Natural language processing in the human brain

A Data Management Plan created using DMPonline.be

Creator: Helena Balabin

Affiliation: KU Leuven (KUL)

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

Template: FWO DMP (Flemish Standard DMP)

Grant number / **URL:** 1154623N

ID: 197328

Start date: 01-11-2022

End date: 30-10-2026

Project abstract:

Natural Language Processing (NLP) is a subfield of artificial intelligence focused on processing and interpreting human language. With the recent success of NLP approaches across a diverse range of application areas, such as machine translation and sentiment analysis, the question arises as to how far the seemingly human-like performances indeed resemble the processing of language in the human brain. Neurological studies of connected speech are central for providing insights into the possible connection between language representations in the human brain and language models used in NLP. By employing state-of-the-art NLP methods, numerical feature representations will be generated for sentences and multi-sentence units. These NLP features will be compared to functional brain connectivity patterns to study their link to human language processing.

Last modified: 06-03-2023

Natural language processing in the human brain
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	
fMRI stimuli	stimuli shown in the planned fMRI experiment	reuse existing data (publicly available data)	digital	aggregated data	.json files with links to .jpg images files and text data, presented as . (psy)exp files	<100GB	N/A
functional magnetic resonance imaging (fMRI) data	functional imaging data collected during the planned fMRI experiment	generate new data	digital	experimental	.dcm (DICOM) and .nii (NIfTI) format for both raw and preprocessed data	<5TB	N/A
informed consent	informed consents from the subjects participating in the fMRI experiment	generate new data	physical	N/A	N/A	N/A	a folder with filled out informed consents
analyses code	code to perform the analyses	generate new data	digital	software	.py, .ipynb, .m python and MATLAB files	<1GB	N/A
F-PACK/ARCK/NLP-FineTuning transcripts	transcripts from interviews conducted with cognitively healthy older adults	reuse existing data (internal lab data)	digital	observational	.txt	<100GB	N/A
F-PACK/ARCK biomarker/neuropsychological data	data containing the amyloid load status and neuropsychological test scores from cognitively healthy older adults	reuse existing data (internal lab data)	digital	observational	.xlsx and .csv	<1GB	N/A

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:

Stimuli data is coming from the Visual Genome dataset (see DOI: 10.1007/s11263-016-0981-7, http://visualgenome.org/)
F-PACK/ARCK/NLP-FineTuning data is coming from prior studies conducted at the laboratory for cognitive neurology (LCN) at KU Leuven (for F-PACK, see DOI: 10.1186/s13195-021-00798-4)

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

The fMRI data (.dcm and .nii files) will be collected in an fMRI experiment consisting of scans from cognitively healthy volunteers, therefore human subject data will be collected in this project.

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

Informed consents for the planned fMRI study containing personal data will be filled out and stored physically, however, personal data is not part of the planned analyses. With regards to the fMRI data itself, all analyses will be performed based on using subject identifiers rather than any personal data from the subjects themselves. The fMRI scans will be defaced/anonymized in later stages of the analyses as well. The planned fMRI experiments will undergo an ethical review procedure at UZ/KU Leuven (mandatory for any fMRI

experiment at UZ/KU Leuven), including a GDPR questionnaire as part of the PRET application and well as a UZ Leuven CTC application.
With regards to the transcripts, interviews are conducted in a way that elicit responses from participants that do not contain sensitive/personal data. Analyses will again be performed using subject identifiers rather than any personal data from the subjects themselves. All procedures regarding the collection of data from the F-PACK/ARCK/NLP-Finetuning cohorts are ethically approved (see study S65221).

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

N/A

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.

- No

N/A

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

N/A

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).

The planned fMRI study will be stored using DICOM files that will be converted to NIfTI stored according to the Brain Imaging Data Structure (BIDS) standard to ensure accompanying metadata, documentation and to follow reproducibility guidelines. All code will be written in python (and MATLAB) using private Git repositories as a version control system (VCS) to ensure reproducibility and clear documentation of the code.
All data from the F-PACK/ARCK/NLP-FineTuning cohorts is collected in established data collection procedures (e.g., using manuals) at LCN.

