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	**Freek Van de Velde (0000-0003-3050-2207), supervisor**
Contributor name(s) (+ ORCID) & roles	**Dirk Speelman (0000-0003-1561-1851), co-supervisor** **Anthe Sevenants (0000-0002-5055-770X), researcher**
Project number[1] & title	3H220497, Connecting morphosyntax and lexical semantics with Elastic Net regression
Funder(s) GrantID[2]	ZKE1867 - G059922N
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 project proposes to use regularization methods from machine learning, more specifically Elastic Net regression (and its siblings Ridge and Lasso), to look into lexical semantic effects in morphosyntactic alternances. These regularization techniques apply shrinkage to the coefficients and can thus be used for variable selection, especially when the number of predictors is very large. In variationist studies, this is often the case if one wishes to enter lexemes associated with a construction into a regression model to predict constructional variants. We combine the Elastic Net regulator with k-fold cross-validation - a standard procedure - to avoid overfitting. Our approach mitigates the various drawbacks present in alternative approaches that are currently used in variationist linguistics, like random factors in mixed models and collostructional analysis. We look at ten multifactorially driven alternances from Dutch. The project offers a transparent pipeline that can easily be extrapolated to other case studies, and to other languages.

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
Corpora and datasets	We will both use existing text corpora, viz. SoNaR (Oostdijk N., M. Reynaert, V. Hoste & I. Schuurman (2013). The Construction of a 500-Million-Word Reference Corpus of Contemporary Written Dutch. In: P. Spyns & J.Odijk (red.), *Essential Speech and Language*	☒ 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	

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

	Technology for Dutch. Theory and Applications of Natural Language Processing. Heidelberg: Springer, 219-247. and Dutch C-CLAMP (Piersoul, Jozefien, Robbert De Troij & Freek Van de Velde. 2021. ‘150 years of written Dutch: the construction of the Dutch Corpus of Contemporary and Late Modern Periodicals’. *Nederlandse*						

	Taalkunde 26(3): 339-362. We will reuse datasets from different Dutch alternation studies, carried out as part of PhD projects or as part of MA theses or as part of individual research within the research unit (QLVL). Exactly which datasets are suitable for the research project at hand will be established.						
Feature matrices	The matrices of features that are deduced from the corpus	New data	Digital	Compiled/ aggregated data	.npy (NumPy binary format)	< 100 MB	

	data.						
Jupyter Notebooks	The Python files used to create the feature matrices.	New data	Digital	Software	.ipynb (IPython Notebook)	< 100 MB	
R scripts	The R scripts used to analyse the data statistically.	New data	Digital	Software	.R (R script source file)	< 100 MB	

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.	Can only be decided after inspection of the datasets (see above)

[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 source code scripts will be commented as much as possible to explain the inner workings of the scripts written for the research workflow.** • **If the scripts are hosted on a collaboration platform such as GitHub, the repository's README file will explain the nature of the scripts and how to use them.**
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: The GitHub README's contents should make the data discoverable with search engines.

4. Data Storage & Back-up during the Research Project	
Where will the data be stored?	• GitHub (researcher's personal account and/or P.I. personal account) • The researcher's laptop • The researcher's personal home server • The QLVL server • OneDrive
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.*	The GitHub repositories provide *version control*, which means that anyone can look at the incremental history of how the data were created. In addition, the GitHub platform will ensure that the data will be available far into the future (they act as a de facto cloud host). We also back up all data by the home institution (KU Leuven) subscription to OneDrive
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: KU Leuven ensures a personal data limit exceeding the current projected size of all data If no, please specify:

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

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 researcher’s GitHub account uses Multi-Factor Authentication (MFA), which means attackers cannot easily break into the account using only a password. In addition, the password for the GitHub account is behind a password manager (which also uses MFA). • The researcher’s laptop is secured using a strong password. • The researcher’s personal server is not directly connected to the internet. • The institution’s OneDrive is also protected
What are the expected costs for data storage and backup during the research project? How will these costs be covered?	There are no expected costs for data storage. GitHub data storage is free. Should more data storage be needed beyond the GitHub provided space, the (free) KU Leuven offerings or CLARIN will be used.

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 data should be able to be preserved for five years or longer.
Where will these data be archived (stored and curated for the long-term)?	All data should remain available on GitHub. In addition, the data will be stored on the university's central servers (with automatic back-up procedures) for at least 10 years, conform the KU Leuven RDM policy.

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?	The expected costs will be determined once more data has been generated for the project. We do not expect the costs to exceed 1000 euros, based on experience with other research projects. The bench fee should be able to cover these 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:
If access is restricted, please specify who will be able to access the data and under what conditions.	
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.	To be decided later

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'.*	To be decided later
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]	To be decided later
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: To be decided later
What are the expected costs for data sharing? How will these costs be covered?	To be decided later

[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?	Anthe Sevenants
Who will manage data storage and backup during the research project?	Anthe Sevenants
Who will manage data preservation and sharing?	Anthe Sevenants
Who will update and implement this DMP?	Anthe Sevenants