

SSHOC Workshop

Sharing Datasets of Pathological Speech



SPEAKERS



Henk van den Heuvel

CLST, Radboud University, Nijmegen, The Netherlands

- Researcher
- Director of the Centre for Language and Speech Technology
- Head of the Humanities Lab and research data officer at the Faculty of Arts
- Coordinating member of the CLARIN Knowledge Centre for Atypical Communication
- DELAD Committee Member



Nicola Bessell

Department of Speech and Hearing Sciences, University College Cork, Cork, Ireland

- Lecturer in Speech and Hearing Sciences
- Research in typical and atypical phonetics and phonology, dialectology, language change
- Active in collection of language corpora for typical and atypical speech
- UCC Social Research Ethics Committee member
- DELAD Committee Member



SPEAKERS



Paul Trilsbeek

The Language Archive, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

- Archivist
- Head of The Language Archive
- Member of the CoreTrustSeal Boardcv



Libby Bishop

GESIS-Leibniz Institute for Social Sciences, Cologne, Germany

- Coordinator for International Data Infrastructures
- Manages relationships between GESIS and international data infrastructures
- Social Sciences and Humanities Open Science Cloud Project – Leader on Task 5.4 Remote Access to Secure Data
- CAVA: Human Communication: an Audio-Visual Archive–UCL–Advisory Group (c. 2009)

SPEAKERS



Katarzyna Klessa

Adam Mickiewicz University in Poznan, Poland

- Researcher
- Member of the Programme Board of CLARIN-PL
- Head of the project: Infant-directed and adult-directed speech: preliminary investigation with Carstens AG501 Articulograph (Polish National Science Center)
- Visiting scholar at the University of Texas at Austin, USA within the COLING project (Minority Languages, Major Opportunities. Collaborative Research, Community Engagement and Innovative Educational Tools; EU Horizon 2020)
- Participant of the project MuMoStance: Multimodal Stancetaking: Expressive Movement and Affective Stance. Political Debates in German Bundestag and Polish Sejm

NOTES



-  **This webinar is being recorded.** All participants will receive a link to the recording later.
-  **Slides are available:** See the chat box for the link.
-  **Questions?** Put them in the chat box. We'll put questions to the speakers at the end of the webinar.

Project:



SSHOC

social sciences & humanities open cloud



Horizon 2020
European Union Funding
for Research & Innovation

Type of action & funding:
Research and Innovation action
(INFRAEOSC-04-2018)

Partners: 45

(20 beneficiaries + 25 LTPs)

SSH ESFRI Landmarks and Projects
& international SSH data infrastructures

Project budget:

€ 14,455,594.08

Duration: 40 months

(January 2019 – 30 April 2022)

Project website:

www.SSHopencloud.eu



Objectives:

- creating the social sciences and humanities (**SSH**) part of European Open Science Cloud (**EOSC**)
- maximising **re-use** through **Open Science** and **FAIR** principles (standards, common catalogue, access control, semantic techniques, training)
- interconnecting existing and new infrastructures (clustered cloud infrastructure)
- establishing appropriate **governance model** for SSH-EOSC

EXPECTED IMPACT



The Social Sciences and Humanities are seamlessly integrated in the European Open Science Cloud



Availability of an EU-wide, easy-to-use SSH Open Marketplace, where tools and data are openly accessible



EU-wide availability of high quality "cloud ready" SSH tools and high quality SSH data



EU-wide availability of trusted and secure access mechanisms for SSH data, conforming to EU legal requirements



State of the art Research Infrastructure in several pilot domains advanced through dedicated SSH data pilots cluster projects



Maximising reuse through Open Science and FAIR principles (standards, common catalogue, access control, semantic techniques, training)

Programme for this webinar

- **Henk van den Heuvel:** [The DELAD initiative](#) for sharing corpora of speech disorders
- **Nicola Bessell:** GDPR & Ethics of special category data
- **Paul Trilsbeek:** Storing and sharing CSD at the The Language Archive
- **Libby Bishop:** Remote Access to sensitive data
- **Libby Bishop:** The CAVA project
- **Katarzyna Klessa:** the case of the [Polish Cued Speech Corpus of Hearing-Impaired Children](#)

