Henk van den Heuvel, Nelleke Oostdijk, Caroline Rowland, and Paul Trilsbeek

The CLARIN Knowledge Centre for Atypical Communication Expertise

Abstract: In this chapter we introduce the CLARIN Knowledge Centre for Atypical Communication Expertise. The mission of ACE is to support researchers engaged in languages which pose particular challenges for analysis; for this, we use the umbrella term "atypical communication". This includes language use by second-language learners, people with language disorders or those suffering from language disabilities, and languages that pose unique challenges for analysis, such as sign languages and languages spoken in a multilingual context. The chapter presents details about the collaborations and outreach of the centre, the services offered, and a number of showcases for its activities.

Keywords: knowledge centre, atypical communication, sensitive data, data sharing solutions

1 Introduction

Over the past years the European Research Infrastructure for Language Resources and Technology (CLARIN; see clarin.eu) has taken shape (Hinrichs and Krauwer 2014; de Jong et al. 2018; Krauwer and Maegaard 2022). The infrastructure is directed towards researchers in the humanities and social sciences. It provides users with access to distributed data and tools through a single sign-on online environment (de Jong 2019). Apart from its technical infrastructure and accompa-

Acknowledgements: This publication was supported by CLARIN ERIC, and by the SSHOC Project (Grant Agreement 823782 under H2020).

Henk van den Heuvel, CLS/CLST, Radboud University, Nijmegen, the Netherlands,
e-mail: henk.vandenheuvel@ru.nl
Nelleke Oostdijk, CLS/CLST, Radboud University, Nijmegen, the Netherlands,
e-mail: nelleke.oostdijk@ru.nl
Caroline Rowland, Donders Institute for Brain, Cognition & Behaviour, Nijmegen & The Language Archive, MPI for Psycholinguistics, Nijmegen, the Netherlands, e-mail: caroline.rowland@mpi.nl
Paul Trilsbeek, The Language Archive, MPI for Psycholinguistics, Nijmegen, the Netherlands,
e-mail: paul.trilsbeek@mpi.nl

https://doi.org/10.1515/9783110767377-014

nying protocols, CLARIN has been investing in what is referred to as the Knowledge Sharing Infrastructure (KSI).[1] The goal of the KSI is to share knowledge and expertise about the technical infrastructure, the way it operates, and how it can be used, between all stakeholders – from resource and technology providers to end users. In the CLARIN networked organizational structure, the Knowledge (K-)Centres play a central role in this. K-centres advise on issues pertaining to data collection and data management, provide information regarding available resources and services, where to find them, and how to access them, and provide support for various methodologies and applications. K-centres can also offer training courses in their respective fields of expertise.

At present there are over 20 certified K-centres.[2] One of the later additions is the K-centre for Atypical Communication Expertise[3] (ACE for short) which has been established at the Centre for Language and Speech Technology (CLST) at Radboud University.[4] The mission of ACE is to support researchers engaged in languages which pose particular challenges for analysis; for this, we use the umbrella term "atypical communication". This includes language use by second-language learners, people with language disorders or those suffering from language disabilities, and languages that pose unique challenges for analysis, such as sign languages and languages spoken in a multilingual context. It involves multiple modalities (text, speech, sign, gesture) and encompasses different developmental stages. The target audience for ACE includes linguists, psychologists, neuroscientists, computer scientists, speech and language therapists, and education specialists. A recent overview publication about the centre can be found in van den Heuvel, Oostdijk et al. (2020). This chapter is an extension of this publication, elaborating on latest developments.

In Section 2 we will address the collaborations in which the ACE centre is engaged. In Section 3 we highlight the services offered by the centre. Section 4 presents a number of resources as showcases for our work. In Section 5 we illustrate the potential of collaboration in making resources accessible via two CLARIN data centres. Finally, in Section 6 our outreach strategies are outlined.

[1] https://www.clarin.eu/content/knowledge-infrastructure

[2] https://www.clarin.eu/content/knowledge-centres

[3] https://ace.ruhosting.nl/

[4] https://www.ru.nl/clst/

2 Collaboration

Within Radboud University the Knowledge centre has CLST[5] as its core but it also has close links to researchers and research groups within the Centre for Language Studies,[6] with ample expertise in the fields of language acquisition,[7] language learning and therapy,[8] and sign language.[9]

Within CLARIN,[10] CLST has the status of C-centre and as such provides metadata to the infrastructure and enables access to tools and web applications through the Federated Identity services that CLARIN offers.

