Help from DELAD and the CLARIN Knowledge Centre for Atypical Communication Expertise (ACE) in sharing CSD

Henk van den Heuvel, CLST Radboud University

DELAD-CLARIN Workshop 27-28 January 2021

DELAD initiative

- Initiative to collect and share corpora of speech with disorders (CSD):
- http://delad.net/

- Partners are a mix of researchers, infrastructure specialists, legal experts

- DELAD organises annual workshops since 2015 where these groups convene

- Since 2017 under CLARIN header and support

Topics addressed:

- Examples of CSD
- Guidelines for collecting and sharing CSD
- Ethics and legal aspects
- Levels of anonymisation
- Layered access of data
- Integration of CSD in the CLARIN infrastructure
- Formats
- Relevant metadata

DELAD Background

- Two DELAD workshops in Linköping, Sweden, funded by Riksbankens JF
  - Organised by Martin Ball & Nicole Müller
  - Partners from Europe, USA, Canada
- October 2015
  - What do partners have in terms of CDS
  - Can this be shared with others
- June 2016:
  - First connections to CLARIN infrastructure incl CMU/ CHILDES & Talkbank
  - Funding options: EU infra, Digging into data, VW Stiftung
- First CLARIN-DELAD workshop, Cork, 15-17 Nov. 2017
  - Type I workshop
  - New researchers, new data
  - Upcoming GDPR
  - Requirements CLARIN infrastructure
  - Funding options

Issues:

- IPR/ ethics
- Formats
- Annotations / transcriptions
- Other research data
- Metadata
- Levels of anonimization
- Levels of public access
- Maintenance
- One portal

DELAD Background

- Second CLARIN-DELAD Workshop, Utrecht NL, SURF premises, 28-29 Jan. 2019
  - Type II workshop
  - Involve further research partners to DELAD initiative
  - Ethical & Legal aspects of CDS sharing (CLIC members)
  - Specs & priorities of a CDS centered webportal in CLARIN context
  - Possible role of Talkbank
  - How to best spend 3 PM ICT developer
- Outcomes:
  - Reaffirmation that CLARIN is the Data Trust, to provide the data fence around CDS.
  - DELAD should apply to become a Task Force within CLARIN focusing on practical issues on sharing CDS.
  - As a result, the DELAD website should be updated with a CLARIN flag and contain relevant guidelines for collecting, sharing and storing CDS.
  - Talkbank is seen as a good CLARIN site to host CDS, especially if a European storage cloud and stricter access policy can be realised.
  - Collaboration DELAD, ACE, TLA, Talkbank in data curation
  - Work together on contributions for the CLARIN AC 2019 in Leipzig.

CLARIN in six bullets

- **CLARIN** is the Common Language Resources and Technology Infrastructure
- **ESFRI** ERIC status since 2012, Landmark since 2016
- that provides easy and sustainable access for scholars in the **humanities and social sciences** and beyond
- to **digital language data** (in written, spoken, video or multimodal form)
- and **advanced tools** to discover, explore, exploit, annotate, analyse or combine them, wherever they are located
- through a **single sign-on** online environment

CLARIN Centres

24+3 countries

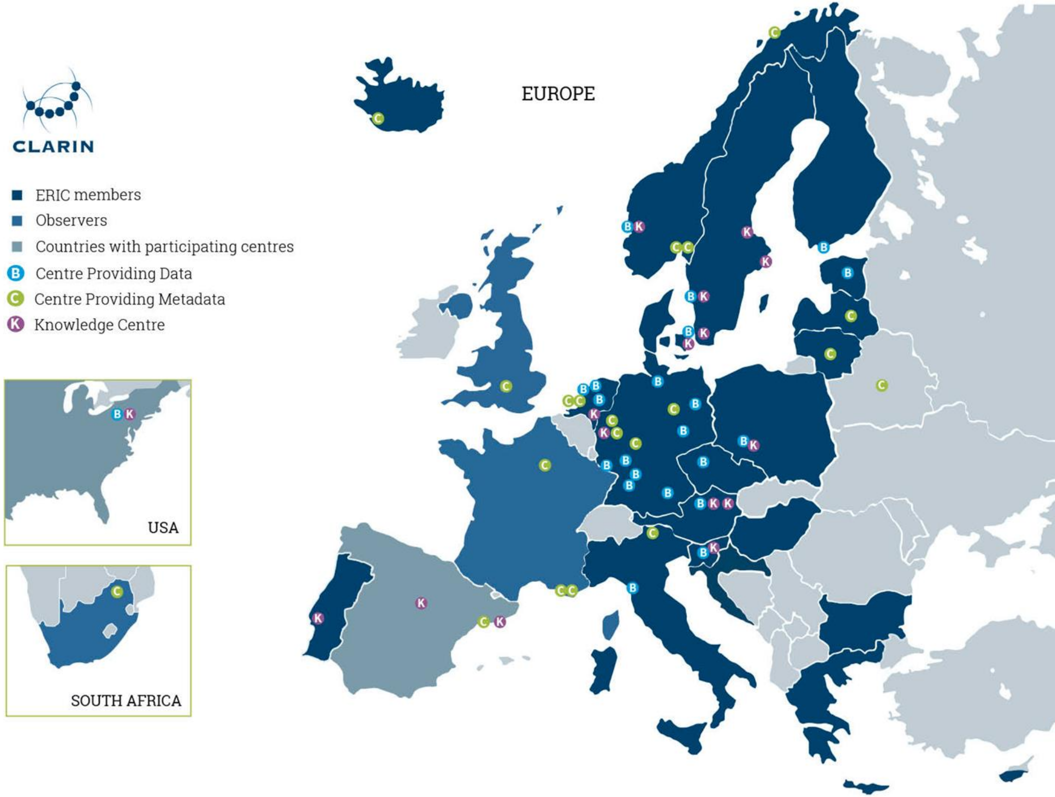
24 B-centres

K-centres

- increased coverage of topics
- inclusion in Tour de CLARIN

Website as channel for multiple audiences.