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

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

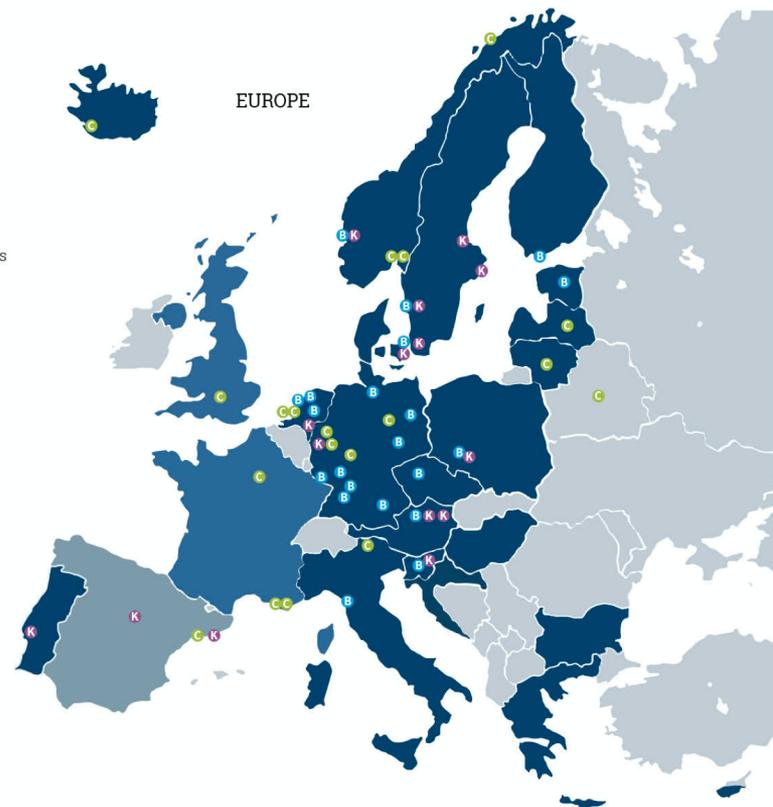
- increased coverage of topics
- inclusion in Tour de CLARIN

Website as channel for multiple audiences.

www.clarin.eu/covid-19



- ERIC members
- Observers
- Countries with participating centres
- B Centre Providing Data
- C Centre Providing Metadata
- K Knowledge Centre



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>
- Next DELAD workshop: 25-29 January 2021, online
 - Access options for CSD
 - Space for researchers to present their datasets and solutions/wishes for sharing them
 - GDPR-issues & DPIA for selected cases, role play
 - Contact person: Henk van den Heuvel, h.vandenheuvel@let.ru.nl



Ethics and GDPR considerations when collecting for corpora of speech disorders

Nicola Bessell
Department of Speech and Hearing Sciences
University College Cork
Ireland

A TRADITION OF
INDEPENDENT
THINKING



University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

GDPR applies
to processing
of personal
data, which is

Data about living persons from which
they can be identified including

- Personal identifiers
- Audio recordings
- Video recordings

Lawful basis (Article 6) is:

'... performance of a task in the public
interest.'

What about
health-related
data?
Article 9
Recital 35

...all data pertaining to the health status of a data subject...any information on disease, disability, medical history...independent of its source...

Additional rationale required for processing health-related data. Must be

- necessary for research purposes
- archiving purposes must be in the public interest

Processing health data requires safeguards

Research ethics

Data minimization: collect only the personal data that's necessary for the research purpose

Anonymize or pseudonymize where possible

Ensure security of data handling and storage

Process to ensure participant control of data

Transparency requires clear language and includes

- Informed consent
- Information about the organization, research unit and members
- Specifics of project
- Specifics of use of data (present and potential)
- Details of sharing of data
- Details of storage of data

Fairness

- Right to refuse
- Right to withdraw

Consent forms must address

Use of data

- Seek explicit consent for current purposes and purposes of any reuse of data
- Outline how confidentiality protected

Dissemination or sharing of data

- On what terms
- Include any future use by research teams

Archiving of data (Article 89)

- Archival period can be specified

Example from <https://ukdataservice.ac.uk/manage-data/legal-ethical/consent-data-sharing/audio-visual.aspx>

- I have read and understood the project information and agree to take part in the study.
- I have had the opportunity to ask questions about the study.
- I understand that my taking part is voluntary and I can withdraw from the project whilst it is ongoing.
- I agree to be interviewed and my contributed information can be used in research outputs and publications by the UK Data Archive, by Knowledge Exchange and partners, whereby I may be quoted by name (I can indicate off-the-record information during the interview).

The interviews will be archived at the UK Data Archive and Knowledge Exchange and disseminated so other researchers can reuse this information for research and learning purposes:

- 
- I agree for the audio recording of my interview to be archived and disseminated for reuse
 - I agree for the transcript of my interview to be archived and disseminated for reuse
 - I agree for any photographs of me taken during interview to be archived and disseminated for reuse
 - I agree to be contacted in future by Knowledge Exchange and partners to participate in data sharing promotion events.

