UNIVERSITY OF GOTHENBURG

SPRÅKBANKENTEXT





TalkBank resources for the study of multilingualism

Brian MacWhinney

Carnegie Mellon University

TalkBank Principles

- Focus on spoken language in its MANY forms
- 14 subject areas: 5 multilingualism areas, 6 clinical areas, 5 child language areas
- Consistent transcription format – CHAT https://talkbank.org/manuals/CHAT.pdf
- Format linked to analysis programs (CLAN), database search (TalkBankDB), and transcript browser (TalkBankBrowser)
- Interoperability with other formats: SALT, Elan, Praat, R
- Version control and replicability through TalkBankDB
- Open access – open Science – **over 8000 published articles**
- Survivability – CMU Library, Containerization, Governing Board

Ground Rules

- Ethics of data collection
- Ethics of data-sharing
- Documentation
- Policies, including GDPR compliance
- All inside the Ground Rules

Interoperability Examples

- Play cinco.cha in CLAN
- CHAT2ELAN +emp4 cinco.cha
- Play cinco.c2elan.eaf in ELAN
- ELAN2CHAT *.eaf – open output
- CHAT2PHON
  - Open File (or folder) in Phon
  - Run auto-segmenter (View- IPALookUp – Autotranscribe)
  - Fix “actual” tier to reflect variation from standard

