

The DELAD CLARIN workshop: sharing corpora of disordered speech

CLARIN Workshop Cork 15-17 November 2017



DELAD initiative

DELAD (=SHARED in Swedish) is an initiative to establish a digital archive of disordered speech and share this with interested researchers.

Website: <http://delad.ruhosting.nl/>



The website has a number of objectives:

- to raise awareness for the DELAD enterprise
- to bring interested researchers together to promote the initiative
- to start an inventory of relevant data
- to stimulate and facilitate the bilateral exchange of data

DELAD Background

- Two workshops in Linköping, Sweden, funded by Riksbankens JF
 - Partners from Europe, USA, Canada
- October 2015
 - What do partners have in terms of CDS
 - Can this be shared with others
 - Issues:
 - IPR/ ethics
 - Formats
 - Annotations/transcriptions
 - Other research data
 - Metadata
 - Levels of anonymization
 - Levels of public access
 - Maintenance
 - One portal
- June 2016:
 - First connections to CLARIN infrastructure incl CMU/ CHILDES & Talkbank
 - Funding options: EU infra, Digging into data, VW Stiftung



DELAD CLARIN Proposal



- Announced via <https://www.clarin.eu/news/call-clarin-workshop-proposals-2-types-2017>. Type I
- Submitted: 31 March. Approved: 14 April
- Applicants:
 - Henk van den Heuvel; CLST Radboud University, the Netherlands
 - Martin Ball; Linköping University, Sweden
 - Alice Lee; University College Cork, Ireland
 - Nicole Müller; University College Cork, Ireland
 - Satu Saalasti, Faculty of Medicine, University of Helsinki, Finland
- Motivation:
 - The relevance of including CDS in the CLARIN infrastructure has been addressed during several meetings of the CLARIN General Assembly. On the one hand CDS 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 CDS in an interoperable and consistent way. 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 with a wealth in best practices. CDS 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.



DELAD CLARIN+ Proposal



- Goals of workshop:
 - Involve further research partners to DELAD initiative
 - Focus on integration of the initiative into the CLARIN infrastructure
 - Specs of a CDS centered webportal in CLARIN context
 - Compile a list of relevant further actions to realise a DELAD CDS portal focusing at increasing user involvement to the initiative
 - Funding options: national, European



Using clinimetical patient data for scientific research

Creating corpora of disordered speech in the clinical setting: opportunities and challenges

Marina Ruiter (Sint Maartenskliniek, Nijmegen; CLS, Radboud University Nijmegen)

Henk van den Heuvel (CLST, Radboud University Nijmegen)