Example from

<https://ukdataservice.ac.uk/manage-data/legal-ethical/consent-data-sharing/audio-visual.aspx>

Use of the information I provide beyond this project		
I agree for the data I provide to be archived at the UK Data Archive. ²	<input type="checkbox"/>	<input type="checkbox"/>
I understand that other authenticated researchers will have access to this data only if they agree to preserve the confidentiality of the information as requested in this form.	<input type="checkbox"/>	<input type="checkbox"/>
I understand that other authenticated researchers may use my words in publications, reports, web pages, and other research outputs, only if they agree to preserve the confidentiality of the information as requested in this form.	<input type="checkbox"/>	<input type="checkbox"/>

Local or
national
organisations
and national
laws

National/public health services

Private health services

Charity run services

Associations and advocacy groups

National Data Protection Guidelines

National Health Research
Regulations

Covid 19

Benefits of research (in the public interest) must outweigh risks to individual

Use collection methods other than face-to-face

Institution specific restrictions may apply: research with vulnerable populations may require an amendment request to existing ethics

Contact Data Protection Officer

Storing and sharing corpora of speech disorders at The Language Archive

PAUL TRILSBEEK
THE LANGUAGE ARCHIVE
MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS

The Language Archive

- Exists since the late 90's, repository that contains data collections resulting from various language-related research disciplines, such as language acquisition, psychology of language, documentary linguistics
- Currently holds more than 350 collections covering over 250 different languages from around the world
- CLARIN B-Type centre, holds CoreTrustSeal certification
- Contains the “DOBES Archive”, resulting from the DOBES Documentation of Endangered Languages funding programme
- 64 TLA collections added to the UNESCO Memory of the World register in 2015
- Research data repository and archive of cultural heritage at the same time

Archiving and sharing of (clinical) speech data

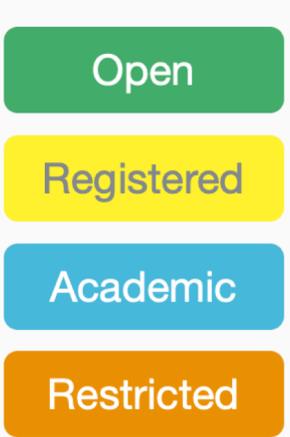
- Anonymisation not possible without invalidating the data for many research purposes -> no transformations of people's voices -> GDPR applies
- GDPR has explicit provisions for “processing” personal data for research purposes (archiving is seen as “processing”)
- Several legal grounds for archiving: informed consent, “public interest”, “legitimate interest” (justification required)
- Agreements needed for archiving and sharing:
 - Deposit / Processing agreement
 - Data use agreement / License (note that many existing licenses are “perpetual” and may therefore be in conflict with the GDPR under certain conditions)

Appropriate technical solution

- “Data protection by design and by default”:
 - Up-to-date systems and software
 - Secure transport of data (HTTPS)
 - Elaborate system of access policies and authorisation
- All archived copies reside within the EU (Netherlands, Germany) at trusted data centres within the Max Planck Society
- Encryption: No strict requirement in the GDPR but mentioned as an appropriate protection measure
 - Conflict: significant risk with respect to long-term preservation, reduces possibilities for working with data in the archive
 - Audio-visual data in TLA is not encrypted “at rest” at the moment
 - Trying to develop a solution that allows us to use encryption while mitigating the issues above

Access policies

- Access levels:
 - Open: Completely public, no registration required
 - Registered: Accessible to any registered user
 - Academic: Accessible to any academic user
 - Restricted: Access permission needs to be requested on an individual basis, depositor typically decides
- Depositor determines appropriate access level
- Different access policies can be defined for different files within a collection, if necessary
 - E.g. anonymous transcriptions could be “open” and audio recordings “restricted”



Open

Registered

Academic

Restricted

Authorisation

- Account registration: validity of a new user is manually verified by archive staff, optionally their academic status as well
- Login via own academic account (“Shibboleth” federated login): account is active immediately and academic status is automatically assigned
- Access to specific restricted materials is granted to individual users once a request is approved
- Access request: Users needs to specify intended usage of the resources

TalkBank exchange

- Arrangement with TalkBank at Carnegie Mellon University in Pittsburgh, U.S.A. for sharing non-sensitive (anonymous/anonymised) parts of collections of “atypical” speech
- Anonymous metadata and transcriptions/annotations are made publicly available in the appropriate part of the TalkBank system
- Both non-sensitive and sensitive files (audio, video) are stored at The Language Archive
- Links to the collections at TLA are provided on the TalkBank collection overview pages