www.clarin.eu/covid-19



DELAD initiative

- Collaboration with CLARIN Knowledge Centre for Atypical Communication Expertise (ACE)

- For data storage hosting and sharing DELAD cooperates with ACE:
  - The Language Archive at MPI Nijmegen: https://archive.mpi.nl/tla/
  - Talkbank at CMU: https://talkbank.org/

- Use case: https://phonbank.talkbank.org/access/Clinical/PCSC.html

A CLARIN Knowledge Centre for Atypical Communication Expertise

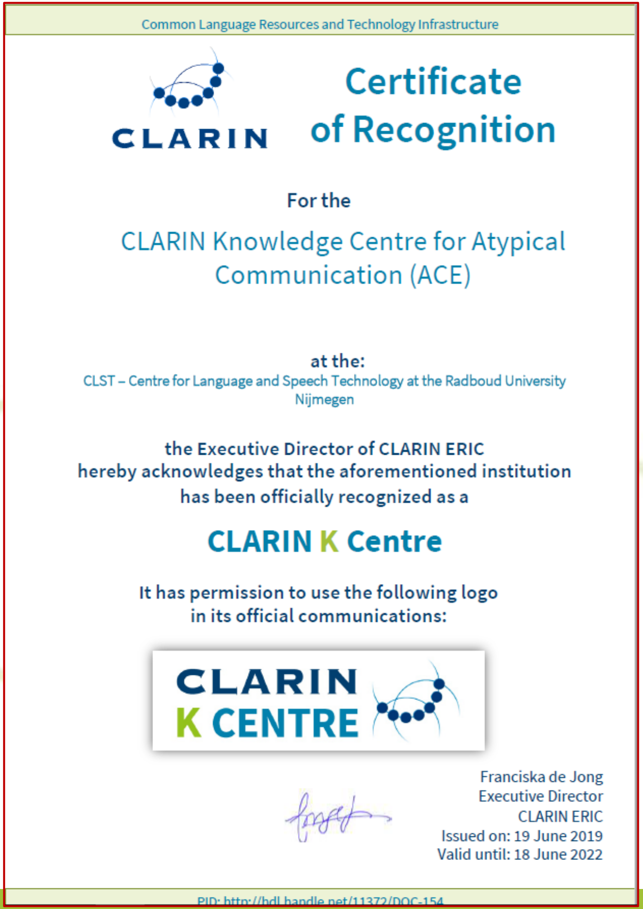
- In short: K-ACE
- Hosted at CLST, Radboud University
- Launch in November 2019

- Description:
  - Atypical communication encompasses language and speech as encountered during (second) language acquisition and development, and in language disorders, but also more broadly in bilingual language development and in sign language. Our centre is specialised in this type of research and concomitant infrastructural issues related to data acquisition, processing and sharing, which is typically highly characterised by sensitivity issues.

Close cooperation with DELAD group

Website: https://ace.ruhosting.nl/

DELAD: http://DELAD.net



Common Language Resources and Technology Infrastructure

CLARIN

Certificate of Recognition

For the

CLARIN Knowledge Centre for Atypical Communication (ACE)

at the:

CLST – Centre for Language and Speech Technology at the Radboud University Nijmegen

the Executive Director of CLARIN ERIC
hereby acknowledges that the aforementioned institution
has been officially recognized as a

CLARIN K Centre

It has permission to use the following logo
in its official communications:

CLARIN K CENTRE

Franciska de Jong
Executive Director
CLARIN ERIC
Issued on: 19 June 2019
Valid until: 18 June 2022

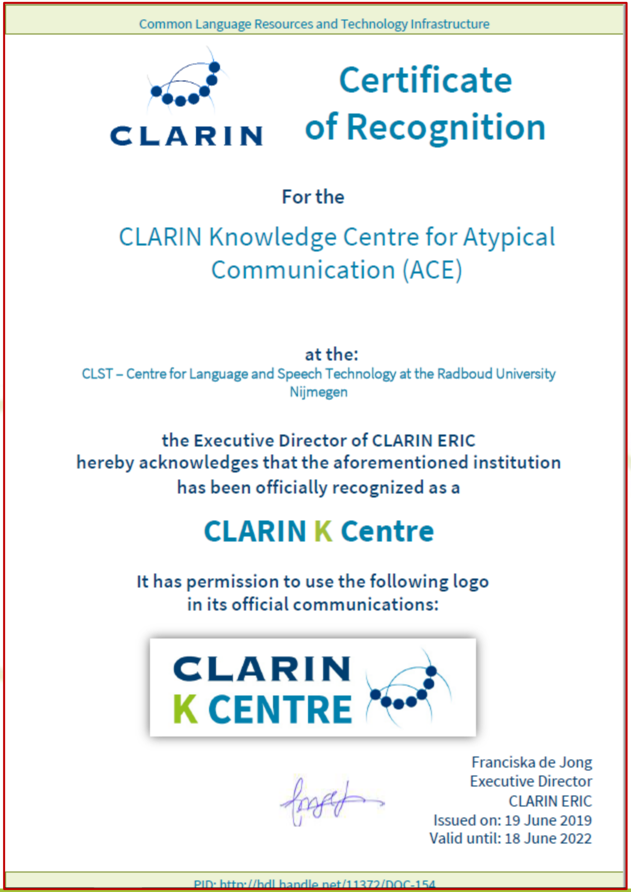
PID: http://hdl.handle.net/11372/DOC-154

A CLARIN Knowledge Centre for Atypical Communication Expertise

- Disciplines of target audience:
  - linguists
  - psychologists
  - neuroscientists
  - computer scientists
  - speech and language therapists
  - education specialists

- Services offered:
  - A website

    with information and guidelines about:
    - consent (forms)
    - hosting corpora and datasets containing atypical communication
    - where to find corpora and datasets containing atypical communication
    - including FAQ
  - Helpdesk/consultancy for questions on these topics
  - Technical assistance for designing, creating, annotating, formatting and metadating these resources
  - Dissemination in the form of flyers, presentations, workshop contributions etc.



