Journal of Open Humanities Data

Event Annotations of Prose

DATA PAPER

MICHAEL VAUTH
EVELYN GIUS
*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

This dataset covers 41,341 manual event annotations of six German prose texts from the 19th and early 20th century comprising 290,997 tokens. For each text, the dataset includes annotations by two annotators and gold standard annotations. These annotations were used for the automation of narratological event annotations (Vauth, Hatzel, Gius, & Biemann, 2021), a reflection of inter annotator agreements in literary studies (Gius & Vauth, 2022) and the development of an event based plot model (Gius & Vauth, accepted).

CORRESPONDING AUTHOR:
Michael Vauth
Institut für Sprach- und Literaturwissenschaft, Technical University Darmstadt, Germany
michael.vauth@gmx.de

KEYWORDS:
annotation; narratology; computational literary studies; event; prose

TO CITE THIS ARTICLE:
Vauth, M., & Gius, E. (2022). Event Annotations of Prose. *Journal of Open Humanities Data*, 8: 19, pp. 1–6. DOI: https://doi.org/10.5334/johd.83

(1) OVERVIEW

REPOSITORY LOCATION

Our dataset is located in a Github repository within the forTEXT organisation: https://github.com/forTEXT/EvENT_Dataset. Additionaly, this repository is published as a Zenodo dataset (Vauth & Gius, 2022).

CONTEXT

The annotations were produced as part of the research project EvENT, located at the Technical University Darmstadt and the University of Hamburg. The EvENT Project is part of the priority programme *Computational Literary Studies* (CLS), funded by the German Research Foundation (DFG). For further informations see the programme website: https://dfg-spp-cls.github.io/. We developed an event annotation tagset that is based on narrative theory, where events are considered the smallest units of narratives (Vauth & Gius, 2022). The event tagset has been used for annotating the texts, assigning to each subclause one of the four categories (non-event, stative event, process event and change of state). Depending on the event types, additional properties have been assigned.

(2) METHOD

The dataset is created by manual annotation using the CATMA tool (Gius et al., 2022) for the manual annotations and the GitMA package (Vauth et al., 2022) for annotation data processing.

STEPS

The annotation procedure includes the following steps:

- **Corpus collection:** The six texts are collected from the Textgrid Corpus (Textgrid, 2021) and the d-Prose corpus (Gius, Guhr, & Adelmann, 2021). We selected narratives representing the literary developments between 1800 and 1920. In order to represent the most common narrative genres of this time period, we included short stories, novellas and novels. The corpus consists of:
  - Ludwig Tieck (1797): *Der blonde Eckbert*
  - Heinrich von Kleist (1807): *Das Erdbeben in Chili*
  - Annette von Droste Huelshoff (1842): *Die Judenbuche*
  - Theodor Fontane (1894): *Effi Briest*
  - Marie von Ebner-Eschenbach (1896): *Krambambuli*
  - Franz Kafka (1915): *Die Verwandlung*
- **Annotation Guidelines:** We developed guidelines for the annotation of narratological event types (Vauth & Gius, 2021).
- **Manual Annotation Process:**
  - **Pilot annotations:** The annotation guidelines were developed and improved by extensive pilot annotations.
  - **Annotator training:** Annotators were first trained by annotating and discussing a training text.
  - **Systematic annotations:** Every text has been annotated by two independent annotators (see Table 1). The annotation process was accompanied by regular meetings to discuss cases of doubt. For the documentation of these cases, the annotators used a dedicated tag.
  - **Gold standard annotations:** Based on the double annotations of every text, gold standard annotations were created by one annotator who resolved inconsistent annotations (Table 3). Here again, cases of doubt were discussed. In this process, the GitMA package (Vauth et al., 2022) was developed for supporting the extraction, comparison and integration of annotations in CATMA.

QUALITY CONTROL

The multi annotator approach with comprehensive training of annotators and the feedback loops described above were designed for controlling the quality of manual annotations. The main annotation task was the classification of the event types based on four categories

- non_event
- stative_event
- process
- change_of_state.

Here, we accomplished an agreement greater than 0.55 Krippendorff's α for the six texts. The evaluation results of inter annotator agreement (IAA) for the final annotations is documented in Table 1.

	ECKBERT	EFFI BRIEST	ERDBEBEN	JUDENBUCHE	KRAMBAMBULI	VERWANDLUNG
event type	0.73	0.57	0.75	0.61	0.66	0.73

Table 1 Inter Annotator Agreement (Krippendorff's α) for event types.

	ECKBERT	EFFI BRIEST	ERDBEBEN	JUDENBUCHE	KRAMBAMBULI	VERWANDLUNG
unpredictable	–0.25	–0.30	–0.08	–0.35	–0.21	–0.55
mental	0.79	0.33	0.58	0.39	0.46	0.79
representation_type	0.94	0.87	0.86	0.91	0.86	0.67
persistent	0.09	0.13	0.28	–0.14	0.25	–0.89
iterative	0.62	0.20	–0.29	0.35	0.07	0.70
intentional	0.75	0.24	0.45	0.43	0.32	0.70
non_event_type	0.66	0.68	0.80	0.71	0.80	0.69

Table 2 Inter Annotator Agreement (Krippendorff's α) for additional event properties. For a detailed description and examples see Vauth and Gius (2021).