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

For the planned fMRI study, metadata is available in two formats through the BIDS standard: (1) a dataset_description.json file with metadata regarding the study and the task and (2) a .json file for each fMRI BOLD sequence with all the scanning parameters. For the code, all metadata (file organization etc.) will be listed in the README.md file of the respective Git repository

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

Where will the data be stored?

The data will be stored on external drives from LCN maintained by the IT department of the biomedical group at KU Leuven (GBW ICT). Data is also stored and processed on an internal server at LCN. During the fMRI experiment, an external hard drive is used to retrieve the data from the scanner, then the data is transferred to the LCN drives and deleted from the external hard drive.

How will the data be backed up?

The data and results of the analyses will be regularly backed up on a separate back-up drive, using an established archiving routine at LCN, overseen by Prof. Patrick Dupont. Code is backed up through its Git repository.

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

The storage capacity of the two main external drives and the separate back-up drive used at LCN is 8.9PB in total. Therefore, no storage bottlenecks are expected for this project.

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

The data is going to be stored on the drives and the server at LCN. Both the drives and the server are only accessible to LCN members, and the data collected in this project is only accessible to the members that are directly working with the data. No data will be stored on other devices (laptops etc.), all analyses will be conducted through remote access to the LCN server. During the fMRI experiment, an external hard drive is used to retrieve the data from the scanner, then the data is transferred to the LCN drives and deleted from the external hard drive.

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

The costs regarding data storage, backup and preservation are covered by data storage resources available at LCN, managed by Prof. Rik Vandenberghe and Prof. Patrick Dupont.

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...).

The data collected and results obtained from the analyses of the planned fMRI experiments (.dcm and .nii files) will be retained for at least five years after the end of this research project. Moreover, data from the F-PACK, ARCK and NLP-FineTuning cohorts is also retained for at least five years. Code will also be retained for at least five years through its accompanying private Git repository, which will be made accessible to LCN members.

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

Data is archived for long-term preservation on a separate archiving drive and at LCN (see previous sections).

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

The costs regarding data storage, backup and preservation are covered by data storage resources available at LCN, managed by Prof. Rik Vandenberghe and Prof. Patrick Dupont.

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

The preprocessed (therefore anonymized/defaced) data of the fMRI experiment will be made publicly available through OpenNeuro (https://openneuro.org/). The code for the analyses will be made publicly available in a public GitHub repository hosted by LCN.
The F-PACK/ARCK/NLP-FineTuning data will not be made publicly available.

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

Access to the raw data from the fMRI experiment will only be given to LCN members directly working on this data. Access to the F-PACK/ARCK/NLP-Finetuning data is also only given to LCN members who are directly working on the data.

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
- Yes, Ethical aspects

F-PACK, ARCK and NLP-FineTuning data will not be shared due to ethical aspects and agreements made in the respective ethical review procedures for the data collection in these three cohorts.

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

Processed (defaced/anonymized) fMRI data will be made available through OpenNeuro: https://openneuro.org/
Code will be made available through GitHub: https://github.com/

When will the data be made available?

Preprocessed (defaced/anonymized) fMRI data and code to analyze it will be made available upon publication of the research results of the fMRI experiments.

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

Datasets (the preprocessed/defaced/anonymized fMRI data in this case) hosted on OpenNeuro fall under the Creative Commons CC0 license, and the code hosted on GitHub will be hosted with a MIT license.

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

The preprocessed (defaced/anonymized) fMRI dataset will have a DOI based on its accompanying publication.

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

There are no expected costs for data sharing, since OpenNeuro and GitHub are hosting the necessary storage resources.

6. Responsibilities

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

Helena Balabin

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

Helena Balabin, Patrick Dupont

Who will manage data preservation and sharing?

Helena Balabin, Patrick Dupont, Rik Vandenberghe

Who will update and implement this DMP?

Helena Balabin