Common Language Resources and Technology Infrastructure

CLARIN

Certificate of Recognition

For the

CLARIN Knowledge Centre for Atypical Communication (ACE)

at the:

CLST – Centre for Language and Speech Technology at the Radboud University Nijmegen

the Executive Director of CLARIN ERIC
hereby acknowledges that the aforementioned institution
has been officially recognized as a

CLARIN K Centre

It has permission to use the following logo
in its official communications:

CLARIN K CENTRE

Franciska de Jong
Executive Director
CLARIN ERIC
Issued on: 19 June 2019
Valid until: 18 June 2022

PID: http://hdl.handle.net/11372/DOC-154

A CLARIN Knowledge Centre for Atypical Communication Expertise

- Collaboration with **TLA, MPI Nijmegen**

- CLST will act as a knowledge centre for researchers wishing to create and deposit collections on atypical communication, and wanting advice on where to find such resources
- TLA as a CLARIN B centre, will provide the archive. TLA is specialised in hosting the type of video and audio data that is collected in the atypical communication context that CLST as K-centre will address

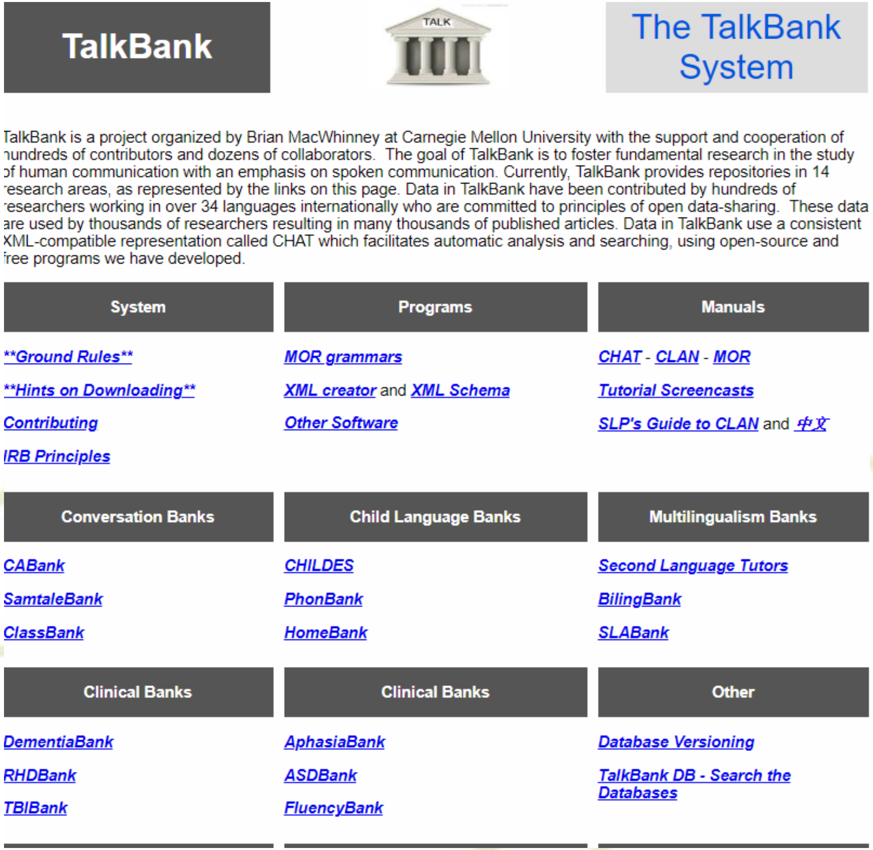
It is relevant to note that the TLA encourages to store video and speech resources not only from the Netherlands but also from other countries and in other languages. This makes the cooperation of CLST with TLA in this field very relevant for CLARIN and the DELAD community.

A CLARIN Knowledge Centre for Atypical Communication Expertise

- Collaboration with **Talkbank / Clinical Banks, CMU**
  - Brian MacWhinney

- Large visibility in communities for LA research and SLI

- Has everything in place for hosting datasets:
  - Curation of data
  - Generate metadatafiles
  - Landing pages for datasets
  - Search through (open) datasets
  - Tools for CHAT transcripts
  - Part of CLARIN (B Centre)
  - Cooperation with TLA (mirror site)
    - Allows storage at TLA only
    - Using their (safer) authentication methods



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. 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	MOR grammars	CHAT - CLAN - MOR
Hints on Downloading	XML creator and XML Schema	Tutorial Screencasts
Contributing	Other Software	SLP's Guide to CLAN and 中文
IRB Principles		

