Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3317–3321
Marseille, 11–16 May 2020

Corpora of Disordered Speech in the Light of the GDPR: Two Use Cases from the DELAD Initiative

Henk van den Heuvel[1], Aleksei Kelli[2], Katarzyna Klessa[3], Satu Salaasti[4]
[1]CLS / CLST, Radboud University, Nijmegen, the Netherlands; [2]School of Law, University of Tartu, Estonia;
[3]Inst. of Applied Linguistics, Adam Mickiewicz University in Poznan, Poland
[4]Department of Psychology and Logopedics, University of Helsinki, Finland

h.vandenheuvel@let.ru.nl, aleksei.kelli@ut.ee, klessa@amu.edu.pl, satu.saalasti@helsinki.fi

Abstract

Corpora of disordered speech (CDS) are costly to collect and difficult to share due to personal data protection and intellectual property (IP) issues. In this contribution we discuss the legal grounds for processing CDS in the light of the GDPR, and illustrate these with two use cases from the DELAD context. One use case deals with clinical datasets and another with legacy data from Polish hearing-impaired children. For both cases, processing based on consent and on public interest are taken into consideration.

Keywords: personal data, special categories of personal data, GDPR, language and speech disorders

1. DELAD and Corpora of Disordered Speech

Corpora of disordered speech (CDS) are hard to obtain. They are costly to collect and difficult to share due to personal data protection and intellectual property[1] (IP) 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. A strong need is felt by the research community to bring together existing and new CDS in an interoperable and consistent way that is both legal and ethically safeguarded. The CLARIN infrastructure is regarded as indispensable for this purpose. The CHILDES Talkbank, CMU also being a CLARIN Centre, is an important asset of this infrastructure following US legislation. CDS can be federatively archived at local CLARIN centres whereas they can be made findable through a central portal via their (harvested) metadata.

DELAD[2] (=SHARED in Swedish) is an initiative to establish a digital archive of disordered speech and share this with interested researchers within CLARIN. DELAD has organised four workshops over the years 2015-2019, the latter two of which were held under the umbrella of CLARIN ERIC. Topics addressed in these workshops were: Guidelines for collecting and sharing CDS (in the light of the GDPR[3]), levels of anonymisation, layered access, integration of CDS in the CLARIN infrastructure, formats, and relevant metadata. The DELAD community consist of researchers involved in collecting and analysing CDS, research data and infrastructure specialists, and legal experts. DELAD has chosen the CLARIN infrastructure as primary space for storing and sharing CDS. More specifically, DELAD has linked up with CLARIN's Knowledge Centre for Atypical Communication Expertise (ACE)[4] (Van den Heuvel, et al., 2020) for making CDS available through The Language Archive (TLA)[5] at the Max Planck Institute in Nijmegen (being a CLARIN Data Centre) and CMU's Talkbank[6] (Clinical Banks). In the last workshop in Utrecht in January 2019 CLARIN's Legal and Ethical Issues Committee (CLIC7) was invited to engage with the participants on the GDPR applied to a range of use cases. This paper is a reflection and an elaboration of this discussion.

2. The GDPR and Its Implications for Collecting, Processing and Sharing CDS

CDS contain personal data. Personal data is defined as "any information relating to an identified or identifiable natural person ('data subject')" (GDPR Art. 4 (1)). The GDPR differentiates between 'regular' and special categories of data (sensitive data). Special categories of personal data include inter alia data concerning health (Art. 9 (1)). It can be concluded that CDS contain special categories of personal data.

Processing[8] of sensitive data is subject to more stringent requirements. As a general principle, the processing of sensitive data is prohibited (GDPR Art. 9 (1)). There are specific legal grounds when processing is allowed. Within the context of this paper, consent and research in public interest are relevant. However, there are no clear guidelines how to choose between these two grounds of processing (for further discussion, see Linden et al. 2019).

[1] Due to the focus of this paper IP issues are not addressed here. However, they might have significant impact on the use and dissemination of research data.

[2] http://delad.net

[3] EU General Data Protection Regulation, https://gdpr-info.eu/

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

[5] https://tla.mpi.nl/

[6] https://talkbank.org/

[7] https://www.clarin.eu/governance/legal-issues-committee

[8] The GDPR defines processing extensively so that it covers all possible operations (collecting, structuring, changing, using, deleting and so forth) with personal data (See Art. 4 (2)).

