Journal of Open Humanities Data

"d-Prose 1870–1920" a Collection of German Prose Texts from 1870 to 1920

DATA PAPER

EVELYN GIUS
SVENJA GUHR
INNA UGLANOVA
Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

The corpus "d-Prose 1870-1920" is a heterogeneous collection of 2511 German-language literary prose texts published between 1870 and 1920. It contains narrative texts from trivial and high literature with a minimum length of 1000 words. The texts are available as UTF-8 text files with an associated metadata table. These data are primarily of interest to literary scholars, linguists, but also for all those who are involved in the modelling of textual and cultural phenomena. The corpus can be used to test literary, literary-historical, cultural, and linguistic hypotheses.

CORRESPONDING AUTHOR:
Evelyn Gius
Technical University of Darmstadt, Darmstadt, Germany
evelyn.gius@tu-darmstadt.de

KEYWORDS:
prose; 19th century; German literature; text corpus; text collection; narrative texts

TO CITE THIS ARTICLE:
Gius, E., Guhr, S., & Uglanova, I. (2021). "d-Prose 1870–1920" a Collection of German Prose Texts from 1870 to 1920. *Journal of Open Humanities Data*, 7: 11, pp. 1–5. DOI: *https://doi.org/10.5334/johd.30*

1 OVERVIEW

REPOSITORY LOCATION

10.5281/zenodo.4315208

CONTEXT

"d-Prose 1870–1920" was created as part of the project "Gender and Illness" in the cooperation project "Automated modelling of hermeneutic processes – The use of annotation in social research and the humanities for analyses on health".[1] The focus of the project was on the description of illness from a gender perspective. The research focused on how the presentation, perception, and dealing with illness differ among characters depending on their gender.

2 METHOD

STEPS

The text files were taken from the KOLIMO corpus (Herrmann & Lauer, 2017), which in turn is based on the repositories TextGrid,[2] Deutsches Textarchiv,[3] and Gutenberg-de.[4] The KOLIMOtoText tool (Adelmann, 2020d) served as the basis for the corpus creation, extracting all texts in the KOLIMO corpus published between 1870 and 1920. After this, a manual check was performed in order to exclude all non-prose texts as well as all texts that have not been originally published in German (i.e., translations). Since the received text collection contained duplicate texts due to the occurrence of different editions with different titles, duplicate texts were removed in a further step using the author-name-title comparison program ANTComp (Adelmann, 2020a) and the full-text comparison program BatchSED (Adelmann, 2020b). All texts were manually cleaned of paratexts (i.e., author names, dedications, prefaces, etc.) and supplemented by metadata: author's name and pseudonyms, author's gender, author's date of birth and death, title of the work, repository source, file name, number of words and types, as well as publication date (extracted from the metadata of the original repositories, checked and if necessary corrected or extended by data from the literary encyclopedias Killy (Kühlmann et al., 2016) and Kindler (Arnold, 2009)).

SAMPLING STRATEGY

In order to cover a broad variety of phenomena in literary prose texts, a certain degree of heterogeneity in form and content was aimed for in the sampling of the corpus. Beyond that, there were no content-related restrictions of the text selection. Thus, criteria for text selection were only date of first publication, text language, genre, and text length (for a more detailed discussion of sampling criteria, see Gius, Krüger, & Sökefeld, 2019). As a result, the corpus includes the works of 334 authors from at least three different literary movements (i.e., naturalism, realism, and modernism). It contains approximately equal proportions of long texts (novels) as well as shorter prose forms (cf. ***Table 1***).

QUALITY CONTROL

The steps of automatic processing described above (Adelmann, 2020a, 2020b, 2020d) were evaluated by manual control of random samples of the output. This lead to an iterative improvement of the results. For the de-duplication process (Adelmann, 2020b) a manual evaluation of all duplicate pairs identified by the process (i.e., about 1,000 text pairs) was performed (Adelmann & Gius, 2020). After the automated cleaning, a manual cleaning was performed. All texts were manually cleaned of paratexts such as author name, author biography, dedication, preface, remarks, etc. in a collaborative approach. In this approach, the data curators worked in a review process so that each text was double-checked. The same procedure was used for enrichment with metadata. A first data curator added the respective meta information and a second data curator reviewed the entered information.