For hosting data and corpora for atypical communication and making these accessible in a FAIR manner, CLST has established a close collaboration with The Language Archive (TLA). TLA is situated at the Max Planck Institute for Psycholinguistics (MPI) in Nijmegen. As a CLARIN B-centre[11] the goal of TLA is to provide a unique record of how people around the world use language in everyday life. They focus on collecting spoken and signed language materials in audio and video form along with transcriptions, analyses, annotations, and other types of relevant material such as photos and accompanying notes. TLA offers storage of sensitive data (speech, audio, and transcripts) and supports the CMDI[12] metadata framework (see also Windhouwer and Goosen 2022). TLA also supports strong authentication procedures, layered access to data, and persistent identification.

For corpora of speech from people with language disorders the ACE centre works closely together with the DELAD initiative.[13] DELAD stands for Database Enterprise for Language And speech Disorders.[14] DELAD is an initiative for sharing corpora of speech of individuals with communication disorders (CSD) among researchers. This is done in a way that is compliant with EU's General Data Protection Regulation (GDPR),[15] at secure repositories in the CLARIN infrastructure (see also Kamocki, Kelli, and Lindén 2022). DELAD organizes workshops focusing on

[5] https://www.ru.nl/clst/ and https://www.ru.nl/cls/our-research/research-groups/language-speech-technology/

[6] https://www.ru.nl/cls/

[7] https://www.ru.nl/cls/our-research/research-groups/first-language-acquisition/

[8] https://www.ru.nl/cls/our-research/research-groups/language-speech-learning-therapy/

[9] https://www.ru.nl/cls/our-research/research-groups/sign-language-linguistics/

[10] https://www.clarin.eu/content/clarin-centres;
http://roadmap2018.esfri.eu/projects-and-landmarks/browse-the-catalogue/clarin-eric/

[11] https://tla.mpi.nl/resources/

[12] https://www.clarin.eu/content/component-metadata

[13] http://delad.net/

[14] It is also Swedish for "shared".

[15] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

how such corpora can be made shareable with other researchers. (For more information, see Lee et al. 2021). For CSD in particular, DELAD fosters a close collaboration through the ACE centre with CMU's TalkBank / Clinical banks.[16] Our collaboration allows for data to be registered at TalkBank and metadata and landing pages to be obtained at the TalkBank website, whereas the storage of data and the authentication of access to the "raw" data (typically audio and video data) is handled at TLA. Examples of such collaboration are presented in Section 5.

For granting access to sensitive data, the ACE centre is also involved in the SSHOC project,[17] in which one of the tasks is devoted to making an inventory of systems and technologies suitable for conduct research on sensitive data, such as video and audio recordings from data subjects with, for example, speech pathologies. This is relevant for offering various ways of accessing sensitive data stored at central repositories, where they can be downloaded, or at shielded repositories, where they can only be remotely accessed. It is essential for the latter option of remote secure access that the data does not leave a safeguarded place. A user cannot download the data but has to access a secure network where analysis of the data can take place, typically using tools available within the secure network. The user can only download analysis results, which may be subject to inspection by the network or data provider. In this way data leakage is avoided, as well as data corruption. This makes exploration of this type of access very relevant for the sensitive data the ACE centre is often dealing with. In Section 3 we will further address the challenges that the General Data Protection Regulation (GDPR) poses in sharing this type of sensitive data.

In 2021 a new collaboration in the area of sign language was set up with other CLARIN K-centres. This happened on the occasion of a K-centre meeting organized by CLARIN in late 2020. In this meeting it was concluded that eight K-centres were involved in the data collection and research on sign language. As a follow-up, these eight K-centres virtually convened a couple of times in 2021. In these meetings they exchanged information regarding the research topics and infrastructure area in which they were active. Further, the resources of each centre, as offered through CLARIN, were included in their websites, and this was the basis of further ideas for collaboration and proposals for funding. In 2021 this resulted in a Resource Family project for Sign Languages, funded by CLARIN-ERIC[18] and carried out by four of the K-centres specializing in sign languages, and supported by all (see also Lenardič and Fišer 2022). This project will be completed in 2022.

