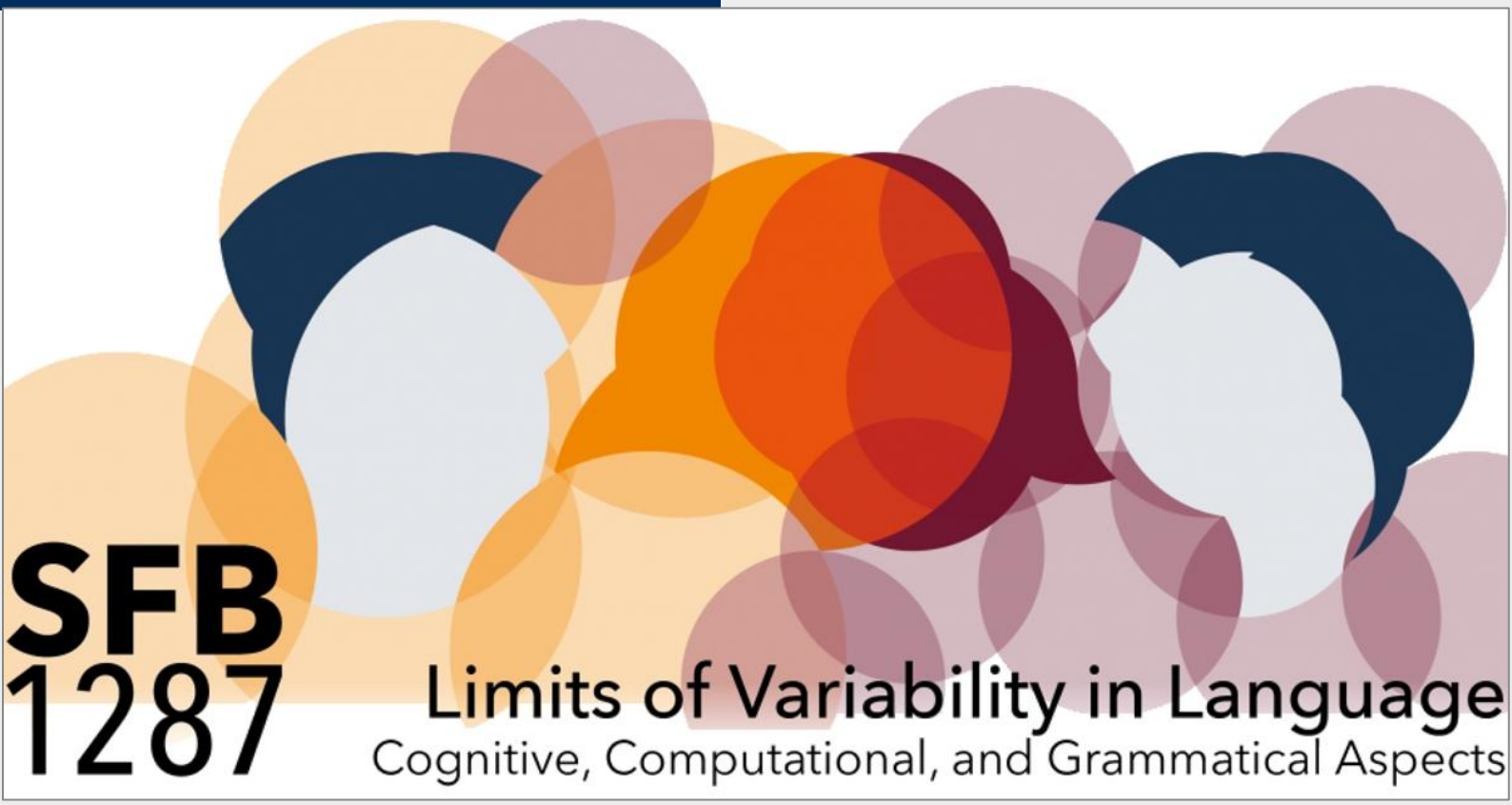

DATA MANAGEMENT PLAN

CRC 1287 | 2025 | Phase 2

Project C07

TABLE OF CONTENTS

GENERAL INFORMATION

Overview

Project number

C07

Name of Experiment / Acronym / Number

Corpus Study

Involved persons

Rowena Garcia, Turgut Agabeyoglu, Liubov Karpova, Arina Shandala

PI or responsible person (head of the study)

Natalie Boll-Avetisyan

Subject area

Linguistics, Language acquisition, Phonotactics, Corpus studies

Method / Type of data

We perform corpus-based lexical statistics to investigate the likelihood of phoneme combinations of our interest (e.g., word-initial consonant-consonant-vowel clusters in German and Russian infant-directed and adult-directed corpora). As data we use modified versions of existing lexical databases (e.g. CELEX German Lemma database, Lexeme database; Stimulstat Russian Lema database, Childes files of German).