1 hermA Project. Available from *https://www.herma.uni-hamburg.de/en.html*. Accessed 2021-02-24.

2 Textgrid Repository. Available from *https://textgridrep.org/*. Accessed: 2021-02-24.

3 Deutsches Textarchiv. Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache. Available from *http://www.deutschestextarchiv.de/*. Accessed: 2021-02-24.

4 Project Gutenberg. Available from *https://www.projekt-gutenberg.org*. Accessed: 2021-02-24.

number of texts	2511	
number of texts written by female authors	346	
number of texts written by male authors	2165	
number of authors	334	
number of female authors	72	
number of male authors	262	
text size average	31146 words; 4753 types	
standard deviation	58117 words; 5323 types	
shortest text	1006 words	
longest text	990351 words	
number of texts per decade	1870–1879	226
	1880–1889	327
	1890–1899	542
	1900–1909	623
	1910–1920	793

Table 1 The average text size (in words and types) was calculated by Voyant Tools (Sinclair & Rockwell, 2021). The statistical data refer to the content of version 2.0 of the corpus "d-Prose 1870–1920".

3 DATASET DESCRIPTION

OBJECT NAME

d-Prose 1870-1920

FORMAT NAMES AND VERSIONS

plain txt-files (UTF-8); spread sheet (xslx) with metadata. Version 2.0

CREATION DATES

2017-05-01 – 2021-06-22

DATASET CREATORS

Adelmann, Benedikt (Developer, University of Hamburg); Gius, Evelyn (Conceptualization, Project administrator, Supervisor, Technical University of Darmstadt); Guhr, Svenja (Data curator, Supervisor, Technical University of Darmstadt); Kurz, Laura (Data curator, Technical University of Darmstadt); Otte, Felicitas (Data curator, University of Hamburg); Schlesiger, Nicole (Data curator, Technical University of Darmstadt); Schreiber, Annekea (Data curator, Technical University of Darmstadt); Sökefeld, Carla (Data curator, University of Hamburg); Krüger, Katharina (Project member, University of Hamburg); Murawska, Anna Aline (Project member, University of Hamburg); Uglanova, Inna (Validation, Application, Technical University of Darmstadt).

LANGUAGE

German

LICENSE

Creative Commons Attribution Non Commercial Share Alike 4.0 International

REPOSITORY NAME

Zenodo.org

PUBLICATION DATE

2020-12-15 (2021-06-22 (V.2.0))

4 REUSE POTENTIAL

"d-Prose 1870–1920" is primarily of interest to literary scholars and linguists, as well as to scholars involved in the modelling of textual and cultural phenomena. The corpus can be used to test literary, literary-historical, cultural and linguistic hypotheses. It may be especially helpful for the study of the literary developments and artistic movements of the represented period of time. The corpus has sufficient volume to allow for the application of machine learning techniques like clustering, classification, or topic modelling (see Uglanova & Gius, 2020, for an example based on topic modelling). It can be used as analysis material for didactic purposes too, such as developing core knowledge and skills for working with a literary corpus (studying patterns of usage, historical dynamics, typical and unique contexts of language or literary phenomena, author's neologisms, collocates across metadata, etc.). The dataset can be extended and included as a sub-corpus for more global tasks. Due to the fact that the corpus is heterogeneous in its structure, it can be divided into smaller subcorpora tailored to specific research objectives. In particular, subcorpora can be created on the basis of publication year, gender of the author, size (short stories, novellas, novels), or, with some additional work on/with the metadata by genre (historical novel, adventure novel, social novel, etc.), literary movements (realism, modernism, naturalism), or other criteria. For more special tasks, the corpus can be pre-processed with a linguistic analysis pipeline, specifically developed for this dataset by Adelmann (2020c). The pipeline is optimized for the corpus and provides tokenization, part-of-speech and morphological tagging, lemmatization, and dependency parsing. It is accompanied by detailed instructions that can easily be used by users with minimal technical knowledge. The data format makes it easy to use linguistic, web-based technology without further extraction and conversion. Additionally it can be used with popular GUI-based open-source software for data mining and analysis such as Voyant Tools (Sinclair & Rockwell, 2021), the manual annotation tool CATMA (Gius et al., 2021) the concordance program AntConc (Anthony, 2020), and others.

