The SAGE Encyclopedia of Human Communication Sciences and Disorders

Databases in Communication Disorders

Contributors: Davida Fromm
Edited by: Jack S. Damico & Martin J. Ball
Book Title: The SAGE Encyclopedia of Human Communication Sciences and Disorders
Chapter Title: "Databases in Communication Disorders"
Pub. Date: 2019
Access Date: May 1, 2019
Publishing Company: SAGE Publications, Inc.
City: Thousand Oaks,
Print ISBN: 9781483380834
Online ISBN: 9781483380810
DOI: http://dx.doi.org/10.4135/9781483380810.n177
Print pages: 542-544

Computerized databases for the study of communication disorders have grown over the past several decades. Many of these databases involve the collection of large amounts of spoken or written language accessed over the Internet. Conditions for access to these collections vary widely. The TalkBank System is currently the world's largest repository of shared databases for spoken language and is freely open to researchers, educators, and clinicians. Its many component banks provide language corpora and resources for research and education in fields such as communication sciences and disorders, linguistics, psychology, education, and computer science. These databases include language samples of children and adults of all ages, in English and many other languages, and with and without clinical impairments.

Each of the TalkBank language banks has a separate web page with an index of its collection, providing corpus names, corpus contributors' names, and short descriptions of the corpus contents. Each of the corpus names links to a page with more detailed information about the corpus as well as links for downloading the transcripts and media. TalkBank's browsable database allows users to view video samples (or listen to audio samples) along with their corresponding transcripts. This entry explains how TalkBank works and then describes eight of the databanks available on the TalkBank website.

About TalkBank

Language samples in the TalkBank system are transcribed in CHAT (Codes for the Human Analysis of Transcripts) format and linked to the digitized audio and video files. CHAT is designed to operate closely with a set of programs called CLAN (Computerized Language Analysis), which permit the analysis of a wide range of linguistic and discourse structures. Automatic morphosyntactic analysis programs are built into CLAN for English and 10 other languages, producing a "%mor" line and "%gra" line below each speaker line in the transcript.

The %mor line provides complete morphosyntactic tagging for the utterance above, and the %gra line provides a grammatical dependency analysis. CLAN commands can be used to perform computations such as total words, total utterances, type-token ratios, word frequency analyses, mean length of utterance (in words or morphemes), words per minute, proposition density, lexical diversity, part-of-speech analyses, and grammatical complexity. Two omnibus utility programs, EVAL and KIDEVAL, allow users to compare an individual participant's expressive language data with means and standard deviations for comparison databases from AphasiaBank and CHILDES (Child Language Data Exchange System), respectively.

EVAL produces over 30 morphosyntactic and fluency measures, and KIDEVAL produces over 40 measures, including DSS (decision support systems), IPSyn (Index of Productive Syntax), and Brown's morphemes. These programs are also used to compare an individual's performance changes over time. Manual coding can be added to the transcripts for word-level (e.g., paraphasias) and utterance-level (e.g., speech acts, main concepts, coherence) behaviors, which can then be automatically searched and analyzed.

Together, the TalkBank databases and TalkBank software tools provide resources for the study of language, language acquisition, and language disorders. The repositories and corpora that relate most directly to the study of communication disorders are described next in alphabetical order, beginning with the child language sites followed by the adult language sites. Several of the adult language repositories contain corpora that use similar discourse tasks, allowing for comparisons across diagnostic categories. Some databases are password protected and available on request and provision of the necessary professional credentials.

TalkBank Databanks

ASDBank

ASDBank is one of the smaller repositories in the TalkBank system, comprising data from both children and adults with autism spectrum disorder (ASD). Researchers have contributed transcripts and some accompanying media files (mostly audio) for English, Dutch, French, Greek, Mandarin, and Spanish corpora. Populations represented include individuals with ASD, Down syndrome, attention-deficit hyperactivity disorder (ADHD), and Asperger's syndrome, as well as infants at risk for ASD and typically developing individuals. Samples vary from naturalistic interactions during preschool activities or with a parent to structured storytelling, book-reading activities, and clinical interactions. Researchers have used these corpora to analyze topics such as word learning, language production and comprehension asymmetries, and pragmatic skills. Some corpora include longitudinal data, allowing for the study of language acquisition.