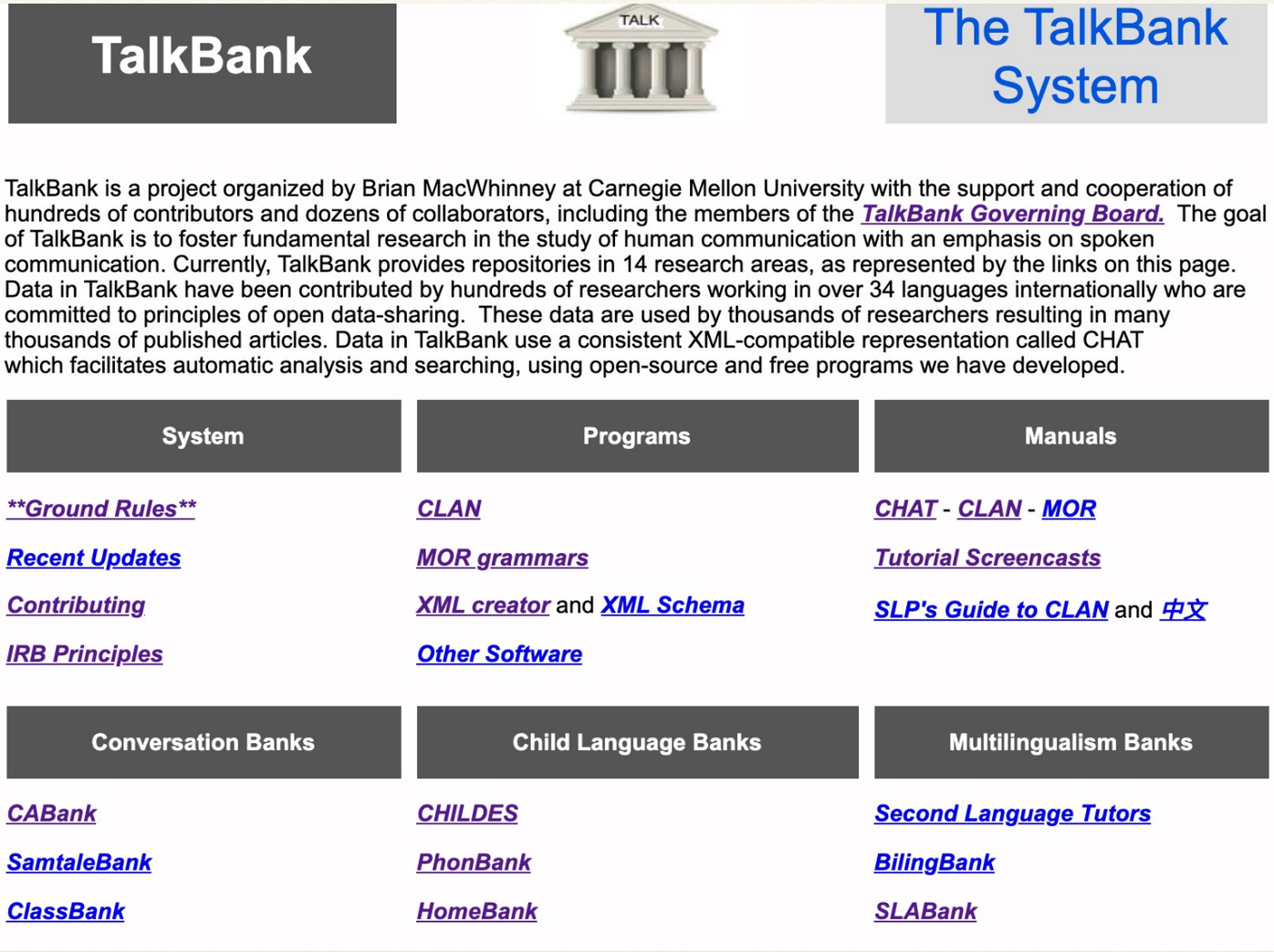
TalkBank Website

https://talkbank.org/

Visit each databank!

Explore the links!

Tutorials



TalkBank

The TalkBank System

TalkBank is a project organized by Brian MacWhinney at Carnegie Mellon University with the support and cooperation of hundreds of contributors and dozens of collaborators, including the members of the ***TalkBank Governing Board.*** The goal of TalkBank is to foster fundamental research in the study of human communication with an emphasis on spoken communication. Currently, TalkBank provides repositories in 14 research areas, as represented by the links on this page. Data in TalkBank have been contributed by hundreds of researchers working in over 34 languages internationally who are committed to principles of open data-sharing. These data are used by thousands of researchers resulting in many thousands of published articles. Data in TalkBank use a consistent XML-compatible representation called CHAT which facilitates automatic analysis and searching, using open-source and free programs we have developed.

System	Programs	Manuals
Ground Rules*	*CLAN*	*CHAT* - *CLAN* - *MOR*
Recent Updates	*MOR grammars*	*Tutorial Screencasts*
Contributing	*XML creator* and *XML Schema*	*SLP's Guide to CLAN* and *中文*
IRB Principles	*Other Software*	

Conversation Banks	Child Language Banks	Multilingualism Banks
CABank	*CHILDES*	*Second Language Tutors*
SamtaleBank	*PhonBank*	*BilingBank*
ClassBank	*HomeBank*	*SLABank*

5 Multilingualism Banks

- All reachable from https://talkbank.org
- Transcripts for everything, most linked to audio/video
- https://slabank.talkbank.org/ – adult L2 corpora
- https://biling.talkbank.org – adult code-switching
- https://sla.talkbank.org/ – adult eCALL methods
- https://childes.talkbank.org – childhood bilinguals
- https://phonbank.talkbank.org – child bilingual phonology

6 Language Disorder Banks

- https://aphasia.talkbank.org aphasia, apraxia
- https://dementia.talkbank.org dementia
- https://fluency.talkbank.org/ child stuttering
- https://rhd.talkbank.org right hemisphere damage
- https://tbi.talkbank.org traumatic brain injury
- https://childes.talkbank.org (Clinical-MOR)

Other Banks

- https://ca.talkbank.org Conversation Analysis
- https://class.talkbank.org Classroom Discourse
- In Progress or Possible
  - PsychBank
  - VoiceBank
  - GestureBank

Web Resources

- Browsable Database: https://slabank.talkbank.org
- Downloadable CLAN program and manuals — all free
- Tutorials at https://talkbank.org/screencasts
- YouTube examples and Grand Rounds https://childes.talkbank.org/topics/youtube.html (good for student projects)
- TalkBankDB at https://talkbank.org/DB (also good for student projects)

Browsable Database Example

Start at https://talkbank.org

Go to slabank

Select Browsable Database

Read instructions (once)

Pick a language: French

Pick a corpus: FLLOC/LingDev/Loch-11

Pick a transcript: 46*, note icons, info

Go to line 3, play, stop, etc.

