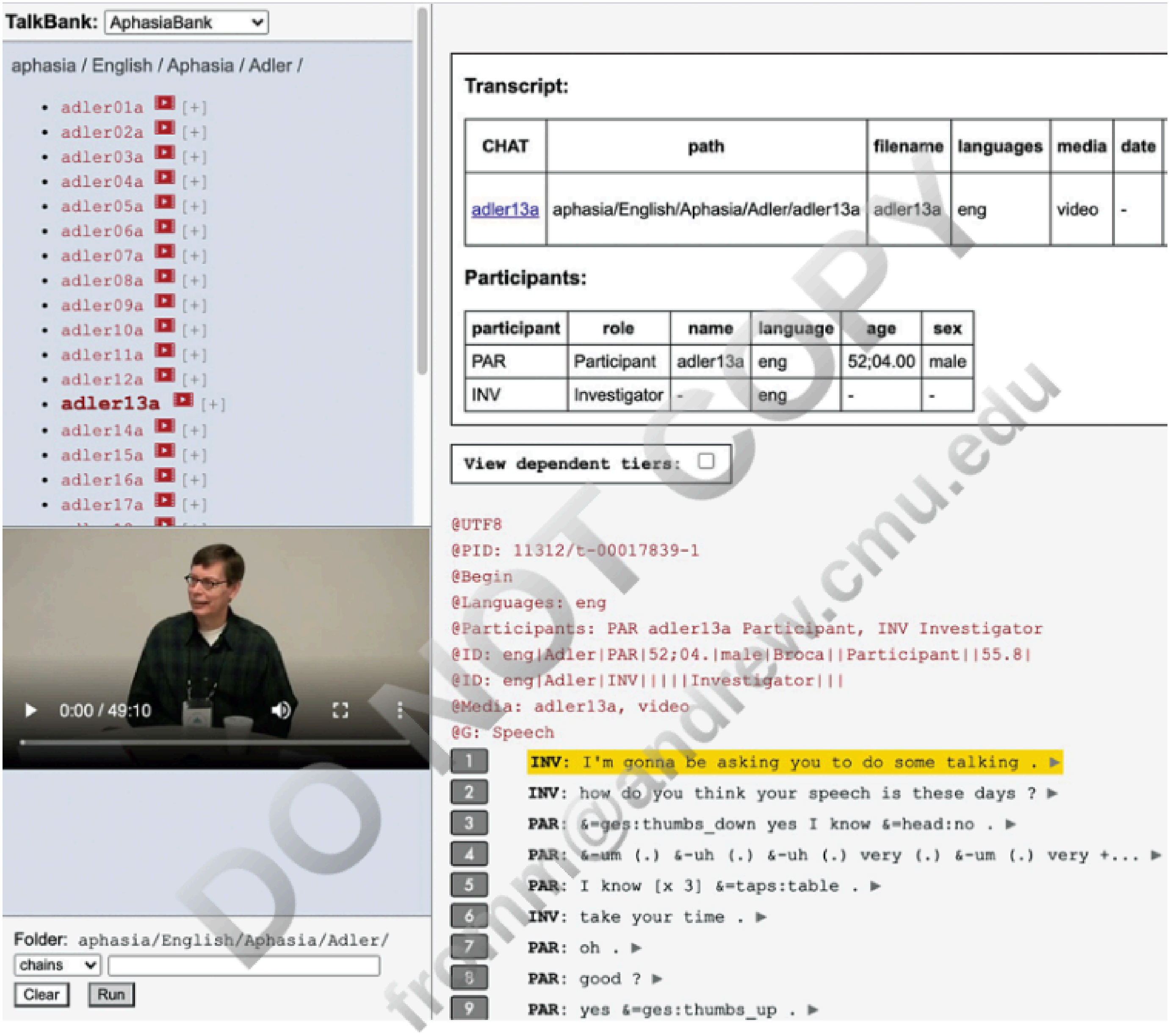
In addition to the various sets of *Grand Rounds* pages and *Classroom Activities* that have already been mentioned, TalkBank has several tools that can be used with all databanks for a variety of research and teaching applications.

