DMP title

Project Name FWO junior postdoc fellow DMP - DMP title

Project Identifier 165422/12W5522N

Grant Title 165422/12W5522N

Principal Investigator / **Researcher** Isabeau De Smet

Project Data Contact 0498407513; isabeau.desmet@kuleuven.be

Description This project explores the interplay between semantics and language change through the assessment of the fourth statement of Kurylowicz (K4). The statement goes as follows: "When as a consequence of a morphological change, a form undergoes differentiation, the new form takes over its primary ('basic') function, the old form remains only in secondary ('derived') function." A famous illustration lies in the English word pair 'brethren-brothers'. The new form 'brothers' took over the primary function, while the old form 'brethren' lived on taking a specialized function 'fellow members of a religious order' Yet, so far, the evidence for this law remains mostly anecdotal. The assessment of this statement can shed light not only on this interplay between semantics and language change, but also on broader linguistic phenomena as iconicity, isomorphism, exaptation and prototype theory. I will carry out multiple quantitative case studies on Dutch, with different time scopes (from Middle Dutch until now). I will tackle 4 research questions. 1) When a doublet arises, does the new form inevitably take on the primary function, while the old form remains in the secondary function? 2) Is K4 limited to morphological cases, or is it a more general principle that can be applied to other domains as well? 3) If cases occur where K4 does not apply, can these cases be grouped under a common denominator and can a general explanation be found? 4) What are the cognitive mechanisms behind K4?

Institution KU Leuven

1. General Information

Name applicant

Isabeau De Smet

FWO Project Number & Title

165422/12W5522N

Diachronic semantic tendencies in lexical, morphological and syntactic variation

Affiliation

- KU Leuven

2. Data description

Will you generate/collect new data and/or make use of existing data?

- Generate new data
- Reuse existing data

Describe in detail the origin, type and format of the data (per dataset) and its (estimated) volume. This may be easiest in a table (see example) or as a data flow and per WP or objective of the project. If you reuse existing data, specify the source of these data. Distinguish data types (the kind of content) from data formats (the technical format).

Existing data

Existing data			
Type of data	**Format**	**Volume**	**Source**
Corpus Gysseling (Linguistic corpus with texts from 1300-1400)	.txt	35.5 MB	Downloaded from https://ivdnt.org/
Corpus Middle Dutch (Linguistic corpus with texts from 1400-1500	.xml	18.8 MB	Downloaded from https://ivdnt.org/
Selection of texts from dbnl	.txt	65.2 MB	Downloaded from https://www.dbnl.org/
Corpus Brieven als Buit (Linguistic corpus 17th-19th century)	.xml	63.1 MB	Downloaded from https://ivdnt.org/
Historical dictionaries	online application	online application	https://gtb.ivdnt.org/search/
Catalogue of semantic shifts	online application	online application	https://datsemshift.ru/
Age of acquisition norms (numerical data)	.xlsx	1.9 MB	Downloaded from http://crr.ugent.be/archives/1602

Concreteness norms (numerical data)	.xlsx	1.7 MB	Downloaded from http://crr.ugent.be/archives/1602
Dataset PhD researcher: preterites and past participles from 800-2000	.csv	958 MB	Dataset created during PhD. Available on personal laptop and Box.
Diamant (a diachronic semantic computational lexicon of Dutch, text data)	online application	online application	https://diamant.ivdnt.org/diamant-ui/
New data			
Type of data	**Format**	**Volume**	**How created**
7 morphological case studies (text data)	.csv	7 files of max. 10 MB	Extraction of text data from corpora mentioned above for 7 morphological alternations (through free AntConc software). Enriched with manual annotations and data from 'Concreteness norms' and 'Historical dictionaries' (see above).
7 syntactic case studies (text data)	.csv	7 files of max. 10 MB	Extraction of text data from corpora mentioned above for 7 syntactic alternations (through free AntConc software). Enriched with manual annotations and data from 'Concreteness norms' and 'Historical dictionaries' (see above).
1 Lexical case study (text data)	.csv	1 file of max. 10 MB	Data gathered from DIAMANT (see above), enriched with data from 'Concreteness norms' and 'Age of Acquisition norms' (see above).
Metadata	.txt	1 file of max 1 MB	Notes on data extraction, data annotation.

