Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3312–3316
Marseille, 11–16 May 2020

The CLARIN Knowledge Centre for Atypical Communication Expertise

Henk van den Heuvel[1], Nelleke Oostdijk[1], Caroline Rowland[2,3], Paul Trilsbeek[2]

[1]CLS/CLST, Radboud University, Nijmegen
[2]The Language Archive, MPI for Psycholinguistics, Nijmegen
[3]Donders Institute for Brain, Cognition & Behaviour, Nijmegen
{h.vandenheuvel, n.oostdijk}@let.ru.nl, {caroline.rowland, paul.trilsbeek}@mpi.nl

Abstract

This paper introduces a new CLARIN Knowledge Center which is the K-Centre for Atypical Communication Expertise (ACE for short) which has been established at the Centre for Language and Speech Technology (CLST) at Radboud University. Atypical communication is an umbrella term used here to denote language use by second language learners, people with language disorders or those suffering from language disabilities, but also more broadly by bilinguals and users of sign languages. It involves multiple modalities (text, speech, sign, gesture) and encompasses different developmental stages. ACE closely collaborates with The Language Archive (TLA) at the Max Planck Institute for Psycholinguistics in order to safeguard GDPR-compliant data storage and access. We explain the mission of ACE and show its potential on a number of showcases and a use case.

Keywords: infrastructure, atypical communication, language resources.

1. Background and Aims

Over the past years the European Research Infrastructure for Language Resources and Technology (CLARIN; clarin.eu) has taken shape (Hinrichs et al., 2014; De Jong et al., 2018). The infrastructure is directed towards researchers in the humanities and social sciences. It provides users access to distributed data and tools through a single sign-on online environment (De Jong, 2019). Apart from the technical infrastructure and accompanying protocols, CLARIN has been investing in what is referred to as the Knowledge Sharing Infrastructure (KSI)[1]. The KSI should ensure that knowledge and expertise as regards the technical infrastructure, the way it operates and how it can be used, is shared 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 the dissemination of (specialized) knowledge and expertise. K-centres can advise on issues pertaining to data collection and data management, can provide information as regards available resources and services, where to find 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 20 certified K-centres[2]. One of the latest additions is the K-Centre for Atypical Communication Expertise (ACE for short) which has been established at the Centre for Language and Speech Technology (CLST) at Radboud University. The mission of ACE is to support researchers engaged in investigating what can be characterised as atypical communication. Atypical communication is an umbrella term used here to denote language use by second language learners, people with language disorders or those suffering from language disabilities, but also to languages that pose particularly difficult issues 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. The website of ACE is: https://ace.ruhosting.nl/

Data originating in a context of atypical communication are particularly sensitive as regards privacy and ethical issues. While collecting, storing, processing and using such data, researchers are bound by strict rules and procedural requirements imposed by ethical committees and the GDPR (see e.g. van den Heuvel et al., 2020).

At all stages appropriate measures must be in place so as to prevent unwanted disclosure. In some cases, this requires that the original data remain stored in a dark archive and cannot be copied or distributed in any form. ACE can advise resource owners and users on how they can preserve sensitive data in a safe manner, from the point where the raw data come into existence up to the moment where the data and information obtained from it are shared with others.

Atypical communication data are also special when it comes to the methods and tools for processing and using the data. Often guidelines and tools that have been developed and are used for standard data cannot be used or require adaptations or special settings; in some other cases dedicated tools are available. ACE is well-positioned to inform researchers who want to work with language development data, data of adults and children with speech disorders, or users of sign language on the availability of such tools and guidelines. ACE can advise on what is feasible and how to go about it.

2. A Fruitful Partnership

Within Radboud University the Knowledge Centre has CLST[3] as its core but it has close links to researchers and research groups within the Centre for Language Studies[4]