General GDPR considerations

- As with any complex law, there's room for different interpretations. No case law yet for the GDPR in relation to academic data
- Different local implementations make things even more complex in an international setting
- Different approaches:
 - Better safe than sorry: stop sharing. Clearly in conflict with current views on proper scientific conduct (“FAIR” principles)
 - Comply to the best of our ability, knowing that there's a chance that we may not always be 100% compliant, in particular with older pre-GDPR collections
- The incentive for the EC to create the GDPR was not to make it hard or impossible for the scientific community to collect, share and preserve personal data for research purposes

M A X
P L A
N C K

MAX PLANCK INSTITUTE
FOR PSYCHOLINGUISTICS

WWW.MPI.NL

Paul.Trilsbeek@mpi.nl

archive.mpi.nl



SSHOC
social sciences & humanities open cloud

From the CAVA to the Cloud? The quest for remote access to (sensitive) data

Libby Bishop, PhD
SSHOC Webinar: Sharing Datasets of Pathological Speech
14 Sep 2020
virtual



CAVA

Human Communication Audio-Visual Archive

Co-funded by UCL and the JISC (Joint Information Systems Committee)
April 2009 – March 2010

Martin Moyle, UCL Library Services

Merle Mahon, Developmental Science, UCL

Suzanne Beeke, Language and Communication, UCL

CAVA project aims

- establish a digital video repository for human communication sciences, populated with an existing body of rights-cleared content (naturally occurring speech) owned by UCL researchers
- house this within the UCL Library Services Digital Collections service which uses the Ex Libris DigiTool repository platform
- catalogue each video to a discipline-specific descriptive standard, IMDI
- deposit transcripts and other supporting material wherever available
- develop procedures and processes for managing access (restricted to bona fide researchers)
- look at options for long-term digital preservation of the master files, with help of UK Data Archive

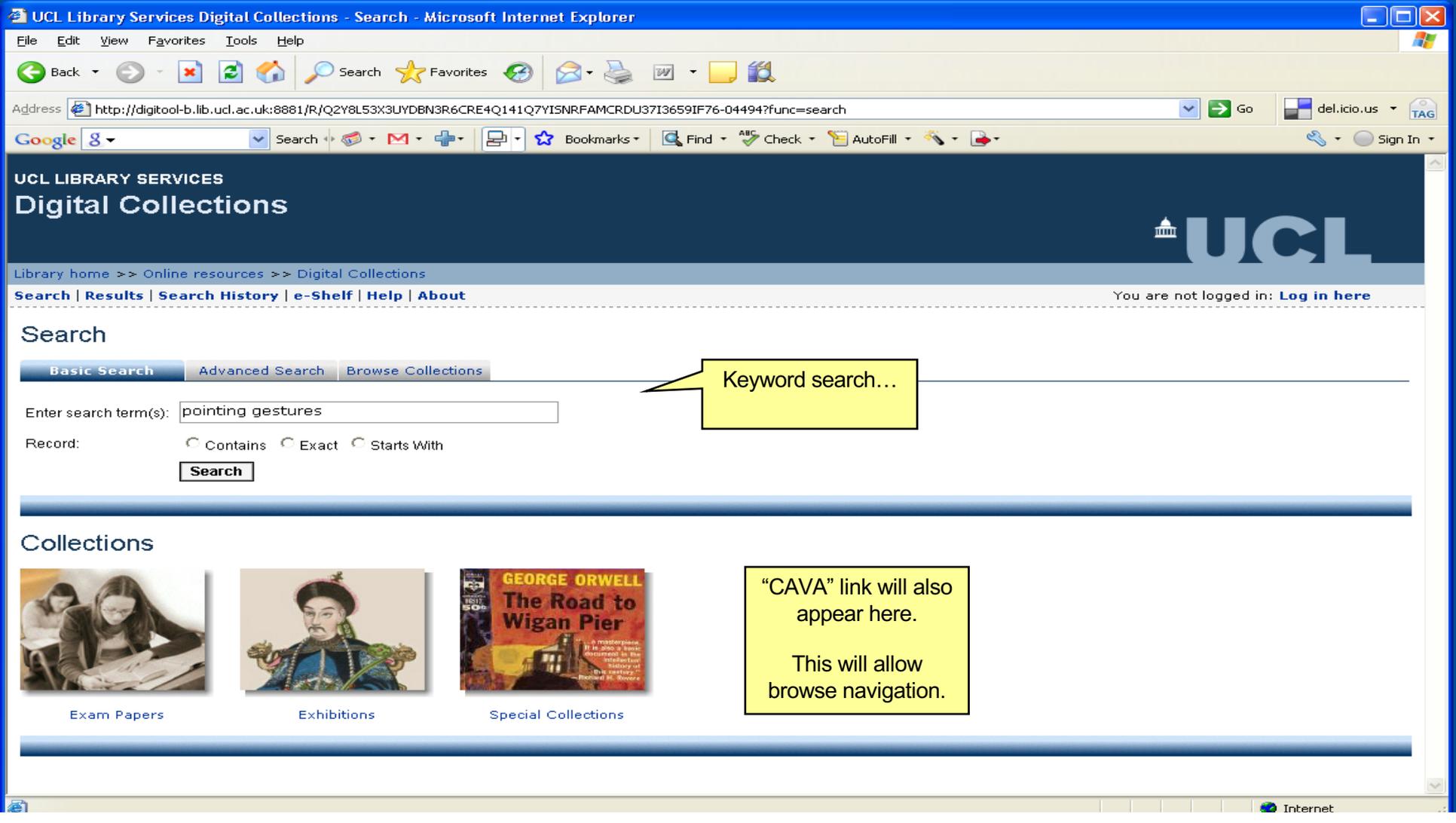
Data from past projects	Hours
Deaf children & teachers UCL/ Mahon/Department of Health	45
Deaf children & parents UCL/Mahon/ESRC	6
Children with language disorder & teachers Institute of Education/Radford/PhD	14
Persons with autism-teacher interaction Roehampton/Rae, Dickerson & Stribling/ESRC	16
Typically developing toddlers & parent UCL/Corrin/PhD	60
Typically developing toddlers & parent Canterbury/Forrester/ESRC	12
Children using AAC &peer UCL/Clarke/PhD	4
People with MND & spouse UCL/Bloch/PhD	6
Data expected from ongoing projects	
Aphasia therapy UCL/Beeke/Stroke Association	13
Adults with neurological disease UCL/ S.Bloch/NHS HIHR/PI	45
British Sign Language Corpus project UCL/Schembri/ESRC	360
Deaf children UCL Mahon/British Academy	7

