Journal of Open Humanities Data

The ChildPoeDE Corpus: 1082 German Children's Poems for Computational and Experimental Studies on Poetry Reception

DATA PAPER

MARINA LEHMANN
ANNE HEUMANN
MONIEK M. KUIJPERS
GERHARD LAUER
JANA LÜDTKE

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

ubiquity press

ABSTRACT

We introduce childPoeDE: the first corpus of German poetry for children comprising poems which are still read today and cover a wide range of topics and authors. ChildPoeDE contains poem texts and both poem-level and token-level metadata. Poem-level metadata includes information about the anthologies and authors, quantitative text features, rhyme and lexical richness. Token-level metadata covers word length, position and frequency, parts-of-speech, onomatopoeia and sonority. This corpus can be used for computational text analysis, but also as a source for stimulus material in experimental studies. The corpus metadata is freely accessible via Zenodo. The poem texts are protected by copyright.

CORRESPONDING AUTHOR:

Marina Lehmann

Department of Book Studies, Johannes Gutenberg-Universität Mainz, Mainz, Germany

marina.lehmann@uni-mainz.de

KEYWORDS:

German poetry for children; text corpus; text analysis; computational literary studies; stimulus material

TO CITE THIS ARTICLE:

Lehmann, M., Heumann, A., Kuijpers, M. M., Lauer, G., & Lüdtke, J. (2023). The ChildPoeDE Corpus: 1082 German Children's Poems for Computational and Experimental Studies on Poetry Reception. *Journal of Open Humanities Data*, 9: 6, pp. 1–6. DOI: https://doi.org/10.5334/johd.102

(1) OVERVIEW

REPOSITORY LOCATION

Zenodo: https://zenodo.org/record/7936860

CONTEXT

Few of the German literary corpora available today focus on children's and young adult literature (e.g. Schroeder et al., 2015) and none of them specifically on poetry. The childPoeDE corpus tries to fill this gap by providing a set of 1082 poems for children, which can be used in computational and experimental studies. It was built within a project on sentiment analysis in children's and young adult literature (Deutsche Forschungsgemeinschaft, 2023). Seven anthologies – published between 1991 and 2018 – form the basis of this corpus. The poems were written by 356 different authors (84 female, 271 male, 1 unknown) and cover a wide range of topics from animals and nature to family life and children's dreams, from everyday situations to adventures, from humorous to serious. The texts also vary in style. Some poems follow traditional formats with strict rhyme and metre, while others show free, experimental or playful characteristics, such as onomatopoeia.

(2) METHOD

STEPS

Based on recommendations by five experts on German children's literature and poetry – all professors or research associates from the fields of German Studies, Didactics or Pedagogy – we selected the following seven anthologies as sources: *Großer Ozean* (Gelberg, 2000), *Die schönsten Kindergedichte* (Kruse, 2003), *Sieben Ziegen fliegen durch die Nacht* (Gutzschhahn, 2018), *Sieben kecke Schnirkelschnecken* (Sailer, 2010), *Im Mondlicht wächst das Gras* (Andresen, 1991), *Ich liebe dich wie Apfelmus* (Fried & Hein, 2006) and *So viele Tage wie das Jahr hat* (Krüss, 1998). The experts were asked to suggest anthologies based on the following criteria: the poems should still be widely read today, aimed mostly at primary school children, written between 1800 and 2018, cover a wide range of poetry, including classics but also less known poetry, written in German and focus on text (not on pictures) to convey meaning. We chose these seven anthologies because they were named multiple times by different experts. Since we intended to provide data suitable as stimulus material in contemporary studies, we only included anthologies with editions published in the last 25 years. We used OCR software from Tesseract and Adobe.

Further, we collected poem-level and token-level metadata (csv). Some poem-level information was added manually (author, title, anthology, anthology count, publisher, publication year, ISBN). From the Integrated Authority File (GND), we retrieved additional data about the authors (GND id, author gender, year of birth, year of death) to ensure their accurate identification. We decided to provide the author's year of birth and year of death as an indication of when the poem could have been published, as it was difficult and, in some cases, impossible to find original publication dates for single poems. Most features, however, were extracted with our own Python script (poemtool.py) (e.g. word/stanza/line counts, data on case, punctuation, layout, rhyme and sonority). To determine rhyme patterns, we used *rhymetagger* (Plecháč, 2018). Calculations for the sonority score are based on Jacobs (2017) and Stenneken et al. (2005). We also calculated the lexical density and type-token ratio (TTR) for each poem to provide information on lexical richness. Along with the standard TTR, we computed Moving-Average-TTRs (MATTRs) to account for different text lengths (Covington & McFall, 2010). As MATTRs are usually computed for longer texts, we used different window sizes. All TTR and MATTR values were calculated using the R-package *quanteda* (Benoit et al., 2018). Data on onomatopoeia was annotated manually. The token-level metadata file additionally provides data on word length, word position and parts-of-speech in different levels of granularity. Part-of-speech information was generated with *TreeTagger* (Schmid, 1995). We also published a frequency table with absolute and relative frequencies for all tokens present in the corpus. Figure 1 represents the childPoeDE corpus in descriptive statistics. It includes frequency tables for the features special layout, rhyme and onomatopoeia, histograms with boxplots for poem length (measured in the number of stanzas and lines), poem sonority, TTR, lexical density and rhyming degree, a word cloud of the most frequent content words, a pie chart on gender distribution and a table with the ten most frequent authors and the number of poems they contributed to the corpus.