[16] https://talkbank.org/

[17] https://sshopencloud.eu/

[18] https://www.clarin.eu/resource-families

3 Services offered

The mission of the ACE centre is to support researchers engaged in languages which pose particular challenges for collection, annotation and analysis, storage and sharing. This includes language use by second-language learners, people with language disorders or those suffering from language disabilities, and languages that pose unique challenges for analysis, such as sign languages and languages spoken in a multilingual context. It often involves multiple modalities (text, speech, sign, gesture) and encompasses different developmental stages.

Researchers working with these types of data face two particular challenges. First, such data often come with unique privacy and ethical challenges, and researchers need to take particular care to follow the strict rules and procedural requirements imposed by ethical committees and by governments or other relevant organizations. In the European Union, this includes the GDPR (see, for example, van den Heuvel, Kelli et al. 2020). At all stages appropriate measures must be in place to gain informed consent and to prevent unwanted disclosure.

For example, children and people with severe learning disabilities may not be able to give informed consent themselves for data collection and sharing, but rely on consent given by an advocate. In these cases, researchers may not wish to share data widely but to restrict access to registered users, even if the advocate has given consent for sharing (for example to restrict access those who have agreed in writing to keep the participants' identity anonymous and use the data only for academic purposes). With particularly sensitive data, or data in which participants have not given consent for sharing, the original non-anonymized data may need to remain stored in a dark archive, not to be copied or distributed in any form. Resource owners and users thus often need advice about how they can preserve sensitive data in a safe manner, from the point where the raw data came into existence up to the moment where the data and information obtained from it are shared with others.

Moreover, atypical communication data poses unique challenges when it comes to choosing tools and methods for annotation and analysis. Guidelines and tools that have been developed for "standard" data are often inappropriate or require adaptation. Researchers require information about the availability of relevant tools and guidelines such as those presented in Crasborn (2015).

The ACE centre provides the information and advice needed to meet these challenges in three ways. First, it provides advice on data collection and data management. This includes general advice available on the website about relevant issues (for instance, examples of GDPR-compliant consent forms), a helpdesk for specific questions, and individually tailored consultancy for larger projects. For example, the procedure of gaining consent for data collection, analysis, and sharing often requires particular attention when the data itself is very sensitive (for example, videotaped conversations with children with learning disabilities). In these cases, the procedure for gaining informed consent often requires carefully managed conversations as well as participant information sheets and consent forms written in clear, plain language. This ensures that the person giving consent is made fully aware of how the data may be shared and reused, and the manner in which it is kept secure and confidential. Such well-designed procedures not only protect the participants but also maximize the opportunity for data sharing since participants are often more willing to allow data sharing when they understand the conditions under which their data will be stored, protected, and reused.

Second, the ACE centre provides information about the methods and tools available for processing and using the data, and advice about which might best fit particular use cases. For example, the ELAN tool developed by the Language Archive team (ELAN 2020) is particularly well suited to the annotation of sign language data, since it is designed for use with video data, and has a flexible tier system that means that researchers can capture simultaneous face, hand, body, eye, and mouth movements (see for example, the corpus of Dutch Sign Language hosted at The Language Archive here[19]). For projects focussed on the acoustic properties of speech, the PRAAT annotation system may be more appropriate (Boersma and Weenink 2021), since it provides a suite of powerful tools for speech analysis, synthesis, manipulation, and labelling. For projects that require detailed morphosyntactic analysis, the CHILDES CLAN system (MacWhinney 2000) may be more suitable, since it contains an automatic morphosyntactic tagger for a number of languages (see, for example, the VALID collection of data on language impairments in Dutch here[20]). Note that many annotation systems are interoperable, meaning that one could, for example, annotate speech and gesture in ELAN and then convert the file to CLAN format for morphosyntactic tagging.