DO NOT COPY
fromm@andrew.cmu.edu

Printed by: fromm@andrew.cmu.edu. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.

11.4.1 Browsable Database

The Browsable Database provides direct playback of the transcripts and media in a databank without having to download anything. A directory in the upper left corner of the screen allows users to select the language, the corpus, and the file of interest. As the media file plays, yellow highlighting appears on the corresponding transcript line, as illustrated in Figure 11.2. This facility is particularly useful for scanning over a corpus and for providing easy access for student work. The Browsable Database facility also provides the platform for the Collaborative Commentary system, and it can be called up to display the lines that correspond to specific string matches in a TalkBankDB search.



TalkBank: AphasiaBank

aphasia / English / Aphasia / Adler /

- adler01a [+]
- adler02a [+]
- adler03a [+]
- adler04a [+]
- adler05a [+]
- adler06a [+]
- adler07a [+]
- adler08a [+]
- adler09a [+]
- adler10a [+]
- adler11a [+]
- adler12a [+]
- **adler13a** [+]
- adler14a [+]
- adler15a [+]
- adler16a [+]
- adler17a [+]

Folder: aphasia/English/Aphasia/Adler/

chains

Clear Run

Transcript:

CHAT	path	filename	languages	media	date
adler13a	aphasia/English/Aphasia/Adler/adler13a	adler13a	eng	video	-

Participants:

participant	role	name	language	age	sex
PAR	Participant	adler13a	eng	52;04.00	male
INV	Investigator	-	eng	-	-

View dependent tiers: □

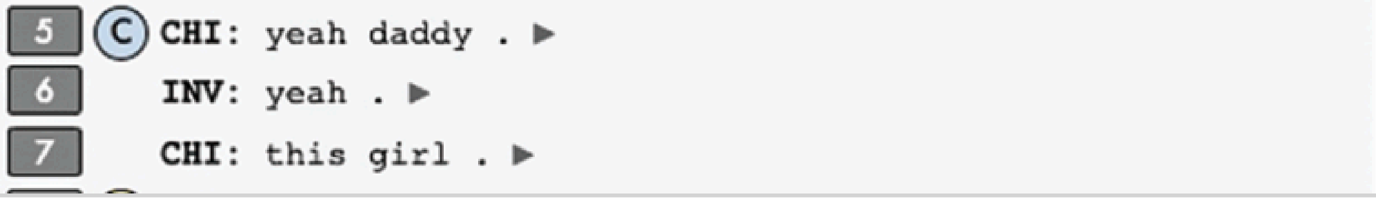
```
@UTF8
@PID: 11312/t-00017839-1
@Begin
@Languages: eng
@Participants: PAR adler13a Participant, INV Investigator
@ID: eng|Adler|PAR|52;04.|male|Broca||Participant||55.8|
@ID: eng|Adler|INV|||||Investigator|||
@Media: adler13a, video
@G: Speech
1  INV: I'm gonna be asking you to do some talking . ▶
2  INV: how do you think your speech is these days ? ▶
3  PAR: &=ges:thumbs_down yes I know &=head:no . ▶
4  PAR: &-um (.) &-uh (.) &-uh (.) very (.) &-um (.) very +... ▶
5  PAR: I know [x 3] &=taps:table . ▶
6  INV: take your time . ▶
7  PAR: oh . ▶
8  PAR: good ? ▶
9  PAR: yes &=ges:thumbs_up . ▶
```

DO NOT COPY
fromm@andrew.cmu.edu

Figure 11.2 Browsable database © Davida Fromm.

