Journal of Open
Humanities Data

Knowledge Distribution in German Drama: An Annotated Corpus

DATA PAPER

]u[ubiquity press

MELANIE ANDRESEN
BENJAMIN KRAUTTER
JANIS PAGEL
NILS REITER

*Author affiliations can be found in the back matter of this article

ABSTRACT

What do characters in theater plays know about character relations, and how does the distribution of knowledge evolve over a play's course? We present a dataset of 30 German plays annotated with information about the distribution of knowledge about character relations (such as "A learns from B that C is the parent of D"). All plays were manually annotated by two independent annotators in the *Q:TRACK* project, which aims to systematically model character knowledge. The dataset is available on GitHub and Zenodo and can be reused, for example, for systematic studies of knowledge in plays or for analyzing annotator disagreements.

CORRESPONDING AUTHOR:
Melanie Andresen
Institute for Natural Language Processing, University of Stuttgart, Stuttgart, Germany
melanie.andresen@ims.uni-stuttgart.de

KEYWORDS:
computational literary studies; drama; annotation; knowledge; character relations

TO CITE THIS ARTICLE:
Andresen, M., Krautter, B., Pagel, J., & Reiter, N. (2024). Knowledge Distribution in German Drama: An Annotated Corpus. *Journal of Open Humanities Data*, 10: 6, pp. 1–7. DOI: https://doi.org/10.5334/johd.167

1 OVERVIEW

REPOSITORY LOCATION

The dataset can be found on GitHub at https://github.com/quadrama/knowledge-annotation as well as on Zenodo at https://doi.org/10.5281/zenodo.8319261.

CONTEXT

This dataset was collected in the project *Q:TRACK – Quantitative Drama Analytics: Tracking Character Knowledge*. *Q:TRACK* targets the fact that a play's dramatic characters can have different levels of awareness of certain information. Hence, the transmission and distribution of knowledge is one central object of study for drama analysis. In his *Poetics*, Aristotle emphasizes the importance of so-called *anagnorisis*. Aristotle's concept of *anagnorisis* refers to recognition scenes, where a character, for instance, recognizes a long-lost relative and all previous events appear in a new light (Aristotle, 1995). The "discrepant awareness" (Evans, 1960, p. VIII) of different characters and/or characters and the audience can propel the plot of a play, creates suspense and thus greatly contributes to the play's effect (Anz, 1998; Cave, 1988; Pfister, 1988). Therefore, the project aims to systematically model and track the distribution of knowledge in plays through annotation.

This adds to existing research in the field of *computational literary studies* where characters and their relationships in plays have recently gained attention (Fischer, Trilcke, Kittel, Milling, & Skorinkin, 2018; Krautter & Vauth, 2023; Lee & Lee, 2017; Trilcke, 2022). The knowledge distribution can also be used to specify the character interactions with regard to network analysis (Krautter, 2023). Character relationships are also covered in a dataset by Massey, Xia, Bamman, and Smith (2015) that is based on English narratives. However, they do not distinguish between the diverging and developing knowledge of individual characters and work with text summaries only.

We restricted the annotation to the domain of knowledge about character relations, as this domain is key in many plays. In Johann Gottlob Benjamin Pfeil's tragedy *Lucie Woodvil* (1756), for instance, the main character Lucie learns too late that her lover and father of her unborn child is also her brother. The annotated relations in our dataset include family relations (`parent_of(A, B)`, `child_of(B, A)`, `siblings(B, C) …`), love relations (`in_love_with(B, D)`, `engaged(B, D)`, `spouses(B, D) …`) questions of identity (`identity (A, E)`, `has_name(A, 'name')`) and death (`dead(A)`, `murderer_of(B, A)`). In addition to the knowledge itself, the annotations contain information about the source and target of each knowledge transfer. This results in the following tag structure:

```
transfer(SOURCE, TARGET, KNOWLEDGE, ATTRIBUTES)
```

`SOURCE` is the character that passes on the knowledge, `TARGET` is one or several characters that receive the knowledge, and `KNOWLEDGE` specifies the knowledge itself as one of the character relations described above. Optional attributes allow to include additional information, e. g. if `SOURCE` is lying or if the information is still uncertain (see details in Andresen, Krautter, Pagel, & Reiter, 2021, in German).

