The emergence of syntactic complexity

A Data Management Plan created using DMPonline.be

Creators: Karen Lahousse, n.n. n.n.

Affiliation: KU Leuven (KUL)

Funder: Fonds voor Wetenschappelijk Onderzoek - Research Foundation Flanders (FWO)

Template: FWO DMP (Flemish Standard DMP)

Principal Investigator: n.n. n.n., Karen Lahousse

Grant number / URL: G021323N

ID: 200476

Start date: 18-09-2023

End date: 08-09-2027

Last modified: 22-06-2023

The emergence of syntactic complexity
FWO DMP (Flemish Standard DMP)

1. 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.

				Only for digital data	Only for digital data	Only for digital data	Only for physical data
Dataset Name	Description	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)	Physical volume
		Please choose from the following options: • Generate new data • Reuse existing data	*Please choose from the following options:* • Digital • Physical	*Please choose from the following options:* • Observational • Experimental • Compiled/aggregated data • Simulation data • Software • Other • NA	*Please choose from the following options:* • .por, .xml, .tab, .cvs,.pdf, .txt, .rtf, .dwg, .gml, ... • NA	*Please choose from the following options:* • <100MB • <1GB • <100GB • <1TB • <5TB • <10TB • <50TB • >50TB • NA	
Videos	Recordings of the York-DeCat corpus (see De Cat 2007 for details)	Reuse existing data	digital	observational	mp4	<100GB	
Transcriptions	The project will generate transcriptions of the 25 not yet transcribed hours.	Generate new data	digital	other	.txt	<1GB	
OA corpus	An open access corpus of adult-child interactions. = Fully coded OA version of the York-De Cat corpus (with 10 syntactic and 7 information-structural codes).	Combination of exiting and new data.	digital	Compiled data	video + transcriptions + code in R	<100GB	
Linguistic analysis	Quantitative results of linguistic analysis from corpus data and the experiments.	Generate new data	Digital	Compiled (code + statistical tests)	to be determined	<1GB	
Setup of experiment	Input of the experiment: context + prompt sentences (tool: open-source software PsychoPy)	Generate new data	Digital	Compiled (software code + content)	to be determined	<1GB	
Videos of experiments	Videos of children reacting to the prompts in the experiment	Generate new data	Digital	Observational	video	<100GB	
Sheets of informed consent	Informed consent form of parents of child participants	Generate new data	Physical	NA	NA	NA	120 sheets of paper + scanned version in pdf
Results pretests	Results of baseline pre-tests to measure syntactic competence and working memory and scalar implicatures.						

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 York-DeCat corpus (see De Cat 2007 for details) is not yet online, but one of the objectives of this project is to create open access material:
(i) Manual transcription + coding of part of the York&Cat corpus;
(ii) conversion of the Varbrul-coded York&Cat corpus into R.

Design of an open access, fully coded, longitudinal corpus of spontaneous child-adult interaction. The York & Cat corpora will be coded for 10 syntactic and 7 IS-related factors relevant to this project and of potential interest for other linguists. Transcriptions for 55/80h are already available via CHILDES. This part of the corpus has been deep-coded (using Varbrul format, 2002 version), but the coded version is not available to the linguistic community. The coded version will be converted into R-format (using Python scripts); all Python and R scripts will be made available in open access. Out of the 55h of transcribed data, all child utterances have been coded (N=17,815), as well as 5,613 adult utterances (6h of recordings). A larger subset of adult utterances will be coded (12h), and the remaining 25h will be transcribed and coded.

The videos of the York-De Cat corpus are currently owned and stored by Cécile De Cat (University of Leeds), co-PI of the project.

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- Yes, human subject data

Existing + newly generated videos of living persons, to be made available in OA as far as privacy legislation allows.

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.

- Yes

Videos: existing videos in the York-De Cat corpus + newly generated videos (experiment)
Pre-tests: name, address, email + baseline pre-tests of individual children to measure syntactic competence and working memory and scalar implicatures.

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.

- Yes

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/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.

- Yes

Research collaboration agreement (FWO) with University of Leeds: University of Leeds (Cécile De Cat) is owner of the York-De Cat corpus.

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.

- Yes

Existing videos of York-De Cat corpus are owned by Cécile De Cat (University of York).

2. 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).

We will add a README.txt file to the data (following template README of the KU Leuven).

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

- Yes

Transcription of the data according to conventions of CHILDES (http://childes.psy.cmu.edu)
Coding protocol for the linguistic analysis generated by the project.

3. Data storage & back-up during the research project

Where will the data be stored?

During the research project, we are currently considering to store the data either on the KU Leuven-supported Onedrive cloud service or, preferably, on SharePoint (not linked to one person, more communication and collaboration options via MS Teams).

How will the data be backed up?

Automatically (Onedrive and SharePoint offer versioning).

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely.
If no or insufficient storage or backup capacities are available, then 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?

Bitlocker (KU Leuven) + we will make use of the most up to date security features offered by KU Leuven and Leeds University.

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

Probably nothing.

4. Data preservation after the end of the research project

Which data will be retained for at least five 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...).

Data with raw personal data will be destroyed when necessary. We will determine the exact procedure in the course of our PRET-review.
All other data will be retained for at least 10 years following KU Leuven policy.

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

As much as possible on a data repository (RDR KU Leuven for the newly generated data, data repository of University of Leeds for the OA York-De Cat corpus).

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

Probably none

5. Data sharing and reuse

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

- Yes, in an Open Access repository

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 in the comment section per dataset or data type where appropriate.

- Yes, Privacy aspects

In case of any pseudomized data, we will offer it in restricted access.

Where will the data be made available? If already known, please provide a repository per dataset or data type.

As much as possible on a data repository (RDR KU Leuven for the newly generated data, data repository of University of Leeds for the OA York-De Cat corpus). We will consider publishing parts of the data in the OA Childes database.

When will the data be made available?

The OA De Cat-York corpus will be published as soon as possible.
The other data generated during the project will be published after the project.

Replication data for articles published during the project may be published in a repository as required by the journal.

Which data usage licenses are you going to provide? If none, please explain why.

Creative commons Attribution (CC-BY-4.0.) (to be checked by University of Leeds for the existing corpus).

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.

- Yes

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

Probably none.

6. Responsibilities

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

Karen Lahousse (KU Leuven) & Cécile De Cat (University of Leeds)

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

Karen Lahousse (KU Leuven) & Cécile De Cat (University of Leeds)

Who will manage data preservation and sharing?

Karen Lahousse (KU Leuven) & Cécile De Cat (University of Leeds)

Who will update and implement this DMP?

Karen Lahousse (KU Leuven) & Cécile De Cat (University of Leeds)