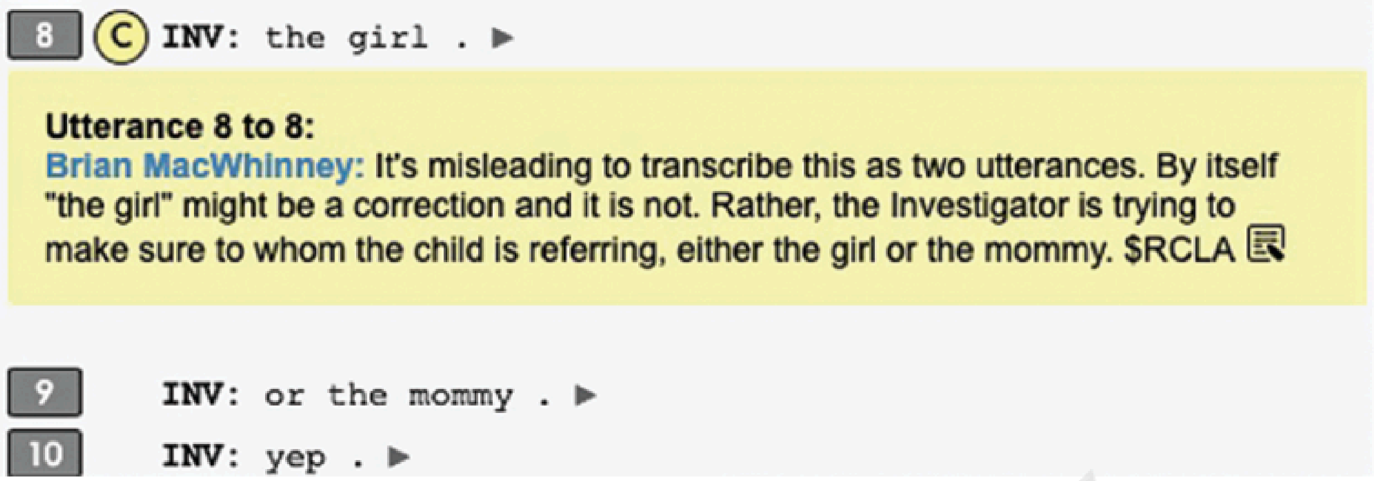
11.4.2 Collaborative Commentary

Collaborative Commentary (CC) is a tool that works from the Browsable database. It allows users to enter comments in relation to single utterances or a range of utterances, as illustrated in Figure 11.3.



```
5  Ⓒ CHI: yeah daddy . ▶
6    INV: yeah . ▶
7    CHI: this girl . ▶
```

Printed by: fromm@andrew.cmu.edu. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.



8 Ⓒ INV: the girl . ▶

Utterance 8 to 8:
Brian MacWhinney: It's misleading to transcribe this as two utterances. By itself "the girl" might be a correction and it is not. Rather, the Investigator is trying to make sure to whom the child is referring, either the girl or the mommy. $RCLA

9 INV: or the mommy . ▶

10 INV: yep . ▶

Figure 11.3 **Collaborative commentary example** © Davida Fromm.

CC allows researchers, instructors, and clinicians to form commentary groups directed by a single manager but composed of multiple group members. Members can be co-workers, colleagues, or students. They can insert analytic comments or codes directly into the online transcript display with each comment or code being tagged to a specific utterance. For example, clinical researchers can collectively evaluate behaviors and refine descriptions, research teams can measure and establish coding reliability, and students can learn to identify a variety of behaviors (e.g., paraphasias, circumlocutions, agrammatism). These comments

DO NOT COPY
fromm@andrew.cmu.edu

Printed by: fromm@andrew.cmu.edu. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.

are stored in a separate but linked postgreSQL database organized by group. Each group has access to comments from its own members, but not to those from other groups. The group manager controls the process of adding members and setting the group password. Within each group, it is possible to search for specific codes and to click on those to open the relevant segment using the Browsable Database. In this way, users can study each comment in detail and can create reports and statistics based on the comments and codes.

Collaborative Commentary provides innovative methods for analyzing spoken language. Using aphasia data as an example, CC will allow researchers to sharpen their coding and interpretation of the details of the successes and difficulties that persons with aphasia face during conversational interaction. The interpretation of the scope and causes of these difficulties can directly inform assessment, classification, and treatment. Inevitably, there will be variance between analysts in the interpretation of patterns and their causes. To quantify and analyze this variation, systematic group-specified codes can be used to draw in additional cases and examples from the larger database. In this way, creators of competing or cooperating interpretations can create a portfolio of documentation for their positions. For instructors, this type of immediate access to samples of interactions with people with aphasia can greatly enhance their students' learning.

