

README

Semi-structured interviews with linguists about Open Science

Elen Le Foll

2025-11-17

1 General information

1. Title of Dataset: Semi-structured interviews with linguists about Open Science
2. DOI: [10.5281/zenodo.17588655](https://doi.org/10.5281/zenodo.17588655)
3. Contact Information:
 - Name: Elen Le Foll
 - Institution: Department of Romance Studies, University of Cologne
 - Email: elefoll@uni-koeln.de
 - ORCID: <https://orcid.org/0000-0002-5839-8010>
4. Date of data collection: See metadata field Date of Collection.
5. Geographic location: See metadata section Geographic Coverage.
6. Description of dataset: This dataset contains the anonymised transcripts of 26 semi-structured interviews conducted with linguistics students and researchers between late February and early April 2025. The interviews were conducted by Elen Le Foll via Zoom and last around half-an-hour (range: 13–52 min). All were conducted in English except one (A01 in German). They were first transcribed using a local installation of Whisper (Radford et al. 2022). The automatic transcriptions were subsequently manually checked, corrected, and anonymised by two student research assistants, Vishar Kavehamoli and Julia Weinberger, and Elen Le Foll. The coding was performed by Elen Le Foll using OpenQDA (Belli et al. 2025). The repository features a CSV file that contains all the coded passages and their corresponding categories, as exported from OpenQDA. Note that passages may be listed several times, if they were assigned several codes.

2 Sharing/access information

(See metadata record for dataset.)

1. Licenses/restrictions: CC-BY-SA 4.0
2. Links to publications that cite or use the data: See metadata field Related Works.
3. Links/relationships to related data sets: See metadata field Related Datasets.
4. Data sources: See metadata field Data Sources.
5. Recommended citation: See citation generated by repository.

3 Data and file overview

The dataset consists of two folders containing the following files:

- Documentation:
 - `00_README.md` (markdown file)
This file (= the current document) contains the documentation of the dataset.
 - `00_Interview_Structure.odt` (OpenDocument Text)
This file contains the pre-formulated questions that guided the interviews in both English and German.
 - `00_Consent_Form.odt` (OpenDocument Text)
The form of consent that all participants (digitally) signed and returned as a PDF.
- Data:
 - Each anonymised interview transcript is contained in a single plain-text file, encoded in Unicode UTF-8. Both the questions and answers are transcribed with one line per speaker turn. There are 26 files in total. They are numbered chronologically, `A01.txt` was the first interview conducted.
 - `OS_Interviews_Coded_Passages_OpenQDA_12_11_2025.csv` (CSV)
This file contains all the coded passages and their corresponding categories, as exported from OpenQDA. Note that passages may be listed several times, if they were assigned several codes.

4 Data collection

A first call to participate in an interview about Open Science in linguistics was sent via the mailing list of [ReproducibiliTea in the HumaniTeas](#) on 24 February 2025. However, only four linguists volunteered as a result of this call. I subsequently sent personalised e-mails to a further 27 linguists of different career statutes and from a broad range of subdisciplines with the aim of

interviewing some 20 linguists in total. The positive response rate to these personalised e-mails was higher than expected and quickly led to the recruitment of 22 additional interviewees.

I drafted interview questions in both English and German. However, the interviews were only semi-structured; the exact wording, order and number of questions varied depending on the interviewees’ responses. I conducted all 26 interviews online over Zoom from February to April 2025. The interviewees consented to the interviews being recorded and to the anonymised transcription being published on an open repository. Most interviews lasted around half-an-hour (range: 13–52 min). All were conducted in English except one (A01 in German).

5 Data processing

The interviews were first transcribed with the help of a locally-run instance of the speech-to-text large language model Whisper (Radford et al. 2022). These automatic transcriptions were subsequently checked, corrected, and anonymised by two student research assistants and myself. No attempt was made to transcribe pauses, false starts, or paralinguistic sounds. Moreover, no attempt was made to systematise the use of punctuation. The punctuation in these transcripts merely serves to improve their readability.

To protect the identity of the interviewees, mentions of all concrete projects were anonymised as PROJECT. Other aspects that were anonymised include mentions of specific institutions (INSTITUTION), cities (CITY), countries/regions/languages (COUNTRY), and colleagues and supervisors (PERSON). To protect the anonymity of the interviewees, personal metadata is not linked to any transcript. The following section only provides summary statistics.

