What Can I Do with this Data Point? Towards Modeling Legal and Ethical Aspects of Linguistic Data Collection and (Re-)use

Annett Jorschick[†‡], **Paul T. Schrader**[§‡], **Hendrik Buschmeier**[†‡]
[†] Bielefeld University, Faculty of Linguistics and Literary Studies
[§] Bielefeld University, Faculty of Law
[‡] Bielefeld University, SFB 1646 'Linguistic Creativity in Communication'
Bielefeld, Germany
{annett.jorschick|paul.schrader|hbuschme}@uni-bielefeld.de

Abstract

Linguistic data often inherits characteristics that limit open science practices such as data publication, sharing, and reuse. Part of the problem is researchers' uncertainty about the legal requirements, which need to be considered at the beginning of study planning, when consent forms for participants, ethics applications, and data management plans need to be written. This paper presents a newly funded project that will develop a research data management infrastructure that will provide automated support to researchers in the planning, collection, storage, use, reuse, and sharing of data, taking into account ethical and legal aspects to encourage open science practices.
Keywords: linguistic data collection, language resources, open data, informed consent, legal tech

1. Introduction

A growing emphasis on transparency in research processes and improved quality assurance in scientific research has underscored the importance of 'open science' and the publication of accompanying research data. New standards for data sharing and reuse have been established, e.g., in the development of the FAIR data principles (Wilkinson et al., 2016). However, linguistic data often has inherent characteristics that makes sharing difficult from a legal and ethical perspectives.

Linguistic data encompass a wide range of data types characterized by their diversity. These data comprise different modalities (from written texts, to spoken audio, to multimodal video recordings), different settings (e.g., generic data collections, field studies, case studies, experimental setups), different topics of interest (e.g., examining and modeling of language families, production and recognition processes, language acquisition, diagnostics and therapy of speech disorders), and involve different groups of participants (e.g., 'standard' speakers and listeners, people with speech disorders, vulnerable speakers such as children). Crucially, linguistic data often contains personal information or comes from vulnerable speaker groups and thus requires careful handling in collection, storage, sharing, and reuse practices. Furthermore, anonymization of linguistic data is not always possible, as researchers may be particularly interested in aspects that are inherently personal (e.g., multimodal behaviors such as gestures, facial expressions, or gaze direction shown in video recordings of participants interacting in face-to-face dialogue). However, beyond the issues of transparency, quality assurance and reproduciblity, being able to share linguistic data would be very useful more generally. The reuse of linguistic data is desirable because it is a valuable resource. Collecting, preparing, and annotating linguistic data is time-consuming and labor-intensive given the meticulous collection methods. More importantly, from a scientific (and also cultural) perspective, linguistic data are valuable because they provide a unique record of linguistic diversity (across languages, speakers, geography and time), the existence and preservation of which is a prerequisite for various areas of linguistics.

Given the sensitivity of specific linguistic data, it is critical to address the legal aspects of data collection and use, and to ensure that data publication is covered by participant consent. Because legal, ethical, and privacy considerations are often intertwined, they should be taken seriously from the outset. Addressing these complexities requires careful planning of data collection, including the preparation of appropriate consent forms and technical and organizational data security measures. However, it is often challenging for researchers to anticipate all relevant considerations (especially also those for re-use), adequately address the regulatory framework, and effectively manage repetitive procedures. There is a need to explore the potential of automation to support researchers in meeting these challenges.

In this paper, we describe the infrastructure project INF ('User-oriented research infrastructure assisting linguistic data collection and (re-)use') of the newly funded Collaborative Research Center SFB 1646 'Linguistic Creativity in Communication' at Bielefeld University, Germany. The project will develop and implement a research data management infrastructure for the collection, storage, use, reuse,

LEGAL2024 @LREC-COLING-2024, pages 47–51

and sharing of different types of linguistic data, along with automated support for the ethical and legal considerations and challenges involved. Each data 'point' (which in this case could range from a single response to all the data generated by a participant in an experiment) is thoroughly characterized (via metadata) to enable researchers to quickly determine permissible operations such as analysis, automated processing, and sharing. To provide maximum support to researchers, all steps in the data life cycle and data organization are automated as much as possible. This includes automated generation of customized consent forms, open science guidelines, and methodological support checklists. By automating these aspects, researchers will be equipped with the tools and resources they need to navigate the complex landscape of data management and (re)use with confidence and ease. This not only facilitates compliance with ethical and legal standards, but also fosters a culture of transparency and reproducibility within the scientific community.