[1] https://www.clarin.eu/content/knowledge-sharing
[2] https://www.clarin.eu/content/knowledge-centres
[3] https://www.ru.nl/clst/ and https://www.ru.nl/cls/our-research/research-groups/language-speech-technology/
[4] https://www.ru.nl/cls/

with ample expertise in the fields of language acquisition[5], language learning and therapy[6], and sign language[7].
Within CLARIN, CLST has the status of C Centre and Trust 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[8] 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[9] metadata framework. TLA also supports strong authentication procedures, layered access to data, and persistent identification.
For corpora of speech from people with language disorders ACE works closely together with the DELAD initiative[10]. Especially for this type of resources there is a close collaboration with CMU's Talkbank / Clinical banks[11]. Our collaboration allows that data can be registered at Talkbank and obtains its metadata and landing page at the Talkbank website whereas the storage of and authentication of access to the 'raw' data (typical audio and video) data is handled at TLA (see also Section 4).
For giving access to critical data ACE is also involved in the SSHOC project[12] in which Task 5.4 is devoted to making an inventory of systems and technologies suitable to conduct research on critical data which is relevant for offering various ways of accessing critical data stored at central repositories where they can be downloaded or at shielded repositories where they can only be remotely accessed.

3. Services Offered

ACE will offer the following services through its website:

- Information and guidelines about:
  - consent (forms)
  - hosting corpora and datasets containing atypical communication
  - where to find corpora and datasets containing atypical communication
- Helpdesk/consultancy for questions on the above topics
- Technical assistance for designing, creating, annotating, formatting and metadating resources of atypical communication
- Outreach: presentations, workshops contributions, etc.

The items in the list above meet a number of demands for which researchers have expressed a need. Typically, assistance in designing and collecting corpora containing atypical communication with consent forms that are GDPR-proof, is considered of great value, as are references to available guidelines and tools for annotating such resources. How to make the resources accessible and share them with other researchers is another issue for which special expertise is requested.
ACE is happy to advise on such issues and also to participate in projects where the acquisition and/or creation of such data collections is foreseen.

4. Show Cases

The website of ACE presents a number of show cases. We mention the rich corpora of speech from children and adults with language disorders collected in the VALID project (Klatter at 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 https://validdata.org/clarin-project/datasets/. This page also contains a link to the persistent identifier of the curated datasets at TLA[13].
Another show case is the P-MoLL dataset[14], which is accessible to all registered users of TLA. The project P-Moll (=Modalität von Lernervarietäten im Längsschnitt) was run at the Free University in Berlin by Prof. Norbert Dittmar 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 speakers and consists of approximately 100 hours of audio, 16 hours of video and 520,000 transcribed words (Dittmar et al., 1990).

[5] https://www.ru.nl/cls/our-research/research-groups/first-language-acquisition/
[6] https://www.ru.nl/cls/our-research/research-groups/language-speech-learning-therapy/
[7] https://www.ru.nl/cls/our-research/research-groups/sign-language-linguistics/
[8] https://tla.mpi.nl/resources/
[9] https://www.clarin.eu/content/component-metadata
[10] http://delad.net/
[11] https://talkbank.org/
[12] https://sshopencloud.eu/
[13] https://hdl.handle.net/1839/00-8C315BC1-AD5E-4348-9A79-A41FE3DE1150
[14] https://hdl.handle.net/1839/00-0000-0000-0000-4EAB-A

