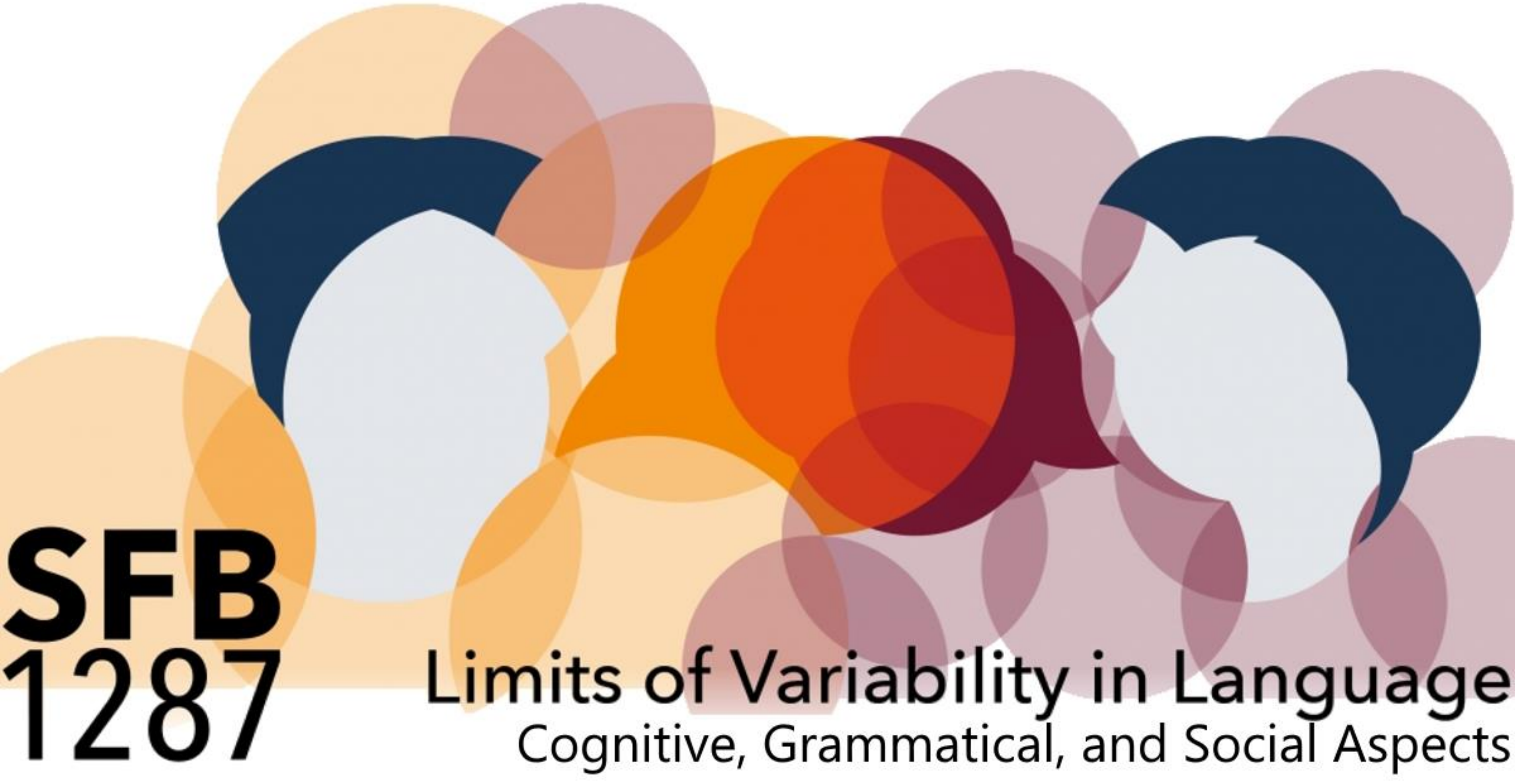
# DATA MANAGEMENT PLAN

**SFB 1287 / 2021 / Phase1 / Project C02**


SFB
1287
Limits of Variability in Language
Cognitive, Grammatical, and Social Aspects

