FWO DMP Template - Flemish Standard Data Management Plan

Project supervisors (from application round 2018 onwards) and fellows (from application round 2020 onwards) will, upon being awarded their project or fellowship, be invited to develop their answers to the data management related questions into a DMP. The FWO expects a **completed DMP no later than 6 months after the official start date** of the project or fellowship. The DMP should not be submitted to FWO but to the research co-ordination office of the host institute; FWO may request the DMP in a random check.

At the end of the project, the **final version of the DMP** has to be added to the final report of the project; this should be submitted to FWO by the supervisor-spokesperson through FWO's e-portal. This DMP may of course have been updated since its first version. The DMP is an element in the final evaluation of the project by the relevant expert panel. Both the DMP submitted within the first 6 months after the start date and the final DMP may use this template.

The DMP template used by the Research Foundation Flanders (FWO) corresponds with the Flemish Standard Data Management Plan. This Flemish Standard DMP was developed by the Flemish Research Data Network (FRDN) Task Force DMP which comprises representatives of all Flemish funders and research institutions. This is a standardized DMP template based on the previous FWO template that contains the core requirements for data management planning. To increase understanding and facilitate completion of the DMP, a standardized **glossary** of definitions and abbreviations is available via the following link.

1. General Project Information	
Name Grant Holder & ORCID	**Veronica Juliana Schmalz 0000-0002-1636-6133**
Contributor name(s) (+ ORCID) & roles	**Piet Desmet (supervisor) 0000-0002-9849-0874** **Paul Van Eecke (co-supervisor) 0000-0001-9153-9092**
Project number[1] & title	**1108723N A COMPUTATIONAL MODEL OF THE USAGE-BASED ACQUISITION OF ABSTRACT CONSTRUCTIONS AND GRAMMATICAL CATEGORIES**
Funder(s) GrantID[2]	
Affiliation(s)	☑ KU Leuven ☐ Universiteit Antwerpen ☐ Universiteit Gent ☐ Universiteit Hasselt ☐ Vrije Universiteit Brussel ☐ Other: Provide ROR[3] identifier when possible:

[1] "Project number" refers to the institutional project number. This question is optional since not every institution has an internal project number different from the GrantID. Applicants can only provide one project number.
[2] Funder(s) GrantID refers to the number of the DMP at the funder(s), here one can specify multiple GrantIDs if multiple funding sources were used.
[3] Research Organization Registry Community. https://ror.org/

Please provide a short project description	This research project aims to computationally model the acquisition of usage-based construction grammars, focusing on the learning of modular constructions and networks of grammatical categories. Methodologically, it is rooted in usage-based, constructivist theories of language acquisition, particularly in research on pattern finding and intention reading. I will computationally operationalise the key insights behind these theories through syntactico-semantic pattern finding operators that facilitate the learning of construction grammars. I will build further on existing work on learning item-based constructions from semantically annotated corpora and advance the state of the art by extending this to modular constructions. Concretely, this will involve the implementation of novel learning operators for the generalization of new linguistic observations with respect to previously acquired constructions. The results will offer a detailed mechanistic insight into the usage- based acquisition of construction grammars. Theoretically, they will contribute to a better understanding of the assumptions and consequences of both usage-based theories of language acquisition and constructionist approaches to language. Practically, they will facilitate the valorisation of construction grammars in language technology applications such as question answering systems, tools for corpus analysis, and intelligent tutoring systems.

2. 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[4].