11.4.3 TalkBankDB

TalkBankDB (https://talkbank.org/DB) is a web-based postgreSQL system that provides fuller and more direct access to and quantitative analysis of the entire TalkBank database. It allows for large segments of the database to be downloaded 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. 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. Alternatively, users can download data sets directly from Python or R.

With the entirety of TalkBank freely accessible from a simple web interface, resources that were previously found 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. There is also a GitHub account where users can upload scripts and analyses. These various options allow TalkBankDB to provide a single point where users can explore, share their research, and see what others are doing in the TalkBank community.

11.5 Summary

The assessment and treatment of discourse is now receiving intense attention from a wide range of disciplines. Researchers who have joined as members come from the departments of Biostatistics, Computer Science, Electrical Engineering, English, Geriatrics, Informatics, Linguistics, Medicine, Neurology, Psychology, and Speech and Hearing Sciences. The types of analyses that have been and can be applied to the TalkBank clinical corpora are equally broad. The need for high-quality, accessible shared databases to make this work possible cannot be overstated.

Many fundamental questions continue to plague the assessment and treatment of discourse, grammar and phonology in the clinical arena (Dietz & Boyle, 2018; Kurland & Stokes, 2018). For example, in the study of TBI, Snow and Douglas (2000) have laid out issues regarding sampling (e.g., which genres, how many, how elicited, with whom), transcribing (e.g., cost-benefit considerations), measuring (e.g., microlinguistic vs. macrolinguistic analyses), and criteria for comparison. The combination of shared databases, standard protocols, consistent transcription formats, and automated analyses now provides concrete methods for addressing these issues. Technological and methodological advances and an increased acceptance of the importance of data-sharing are helping clinical researchers advance our understanding and better inform our approaches to clinical management.

ACKNOWLEDGMENTS

This work is supported by NIDCD grant (R01-DC008524) for AphasiaBank, NIDCD grant DC015494 for FluencyBank, NICHD grant HD051698 for PhonBank, and NICHD grant (R01-HD082736) for CHILDES. We also gratefully acknowledge the teachers, researchers and clinicians who have contributed data to the shared database and the participants who have consented to participate and share their data. We are also deeply grateful to Audrey Holland and Margaret Forbes for their contributions to all these databases and their support over the years.

DO NOT COPY fromm@andrew.cmu.edu

Printed by: fromm@andrew.cmu.edu. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.

REFERENCES

Becker, J. T., Boller, F., Lopez, O. L., Saxton, J., & McGonigle, K. L. (1994). The natural history of Alzheimer's disease: Description of study cohort and accuracy of diagnosis. *Archives of Neurology*, *51*(6), 585–594.

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

Boyle, M. (2015). Stability of word-retrieval errors with the AphasiaBank stimuli. *American Journal of Speech Language Pathology*, *24*(4), 953–960. https://doi.org/10.1044/2015_AJSLP-14-0152

Casilio, M., Rising, K., Beeson, P. M., Bunton, K., & Wilson, S. M. (2019). Auditory-perceptual rating of connected speech in aphasia. *American Journal of Speech-Language Pathology*, *28*(2), 550–568. https://doi.org/10.1044/2018_AJSLP-18-0192

Chang, S.-E., Garnett, E. O., Etchell, A., & Chow, H. M. (2019). Functional and neuroanatomical bases of developmental stuttering: Current insights. *The Neuroscientist*, *25*(6), 566–582. https://doi.org/10.1177/1073858418803594

Clough, S., & Gordon, J. K. (2020). Fluent or nonfluent? Part A. Underlying contributors to categorical classifications of fluency in aphasia. *Aphasiology*, *34*(5), 515–539. https://doi.org/10.1080/02687038.2020.1727709

Coelho, C., Youse, K., Le, K., & Feinn, R. (2003). Narrative and conversational discourse of adults with closed head injuries and non-brain-injured adults: A discriminant analysis. *Aphasiology*, *17*(5), 499–510. https://doi.org/10.1080/02687030344000111

Cummings, L. (2019). On making a sandwich: Procedural discourse in adults with right-hemisphere damage. In A. Capone, M. Carapezza, & F. Lo Piparo (Eds.), *Further advances in pragmatics and philosophy: Part 2 theories and applications* (pp. 331–355). Springer.

