STEFANIA'S PLAN

A Data Management Plan created using DMPonline.be

Creator: Stefania Marzo

Affiliation: KU Leuven (KUL)

Template: KU Leuven BOF-IOF

ID: 205948

Last modified: 26-03-2024

STEFANIA'S PLAN

RESEARCH DATA SUMMARY

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

Dataset Name	Description	New or Reused	Digital or Physical	Digital Data Type	Digital Data Format	Digital Data Volume (MB, GB, TB)
Production (**CitéTALK**)	**Audio-recordings** • **self-recordings** at home and during leisure activities (1h/speaker) • **picture description tasks** *(10 min/speaker)*	☒ Generate new data ☐ Reuse existing data	☒ Digital ☐ Physical	☒ Observational ☐ Experimental ☒ Compiled/ aggregated data ☐ Simulation data ☐ Software ☐ Other ☐ NA	☐ .por ☐ .xml ☐ .tab ☐ .csv ☐ .pdf ☒ .txt ☐ .rtf ☐ .dwg	☐ < 100 MB ☐ < 1 GB ☐ < 100 GB ☐ < 1 TB ☐ < 5 TB ☒ < 10 TB ☐ < 50 TB ☐ > 50 TB

	• ***Sociolinguistic interviews*** *(20 min/speaker)* *= in total* ***1,5 hours of spoken data per speaker or 90 hours in total (60 speakers)***				☐ .tab ☐ .gml ☒ other: - WAV files (audio) - Computational scripts for R (statistics) ☐ NA	☐ NA
PRAAT Annotated corpus (**CitéTALK PRAAT**)	Phonetic annotations of dataset 1 (CitéTALK): exclusively for phonetic features in the softwaare PRAAT	☒ Generate new data ☐ Reuse existing data	☒ Digital ☐ Physical	☒ Observational ☐ Experimental ☒ Compiled/ aggregated data ☐ Simulation data ☐ Software ☐ Other ☐ N	☐ .por ☒ .xml ☐ .tab ☐ .csv ☐ .pdf ☒ .txt ☐ .rtf ☐ .dwg ☐ .tab ☐ .gml ☐ other: ☐ NA:	☐ < 100 MB ☐ < 1 GB ☐ < 100 GB ☒ < 1 TB ☐ < 5 TB ☐ < 10 TB ☐ < 50 TB ☐ > 50 TB ☐ NA
ELAN Annotated corpus (**CitéTALK ELAN**)	Transcriptions of dataset 1 (CitéTALK): transcriptions and annotation of phonetic and morphologic features	☒ Generate new data ☐ Reuse existing data	☒ Digital ☐ Physical	☒ Observational ☐ Experimental ☒ Compiled/ aggregated data ☐ Simulation data ☐ Software ☐ Other ☐ N	☐ .por ☐ .xml ☐ .tab ☐ .csv ☐ .pdf ☒ .txt ☐ .rtf ☐ .dwg ☐ .tab ☐ .gml ☐ other: ☐ NA:	☐ < 100 MB ☐ < 1 GB ☐ < 100 GB ☐ < 1 TB ☒ < 5 TB ☐ < 10 TB ☐ < 50 TB ☐ > 50 TB ☐ NA

Annotated corpus (**CitéTALK EXCEL**)	Excel sheets with detailed transcriptions of dataset 1 (CITETALK), annotated for several variables for 90 informants in different contexts (home/leisure/interviews/tasks) the excel sheets are exported files from ELAN	☒ Generate new data ☐ Reuse existing data	☒ Digital ☐ Physical	☒ Observational ☐ Experimental ☒ Compiled/ aggregated data ☐ Simulation data ☐ Software ☐ Other ☐ N	☐ .por ☒ .xml ☐ .tab ☐ .csv ☐ .pdf ☒ .txt ☐ .rtf ☐ .dwg ☐ .tab ☐ .gml ☐ other: ☐ NA:	☐ < 100 MB ☒ < 1 GB ☐ < 100 GB ☐ < 1 TB ☐ < 5 TB ☐ < 10 TB ☐ < 50 TB ☐ > 50 TB ☐ NA
Attitudes (**CitéEXPERIENCE**)	**Two online experiments** (tool to be determined) Where evaluations and perceptions of linguistic features are measured + the output files (csv) 500 respondents	☒ Generate new data ☐ Reuse existing data	☒ Digital ☐ Physical	☐ Observational ☒ Experimental ☐ Compiled/ aggregated data ☐ Simulation data ☐ Software ☐ Other ☐ NA	☐ .por ☐ .xml ☐ .tab ☒ .csv ☐ .pdf ☒ .txt ☐ .rtf ☐ .dwg ☐ .tab ☐ .gml ☐ other: ☐ NA:	☐ < 100 MB ☐ < 1 GB ☐ < 100 GB ☐ < 1 TB ☒ < 5 TB ☐ < 10 TB ☐ < 50 TB ☐ > 50 TB ☐ NA