The CAVA repository

- Will use UCL's DigiTool repository platform
 - <http://digital-collections.lib.ucl.ac.uk>
- Metadata is openly searchable; video resources will have access restrictions
- Built-in technical metadata extraction (using JHOVE), checksums, change history metadata; access control capabilities (IP and/or username)
- Front-end: quick overview...

Consent and technical issues

- Consent
 - Retrospective
 - Prospective
 - guidelines for depositors based on recent successful approval via UCL Research Ethics Committee/NHS multi-site ethics process
 - renewed consent at 18 years old
- Access controls (item-specific) - Application, authorisation and authentication
- What technical work is needed to underpin access and rights management procedures?
- Long-term preservation
- The ideal would be to retain the uncompressed master files in a managed environment, but these exceed current storage capacity.



Search

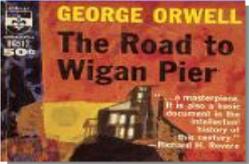
[Basic Search](#) | [Advanced Search](#) | [Browse Collections](#)

Enter search term(s):

Record: Contains Exact Starts With

Keyword search...

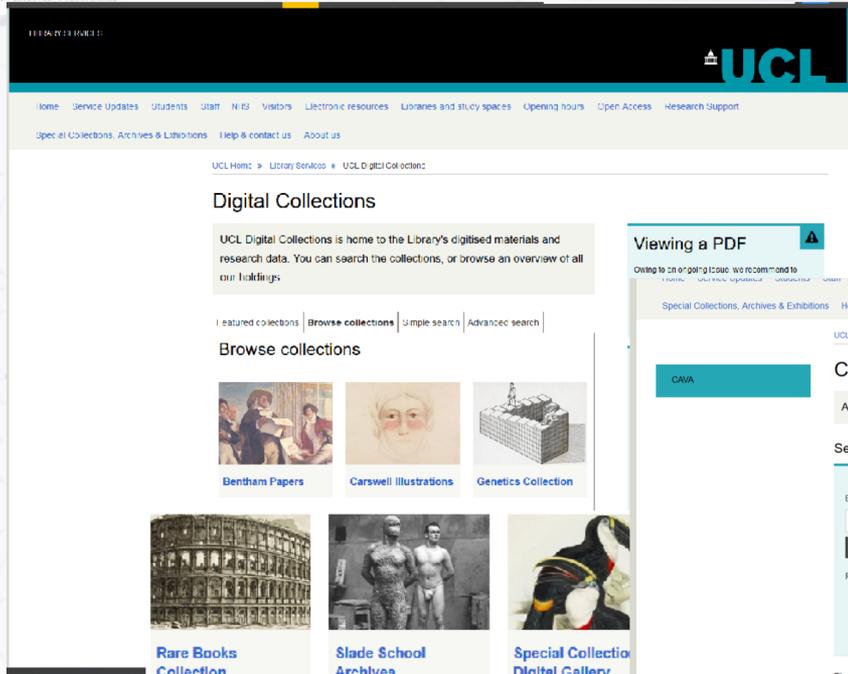
Collections

 <p>Exam Papers</p>	 <p>Exhibitions</p>	 <p>Special Collections</p>
---	--	--

"CAVA" link will also appear here.