Third, the ACE centre provides advice on secure long-term data storage, including options for data sharing and the reuse of data. This includes technical assistance for designing, creating, annotating, formatting, and meta-dating, which is crucial because it can be very difficult to interpret, and navigate, unlabelled or badly labelled data collections. For this, the partnership with the archiving experts at The Language Archive is particularly useful. For example, TLA hosts

[19] https://archive.mpi.nl/tla/islandora/object/tla%3A1839_00_0000_0000_0004_DF8E_6

[20] https://archive.mpi.nl/tla/islandora/object/tla%3A1839_00_8C315BC1_AD5E_4348_9A79_A41FE3DE1150

on its website a number of screencasts providing advice about how to create data collections that are labelled and structured in such a way that it facilitates their reuse by other researchers, as well as a detailed deposit manual.[21] For researchers who are not collecting new data themselves but wish to reuse data, the ACE centre also provides information about where to find relevant corpora and datasets.

4 A collection of showcases

The website of ACE presents a number of showcases. We have already alluded to rich corpora of speech from children and adults with language disorders collected in the VALID project (Klatter et al. 2014) and stored at TLA. Within VALID, four existing digital datasets were curated in order to make them available for scientific research in CLARIN-compatible format. The datasets included are:

- SLI RU-Kentalis database, containing around 40 hours of audio and 150,000 transcribed words;
- Bilingual Deaf Children RU-Kentalis database, containing around 9 hours of video and 19,500 transcribed words;
- ADHD and SLI Corpus UvA database, containing around 26 hours of video and 23,000 transcribed words;
- Deaf Adults RU database, containing results of a writing task in ScriptLog format.

More information about these datasets can be found at VALID's web page,[22] which also contains a link to the persistent identifier of the curated datasets at TLA.[23]

Another showcase is the P-MoLL dataset,[24] which is accessible to all registered users of TLA. The project P-MoLL (Modalität von Lernervarietäten im Längsschnitt) was led by Prof. Norbert Dittmar at the Free University in Berlin from 1987 to 1992. It dealt with the study of the acquisition of modality in German as a second language by untutored adult immigrants with Polish or Italian as their native language. The longitudinal data collection covers about two and a half years of the learners' acquisition process. It contains their oral speech production from different elicitation tasks and free conversations with native speak-

[21] https://archive.mpi.nl/forums/c/tla/archiving-info/9
[22] https://validdata.org/clarin-project/datasets/
[23] https://hdl.handle.net/1839/00-8C315BC1-AD5E-4348-9A79-A41FE3DE1150
[24] https://hdl.handle.net/1839/00-0000-0000-0000-4EAB-A

ers and consists of approximately 100 hours of audio, 16 hours of video, and 520,000 transcribed words (Dittmar et al. 1990).

Another example of a well-documented dataset on second-language learning is the LESLLA corpus. LESLLA stands for Literacy Education and Second Language Learning for Adults.[25] The corpus contains speech data of 15 low-educated learners of Dutch as a second language. All of them are women; eight are Turkish, seven Moroccan. (Turks and Moroccans are the two largest immigrant groups in the Netherlands.) At the time of the recordings, they were between 22 and 45 years old. Participants had to carry out five tasks, which all involved spoken language but varied from strictly controlled to semi-spontaneous. In total, the corpus contains around 30 hours of audio and about 180,000 transcribed words. An extensive description of the curated corpus can be found in Sanders, van de Craats et al. (2014). This corpus is also accessible at TLA.[26]

The LeaP (Learning Prosody in a Foreign Language) corpus[27] (Gut 2012) was collected with the goal of studying the acquisition of prosody by non-native speakers of German and English. The German and English parts of the corpus contain audio recordings of 62 and 50 different speakers, respectively, with a wide variety of native languages. The audio recordings (over 12 hours in total) have been transcribed and annotated by hand, resulting in approximately 72,000 transcribed and annotated words. Part-of-speech tagging and lemmatization were carried out automatically. A detailed description of the corpus can be found in the manual that is included.

The Dutch Bilingual Database[28] (Muysken et al. 2008) is another rather substantial collection of data fitting within the scope of ACE and hosted at TLA. It results from a number of projects and research programmes that were directed at investigating multilingualism and comprises data originating from Dutch, Sranan, Sarnami, Papiamentu, Arabic, Berber, and Turkish speakers. In total, it contains over 500 hours of audio recordings, 10 hours of video recordings, and approximately 615,000 transcribed words. It is accessible to any academic user.