The GDPR provides that processing of sensitive data is allowed if "the data subject has given explicit consent to the processing of those personal data for one or more specified purposes" (Art. 9 (2) a). Consent as a key concept of the GDPR is "any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her" (Art. 4 (11)). There is no valid consent if it is not given freely (some circumstances forced the data subject to consent), or the data subject does have enough information (for further explanation see WP29 2017).

Consent for processing sensitive data must be explicit and given for specified purposes. The guidelines concerning consent explain that "the data subject must give an express statement of consent. An obvious way to make sure consent is explicit would be to expressly confirm consent in a written statement" (WP29 2017: 18).

Processing based on consent has advantages. Firstly, it leads to a higher degree of privacy protection. Secondly, the data subject can consent to public dissemination of his/her data. In order to limit potential disputes and liabilities relating to processing of personal data, possible uses of personal data (including commercial uses and sharing) can be described in consent forms.

Consent as a legal ground for processing personal data for research purposes has also challenges. For example:

1) it involves a certain amount of uncertainty. According to the GDPR, the data subject can withdraw his or her consent at any time without detriment (Art. 7 (3), WP29 2017: 21);

2) the acquisition and management of consents involves considerable administrative burden;

3) it is not always suitable for legacy data.[9]

The other legal ground to process sensitive personal data (speech disorders data) for research is to rely on public interest. The GDPR provides that sensitive personal data can be processed if it "is necessary for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes" (Art. 9 (2) j).

Processing without consent (for further discussion on general framework and national examples, see Kelli et al. 2019) does not mean that the data subject's rights do not have to be honoured. For instance, according to the GDPR, the data subject has to be informed of processing. This gives rise to a question concerning the processing of legacy data because the fulfilment of this obligation is complicated or impossible. The GDPR foresees the situation and provides that the data subject does not have to be informed if "the provision of such information proves impossible or would involve a disproportionate effort, in particular for processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes" (Art. 14 (5) b).

The GDPR establishes a general framework for processing data for research purposes. The GDPR defines research extensively so that it includes "technological development and demonstration, fundamental research, applied research and privately funded research" (Recital 159). The GDPR also requires processing for research purposes to be subject to appropriate safeguards which ensure technical and organisational measures (Art. 89).

The use of an appropriate legal ground for processing personal data is just one aspect. To avoid the violation of the data subject's rights, other GDPR requirements have to be followed. The controller[10] has to follow the principles relating to processing of personal data such as lawfulness, fairness and transparency, data minimisation, accuracy, integrity and confidentiality and accountability (GDPR Art. 5).

The data controller has to follow the principle of data protection by design and by default (see GDPR Art. 25, for further discussion see EDPB 2019).

3. Corpora of Disordered Speech: Example Use Cases

3.1 Use case 1: Clinical datasets requiring personal data from different sources

Collecting new datasets from individuals with diagnosed speech disorders is a multidisciplinary effort, and involves data from multiple sources. Prior to data collection, a solid inclusion and exclusion criteria with relevant diagnostic procedures is established. Finally, from the potential individuals only a subset will be willing to use their valuable time for taking part in scientific research, and often the number of participants remains small. Therefore, enabling the use of these datasets by other researchers beyond the data generators would not only benefit the scientific community but also help individuals with speech disorders. Shared data may have remarkable impact for reproducibility and confirming results, but in regard to clinical datasets it can only happen, if there are effective ways to share such sensitive data.

A one relevant example of such dataset discussed within the DELAD initiative is a study investigating the use of ultrasound visual feedback (UVF) as an intervention method for children and adults with persistent speech sound disorders. Accumulating evidence suggests that UVF has the potential to increase treatment efficacy, but most of the studies have been done in English speaking countries (e.g. Sudgen, Lloyd, Lam, & Cleland, 2019; Preston et al., 2017), and with relatively small number of participants. Therefore, there is a growing need to provide comparable data from other languages, and this is now in planning at the University of Helsinki, Finland.

Intervention study designs such as this involve collecting multiple types of data. Because the underlying cause of speech sound disorder may vary from phonological delay to apraxia of speech (Dodd et al. 2018, Waring and Knight 2013), careful assessment of the underlying linguistic-cognitive abilities is important. The hospital or clinic where the individual has been diagnosed will provide some of this information, and some will be collected as part of the study protocol. These data are usually stored as anonymised spreadsheet data. However,

[9] For the purpose of this article legacy data refers to personal data collected long time ago and there is not enough information for the identification and contacting data subjects. However, at the same time it is still protectable personal data.

