# It's all frequency? - testing usage-based theories of language change using agent-based models

*A Data Management Plan created using DMPonline.be*

**Creator:** Anthe Sevenants

**Affiliation:** KU Leuven (KUL)

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

**Template:** FWO DMP (Flemish Standard DMP)

**Grant number / URL:** 11P8324N

**ID:** 205621

**Start date:** 01-11-2023

**End date:** 31-10-2027

**Project abstract:**

Usage-based theory of language has been conducive to some highly impactful explanatory mechanisms of language change, such as the reducing effect, syntactic priming effect and analogy, yet critics suggest that these mechanisms are often deployed on an ‘ad-hoc’ basis. The proposed project emanates from the idea that concrete implementations of such theories could take a more central place in usage-based linguistics, and serve as a means to test and refine usage-based models of the theories mentioned above. The overarching goal of this project is to use agent-based computer simulations to implement the usage-based frequency mechanisms mentioned above in a computational simulation, such that more easily falsifiable hypotheses can be inferred. By testing what minimal conditions are necessary for the three general effects mentioned above to occur in the simulation models, I will be able to fill in their blank spots and assess their potential for language change on the diachronic level. Through these investigations, I will also demonstrate how useful computer simulations can be in this context and exemplify how they can be used to solidify (or undermine) theoretical claims.

**Last modified:** 18-03-2024

# It's all frequency? - testing usage-based theories of language change using agent-based models

## 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:*<br>- Generate new data<br>- Reuse existing data | *Please choose from the following options:*<br>- Digital<br>- Physical | *Please choose from the following options:*<br>- Observational<br>- Experimental<br>- Compiled/aggregated data<br>- Simulation data<br>- Software<br>- Other<br>- NA | *Please choose from the following options:*<br>- .por, .xml, .tab, .csv,.pdf, .txt, .rtf, .dwg, .gml, …<br>- NA | *Please choose from the following options:*<br>- <100MB<br>- <1GB<br>- <100GB<br>- <1TB<br>- <5TB<br>- <10TB<br>- <50TB<br>- >50TB<br>- NA | |
| Corpora and datasets | We might use certain corpora to base frequency information on. Examples could be the SoNaR Corpus (Oostdijk et al. 2013) or the Frown corpus (Mair 1999), however this is currently undecided. | Reuse existing data | Digital | Observational | .txt<br>.xml | < 1 TB | |
| Python scripts | Software libraries or other code run in regular Python. These will be published on GitHub with a full version history. | Generate new data | Digital | Software | .py | < 100 MB | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Jupyter notebooks | Software experiments run in interactive Python will be saved as Jupyter notebooks. These will be published on GitHub with a full version history. | Generate new data | Digital | Software | .ipynb | < 100 MB | |
| R scripts | The R scripts used to analyse the data statistically. These will be published on GitHub with a full version history. | Generate new data | Digital | Software | .R | < 100 MB | |
| NetLogo scripts | NetLogo is a mature platform for running computer simulations of emergent behaviour. I do not intend to use NetLogo for my own simulations, but I do want to use it to practice building simulations. These 'practice' simulations will also be available on GitHub. | Generate new data | Digital | Software | .nlogo | < 100 MB | |
| Docker containers | I run all my software under Docker in order to ensure a fully reproducible environment. I will provide both the Dockerfiles, the Docker-compose manifests and built containers publicly. This ensures *everything* I run software-wise can be reproduced in the future by *anyone*. | Generate new data | Digital | Software | Dockerfile<br>docker-compose.yml | < 100 MB | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Literature summaries | I tend to summarise every book, article … I read for my literature review. These summarisations are written in a (custom) Markdown format. All my notes are available publicly on GitHub, both raw *and* rendered (automatically through GitHub Actions). | Generate new data | Digital | Compiled data | .md | < 100 MB | |

**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:**

(Will be decided in the future, if any)

**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.**

- No

**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.**

- No

**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

**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

**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

### 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).**

- 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.
- I run all my software under Docker in order to ensure a fully reproducible environment. I will provide both the Dockerfiles, the Docker-compose manifests and built containers publicly. This ensures *everything* I run software-wise can be reproduced in the future by *anyone*.

**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

Whenever I create a dataset (for whatever purpose), I just upload it online to Zenodo, which handles the metadata standard for me.

### 3. 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 QLVL server ("volt", in the CCL network)
- OneDrive

**How will the data be backed up?**

- 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

KU Leuven ensures a personal data limit exceeding the current projected size of all data.

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

- 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 *and* Windows BitLocker.
- 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.

### 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...).**

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.

### 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.**

/

**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.**

- No

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

- Datasets will very likely go to Zenodo
- Software is continually pushed to GitHub

**When will the data be made available?**

- Software is available immediately (whenever it is pushed).
- Dataset availability must be decided at a later date. There is no real system to it (likely upon publication of research results).

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

- I tend to use AGPL v3 for everything.

**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?**

There are no expected costs for data sharing. Should there be costs anyway, the bench fee should be able to cover them.

### 6. 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