Dalton, S. G., Kim, H., Richardson, J., & Wright, H. H. (2020). A compendium of core lexicon checklists. *Seminars in Speech and Language*, *41*(01), 045–060. https://doi.org/10.1055/s-0039-3400972

de la Fuente Garcia, S., Ritchie, C. W., & Luz, S. (2020). Artificial intelligence, speech, and language processing approaches to monitoring Alzheimer's disease: A systematic review. *Journal of Alzheimer's Disease*, *78*(4), 1547–1574. https://doi.org/10.3233/JAD-200888

Dietz, A., & Boyle, M. (2018). Discourse measurement in aphasia: Consensus and caveats. *Aphasiology*, *32*(4), 487–492. https://doi.org/10.1080/02687038.2017.1398803

Fergadiotis, G., Wright, H., & West, T. M. (2013). Measuring lexical diversity in narrative discourse of people with aphasia. *American Journal of Speech-Language Pathology*, *22*(2), 397–408. https://doi.org/10.1044/1058-0360(2013/12-0083)

Ferreira, F. (2007). Prosody and performance in language production. *Language and Cognitive Processes*, *22*(8), 1151–1177. https://doi.org/10.1080/01690960701461293

Forbes, M., Fromm, D., & MacWhinney, B. (2012). AphasiaBank: A resource for clinicians. *Seminars in Speech and Language*, *33*(3), 217–222. https://doi.org/10.1055/s-0032-1320041

Fridriksson, J., Hubbard, I., Hudspeth, S. G., Holland, A., Bonilha, L., Fromm, D., & Rorden, C. (2012). Speech entrainment enables patients with Broca's aphasia to produce fluent speech. *Brain*, *135*(Pt 12), 3815–3829. https://doi.org/10.1093/brain/aws301

Fromm, D., Greenhouse, J., Pudil, M., Shi, Y., & MacWhinney, B. (2021). Enhancing the classification of aphasia: A statistical analysis using connected speech. *Aphasiology*, *36*(12), 5. https://doi.org/10.1080/02687038.2021.1975636

Fromm, D., Katta, F., Paccione, M., Hecht, F., Greenhouse, J. B., MacWhinney, B., & Schnur, T. T. (2021). A comparison of manual vs. automated Quantitative Production Analysis of connected speech. *Journal of Speech, Language, and Hearing Research*, *64*(4), 1271–1282. https://doi.org/10.1044/2020_JSLHR-20-00561

Fromm, D., MacWhinney, B., & Thompson, C. K. (2020). Automation of the northwestern narrative language

DO NOT COPY fromm@andrew.cmu.edu

Printed by: fromm@andrew.cmu.edu. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.

Fromm, D., MacWhinney, B., & Thompson, C. K. (2020). Automation of the northwestern narrative language analysis system. *Journal of Speech, Language and Hearing Research*, *63*(6), 1835–1844. https://doi.org/10.1044/2020_JSLHR-19-00267

Gerlach, H., Totty, E., Subramanian, A., & Zebrowski, P. (2018). Stuttering and labor market outcomes in the United States. *Journal of Speech, Language, and Hearing Research*, *61*(7), 1649–1663. https://doi.org/10.1044/2018_JSLHR-S-17-0353

Gordon, J. K., & Clough, S. (2020). How fluent? Part B. Underlying contributors to continuous measures of fluency in aphasia. *Aphasiology*, *34*(5), 643–663. https://doi.org/10.1080/02687038.2020.1712586

Herd, P., Carr, D., & Roan, C. (2014). Cohort profile: Wisconsin longitudinal study (WLS). *International Journal of Epidemiology*, *43*(1), 34–41. https://doi.org/10.1093/ije/dys194

Holland, A., Forbes, M., Fromm, D., & MacWhinney, B. (2019). Communicative strengths in severe aphasia: The famous people protocol and its value in planning treatment. *American Journal of Speech Language Pathology*, *28*(3), 1010–1018. https://doi.org/10.1044/2019_AJSLP-18-0283

Hsu, C.-J., & Thompson, C. K. (2018). Manual versus automated narrative analysis of agrammatic production patterns: The northwestern narrative language analysis and computerized language analysis. *Journal of Speech, Language, and Hearing Research*, *61*(2), 373–385. https://doi.org/10.1044/2017_JSLHR-L-17-0185

Johnson, M., Randolph, E., Fromm, D., & MacWhinney, B. (2019). Comparisons of narrative discourse in Right Hemisphere Brain Damage (RHD), aphasia, and healthy adults. Poster presented at the American Speech-Language-Hearing Association convention, Orlando, FL.