This will allow browse navigation.

Ten years later – still going



UCL Home » Library Services » UCL Digital Collections

Digital Collections

UCL Digital Collections is home to the Library's digitised materials and research data. You can search the collections, or browse an overview of all our holdings.

Featured collections | **Browse collections** | Simple search | Advanced search

Browse collections



Bernham Papers



Carswell Illustrations



Genetics Collection



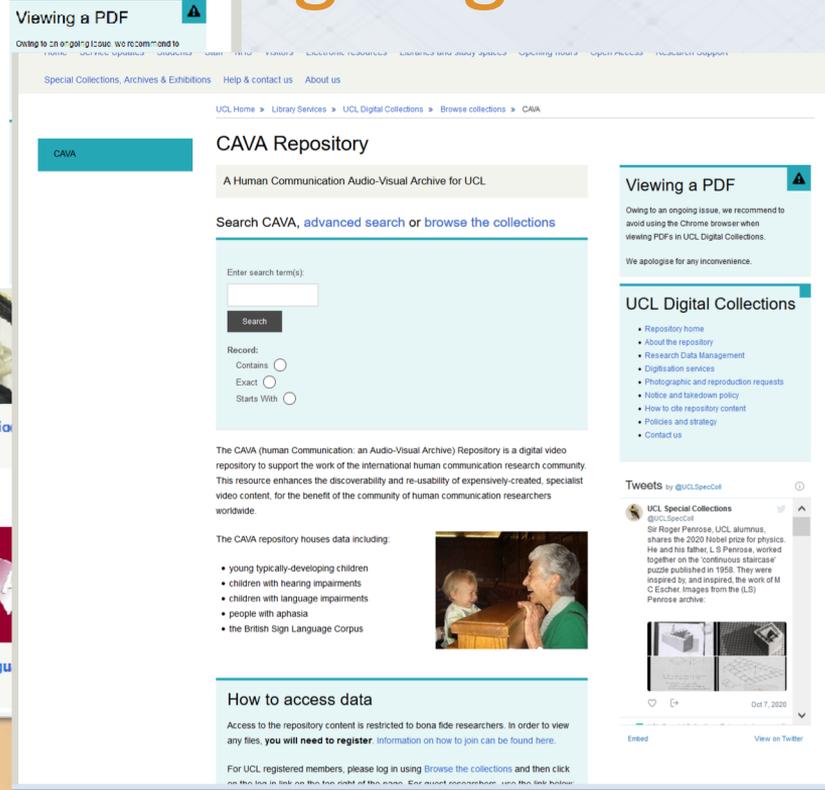
Rare Books Collection



Slade School Archives



Special Collection Digital Gallery



UCL Home » Library Services » UCL Digital Collections » Browse collections » CAVA

CAVA Repository

A Human Communication Audio-Visual Archive for UCL

Viewing a PDF

Owing to an ongoing issue, we recommend to avoid using the Chrome browser when viewing PDFs in UCL Digital Collections.

We apologise for any inconvenience.

UCL Digital Collections

- Repository home
- About the repository
- Research Data Management
- Digitisation services
- Photographic and reproduction requests
- Notice and takedown policy
- How to cite repository content
- Policies and strategy
- Contact us

Tweets by @UCLSpecColl

UCL Special Collections @UCLSpecColl
Sir Roger Penrose, UCL alumnus, shares the 2020 Nobel prize for physics. He and his father, L.S. Penrose, worked together on the 'continuous staircase' puzzle published in 1968. They were inspired by, and inspired, the work of M.C. Escher: images from the (L.S.) Penrose archive.

Oct 7, 2020

Embed View on Twitter

Search CAVA, advanced search or browse the collections

Enter search term(s):

Search

Records:

- Contains
- Exact
- Starts With

The CAVA (human Communication: an Audio-Visual Archive) Repository is a digital video repository to support the work of the international human communication research community. This resource enhances the discoverability and re-usability of expensively-created, specialist video content, for the benefit of the community of human communication researchers worldwide.

The CAVA repository houses data including:

- young typically-developing children
- children with hearing impairments
- children with language impairments
- people with aphasia
- the British Sign Language Corpus

How to access data

Access to the repository content is restricted to bona fide researchers. In order to view any files, you will need to register. Information on how to join can be found here.