A number of projects have developed approaches to support the ethics application process and the creation of consent forms. Hannesschläger et al. (2020) and van den Heuvel et al. (2020) describe tools that facilitate the creation of GDPR-compliant (GDPR, 2016) consent forms (see sec. 2.1 for a detailed review of the relevant legal norms). 'Ethiktool' (Bendixen et al., 2023) is a computer program that supports researchers in creating applications to internal ethics review boards (IRBs) as well as participant information.

The diversity of linguistic data collection efforts, however, may present particular challenges, which complicate the use of standardized text modules for consent forms, ethics applications, or data management plans. Moreover, consent forms composed of automated components are often lengthy and legal texts are difficult to comprehend, particularly for participants who struggle with language. This is a fundamental problem, as participants need to be well-informed to provide informed consent. This complexity is further exacerbated when considering data reuse and sharing. The results of the replication crisis (Open Science Collaboration, 2015) demonstrate that sustainable research and open science practices are not merely optional methods but essential additional research goals. As explained above, adherence to the FAIR principles (Wilkinson et al., 2016) is particularly relevant for linguistic data.

The project aims to address these methodological, legal and technical issues through an interdisciplinary effort – comprising (psycho-)linguists (for the expertise in linguistic data collection and preparation), legal experts (for adherence to regulatory frameworks), and computational scientists (for modeling and building the technical infrastructure).

2. Concept

The project will consider the ethical and legal regulations (as will be outlined in sec. 2.1), as well as the technical framework (see sec. 2.2), to model the life cycle of linguistic data in a research data management infrastructure composed of three fundamental modules: (1) a data collection setup wizard guiding researchers to setup studies, which automatically generates customized consent forms, potentially guiding the creation of ethics applications, and data management plans; (2) a computational platform where all information about studies and their associated data are collected; (3) a search engine allowing data queries and sharing in a way that is consistent with the individualized consent (opt-in/out) provided by participants.

The data life cycle begins as early as the planning phase of the linguistic study. As outlined above, the creation of appropriate informed consent forms and the data management plan necessitate information about the use, sharing, and potential reuse of the data. To assist researchers in this initial stages of study planning, we will develop a 'Wizard' tool that guides users through the setup and design process, aiming to maximize the application of the FAIR principles (Wilkinson et al., 2016) wherever feasible. The main outcome of this wizard is a highly customized informed consent form tailored to the specific study and adapted to the requirements of it's participants, utilizing combinable text blocks, that are ethically and legally sound and coherent.

The underlying technical platform stores information collected from each study (including individual consent information), converts details into a standardized metadata format (Broeder et al., 2012), and subsequently integrates the information gathered during the study design phase with the collected data in order to model the permissible operations that can be done with a data point based on participants' consent.

To facilitate data reuse and collaboration, a search engine will be implemented to query and retrieve available data resources. It will enable researchers to easily identify and access relevant data 'points' for their specific research needs. Advanced search functionalities and metadata indexing will enable users to filter and refine search results based on various detailed criteria, such as data type, topic, accessibility status, and permissions for usage.

To achieve this goal, a careful analysis of the various legal norms (see sec. 2.1) will identify potential contradictions and redundancies between requirements that researchers may have. In addition, the objective requires a comprehensive review of legal doctrines, their interpretations, and relevant judicial precedents. Based on this, text blocks are created

that can be assembled into individualized consent forms. It is important to ensure the use of easy-to-understand language to enhance participant's comprehension, while avoiding excessive detail that may discourage participation due to length. In addition, as text blocks are combined within the consent form, it is critical to avoid redundancy, so the computational generation process must cross-check building blocks as they are uniquely assembled and resolve redundancies to maintain clarity and conciseness in the final document.

2.1. Legal and Ethical Regulations

The creation of a wizard tool to support researchers through the automated creation of data protection declarations and subsequent storage in a data management system requires a precise analysis of the legal and ethical framework conditions. Data collection in the field of empirical research is subject to various regulations.

