JLAR 2 (2024)
10.15460/jlar.2024.2.2.1575

A new series of corpus presentations

Konstantin Sering[1]*, David Bowie[2] & Annette Gerstenberg[3]

[1]University of Tübingen, Germany; Series Editor
[2]University of Alaska Anchorage, USA
[3]University of Potsdam, Germany

Abstract. The *Journal of Language and Aging Research* (JLAR) presents a new series of corpus presentations. The series is meant to feature research data suitable for analyzing the processes and manifestations of linguistic aging. Dedicated to the idea of open science, data are presented with regard to the principles of findability, availability, interoperability, and reusability (FAIR). Limitations of open access due to ethical concerns and possible other limitations are discussed. The series of corpus presentations allows for a standardized introduction of newly conceived or established datasets, promotes the exchange, on an empirical basis, of Language and Aging Research (LAR), stimulates the discussion of standards in Research Data Management (RDM), and motivates the further development and sharing of resources and techniques within the community. Faced with age-related bias in artificial intelligence and with copyright issues, settling the provenance of datasets representing older adults and guaranteeing their originality holds immense significance.

Keywords. corpus presentation, open science, research data management (RDM), FAIR principles, language and aging research (LAR)

Submitted: 2024-10-01
Accepted: 2024-10-11
Published: 2024-10-31
DOI: 10.15460/jlar.2024.2.2.1575

1 The new series of corpus presentations

With this new series of corpus presentations in the *Journal of Language and Aging Research* (JLAR) we aim to feature empirical data used in Language and Aging Research (LAR), inspire colleagues, and have resources made available as far as possible within the community. We also hope to promote standards in Research Data Management (RDM), e.g., the principles of findability, availability, interoperability, and reusability (FAIR, Wilkinson et al. 2016; GO FAIR International Support and Coordination Office (GFISCO) 2024). In this way, the corpus presentation series encourages professional

*. Corresponding author, konstantin.sering@uni-tuebingen.de

Please cite as Sering, Konstantin, David Bowie, and Annette Gerstenberg. 2024. "A new series of corpus presentations." *Journal of Language and Aging Research* 2(2): 1–5. 10.15460/JLAR.2024.2.2.1575.

JLAR 2 (2024)
10.15460/jlar.2024.2.2.1575

research data management, helps to plan new projects with regard to reuse and interoperability, and offers a reference paper for corpora, with state of the art descriptors and a stable point of reference. For corpora that are not currently publicly available it allows for a dated and visible mention in the scholarly record.

There is value in this for the LAR community in bringing together an overview of a variety of resources and of solutions for ethical issues, and making existing data (more) findable and available for researchers. Also, having a record of a corpus that already exists for a particular community will be of great value for researchers who want to do work dealing with the same or a similar community.

Furthermore, the series of corpus presentations aims to make challenges and limitations to open data within the LAR community transparent and should facilitate best practices on anonymization and pseudonymization, as well as providing a reference on what kind of data can and cannot be legally accessible to the research community and the general public. Especially because the LAR community regularly works with vulnerable groups of people and is in its very nature interested in the idiosyncrasies of language use, it faces the challenge of properly protecting recordings of those who provide the data we work with.

A last goal is for the LAR community to develop standards on how to implement proper research management standards. This includes, e.g., providing a "how to cite" standard for our corpora including a persistent Digital Object Identifier (DOI), approaching the question of which metadata is most relevant, and making progress in data driven research. A data format that focuses on the presentation of corpus data helps to track publications related to it. The corpus presentations follow a structure to make sure that all the papers cover the necessary elements as summarized in "Author instructions: Corpus or Collection Presentation" (Sering, Bowie, and Gerstenberg 2024, as an addition to JLAR Editors 2023).

2 Questions and answers

2.1 Why not a general data paper journal?

There are two main reasons why corpora from the LAR community find a better home in JLAR corpus presentations than an open access data paper journal. The first is that a presentation within JLAR allows us to build up a collection of descriptions of data resources that are used by and useful to the LAR community in particular. The other, more practical reason is that there are different types of ethical and privacy related challenges that the LAR community faces compared to communities with a more general target group.

