CLARIN

Common Language Resources and Technology Infrastructure

Tour de CLARIN

The TalkBank Knowledge Centre

Edited by **Darja Fišer** and **Jakob Lenardič**

Foreword

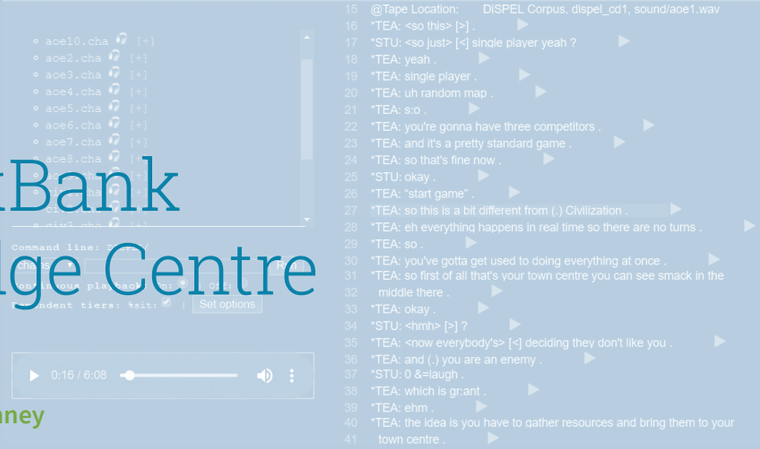
Tour de CLARIN highlights prominent user involvement activities of CLARIN National Consortia and Knowledge Centres with the aim to increase their visibility, reveal the richness of the CLARIN landscape, and display the full range of activities throughout the CLARIN network that can inform and inspire other consortia and knowledge centres as well as show what CLARIN has to offer to researchers, teachers, students, professionals and the general public interested in using and processing language data in various forms.

The brochure presents the TalkBank Knowledge Centre and is organized in two sections:

- Section One presents the members of the Knowledge Centre and their work
- Section Two includes an interview with a renowned researcher from the digital humanities or social sciences who has successfully used the Knowledge Centre’s infrastructure in their research

The TalkBank Knowledge Centre

The TalkBank Knowledge Centre



Introduction

Written by **Brian MacWhinney**

TalkBank, which was recognized as a CLARIN Knowledge Centre in 2016, is the world's largest open access integrated repository for spoken language data.[1] It provides language corpora and other audio resources to support researchers in Psychology, Linguistics, Education, Computer Science, and Speech Pathology. The National Institutes of Health and the National Science Foundation have provided support for the construction of five components of TalkBank:

- AphasiaBank for the study of language in aphasia in six languages;
- CHILDES for the study of child language development in 42 languages from infancy to age six;
- FluencyBank for the study of language fluency and disfluency in stuttering, aphasia, second language learning, and normal processing;
- HomeBank for the study through automatic speech recognition of untranscribed daylong recordings in the home and elsewhere; and
- PhonBank for the analysis of children's phonological development in 18 languages.

The five components, which involve multiple corpora collected and encoded according to the same principles contributed by individual researchers from all over the world, form very large collections that are being used extensively to study the cognitive, neurological, developmental, and social bases of language processing and structure. In addition to our support for these five areas, TalkBank also promotes the growth of corpora in nine other related areas:

- ASDBank for the study of language in autism spectrum disorder;
- BilingBank for the study of bilingualism and multilingualism;
- CABank for the study of conversation using the methods of Conversation Analysis;
- ClassBank for the study of language in the classroom;
- DementiaBank for the study of language in dementia;
- RHDBank for the study of language in right hemisphere damage;
- SamtaleBank for the study of conversations in Danish;
- SLABank for the study of second language learning; and
- TBIBank for the study of language in traumatic brain injury.