[10] The GDPR defines the controller as "the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data" (Art. 4 (7)).

audio and video recordings of the treatment cannot be anonymised. For example, a ten-session treatment including 10–15 minute training with real-time ultrasound feedback of articulation needs to be recorded for reliable assessment of efficacy. Furthermore, word lists are recorded for accurate acoustic analysis for quantifying the change. Therefore, a secure system that would meet the high requirements of ethical guidelines for medical research is needed.

The technological requirements for such reliable data sharing systems are under development. For example, the Language Bank of Finland (Kielipankki) already serves the researchers and students who use text and speech corpora of typical speech. The Language Bank of Finland is hosted by CSC - IT Center for Science on servers in CSC data centers in Finland, and overcoming questions related to sharing sensitive data are under consideration. The dataset containing sensitive information will be licenced and stored in a secure environment without a direct connection to the internet. Limiting access to data with authorized access via a secured system will minimize the possibility for data misuse. The authorized user can access the data only on a virtual machine via Remote Desktop and would not be able to download the data[11]. However, even the best possible technological advances will become useful only after an agreement between the committees assessing the ethical guidelines of medical research and the participants. During ethical assessment of the study, special attention was drawn to how the experimental design and data sharing protocol is explained to the subjects. Importantly, it should be understandable also for children. The consent of the participants is asked with an opt-in form, where participants specify, if all of the data (audio, video, test results) or only some part of it can be shared for scientific purposes. The consent forms are stored in a secured place, separate from the data. In pilot measurements, participants and their parents were willing to opt-in data sharing, and the data collection will continue.

3.2 Use case 2: Archival recordings and metadata of hearing impaired children

An example of legacy corpora discussed within the DELAD initiative are the collections of the audio recordings of Polish hearing-impaired children collected in the years 1990s and later until 2006 in the region of Greater Poland (Wielkopolska), in Kalisz and Poznań (see the DELAD group progress report for 2019, and the presentation delivered by Anita Lorenc and Katarzyna Klessa). The data were first analysed and described in several academic dissertations (e.g., Andruszka et al., 2000; Francuzik & Szalkowska, 2001; Kleśta, 2002; Stankiewicz & Włoch, 2001; Trochymiuk, 2006). The results of further explorations based on the data were reported in a number of publications (e.g., Łobacz et al., 2002; 2003; Kleśta, 2004; 2006; Lorenc, 2012).

The existing collections consist of elicited speech (word lists) from above 60 children educated in two different schools in which two different methods of teaching were implemented at the time of the recordings. In Kalisz (cf. Trochymiuk, 2006), the so-called *Cued Speech* method was used according to which the speech is accompanied by special rhythmic movement of one hand enhancing the pronunciation (Cornett, 1967; Polish adaptation by Krakowiak, 1986). In the Poznań school (e.g. Kleśta, 2002), the starting point for speech practice was usually the phone-level articulatory training, while the approach to speech elicitation and pronunciation control was defined individually for each child. The majority of school subjects were taught by means of the *Total Communication* method making use of all accessible modalities: both vocal and visual channels, linguistic and non-linguistic signals, various artefacts and teaching aids (e.g., Holcomb, 1972, cited after: Evans, 1982).

The speakers whose voices are included in the collections were children during the recordings. At that time, the legislation in Poland was much more lenient than afterwards and thus no written consents were issued. The recording sessions were conducted based on the oral agreements between the researchers and the schools headmasters. Correspondingly, no direct contact information to reach the speakers is available at present which makes it practically impossible to develop appropriate consent documents *post factum*.

What makes the case of the disordered speech archives more problematic than other types of legacy data, is the sensitive nature of metadata collected along with the recordings. The additional information for the abovementioned corpora of the speech of hearing-impaired children involves details important from the point of view of speech therapists and phoneticians such as: standard medical facts about the hearing loss reasons and degree, the personal history of hearing aids, therapy and education stages, other possible disorders or diseases, family information, speech evaluation results, and more. The character of the metadata can become an obstacle for data sharing because part of the information might be regarded as especially sensitive which would obviously entail a need to implement varied data and metadata access levels.

Currently, one of the collections (Polish Cued Speech Corpus of 20 Hearing Impaired Children, cf. Trochymiuk, 2006) is being curated and will soon be made findable and accessible through CMU's Talkbank and stored at the TLA. For the initial version of the shared resource, the dataset will include: the original audio recordings, prompt texts for the recorded utterances, and the basic speaker information (gender, age). The access to more sensitive information, medical and family facts remains restricted.