If parts of the corpus are comprised of audio and video recordings it is inherently difficult to properly anonymize the data and, therefore, it might not be possible to put the raw corpus data under a free and open access license. With the new corpus presentation series we try to address this problem by including presentations of corpora that are only partly open access. The format is an invitation for the community to move in the direction of making materials like code books and transcription guidelines open access alongside the corpus itself. This should help the LAR community to establish good practices of data collection, annotation, and enhancement.

JLAR 2 (2024)
10.15460/jlar.2024.2.2.1575

2.2 Is JLAR a repository now?

The purpose of a corpus presentation is to document, highlight, and present an existing corpus. The corpus preparation can be still in progress but substantial amounts of the data should be already collected. Additionally, the corpus presentation should showcase and document the overall structure of the corpus and the process of data collection as well as provide an example analysis showing how the corpus data can be used. This type of publication therefore becomes an artifact that bundles different resources of and about a corpus, which should be stored — as far as it is possible — in open repositories with an open access license.

An important complement to the corpus data is the documentation such as informed consent forms, questionnaires, stimuli, guidelines for interviewers or transcribers, scripts, and codebooks. They also deserve to be consistently referenced and made publicly available, on the project homepage or with public repositories such as Zenodo (European Organization for Nuclear Research and OpenAIRE 2013) or the Open Science Framework (OSF, Center for Open Science 2011–2024). (Alternatively, many universities, often through university libraries, provide the opportunity for publishing research data.) Research data repositories usually allow for "versioning", with a first version 1.0 and following updates listed as new versions with corresponding dates.

In conclusion, JLAR does not want to be and will not be a repository, but rather it aims to collect human readable metadata in terms of presentations of corpora. The corpus itself and additional materials should, as much as possible, be published in available repositories.

2.3 What are the submission requirements?

Corpus presentations should at least partly suitable for and clearly relate to LAR. Therefore, it should either focus on an older age group or include longitudinal data that can be analyzed with regard to language change and variation due to aging.

We are aware of the sensitive nature of personal data, and that some parts of the (raw) corpus data may not be allowed to be shared, or to be shared only if very specific research questions are declared, and / or a commitment to refrain from deanonymization is signed. As a result, corpora presented in the new series are not necessarily restricted to open access resources only, and can present a precise explanation of restricted access resulting from, e.g., ethical considerations such as those detailed by The British Association for Applied Linguistics (2021). Further, descriptions of these limitations — together with some basic consideration of the reasons for such practices — should appear as part of the corpus presentation.

The new series is thus open for corpus data that is not yet publicly available, and even for data that will never be. Nevertheless, JLAR is committed to Open Science and therefore as much as possible of the metadata and data derived from the corpus should be made publicly available in public repositories. These public repository items then can be referenced in the corpus presentation.

2.4 Where are the author guidelines, and how are these submitted?

Besides the general author guidelines for publications (JLAR Editors 2023), there are additional specific guidelines for corpus presentations (Sering, Bowie, and Ger-

JLAR 2 (2024)
10.15460/jlar.2024.2.2.1575

stenberg 2024), which list in a comprehensive way all the aspects that should be addressed in one of them.

The submission of a corpus presentation is done in the same way as for other publications, that is, through the JLAR website. Submissions of corpus presentations, however, must be pre-approved by the journal or section editors, and will be reviewed by the editors before publication. Corpus presentations will normally be no longer than 3,500 words.

3 Key features

Corpus presentations in this new series open the opportunity to publish a reference paper for LAR data sets and corpora detailing the specifics from the "machine room."