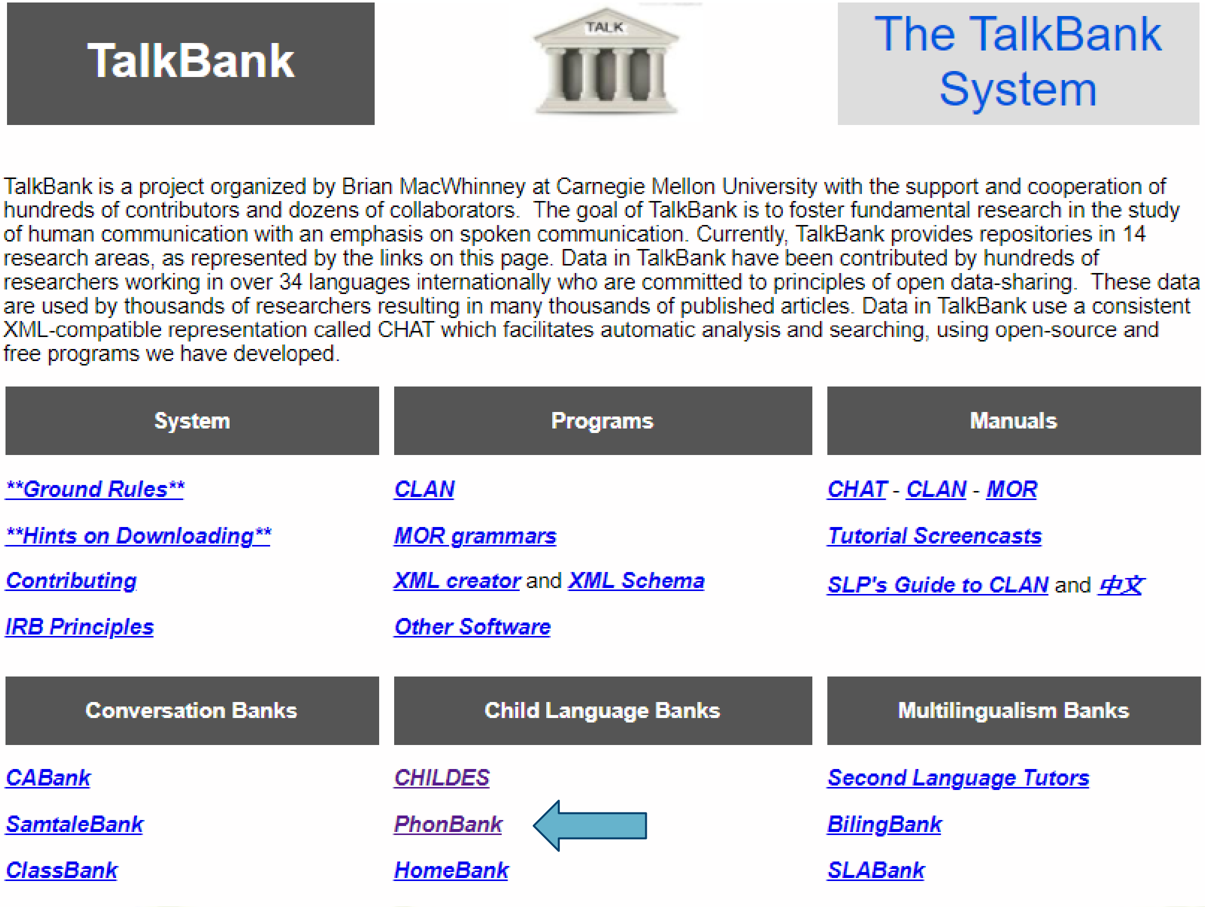
Conversation Banks	Child Language Banks	Multilingualism Banks
CABank	CHILDES	Second Language Tutors
SamtaleBank	PhonBank	BilingBank
ClassBank	HomeBank	SLABank

Clinical Banks	Clinical Banks	Other
DementiaBank	AphasiaBank	Database Versioning
RHDBank	ASDBank	TalkBank DB - Search the Databases
TBIBank	FluencyBank	

Example for a curation within DELAD via ACE

- Polish Cued Speech Corpus of Hearing Impaired Children (A. Lorenc)
- Legacy data involving hearing impaired children from Poland
- DELAD initiative to make thses accessible through Talkbank/Phonbank and The Language Archive (raw data)

- Next five slides from Katarzyna Klessa



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. 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