As already mentioned, the novel requirements might be difficult or even impossible to be fully implemented for archival recordings. However, given that the data collection procedure was legal at the time of the recordings (even if conducted according to guidelines that were less stringent than the present ones), we assume that the resources (or part of them) can still be shared. See further section 4.

4. Discussion and Conclusion

The addressed cases concern the use of sensitive data (health data) which processing is more restricted. One option to avoid the applicability of personal data

[11] See: https://www.csc.fi/web/blog/post/-/blogs/csc-for-sensitive-data-because-your-data-is-worth-it-and-should-be-kept-that-wa-2

protection requirements is to anonymise the data. However, this is complicated for audio and video data.
Due to the delicate nature of the data it is crucially important to follow the principle of confidentiality and integrity. It means that the involved research institutions must ensure "appropriate security of the personal data, including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures" (GDPR Art. 5 f)). Data protection by design and by default must be followed as well (GDPR Art. 25).
Processing of the data can be based on consent or public interest research. The use of consent model entails legal uncertainty. The data subject can withdraw the consent at any time. Consent has to be freely given, specific and explicit.
Public interest research is another ground. The problem with this ground is that the collected data cannot be made widely available. There are limitations of commercial use as well.
For the first case, the consent model seems to be more appropriate. The group is small and the collection of data will start after appropriate ethical permission is obtained. Since there is a direct contact with the participants anyway, then it is also possible to acquire consents. Consents should include the description of all possible uses of personal data (e.g. commercial use, data sharing and so forth). The risk is that some participant may decide later that they do not want be part of the research (withdraws the consent). However, they cannot request the deletion of the collected data (see, GDPR Art. 17 3 (d)).
For the second case, the consent model is not suitable since the case concerns legacy data. It is unrealistic to obtain consents for processing the data since the contact information is not available; the data subjects are not fully identifiable and so forth. In this scenario the main way forward is to rely on the ground of research in public interest. The reliance on this ground requires informing data subjects about the data processing. However, this obligation is limited if it is not possible or it is disproportionally complicated to contact the data subjects (GDPR Art. 14, for further explanation, see also recital 62). Data sharing should always be preceded by careful inspection of the archive contents in order to secure GDPR-compliance. In the case of archival data curation and inclusion in public repositories, the GDPR data minimisation rule could be useful as one of the guidelines for the selection of publishable and shareable information and establishing the access right levels. Advice regarding specific cases can be sought from Data Protection Officers appointed in many European companies, institutions, and universities..
Apart from the legal grounds used for processing personal data, it is crucial to comply with other data protection requirements such as fairness and transparency, data minimisation, integrity and confidentiality and so forth (see GDPR Art. 5).

5. Bibliographical References

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). OJ L 119, 4.5.2016, p. 1-88. Available at http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1515793631105&uri=CELEX:32016R0679 (15.11.2019)

WP29. Guidelines on Consent under Regulation 2016/679. Adopted on 28 November 2017. As last Revised and Adopted on 10 April 2018. Available at https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=623051 (13.3.2019)

Andruszka, A., Dembińska, S., Goralewska, M., Pawełek, P., Piasecka, J., (2000). Fonetyczna charakterystyka wyrazów w wypowiedziach dzieci niesłyszących (Phonetic characteristics of words in the speech of hearing-impaired children), MA Thesis, Adam Mickiewicz University in Poznań.

Cornett, R. O. (1967). Cued speech. *American Annals of the Deaf*, 3-13.

Dodd, B., Ttofari-Eecen, K., Brommeyer, K., Ng, K., Reilly, S., & Morgan, A. (2018). Delayed and disordered development of articulation and phonology between four and seven years. *Child Language Teaching and Therapy*, 34(2), 87-99.

EDPB (European Data Protection Board). Guidelines 4/2019 on Article 25 Data Protection by Design and by Default. Adopted on 13 November 2019. Available at https://edpb.europa.eu/sites/edpb/files/consultation/edpb_guidelines_201904_dataprotection_by_design_and_by_default.pdf (27.2.2020)

Evans, L. (1982). Total communication: Structure and strategy. Gallaudet University Press.

Francuzik (Klessa), K., Szalkowska, E., (2001). Częstotliwości formantowe samogłosek w mowie dzieci niesłyszących (Formant requencies of the speech of hearing-impaired children), MA Thesis, Adam Mickiewicz University in Poznań.

Holcomb, R.K. Three years of the total approach 1968-71 (1972). *Report of the Proceedings of the Forty-Fifth Meeting of the Convention of American Instructors of the Deaf. Washington, D.C.*: U.S. Government Printing Office, 165-183.