2 METHOD

This dataset was created by manual annotation using the tool CorefAnnotator (Reiter, 2018).

STEPS

The dataset comprises the 30 German plays listed in Table 1, with a total size of 736,808 tokens (including all utterances as well as stage directions). The plays were retrieved in the TEI-XML format from the Drama Corpora Project (Fischer et al., 2019) and imported into *CorefAnnotator*. The data were annotated in three rounds:

1. In the initial round, 16 plays were annotated by two annotators following a preliminary guideline. Issues were discussed with one of the authors and, where necessary, with the

ID	AUTHOR	TEXT	YEAR
1	Brentano, C.	Ponce de Leon	1803
2	von Eichendorff, J.	Die Freier	1833
3	Gellert, C. F.	Die zärtlichen Schwestern	1747
4	Goethe, J. W.	Die natürliche Tochter	1803
5	Goethe, J. W.	Iphigenie auf Tauris	1787
6	Goethe, J. W.	Stella	1776
7	Goethe, J. W.	Clavigo	1774
8	Gottsched, L. A. V.	Das Testament	1745
9	Grillparzer, F.	Die Ahnfrau	1817
10	von Günderode, K.	Magie und Schicksal	1805
11	von Günderode, K.	Udohla	1805
12	Hauptmann, G.	Vor Sonnenaufgang	1889
13	Hebbel, F.	Maria Magdalene	1844
14	von Hofmannsthal, H.	Der Rosenkavalier	1911
15	von Hofmannsthal, H.	Elektra	1903
16	von Kleist, H.	Familie Schroffenstein	1803
17	Klinger, F. M.	Die Zwillinge	1776
18	Lenz, J. M. R.	Der Hofmeister	1774
19	Lessing, G. E.	Nathan der Weise	1779
20	Lessing, G. E.	Emilia Galotti	1772
21	Lessing, G. E.	Miß Sara Sampson	1755
22	Pfeil, J. G. B.	Lucie Woodvil	1756
23	Schiller, F.	Die Braut von Messina	1803
24	Schiller, F.	Die Räuber	1781
25	Schiller, F.	Maria Stuart	1800
26	Schlegel, J. E.	Canut	1746
27	Schnitzler, A.	Komtesse Mizzi oder Der Familientag	1909
28	Wagner, H. L.	Die Kindermörderin	1776
29	Wagner, R.	Die Walküre	1853
30	von Weißenthurn, J.	Das Manuscript	1817

Table 1 List of all plays included in the corpus.

whole team. This process resulted in the final annotation guideline (Andresen et al., 2021, in German).

2. In the second round, the other 14 plays were annotated independently following the guideline. These plays were used to calculate the inter-annotator agreement using the measure gamma by Mathet, Widlöcher, and Métivier (2015), as presented (and discussed critically) in Andresen, Krautter, Pagel, and Reiter (2022b).
3. In a final round, every play was discussed and double checked by at least one annotator. In this round, three more relations were added for murder, death and pregnancy.

The final version of the corpus (round 3) comprises 37 files, as for seven plays, both annotators performed the last step of finalizing the annotations, resulting in two final versions for these plays. We decided to keep two versions instead of creating a single gold standard, because in many cases more than one way of annotating the play was justified (see below). In total, there are 1277 annotated text passages, which corresponds to an average number of 34.5 annotations per text, with a considerable standard deviation of 18.8.

SAMPLING STRATEGY

The plays were manually selected to cover

- plays of which we knew that knowledge about character relations is important for the plot, (?) as well as plays where this was not the case,
- tragedies as well as comedies,
- plays from different literary epochs (1740–1900).

Accordingly, the dataset is not designed to be representative of a specific group of texts, but to cover a wide range of relevant phenomena.