- **Ground Rules**
- **Hints on Downloading**
- Contributing
- IRB Principles

Programs

- CLAN
- MOR grammars
- XML creator and XML Schema
- Other Software

Manuals

- CHAT - CLAN - MOR
- Tutorial Screencasts
- SLP's Guide to CLAN and 中文

Conversation Banks

- CABank
- SamtaleBank
- ClassBank

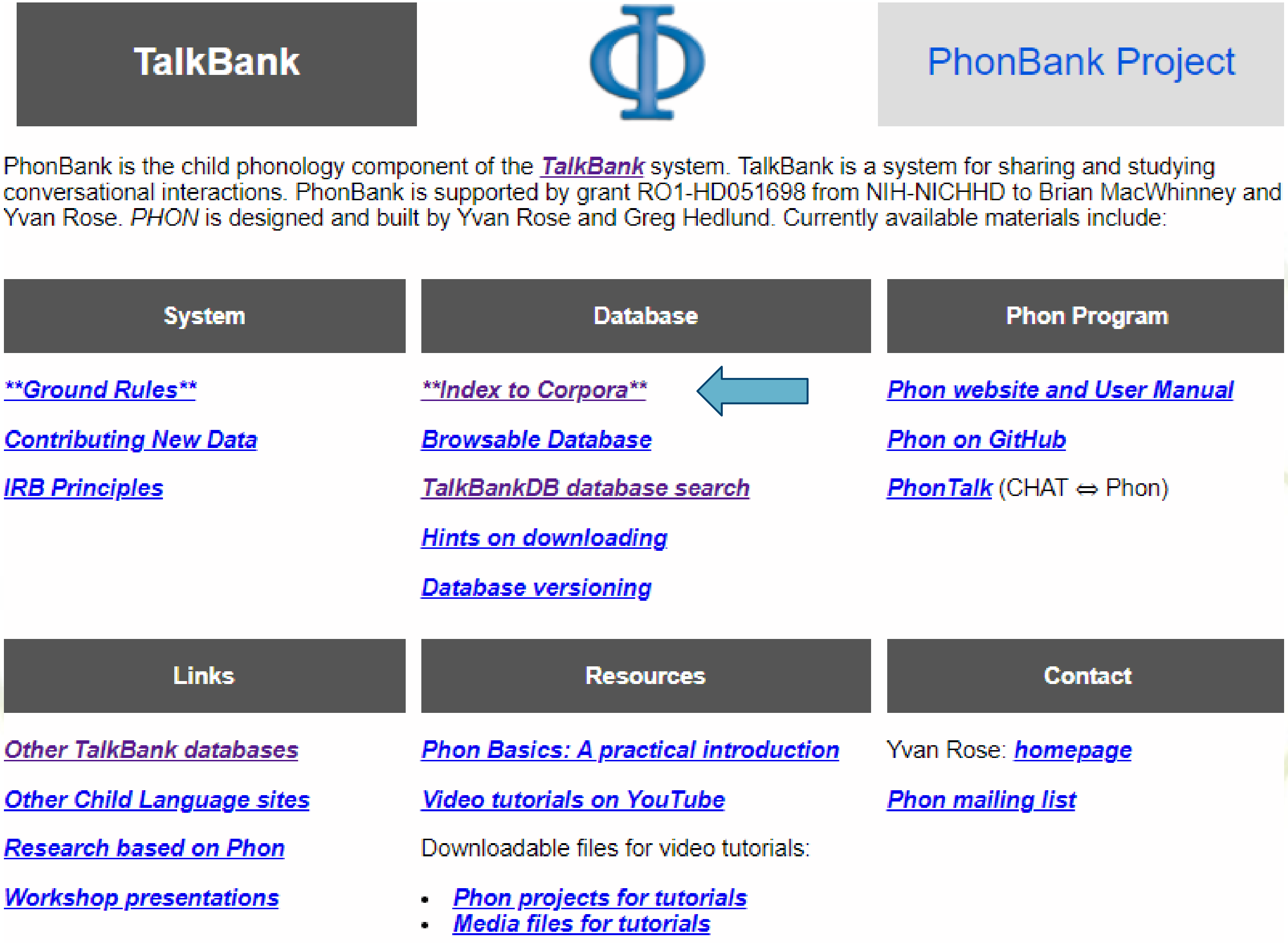
Child Language Banks

- CHILDES
- PhonBank
- HomeBank

Multilingualism Banks

- Second Language Tutors
- BilingBank
- SLABank

https://talkbank.org/



TalkBank

PhonBank Project

PhonBank is the child phonology component of the TalkBank system. TalkBank is a system for sharing and studying conversational interactions. PhonBank is supported by grant RO1-HD051698 from NIH-NICHHD to Brian MacWhinney and Yvan Rose. *PHON* is designed and built by Yvan Rose and Greg Hedlund. Currently available materials include:

System

Ground Rules

Contributing New Data

IRB Principles

Database

Index to Corpora

Browsable Database

TalkBankDB database search

Hints on downloading

Database versioning

Phon Program

Phon website and User Manual

Phon on GitHub

PhonTalk (CHAT ⇔ Phon)

Links

Other TalkBank databases

Other Child Language sites

Research based on Phon

Workshop presentations

Resources

Phon Basics: A practical introduction

Video tutorials on YouTube

Downloadable files for video tutorials:

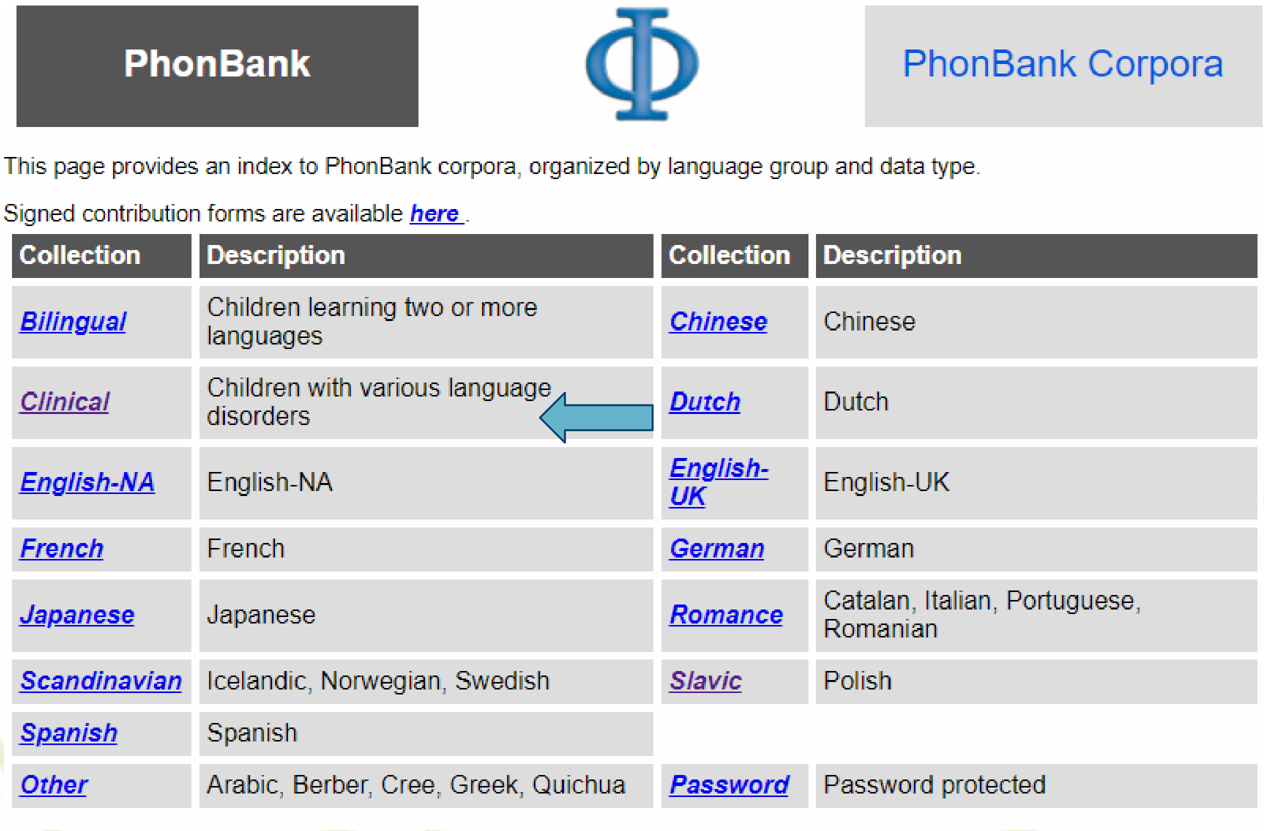
- Phon projects for tutorials
- Media files for tutorials