6 Metadata

6.1 Gender

Gender	n
F	15
M	11

6.2 Country

Country	n
Belgium	1
Germany	21

Country	n
Norway	1
Sweden	1
UK	2

6.3 Affiliation

Of the interviewees with a German affiliation, eight are affiliated with the University of Cologne. Nine other German institutions are represented. Three of the interviewees are currently affiliated with European institutions but come from or have spent many years teaching in countries of the so-called Global South.

6.4 Academic status

Position	n
Associate professor	5
B.A. student	1
Doctoral researcher	3
Full professor	3
Lecturer	2
Librarian	1
M.A. student	2
No longer in academia	2
Post-doc	7

The two interviewees who are no longer in academia had completed a PhD in linguistics very recently and had decided to pursue a job outside of academia.

6.5 Tenure status

Tenured	n
No	11
No longer in academia	2
Student	3
TT	1
Yes	9

TT = tenure-track

6.6 Subdisciplines

At the beginning of the interviews, the interviewees were asked in which subdiscipline(s) of linguistics they situate themselves and their research. Almost all mentioned several subdisciplines, some of which were anonymised either because they were highly specialised or because the combination of the subdisciplines mentioned made it potentially possible to identify the interviewees. Across all 26 interviewees, the following subdisciplines were mentioned (ordered by frequency of mentions): corpus linguistics (11), phonetics (7), applied linguistics (5), sociolinguistics (5), discourse analysis (4), language teaching (4), phonology (4), second language acquisition (4), language learning (3), psycholinguistics (3), theoretical linguistics (3), English linguistics (2), World Englishes (2), cognitive linguistics (2), computational linguistics (2), language documentation (2), pragmatics (2), teacher training (2), typology (2), variational linguistics (2), German linguistics (1), Romance linguistics (1), Slavic linguistics (1), classics (1), cognitive semantics (1), contrastive linguistics (1), corpus phonetics (1), dialectology (1), experimental linguistics (1), generative linguistics (1), heritage language research (1), language testing (1), learner corpus research (1), lexicogrammar (1), linguistic landscaping (1), morphology (1), multilingualism (1), neurolinguistics (1), semantics (1), sociophonetics (1), speech science (1), syntax (1), translation studies (1), variationist linguistics (1).

6.7 Anonymised mentions of individuals

Many interviewees mentioned specific individuals who acted as role models or who otherwise inspired them to find out more about OS. All of these mentions have been anonymised as PERSON in the transcripts because some interviewees also mentioned them being colleagues or otherwise specified their relationship in such a way that may have led to their identification.

The following individuals were specifically commended (some by multiple interviewees), for one or more specific reasons. They are listed here in alphabetical order.

Name	Reason(s) for mention(s)
Benedikt Szmrecsanyi	Pre-registration, sharing data
Bodo Winter	Sharing data and code, promotion of OS in linguistics
Brian McWinney	CHILDES
Clare Patterson	Pre-registration, registered report
Elen Le Foll	OERs, ReproducibiliTea in the HumaniTeas, OS workshops and talks
Emma Marsden	Promotion of OS in applied linguistics
Jan Blommert	Tilburg Working Papers in Linguistics

Name	Reason(s) for mention(s)
Jordan Zlatev	Phenomenological triangulation
Lukas Sönning	Promotion of OS in linguistics, OS workshops and talks
Luke Plonsky	Promotion of OS in applied linguistics
Martin Hilpert	YouTube videos of online teaching materials
Martin Schweinberger	OER, online talks
Martine Grice	Promotion of diamond open access publishing in linguistics
Meng Liu	Promotion of OS in applied linguistics
Monica Bednarek	Sharing of annotation scheme
Naomi Truan	Talk “Open Science and me”
Petra Schumacher	Pre-registration, open data and code, mentoring of PhD students on OS matters
Sigrid Norris	Researcher stance/objectivity
Silbe Andringa	Promotion of diamond open access publishing in linguistics
Stefan Müller	Language Science Press
Timo Röttger	Pre-registration paper, registered reports, sharing data and code, supervision of metascientific theses
University of Lancaster	MOOC on corpus linguistics

References

- Belli, Alessandro, Jan Küster, Florian Hohmann, Philip Sinner, Gino Krüger, Karsten Wolf, and Andreas Hepp. 2025. *OpenQDA*. Zenodo. <https://doi.org/10.5281/zenodo.15024779>.
- Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. “Robust Speech Recognition via Large-Scale Weak Supervision.” <https://doi.org/10.48550/ARXIV.2212.04356>.