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

January 2018

Authors and Organisers

Henk van den Heuvel; CLST Radboud University, the Netherlands
Martin Ball; Bangor University, Wales, UK
Alice Lee; University College Cork, Ireland
Nicole Müller; University College Cork, Ireland
Satu Saalasti, Faculty of Medicine, University of Helsinki, Finland

Date and location of the workshop

The workshop took place as a lunch to lunch workshop on 15-17 November 2017 in room LG103, Brookfield Health Sciences Complex, University College Cork, Cork, Ireland (https://www.ucc.ie/en/about/visitors/)

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 CSD in the CLARIN infrastructure. He is also co-coordinator of CLARIN-NL's Data Curation Service.

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 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.

List of speakers and/or attendees

Country	Category	Name
NL	COORD	Henk van den Heuvel
UK	COORD	Martin Ball
IE	COORD	Alice Lee
IE	COORD	Nicole Müller
FI	COORD	Satu Saalasti
UK	RESEARCH	Heidi Christensen (TAPAS)
NL	RESEARCH	Rob van Son (TAPAS)
CY	RESEARCH	Kakia Petinou
PL	RESEARCH	Anita Lorenc
PL	RESEARCH	Katarzyna Klessa
EU	CLARIN	Dieter Van Uytvanck
FI	CLARIN-FI	Mietta Lennes
FR	IPR	Pawel Kamocki
MALTA	RESEARCH	Nadine Tabone
MALTA	RESEARCH	Kristina Agius
FI	CLARIN-FI	Martin Matthiesen
IRL	RESEARCH	Yvonne Fitzmaurice
IRL	Library	Eoghan Ó Carragáin
IRL	RESEARCH	Ciara O'Toole
IRL	RESEARCH	Pauline Frizelle
IRL	RESEARCH	Anne Marie Devlin
IRL	RESEARCH	Nicola Bessell
IRL	RESEARCH	Deirdre O'Leary
IRL	RESEARCH	Jennifer Harte

The goal of the workshop

Corpora of disordered speech (CSD) are hard to obtain. They are costly to collect and difficult to share due to privacy issues. Moreover, they are often small in size and very dedicated in terms of language impairments addressed. These factors make re-use a challenge on the one hand, and a necessity on the other.

Two of the co-applicants (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 (1) that only a minority of existing CSD can be made accessible due to privacy constraints, and (2) that we now have the knowhow to collect new CSD that can be shared according to FAIR principles, (3) that the CLARIN infrastructure is indispensable for this purpose.

In this workshop we brought together a selected group of experts in language disorders research and CSD to set-up a plan to collect existing and new CSD and to include these in the CLARIN infrastructure.

The action plan resulting from the workshop should include concrete initiatives to set up a repository of existing and new CSD embedded in the CLARIN infrastructure. A proposal for e.g. a Marie Curie ITN would fit excellently in such an action plan.

The workshop was a CLARIN Type I workshop.

Contributions of the workshop to strategic goals of CLARIN ERIC

The relevance of including CSD in the CLARIN infrastructure has been addressed during several meetings of the CLARIN General Assembly. On the one hand CSD are difficult to obtain; on the other hand, due to their small size and dedicated purpose, they should be combined to be suited for re-use. Moreover, they are also very costly to collect. Therefore, a strong need is felt by the research community to bring together existing and new CSD in an interoperable and consistent way. The CLARIN infrastructure is regarded as indispensable for this purpose. The CHILDES Talkbank, CMU[1] also being a CLARIN Centre, is an important asset of this infrastructure with a wealth in best practices. CSD can be 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.

A summary of the main results of the workshop

Difference of this workshop compared to the two previous ones in Linköping

[1] Carnegy Mellon University

- Conclusion of most participants in Linköping was that they could not share data due to ethical/IPR regulations. Therefore we invited others to the Cork workshop, including those starting data collections (such as in the TAPAS project) so that they could benefit from the information to make their data optimally shareable
- This workshop was focussed on CLARIN. We had participants from CLARIN ERIC and the local people of FIN-CLARIN
- Therefore, unlike Linköping, we invited participants from Europe only
- We also focussed on legal/ethical issues anticipating the GDPR 2018 with legal expertise from partners, particular ELRA

As a result of the workshop the following recommendations for a CLARIN CSD portal were formulated:

Data

- Datasets should always be related to language and speech pathologies. Matched normal controls are welcome, even desirable
- Look for best practices for defining categories of pathologies
- Formats: follow CLARIN guidelines for standard formats
- Versioning:
    - Persistent identifiers as the key to proper versioning
    - Changes should be made in a way that it should be traced, e.g. in new annotation layers
    - If data is changed by someone else than the author, the changed version should be a new submission, preferably in contact with the author
    - Look for best practices, e.g. Data Alliance recommendations
    - CLARIN has versioning systems in place

Anonymisation:

- Direct identifiers: name, date of birth, address, phone number - should definitely be pseudonymous, i.e., coded.
- Sensitive information / indirect identifiers: economic status (official scales should be used), area code, religion, ethnicity, sexual preferences, use categories as broad as possible (coarse grained)
- If ethical permissions are there, such information can be revealed but it is recommended to store it separately and/or as encoded
- Leave raw data intact for research. If this is in conflict with sharing for scientific integrity, then national and EU regulations should find a solution for that. Let's not make this our problem

Metadata:

- Look at best practices such as aphasia & dementia bank as part of Talkbank

- In CMDI[2] there already a couple of profiles, e.g. for the VALID datasets
- Find common denominators of existing profiles and use that as a minimum

IPR & Ethics:

- It is important to gather positive examples that can serve as inspiration for others, e.g. the DementiaBank and AphasiaBank at the TalkBank. This can be used as illustration in discussions with ethics boards.
- Gathering best practices and templates (e.g. consent forms, data management plans) and sharing these would also be helpful. For the documents from the US-based TalkBank a European variant could be made.
- The GDPR should not be necessarily seen as a threat when you are already complying with existing requirements (and common sense) with regards to data protection etc.
- Safeguarding data sets from deletion (as e.g. demanded by some ethics boards) should be an absolute priority. 2 proposed ways to achieve a policy change:
    - Involving people with impairments in the process (to give them a voice in the discussion and in the end access to research participation).
    - Involving funding agencies, with the argument of replicability/accountability.

Access

- There is a clear need for a layered approach with different access levels.
- Important question is where the data will reside: locally (close to where it was captured) or somewhere else, e.g. in a central repository. There is no universal answer – this needs to be decided at the local level – but in general a decentralized approach seems to be the most feasible, since it allows depositing the data in a well-known and trusted repository, preferably a national CLARIN data centre, otherwise this is a good opportunity to set up one (C centre for harvesting metadata at least).
- In any case it is good to realize that there is already a lot of technology and know-how to build upon in the context of CLARIN: be it existing B-centres, repository software or expertise on e.g. metadata modelling. Principles like versioning, assigning persistent identifiers and federated login seem non-controversial.
- The need for a local (or national) repository might fit well with the possible creation of new CLARIN consortia (e.g. Ireland and Cyprus).
- A federated data approach is also needed to obtain national funding

CLARIN-DELAD Portal

- Making the metadata available via the Virtual Language Observatory (vlo.clarin.eu) would be important for the discoverability of the data sets. Making datasets available via TalkBank could be an option but is probably complicated given the different legislation in the US.
- The target users should be researchers. Patient access is important too but can be handled via the local institutions.

[2] Component MetaData Infrastructure, see https://www.clarin.eu/content/component-metadata

- The local centre/institution is best positioned to decide on which data sets to include or to act as a gatekeeper in general.
- Maintenance and funding: a decentral approach probably makes the best chance to stand and survive. Of course the wheel does not need to be reinvented, a lot can be used from the existing CLARIN infrastructure and centres. How much can be (re)used will depend on the situation of individual institutions. This is something that needs to be arranged and looked into on a case-by-case basis.

Next steps to be taken within the CLARIN community

⇒ Compile overview of best practices consent forms (also in TAPAS)
Rob, Heidi

⇒ 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

⇒ 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)

⇒ 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

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

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

⇒ Explore organising a CLARIN Type II workshop focusing at some development work for the DELAD infrastructure. Potential topics were discussed on 16 Nov. regarding the DELAD portal including Dynamic selection of content. The workshop could be connected to a TAPAS event.
Dieter, Henk, Satu, Alice, Nicole, Martin

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

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

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

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

⇒ Compile overview of best practices for pseudonimisation
Rob

⇒ Poster presentation at ICPLA in Oct. 2018
Martin B