Corpus Index Example

- Select **Index to Corpora **
- Open Media Folder, show Chrome Downloader
- Download corpus files
- Pick a file: Spanish – Nebrija CAO 07_26

Access

- All multilingualism banks are open access
- Clinical banks have easy to get passwords
- HomeBank requires membership process
- Contacts: macw@cmu.edu, fromm@andrew.cmu.edu
- Google Groups: info-childes, chibolts, phon, fluencybank https://talkbank.org/share/email.html

Opening in CLAN editor

- sequence.cha as an example
- Continuous playback
- Single utterance playback
- Playback from video
- Changing bullets
- F5 linking

CLAN Commands

- Tour of CLAN interface
  - working, mor lib, progs
    - get MOR grammar
- Opening files with the editor
- Commands screencast at https://talkbank.org/screencasts

CLAN

- freq 21110.cha
- kwal +scow 21110.cha
- hits can be triple-clicked to enter transcript and play
- CLAN manual at https://talkbank.org/manuals/CLAN.pdf

Profiling Commands

- Designed for development and clinical analysis as well as comprehensive research analysis
- Children:
  - Packages: KidEval and FluCalc
  - Scales: IPSyn, DSS
- Aphasics:
  - Package: EVAL
  - Scales: C-NNLA, C-QPA

Profiling for L2

- We need an organized comparison corpus
- Could be used to judge
  - learner level
  - specific error types – syntactic, lexical, phonological
  - discourse analysis of story retell
  - fluency (see Skehan Festscrift)

MOR, POST, GRASP

- %mor is crucial for IPSyn, DSS, etc.
- Grammar List at https://talkbank.org/morgrams
- Chain of MOR, PrePost, POST, PostMortem, GRASP
- MOR/POST for 11 languages
  - select ENG grammar
  - mor 21110.cha
  - triple-click for dependency graph (similar to *Sparv*)

What happened

- MOR's a-rules processed the *.cut files in the /lex folder to create a run-time set of allomorphs
- MOR's c-rules ran through possible continuation using each allomorph, checking for MATCHCAT and other conditions
- Result included many ambiguities. The goal is to eliminate all ambiguity between part of speech. Within part of speech ambiguity is not handled
- PREPOST eliminated some ambiguities
- Then POST did trigram statistical disambiguation
- POSTMORTEM did a final ambiguity elimination
- GRASP used POS information and the megrasp.mod file to create a dependency structure
- Finally, we check the result

A-rule example

RULENAME: n-um

LEX-ENTRY:

LEXCAT = [scat n], [pl a]

LEXSURF = $Xum

ALLO:

ALLOSURF = LEXSURF

ALLOCAT = LEXCAT, ADD [allo n0]

ALLO:

ALLOSURF = $Xa

ALLOCAT = LEXCAT, ADD [allo n0], ADD [gen PL]

C-rule example

%9 deverbal participles, including compounds like baby+sitting

if

STARTCAT = [scat OR cop v], ![gen DUP], ![block part]

NEXTCAT = [scat part]

NEXTSURF = ing

MATCHCAT [allo]

then

RESULTCAT = NEXTCAT, STARTCAT [comp], NEXTCAT [tense], DEL [allo], ADD [allo a0], DEL [bare]

RULEPACKAGE = {adv-ly, adj-deriv, denom-deriv}

Bilingual MOR

- *LRT: +" [- spa] <yo no la> [/] yo no la desmentí porque.
- *LRT: what's my word against hers &ladadada .
- *LRT: +" [- spa] todos estamos con un calor and@s working@s .
- All words are tagged implicity; can be made explicit.
- Coding system makes code-switching junctures evident.
- Run English MOR, excluding [- spa], then Spanish MOR including [- spa]

