UNIVERSITÄT BIELEFELD

LINGUISTIC CREATIVITY in Communication
SFB 1646

Project INF: User-oriented research infrastructure assisting linguistic data collection and (re-) use

Project leaders: Hendrik Buschmeier, Annett Jorschick, Paul T. Schrader

Project Summary

Managing linguistic research data is a complex and sensitive process because it often contains personal information, such as audio and video recordings used to study acoustic parameters, speech-accompanying gestures, facial expressions or data from vulnerable speaker groups such as children or people with speech disorders. In addition, there is also a tension between the goals of open science and privacy: Multimodal data, for example, requires extensive and costly preparation to make it usable for research. Because of this reuse and sharing of such carefully prepared corpora is an important goal. This, however, poses ethical, privacy and legal challenges as videos cannot be anonymised without losing important features. Linguistic data therefore require extremely careful management throughout the data lifecycle.

To support researchers in the data management process, the central aim of the INF project is to develop a research-focused, user-oriented data infrastructure that will facilitate the organisation, access and management of different types of empirical linguistic data collected within the CRC. The project data to be organised are: (i) from heterogeneous experimental setups and methods; (ii) from language corpora comprising written, spoken and multimodal data, as well as annotated data; (iii) in different languages; and (iv) involving different speaker groups, including vulnerable participants.

The researchers who are responsible for the data and the human participants who provide their data need to be confident that it is being collected, used and shared in compliance with the law, particularly GDPR (and other relevant aspects). Trust in the infrastructure to be developed in the INF project is based on two principles. Process-orientation: In the planning phase of data collection, researchers will be assisted by a 'wizard' tool to specify the 'components' (i.e., methods and workflow) of their study. This information will be used to automatically generate customised, ethically and legally compliant informed consent documents for participants and checklists for researchers. Technical measures/privacy-by-design: When the collected data is imported into the data management platform, it will be associated with the specific consent information provided by each participant. Based on the rules associated with the components chosen during study design, it will be possible for researchers to know how a data item can be used and who can access it (i.e., which data can be shared with whom and under what conditions). The infrastructure will consist of: (1) a conceptual framework for and a taxonomy of the 'components' (and their inter-dependencies) that covers data collection, use, and sharing reg-

ularities for the research in the CRC, and (2) a user-oriented technical data management platform that helps researchers plan data collections through a wizard that generates consent forms and checklists, imports data and consent information, technically enforces rules for use, reuse and sharing, and is interoperable with existing standards, platforms and repositories, i.e., it follows the 'FAIR' principles and established standards for metadata sharing and querying.

Open Positions

Post-doc position 1 (100%)
Profile: The ideal candidate has a PhD in psycho-linguistics, linguistics, psychology or a related field, with broad working knowledge of empirical methods and applied statistics, as well as data management.
Main research focus within the project: The postdoc will be responsible for developing – in close collaboration with CRC-projects – data collection and management practices for the CRC and creating a data management framework. An additional focus is on providing methodological and statistical support to projects.

PhD position 2 (100%)
Profile:
The ideal candidate has a Master's degree in computer science, computational linguistics, business informatics, data science or a related field. A research interest in open science, data management practices and knowledge of databases, ontologies, domain-specific languages and platform development is an advantage.
Main research focus within the project: The focus of the PhD project will be technical modelling of the legal framework and metadata formats, as well as designing, developing, and evaluating the data management platform. An additional focus is on providing methodological support with respect to data management and Open Science practices to the research

PhD position 3 (65%)
Profile: Sie haben ein rechtswissenschaftliches Hochschulstudium mit mindestens dem Ersten Juristischen Staatsexamen (mit mindestens 7,5 Punkten) abgeschlossen und erkennbares Interesse an datenschutzrechtlichen Fragestellungen im Forschungskontext und sind bereit, sich in ein interdisziplinäres Team einzubringen.

The ideal candidate has a university degree in German law (1. Staatsexamen with 7.5 points), a discernible interest in problems of data protection in the context of scientific research, and is ready to collaborate in an interdisciplinary team.
Main research focus within the project: The focus of the PhD project will be on providing legal support in the form of elaborating the legal framework, preparing consent forms, delivering text modules from practice, and undertaking legal modelling. He or she will combine practice and academic theory. For this

UNIVERSITÄT BIELEFELD

LINGUISTIC CREATIVITY in Communication
SFB 1646

purpose, it is important that exchange with practicing lawyers takes place; it is therefore intended that he or she will work alongside in a law firm specialised in this field.

For further information please contact the project leaders:

Prof. Dr. Hendrik Buschmeier (hbuschme@uni-bielefeld.de), Dr. Annett Jorschick (annett.jorschick@uni-bielefeld.de, Prof. Dr. Paul T. Schrader (paul.schrader@uni-bielefeld.de))