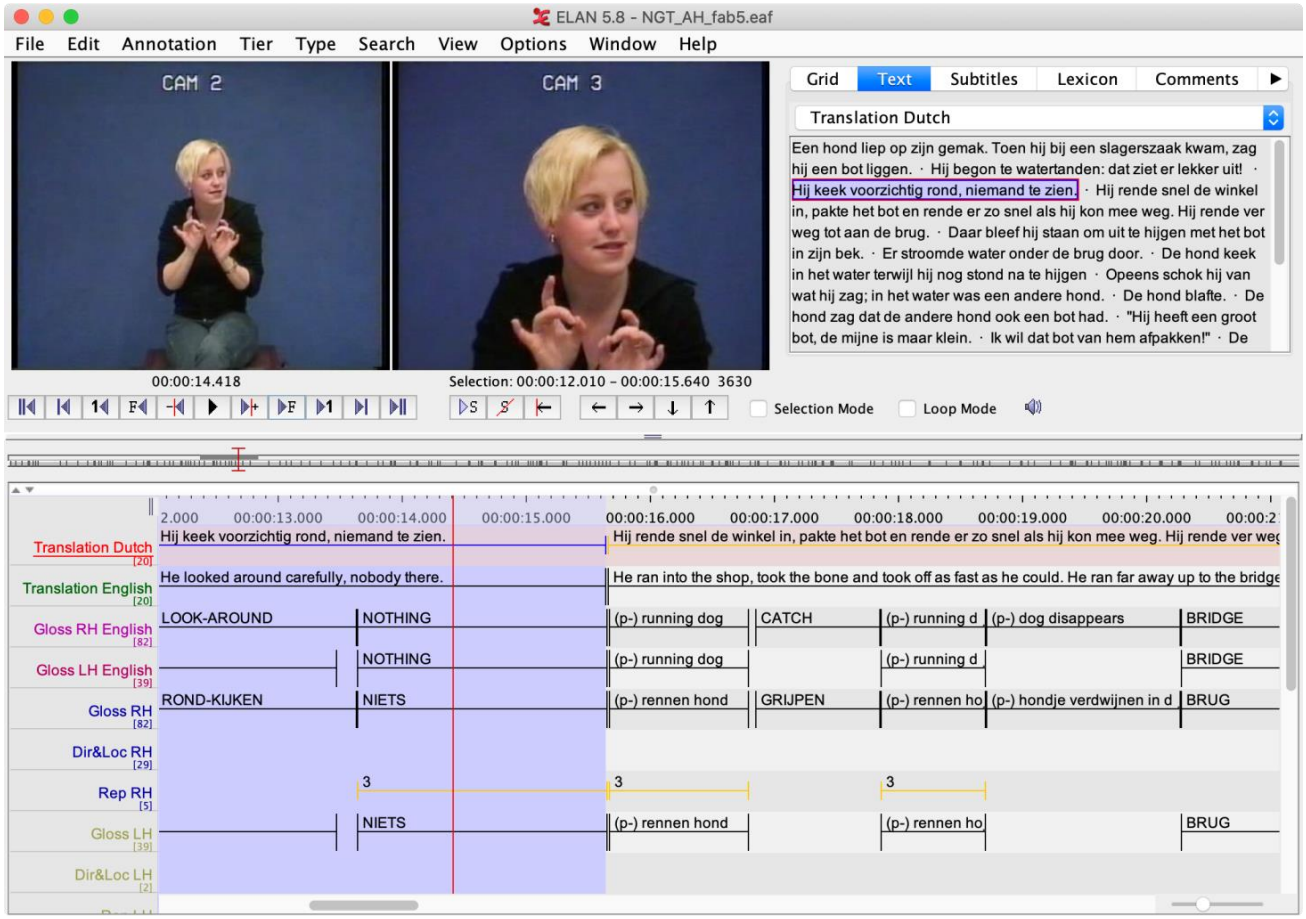

Figure 1. The ELAN multimedia annotation tool that is produced at The Language Archive is widely used for the transcription, annotation and analysis of all sorts of language recordings, including sign languages.

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, see https://www.leslla.org/. The corpus contains speech of 15 low-educated learners of Dutch as a second language. All of them are women; 8 are Turkish, 7 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 & De Lint (2014). This corpus is also accessible at TLA[15].

The LeaP (Learning Prosody in a Foreign Language) corpus[16] 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 more than 12 hours of audio recordings are 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[17] is another rather substantial collection of data fitting in 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[18]. Figure 1 shows an example. The Corpus NGT (Nederlandse Gebarentaal / Dutch Sign Language)[19][20] 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

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

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

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

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

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

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

the latest version. The largest part of the corpus is freely accessible.

One could debate in how far sign language is a form of atypical communication. In our view what makes the language atypical is that it results from a way of dealing with a (hearing) deficiency; the non-atypical part is that sign language as such is a mature version of language as any other. This makes it different from the true atypical variants.

5. Use Case

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 A. Trochimyuk-Lorenc and K. Klessa from the University of Warsaw (Institute of Applied Polish Studies). The corpus is described in Trochymiuk (2003, 2007). 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. Once the CHAT transcripts have been added to the Talkbank database, the Handle persistent identifier to the collection containing the audio files in The Language Archive[21] will be added to the landing page, such that users will be able to download them there.

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[22] metadata schema.

6. Reaching out

The ACE Centre's services will be publicised in a variety of ways. Its launch in December 2019 was announced via a press release published on both the Radboud University and Max Planck Institute websites. After launch, all information about the ACE was made available via its website: https://ace.ruhosting.nl/. Advice and personalised information will be provided via the helpdesk. Centre personnel will further dissemination information and advice via invited presentations and at workshops as well as via webinars and screencasts published on the website. The DELAD network is preparing a workshop with CLARIN in June/July in which the opportunities that ACE offers for researchers studying language disorders will be a main theme.

