Bystanders no more: simple forms and the expression of aspect in the history of English and beyond

A Data Management Plan created using DMPonline.be

Creator: Juliette Kayenbergh

Affiliation: KU Leuven (KUL)

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

Template: FWO DMP (Flemish Standard DMP)

ID: 206260

Last modified: 23-04-2024

Bystanders no more: simple forms and the expression of aspect in the history of English and beyond
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, .csv,.pdf, .txt, .rtf, .dwg, .gml, … • NA	*Please choose from the following options:* • <100MB • <1GB • <100GB • <1TB • <5TB • <10TB • <50TB • >50TB • NA	
available text corpora	Text corpora publicly available (BNC, EEBO, CLMET)	reuse existing data	digital	compiled data	.txt .xml	<100GB	
newly compiled corpora	Dickens corpus (corpus made of books by Dickens and of aligned translations in Dutch, Afrikaans, and German); Bible Corpus (aligned corpus of Bibles in Old English, Middle English, Modern English and Present-Day English)	reuse existing data	digital	compiled data	.txt	<100GB	

supervisor's corpora	Germanic Bibles (aligned corpus of Bible in German, Afrikaans, PDE, and Dutch); Don Quixote Corpus (aligned parallel corpus of Cervantes' Don Quijote and 8 successive English translations from Modern English to Present-day English)	reuse existing data	digital	compiled data	.txt	<100GB	
PERL scripts	scripts written by my supervisor Hendrik De Smet to create concordances from the corpora	reuse existing data	digital	software	.pl	<100MB	
concordances	list of hits generated by the query in the PERL scripts	generate new data	digital	compiled data	.txt	<100MB	
annotated data samples	concordances are pasted in an Excel table and relevant tokens are analyzed	generate new data	digital	aggregated data	.xlsx	<100MB	
data protocols	Word documents explaining the coding annotations of the Excel tables	generate new data	digital	other	.docx	<100MB	

observations/notes about the datasets	additional observations while annotating, thoughts, notes of the discussions with supervisors about the data, question lists in Word documents	generate new data	digital	observational	.docx	<100MB	
R scripts	scripts to create graphs and statistical analyses of the annotated data samples in R	generate new data	digital	software	.R	<100MB	
data visualization	preliminary data visualization of the data in graphs generated in Excel	generate new data	digital	aggregated data	.xlsx	<100MB	
R graphs	graphs generated by the R scripts	generate new data	digital	aggregated data	.pdf/.png/.jpg	<1GB	
annotated literature	when digital articles, papers, books, annotations directly in Adobe Reader or Zotero	generate new data	digital	observational	.pdf	<1GB	
notes about the literature	when printed literature, notes in a Word document	generate new data	digital	compiled data/observational	.docx	<100MB	
bibliographic information	list of all the literature in a Zotero database	generate new data	digital	compiled data	NA	NA	

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:

EEBO: www.proquest.com/eebo/
CLMET: https://perswww.kuleuven.be/~u0044428/clmet.htm
BNC: http://www.natcorp.ox.ac.uk/

+ text corpora provided by my supervisor Hendrik De Smet
+ PERL scripts provided by my supervisor Hendrik De Smet
+ texts for the Bible corpus provided by my supervisor

+ Translations of Dickens in different languages -> some sources still to be determined

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.

- Yes

Certain texts used to compile the corpora are not of the public domain (e.g. Dickens corpus, Don Quijote corpus), so the annotated datasets coming from these corpora will not be made openly accessible at the end of the project (for copyright reasons). The PERL scripts of my supervisor won't be made available either (intellectual property).

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 concordances (.txt files) contain the query that was typed into the PERL scripts to gather data in the text corpora, so I will always know which query was used to retrieve data in the corpora.

For the annotation and coding of the text corpora, I write data protocols in Word documents with each coding category, each coding decision, each color code thoroughly explained and accompanied by an example from the data.

My additional notes, observations on the data, as well as my annotations about the literature (either in the pdfs directly, or in Word documents), are quite straightforwardly understandable.

The scripts will contain clear comments to explain the code.

I will insert README files when I share my data in data repositories, to explain where the reader will find each type of information they need.

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

DataCite will be used because my data will be shared on the RDR data repository. Each published dataset will also receive a DOI to be easier to find and reuse.
(internally, each concordance of the text corpora generated by a Perl-script includes the date/period; author, text type, title of the work the hit comes from. Moreover, all the metadata about the (digital and physical) literature relevant to the project is stored in my Zotero database).

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

Where will the data be stored?

During the project, all the data is stored both on my work computer and on my KU Leuven One Drive for Business account (synchronized).

How will the data be backed up?

During the project, the data is automatically backed up on my KU Leuven One Drive for Business account.

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 KU Leuven Onedrive for Business cloud service provides 2TB storage (thus currently there is enough space to store the data I listed), and I can request up to 5TB if needed.

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

I have purchased a laptop though the ICT service, it has Bitlocker installed and I have a good password (there are thus two steps to unlock my computer). Additionally, One Drive is safe and the online version can only be accessed by logging in through the KU Leuven system, which has to be done with KU Leuven Authenticator (and requires to have my own phone).

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

We do not expect there will be costs for data storage and backup during the research project (and if need be, my bench fee will cover the costs).

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 the data will be preserved for 10 years (according to KU Leuven RDM policy). There is no sensitive data which cannot be preserved, and the storage capacity is no issue.

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

All the data will be preserved on the KU Leuven data repository RDR (with access rights varying according to the copyright issues and the relevance of the data for an open access)

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

We do not expect to pay any cost for the data preservation. RDR is free (and if really needed, my bench fee can cover the fees).

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 annotated datasets of the corpora made of texts available in the public domain, the data protocols, my R scripts and the various visualizations and graphs will be made available in Open Access (but not the annotated datasets of the corpora made of texts not publicly available, and not my supervisor's scripts).
However my personal notes about the literature or my thoughts and observations about the data are not relevant for sharing. For the bibliographic information, that still needs to be decided.

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

Again, there will be no access to the annotated samples of corpora made of texts that are not publicly available and to my supervisor's Perl scripts.

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, Other

The annotated samples of corpora made of texts that are not publicly available will not be shared (for copyright reasons) and my supervisor's Perl scripts won't either.

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

I will use RDR for all the data.

When will the data be made available?

The annotated datasets, the corresponding data protocols, R scripts and visualizations will be made available upon acceptance of the publications.

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

I will use the CC BY-NC-SA 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

Each item published on RDR will be identified by a DOI.

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

We do not expect any costs for data sharing (if need be, my bench fee will cover the costs).

6. Responsibilities

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

Juliette Kayenbergh

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

Juliette Kayenbergh

Who will manage data preservation and sharing?

Hendrik De Smet

Who will update and implement this DMP?

Juliette Kayenbergh