CHILDES

CHILDES began in 1984 and is the primary child language component of the TalkBank system. The collection includes hundreds of corpora from contributors around the world representing dozens of languages. The repository has various types of data (e.g., longitudinal, cross-sectional, case studies) and various types of language samples (e.g., mother–child interactions, narratives, free play) for normally developing children of all ages and for some clinical populations (e.g., children with hearing impairments, traumatic brain injury, specific language impairment).

Over 7,000 published articles are based on the use of data or programs from CHILDES. This work includes the areas of phonology, morphology, syntax, lexicon, narrative, literacy, and discourse. The availability of a shared open database has been essential in the development of analysis and theory. Researchers have been able to use comparable data to evaluate alternative theoretical approaches, such as connectionist models and dual-route models of learning or emergentist and generativist theories of acquisition. CHILDES data and programs have also been widely used to provide materials for teaching courses in language development.

FluencyBank

FluencyBank was recently funded and created for the study of the development of fluency and disfluency in children and language learners. A primary goal is to characterize the developmental pathway of fluency and disfluency in children between the ages of 3 and 7. During this period, many children who show signs of early disfluency become normally fluent, with only a fraction of this population developing stuttering. How and why this occurs developmentally is unclear, largely because data from this period are limited.

A standard protocol is being developed for data collection, and TalkBank methods will be used to conduct a longitudinal study across this period. A newly developed CLAN program called FLUCALC generates two dozen measures that clinicians and researchers can use to assess speech fluency behaviors. This utility tracks and computes the frequency of both stutter-like disfluencies (e.g., blocks, prolongations, part-word and single-word repetitions) and typical disfluencies. FluencyBank also includes data contributed from earlier studies of disfluency from a variety of laboratories and a variety of participants, including normally developing monolingual and bilingual children, children with disfluencies, adults with disfluencies, and second-language learners. Corpora are mostly in English, but the repository also includes a large German corpus and a Dutch case study.

PhonBank

PhonBank is the child phonology component of the TalkBank system. As of May 2017, the repository contained 40 corpora of early child phonological productions across 12 languages, all transcribed in IPA (international phonetic alphabet) along with the target language forms and linked directly to the audio record. The corpora are available in CHAT and Phon formats. CHAT files can be analyzed using CLAN programs, as described earlier.

Files in Phon format can be analyzed using the Phon program, which has all the basic analyses required for tracking growth of segments, features, prosodic patterns, and phonological processes in the study of child phonology. Phon also incorporates the full source code of Praat (a software program for the analysis of speech in phonetics), making it possible to run Praat's acoustic analysis directly inside Phon and storing the results in the Phon transcript.

AphasiaBank

AphasiaBank began in 2007 and is a repository of videotaped language samples, test results, and demographic information from adults with aphasia as well as a comparison group of adults without aphasia. Aphasia results from damage to the language areas of the brain (usually the left hemisphere) and may impair expression, comprehension, reading, and writing. The overarching goal of AphasiaBank is to support the development of methods to help people with aphasia improve their communicative use of language.

A unique feature of this language bank is its standard discourse protocol and elicitation script used to gather language samples, which include free speech, picture descriptions, the Cinderella story narrative, and a procedural discourse task. These protocol samples are mostly in English, but smaller corpora are available for French, Cantonese, German, Italian, Japanese, Spanish, and some bilingual participants. The standard discourse protocol is augmented by a standard test battery and comprehensive demographic data collection on all participants. Aphasia researchers and clinicians have contributed many other videos and transcripts of nonprotocol samples (e.g., conversations, story retells, other picture descriptions) as well as treatment approaches (e.g., script training, group therapy). The database has been used in over a 100 publications, presentations, and theses.