				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
CLEVR	A diagnostic dataset for compositional language and elementary visual reasoning consisting of VQA utterances associated to objects in a scene. Language: English.	☐ 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: json ☐ NA	☐ < 100 MB ☐ < 1 GB ☒ < 100 GB ☐ < 1 TB ☐ < 5 TB ☐ < 10 TB ☐ < 50 TB ☐ > 50 TB ☐ NA	
German Cases acquisition through OEIT	Oral productions from a OEIT administered to 36 students learning German as a second	☐ Generate new data ☒ Reuse existing data	☒ Digital ☐ Physical	☐ Observational ☒ Experimental ☐ Compiled/ aggregated data ☐ Simulation data	☐ .por ☐ .xml ☐ .tab ☒ .csv ☐ .pdf ☐ .txt	☐ < 100 MB ☒ < 1 GB ☐ < 100 GB ☐ < 1 TB ☐ < 5 TB ☐ < 10 TB	

[4] Add rows for each dataset you want to describe.

	language in university. These data have been transcribed and annotated with abstract meaning representations. Language: German.			☐ Software ☐ Other ☐ NA	☐ .rtf ☐ .dwg ☐ .tab ☐ .gml ☒ other: json ☐ NA	☐ < 50 TB ☐ > 50 TB ☐ NA	
The little Prince corpus	Annotation of the novel by A. Saint-Exupéry with abstract meaning representations. The dataset contains 1,562 sentences. Language: English.	☐ 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: json ☐ NA	☐ < 100 MB ☒ < 1 GB ☐ < 100 GB ☐ < 1 TB ☐ < 5 TB ☐ < 10 TB ☐ < 50 TB ☐ > 50 TB ☐ NA	
Abstract Meaning Representation (AMR) Annotation Release 3.0	Sentences from broadcast conversations, discussion forums, newswire, web collections, weblogs annotated with	☐ Generate new data ☒ Reuse existing data	☒ Digital ☐ Physical	☐ Observational ☐ Experimental ☐ Compiled/ aggregated data ☐ Simulation data ☐ Software ☒ Other	☐ .por ☒ .xml ☐ .tab ☐ .csv ☐ .pdf ☐ .txt ☐ .rtf ☐ .dwg	☐ < 100 MB ☐ < 1 GB ☒ < 100 GB ☐ < 1 TB ☐ < 5 TB ☐ < 10 TB ☐ < 50 TB ☐ > 50 TB	

	abstract meaning representations. In total, the corpus contains 59,255 meaning representations. Language: English.			□ NA	□ .tab □ .gml ☒ other: json □ NA	□ NA	

GUIDANCE:

DATA CAN BE DIGITAL OR PHYSICAL (FOR EXAMPLE BIOBANK, BIOLOGICAL SAMPLES, …). DATA TYPE: DATA ARE OFTEN GROUPED BY TYPE (OBSERVATIONAL, EXPERIMENTAL ETC.), FORMAT AND/OR COLLECTION/GENERATION METHOD.

EXAMPLES OF DATA TYPES: OBSERVATIONAL (E.G. SURVEY RESULTS, SENSOR READINGS, SENSORY OBSERVATIONS); EXPERIMENTAL (E.G. MICROSCOPY, SPECTROSCOPY, CHROMATOGRAMS, GENE SEQUENCES); COMPILED/AGGREGATED DATA[5] (E.G. TEXT & DATA MINING, DERIVED VARIABLES, 3D MODELLING); SIMULATION DATA (E.G. CLIMATE MODELS); SOFTWARE, ETC.

EXAMPLES OF DATA FORMATS: TABULAR DATA (.POR,. SPSS, STRUCTURED TEXT OR MARK-UP FILE XML, .TAB, .CSV), TEXTUAL DATA (.RTF, .XML, .TXT), GEOSPATIAL DATA (.DWG,. GML, ..), IMAGE DATA, AUDIO DATA, VIDEO DATA, DOCUMENTATION & COMPUTATIONAL SCRIPT.

DIGITAL DATA VOLUME: PLEASE ESTIMATE THE UPPER LIMIT OF THE VOLUME OF THE DATA PER DATASET OR DATA TYPE.

PHYSICAL VOLUME: PLEASE ESTIMATE THE PHYSICAL VOLUME OF THE RESEARCH MATERIALS (FOR EXAMPLE THE NUMBER OF RELEVANT BIOLOGICAL SAMPLES THAT NEED TO BE STORED AND PRESERVED DURING THE PROJECT AND/OR AFTER).