This is CLST

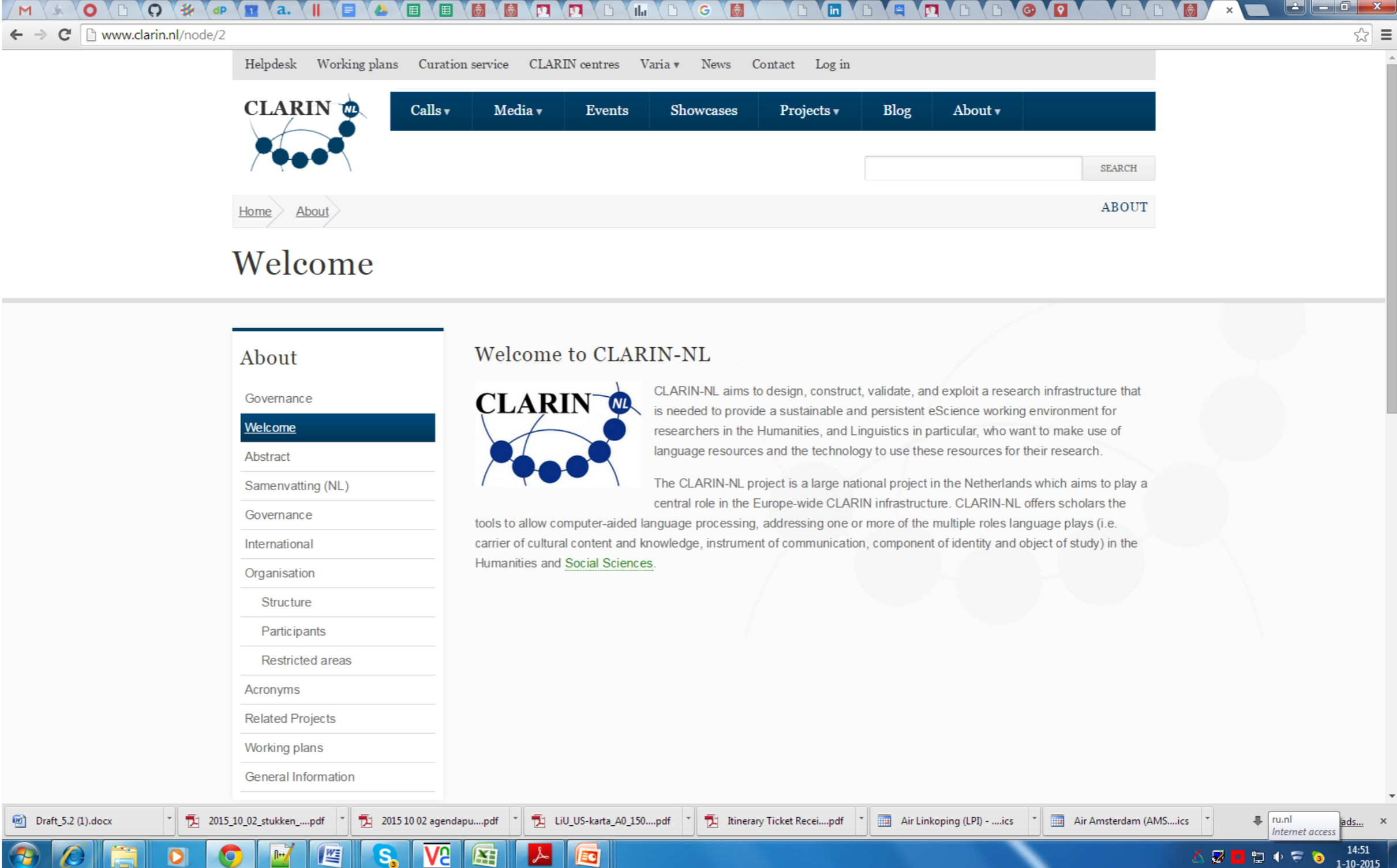


- CLST = Centre for Language and Speech Technology
- Based at Faculty of Arts of the Radboud University
- ± 20 researchers, including PhDs
- Research domains:
 1. Audio and text mining
 2. Language learning and language teaching
 3. Communication in health care
 4. Resources and infrastructure

VALID

=Vulnerability in Acquisition, Language Impairments in Dutch)

Project in
CLARIN-
NL



The screenshot shows the CLARIN-NL website in a browser window. The address bar displays 'www.clarin.nl/node/2'. The navigation menu includes 'Helpdesk', 'Working plans', 'Curation service', 'CLARIN centres', 'Varia', 'News', 'Contact', and 'Log in'. A secondary menu features 'Calls', 'Media', 'Events', 'Showcases', 'Projects', 'Blog', and 'About'. A search bar is located to the right of the secondary menu. Below the navigation, there are breadcrumb links for 'Home' and 'About', and a 'SEARCH' button. The main content area is titled 'Welcome' and features a sidebar with a list of links: 'About', 'Governance', 'Welcome' (highlighted), 'Abstract', 'Samenvatting (NL)', 'Governance', 'International', 'Organisation', 'Structure', 'Participants', 'Restricted areas', 'Acronyms', 'Related Projects', 'Working plans', and 'General Information'. The main text area is titled 'Welcome to CLARIN-NL' and contains the CLARIN-NL logo and a paragraph describing the project's goals: 'CLARIN-NL aims to design, construct, validate, and exploit a research infrastructure that is needed to provide a sustainable and persistent eScience working environment for researchers in the Humanities, and Linguistics in particular, who want to make use of language resources and the technology to use these resources for their research.' Below this, it states: 'The CLARIN-NL project is a large national project in the Netherlands which aims to play a central role in the Europe-wide CLARIN infrastructure. CLARIN-NL offers scholars the tools to allow computer-aided language processing, addressing one or more of the multiple roles language plays (i.e. carrier of cultural content and knowledge, instrument of communication, component of identity and object of study) in the Humanities and [Social Sciences](#).'

VALID: Data collection

(1) The SLI RU-Kentalis data base

Informants: 63 SLI + 24 controls; Characteristics: 56 boys and 31 girls ; 5 – 12; Specific Language Impairment (SLI); Aim data collection: investigation of the expression of spatial relations by children with SLI and normally developing children in their spoken language production.