# TABLE OF CONTENTS

# GENERAL INFORMATION

## (1) Are there requirements regarding the data management from your scholarly/scientific community?

- yes

### (a)If yes, what are the requirements?

- DFG Guidelines on the Handling of Research Data

## (2) What kind of dataset is it?

*The following questions collect information on the data that is produced or used in the project. They also help to estimate the value of the data in terms of potential re-use and long-term preservation. In the case of personal data, the principle of data minimization (Art. 5 EU General Data Protection Regulation) allows the collection of personal data only when there are no other reasonable means to clarify the research question (re-use of existing data would be such a reasonable means). Also, there shall be no more information collected than necessary. The information regarding the data collected, produced, or used in the project is gathered along the datasets. The definition of these datasets is an important conceptual decision that has to be done individually and carefully for each project.*

| | |
|---|---|
| **Name of Experiment / Acronym / Number:** | WP2 "definite pseudocleft/cleft" experiments (Akan, German, French; in part funded by XPrag.de SPP 1727) and 'homogeneity' experiments (English; in part funded by XPrag.de SPP 1727) |
| **PI or responsible person (head of the study):** | Prof. Dr. Malte Zimmermann, Prof. Dr. Edgar Onea |
| **Other persons involved:** | Joseph P. De Veaugh-Geiss, Swantje Tönnis, Emilie Destruel, Agata Renans |
| **Subject area:** | Linguistics > Semantics > definiteness > cross-linguistic comparisons, in particular with understduied language Akan; homogeneity |
| **Method / Type of data:** | - Behavioral response and reaction time data<br>- CSV/ODS files for stimuli<br>- python scripts for experimental software<br>- audio files for experiment |
| **Anonymizable data:** | yes |
| **Participants:** | University students and staff at the following universities: Universität Potsdam/Göttingen, Germany; University of Ghana at Legon, Ghana; University of Albi or University of Pau, France; Prolific (crowd-sourced) |
| **Short description (of the study):** | Nouns in Akan exhibit a three-way definite system: definites, specific indefinites, and bare nouns with either unique or indefinite interpretation; in contrast, German, English, and French have a two way system divided between definiteness and indefinites. Moreover, for all languages parallelism between definite pseudoclefts and clefts has been proposed. |

| | |
|---|---|
| **Short description (of the study):** | The studies explore definiteness in the four languages from different perspectives, specifically, the homogeneity approach to and uniqueness inferences in definite pseudoclefts/clefts. |
| **Other comments:** | Studies conducted in part with funding from “Exhaustivity in it-Clefts” project in the XPrag.de priority program (SPP 1727) |
| **Time for data collection (approximate):** | 4-6 weeks per experiment |
| **Time for data analysis (approximate):** | 3-6 weeks |
| **Related publications:** | De Veaugh-Geiss, J oseph P. 2020. nà-Cleft (non-)exhaustivity: Variability in Akan. Mansucript to be resubmitted to Glossa.<br>Destruel, Emilie & Joseph P. De Veaugh-Geiss. 2019. (Non-) Exhaustivity in French c’est-Clefts. In. Empirical Issues in Syntax and Semantics 12. Christopher Pinon (ed.). Paris: CSSP. 91–120. http://www.cssp.cnrs.fr/eiss12<br>Renans, Agata & Joseph P. De Veaugh-Geiss. 2019. Experimental Studies on it-Clefts and Predicate Interpretation. Semantics & Pragmatics 12(11). 1–50. Retrieved from https://semprag.org/index.php/sp/article/view/sp.12.11 |
| **Keywords (used in publications):** | Akan; nà-clefts; definite pseudoclefts; exhaustivity; experimental studies; c’est-clefts; es-clefts; German; French; it-clefts; exhaustivity; homogeneity; distributive, collective, and mixed predicates; distributive vs. non-distributive interpretation |

| | |
|---|---|
| **Link to preregistration:** | |
| **Funding reference:** | Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)<br>– Project number 317633480<br>– SFB 1287<br><br>Project in part, funded also by “Exhaustivity in it-Clefts” project in the XPrag.de priority program (SPP 1727) |

## (3) Which individuals, groups or institutions could be interested in re-using this dataset? What consequences does the reuse potential have for the provision of the data later?