FUNDING STATEMENT

Work on this corpus was funded by the Hamburg Research Foundation (Landesforschungsförderung Hamburg) with grant number LFF-FV 35.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Evelyn Gius: Conceptualization, Funding acquisition, Project administration, Supervision, Validation, Writing; Svenja Guhr: Data Curation, Supervision of Data Curation, Writing; Inna Uglanova: Validation, Application, Writing.

AUTHOR AFFILIATIONS

Evelyn Gius *orcid.org/0000-0001-8888-8419*
Technical University of Darmstadt, Darmstadt, Germany
Svenja Guhr *orcid.org/0000-0002-7686-3609*
Technical University of Darmstadt, Darmstadt, Germany
Inna Uglanova *orcid.org/0000-0002-8092-3512*
Technical University of Darmstadt, Darmstadt, Germany

REFERENCES

Adelmann, B. (2020a). *ANTComp (author-name-title-comparer). An implementation of a heuristic approach to identifying possible duplicates in a corpus of literary works.* https://github.com/benadelm/ANTComp (Accessed: 2021-02-24).

Adelmann, B. (2020b). *BatchSED (batch-substring-edit-distance). A tool for batch computing substring editing distances between the full texts of novels for duplicate detection.* https://github.com/benadelm/BatchSED (Accessed: 2021-02-24).

Adelmann, B. (2020c). *hermA-Pipeline. The linguistic processing pipeline for German.* https://github.com/benadelm/hermA-Pipeline (Accessed: 2021-02-24).

Adelmann, B. (2020d). *KOLIMOtoText. A tool for extracting the document text of TEI and XHTML files in the KOLIMO corpus.* https://github.com/benadelm/KOLIMOtoText (Accessed: 2021-02-24).

Adelmann, B., & **Gius, E.** (2020). Korpusbereinigung für größere Textmengen. Eine (kurze) Problematisierung und ein Lösungsansatz für Duplikate. In C. Schöch (Ed.), *DHd 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation. Konferenzabstracts* (pp. 331–334). Paderborn: Universität Paderborn. DOI: https://doi.org/10.5281/zenodo.3666690

Anthony, L. (2020). *AntConc. Version 3.5.9.* https://www.laurenceanthony.net/software (Accessed: 2021-02-24).

Arnold, H. (2009). *Kindlers Literatur-Lexikon.* Stuttgart: Metzler.

Gius, E., Krüger, K., & **Sökefeld, C.** (2019). Korpuserstellung als literaturwissenschaftliche Aufgabe. In *DHd 2019 Digital Humanities: multimedial & multimodal Konferenzabstracts* (pp. 164–166). Frankfurt & Mainz.

Gius, E., Meister, J., Petris, M., Meister, M., Bruck, C., Jacke, J., & **Horstmann, J.** (2021). *CATMA 6.1.* https://catma.de/. (Accessed: 2021-02-24) DOI: https://doi.org/10.5281/zenodo.1470118

Herrmann, B., & **Lauer, G.** (2017). *KOLIMO - A corpus of literary modernism for comparative analysis.* https://kolimo.uni-goettingen.de/about (Accessed: 2020-05-24).

Kühlmann, W., Aurnhammer, A., Egyptien, J., Kellermann, K., Kiesel, H., Martus, S., & **Sdzuj, R.** (2016). *Killy Literaturlexikon. Autoren und Werke des deutschsprachigen Kulturraums.* Darmstadt: WBG.

Sinclair, S., & **Rockwell, G.** (2021). *Voyant Tools.* http://voyant-tools.org/ (Accessed: 2021-02-24).

Uglanova, I., & **Gius, E.** (2020). The order of things. A study on topic modelling of literary texts. In F. Karsdorp, B. McGillivray, A. Nerghes, & M. Wevers (Eds.), *Proceedings of the Workshop on Computational Humanities Research (CHR 2020)* (pp. 57–76). Amsterdam, the Netherlands: CEUR Workshop Proceedings. http://ceur-ws.org/Vol-2723/long7.pdf (Accessed: 2021-02-24).

TO CITE THIS ARTICLE:
Gius, E., Guhr, S., & Uglanova, I. (2021). "d-Prose 1870–1920" a Collection of German Prose Texts from 1870 to 1920. *Journal of Open Humanities Data*, 7: 11, pp. 1–5. DOI: https://doi.org/10.5334/johd.30

Published: 08 July 2021

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.