Kelli, Aleksei; Lindén, Krister; Vider, Kadri; Kamocki, Pawel; Birštonas, Ramunas; Calamai, Silvia; Labropoulou, Penny; Gavrilidou, Maria; Pavel Straňák (2019). Processing personal data without the consent of the data subject for the development and use of language resources. In: Inguna Skadina, Maria Eskevich (Ed.). Selected papers from the CLARIN Annual Conference 2018 (72−82). CLARIN Annual Conference 2018, Pisa, 8-10 October 2018. Linköping University Electronic Press, Linköpings universitet. Available at http://www.ep.liu.se/ecp/article.asp?issue=159&article=008&volume= (15.11.2019).

Kleśta, J. (2002). Charakterystyka akustyczna cech segmentałnych mowy dzieci niesłyszących. PhD Thesis, Adam Mickiewicz University in Poznan.

Kleśta, J. (2004). Percepcyjna ocena zrozumiałości mowy realizowanej przez dzieci niesłyszące poddawane kształceniu w szkole specjalnej. *Investigationes Linguisticae, 10*, 28-41.

Kleśta, J. (2006). Analiza akustyczna polskich spółgłosek nosowych realizowanych przez dzieci niesłyszące. *Investigationes Linguisticae, 13*, 118-134.

Krakowiak, K. (1986). Fonogesty. Gesty wspomagające odczytywanie wypowiedzi z ust. Poradnik dla logopedów, nauczycieli i rodziców dzieci niesłyszących. *(Cued speech. Gestures supporting lip-reading. A handbook for therapists, teachers and parents of hearing-impaired children)* Lublin: IKN ODN.

Lindén, Krister; Kelli, Aleksei; Nousias, Alexandros (2019). To Ask or not to Ask: Informed Consent to Participate and Using Data in the Public Interest. Proceedings of CLARIN Annual Conference 2019: CLARIN Annual Conference, Leipzig, Germany, 30 September – 2 October 2019. Ed. K. Simov and M. Eskevich. CLARIN, 56−60. Available at https://office.clarin.eu/v/CE-2019-1512_CLARIN2019_ConferenceProceedings.pdf (15.11.2019).

Łobacz, P., Francuzik (Klessa), K., Szalkowska, E. (2002). Acoustic-phonetic description of Polish vowels: teen-age deaf speech, *Psychology of Language and Communication* Vol. 6 (2), 3-30.

Łobacz, P., Grygiel, W., Baranowska, E., Francuzik (Klessa), K., (2003). Klasyfikacja samogłosek polskich za pomocą sieci neuronowych w wymowie dzieci niesłyszących. (Classification of Polish Vowels Using Neural Networks), Audiofonologia (Vol. XXIII/2003), 7-31.

Lorenc A., 2012, Samogłoski w wymowie dzieci z uszkodzeniami narządu słuchu (Vowels in the speeach of hearing-impaired children), In Krakowiak K., Dziurda-Multan A. (Eds.), Wychowanie dzieci z uszkodzeniami słuchu – nowe wyzwania dla rodziców i specjalistów, Lublin: Wydawnictwo KUL, 77–113.

Preston, J. L., Leece, M. C., & Maas, E. (2017). Motor-based treatment with and without ultrasound feedback for residual speech sound errors. *International Journal of Language & Communication Disorders*, 52(1), 80-94.

Stankiewicz, K., Włoch, K., (2001). Cechy prozodyczne w mowie osób niesłyszących (Prosodic features of the speech of hearing impaired persons), MA Thesis, Adam Mickiewicz University in Poznań.

Sugden, E., Lloyd, S., Lam, J., & Cleland, J. (2019). Systematic review of ultrasound visual biofeedback in intervention for speech sound disorders. *International journal of language & communication disorders.*

Trochymiuk (Lorenc), A. (2006). Wymowa dzieci niesłyszących posługujących się fonogestami. PhD Thesis, Adam Mickiewicz University in Poznan.

Waring, R., & Knight, R. (2013). How should children with speech sound disorders be classified? A review and critical evaluation of current classification systems. *International Journal of Language & Communication Disorders*, 48(1), 25-40.

Van den Heuvel, H., Oostdijjk, N., Rowland, C., Trilsbeek, P. (2020). The CLARIN Knowledge Centre for Atypical Communication Expertise. In: Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC'20).

6. Language Resource References

Ultrasound visual feedback in treatment of persistent speech sound errors: deposit under the Language Bank of Finland, using CSC Finland's upcoming Remote Desktop for sensitive data.

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