Further, TLA also hosts a wealth of sign language corpora. Many of these are carefully annotated using the ELAN annotation software.[29] The Corpus NGT (Nederlandse Gebarentaal / Dutch Sign Language;[30,31] see Crasborn and Zwitser-

[25] https://www.leslla.org/

[26] https://hdl.handle.net/1839/00-37EBCC6D-04A5-4598-88E2-E0F390D5FCE1

[27] https://hdl.handle.net/1839/00-0000-0000-000A-3D5E-1

[28] https://hdl.handle.net/1839/00-0000-0000-0001-4AF0-7

[29] https://tla.mpi.nl/tools/tla-tools/elan/

[30] https://hdl.handle.net/1839/00-0000-0000-0004-DF8E-6

[31] https://www.ru.nl/corpusngtuk/

lood 2008; Crasborn, Zwitserlood, and Ros 2008) is a highly systematically collected dataset of 92 signers of Dutch Sign Language. It contains over 72 hours of dialogues recorded on video from different angles, using a variety of tasks and genres. A significant part of the recordings has been manually annotated using ELAN, with approximately 200,000 annotation tokens in the latest version. Most of the corpus is freely accessible.

Note that many of the language datasets that come under the scope of the ACE centre are not datasets of atypical communication systems. For example, sign languages are not atypical forms of communication. They are mature, complex languages that evolved spontaneously in deaf communities in the same way that spoken languages evolved in hearing communities. However, the collection, analysis, and storage of sign language data poses particular challenges that are often not addressed by standard systems and tools. Thus, the ACE centre also provides resources to support researchers working with sign languages.

5 Exploiting collaborative potential

In this section we address corpora that are made accessible by exploiting the potential of the collaborations in the ACE centre. In Section 2 we mentioned our collaboration with CMU's TalkBank. As a use case for the curation of a dataset, registering it at the TalkBank and storing the primary data (only) at TLA, we processed the Polish Cued Speech Corpus of Hearing-Impaired Children. The corpus contains legacy data of 20 hearing impaired children aged between 8 and 12 years (11 girls and 9 boys) and was kindly provided by Anita Trochimyuk-Lorenc and Katarzyna Klessa from the University of Warsaw (Institute of Applied Polish Studies). The corpus is described in Trochymiuk (2003, 2005). The curation of this dataset involved the creation of CMDI metadata records as well as the creation of a script for normalizing filenames and for converting the text files into CHAT format – including the required metadata headers that could partially be derived from the filenames. A landing page for this collection has been created at TalkBank.[32] The CHAT transcripts have been added to the TalkBank database, and the Handle persistent identifier for the collection containing the audio files in The Language Archive[33] has been added to the landing page, such that users will be able to download them there. Thus, we have created a situation where the corpus can be found via the TalkBank (which is a popular repository for research-

[32] https://phonbank.talkbank.org/access/Clinical/PCSC.html

[33] https://hdl.handle.net/1839/77ea572d-f4c4-48d8-b67b-956f946b59c5

ers of second-language acquisition and special language impairments) whereas the sensitive audio data is on European servers with the appropriate protection measures and licensing arrangements.

Since the structures and systems of the TalkBank and TLA repositories differ quite significantly, a script was created to extract specific file types from collections in the Fedora Commons repository system at TLA and to put those into a structure that can be easily ingested into the TalkBank repository. The script also transforms TLA's metadata into TalkBank metadata, which is relatively straightforward as both are based on the IMDI[34] metadata schema.

A second use case is the archiving of a collection of materials related to the Arezzo neuropsychiatric hospital. This collection consists of recordings and transcripts of interviews by historian Anna Maria Bruzzone with patients of the hospital in the 1970s, as well as a diary written by a patient with schizophrenia from the same hospital. Many of the interviews have been published (Bruzzone 2021). However, the corresponding audio recordings are currently not accessible through an archive. While most patients have passed away now and therefore may technically not be protected under the GDPR, the recordings should be handled with care and with consideration for the patients' relatives. The archiving of this collection is still in the early stages, where the researchers from the University of Siena, which inherited the collection, are determining which parts can be shared anonymously and which parts need more restricted access policies (Nodari, Calamai, and van den Heuvel 2021). A dynamic process is foreseen in which material flagged as not accessible can be released once the required consent is obtained. Moreover, Calamai and colleagues are preparing a fine-grained metadata profile for these recordings, which will be an important additional feature of this collection. As with the Polish Cued Speech Corpus of Hearing-Impaired Children, we will create a landing page at TalkBank and store derived data such as transcriptions there, whereas the original audio recordings will be stored on the servers of the TLA.