Contact

Yvan Rose: homepage

Phon mailing list

https://phonbank.talkbank.org/



PhonBank

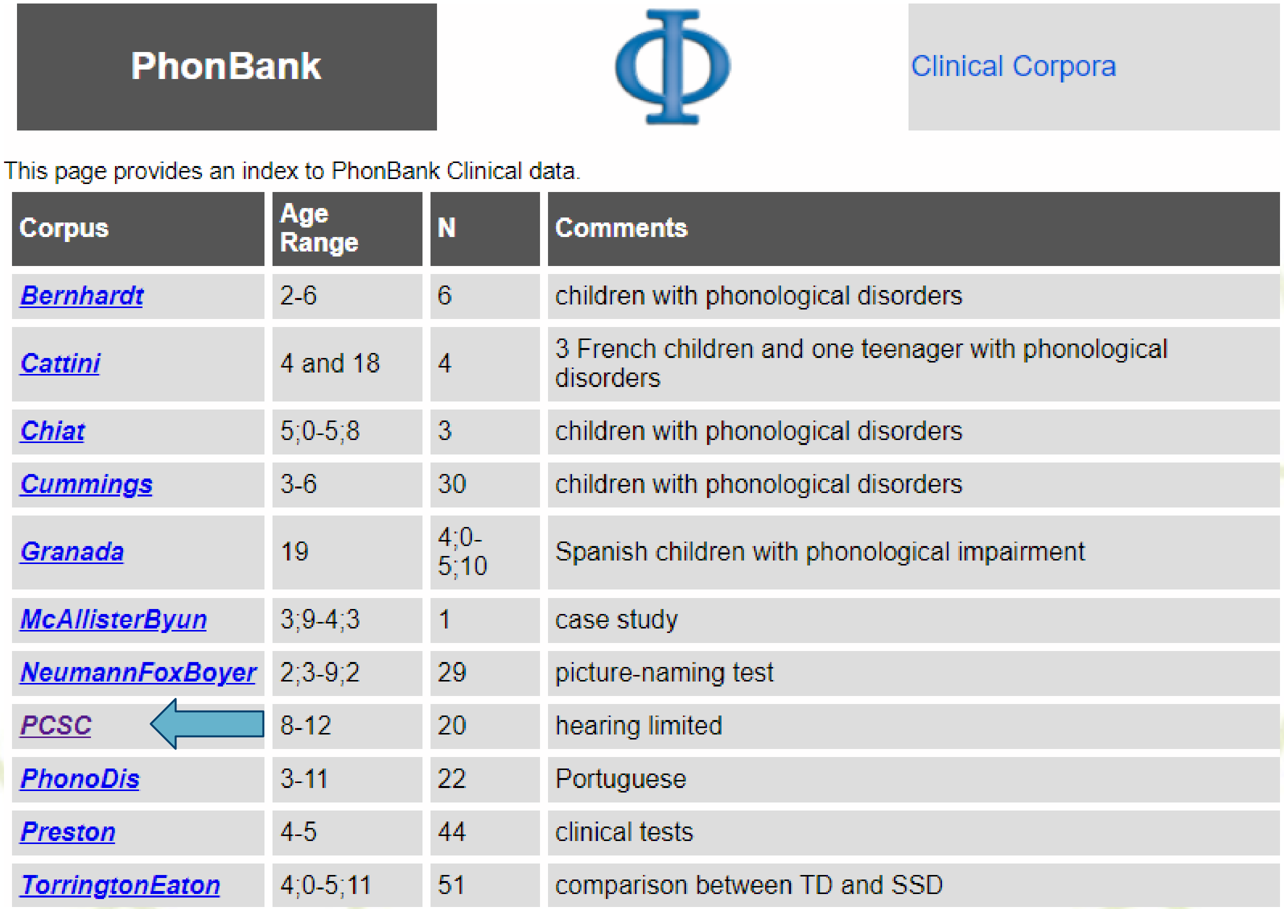
PhonBank Corpora

This page provides an index to PhonBank corpora, organized by language group and data type.

Signed contribution forms are available *here* .

Collection	Description	Collection	Description
Bilingual	Children learning two or more languages	*Chinese*	Chinese
Clinical	Children with various language disorders	*Dutch*	Dutch
English-NA	English-NA	*English-UK*	English-UK
French	French	*German*	German
Japanese	Japanese	*Romance*	Catalan, Italian, Portuguese, Romanian
Scandinavian	Icelandic, Norwegian, Swedish	*Slavic*	Polish
Spanish	Spanish		
Other	Arabic, Berber, Cree, Greek, Quichua	*Password*	Password protected

