Relatively complex or relatively simple? Toward new ways of analyzing language variation

A Data Management Plan created using DMPonline.be

Creators: Thomas Van Hoey, Benedikt Szmrecsanyi https://orcid.org/0000-0001-8844-6602

Affiliation: KU Leuven (KUL)

Template: KU Leuven BOF-IOF

Principal Investigator: Benedikt Szmrecsanyi https://orcid.org/0000-0001-8844-6602

Data Manager: Thomas Van Hoey

Grant number / URL: 3H220293

ID: 192901

Start date: 01-10-2022

End date: 30-09-2026

Project abstract:

Pilootonderzoek uit het lab van de PI in Leuven (Gardner et al. 2021) toont dat grammaticale variatie taalproductie niet bemoeilijkt, aangezien keuzecontexten geen onvloeiendheden (lange pauzes en um's en uh's) aantrekken in een groot corpus van gesproken Engels. Dit is verbluffend gezien vastgeroeste opvattingen in de theoretische taalkunde. Voortbouwend op deze bevindingen zal het project een basis leggen voor grootschalig vervolgonderzoek. We zullen de dataset uit Gardner et al. (2021) uitbreiden en de analyse verfijnen om verschillende dringende onopgeloste vragen aan te pakken: Gedragen niet-canonieke onvloeiendheden (bijv. discourse markers zoals like) zich bijvoorbeeld anders dan um en uh? Zijn verschillende soorten keuzecontexten meer of minder vatbaar voor het aantrekken van onvloeiendheden? Zijn beperktere keuzes “gemakkelijker” dan vrijere keuzes? Onze resultaten zullen grote implicaties hebben voor taalkundige theorievorming.

Last modified: 16-01-2023

Relatively complex or relatively simple? Toward new ways of analyzing language variation

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 / ID	Description	New or reuse	Digital or Physical data	Data Type	File format	Data volume	Physical volume
		Indicate: ***N****(ew data) or* ***E****(xisting data)*	Indicate: **D**(igital) or **P**(hysical)	Indicate: **A**udiovisual **I**mages **S**ound **N**umerical **T**extual **M**odel **SO**ftware Other (specify)		Indicate: <1GB <100GB <1TB <5TB >5TB NA	
Switchboard corpus	Corpus	E	D	T	XML	<100 GB	
Switchboard subcorpus	Corpus	N	D	T	Rds	<1 GB	
R models	R models	N	D	M	Rmd	<1 GB	
Annotated subset	Annotated set based on the subcorpus	N	D	T	csv or tsv	< 1 GB	
Switchboard audiofiles	audio files	E	D	S	wav	< 100 GB	

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:

The switchboard corpus can be found here: https://groups.inf.ed.ac.uk/switchboard/index.html
The associated LDC number is catalogue number LDC2009T26.

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.

- No

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).

- No

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.

- No

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).

/1/ Documentation sheets for the project as a whole. These encompassing documents sum up the work flow and methodology of the project, the types of data used, where these can be found (plus document name), and what the relation between these data types is. Amongst others, the following documentation sheets will be created:

- Sheet with descriptive information: the project title, the main contributors, the start and end date of the project, the funding for the project

- Sheet with methodological information for the project as a whole: the aims and research questions, the design, the final sample (subcorpus), the software used for data collection and data analyses, the types of data gathered, a summary of the analyses and results, references to publications or other means of dissemination.

- Sheet with administrative data: rights management, technical information concerning formats, i.e., a README file.

/2/ Methodology reports will be created, in the form of notebooks such as R markdown files, Quarto documents or Jupyter notebooks. These allow for replicable analyses, and can be shared in repositories in the future. Furthermore, they are accompanied by comments detailing certain steps.

/3/ For all documents, special care is awarded to the document names and folder structure. Consistent and straightforward names will be selected for the documents. Additionally, readme-files are included in every folder to guide the reader to the relevant overview documents containing documentation to increase accessibility and usability of the data.

Will a metadata standard be used to make it easier to find and reuse the data ?
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.

- Yes

The Switchboard corpus is stored in xml format (Nite xml). We will follow that format, as well as translate into more accessible (simpler) formats that can be shared in csv/tsv files or rds files.

Data Storage & Back-up during the Research Project

Where will the data be stored?

- OneDrive (KU Leuven)
- Other (specify below)

During the project, we will store data on OneDrive (provided by KU Leuven). Afterwards, we will use repositories like RDR (https://www.kuleuven.be/rdm/en/rdr) or OSF (https://osf.io/).

How will the data be backed up?

- Standard back-up provided by KU Leuven ICTS for my storage solution

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

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

Data are stored on devices that are secured with Bitlocker provided by KU Leuven, as well as in OneDrive, which is secured through 2FA via KU Leuven.

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

The size of the data will not exceed what has been provided through KU Leuven in terms of OneDrive or personal computers protected by bitlocker.

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...).

- All data will be preserved for 10 years according to KU Leuven RDM policy

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

- KU Leuven RDR
- Other (specify below)

Depending on the needs, the KU Leuven repository RDR, or the Open Science Framework OSF.

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

Both of these repositories are funded and free of charge.

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.

- Yes, as open data

- Switchboard corpus: The existing dataset is already available under a Creative Commons Attribution-Noncommercial Share Alike 3.0 license.
- Switchboard subcorpus: This generated dataset will be made available in a repository in a new format suitable to the research conducted. This dataset will be made available for future reanalysis and replication.
- R models: These models will be made available for future reanalysis and replication.
- Annotated subset: This annotated subset will be made available for future reanalysis and replication.
- Switchboard audiofiles: These existing files are already available (see above).

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

NA

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.

- No

Where will the data be made available?

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

- KU Leuven RDR (Research Data Repository)
- Other data repository (specify below)

RDR or OSF. We will choose either depending on the needs and potential these repositories offer.

When will the data be made available?

- Upon publication of research results

Which data usage licenses are you going to provide?

If none, please explain why.

- CC-BY 4.0 (data)
- Other (specify below)

The data will be shared with either a CC-BY 4.0 licence or CC-BY-NC-4.0 licence.

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.

- Yes, a PID will be added upon deposit in a data repository

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

No costs, the repositories are well funded and free of charge.

Responsibilities

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

The postdoctoral assistant Thomas Van Hoey.

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

The postdoctoral assistant Thomas Van Hoey.

Who will manage data preservation and sharing?

The postdoctoral assistant Thomas Van Hoey.

Who will update and implement this DMP?

The postdoctoral assistant Thomas Van Hoey.