R-scripts for statistical analysis (**CitéSCRIPTS**)		☒ Generate new data ☒ Reuse existing data	☒ Digital ☐ Physical	☐ Observational ☐ Experimental ☒ Compiled/ aggregated data ☐ Simulation data ☐ Software ☐ Other ☐ NA	☐ .por ☒ .xml ☐ .tab ☒ .csv ☐ .pdf ☒ .txt ☐ .rtf ☐ .dwg ☐ .tab ☐ .gml ☒ other: scripts in used in R studio ☐ NA:	☐ < 100 MB ☒ < 1 GB ☐ < 100 GB ☐ < 1 TB ☐ < 5 TB ☐ < 10 TB ☐ < 50 TB ☐ > 50 TB ☐ NA

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

Question not answered.

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? If so, refer to specific datasets or data types when appropriate and provide the relevant ethical approval number.

Yes, we are using human subject data.
We plan to collect linguistic (**CitéTALK**) and attitudinal (**CitéEXPERIENCE**) data from speakers, including specific types of data described below. We have already received ethical approval for the majority of the data collection, which includes participants aged 16 and over. An amendment is being prepared to extend the data collection to younger people, specifically those between the ages of 13 and 16.

Will you process personal data? If so, please refer to specific datasets or data types when appropriate and provide the KU Leuven or UZ Leuven privacy register number (G or S number).

For each data type (self-recordings, interviews, tasks and attitude experiments), the project team will produce detailed documentation consisting of metadata. We will collect the following personal demographic metadata: Year of birth, gender, place of living during childhood, educational qualification, current place of living, citizenship and ethnic background (migration or not), place of birth and living of grand-parents/partners, social network, and language use at home, at work and among friends.
Also, each respondent will receive a unique identifier.

Privacy Registry Reference: G-2023-6774-R2(AMD)

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.

No.

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material or Data transfer agreements, Research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

No.

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

Ideally, the production corpus (**CitéTALK**) will be shared as a corpus (at the end of the project). This issue will be dealt with in the next months.

DOCUMENTATION AND METADATA

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g. in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, codebook.tsv etc. where this information is recorded).

A comprehensive codebook is currently under development, which will be a detailed explanation of each socio-demographic parameter, each language variable, and each language annotation (across all language levels). This will include a code for all parameters, covering all levels and detailing how everything is annotated. In addition, each speaker is assigned a unique ID for both the production and attitude datasets. The process for the production dataset is slightly more complex, as each ID is also pseudonymized; this metadata set is keyed securely. In addition, the data collection protocol, analysis methods, and decisions made are documented in a readme file.

Will a metadata standard be used to make it easier to find and reuse the data? NO
If so, please specify which metadata standard will be used.

If not, please specify which metadata will be created to make the data easier to find and reuse.

Question not answered.

DATA STORAGE & BACK-UP DURING THE RESEARCH PROJECT

Where will the data be stored?

- We will use the desktop file storage offered at KU Leuven, including central network folders KU Leuven OneDrive (this storage space is safe and automatically backed up).
- We will use two external hard drives for local storage.
- We are considering the use of back-end storage at KU Leuven for a supplementary and safe backup of the audio files (larger volume types): during the project
- We will also *additionally publish our anonymized datasets on Open Science Framework (OSF).*

How will the data be backed up?

- We will use two external hard drives for local storage.
- Storage on the KU Leuven OneDrive – it is safe and automatically backed up.
- Back-end storage at KU leuven for supplementary backup (larger volume):

Is there currently sufficient storage & backup capacity during the project?