QUALITY CONTROL

All plays were annotated by two people independently, making it possible to calculate the inter-annotator agreement. The agreement is rather low for many of the plays, see Table 2. This is due to the high complexity and interpretation dependency of the task. In many cases more than one way of modeling the data is plausible. Also, measuring inter-annotator agreement in a way that makes the scores comparable to other studies is challenging for annotations without predefined annotation spans. See Andresen et al. (2022b) for a more in-depth discussion and the repository for more detailed scores. We publish several versions of each annotation as well as the annotation guidelines (Andresen et al., 2021, in German) for comparability and transparency.

TEXT	UNLABELED	LABELED
Brentano: Ponce de Leon	0.576	0.355
Eichendorff: Die Freier	0.573	0.375
Gellert: Die zärtlichen Schwestern	0.474	0.476
Goethe: Clavigo	0.427	0.438
Gottsched: Das Testament	0.401	0.290
Günderrode: Magie und Schicksal	0.536	0.428
Günderrode: Udohla	0.467	0.194
Hauptmann: Vor Sonnenaufgang	0.644	0.493
Lessing: Miß Sara Sampson	0.531	0.362
Schiller: Maria Stuart	0.651	0.496
Schlegel: Canut	0.519	0.431
Wagner: Die Kindermörderin	0.493	0.410
Wagner: Die Walküre	0.602	0.400
Weißenthurn: Das Manuscript	0.634	0.510
mean	0.538	0.404

Table 2 IAA scores (Gamma) for the 14 texts of annotation round 2. For the unlabeled scores, only the position of annotations is taken into account. For the labeled scores, position and labels are considered.

3 DATASET DESCRIPTION

Object name quadrama/knowledge-annotation

Format names and versions CSV, JSON, ca2z (a compressed data format used by the *CorefAnnotator*)

Creation dates 2020-11-01 until 2023-08-02

Dataset creators Melanie Andresen (University of Stuttgart), Benjamin Krautter (University of Cologne), Janis Pagel (University of Cologne), Nils Reiter (University of Cologne), Christian Lantzinger (student assistant, University of Stuttgart), and Jonas Hirner (student assistant, University of Stuttgart).

Language The plays in the dataset are in German, the annotation labels and variable names are in English.

License CC-BY-4.0

Repository name GitHub, Zenodo

Publication date 2023-09-05

4 REUSE POTENTIAL

The dataset can be reused in a number of ways. Literary scholars might take the data as a starting point for a systematic analysis of knowing and not-knowing, knowledge distribution and knowledge transmission between characters in one or several individual plays. This is often considered a crucial piece of information for the interpretation of dramatic texts (Gutjahr, 2012; Kiss, 2010). Horstmann (2018, pp. 184–209) has proposed to narratologically reinforce theater studies by including focalization, understood as relations of knowledge, into the analysis. Analyses of individual plays can be supported by the visualization of the data as we have suggested in Andresen, Krautter, Pagel, and Reiter (2022a) and Andresen et al. (2022b).

Quantitative analyses of the frequency of specific types of knowledge transfers, for instance, are limited by the size of the dataset, but are still possible on a small scale. This allows insights into which relations are discussed most often, which characters are the most important for knowledge transfer and similar questions. The annotations could also be aligned with the attempt to model character relationships based on topic modeling as presented in Iyyer, Guha, Chaturvedi, Boyd-Graber, and Daumé III (2016).

To solve the problem of data scarcity in the long term, the dataset can be used as training and/or test data for attempts to automate this type of annotation, for instance by prompting large language models (Liu et al., 2023; Ziems et al., 2023). As we provide the annotations of two annotators for most plays, the data can also be used to investigate annotation disagreement. One may investigate if annotation disagreements point to ambiguous and potentially crucial text passages or look into the causes of disagreements (Andresen, Vauth, & Zinsmeister, 2020; Gius & Jacke, 2017).

