Report CLARIN Workshop DELAD: Database Enterprise for Language And speech Disorders

Date and location of the workshop

The workshop took place as a lunch to lunch workshop on 28-30 January 2019 in room 3.1 at the SURF Offices, Hoog Overborch, Utrecht (https://www.surf.nl/en/about-surf/contact/directions-to-surf-surfmarket-and-surfnet/index.html) For CLARIN this was a Type II workshop.

The Organizing Team

Henk van den Heuvel; CLST Radboud University, the Netherlands
Satu Saalasti, University of Helsinki, Finland
Martin Ball; Bangor University, UK
Alice Lee; University College Cork, Ireland
Nicole Müller; University College Cork, Ireland
Aleksei Kelli; University of Tartu, Estonia

Information about the organizing team

Dr Henk van den Heuvel has been involved in the collection, compilation and validation of many spoken and written language resources at the national and international level. He has been project leader and project participant in CLARIN-NL projects amongst which the VALID project (http://validdata.org/) which aimed to include a number of Dutch Corpora of Disordered Speech (CDS) in the CLARIN infrastructure. He is also co-coordinator of CLARIN-NL's Data Curation Service now integrated into CLARIAH.

Dr Satu Saalasti is a University Lecturer of Logopedics in the Department of Psychology and Logopedics at the University of Helsinki. Her research interests include multisensory perception of speech in individuals with autism spectrum disorders and neural mechanisms underlying real-life language. She has studied the brain mechanisms underlying lipreading, listening and reading in her post doctoral studies by combining functional magnetic resonance imaging and methods of computational linguistics.

Prof. Martin Ball has held senior positions in the UK, the US, and Sweden. He has published widely in the areas of clinical phonetics and phonology and is the editor-in-chief of the journal *Clinical Linguistics and Phonetics*. He was the PI in a grant from Riksbankens Jubileumsfond that funded the preliminary workshops in this projects.

Dr Alice Lee is a speech and language therapist and Lecturer in Speech and Hearing Sciences at University College Cork, Ireland. Her research interest includes perceptual and instrumental investigations of speech disorders associated with structural anomalies and neurological impairment. Currently, she is a PI of a project on speech prosody skills in children with spina bifida funded by Health Research Board, Ireland. She is supervising a PhD student on speech difficulties in individuals with Down syndrome. She is the Editor of *Journal of Clinical Speech and Language Studies* – the official journal of the Irish Association of Speech and Language Therapists.

Prof. Nicole Müller is the Head of Department of Speech and Hearing Sciences at University College Cork, Ireland. She has held academic positions in the UK, the US, and in Sweden. Her research interests are in the areas of bilingualism, and adult acquired deficits in communication and cognition. She is currently supervising doctoral work in clinical linguistics, dementia, and Parkinson's disease. As the former chairperson of a university research ethics board, she has many years experience in monitoring research ethics.

Dr Aleksei Kelli is Professor in Intellectual Property at the Faculty of Law at the University of Tartu, Estonia. Dr Kelli is responsible for legal issues concerning digital language resources at the University of Tartu and the Institute of the Estonian Language. He is Chair of Legal Committee of CLARIN ERIC (Common Language Resources and Technology Infrastructure). Aleksei acted as the Head of an Expert Group on the Codification of the Intellectual Property Law (2012-2014, the Ministry of Justice of Estonia). He was responsible for a project dedicated to open science (Estonian Research Foundation, 2016-2017).

List of speakers and attendees

Country	Category	Name	Affiliation
NL	COORD	Henk van den Heuvel	Radboud University, Nijmegen
IE	COORD	Alice Lee	University College Cork
FI	COORD	Satu Saalasti	University of Helsinki
EE	CLIC	Aleksei Kelli	University of Tartu
EL	CLIC	Alexandros Nousias	Open Data Institute-Athens
FI	CLIC	Krister Linden	University of Helsinki
UK	RESEARCH	Stuart Cunningham	University of Sheffield
NO	RESEARCH	Pernille Hansen	University of Oslo
NL	RESEARCH	Marina Ruiter	Radboud University, Nijmegen
CY	RESEARCH	Kakia Petinou	Cyprus University of Technology
PL	RESEARCH	Anita Lorenc	Warsaw University
PL	RESEARCH	Katarzyna Klessa	Adam Mickiewicz University, Poznań
EU	CLARIN	Dieter Van Uytvanck	CLARIN-ERIC
EU	CLARIN	Maria Eskevich	CLARIN-ERIC
FI	FIN-CLARIN	Mietta Lennes	University of Helsinki
FI	FIN-CLARIN	Martin Matthiesen	CSC-IT Center for Science Ltd
UK	RESEARCH	Joanne Cleland	University of Strathclyde
IE	RESEARCH	Nicola Bessell	University College Cork
BE	RESEARCH	Veronique Delvaux (with George)	University Mons
BE	RESEARCH	George Christodoulides (with Veronique, U Mons)	University Mons