*It is important to specify whether the data will be permitted for reuse. But legal impediments, such as privacy, and copyright must be taken into account.*

- ❖ Raw data are not anonymous; reuse of raw data by other researchers/scientists is NOT possible because the identity of the participants could be revealed (personal data).
- ❖ However, after data preparation the (pesudo-) anonymized data can be released to others.

# TECHNICAL INFORMATION

## (4) Where is the dataset stored during the project?

*Please delete all project-files from source that is not part of the University of Potsdam.*

| | |
|---|---|
| **for raw data:** | Box.UP - Cloud (University of Potsdam)<br>Researcher´s Computer |
| **for analysis data:** | SFB1287 - File Server<br>Box.UP - Cloud (University of Potsdam)<br>Git.UP - Cloud (University of Potsdam)<br>Researcher´s Computer |
| **for further documentation, related code, or software:** | SFB1287 - File Server<br>Box.UP - Cloud (University of Potsdam)<br>Git.UP - Cloud (University of Potsdam)<br>Researcher´s Computer |

### (a)If data is stored on lab or personal computers, please describe the backup strategy.

❖ syncronization with BoxUP / Regular full-system backups using backup program (deja-dup).

# (5) Which file formats are used?

*When choosing a data format, one should consider the consequences for collaborative use, long-term preservation as well as reuse. It is advisable to use formats that are standardised, open, non-proprietary, and well-established in the respective scholarly community. A table with recommended file formats can be found in found in Kristin Briney, Data Management for Researchers, Pelargic, 2015, pages 133-134.*

- Responses / Reaction time data (raw and processed): *.csv-/*.tsv-files
- Presentation: *.csv-/*.ods-files for stimuli; * .py-scripts and *.html files for experiment (Python/Onexp); *.wav-files for auditory stimuli; *.jpg- files for images
- Analysis scripts: *.Rmd-code

# PUBLICATION

## (6) Will this dataset be published or shared?

*anonymizable data*

- *ID will be removed (and or code-list will be destroyed) [legally correct: code list will be destroyed as soon as possible without jeopardizing experiment; exception: follow-up study planned, if so, talk to UP data protection officer (Dr. David Kneis; mailto:datenschutz@uni-potsdam.de) on how to do this correctly]*
- *Publication of anonymized data and code on OSF or RADAR (University of Potsdam) (or as required by the Journal)*

*non-anonymizable data*

- *on RADAR (University of Potsdam) but not accessible from the outside world*

❖ yes

### (a)If yes, under which terms of use or license will the dataset be published or shared?

*The options refer to the licenses of the Creative Commons family. If data is anonymised / pseudonymized, it's probably not legally required, but might be good in terms of research ethics to adjust consent forms / subject information sheets.*

✓ **Principal investigator of the study assures that the consent form / subject information sheets support publishing of the data.**

**for data:** CC 0 (Public Domain) (recommended)

| | |
|---|---|
| **for scripts:** | CC 0 (Public Domain) |
| **for software:** | none (shared only internally) |

## (b)If yes, when will the data be published?

*Recommended procedure: Upload data and obtain digital identifier (e.g., DOI, OSF link) when submitting the first paper; thus, you can cite the data in the paper. If necessary, restrict public access (embargo) until last paper published (max. 2 years).*

- ❖ when the paper is first submitted

## (c)If no, please explain why not. Please differentiate between legal and contractual reasons and voluntary restrictions.