(2) The UU SLI-Dyslexia project data base

Informants: two longitudinal cohorts: (a) baby's, from 19 months to approximately 37 months; N ≈ 110. (b) toddlers; 3;2 (years; months) at the onset; about 5;0 at the last test session; N ≈ 140; Characteristics: baby cohort: ~70 children at familial risk (FR) of dyslexia; ~40 controls; toddler cohort: ~70 FR children, ~40 controls, ~30 children (tentatively) diagnosed with Specific Language Impairment (SLI).

(3) The bilingual deaf children RU-Kentalis database

Informants: 11 deaf children, longitudinal; Characteristics: 5 boys and 6 girls ; 3 – 6; prelingual deafness (hearing loss of minimally 80dB Fletcher Index on the best ear), no mental restrictions; Aim data collection: investigation of the bilingual language and communication development of young deaf children in Sign Language of the Netherlands and Dutch.

(4) The ADHD and SLI corpus UvA database

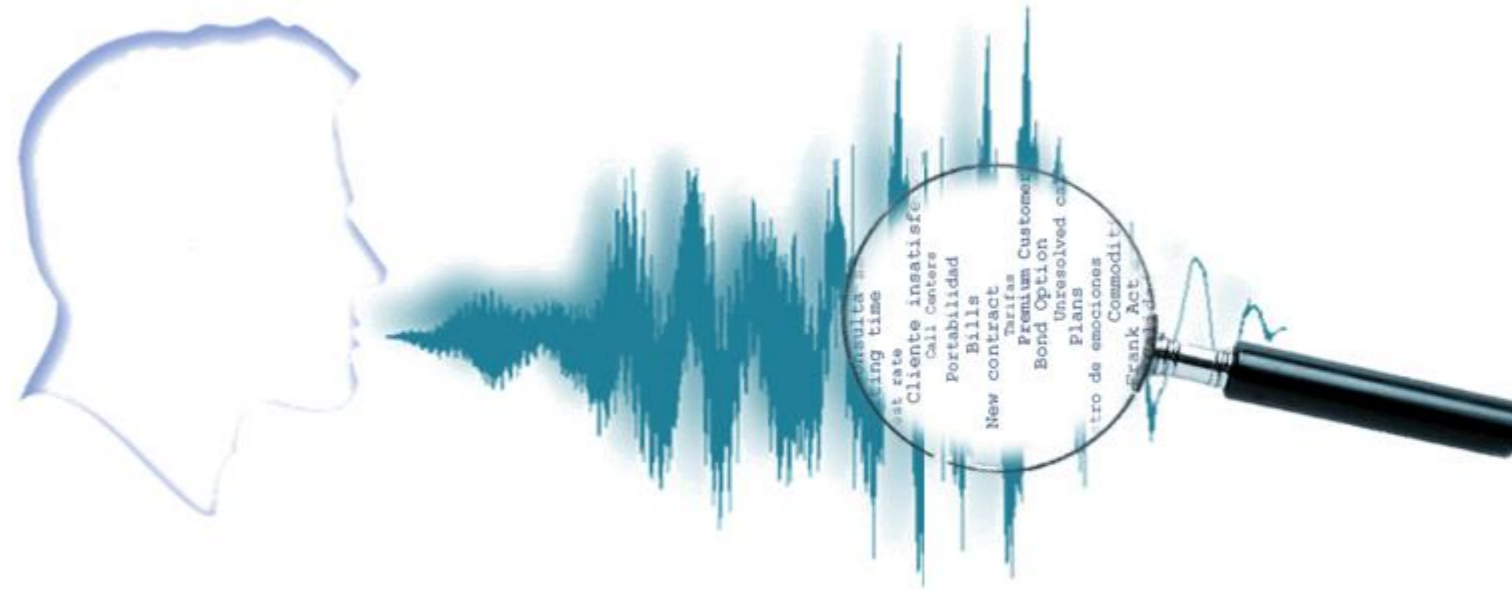
Informants: 26 Dutch children with ADHD, 19 Dutch children with SLI, 22 children Dutch controls; Characteristics: ages between 7 and 8 years; 80% male, 20% female; intelligence within normal ranges; Aim data collection: to compare the language and executive functioning profiles of children with ADHD to children with SLI and TD children.

(5) The deaf adults RU database

Informants: 46 deaf Dutch adults + 38 hearing Turkish adults + 24 hearing Moroccan adults + 10 Dutch controls; Characteristics: males: 22 deaf + 31 Turkish/Moroccan; females: 24 deaf + 31 Turkish/Moroccan; Aim data collection: investigation of the acquisition of Dutch by deaf Dutch adults (late L1/early L2) and comparison to hearing foreign L2-learners of Dutch (late L2) on morphosyntactic aspects.