Workshop goal

Ball & Müller organised two workshops in Linköping, Sweden, in 2015 and 2016, in which available resources and issues related to accessibility were inventorised. These workshops resulted in the DELAD initiative (see http://delad.ruhosting.nl/). From these workshops it was concluded that (1) only a minority of existing Corpora of Disordered Speech (CDS) can be made accessible due to privacy constraints, (2) we now have the knowhow to collect new CDS that can be shared according to FAIR principles, and (3) the CLARIN infrastructure is indispensable for this purpose.

In CLARIN context, we organised a first workshop in Cork, November 2017, to set-up a plan to collect existing and new CDS and to include these in the CLARIN infrastructure. A report with video-lectures of this workshop is available: https://www.clarin.eu/blog/report-clarin-delad-workshop.

The goal of this second workshop was to review the status of the actions set out in the first workshop, exchange deeper insights on ethical and legal aspects (including IPR) of CDS collection against the

background of the GDPR, and come up with a plan for primary special needs for the CLARIN infrastructure to host CDS, and a feasible plan to spend 3 PM ICT developer time on this infrastructure.

Contributions of the workshop to strategic goals of CLARIN ERIC

Corpora of Disordered Speech (CDS) are difficult to find, very costly to collect, and for privacy reasons hard to share. On the other hand, due to their small size and dedicated purpose, they should be combined to be suited for re-use. A strong need is felt by the research community to bring together existing and new CDS in an interoperable and consistent way that is both legal and ethically safeguarded. The CLARIN infrastructure is regarded as indispensable for this purpose. The CHILDES Talkbank, CMU also being a CLARIN Centre, is an important asset of this infrastructure following US legislation. CDS can be federatively archived at local CLARIN centres whereas they can be made findable through a central portal via their (harvested) metadata. CLARIN precisely offers the standards, best practices and services which are needed for this. Through the collaboration with CLARIN's CLIC in this workshop, the requirements following from the GDPR for sharing CDS were also substantially addressed.

Agenda of the Workshop

Date	Time	Topic	Agenda
28 Jan	**13:00-14:00**	**LUNCH**	
	14:00-14:20	**Overview of state of affairs from action points resulting from workshop in Cork (Nov 2017)**	**Welcome and introduction** **Presentations by** **★ Overview of the action points (Henk van den Heuvel)** **★ Update on state of affairs / suggestions (All)**
	14:20-17:30	**Presentations by researchers specifying relevant requirements for the infrastructure for sharing their CDS**	**Presentations by** **★ [14:20-14:40] Nicola Bessell & Alice Lee** **★ [14:40-15:00] Satu Saalasti** **★ [15:00-15:20] Joanne Cleland (Skype/Zoom)** **★ [15:20-15:40] Kakia Petinou (Skype/ Zoom)** **★ [15:40-16:00] Stuart Cunningham** **16:00 COFFEE** **★ [16:00-16:20] Pernille Hansen** **★ [16:20-16:40] Marina Ruiter & Henk van den Heuvel** **★ [16:40-17:00] Anita Lorenc & Katarzyna Klessa** **★ [17:00-17:20] Veronique Delvaux & George Christodoulides**
	19:00	**DINNER**	
29 Jan	**09:30-10:30**	**Legal and ethical aspects of collecting, hosting and sharing CDS (together with CLARIN's Legal Issues Committee – CLIC]**	**Presentations by** **★ [09:30-09:50] Alexandros Nousias, "Open data and privacy"** **★ [09:50-10:10] Aleksei Kelli, "To ask or not to ask: How about consent for CDS?"** **★ [10:10-10:30] Krister Linden, "Safeguards between consent and data sharing"** **10:30 COFFEE BREAK**