Since the Helsinki Declaration in 1964 (World Medical Association, 2013) and the Belmont report in 1978 (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979), ethical considerations have become an increasingly important aspect of empirical research. In cooperation with the Ethics Committee of Bielefeld University, we will use the German Psychological Society's guidelines (DGPs, 2022) as the institutional basis for modeling consent forms. These guidelines are used as a framework for ethics applications in many universities in Germany. Moreover, the German higher education framework act (Hochschulrahmengesetz; HRG, 1999) encompasses provisions that specifically address research activities conducted within academic institutions, ensuring that research adheres to legal standards and ethical principles.

In addition to ethical considerations, ensuring compliance with various legal frameworks is essential for maintaining data security, proper data storage, facilitating data reuse, and regulating data sharing practices. Data protection regulations include, in particular, the European General Data Protection Regulation, which has been in force since 2018 (GDPR, 2016). As an European regulation, it is directly legally valid in the member states and also applies to public institutions. The GDPR is supplemented by the national BDSG (BDSG, 2017) and the DSG-NRW (DSG-NRW, 2018). The BDSG only applies if the law of the European Union, in particular the GDPR, does not apply directly (§ 1 V BDSG). A further restriction of the scope of application of the law can be found in § 1 I S. 1 Nr. 2 BDSG: The data protection law of the federal states (DSG-NRW) applies to the processing of personal data by public organizations of the federal states. However, the GDPR forms the overarching legal framework. As part of project INF, linguistic research data is to be collected, stored and processed. In certain cases, data processing without explicit consent may be lawful if it is necessary for the performance of a task carried out in the public interest (Art. 6 I lit. e) GDPR). This may apply in accordance with § 17 I DSG-NRW for data processing in the scientific field. Under certain circumstances, consent may not be required. The rights of access (Art. 15 GDPR), rectification (Art. 16 GDPR), restriction of processing (Art. 18 GDPR) and objection (Art. 21 GDPR) may also be restricted if the exercise of these rights is likely to render impossible or seriously impair the realization of the research or statistical purposes and the restriction of these rights is necessary for the fulfillment of these purposes (§ 17 V DSG-NRW). However, the consent of the study participants is generally required (Art. 6 I lit. a) GDPR). This consent is subject to certain legal requirements. Above all, it must be given freely, specifically and for a particular purpose (Art. 6 I lit. a) GDPR). Any use of the data outside of this purpose limitation is generally not permitted and may only take place with the renewed consent of the data subject.

In the scientific field, the purpose of the processing of personal data cannot usually be fully specified at the time of data collection. In order to take sufficient account of the special requirements of data collection in scientific research, Art. 5 I lit. b) GDPR provides for a relaxation of this rigid purpose limitation. However, data subjects still have the option of restricting their consent to certain research areas or parts of research projects (Recital 33 GDPR). Special requirements apply to consent in relation to data processing of particularly sensitive groups such as children (Art. 8 GDPR) or the processing of particularly sensitive data such as health data (Art. 9 GDPR). For example, when collecting health data, the consent to be given must contain an explicit reference to the processing of sensitive data. This information should enable the data subject to make an individual risk assessment. In general, it should be noted that consent can be withdrawn at any time and the possibility of withdrawal must be pointed out (Art. 7 III GDPR). With regard to the legally compliant use of the planned wizard, the legal requirements described here must be included in the module to be programmed. Particular attention is paid to the specific requirements for data collection in the field of scientific research.

2.2. Technical Framework

The establishment of a robust research data management infrastructure as outlined above involves several key stages. First, the integration of the framework into the infrastructure of Bielefeld University has to be considered, while ensuring compatibility with national and international data and

metadata standards and repositories. Second, security standards have to be implemented to ensure privacy and integrity of (meta)data. Finally, the source code of the infrastructure will be made openly available to encourage adoption and facilitate collaboration within the scientific community.

In order to ensure a integration across the university and compatibility between different systems, a careful consideration of existing services and interfaces is essential. This includes harmonizing research data management processes with existing platforms at Bielefeld University (such as RDMO, Gitlab.UB, PUB). These platforms support various aspects of research data management, from planning and documentation to version control and (data) publication. By integrating these interfaces, researchers will be able to streamline data management workflows and improve accessibility and reproducibility within the university.

In establishing effective research data management practices, it's important to consider both regional and global standards and initiatives. This includes adherence to metadata schemata for interoperability (Broeder et al., 2012) and engagement with platforms such as the Registry of Research Data Repositories. Additionally, infrastructures such as CLARIN, CLARIN-D and CMDI enhance accessibility and usability of linguistic resources. By aligning with these initiatives, the linguistic research data published in the platform/infrastructure will increase its visibility and impact on an international scale.