https://phonbank.talkbank.org/access/



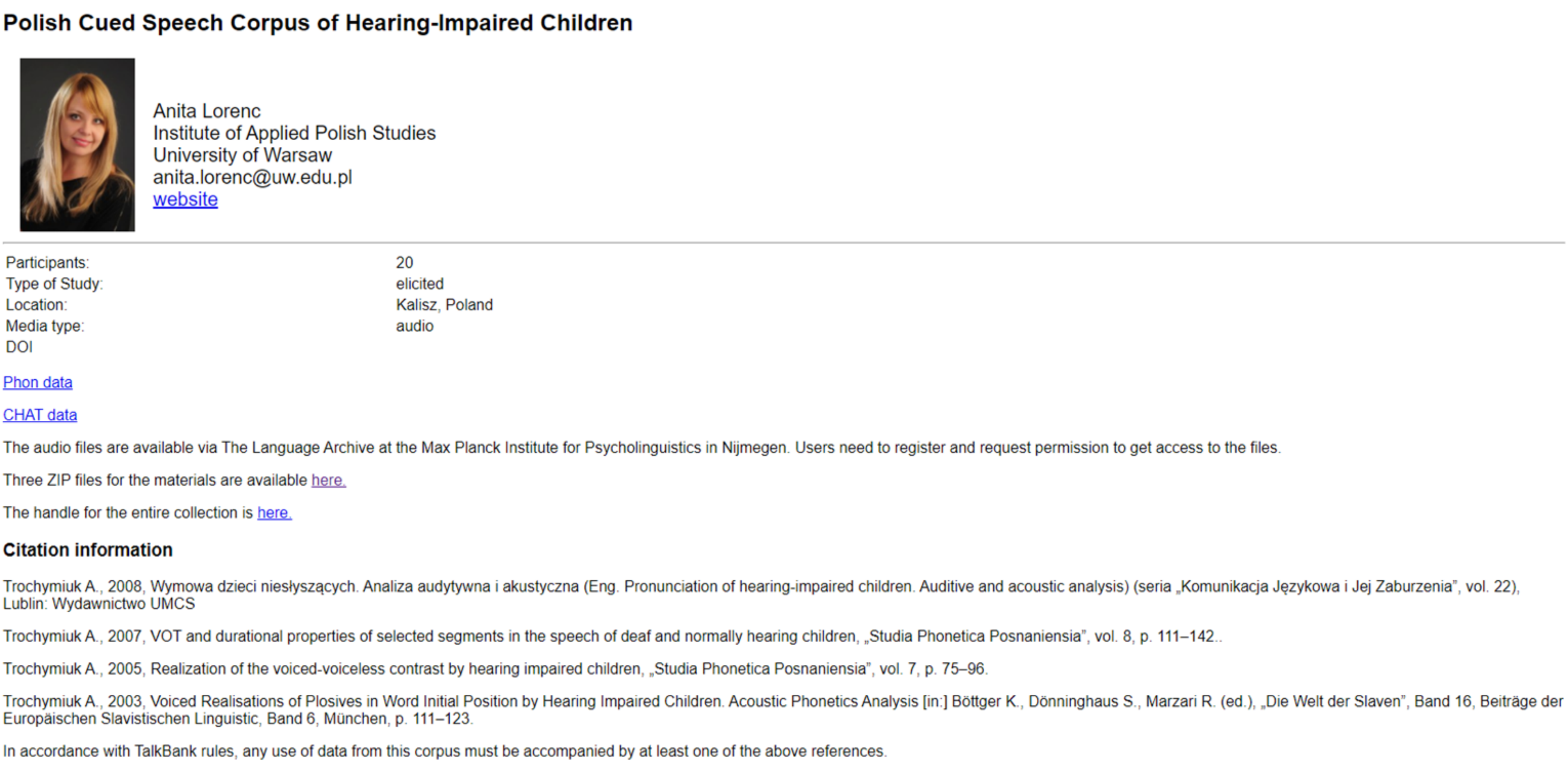
PhonBank

Clinical Corpora

This page provides an index to PhonBank Clinical data.

Corpus	Age Range	N	Comments
Bernhardt	2-6	6	children with phonological disorders
Cattini	4 and 18	4	3 French children and one teenager with phonological disorders
Chiat	5;0-5;8	3	children with phonological disorders
Cummings	3-6	30	children with phonological disorders
Granada	19	4;0-5;10	Spanish children with phonological impairment
McAllisterByun	3;9-4;3	1	case study
NeumannFoxBoyer	2;3-9;2	29	picture-naming test
PCSC	8-12	20	hearing limited
PhonoDis	3-11	22	Portuguese
Preston	4-5	44	clinical tests
TorringtonEaton	4;0-5;11	51	comparison between TD and SSD

https://phonbank.talkbank.org/access/Clinical/



Polish Cued Speech Corpus of Hearing-Impaired Children

Anita Lorenc
Institute of Applied Polish Studies
University of Warsaw
anita.lorenc@uw.edu.pl
website

Participants:	20
Type of Study:	elicited
Location:	Kalisz, Poland
Media type:	audio
DOI	

Phon data

CHAT data

The audio files are available via The Language Archive at the Max Planck Institute for Psycholinguistics in Nijmegen. Users need to register and request permission to get access to the files.

Three ZIP files for the materials are available here.

The handle for the entire collection is here.

Citation information

Trochymiuk A., 2008, Wymowa dzieci niesłyszących. Analiza audytywna i akustyczna (Eng. Pronunciation of hearing-impaired children. Auditive and acoustic analysis) (seria „Komunikacja Językowa i Jej Zaburzenia", vol. 22), Lublin: Wydawnictwo UMCS

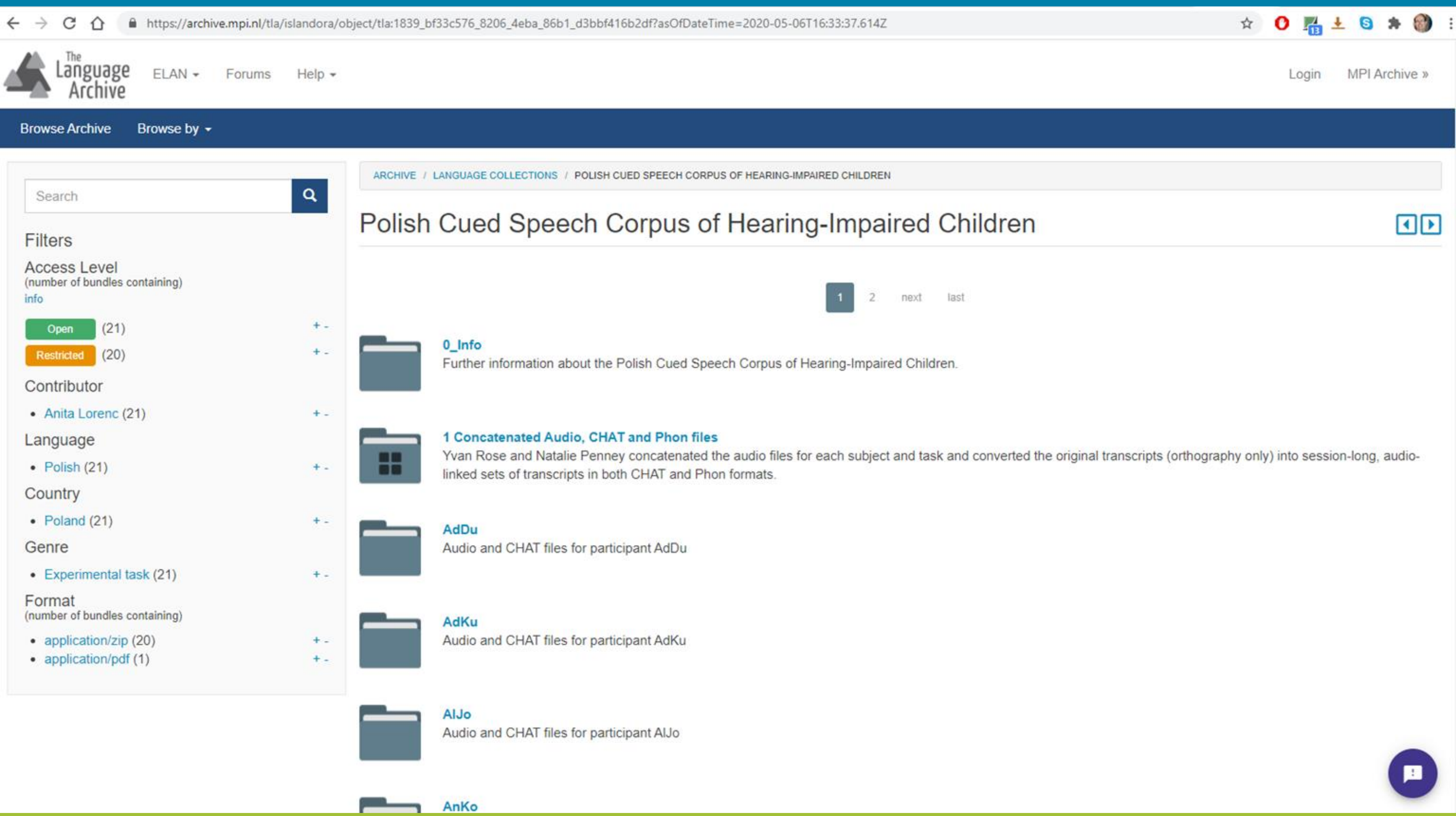
Trochymiuk A., 2007, VOT and durational properties of selected segments in the speech of deaf and normally hearing children, „Studia Phonetica Posnaniensia", vol. 8, p. 111–142..