For UCL registered members, please log in using [Browse the collections](#) and then click on the log in link on the top right of the page. For guest researchers, use the link below.



Post-project issues

2010

- Sustaining maintenance and growth?
 - ongoing costs of access management, storage, licences, support
- Continuing deposit by UCL researchers is foreseen
- Long term future (possible with UKDA)
 - Deposit by non-UCL researchers?
 - Exit strategy will be required...

2020

- Still functioning!
 - Modest level
 - Permission system works, does not scale
- New data, only as resources permit
- Considering migration options
- “Worth trying”

Fast Forward 2020 – Remote Access Landscape

- 🌀 My context: social science research data (+)
- 🌀 Many successful infrastructures
 - 🌀 NordMan -Nordic Microdata Access Network
 - 🌀 German RDCs – LIfBi - Leibniz Institute for Educational Pathways
 - 🌀 IDAN – International Data Access Network
 - 🌀 UK Data Service Secure Lab
 - 🌀 GESIS – Secure Data Center
 - 🌀 ICPSR – Virtual Data Enclave
- 🌀 But lots of infrastructures is a bit like lots of metadata standards...more ≠ better

SSHOC WP5: Innovations in Data Access

 Goal: to make data access FAIR & intelligently open

-  1 Access to biomedical data
-  2 Hosting and sharing data repositories
-  3 Legal Issues of innovative data access
-  4 Remote access to Sensitive Data
-  5 Cross-national survey data (ESS pilot)
-  6 Open Data in Heritage Science and Archaeology
-  7 Archaeology case study

“Two roads diverged in a wood...”

- The CAVA v.2020
 - Lots of experience to draw on, more stable tech, but...
 - No reliable path to sustainable infrastructure
- European Open Science Cloud/SSHOC
 - *Possible* path to sustainable infrastructure
 - Stop reinventing Safe Enclaves, Labs, Rooms, Centers
 - Demand for services to support Remote Work can only grow (Covid effect)
 - But...
 - Privacy concerns grown, w/ legal sanctions (GDPR)
 - *“Ice cream castles in the air” or “illusions”?*





SSHOC

social sciences & humanities open cloud

Join our community



<https://www.sshopencloud.eu>



info@sshopencloud.eu



[@SSHOpenCloud](https://twitter.com/SSHOpenCloud)



[/in/sshopencloud](https://in.linkedin.com/company/sshopencloud)



The curation and disclosure of pathological speech corpora

Katarzyna Klessa



UNIwersytet
IM. ADAMA MICKIEWICZA
W POZNANIU

Disclosure and curation of corpora

- Legal, research and technical issues play role in the process of disclosure and curation of corpora
- The importance of these issues for:
 - CDS (Corpora of Disordered Speech)
 - especially for CDS created in the past, before the present regulations came into existence: **“archival” corpora**

legal

research

technical



Licence attribution

legal

- Explicit licence attribution is useful even in case of data published without any access restrictions
 - Users will often expect clear information about licensing, preferably compliant with one of the standard licenses (e.g. those suggested by CLARIN centers)
 - GDPR - accepted by EU but not necessarily elsewhere
-

Archival data, metadata vs. new regulations

legal

However:

- In the past, recordings were often made based on oral agreements with speakers (no consent form available now)
- In case of children (now adults) - the consent usually given by their representatives: parents, guardians or school teachers / directors
- Often: no contact with the speakers / representatives at present.
- Present regulations not clear with respect to data collected before the _____ present regulations

Data reusability interoperability ?

research

- Various perspectives regarding data collection & storage:
 - Data re-usability, interoperability perspective: collecting and storing as much (meta)data as possible; data collection is very costly and time-consuming
 - The limitations due to GDPR: e.g. it is recommended to collect only the info necessary for a specific purpose
-

Archival data formats

technical

- File formats can be obsolete, unsupported by many of the present data analysis tools, e.g. audio data collected with Kay Elemetrics (CSL, The Computerized Speech Lab proprietary format)
- Annotation standards change, currently most formats are XML-based but not only
- If allowed to share - data contributors might need help regarding technical issues

How CDS can be found through one organisation and made accessible through another

- DELAD
- K-Centre for Atypical Speech
- TalkBank
- Other?