6 Reaching out

The target audience for the ACE centre encompasses anyone working with datasets that pose particular challenges for research on language and communication. The audience thus includes linguists, psychologists, neuroscientists, computer scientists, speech and language therapists, and education specialists. The ACE centre provides online resources via its website, a helpdesk for specific ques-

[34] https://tla.mpi.nl/imdi-metadata/

tions, and a bespoke consultancy service for researchers who need more individualized advice.

The focus of ACE's outreach programme is its website, https://ace.ruhosting.nl/, where all information is made available, including links to relevant resources hosted on other sites, such as The Language Archive and TalkBank. However, its services are publicized in a variety of other ways. Its launch in December 2019 was announced via a press release published on both the Radboud University and Max Planck Institute websites. Centre personnel are now disseminating further information and advice via invited presentations and at workshops, as well as via webinars and screencasts published on the website (see Draxler et al. 2022).

A first workshop was held as a webinar and was organized under the auspices of the SSHOC project, due to its close links with a task about secure access to sensitive data in that project. The webinar was held on 14 October 2020 with the title *Sharing Datasets of Pathological Speech*.[35] In this webinar the following topics were addressed:

- progress achieved by the DELAD initiative for sharing corpora of speech disorders (CSD) and the role of the ACE centre;
- GDPR and the ethics of special category data relevant for collecting and sharing CSD;
- how storing and sharing CSD is arranged in a GDPR-compliant way at the Language Archive of the Max Plank Institute for Psycholinguistics and the collaboration with the TalkBank at CMU;
- infrastructure requirements for secure remote access to sensitive research data with diverse legal (for example, social media terms of service), ethical (for instance, children as subjects), and technical (typically audio and video) challenges, and assessment of several existing platforms;
- the CAVA audio-visual human communication archive project – a digital video repository to support the work of the international human communication research community, which enhances the discoverability and reusability of expensively created specialist video content;
- the curation and disclosure of pathological speech corpora: how CSD can be found through one organization and made accessible through another; this includes a demonstration using the example of the Polish Cued Speech Corpus of Hearing-Impaired Children, as discussed above.

[35] https://www.sshopencloud.eu/sshoc-webinar-sharing-datasets-pathological-speech

The webinar has been recorded and published on YouTube.[36] The slides are available on Zenodo.[37] A report in the form of webinar notes is available via the Social Sciences and Humanities Open Cloud.[38]

On 27 and 28 January 2021 DELAD organized a workshop entitled *How to Share Your Data in a GDPR-Compliant Way*. In this workshop, a number of researchers presented the corpora they collected and the research carried out with them. The central question here was how the data were or could be shared with other researchers. Further presentations addressed:

- the potential of the ACE centre for hosting CSD of DELAD members;
- exchanging deeper insights on Data Protection Impact Assessments (DPIAs), including role play;
- presenting and discussing voice conversion as a means to pseudonymize speech.

The DPIA and role play was led by a member of CLARIN's Committee for Legal and IPR issues (CLIC).[39] A report on the workshop was published by CLARIN[40] and all materials are available via Zenodo.[41] An educational version of the DPIA role play was recorded, published, and presented at the CLARIN Annual Conference 2021.[42]

The ACE centre was also featured at the TOK day in Nijmegen in December, 2021 (the annual meeting of the TaalOntwikkeling van Kinderen network of researchers and speech and language therapists from the Netherlands and Belgium). Printed materials such as posters, leaflets, and a one-page briefing document will be created ready for dissemination when in-person events resume after the Covid-19 pandemic.