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.	- CLEVR https://cs.stanford.edu/people/jcjohns/clevr/ - German Cases acquisition through OEIT http://doi.org/10.22599/jesla.56 https://doi.org/10.29140/9781914291050 - The little Prince https://amr.isi.edu/download.html - AMR https://doi.org/10.35111/44cy-bp51

[5] These data are generated by combining multiple existing datasets.

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, please describe these issues further and refer to specific datasets or data types when appropriate.	☐ Yes, human subject data ☐ Yes, animal data ☐ Yes, dual use ☒ No If yes, please describe:
Will you process personal data[6]? If so, briefly describe the kind of personal data you will use. Please refer to specific datasets or data types when appropriate. If available, add the reference to your file in your host institution's privacy register.	☐ Yes ☒ No If yes: - Short description of the kind of personal data that will be used: - Privacy Registry Reference:
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 ☒ No If yes, please comment:
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 to what data they relate and what restrictions are in place.	☐ Yes ☒ No If yes, please explain:

[6] See Glossary Flemish Standard Data Management Plan

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 to what data they relate and which restrictions will be asserted.	☐ Yes ☒ No If yes, please explain:

3. 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).	All necessary contextual details will be documented within the source code itself, and in the user guide that accompanies the code. This documentation includes: • documentation of source code inside the source code (purpose of classes and methods, their input/output and parameters) • documentation of how to use these functions in the user guide that accompanies the code. • documentation of experimental setup for all carried out experiments • all code and documents are versioned in a versioning system (Git on gitlab.kuleuven.be).
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. *REPOSITORIES COULD ASK TO DELIVER METADATA IN A CERTAIN FORMAT, WITH SPECIFIED ONTOLOGIES AND VOCABULARIES, I.E. STANDARD LISTS WITH UNIQUE IDENTIFIERS.*	☐ Yes ☒ No If yes, please specify (where appropriate per dataset or data type) which metadata standard will be used: If no, please specify (where appropriate per dataset or data type) which metadata will be created: No new data will be collected.

4. Data Storage & Back-up during the Research Project	
Where will the data be stored?	The data will be stored centrally on storage facilities of the research unit and of the university (KU Leuven).
How will the data be backed up? *What storage and backup procedures will be in place to prevent data loss? Describe the locations, storage media and procedures that will be used for storing and backing up digital and non-digital data during research.*[7] *Refer to institution-specific policies regarding backup procedures when appropriate.*	All source code and thesis texts will be versioned in Git (with a gitlab server at gitlab.kuleuven.be). All other thesis data will be stored on the Box cloud, provided by KU Leuven or on the KU Leuven central drives (I and J).
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 ☐ No If yes, please specify concisely: If no, please specify:
How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons? *Clearly describe the measures (in terms of physical security, network security, and security of computer systems and files) that will be taken to ensure that stored and transferred data are safe.* [7]	The data is accessible by myself, my supervisors, and the KU Leuven IT department that administers the Gitlab and Box server. Only these subjects listed above can access the data after having successfully passed solid authentication measures implemented by KU Leuven.

[7] Source: Ghent University Generic DMP Evaluation Rubric: https://osf.io/2z5g3/

What are the expected costs for data storage and backup during the research project? How will these costs be covered?	The cost of versioning this data is negligible (in the order of several megabytes) and covered by the central Gitlab server's operation costs.

5. 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...).	All source code, a copy of the employed datasets and my thesis will be preserved for at least five years after the end of this research project. We will not redistribute these data sets, but we'll keep a copy on KU Leuven servers for internal records and versioning needs.
Where will these data be archived (stored and curated for the long-term)?	These data will be archived on the data server of the KU Leuven during and for at least 5 years after the end of the research. For that, more than the required storage space is available.
What are the expected costs for data preservation during the expected retention period? How will these costs be covered?	The cost of versioning this data is negligible (in the order of several megabytes) and covered by the central Gitlab server's operation costs.