Security guidelines are important for software development because they ensure the integrity, confidentiality, and availability of data and systems. Standards such as ISO/IEC 27001 (ISO/IEC, 2022) provide a framework for establishing, implementing, maintaining, and continually improving an information security management system. This includes defining security policies, conducting risk assessments, implementing controls, and monitoring and auditing security measures. By following these standards and incorporating security best practices into software development processes, we can mitigate security risks and protect sensitive information from potential threats and vulnerabilities.

3. Embedding and Perspectives

The Collaborative Research Center SFB 1646 'Linguistic Creativity in Communication' comprises 16 research projects using different empirical methods to collect a diverse set of linguistic data. These methods include historical data analysis, corpus collection in various modalities (e.g., auditory, multimodal), and experimental investigations with diverse speaker groups. This rich research environment will allow project INF to comprehensively model linguistic data collection and usage practices, and to generate a wide range of use cases for the building block inventory for consent form generation. At the same time, the research center will be able to immediately benefit from (and influence) the creation of the infrastructure developed within INF that enables its open science and open data objectives.

In this first funding phase of the Collaborative Research Center, INF will focus on the implementation of the data management platform and its integration into the infrastructure of Bielefeld University. Data sharing with national and international infrastructures is planned for the second funding phase and will only be pursued after a thorough evaluation of the software security.

Acknowledgments

INF (537522983) is supported by the Deutsche Forschungsgemeinschaft (DFG) via the Collaborative Research Center SFB 1646 'Linguistic Creativity in Communication' (512393437).

Bibliography

BDSG. 2017. Gesetz zur Anpassung des Datenschutzrechts an die Verordnung (EU) 2016/679 und zur Umsetzung der Richtlinie (EU) 2016/680 (Datenschutz-Anpassungs- und -Umsetzungsgesetz EU – DSAnpUG-EU). *Bundesgesetzblatt Teil I*, 2017(44):2097–2132.

Alexandra Bendixen, Thomas G.G. Wegner, and Wolfgang Einhäuser. 2023. Facilitating ethics application and review for interdisciplinary human-participant research via software-based guidance and standardization. In *1st International Conference on Hybrid Societies*, Chemnitz, Germany.

Daan Broeder, Menzo Windhouwer, Dieter van Uytvanck, Twan Goosen, and Thorsten Trippel. 2012. CMDI: A component metadata infrastructure. In *Proceedings of the LREC 2012 Workshop on Describing Language Resources with Metadaa*, pages 1–4, Istanbul, Turkey.

DGPs. 2022. Berufsethische Richtlinien DGPs / BDP. Föderation Deutscher Psychologenvereinigungen.

DSG-NRW. 2018. Gesetz zur Anpassung des allgemeinen Datenschutzrechts an die Verordnung (EU) 2016/679 und zur Umsetzung der Richtlinie (EU) 2016/680 (Nordrhein-Westfälisches Datenschutz-Anpassungs- und Umsetzungsgesetz EU – NRWDSAnpUG-EU). *Gesetz- und Verordnungsblatt (GV. NRW.)*, 2018(12):243–268.

GDPR. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council. *Official Journal of the European Union*, L 119:1–88.

Vanessa Hannesschläger, Walter Scholger, and Koraljka Kuzman. 2020. The DARIAH ELDAH consent form wizard. In *DARIAH Annual Event 2020: Scholarly Primitives*.

HRG. 1999. Hochschulrahmengesetz. *Bundesgesetzblatt Teil I*, 1999(3):18–34.

ISO/IEC. 2022. ISO/IEC 27001:2022 – Information security, cybersecurity and privacy protection – Information security management systems – Requirements. Standard 27001:2022, International Organization for Standardization (ISO), Geneva, Switzerland.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. 1979. The Belmont Report: Ethical principles and guidelines for the protection of human subjects of research. *Federal Register*, 44(76):23192–23197.

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.

Henk van den Heuvel, Aleksei Kelli, Katarzyna Klessa, and Satu Salaasti. 2020. Corpora of disordered speech in the light of the GDPR: Two use cases from the DELAD initiative. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3317–3321, Marseille, France.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018.

World Medical Association. 2013. World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects. *JAMA*, 310:2191–2194.