	10:45-12:45	**Group discussions in depth about three cases:** **➔ Quick case presentation / reminder** **➔ Discussion in three groups** **➔ Report from groups** **➔ Plenary discussion**	**★ [10:45-11:45] Use Case 1*** **★ [11:45-12:45] Use Case 2***
	12:45-13:30	**LUNCH**	
	13:30-14:30	**Group discussions (cont)**	**★ [13:30-14:30] Use Case 3***
	14:30-17:00	**Overview of the present facilities and future options of the CLARIN infrastructure (CLARIN technical staff & data centres]**	**★ [14:30-14:50] Dieter Van Uytvanck, "The CLARIN Infrastructure"** **★ [14:50-15:10] Henk van den Heuvel, "Requirements – webportal & recommendation, Cork"** **★ [15:10-15:40] Brian MacWhinney/Yvan Rose, "Clinical banks at Talkbank" (Skype/Zoom)** **★ [15:40-16:00] Mietta Lennes & Martin Matthiesen: Developments in sensitive data at Kielipankki - The Language Bank of Finland** **★ Overview priorities from Monday's presentations/cases Monday & Discussion (Dieter Van Uytvanck & Maria Eskevich)**
	18:30	**DINNER**	
30 Jan	**09:30-11:30**	**Set up a priority list for additional facilities for a CLARIN CDS portal**	**★ Make inventory or priorities (Maria Eskevich & Henk van den Heuvel)** **➔ Design** **➔ Legal/ethical** **➔ Technical for data & metadata** **➔ Access to data** **10:30 COFFEE BREAK** **★ Open discussion/Q&A**

	11:30-13:00	**Set up specification for a relevant and feasible 3 PM development contribution to this portal. Further funding options**	**➔ What will be implemented?** **➔ Who can/will do it?** **➔ When will it be done?** **➔ How can further developments be financed?**
	13:00-14:00	**LUNCH**	

A summary for publication on the CLARIN ERIC website describing the development goals and methods

Corpora of Disordered Speech (CDS) are difficult to find, very costly to collect, and for privacy reasons hard to share. On the other hand, due to their small size and dedicated purpose, they should be combined to be suited for re-use. A strong need is felt by the research community to bring together existing and new CDS in an interoperable and consistent way that is both legal and ethically safeguarded.

In CLARIN context we organised a first workshop in Cork, November 2017, to set-up a plan to collect existing and new CDS and to include these in the CLARIN infrastructure. A report with video-lectures of this workshop is available: https://www.clarin.eu/blog/report-clarin-delad-workshop.

The goal of this second workshop was to review the status of the actions set out in the first workshop, exchange deeper insights on ethical and legal aspects (including IPR) of CDS collection against the background of the GDPR, and come up with a plan for primary special needs for the CLARIN infrastructure to host CDS, and a feasible plan to spend 3 PM ICT developer time on this infrastructure.

The presentations and conclusions of the workshop can be found here.

Main outcomes of the workshop:

1. Reaffirmation that CLARIN is the Data Trust, to provide the data fence around CDS.
2. DELAD should apply to become a Task Force within CLARIN focusing on practical issues on sharing CDS.
3. As a result, the DELAD website should be updated with a CLARIN flag and contain relevant guidelines for collecting, sharing and storing CDS.
4. Talkbank is seen as a good CLARIN site to host CDS, especially if a European storage cloud and stricter access policy can be realised.
5. Work together on contributions for the CLARIN AC 2019 in Leipzig.

Items 3 & 4 will be pursued for obtaining the 3 PM ICT developer time.

6. Proposal on how to spend 3 PM CLARIN ICT support

Priorities:

1- Upgrade of DELAD website to be pointer portal to guidelines & data (integrate into CLARIN as TF page)
2- European storage cloud for Talkbank (corpora)
3- Language bank rights as a feature of Talkbank (presupposes CLARIN federated login)
4- Open source Player for audio & video player (for “ANNEX” works with ELAN (or Talkbank Player)

Update of action points resulting from the workshop

Actions previous workshop

⇒ Establish a working group on consent forms directed at justifying sharing data through permission by ethical committees (look for use-case flagship) (look also at best practices from others like Talkbank; Talkbank can be our handle to organise something similar for Europe)
Nicola Bessell, Pawel, Nicole, Alice, Satu

Status:
Forms collected and put together in this folder. Must be extended.
Moreover: Post GDPR updates needed
Pernille Hansen has shared a response text with a justification towards her ethical committee to share data.