[1] http://talkbank.org/

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

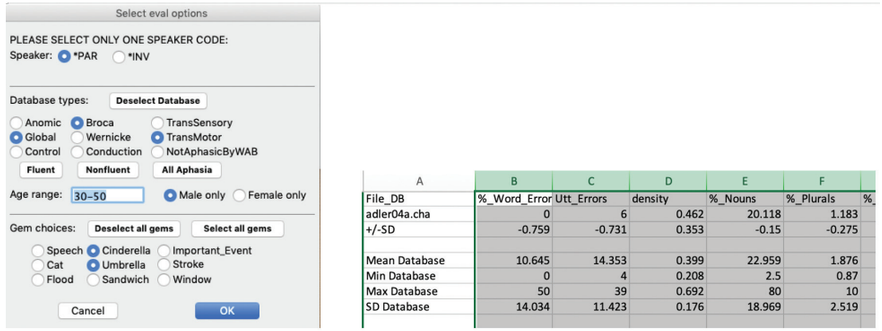
In the physical sciences, the process of data-sharing is taken as a given. However, data-sharing has not yet been adopted as the norm in the Social Sciences and Humanities. This failure to share research results – much of it supported by public funds – represents a big loss to science. Researchers often cite privacy concerns as reasons for not sharing data on spoken interaction. In response to this, TalkBank provides a variety of options in which data can be made available to other researchers, while still preserving participant anonymity, such as password protection and pseudonymization of the participants' first and last names.

CHAT Transcription format

As individual researchers sample from a great diversity of language contexts, they tend to develop idiosyncratic, incompatible methods for language transcription and analysis. In order to provide maximum harmonization across these formats, TalkBank has created an inclusive transcription standard, called CHAT, that recognizes all the features required by different disciplinary analyses. Furthermore, CHAT allows researchers to link transcripts directly to the audio or video, which significantly speeds up transcription and improves accuracy.

CHAT-consistent software

The basic program for analysis of TalkBank data is called CLAN. For language analysis, CLAN automatically computes clinical measures, such as the mean length of the utterance (MLU), the Type-Token Ratio (TRR), Brown's morphemes (for children), and several other values, without errors. Figure 1 illustrates the use of a dialog in CLAN's EVAL program for comparing a transcript from a single participant with those from matched participants in the larger AphasiaBank database.



	A	B	C	D	E	F
	File_DB	%_Word_Error	Utt_Errors	density	%_Nouns	%_Plurals
	adler04a.cha	0	6	0.462	20.118	1.183
	+/-SD	-0.759	-0.731	0.353	-0.15	-0.275
	Mean Database	10.645	14.353	0.399	22.959	1.876
	Min Database	0	4	0.208	2.5	0.87
	Max Database	50	39	0.692	80	10
	SD Database	14.034	11.423	0.176	18.969	2.519

Figure 1: *A dialog (on the left) in CLAN's EVAL program for comparing a transcript from a single participant (Adler04a) with those from matched participants in the larger AphasiaBank database. On the right is a small segment of the Excel output of the analysis with means and standard deviations.*

Much of the morphosyntactic analysis in CLAN depends on the use of automatic part-of-speech taggers and grammatical dependency taggers that we have constructed for Cantonese, Chinese, Danish, Dutch, English, French, German, Hebrew, Japanese, Italian, and Spanish. The TalkBankDB database search engine permits rapid searches of the database, CQL queries, graphic displays, and downloading of data in CSV format for further statistical analysis. A user-friendly guide for using CLAN that does not presuppose technical knowledge was written by Nan Bernstein Ratner (University of Maryland) and Shelley B. Brundage (George Washington University).

Interoperability

The PhonBank component of TalkBank has developed a separate program called Phon, which provides extensive support for the analysis of phonological data. Crucially, the entire code and functionality of the popular PRAAT software for phonetic transcription are now included inside Phon. Compatibility with other common formats, including Anvil, CONNL, DataVyu, ELAN, EXMARaLDA, LENA, Praat, SRT, SALT, and Transcriber is achieved through translation programs inside CLAN. Recently, Christophe Parisse from INSERM/CNRS and Ortolang, also the repository of the French CLARIN observer, has built a powerful new editor called TRJS, which is used for the transcription, editing and visualization of data and corpora of spoken language, and works directly with the CHAT, ELAN, and TEI formats.