Kennedy, M. R., Strand, E. A., Burton, W., & Peterson, C. (1994). Analysis of first-encounter conversations of right-hemisphere-damaged adults. *Clinical Aphasiology*, *22*, 67–80.

Kim, H., & Wright, H. H. (2020). Concurrent validity and reliability of the core lexicon measure as a measure of word retrieval ability in aphasia narratives. *American Journal of Speech-Language Pathology*, *29*(1), 101–110. https://doi.org/10.1044/2019_AJSLP-19-0063

Kuijper, S. J., Hartman, C. A., & Hendriks, P. (2015). Who is he? Children with ASD and ADHD take the listener into account in their production of ambiguous pronouns. *PLoS ONE*, *10*(7), e0132408. https://doi.org/10.1037/abn0000231

Kurland, J., & Stokes, P. (2018). Let’s talk real talk: An argument to include conversation in a D-COS for aphasia research with an acknowledgment of the challenges ahead. *Aphasiology*, *32*(4), 475–478. https://doi.org/10.1080/02687038.2017.1398808

Le, D., Licata, K., & Provost, E. M. (2018). Automatic quantitative analysis of spontaneous aphasic speech. *Speech Communication*, *100*, 1–12. https://doi.org/10.1016/j.specom.2018.04.001

Lin, D., Crabtree, J., Dillo, I., Downs, R. R., Edmunds, R., Giaretta, D., DeGiusti, M., L’Hours, H., Hugo, W., Jenkyns, R., Khodiyar, V., Martone, M. E., Mokrane, M., Navale, V., Petters, J., Sierman, B., Sokolova, D. V., Stockhause, M., & Westbrook, J., Jenkyns, R. (2020). The TRUST Principles for digital repositories. *Scientific Data*, *7*(1), 1–5. https://doi.org/10.1038/s41597-020-0486-7

Luz, S., Haider, F., de la Fuente, S., Fromm, D., & MacWhinney, B. (2021). *Detecting cognitive decline using speech only: The ADReSSo Challenge. arXiv preprint arXiv:2104.09356.*

MacWhinney, B., Fromm, D., Forbes, M., & Holland, A. (2011). AphasiaBank: Methods for studying discourse. *Aphasiology*, *25*(11), 1286–1307. https://doi.org/10.1080/02687038.2011.589893

Minga, J., Fromm, D., Jacks, A., Stockbridge, M., Nelthropp, J., & MacWhinney, B. (2022). The effects of right hemisphere brain damage on question-answering in conversation. *Journal of Speech, Language and Hearing Research*, *65*(2), https://doi.org/10.1044/2021_JSLHR-21-00309

Minga, J., Fromm, D., Williams-DeVane, C., & MacWhinney, B. (2020). Question use in adults with Right-Hemisphere Brain Damage. *Journal of Speech, Language, and Hearing Research*, *63*(3), 738–748. https://doi.org/10.1044/2019_JSLHR-19-00063

Minga, J., Johnson, M., Blake, M. L., Fromm, D., & MacWhinney, B. (2021). Making sense of right hemisphere discourse using RHDBank. *Topics in Language Disorders*, *41*(1), 99–122

DO NOT COPY
fromm@andrew.cmu.edu

Printed by: fromm@andrew.cmu.edu. Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.

discourse using RHDBank. *Topics in Language Disorders*, *41*(1), 99–122. https://doi.org/10.1097/tld.0000000000000244

Norman, R. S., Mueller, K. D., Huerta, P., Shah, M. N., Turkstra, L. S., & Power, E. (2022). Discourse performance in adults with mild traumatic brain injury, orthopedic injuries, and moderate to severe traumatic brain injury, and healthy controls. *American Journal of Speech-Language Pathology*, *31*(1), 67–83. https://doi.org/10.1044/2021_AJSLP-20-00299

Olness, G. S., Ulatowska, H. K., Wertz, R. T., Thompson, J. L., & Auther, L. L. (2002). Discourse elicitation with pictorial stimuli in African Americans and Caucasians with and without aphasia. *Aphasiology*, *16*(4–6), 623–633.