# (7) Which measures of quality assurance are taken for this dataset?

# LEGAL AND ETHICS

## (8) Does this dataset contain personal data?

*The EU General Data Protection Regulation (GDPR) defines in Art. 4 personal data as "any information relating to an identified or identifiable natural person". An identifiable natural person is "one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person".*

*For advice and/or trainings on how to comply with privacy regulations - including proper anonymization & pseudonymization - you can always contact UP's privacy officer, Dr David Kneis, at datenschutz@uni-potsdam.de. As the privacy officer, his perspective is focussed more on the legal side of things than the research ethics or technical aspects.*

❖ yes

## (9) Are the data anonymised?

*Anonymised data: ID will be removed (and or code-list will be destroyed) [legally correct: code list will be destroyed as soon as possible without jeopardizing experiment; exception: follow-up study planned, if so, definitely talk to UP data protection officer on how to do this correctly]*

| | |
|---|---|
| **for raw data:** | no |
| **for analysis data:** | yes |
| **for published data:** | yes |

# (10)Does the project use and/or produce data that is protected by intellectual or industrial property rights?

*Measurement data has no intellectual property, so usually, the answer here will be „no". Data or software can be subject to intellectual or industrial property rights. Applicable laws differ broadly even within EU. According to the German copyright law (UrhG) works of literature, scholarship and the arts that can be regarded as a "personal intellectual creation" are protected by copyright. Mere data, e.g., measured data or survey data, and metadata (except in some cases descriptive metadata) are not protected by copyright. § 2 of the UrhG lists the following kinds of protected works (list is not concluded):*

- *linguistic works such as written works, speeches, and computer programs*
- *works or the fine arts including works of the applied arts as well as sketches of such works*
- *works of photography*
- *descriptions and illustrations of scholarly or technical nature such as drawings, plans, maps, sketches, tables, and three-dimensional representations*

*According to § 3, copyright is also applicable to translations and other modifications or adaptions of work if they are individual intellectual creations of the editor. Finally, according to § 4 copyright also extents to collected editions and database works. Collected editions are: "collections of work, data or other independent elements that are individual intellectual creations based on the selection and arrangement of the elements".*

*Database works are defined as "collected editions, the elements of which are arranged in a systematic or methodical way and can be accessed individually by electronic means or in other ways".*

❖ yes

## (a)If yes, please explain which!

- Images used in Akan experiment reported in De Veaugh-Geiss (2020) according to the terms of the Creative Commons BY-SA license: https: / /creativecommons.org/ licenses/by-sa/2.0/. The only modifications made to them were (i) cropping the images and (ii) converting them to black-and-white. The four photos were taken by Mark Fischer and are available for download at the following

URLs:

- https://www.flickr.com/photos/fischerfotos/16360022790
- https://www.flickr.com/photos/fischerfotos/22986270194
- https://www.flickr.com/photos/fischerfotos/23519903990
- https://www.flickr.com/photos/fischerfotos/23484547082

# STORAGE AND LONG-TERM PRESERVATION

## (11)Does this dataset have to be preserved for long-term?

*The DFG expects primary data that is the basis of a publication to be stored in the researcher's own institution or an appropriate nationwide infrastructure long-term (for at least 10 years).*

- ❖ At least 10 years after the end of the first funding period of the SFB1287.

## (12)What are the reasons this dataset has to be preserved for the long-term?

- ❖ Used in a publication / proof of good scientific practice
- ❖ Re-use (if anonymizable data) in subsequent projects or by others
- ❖ Legal obligations
- ❖ Documentation, because it is relevant to society
- ❖ Self-commitment
- ❖ Proof of good scientific practice
- ❖ By DFG requirements

## (13)Where will the data (including metadata, documentation, and relevant code) be stored or archived after the end of the project?

- ❖ Zenodo, S&P servers with publications, Harvard Dataverse, GitUP
- ❖ SFB1287 - File Server

# REFERENCES

SFB 1287 Limits of Variability in Language
Cognitive, Grammatical, and Social Aspects

**SFB 1287**
https://www.sfb1287.uni-potsdam.de

Universität Potsdam

**University of Potsdam**
https://www.uni-potsdam.de

Funded by
DFG Deutsche Forschungsgemeinschaft
German Research Foundation

**Deutsche Forschungsgemeinschaft e.V.**
https://www.dfg.de