Responsivity to research community needs

TalkBank seeks to be maximally responsive to the needs of individual researchers and their research communities, as well as instructors and clinicians. 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.

Each corpus page includes a link to a facility called the TalkBank browser that allows users to play back linked multimedia corpora directly in their web browser (Figure 2). Users can choose to have continuous playback or playback of specific sections or utterances. For AphasiaBank, FluencyBank, RHDBank, and TBIBank, there are web pages with example videos and instructional commentary designed for use in teaching about language disabilities.

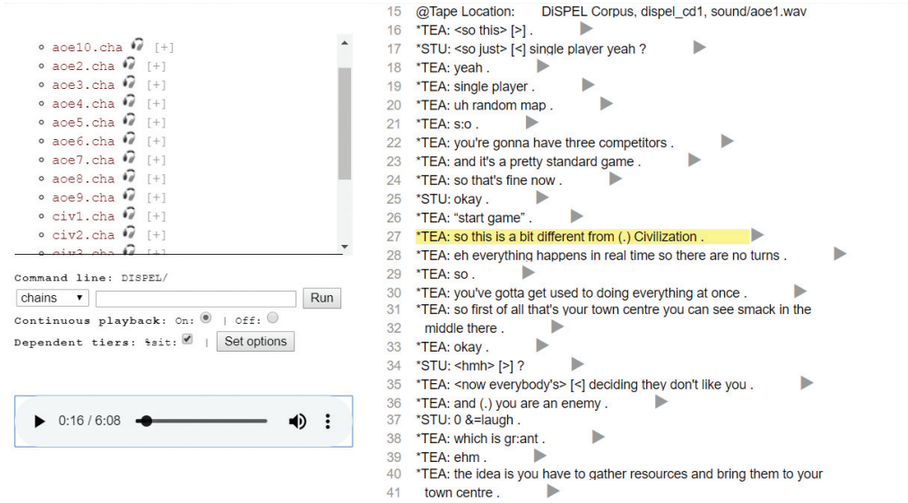


Figure 2: *An audio recording and its assorted transcription in the Dispel Corpus. Using the CLAN editor, the transcriptions have been aligned with the recording, as shown by the yellow highlight. Features inherent to spoken language are also transcribed. For instance, the symbol (.) in the highlighted text stands for a verbal pause made by the speaker.*

TalkBank provides several avenues for user support. In addition to detailed manuals, configuration as a CLARIN Knowledge Centre, and GoogleGroups lists for user support, we have created screencast tutorials that explain how to use the database and the tools. These are hosted both on our own servers and through YouTube. We also conduct presentations and workshops each year at international conferences, such as IASCL, ASHA, LSA, the Academy of Aphasia, LREC, and CLARIN.

International Standards

The sixth basic TalkBank principle is our commitment to international standards for database and language technology. Toward this end, TalkBank has joined the CLARIN federation and is now one of the two members of the CLARIN ERIC infrastructure outside Europe. In 2017, TalkBank received the approval of the Core Trust Seal, which emphasizes the adoption of international standards in data access, protection of confidentiality, organizational infrastructure, data integrity, data storage, data curation, and data preservation. To achieve this, TalkBank maintains incremental GIT repositories for all of its datasets, where researchers interested in replicating earlier analyses can obtain copies of segments of the database from any particular date. In addition, 74 520 language resources in the Virtual Language Observatory (about 10% of all resources listed) derive from TalkBank corpora. Moreover, these resources are all available through open access in a single, consistent, fully documented, and validated format.

Interview | **Nan Bernstein Ratner**

Nan Bernstein Ratner is a Professor at the Department of Hearing and Speech Sciences, University of Maryland, College Park, as well as a Fellow and Honors recipient of the American Speech, Language and Hearing Association. Professor Bernstein Ratner is, along with Brian MacWhinney, one of the PIs of FluencyBank, a shared database for the study of the development of fluency in typical and disordered populations. FluencyBank is part of TalkBank, a CLARIN K-Centre.

Please describe your academic background.