6. 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. *NOTE THAT 'AVAILABLE' DOES NOT NECESSARILY MEAN THAT THE DATA SET BECOMES OPENLY AVAILABLE, CONDITIONS FOR ACCESS AND USE MAY APPLY. AVAILABILITY IN THIS QUESTION THUS ENTAILS BOTH OPEN & RESTRICTED ACCESS. FOR MORE INFORMATION: HTTPS://WIKI.SURFNET.NL/DISPLAY/STANDARDS/INFO-EU-REPO/#INFOEUREPO-ACCESSRIGHTS*	☐ Yes, in an Open Access repository ☐ Yes, in a restricted access repository (after approval, institutional access only, …) ☐ No (closed access) ☒ Other, please specify: all data is already publicly available
If access is restricted, please specify who will be able to access the data and under what conditions.	Not applicable
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.	☐ Yes, privacy aspects ☐ Yes, intellectual property rights ☐ Yes, ethical aspects ☐ Yes, aspects of dual use ☐ Yes, other ☒ No If yes, please specify:
Where will the data be made available? If already known, please provide a repository per dataset or data type.	These data will be made available on the data server of the KU Leuven. For what concerns the datasets and not the code written by me, we will not redistribute these datasets, but we'll keep a copy on KU Leuven servers for internal records and versioning needs.

When will the data be made available? *THIS COULD BE A SPECIFIC DATE (DD/MM/YYYY) OR AN INDICATION SUCH AS 'UPON PUBLICATION OF RESEARCH RESULTS'.*	It is already available.
Which data usage licenses are you going to provide? If none, please explain why. *A DATA USAGE LICENSE INDICATES WHETHER THE DATA CAN BE REUSED OR NOT AND UNDER WHAT CONDITIONS. IF NO LICENCE IS GRANTED, THE DATA ARE IN A GREY ZONE AND CANNOT BE LEGALLY REUSED. DO NOTE THAT YOU MAY ONLY RELEASE DATA UNDER A LICENCE CHOSEN BY YOURSELF IF IT DOES NOT ALREADY FALL UNDER ANOTHER LICENCE THAT MIGHT PROHIBIT THAT.* *EXAMPLE ANSWER: E.G. "DATA FROM THE PROJECT THAT CAN BE SHARED WILL BE MADE AVAILABLE UNDER A CREATIVE COMMONS ATTRIBUTION LICENSE (CC-BY 4.0), SO THAT USERS HAVE TO GIVE CREDIT TO THE ORIGINAL DATA CREATORS."* [8]	We will not modify the corpora in any way, we will not redistribute them - as it would be in conflict with the licensing terms. The corpora can be acquired freely from the original license holders: • CLEVR dataset: license: Creative Commons CC BY 4.0 • AMR, Little Prince Corpus: licensing terms according to the Linguistic Data Consortium, non-membership license. The rest of the data from the project that can be shared will be made available under a creative commons attribution license (cc-by 4.0), so that users have to give credit to the original data creators.
Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, please provide it here. *INDICATE WHETHER YOU INTEND TO ADD A PERSISTENT AND UNIQUE IDENTIFIER IN ORDER TO IDENTIFY AND RETRIEVE THE DATA.*	☐ Yes ☒ No If yes:
What are the expected costs for data sharing? How will these costs be covered?	**Not applicable**

[8] Source: Ghent University Generic DMP Evaluation Rubric: https://osf.io/2z5g3/

7. Responsibilities	
Who will manage data documentation and metadata during the research project?	My co-supervisor, Paul Van Eecke, and I, Veronica Juliana Schmalz.
Who will manage data storage and backup during the research project?	My co-supervisor, Paul Van Eecke, and I, Veronica Juliana Schmalz.
Who will manage data preservation and sharing?	My co-supervisor, Paul Van Eecke, and I, Veronica Juliana Schmalz.
Who will update and implement this DMP?	My co-supervisor, Paul Van Eecke, and I, Veronica Juliana Schmalz.