FUNDING INFORMATION

The project *Q:TRACK* is funded by the German Research Foundation (DFG) within the priority program SPP 2207 Computational Literary Studies (CLS).

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Melanie Andresen: Data Curation, Supervision, Writing – original draft

Benjamin Krautter: Conceptualization, Writing – review & editing

Janis Pagel: Data Curation, Writing – review & editing

Nils Reiter: Project Administration, Supervision, Conceptualization, Funding Acquisition

AUTHOR AFFILIATIONS

Melanie Andresen orcid.org/0000-0002-3913-1273
Institute for Natural Language Processing, University of Stuttgart, Stuttgart, Germany

Benjamin Krautter orcid.org/0000-0003-1804-7520
Department of Digital Humanities, University of Cologne, Cologne, Germany

Janis Pagel orcid.org/0000-0003-4370-1483
Department of Digital Humanities, University of Cologne, Cologne, Germany

Nils Reiter orcid.org/0000-0003-3193-6170
Department of Digital Humanities, University of Cologne, Cologne, Germany

REFERENCES

Andresen, M., Krautter, B., Pagel, J., & **Reiter, N.** (2021). *Wissensvermittlungen im Drama annotieren. Annotationsguideline.* Zenodo. (Version Number: 1.0) DOI: https://doi.org/10.5281/zenodo.5729706

Andresen, M., Krautter, B., Pagel, J., & **Reiter, N.** (2022a). Nathan nicht ihr Vater? – Wissensvermittlungen im Drama annotieren. In *Book of Abstracts of DHd 2022*. DOI: https://doi.org/10.5281/zenodo.6327912

Andresen, M., Krautter, B., Pagel, J., & **Reiter, N.** (2022b). Who knows what in German drama? A composite annotation scheme for knowledge transfer. *Journal of Computational Literary Studies*, *1*(1). DOI: https://doi.org/10.48694/jcls.107

Andresen, M., Vauth, M., & **Zinsmeister, H.** (2020). Modeling Ambiguity with Many Annotators and Self-Assessments. In *Proceedings of the 14th Linguistic Annotation Workshop* (pp. 48–59). Retrieved 12 December 2023, from https://www.aclweb.org/anthology/2020.law-1.5/

Anz, T. (1998). *Literatur und Lust. Glück und Unglück beim Lesen*. München: C.H. Beck.

Aristotle. (1995). Poetics. In S. Halliwell (Ed.), *Aristotle: Poetics* (pp. 27–141). Cambridge and London: Harvard University Press. DOI: https://doi.org/10.4159/DLCL.aristotle-poetics.1995

Cave, T. (1988). *Recognitions. A study in poetics*. Oxford: Clarendon Press.

Evans, B. (1960). *Shakespeare's comedies*. Oxford: Clarendon Press.

Fischer, F., Börner, I., Göbel, M., Hechtl, A., Kittel, C., Milling, C., & **Trilcke, P.** (2019). Programmable Corpora – Die digitale Literaturwissenschaft zwischen Forschung und Infrastruktur am Beispiel von DraCor. In *DHd 2019 Digital Humanities: multimedial & multimodal. Book of Abstracts* (pp. 194–197). DOI: https://doi.org/10.5281/zenodo.2596095

Fischer, F., Trilcke, P., Kittel, C., Milling, C., & **Skorinkin, D.** (2018). To catch a protagonist: Quantitative dominance relations in German-language drama (1730–1930). In *Book of Abstracts of DH 2018* (pp. 193–201). Retrieved 12 December 2023, from https://dh2018.adho.org/wp-content/uploads/2018/06/dh2018_abstracts.pdf#page=193

Gius, E., & **Jacke, J.** (2017). The Hermeneutic Profit of Annotation. On Preventing and Fostering Disagreement in Literary Text Analysis. *International Journal of Humanities and Arts Computing*, *11*(2), 233–254. DOI: https://doi.org/10.3366/ijhac.2017.0194

Gutjahr, O. (2012). Komödie des (Ge)Wissens: Heinrich von Kleists 'Der zerbrochne Krug'. In Y. Lü, A. Stephens, A. Lewis, & W. Voßkamp (Eds.), *Wissensfiguren im Werk Heinrich von Kleists* (pp. 23–39). Rombach Verlag.