[36] https://www.youtube.com/watch?v=qjTJ4ZxzfvI

[37] https://zenodo.org/record/4081602#.X42YC9Azba8

[38] https://www.sshopencloud.eu/news/webinar-notes-sharing-datasets-pathological-speech

[39] The roleplay can be found at https://sites.google.com/rug.nl/privacy-in-research/cases

[40] https://www.clarin.eu/blog/outcomes-fifth-delad-workshop

[41] https://zenodo.org/record/4560478#.YEeAEJ1Ki71

[42] All materials can be found via this link: https://delad.ruhosting.nl/wordpress/dpia-role-play-with-video/

Bibliography

Broersma, Paul & David Weenink. 2021. *Praat: doing phonetics by computer* [Computer program]. Version 6.1.41. http://www.praat.org/ (accessed 25 March 2021).

Bruzzone, Anna Maria. 2021. *Ci chiamavano matti. Voci dal manicomio (1968–1977)*. Milan: Il Saggiatore.

Crasborn, Onno 2015. Transcription and notation methods. In Eleni Orfanidou, Bencie Woll, & Gary Morgan (eds.), *Research methods in sign language studies: A practical guide*, 74–88. Chichester: John Wiley & Sons.

Crasborn, Onno & Inge Zwitserlood. 2008. The Corpus NGT: An online corpus for professionals and laymen. In Onno Crasborn, Thomas Hanke, Eleni Efthimiou, Inge Zwitserlood & Ernst Thoutenhoofd (eds.), *Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages: Construction and exploitation of sign language corpora,* 44–49. Paris: ELRA.

Crasborn, Onno, Inge Zwitserlood & Johan Ros. 2008. The Corpus NGT. A digital open access corpus of movies and annotations of sign language of the Netherlands. Centre for Language Studies, Radboud Universiteit Nijmegen. ISLRN175-346-174-413-3.https://hdl.handle.net/hdl:1839/00-0000-0000-0004-DF8E-6

Dittmar, Norbert, Astrid Reich, Romuald Skiba, Magdalena Schumacher & Heiner Terborg. 1990. Die Erlernung modaler Konzepte des Deutschen durch erwachsene polnische Migranten: Eine empirische Längsschnittstudie. *Informationen Deutsch als Fremdsprache: Info DaF* 17 (2). 125–172.

Dittmar, Norbert, Astrid Reich, Romuald Skiba, Magdalena Schumacher & Heiner Terborg. 2002. The P-MoLL Corpus. https://hdl.handle.net/1839/00-0000-0000-0000-4EAB-A

Draxler, Christoph, Alexander Geyken, Erhard Hinrichs, Annette Klosa-Kückelhaus, Elke Teich & Thorsten Trippel. 2022. How to connect language resources, infrastructures, and communities. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources*. Berlin: deGruyter.

ELAN (Version 6.0) [Computer software]. 2020. Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from https://archive.mpi.nl/tla/elan

Emmerik. Joanne van. 2014. Deaf Adults RU Database. ISLRN 944-022-313-325-3. https://hdl.handle.net/1839/00-97AF29EA-877D-422A-BAF7-25FA269351A6

Gut, Ulrike. 2009. LeaP Corpus. https://hdl.handle.net/1839/00-0000-0000-000A-3D5E-1

Gut, Ulrike. 2012. The LeaP corpus. A multilingual corpus of spoken learner German and learner English. In Thomas Schmidt and Kai Wörner (eds.), *Multilingual corpora and multilingual corpus analysis*, 3–23. Amsterdam: John Benjamins.

Heuvel, Henk van den, Nelleke Oostdijk, Caroline Rowland & Paul Trilsbeek. 2020. The CLARIN knowledge centre for atypical communication expertise. *International Conference on Language Resources and Evaluation (LREC)* 12. 3312–3316.

Heuvel, Henk van den, Aleksei Kelli, Katarzyna Klessa & Satu Salaasti. 2020. Corpora of disordered speech in the light of the GDPR: Two use cases from the DELAD initiative. *International Conference on Language Resources and Evaluation (LREC)* 12. 3317–3321.

Hinrichs, Erhard & Steven Krauwer. 2014. The CLARIN research infrastructure: Resources and tools for ehumanities scholars. *International Conference on Language Resources and Evaluation (LREC)* 9. 1525–1531.

Jong, Franciska de. 2019. CLARIN: Infrastructural support for impact through the study of language as social and cultural data. In Bente Maegaard, Riccardo Pozzo, Alberto Melloni and Matthew Woollard (eds). *Stay tuned to the future: Impact of the research infrastructures for social sciences and humanities* (Lessico intellettuale Europeo 128), 121–129. Rome: Leo Olschki.