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

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

7. Bibliographical References

Crasborn, O. & Zwitserlood, I. (2008) The Corpus NGT: an online corpus for professionals and laymen, In: Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages,O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitserlood & E. Thoutenhoofd, eds. ELDA, Parijs. pp 44-49.

De Jong, F., Maegaard, B., De Smedt, K., Fišer, D. and Van Uytvanck,D. (2018). CLARIN: Towards FAIR and Responsible Data Science Using Language Resources. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, May 2018, pp. 3259-3264.

De Jong, F. (2019). CLARIN – Infrastructural support for impact through the study of language as social and cultural data. In B. Maegaard, R. Pozzo, A. Melloni and M. Woollard (Eds.), *Stay Tuned to the Future. Impact of the research infrastructures for social sciences and humanities*, Lessico Intellettuale Europeo, LIE-CXXVIII, 121-129.

Dittmar, N., Reich, A., Skiba, R., Schumacher, M., & Terborg, H. (1990). Die Erlernung modaler Konzepte des Deutschen durch erwachsene polnische Migranten: Eine empirische Längsschnittstudie. In: Informationen Deutsch als Fremdsprache: Info DaF 17(2), pp. 125-172.

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. Amsterdam: Benjamins, pp. 3-23.

Hinrichs, Erhard & Steven Krauwer (2014). The CLARIN Research Infrastructure: Resources and Tools for E-Humanities Scholars. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, May 2014, pp. 1525–31.

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 of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, May 2014, pp. 1525–31.

Trochymiuk A. (2003). Voiced Realisations of Plosives in Word Initial Position by Hearing Impaired Children. Acoustic Phonetics Analysis. In K. Böttger., S. Dönninghaus., & R. Marzari. (Eds), *Die Welt der Slaven*, Band 16, Beiträge der Europäischen Slavistischen Linguistic, Band 6, München, pp. 111–123

Trochymiuk A. (2005). Realization of the voiced-voiceless contrast by hearing impaired children, *Studia Phonetica Posnaniensia*, vol. 7, pp. 75–96.

Sanders, E., Van de Craats, I, De Lint, V. (2014). The Dutch LESLLA Corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, May 2014, pp. 2715-2718.

Van den Heuvel, H., Kelli, A., Klessa, K., Salaasti, S. (2020). Corpora of disordered speech in the light of the GDP: Two use cases from the DELAD initiative. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC2020).*

8. Language Resource References

Crasborn, O., Zwitserlood, I. & Ros, J. (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. URL: http://hdl.handle.net/hdl:1839/00-0000-0000-0004-DF8E-6. ISLRN: 175-346-174-413-3.

Dittmar, N., Reich, A., Skiba, R., Schumacher, M., & Terborg, H. (2002) The P-MoLL corpus. Distributed by The Language Archive: https://hdl.handle.net/1839/00-0000-0000-0000-4EAB-A

Emmerik. J. van (2014) Deaf adults RU database. ISLRN 944-022-313-325-3. https://hdl.handle.net/1839/00-97AF29EA-877D-422A-BAF7-25FA269351A6

Gut, U. (2009) LeaP corpus. Distributed by The Language Archive: https://hdl.handle.net/1839/00-0000-0000-000A-3D5E-1

Kolen, E. (2014) Bilingual deaf children RU-Kentalis database. ISLRN 941-351-623-486-4. https://hdl.handle.net/1839/00-F6BC06C4-B2AD-4ED8-8527-AB81F4EF4E8F.

Lorenc, A. (2019) Polish Cued Speech Corpus of Hearing-Impaired Children, Distributed by The Language Archive: https://hdl.handle.net/1839/77ea572d-f4c4-48d8-b67b-956f946b59c5

Made, A. 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 et al. (2008) Dutch Bilingual Database. Distributed by The Language Archive : https://hdl.handle.net/1839/00-0000-0000-0001-4AF0-7

Parigger, E. (2014) ADHD and SLI corpus UvA database. ISLRN 456-360-189-350-0. https://hdl.handle.net/1839/00-2766F32F-4305-4F13-A02C-F4A8F521642

Sanders, E., Van de Craats, I, De Lint, V. (2014) The curated Dutch LESLLA corpus. Distributed by CLARIN via The Language Archive: https://hdl.handle.net/1839/00-37EBCC6D-04A5-4598-88E2-E0F390D5FCE1