VALID: Data types

- Audio & Video
- Transcriptions
- Conversation analyses
- Test scores



See: Klatter, J., Van Hout, R., Heuvel, H. van den, Fikkert, P., Baker, A., De Jong J., Wijnen, F., Sanders, E., Trilsbeek, P. (2014). Vulnerability in Acquisition, Language Impairments in Dutch: Creating a VALID Data Archive. In: Proceedings LREC 2014, Reykjavik, 26-31 May 2014. pp. 357-364

VALID: Data conversion/formats

- CLARIN-NL has a restricted set of permitted formats for various file types. For audio and video it was convenient to stick to wav/mp3 and mpg standard formats.
- SPSS files were converted into CSV text files, as were excel files .
- Transcriptions in CHAT, Praat and text formats
- Where deemed appropriate the directory structure of the database was modified, e.g. we put the recordings and transcripts of one participant all in the same subdirectory. Names of directories and files were made as uniform and consistent as possible

VALID: IPR / Consent forms

- Consent forms were signed by participants (or their legal caretakers) for all five databases. The forms arrange that:
 - metadata and transcriptions are anonymized and
 - are accessible for researchers without further consent.
- In some cases, informants were contacted again to check if their initial permission extends to the audio and video recordings upon a motivated request by a researcher.
- Researchers using the database must declare that they have received an aggregated or anonymized form, with an acknowledgment of the data archive plus the specific database(s) used.



VALID: Metadata

- CLARIN works with the Component Metadata Infrastructure (CMDI)
 - VALID obtained a separate CMDI profile which borrowed many building blocks from the [LESLLA](#) profile
 - All metadata categories in the VALID CMDI profile are registered at ISOcat. This profile will be made available via the CMDI registry.
 - An excel file was used as intermediary for the researchers to fill the metadata (see [excel file](#)).
 - Particular attention was given to map the original metadata to the corresponding CMDI category. For each database a file was kept in which these mappings were documented. The files with the original metadata and the file with the mapping information were added as CSV files to the curated database after they were anonymized.
 - A Python script converted the resulting excel file to CMDI metadata files.
-
- <http://www.isocat.org>
 - <http://www.clarin.eu/cmd>

VALID: Persistent Identifiers & Accessibility

- Persistent Identifier (PID): An identifier (URL) that warrants a stable link to a data element
- Every single file and metadata obtained a Handle Persistent Identifier
- This ensures a stable reference that can be used in publications or in other online resources.

- Accessibility is possible through the VALID website
 - <http://validdata.org>
 - and directly through The Language Archive of the [MPI](#)
- Metadata and transcription files are freely accessible
- The audio and video files within VALID are only accessible upon request
 - (Since they were not anonymized)

A VALID future

- VALID is a pilot / a launching platform
- Extension to include more data of pathological speakers is envisaged
- National and international funding is being looked for



Data derived from patient care: a potential large source of research information

“ When people go to their doctor or health professional, they are seeking treatment because they are unwell, and not to become subjects of research”

but

‘Research using personal data has benefited public health by identifying the causes and changing patterns of disease, improving therapeutic practice and the use of health care services, and by indicating promising areas of research’

(Academy of Medical Sciences, 2006)

How may data on distorted speech be derived from patient care?

Illustrated for rehabilitation services:

- Rehabilitation of patients who have severe neurological or musculoskeletal impairments.
- Ultimate goal: to achieve the highest possible level of medical, psychological, social and vocational function either by restoration or compensation of the impairment by means of interdisciplinary treatment.

Rehabilitation may include speech and language therapy (SLT) for the following neurogenic communication disorders resulting in disordered speech:

- **Aphasia**: an impaired ability to understand or produce language, as a result of brain damage.
[example of Broca's aphasia](#)
- **Apraxia of speech**: An disorder in the planning or programming of the movements needed for speech, even though the muscles are not weak and the person 'knows' the word.
[example of apraxia of speech](#)

Assessment of neurogenic communication disorders in the clinical setting: usual care

Pre-therapy assessments in order to:

- Establish diagnosis and prognosis
- Describe and understand all speech language (and other cognitive) components underlying the communication disorder

Resulting personal data (among others):

- Raw test scores
- Audio or video files of (hetro)anamnesis
- Audio files of test items and (semi)spontaneous speech
- Video files of test items and (semi)spontaneous speech

Speech and language therapy (SLT)

Post-therapy assessments (= administrating the pre-therapy assessments again):

- Establish the therapy-induced change (in addition to spontaneous recovery) in function, activity and participation.