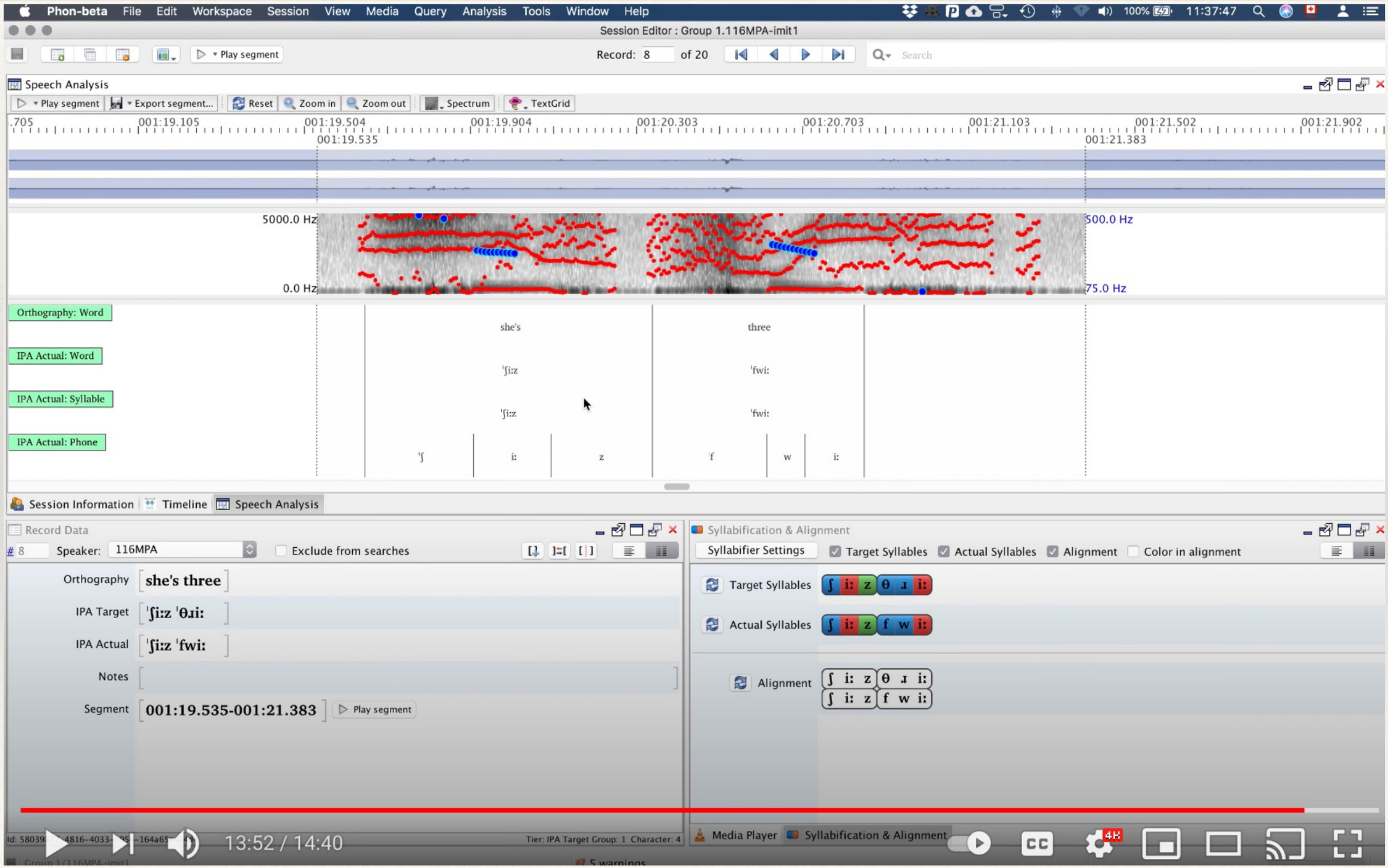
Error Analysis

- [*p] phonological p:w, p:n, p:m
- [* s] semantic s:r, s:ur. s:uk, s:per
- [* n] neologism n:k, n:uk, n:k:s, n:uk:s
- [* d] dysfluency
- [* m] morphology m:a:0es etc.
- [* f] formal lexical
- [+ gram] [+ jar] [+ es] [+ per] [+ cir]
- Adler06a.cha in Grand Rounds

TalkBankDB at https://talkbank.org/DB

- Similar to *Korp* and *Strix*, but with emphasis on versioning and collaborative commentary
- Browse Manual
- Select: CHILDES, Eng-NA, Brown, Eve, all
- Click to go to any transcript
- Tabs for utterances, participants, types, tokens, CQL
- Download all utterances – 26432 rows in Excel
- 3-grams => I want some 39 times
- R API (not yet in a package) to pull directly

Phon at https://phon.ca



https://sla.talkbank.org

- German Wikipedia articles (like *Larka*)
- Pinyin Tutor
- English Preposition Tutor, Article Tutor (CogLing)
- German Gender, Prefixes, Case
- Virtual Reality
- Captioned Video – CapVid French 09
- Language Learning in the Wild
  - snakke svenska med mig
  - Language Tours
  - Icelandic Village

SweLL/CLARIN/TalkBank

- Data-Sharing Issue
- Software sharing, interoperability – Korp
- CLARIN – Federated Content Search
- CLARIN – SSO authentication

Integration

- Putting all of SLABank or TalkBank into Språkbanken
- But SLABank doesn't have corrections
- What format is required for Korp
- What about AntConc

In Summary

- TalkBank has a lot of data in a consistent format
- TalkBank has many powerful methods for data analysis
- The five multilingualism data banks provide rich data on spoken language learning and usage
- It is good to be aware of what is available, but
- You don’t need to learn everything – just what you need for the work you are doing

Thanks from Andrew, Franklin, John, Davida, Leonid, and Brian