Horstmann, J. (2018). *Theaternarratologie. Ein erzähltheoretisches Analyseverfahren für Theaterinszenierungen*. Berlin and Boston: De Gruyter. DOI: https://doi.org/10.1515/9783110597868

Iyyer, M., Guha, A., Chaturvedi, S., Boyd-Graber, J., & **Daumé III, H.** (2016). Feuding Families and Former Friends: Unsupervised Learning for Dynamic Fictional Relationships. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1534–1544). DOI: https://doi.org/10.18653/v1/N16-1180

Kiss, O. (2010). Reinventing the Plot: J. C. Gottsched's 'Sterbender Cato'. *Deutsche Vierteljahrsschrift für Literaturwissenschaft und Geistesgeschichte*, *84*(4), 507–525. DOI: https://doi.org/10.1007/BF03375820

Krautter, B. (2023). Kopräsenz-, Koreferenz- und Wissens-Netzwerke. Kantenkriterien in dramatischen Figurennetzwerken am Beispiel von Kleists 'Die Familie Schroffenstein' (1803). *Journal of Literary Theory*, *17*(2), 261–289. DOI: https://doi.org/10.1515/jlt-2023-2012

Krautter, B., & **Vauth, M.** (2023). Konstellationen kommunikativer Macht. Hypothesengeleitete Netzwerkanalyse in der Literaturwissenschaft. In H. Schwab (Ed.), *Figurenkonstellation und Gesellschaftsentwurf: Annäherungen an eine narratologische Kategorie und ihre Deutungspotentiale* (pp. 205–238). Universitätsverlag Winter.

Lee, J., & **Lee, J.** (2017). Shakespeare's tragic social network; or why all the world's a stage. *Digital Humanities Quarterly*, *11*(2). Retrieved 12 December 2023, from http://digitalhumanities.org:8081/dhq/vol/11/2/000289/000289.html

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & **Neubig, G.** (2023). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*, *55*(9), 195:1–195:35. DOI: https://doi.org/10.1145/3560815

Massey, P., Xia, P., Bamman, D., & **Smith, N. A.** (2015). *Annotating Character Relationships in Literary Texts*. arXiv. DOI: https://doi.org/10.48550/arXiv.1512.00728

Mathet, Y., Widlöcher, A., & **Métivier, J.-P.** (2015). The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics*, *41*(3), 437–479. DOI: https://doi.org/10.1162/COLI_a_00227

Pfister, M. (1988). *The theory and analysis of drama* (J. Halliday, Trans.). Cambridge and New York: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511553998

Reiter, N. (2018). CorefAnnotator – a new annotation tool for entity references. In *EADH 2018*. Retrieved 12 December 2023, from https://eadh2018.exordo.com/programme/presentation/118

Trilcke, P. (2022). Small Worlds, Beat Charts und die Netzwerkanalyse dramatischer Texte. Reflexionen aus dem Rabbit Hole. In F. Jannidis (Ed.), *Digitale Literaturwissenschaft: DFG-Symposion 2017* (pp. 563–596). J.B. Metzler. DOI: https://doi.org/10.1007/978-3-476-05886-7_23

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & **Yang, D.** (2023). Can large language models transform computational social science? *Computational Linguistics*, *49*(4), 1–53. DOI: https://doi.org/10.1162/coli_a_00502

TO CITE THIS ARTICLE:
Andresen, M., Krautter, B., Pagel, J., & Reiter, N. (2024). Knowledge Distribution in German Drama: An Annotated Corpus. *Journal of Open Humanities Data*, 10: 6, pp. 1–7. DOI: https://doi.org/10.5334/johd.167

Submitted: 11 October 2023
Accepted: 08 December 2023
Published: 12 January 2024

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