If no or insufficient storage or backup capacities are available, explain how this will be taken care of.

Yes: external hard drivers are used + KU Leuven OneDrive + backup solution at KU Leuven (back-end storage)

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

All PCs purchased through the Faculty have Bitlocker pre-installed and Back-end storage at KU Leuven is also safe.

SAFE STORAGE OF PSEUDONIMISED SPOKEN DATA

Spoken data:

- A list of 60 names (boys and girls) is created. Each name on the list is assigned a random number between 1 and 60.
- The names of the respondents are assigned a random sequence number.
- The sequence numbers are linked, and thus a random (girl or boy) name is assigned to each respondent.

Storage mode: the PI keeps the key in a separate folder created specifically for this document. The keys are stored in a password-protected Word document. The key and the data are not in the same place.

The online questionnaires:
The questionnaires are completely anonymous and thus don't have to be pseudonomised or stored with a key.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

The two external hard drives for local storage will be covered by the project budget (+/-€ 150 for 5TB = +/- € 300 for the backup of the dataset)
Back-end storage will also be covered by the project: around 225 euro per TB – only mp3 versions of the audio files will be stored.

DATA PRESERVATION AFTER THE END OF THE RESEARCH PROJECT

Which data will be retained for 10 years (or longer, in agreement with other retention policies that are applicable) after the end of the project?

In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

It is intended that all data will be preserved for longitudinal research (by the PI Stefania Marzo), that is, kept for more than 10 years. This approach is taken because the project marks the beginning of long-term research aimed at studying language change throughout the life span of speakers.

Where will these data be archived (stored and curated for the long-term)?

We will rely on a file storage option offered at KU Leuven (Archive storage per TB).
This type of storage is used to cost-efficiently store large amounts of data for long periods of time.
See: File storage - ICTS Servicecatalogus (kuleuven.be)

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

Around € 250 euro / year. These costs will be covered and anticipated (for 10 years) by the project budget

DATA SHARING AND REUSE

Will the data (or part of the data) be made available for reuse after/during the project?
Please explain per dataset or data type which data will be made available.

Production (**CitéTALK**)	Raw material (audio data):	not shared
PRAAT Annotated corpus (**CitéTALK PRAAT**)	Phonetic annotations	Not shared
ELAN Annotated corpus (**CitéTALK ELAN**)	Exported transcriptions (not Elan files)	Shared (only the exported transcriptions in excel) at the end of the project: platform to be discussed
Annotated corpus (**CitéTALK EXCEL**)	Excel sheets with transcriptions and annotation	Shared at the end of the project: platform to be discussed
Attitudes (**CitéEXPERIENCE**)	Datasets (anonyme) with csv files	Shared via Open Science Framework (OSF)
R-scripts for statistical analysis (**CitéSCRIPTS**)	Scripts for statistical analyses and output of the analyses	Shared via OSF

If access is restricted, please specify who will be able to access the data and under what conditions.

Question not answered.

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)?

Please explain per dataset or data type where appropriate.

Privacy and ethical issues.

Where will the data be made available?

If already known, please provide a repository per dataset or data type.

We are considering making the transcribed corpus of spoken data available through the use of an online tool or repository for access (to be determined).

When will the data be made available?

This initiative will be in place at the end of the project, after the dissertation defense and publication of the first papers. This approach ensures academic integrity and strategic publication of research results, allowing for peer review and scholarly discourse prior to public access.

Which data usage licenses are you going to provide?

If none, please explain why.

Copyright for analyses and transcriptions is held by PI under the license CC-BY-NC 4.0. Participants retain full access to their data if they wish.

Do you intend to add a persistent identifier (PID) to your dataset(s), e.g. a DOI or accession number? If already available, please provide it here.

No.

What are the expected costs for data sharing? How will these costs be covered?

This aspect is still to be determined, but possible costs will be covered by the project.

RESPONSIBILITIES

Who will manage data documentation and metadata during the research project?

Stefania Marzo (PI)

Who will manage data storage and backup during the research project?

Stefania Marzo (PI)

Who will manage data preservation and sharing?

Stefania Marzo (PI)

Who will update and implement this DMP?

Stefania Marzo (PI)