- With its related instructions for authors (Sering, Bowie, and Gerstenberg 2024), the series is meant to set standards for the creation, organization, and distribution of research data used in LAR.
- The contribution to a balanced representation of data from older individuals holds immense significance (Pichler, Wagner, and Hesson 2018).
- The reuse not only of research data, but also of techniques and scripts for developing, using, and disseminating resources becomes possible in a transparent manner with a citable stable reference acknowldging the authors' work.
- Providing a space for such a focus opens the floor for the discussion of standards in RDM related to LAR and best practice models.
- The need for suitable repositories is highlighted, and institutional exchange on issues such as storage and sustainable RDM may be enhanced.
- Apart from specialized venues (see, e.g., Tomaschek et al. 2021), this focus on technical issues, tools used, and legal issues can usually be set forth only in a very limited form in other types of presentation.[2]
- The significant effort behind the creation of resources is acknowledged and made visible. This advances the recognition of the underlying research output as academic achievement in academic reviews.
- Artificial intelligence tools in the humanities are faced with a number of challenges such as age-related bias (Chu et al. 2023). Settling the provenance of research data and guaranteeing its originality has become increasingly important.

In the growing field of LAR, the empirical basis for work is impressive and multifaceted (Bowie and Gerstenberg 2023). The new series of corpus presentations aims at contributing to its establishment and to motivate the ambition to further develop — and share — resources and techniques.

2. The authors wish to thank Simon Oppermann, who pointed out this desideratum at the 6th Conference for Language and Aging Research, held at the University of Tübingen April 10–12, 2024.

References

Bowie, David, and Annette Gerstenberg. 2023. "Language and aging research: Contradictions and aspirations." *Journal of Language and Aging Research* 1 (1): 1–6. https://doi.org/10.15460/jlar.2023.1.1.1243.

Center for Open Science. 2011–2024. *Open Science Framework (OSF).* Charlottesville: Center for Open Science. https://osf.io/.

Chu, Charlene H., Simon Donato-Woodger, Shehroz S. Khan, Rune Nyrup, Kathleen Leslie, Alexandra Lyn, Tianyu Shi, Andria Bianchi, Samira Abbasgholizadeh Rahimi, and Amanda Grenier. 2023. "Age-related bias and artificial intelligence: A scoping review." *Humanities and Social Sciences Communications* 10 (510): 1–17. https://doi.org/10.1057/s41599-023-01999-y.

European Organization for Nuclear Research and OpenAIRE. 2013. *Zenodo.* Geneva: CERN Data Centre / Invenio / CERN. https://doi.org/10.25495/7gxk-rd71.

GO FAIR International Support and Coordination Office (GFISCO). 2024. *FAIR principles.* Leiden, Paris, and Hamburg: GFISCO. https://www.go-fair.org/fair-principles/.

JLAR Editors. 2023. *Journal of Language and Aging Research: Information for authors.* V. 2.0, November 16, 2023. Zenodo / SUB Hamburg. https://doi.org/10.5281/zenodo.14021406.

Pichler, Heike, Suzanne Evans Wagner, and Ashley Hesson. 2018. "Old-age language variation and change: Confronting variationist ageism." *Language and Linguistics Compass* 12 (6): e12281. https://doi.org/10.1111/lnc3.12281.

Sering, Konstantin, David Bowie, and Annette Gerstenberg. 2024. *Author instructions: Corpus or collection presentation.* Zenodo, November. https://doi.org/10.5281/zenodo.14034372. https://doi.org/10.5281/zenodo.14034372.

The British Association for Applied Linguistics. 2021. *Recommendations on Good Practice in Applied Linguistics, 4th Edition.* BAAL. https://www.baal.org.uk/wp-content/uploads/2021/03/BAAL-Good-Practice-Guidelines-2021.pdf.

Tomaschek, Fabian, Denis Arnold, Konstantin Sering, and Friedolin Strauss. 2021. "A corpus of Schlieren photography of speech production: Potential methodology to study aerodynamics of labial, nasal and vocalic processes." *Language Resources and Evaluation* 55 (4): 1127–1140. https://doi.org/10.1007/s10579-021-09550-8.

Wilkinson, Mark D., Michel Dumontier, I. Jsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific Data* 3:1–9. https://doi.org/10.1038/sdata.2016.18.