Table 2 shows additional event classifications that are also grounded in narrative theory and depend on the event type classification. These categories are implemented as properties for defined event types. For instance, only process events and changes of state can be iterative. As the lower IAA values for some categories indicate, some of these categories are highly interpretative. The strongly varying agreement values are also due to the fact that different classification systems are provided for these event properties:

- **unpredictable**: 0, 1, 2, 3, 4
- **mental**: yes, no
- **representation_type**: (any combination of) narrator_speech, character_speech, thought_representation
- **persistent**: 0, 1, 2, 3, 4
- **iterative**: yes, no
- **intentional**: yes, no
- **non_event_type**: conditional_sentence, subjunctive_sentence, modalised_statement, negation, generic_sentence, ellipsis, imperative_sentence, question, request

(3) DATASET DESCRIPTION

OBJECT NAME

Annotations_EvENT.json

FORMAT NAMES AND VERSIONS

JSON

Table 3 Number and extension (in tokens) of gold standard annotations per text. For tokenization we used the German tokenizer in the NLTK toolkit version 3.7 (Bird et al., 2009).

	ERDBEBEN		VERWANDLUNG		ECKBERT		KRAMBAMBULI		JUDENBUCHE		EFFI BRIEST		ALL TEXT	
	COUNT	TOKEN	COUNT	TOKEN	COUNT	TOKEN	COUNT	TOKEN	COUNT	TOKEN	COUNT	TOKEN	COUNT	TOKEN
non_event	167	1,086	757	5,938	212	1,488	116	712	856	4,732	2,887	16,655	**4,995**	**30,611**
stative_event	136	1,046	455	3,830	243	1,667	82	637	476	3,502	1,675	11,656	**3,067**	**22,338**
process	400	3,459	1,126	9,748	450	3,225	268	1,990	1,120	8,146	2,061	15,180	**5,425**	**41,748**
change_of_state	9	63	26	216	25	163	4	39	39	324	43	362	**146**	**1,167**

CREATION DATES

2020-12-01 – 2022-03-31

DATASET CREATORS

Evelyn Gius, Michael Vauth, Michael Weiland (student assistant), Gina Maria Sachse (student assistant), Angela Nöll (student assistant) (all contributors are affiliated to Technical University Darmstadt).

LANGUAGE

German (texts) and English (annotation categories)

LICENSE

GPL-3.0 License.

REPOSITORY NAME

EvENT_Dataset

PUBLICATION DATE

2022-04-01

(4) REUSE POTENTIAL

The dataset is reusable for several natural language processing (NLP) tasks focused on the detection of events. Based on the manual annotations in the dataset we accomplished the automation of narratological event type recognition (Vauth et al., 2021). In general, the event annotations can be used as features for the detection of phenomena related to narrative text structures.

Furthermore, based on the event annotations we developed and evaluated an approach to model the narrativeness/eventfulness and to identify the most 'tellable' parts in a narrative (Gius & Vauth, accepted). In a next step, the modelling of narrativity will be used in text comparisons.

FUNDING STATEMENT

The EvENT project is funded by the German Research Foundation (DFG) within the priority programme SPP 2207 Computational Literary Studies (CLS).

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Evelyn Gius: conceptualization, project administration, funding acquisition, supervision, writing – review and editing; Michael Vauth: conceptualization, data curation, project administration, writing – original draft.

AUTHOR AFFILIATIONS

Michael Vauth orcid.org/0000-0002-3668-6273
Institut für Sprach- und Literaturwissenschaft, Technical University Darmstadt, Germany
Evelyn Gius orcid.org/0000-0001-8888-8419
Institut für Sprach- und Literaturwissenschaft, Technical University Darmstadt, Germany

REFERENCES

Bird, S., Klein, E., & **Loper, E.** (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Gius, E., Guhr, S., & **Adelmann, B.** (2021, June). d-prose 1870–1920. *Zenodo*. DOI: https://doi.org/10.5281/zenodo.5015008

Gius, E., Meister, J. C., Meister, M., Petris, M., Bruck, C., Jacke, J., Schumacher, M., Gerstorfer, D., Flüh, M., & **Horstmann, J.** (2022, January). Catma. *Zenodo*. DOI: https://doi.org/10.5281/zenodo.6046763

Gius, E., & **Vauth, M.** (2022). *Inter Annotator Agreement und Intersubjektivität – Ein Vorschlag zur Messbarkeit der Qualität literaturwissenschaftlicher Annotationen.* DHd 2022 Kulturen des digitalen Gedächtnisses. 8. Tagung des Verbands "Digital Humanities im deutschsprachigen Raum" (DHd 2022), Potsdam. DOI: https://doi.org/10.5281/zenodo.6328209

Gius, E., & **Vauth, M.** (accepted). Towards an Event Based Plot Model. A Computational Narratology Approach. *Journal of Computational Literary Studies.*

TextGrid. (2021). *Die digitale bibliothek bei textgrid*. Retrieved 2021-11-10, from https://textgrid.de/de/digitale-bibliothek

Vauth, M., & **Gius, E.** (2021, July). Richtlinien für die Annotation narratologischer Ereigniskonzepte. *Zenodo*. DOI: https://doi.org/10.5281/zenodo.5078175

Vauth, M., & **Gius, E.** (2022, April). fortext/event dataset: v.1.0. *Zenodo*. DOI: https://doi.org/10.5281/zenodo.6406569

Vauth, M., Hatzel, H. O., Gius, E., & **Biemann, C.** (2021). Automated Event Annotation in Literary Texts. *Computational Humanities Research*, 333–345. Retrieved from http://ceur-ws.org/Vol-2989/shortpaper18.pdf

Vauth, M., Meister, M., Hatzel, H. O., Gerstorfer, D., & **Gius, E.** (2022, March). Gitma. *Zenodo*. DOI: https://doi.org/10.5281/zenodo.6330464

TO CITE THIS ARTICLE:
Vauth, M., & Gius, E. (2022). Event Annotations of Prose. *Journal of Open Humanities Data*, 8: 19, pp. 1–6. DOI: https://doi.org/10.5334/johd.83

Published: 12 August 2022

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