>

I began as a Child Study major at Tufts University, which offered a large number of language classes. After graduation, I originally planned to seek a PhD in Linguistics, my advisor joked that linguists had a hard time finding jobs. She recommended something “applied” that involved language, so I started a federally subsidized MA in speech-language pathology (SLP) from Temple University in Philadelphia. I soon felt that SLPs weren’t making good use of basic language acquisition research. For instance, we were just beginning to explore the ramifications of Roger Brown’s work for clinical practice. Consequently, I decided to do a PhD in Applied Psycholinguistics at Boston University. While at Temple, I wrote an argumentative term paper on why stuttering might be a language disorder with a physiological origin, which turned into a thesis that got published and well-received. But my PhD advisors Paula Menyuk and Jean Berko Gleason, the inventor of the famous Wug Test, still wanted me to pursue first language acquisition and I’ve been a split personality ever since, straddling child language development/disorder and fluency/fluency disorders. Now that I work as Professor at the Department of Hearing and Speech Sciences at the University of Maryland, I am able to combine these interests. As time goes by, they seem less and less separable – fluency and language share interesting interactions.

>

What was the motivation for the FluencyBank project?

>

It is a well-kept secret that even researchers, let alone clinicians, have a lot of trouble accurately transcribing disfluency behaviours like stuttering. What you hear and where you hear it happen can be very variable. Furthermore, fluency researchers were generally very siloed, so there was little collaborative research combining data from different projects. Most of the studies in stuttering also involved too few participants, and there weren't enough longitudinal studies. In response to this, we started the FluencyBank project[2] under the TalkBank initiative because we wanted to make our data available as part of a large-scale interoperable multi-media archive which gives access to utilities specialized for processing audio materials.

There was also a lack of a structured approach to analysing stuttering and related disfluency profiles. Researchers didn't agree on how to code these behaviours, nor were they able to combine their data because everyone made up their own codes for annotation. In this sense, FluencyBank, like the entire TalkBank initiative, was created as an open site where annotation follows a uniform standard to enable multiple data sets to be combined for greater power. Although past work that wasn't consented directly for use in FluencyBank is being kept password-protected and researchers must explain what they want to do with the data to obtain access, we aim to make all the ongoing data contributions open access, which is also in line with TalkBank as a CLARIN K-Centre. All of our teaching materials are open-access now; they are being used across the globe to teach SLP students about the behavioural, affective and cognitive features of stuttering in adults and now children.

>

Could you describe a tool offered by TalkBank that's especially important for your research?

>

The most important tool that TalkBank offers is the transcription program CLAN and its media linkage capacity. Its key advantage is that it offers an easy way to chop up the audio or video signal into very small segments and link them to lines of transcription. Researchers using this program can more reliably process what they have transcribed while listening to the relevant segment.

We think this has real implications for improving the reliability of fluency transcription. For years, I have taught a class of graduate clinicians how to code for stuttering and I would ask my students to transcribe a sample that is available through FluencyBank. Even though the segment is very short, only about 250 words long, my students strongly disagreed on how many stutters or typical disfluencies it contained. Since this sort of disagreement is common among researchers and experienced clinicians as well, we now have a study in progress in which we're trying to compare the accuracy of the CLAN transcriptions with the traditional practice where clinicians simply play the audio and write down their observations. We're doing this to raise awareness as well as to help clinicians do a better job in analysing and understanding their data.

>

How does stuttering differ from other types of disfluency? How can TalkBank help?

>

Generally, articulation and language disorders are there from the very beginning and can be noticed as soon as a child starts speaking. Stuttering, however, is unique in that it seemingly appears out of nowhere in otherwise clinically typical children between the ages of two and four years. This has spurred wide speculation in the literature as to the exact nature of this disorder. For a long time, environmental factors, such as traumatic events, were claimed to precipitate stuttering. For instance, Freud claimed that parents are to blame for stuttering and neo-Freudians promoted the view that children who stutter are suffering from some kind of psychological neurosis, despite the fact there were no data to suggest this was true. Unfortunately, this belief persists in minds of parents world-wide and is difficult to eradicate.