Perez, M., Aldeneh, Z., & Provost, E. M. (2020). *Aphasic speech recognition using a mixture of speech intelligibility experts. arXiv preprint arXiv:2008.10788*.

Power, E., Weir, S., Richardson, J., Fromm, D., Forbes, M., MacWhinney, B., & Togher, L. (2019). Patterns of narrative discourse in early recovery following severe Traumatic Brain Injury. *Brain Injury*, *34*(1), 98–109. https://doi.org/10.1080/02699052.2019.1682192

Richardson, J., & Dalton, S. G. H. (2020). Main concepts for two picture description tasks: An addition to Richardson and Dalton, 2016. *Aphasiology*, *34*(1), 119–136. https://doi.org/10.1080/02687038.2018.1561417

Richardson, J., Grace Dalton, S., Greenslade, K., Jacks, A., Haley, K., & Adams, J. (2021). Main concept, sequencing, and story grammar (MSSG) analyses of Cinderella narratives in a large sample of persons with aphasia. *Brain Sciences*, *11*(1), https://doi.org/10.3390/brainsci11010110

Sekine, K., Rose, M. L., Foster, A. M., Attard, M. C., & Lanyon, L. E. (2013). Gesture production patterns in aphasic discourse: In-depth description and preliminary predictions. *Aphasiology*, *27*(9), 1031–1049. https://doi.org/10.1080/02687038.2013.803017

Smith, K. G., & Clark, K. F. (2019). Error analysis of oral paragraph reading in individuals with aphasia, *Aphasiology*, *33*(2), 234–252. https://doi.org/10.1080/02687038.2018.1545992

Smith, M., Cunningham, K. T., & Haley, K. L. (2019). Automating error frequency analysis via the phonemic edit distance ratio. *Journal of Speech, Language, and Hearing Research*, *62*(6), 1719–1723. https://doi.org/10.1044/2019_JSLHR-S-18-0423

Snow, P. C., & Douglas, J. M. (2000). Subject review: Conceptual and methodological challenges in discourse assessment with TBI speakers: Towards an understanding. *Brain Injury*, *14*(5), 397–415. https://doi.org/10.1080/026990500120510

Stark, B. (2019). A comparison of three discourse elicitation methods in aphasia and age-matched adults: Implications for language assessment and outcome. *American Journal of Speech-Language Pathology*, *28*(3), 1067–1083. https://doi.org/10.1044/2019_AJSLP-18-0265

Stubbs, E., Togher, L., Kenny, B., Fromm, D., Forbes, M., MacWhinney, B., McDonald, S., Tate, R., Turkstra, L., Power, E. (2018). Procedural discourse performance in adults with severe traumatic brain injury at 3 and 6 months post injury. *Brain Injury*, *32*(2), 167–181. https://doi.org/10.1080/02699052.2017.1291989

Tichenor, S., & Yaruss, J. S. (2020). Repetitive negative thinking, temperament, and adverse impact in adults who stutter. *American Journal of Speech-Language Pathology*, *29*(1), 201–215. https://doi.org/10.1044/2019_AJSLP-19-00077

van Nispen, K., van de Sandt-Koenderman, M., Sekine, K., Krahmer, E., & Rose, M. L. (2017). Part of the message comes in gesture: How people with aphasia convey information in different gesture types as compared with information in their speech. *Aphasiology*, *31*(9), 1078–1103. https://doi.org/10.1080/02687038.2017.1301368

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg,N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., … Bourne, P. E. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 1–9. https://doi.org/10.1038/sdata.2016.18

Yairi, E., & Ambrose, N. G. (1999). Early childhood stuttering I: Persistency and recovery rates. *Journal of Speech, Language, and Hearing Research*, *42*(5), 1097–1112.

DO NOT COPY
fromm@andrew.cmu.edu