Participants (of the study)

-

Short description (of the study)

Among the first things that infants must learn in language acquisition are the phoneme repertoire and the permissible phoneme combinations (i.e., phonotactics) in their ambient language. It has been proposed that phonotactic acquisition is facilitated by universal well-formedness principles (Berent et al., 2007). One of these is the Sonority Sequencing Principle (SSP; Selkirk, 1984; Steriade, 1982), which claims that syllables ideally rise in sonority from the edges towards the nucleus (Clements, 1990). Steeper sonority rises (big difference in sonority index between consonant2 and consonant1; e.g., /blik/) are more well-formed than flatter ones (small difference; e.g., /bnik/), while sonority plateaus (0 difference; e.g., */bdik/) and falls (negative difference; e.g., */lbik/) are typically ill-formed. Studies have shown that even newborn listeners prefer SSP-conforming structures compared to SSP-violating ones (Gomez et al., 2014). Typically, languages are categorically described as either SSP-obeying (e.g., German; Wiese 1988), or SSP-violating (e.g., Russian; Selkirk, 1984). The present study asks whether both SSP-obeying and -violating languages would show gradient effects of the SSP, and specifically, whether infant-directed speech (IDS) provides enough cues for acquiring an SSP generalization.

Comments (optional)

-

Data Management Requirements

Are there requirements regarding the data management from your scholarly / scientific community?

yes

If yes, what are the requirements?

- DFG Guidelines on the Handling of Research Data

- Handlungsempfehlungen zum Umgang mit Forschungsdaten University of Potsdam

- Technische und organisatorische Maßnahmen (TOM) gemäß Art. 32 Abs. 1 DSGVO

- Data Management in Psychological Science

Financial Support

Who is funding the project?

DFG - Deutsche Forschungsgemeinschaft e.V. (German Research Foundation) - https://www.dfg.de/en/

In which funding line and / or which funding program is the project funded?

Collaborative Research Centre 1287 - Project number 317633480

DATASET INFORMATION

Data Origin

Is the dataset being created or re-used?

reused

If re-used, who created the dataset and under which address, PID or URL is the data set available?

- German adult-directed corpus: https://doi.org/10.35111/gs6s-gm48

- German infant-directed corpus: https://osf.io/vpdu6/

- Russian corpus: https://ruscorpora.ru/

Data Collection

When does data collection start? (approximately / tentatively)

-

When does data collection end? (approximately / tentatively)

-

Data Handling

Where is the dataset stored during the project?

- CRC file server

- Box.UP university cloud

- researcher's computer

- computer in the laboratory

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

CRC 1287 File-Server

Which file formats are used?

.txt or .csv for the data and .R for the program code

Which measures of quality assurance are taken for this dataset?

good documentation, code review, and use of common file formats

Data Analysis

When does data analysis start? (approximately / tentatively)

31.11.2021

When does data analysis end? (approximately / tentatively)

31.05.2022

Data Reuse

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?

Those who are creating German or Russian experimental stimuli who would like to control for frequencies might find the data set useful. The corpus frequencies themselves should not be problematic to reuse, since they are not identifiable (cannot be traced to the original participants). StimulStat & Childes data is freely available, hence it should be licit to make our transformations of the data available, too. Celex is not.

LEGAL AND ETHICS

Personal Data

Does this dataset contain personal data?

no

If yes, are these data anonymised?

-

Property Rights

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

no

If yes, please explain which data protected by intellectual or industrial property rights?

-

PUBLICATION

Publishing or Sharing Data

Will this dataset be published or shared?

no

If yes, the principal investigator of the study ensured that the consent form / subject information sheets support publishing of the data?

-

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

-

If yes, when will the data be published?

-

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

-

STORAGE AND LONG-TERM PRESERVATION

Archive

Does this dataset have to be preserved for long-term?

yes

How long does the data need to be stored?

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).

What are the reasons this dataset must be preserved for the long-term?

- Use in a publication / Evidence of good scientific practice
- Reuse (if anonymizable data) in subsequent projects or by others
- Legal obligations
- Documentation because it is socially relevant
- Self-commitment
- Evidence of good scientific practice
- DFG requirements

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

- CRC 1287 File-Server

- OSF

- Research Data Server from Project IN-FDM-BB (a.t.m. not available)

coordinated by:

Universität Potsdam

https://www.uni-potsdam.de

funded by:

DFG Deutsche Forschungsgemeinschaft

https://www.dfg.de/

in cooperation with:

RUHR UNIVERSITÄT BOCHUM RUB

https://www.ruhr-uni-bochum.de