We now know that stuttering has neurophysiological origin and genetic predisposition. Contemporary neurological studies using brain imaging techniques suggest that there's more limited brain connectivity between the regions associated with language planning and motor execution in stutterers compared to typically fluent speakers. The underlying cause of stuttering, however, remains a mystery, so it's valuable to compare it to other forms of disfluency in terms of typology, distributions, and response to linguistic variables, such as the complexity of the intended targets.

TalkBank is an especially good environment for such comparative studies, because the FluencyBank data are interoperable with other similar collections, such as CHILDES and Phon. CLAN offers a wonderful utility called KidEval, which performs a plethora of useful statistical analyses in English and some other languages, such as clause density, counts of important morphemes that are acquired over early childhood and often missing in disordered children's speech, or mean utterance length in morphemes/words, in addition to lexical diversity measures. It then exports the analysis to an Excel spreadsheet and even compares findings to hundreds of children of the same age and sex in the CHILDES Archive. This is important for our work in fluency

[2] https://fluency.talkbank.org/

development and disorder because we now know that linguistic complexity, defined in multiple ways, can impact the fluency of a child's speech. For example, in prior research we have found that it is more likely that someone will stutter on a word like boys than on boy, even though both are phonologically equally complex.

>

What makes the application of language technologies for the analysis of speech challenging for data collection and research, and how do you overcome these challenges in FluencyBank?

>

We would love to be able to automatically differentiate stuttering from the other disfluencies, which is even more challenging in the case of children in comparison to adults, because many children don't show the active struggle in speaking and secondary behaviours that make stuttering in adults so much more obvious. There also aren't any robust pre-existing models of kids' rate and fluency development, and how typically developing children's fluency might be distinguished from that of kids with language impairment (although we have some studies suggesting that kids with language impairment are less fluent than typical kids), kids who are grappling with trying to learn to talk in more than one language, as well as kids who stutter.

It is both tedious and frustrating to document the distributional patterns of fluency in speech samples. Through my career I have repeatedly seen SLPs who make mistakes even just counting the number of words in a read paragraph. However, we have greatly streamlined this process with FluCalc, which is a computational measure in CLAN that gives a detailed breakdown of disfluency behaviours, both over intended words and syllables. Crucially, FluCalc does this by comparing the disfluency behaviours against a weighted score, which on the one hand distinguishes disfluencies that are considered more atypical (i.e., clinically relevant) from those that are considered typical (i.e., disfluency that can be found in otherwise non-disfluent speakers, who may repeat words or phrases when anxious or tired), as well as ranks the atypical disfluencies according to their pathological severity on the basis of a criterion-referenced cut-off point.

For instance, a type of atypical disfluency is the prolongation of a word-initial consonant, such as when a person articulates a word like really as /r-r-r-r-r-eally/, repeating the /r/ sound. FluCal would mark this as more severe than repeating the entire word (really really big), which speakers do all the time in everyday communication when they want to emphasise something. By contrast, blocks are a terrifying form of stuttering where a speaker opens his or her mouth but nothing comes out. A typical speaker would only experience a behaviour like this in a nightmare; thus, they are given higher weight because they would rarely appear in a typical speaker's speech. FluCalc implements a weighted score that examines what types of disfluencies you see in a person's speech, and how many repetitions, or how long a prolongation is, as measures of severity. In the research community there is now an agreement that a child can be considered as stuttering if they receive a weighted score higher than 4% on a speech sample, and FluCalc can calculate this percentage automatically, which is especially important for teachers, clinicians, and doctors.

>

Could you describe some of the recent results achieved in the project?

>

Recently, we teamed up with Purdue University, where Anne Smith and Christine Weber had previously prepared a large-scale longitudinal study in which they followed a large sample of kids who stutter and compared them with their typically fluent peers. Since TalkBank utilities gave us the ability to map multiple language measures easily from the Purdue participants' language samples, we were able to use growth modelling to show that children's expressive language skill was a statistically relevant predictor of recovery from stuttering during early childhood – that is, the better a child's general language skills were, the more it was likely that they would outgrow stuttering on their own over a three-year window of observation (Leech et al. 2019[3]).