Using various platforms for better visibility



K-ACE
CENTRE

HOME MORE ABOUT ACE SERVICES SHOW CASES PUBLICATIONS REACH

K-Centre for Atypical
Communication
Expertise

Contact us



<https://delad.net>

DELAD

sharing corpora of
disordered speech
among researchers

CONTACT US

<https://ace.ruhosting.nl/>

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	Programs	Manuals
<u>**Ground Rules**</u>	<u>CLAN</u>	<u>CHAT - CLAN - MOR</u>
<u>**Hints on Downloading**</u>	<u>MOR grammars</u>	<u>Tutorial Screencasts</u>
<u>Contributing</u>	<u>XML creator</u> and <u>XML Schema</u>	<u>SLP's Guide to CLAN</u> and <u>中文</u>
<u>IRB Principles</u>	<u>Other Software</u>	
Conversation Banks	Child Language Banks	Multilingualism Banks
<u>CABank</u>	<u>CHILDES</u>	<u>Second Language Tutors</u>
<u>SamtaleBank</u>	<u>PhonBank</u> 	<u>BilingBank</u>
<u>ClassBank</u>	<u>HomeBank</u>	<u>SLABank</u>

<https://talkbank.org/>



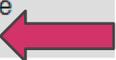
PhonBank is the child phonology component of the [TalkBank](#) system. TalkBank is a system for sharing and studying conversational interactions. PhonBank is supported by grant RO1-HD051698 from NIH-NICHD to Brian MacWhinney and Yvan Rose. *PHON* is designed and built by Yvan Rose and Greg Hedlund. Currently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

System	Database	Phon Program
<p><u>**Ground Rules**</u></p> <p>Contributing New Data</p> <p>IRB Principles</p>	<p><u>**Index to Corpora**</u> ←</p> <p>Browsable Database</p> <p>TalkBankDB database search</p> <p>Hints on downloading</p> <p>Database versioning</p>	<p>Phon website and User Manual</p> <p>Phon on GitHub</p> <p>PhonTalk (CHAT ↔ Phon)</p>
Links	Resources	Contact
<p>Other TalkBank databases</p> <p>Other Child Language sites</p> <p>Research based on Phon</p> <p>Workshop presentations</p>	<p>Phon Basics: A practical introduction</p> <p>Video tutorials on YouTube</p> <p>Downloadable files for video tutorials:</p> <ul style="list-style-type: none"> • Phon projects for tutorials • Media files for tutorials 	<p>Yvan Rose: homepage</p> <p>Phon mailing list</p>



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

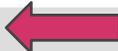
Signed contribution forms are available [here](#).

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

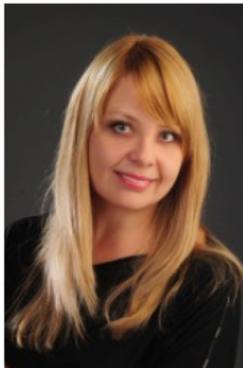
<https://phonbank.talkbank.org/access/>



This page provides an index to PhonBank Clinical data.

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

Polish Cued Speech Corpus of Hearing-Impaired Children

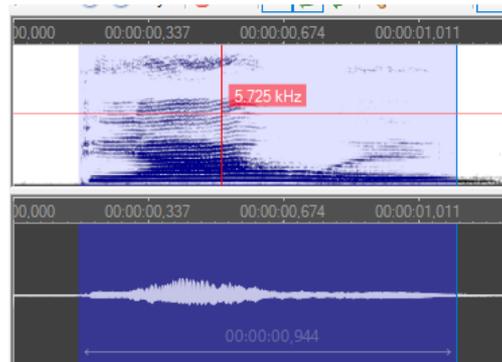


Anita Lorenc
Institute of Applied Polish Studies
University of Warsaw
anita.lorenc@uw.edu.pl
[website](#)

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

[Phon data](#)

[CHAT data](#)



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

Linking to local repositories ?

- Polish CLARIN provides an on-line platform supporting deposit and sharing of various types of language corpora
- <https://clarin-pl.eu/dspace/>
- The platform is intuitive, basic metadata forms are available, licence agreement type can be selected from a list of defaults or adjusted to the depositor's needs
- It is possible to add links to external websites of the project or data samples
- Licence suggestions: <https://clarin-pl.eu/dspace/page/licenses>



The curation & disclosure of archival corpora

- Legal, research and technical issues in dealing with archival corpora
- Datasets created in the past can be findable via different platforms thanks to **collaboration** of partners and technical support from infrastructure representatives
 - Access to some parts of the archival corpora must remain restricted because of unsolved legal issues or the character of data (sensitive information)
- DELAD initiative - as an integrating platform for potential contributors of disordered speech corpora

OFFER YOUR DATASET



Thank you for your attention!

Questions?

Please put them in the chat box.

Slides and a recording will be sent to all registered participants.

Join our community



<https://www.sshopencloud.eu>



@SSHOpenCloud



info@sshopencloud.eu



/in/sshopencloud