R scripts (scripts for analysis in R software). One script for each case study + one overarching script. 16 scripts in total	.R	16 files of max. 1 MB	Created in analysis software R.

3. Legal and ethical issues

Will you use personal data? If so, shortly describe the kind of personal data you will use. Add the reference to your file in KU Leuven's Register of Data Processing for Research and Public Service Purposes (PRET application). Be aware that registering the fact that you process personal data is a legal obligation.

- No

Privacy Registry Reference:

Short description of the kind of personal data that will be used:

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, add the reference to the formal approval by the relevant ethical review committee(s)

- No

Does your work possibly result in research data with potential for tech transfer and valorisation? Will IP restrictions be claimed for the data you created? If so, for what data and which restrictions will be asserted?

- No

Do existing 3rd party agreements restrict dissemination or exploitation of the data you (re)use? If so, to what data do they relate and what restrictions are in place?

- No

4. Documentation and metadata

What documentation will be provided to enable reuse of the data collected/generated in this project?

- An excel sheet will be created that specifies the location and names of all the datasets used and created in this project, as well as which files belong to which case

study.

- For each case study, a text file (.txt) will be provided explaining in detail how the data were extracted and annotated.

- All steps of the analyses will be annotated in the R scripts.

Will a metadata standard be used? If so, describe in detail which standard will be used. If no, state in detail which metadata will be created to make the data easy/easier to find and reuse.

- No

No relevant metadata standards exist for this project. We will use the standard terminology in historical and variational linguistics in our descriptions in the metadata.

- An excel sheet will be created that specifies the location and names of all the datasets used and created in this project, as well as which files belong to which case study.

- For each case study, a text file (.txt) will be provided explaining in detail how the data were extracted and annotated.

- All steps of the analyses will be annotated in the R scripts.

5. Data storage and backup during the FWO project

Where will the data be stored?

The data will be stored on the encrypted hard drive of the researcher's KU Leuven laptop, as well as on the cloud service OneDrive.

How is backup of the data provided?

The data will be stored on OneDrive with automatic daily back-up procedures.

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

We estimate we will need max. 2000 MB in capacity. The personal computer of the researcher has a storage capacity of 512 GB which is more than enough.

For every KU Leuven staff member, a OneDrive capacity of 2TB is provided, which is again more than enough for both storage of the data and backups.

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

The needed storage and back up capacity is free. No additional costs are expected.

Data security: how will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

We will not be working with sensitive personal data or other special categories that require extra security procedures. Data are stored on the encrypted hard drive of the researcher, as well as on OneDrive (which has a password protection).

6. Data preservation after the FWO project

Which data will be retained for the expected 5 year period after the end of the project? In case only a selection of the data can/will be preserved, clearly state the reasons for this (legal or contractual restrictions, physical preservation issues, ...).

All data will be retained for a period of minimally 10 years.

Where will the data be archived (= stored for the longer term)?

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 retention period of 5 years? How will the costs be covered?

The cost of storage at the KU Leuven servers lies at €54 for 100 GB. In view of the expected size of the database (< 100 GB), estimated cost will be max. €540 for ten year storage. This will be paid for by the FWO bench fee.

7. Data sharing and reuse

Are there any factors restricting or preventing the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)?

- No

Which data will be made available after the end of the project?

All data will be made available to fellow researchers upon request by email.

Where/how will the data be made available for reuse?

- Upon request by mail

When will the data be made available?

- Upon publication of the research results

Who will be able to access the data and under what conditions?

All data will be made available to fellow researchers upon request.

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

We do not expect any costs for data sharing.

8. Responsibilities

Who will be responsible for data documentation & metadata?

Isabeau De Smet

Who will be responsible for data storage & back up during the project?
Isabeau De Smet

Who will be responsible for ensuring data preservation and reuse ?
Isabeau De Smet & the PI of the project Freek Van de Velde

Who bears the end responsibility for updating & implementing this DMP?
Isabeau De Smet