Jong, Franciska de, Bente Maegaard, Koenraad De Smedt, Darja Fišer & Dieter Van Uytvanck. 2018. CLARIN: Towards FAIR and responsible data science using language resources. *International Conference on Language Resources and Evaluation (LREC)* 11. 3259–3264.

Kamocki, Paweł, Aleksei Kelli & Krister Lindén. 2022. The CLARIN Committee for Legal and Ethical Issues and the Normative Layer of the CLARIN infrastructure. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources*. Berlin: deGruyter.

Klatter, Jetske, Roeland van Hout, Henk van den Heuvel, Paula Fikkert, Anne Baker, Jan de Jong, Frank Wijnen, Eric Sanders & Paul Trilsbeek. 2014. Vulnerability in acquisition, language impairments in Dutch: Creating a VALID data archive. *International Conference on Language Resources and Evaluation (LREC)* 9. 356–364

Kolen, Esther. 2014. Bilingual Deaf Children RU-Kentalis Database. ISLRN 941-351-623-486-4. https://hdl.handle.net/1839/00-F6BC06C4-B2AD-4ED8-8527-AB81F4EF4E8F

Krauwer, Steven & Bente Maegaard. 2022. CLARIN – how it started. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources*. Berlin: deGruyter.

Lee, Alice, Nicola Bessell, Henk van den Heuvel, Satu Saalasti, Katarzyna Klessa, Nicole Müller & Martin J. Ball. 2021. The latest development of the DELAD project for sharing corpora of disordered speech. *Clinical Linguistics & Phonetics*. https://doi.org/10.1080/02699206.2021.1913514

Lenardič, Jakob & Darja Fišer. 2022. The CLARIN Resource and Tool Families. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources*. Berlin: deGruyter.

Lorenc, Anita. 2019. Polish Cued Speech Corpus of Hearing-Impaired Children. https://hdl.handle.net/1839/77ea572d-f4c4-48d8-b67b-956f946b59c5

MacWhinney, Brian. 2000. *The CHILDES Project: Tools for analyzing talk*, 3rd edn. Mahwah, NJ: Lawrence Erlbaum Associates.

Made, Annika van der. 2014. SLI RU-Kentalis Database. ISLRN 541-534-411-504-6. https://hdl.handle.net/1839/00-712802F3-C245-4EF0-BE9D-D09714DEDE67

Muysken, Pieter, et al. 2008. Dutch Bilingual Database. https://hdl.handle.net/1839/00-0000-0000-0001-4AF0-7

Nodari, Rosalba, Silvia Calamai & Henk van den Heuvel. 2021. Less is more when FAIR. The Minimum Level of Description in Pathological Oral and Written Data. In Monica Monachini & Maria Eskevich (eds.), *CLARIN Annual Conference Proceedings, 2021*, 166–171. Virtual edition.

Parigger, Esther. 2014. ADHD and SLI Corpus UvA database. ISLRN 456-360-189-350-0. https://hdl.handle.net/1839/00-2766F32F-4305-4F13-A02C-F4A8F5216425

Sanders, Eric, Ineke van de Craats & Vanja de Lint. 2014. The curated Dutch LESLLA corpus. https://hdl.handle.net/1839/00-37EBCC6D-04A5-4598-88E2-E0F390D5FCE1

Trochymiuk, Anita. 2003. Voiced realisations of plosives in word initial position by hearing impaired children: Acoustic phonetics analysis. In Katharina Böttger, Sabine Dönninghaus & Robert Marzari (eds.), *Die Welt der Slaven*. Vol. 16 (Beiträge der Europäischen Slavistischen Linguistic 6). 111–123. Munich: Sagner.

Trochymiuk Anita. 2005. Realization of the voiced-voiceless contrast by hearing impaired children. *Studia Phonetica Posnaniensia* 7. 75–96.

Sanders, Eric, Ineke van de Craats & Vanja de Lint. 2014. The Dutch LESLLA corpus. *International Conference on Language Resources and Evaluation (LREC)* 9. 2715–2718.

Windhouwer, Menzo & Twan Goosen. 2022. Component Metadata Infrastructure. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources*. Berlin: deGruyter.