=> Next action: make template with guidelines of issues to take care of consent forms

=> make list of questions/issues for CLIC lawyers about issues that we encounter in ethical/legal matters. Can be Google doc and part of website perhaps (Nicola)

⇒ Compile overview of best practices for metadata profiles (mandatory and optional categories) (include profiles of VALID & Talkbank!)
Henk, + those involved in action for Talkbank,

Status:
TBD

⇒ Make an effort to make TAPAS resources available through a CLARIN centre in some way (starting with metadata only, up to sharing through registration / agreements), and via BEAT
Heidi & Rob (involve Schuller)

Status:
TBD. It is too early for TAPAS to do that.

⇒ Try to set up a CLARIN Centre in your country if there is not one or establish a trusted cooperation with an existing one in another country
Nadine, Kristina, Nicola, Alice, Nicole, Martin, Kakia

Status:
There is a Polish CLARIN now with Dspace (adopted and adapted from LINDAT)
Kakia will try connect to the Greek CLARIN
CLARIN UK and FR are observers,
CLST Nijmegen is in application phase for B centre for this type of data (which is not a warrant for hosting special category data as such).

⇒ Explore including DELAD as highly relevant pilot/use case into ICT-29 call for May 2018, together with ELRA & CLARIN
Henk, Dieter

Status:
Realised in SSHOC project: H2020-INFRAEOSC-2018-2-823782, see also http://www.ilc.cnr.it/en/content/sshoc

⇒ Explore MSC EJD (or ETN) proposal by January 2019: Research into training and education and assessment of therapy (building CDS as side products) with multilingual aspect. Distinguish it from TAPAS as a speech technology project. Explore options of joint doctorates at your universities.
All

Status:
Proposal made in Cork, but no expression of interest. This works better through personal contacts. Link to text will be provide by Nicola and Alice

⇒ Explore options to connect a next workshop to a TAPAS meeting (Henk, Stuart)
Fall 2019?

⇒ Investigate possibility of COST proposal as follow up for these type of workshops as a preparatory action for a EJD proposal
Kakia, Nadine, Kristina

Status:
No actions taken. A related action is: https://www.cost.eu/actions/IS0804/#tabs|Name:overview

⇒ Make a resource from Satu Saalasti's lab accessible through FIN-CLARIN
Mietta, Martin M, Satu

Status:
Now in progress

⇒ Compile overview of best practices for layered access to data
Dieter, Martin M

Status:
The works has started, see this folder

⇒ Compile overview of best practices for pseudonimisation
Rob

Status:
TBD

⇒ Poster presentation at ICPLA in Oct. 2018
Martin B
Done: https://www.um.edu.mt/__data/assets/pdf_file/0017/353123/icplamaltaprogramme.pdf
We also had a poster presentation at CLARIN Annual Conference, Pisa, 2018:
https://www.clarin.eu/clarin-annual-conference-2018-bazaar#Workshop%20presentations

New actions:

Next CLARIN AC in Leipzig 2019:
=> Submit a paper about the DELAD initiative: Henk, Satu, Kasia
=> Also Aleksei Kelli would like to write a paper about the Use cases in the light of GPR
=> Make special category data a focus for a session in the next CLARIN Annual Conference
Deadline: mid April

=> DELAD to apply as a Task Force within CLARIN about practical issues on sharing CDS (perhaps as specific family of resources; clear links to CLIC and metadata TF).
We see CLARIN as the Data Trust, to provide the data fence around CDS. The federated data storage approach is very good for this type of data, e.g. Kielipankki (also caters for remote desktop access), Polish D-space
Maria Eskevic will check the procedure for becoming a TF in CLARIN

=> Look for European Ethical Committee to invite to next meeting, enrio.eu

=> Make an overview of CLARIN centres and protecting levels of sensitive data & option to host CDS from other countries (Dieter)

=>How to spend 3 PM CLARIN ICT support

- add facet language /speech disorder to VLO & CMDI (must be done anyhow)

Priorities:

1- Upgrade of DELAD website to be pointer portal to guidelines & data (integrate into CLARIN as TF page)

2- European storage cloud for Talkbank (corpora)

3- Language bank rights as a feature of Talkbank (presupposes CLARIN federated login)

4- Open source Player for audio & video player (for “ANNEX” works with ELAN (or Talkbank Player)

=> Add facet language /speech disorder to VLO & CMDI (Henk, Dieter)