Trochymiuk A., 2005, Realization of the voiced-voiceless contrast by hearing impaired children, „Studia Phonetica Posnaniensia", vol. 7, p. 75–96.

Trochymiuk A., 2003, Voiced Realisations of Plosives in Word Initial Position by Hearing Impaired Children. Acoustic Phonetics Analysis [in:] Böttger K., Dönninghaus S., Marzari R. (ed.), „Die Welt der Slaven", Band 16, Beiträge der Europäischen Slavistischen Linguistic, Band 6, München, p. 111–123.

In accordance with TalkBank rules, any use of data from this corpus must be accompanied by at least one of the above references.

https://phonbank.talkbank.org/access/Clinical/PCSC.html



https://archive.mpi.nl/tla/islandora/object/tla:1839_bf33c576_8206_4eba_86b1_d3bbf416b2df?asOfDateTime=2020-05-06T16:33:37.614Z
The Language Archive
ELAN
Forums
Help
Login
MPI Archive »
Browse Archive
Browse by
Search
1
2
next
last

ARCHIVE / LANGUAGE COLLECTIONS / POLISH CUED SPEECH CORPUS OF HEARING-IMPAIRED CHILDREN

Polish Cued Speech Corpus of Hearing-Impaired Children

Filters

Access Level
(number of bundles containing)
info

- Open (21)
- Restricted (20)

Contributor

- Anita Lorenc (21)

Language

- Polish (21)

Country

- Poland (21)

Genre

- Experimental task (21)

Format
(number of bundles containing)

- application/zip (20)
- application/pdf (1)

0_Info
Further information about the Polish Cued Speech Corpus of Hearing-Impaired Children.

1 Concatenated Audio, CHAT and Phon files
Yvan Rose and Natalie Penney concatenated the audio files for each subject and task and converted the original transcripts (orthography only) into session-long, audio-linked sets of transcripts in both CHAT and Phon formats.

AdDu
Audio and CHAT files for participant AdDu

AdKu
Audio and CHAT files for participant AdKu

AlJo
Audio and CHAT files for participant AlJo

AnKo

Recent developments

- Website revised: http://delad.net
- Polish corpus curated
- Blogs Tour de CLARIN
  - Blog: https://www.clarin.eu/blog/tour-de-clarin-knowledge-centre-atypical-communication-expertise
  - https://www.clarin.eu/blog/tour-de-clarin-interview-katarzyna-klessa-and-anita-lorenc
- LREC papers:
  - Van den Heuvel, H., Kelli, A., Klessa, K., Salaasti, S. (2020) Corpora of disordered speech in the light of the GDPR: Two use cases from the DELAD initiative. Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC2020). pp. 3310-3314, http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.406.pdf
  - Van den Heuvel, H., Oostdijk, N., Rowland, C., Trilsbeek, P. (2020) The CLARIN Knowledge Centre for Atypical Communication Expertise. Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC2020). pp. 3305-3309, http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.405.pdf

Recent developments

- CLARIN Annual Conference, Bazaar, 1 October 2019:
  - Van den Heuvel, H. *DELAD: How to share and access sensitive datasets with language and speech disorders via CLARIN.* https://www.clarin.eu/clarin-annual-conference-2019-bazaar
- CLARIN Annual Conference, Bazaar, 7 October 2020:
  - Van den Heuvel, H. & Hoorn, E. *Sharing corpora of disordered speech and finding relevant use cases*. https://www.clarin.eu/content/clarin-bazaar-2020
- SSHOC webinar, 14 October 2020:
  - Van den Heuvel, H., Bessell, N., Trilsbeek, P., Bishop, E. & K. Klessa. (2020, October). SSHOC Workshop: *Sharing Datasets of Pathological Speech* (Version v1.0). Zenodo. http://doi.org/10.5281/zenodo.4081602

- Upcoming:
  - Lee, A., Bessell, N., Van den Heuvel, H., Saalasti, S., Klessa, K., Müller, N., & M.J. Ball. *The latest development of the DELAD project for sharing corpora of disordered speech*. Accepted for publication in *Clinical Linguistics & Phonetics*

- *What about Remote Secure Access?*