It is estimated that 80% of children who start to stutter stop on their own, for reasons we still don't understand well. Our major clinical and research problem is separating those children from those who won't recover and should get therapy early to ensure that the child can learn to speak more easily and not develop handicapping speaking fears. In light of this fact, we are currently working with the Purdue team to determine if other linguistic factors permit us to distinguish between children likely to recover and those who are likely to be persistent. Because the Purdue data are longitudinal, we can do a cross-sectional analysis that will detangle the persistent stutterers, especially given CLAN's ability to link fluency on the speaking tier with grammatical analysis of a dependent tier.

>

[3] https://doi.org/10.1044/2019_JSLHR-S-18-0318

Could you describe the educational component of FluencyBank?

>

Yes, from the very beginning we thought that we would achieve better awareness of the project if we included a teaching component. All the other Banks in TalkBank have teaching resources. We first went to stutterers' support group meetings and asked the attendees if they wanted to participate in a recorded interview that would be transcribed, annotated and put on the FluencyBank page for educational purposes. All of the participants have consented that the interviews – both the videos and the corresponding transcriptions – are made available as open access under *Voices of Adults and Voices of Children Who Stutter and Clutter* categories in the teaching component in FluencyBank. We have standardized these interviews so that the participants are always asked to talk about the impact that stuttering has had on their lives, their experiences with treatment, and to point out those aspects of their disorder that they want clinicians to understand better.

The teaching component has become widely used in education, and I keep getting thanks from professors of stuttering courses about it. The reason for its popularity partly has to do with the fact that more than half of the training programs world-wide lack a resident stuttering "expert", so they mostly have to resort to descriptions in textbooks, which are of course much less illustrative when it comes to explaining the phenomenon or how best to work with clients/patients. We have also designed a set of exercises aimed at university teachers, and we've received positive feedback from various instructors who use the Voices interviews as homework for their graduate students. Additionally, the latest editions of the two most widely used textbooks on stuttering, which are Barry Guitar's Stuttering: *An Integrated Approach to Its Nature and Treatment* and Walter H. Manning's *Clinical Decision Making in Fluency Disorders,* now explicitly mention FluencyBank as a both clinical and research resource.

>

What are your future goals with the project?

>

We want to get more data. We are already trying to recover and preserve precious data from the "baby boomer" generation of professors who are now retiring. We also want to change the culture of the field to be more like that of child language – that data do more good when shared than when kept close to the vest of their collector. In the case of non-stuttered disfluency, we aim to show that disfluency profiles may inform subtle levels of language impairment or need for remediation that would go undetected by crude language testing, which is known to be non-specific and non-sensitive in identifying older kids with language learning needs. We also seek to show that the elevated disfluency seen in some bilingual children isn't stuttering; it's the natural profile of a child learning to talk in two languages.

For both FluencyBank and CHILDES, we also want to make the research initiative appealing, useful and easy to use for practicing clinicians. Right now, language assessment takes a lot of time and energy – we want to speed it up, make it more informative, and guide more effective therapy goal selection, follow-up and documentation of outcomes. Less time diagnosing the problem and more time available to work towards helping children speak more like their typical peers.

>

COLOPHON

This brochure is part of the ‘Tour de CLARIN’ volume II
(publication number: CLARIN-CE-2019-1537, November 2019).

Coordinated by
Darja Fišer and **Jakob Lenardič**

Edited by
Darja Fišer and **Jakob Lenardič**

Proofread by
Paul Steed

Designed by
Tanja Radež

Online version
www.clarin.eu/Tour-de-CLARIN/Publication

Publication number
CLARIN-CE-2019-1537
November 2019

ISBN
9789082990911

Contact
CLARIN ERIC
c/o Utrecht University
Drift 10, 3512 BS Utrecht
The Netherlands

www.clarin.eu

CLARIN
Common Language Resources and Technology Infrastructure

CLARIN 2019