Normative data from the standard protocol have been used to develop clinician-friendly discourse evaluation tools such as core lexicons and main concept checklists. Other research has addressed topics such as lexical diversity, proposition density, word retrieval, errors, agrammatism, syntax, gestures, story grammar, coherence, listener perceptions, prosody, test–retest reliability, recovery, crosslinguistic differences, treatment outcomes, and the effect of demographic and other clinical variables on language performance. Education resources include a Grand Rounds link with video samples, descriptions of classic aphasia types, and discussion questions about the language samples and potential treatment approaches.

DementiaBank

DementiaBank includes transcripts and media from individuals with various types of dementia as well as individuals with primary progressive aphasia. Although dementia has many potential causes and presentations, it usually involves gradually worsening impairments in memory, communication, reasoning, and orientation. Language symptoms in dementia depend largely on the type and severity of dementia. In general, language production deficits may include word-finding problems, empty speech, paraphasias, circumlocution, perseveration, and reduced output. The largest corpus in this repository contains longitudinal data for four language tasks (Cookie Theft picture descriptions, a sentence construction task, word fluency, and a story retell task) from individuals with Alzheimer's disease (AD) and other types of dementia as well as elderly controls.

These data have been of particular interest to researchers who are using machine learning and linguistic analysis to automatically identify AD from short narrative samples and researchers who are working to improve speech recognition skills in personal assistive robots trained to work with older adults with AD. Other corpora in DementiaBank include conversations and other language tasks from individuals with AD. Recent contributions of discourse data from individuals with primary progressive aphasia are being added to this repository. Corpora have been contributed in German, Mandarin, Spanish, Taiwanese, and English.

RHDBank

RHDBank is another of the more recent databases created for the study of communication in people with right hemisphere disorder (RHD) resulting from brain damage to the right hemisphere in adults. Symptoms of RHD include cognitive-communication deficits that impair pragmatic skills, resulting in difficulties producing and comprehending discourse. Specifically, difficulty with topic maintenance, discourse coherence and cohesion, inference generation, turn-taking, question use, and the integration of contextual nuance are among the deficits commonly seen in people with RHD.

As with the main AphasiaBank corpus, this corpus is based on a standard discourse protocol, demographic data collection, and set of assessment procedures. Data are from individuals with a history of right hemisphere stroke and individuals without neurological impairments. The discourse protocol includes free speech, picture descriptions, the Cinderella story-telling task, a procedural discourse task, a question production task, and a first-encounter conversation.

TBIBank

TBIBank is a shared database of 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. Discourse has been described as disorganized, inappropriate, tangential, unclear, redundant, and self-focused. As with AphasiaBank, this repository includes media files and transcripts from a standard discourse protocol as well as other contributions of conversations, story retells, story generations, picture descriptions, and procedural discourse. The standard discourse protocol was developed for a large longitudinal study of communication recovery after TBI. Demographic data and test results are available for most of the larger corpora.

See also Communication Disorders; Connected Speech; Language; Linguistics

Websites

AphasiaBank: http://aphasia.talkbank.org

ASDBank: http://talkbank.org/access/ASDBank

CHILDES (Child Language Data Exchange System): http://childes.talkbank.org

DementiaBank: http://dementia.talkbank.org

FluencyBank: http://fluency.talkbank.org

PhonBank: http://phonbank.talkbank.org

RHDBank: http://rhd.talkbank.org

TalkBank: https://talkbank.org

TBIBank: http://tbi.talkbank.org

Davida Fromm

http://dx.doi.org/10.4135/9781483380810.n177
10.4135/9781483380810.n177

Further Readings

MacWhinney, B. (2000). The CHILDES Project: Tools for analyzing talk (3rd ed.). Mahwah, NJ: Lawrence.
MacWhinney, B., & Fromm, D. (2016). AphasiaBank as BigData. Seminars in Speech and Language, 37(1), 10–22. https://dx.doi.org/10.1055/s-0036-1571357
MacWhinney, B., Fromm, D., Holland, A., & Forbes, M. (2012). AphasiaBank: Data and methods. In N. Mueller & M. Ball (Eds.), Methods in clinical linguistics (pp. 31–48). New York, NY: Wiley.