Resulting personal data (among others):

- Raw test scores
- Audio files of test items and (semi)spontaneous speech
- Video files of test items and (semi)spontaneous speech

Time

Assessment of neurogenic communication disorders in the clinical setting

Pre-therapy assessments in order to:

- Establish diagnosis and prognosis
- Describe and understand all speech language (and other cognitive) components underlying the communication disorder

Resulting personal data (among others):

- Raw test scores
- Audio or video files of (hetero)anamnesis
- Audio files of test items and (semi)spontaneous speech
- Video files of test items and (semi)spontaneous speech

Speech and language therapy (SLT)

Corpus of disordered speech

Post-therapy assessments (= administering the pre-therapy assessments again):

- Establish the therapy-induced change (in addition to spontaneous recovery) in function, activity and participation.

Resulting personal data (among others):

- Raw test scores
- Audio files of test items and (semi)spontaneous speech
- Video files of test items and (semi)spontaneous speech

Time

Assessment of neurogenic communication disorders in the clinical setting

Pre-therapy assessments in order to:

- Establish diagnosis and prognosis
- Describe and understand all speech language (and other cognitive) components underlying the communication disorder

Resulting personal data (among others):

- Raw test scores
- Audio or video files of (hetero)anamnesis
- Audio files of test items and (semi)spontaneous speech
- Video files of test items and (semi)spontaneous speech

Speech and language therapy (SLT)

Corpus of disordered speech

Post-therapy assessments (= administering the pre-therapy assessments again):

- Establish the therapy-induced change (in addition to spontaneous recovery) in function, activity and participation.

Resulting personal data (among others):

- Raw test scores
- Audio files of test items and (semi)spontaneous speech
- Video files of test items and (semi)spontaneous speech

Time

Corpora of disorder speech based on patient data

The data derived from patient care may be used to establish corpora of speech disorders, especially the **video and audio files** of test items and/or (semi) spontaneous speech. Please note: no other outcomes are gathered than the ones necessary for usual care. You cannot collect any additional data or metadata without entering much more rigid regimes of forced interventions.

One could try to also obtain the questionnaires on [User-Participation](#) (User-P)

Possible advantages for speech and language technologists and pathologists:

- Relatively large numbers of patients can be studied, which allows a better **measure of the inter- and intrapersonal variability in speech characteristics**.
- **Data may be (re)used** for various research purposes

Challenges

Research with personal data faces several challenges, which include:

- **Sensitivity of personal data:** Patient confidentiality should be assured, for example when personal information is disclosed in the (hetero)anamnesis (professional confidentiality)
- **Privacy and autonomy:** patients should be afforded the opportunity to make decisions based on their own values (e.g. 'Use my data to treat me, but not to improve human language technology for others').

Solutions:

- Coded data
- Informed consent
- **Legal and regulatory complexity:** the legislative and regulatory environment 'protects the patient' but is inhibitory to research using personal data, such as requirements on high levels of data security, the burden of creating and managing a research data set is considerable
- **Financial aspects:** Structural costs for making the data shareable. Who will pay for that?

Coded data

Medical research may be carried out on **anonymised data** without consent (provided certain conditions are met).

However, a truly anonymous data set is unlikely to be useful for much research.

That is, research often requires data to be identifiable to some degree (i.e., to ensure that the data cover a valid or representative sample of the population).

It takes a lot of effort (= time and money) to construct pseudonymised (coded) data sets.

Informed consent from the patient is (still) needed for the publication of coded data sets.

Informed consent

The patient should clearly and unambiguously express his/her consent, preferably in writing, for the use of record containing personal medical data for research purposes.

A. Information for consent, includes (among other aspects):

- Aims of the study/project in lay terms, with an outline of the intended benefits
- Summaries of the types of information needed from the patient
- A statement that any aspect of the data collection can be declined without consequences for the medical care received.
- It should be stated as well whether the data can be retained for future research as well (or for one particular research question/ study solely). And in which form.

B. Types of consent:

- Opt-in (active agreement) = often viewed as a better form of consent
- Opt-out (agreement by default unless patients object or choose an alternative) = more efficient
- Extra requirements due to EU's new GDPR as of May 2018

Conclusion

There are both opportunities and challenges to using patient data for scientific research