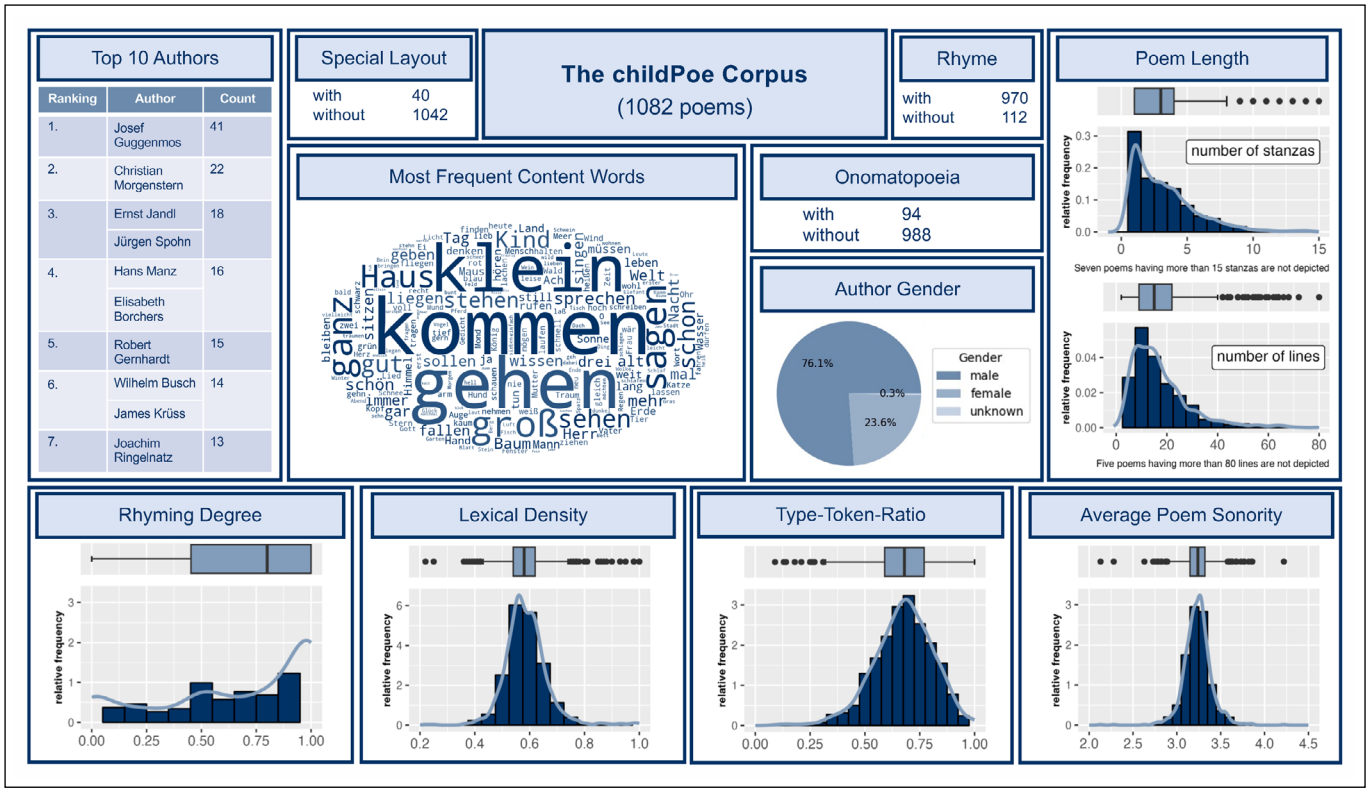

Figure 1 ChildPoeDE corpus – overview of poem-level metadata.

SAMPLING STRATEGY

We included as many poems from the anthologies as possible. However, poems relying on pictures, graphical layout or typography to convey meaning were excluded, as well as poems that used archaic or difficult language (e.g. all poems from "Des Knaben Wunderhorn") and poems consisting of a single repeated word. A list of the omitted poems can be found on Zenodo. If a poem appeared in more than one anthology, this was noted in the column "anthology count" in the poem-level metadata file. The childPoeDE corpus in its current state is a first (yet still imperfect) attempt to collect data of German poetry for children. Ideally, a corpus should be balanced with regards to author gender. We will work towards this in the future. For now, the gender imbalance of the corpus represents the gender imbalance present in the anthologies.

QUALITY CONTROL

All texts were checked for OCR errors. Additionally, whitespace and special characters, such as quotation marks, were normalised. We also harmonised the poems' structure to simplify automatic text processing: Detailed information on normalisation processes and explanations of text features can be found in the README files on Zenodo. The part-of-speech data was checked and manually corrected if necessary. In the end we conducted a quality check by reviewing randomly selected data.

(3) DATASET DESCRIPTION

OBJECT NAME

childPoeDE

FORMAT NAMES AND VERSIONS

TXT, CSV

Version 2.0

CREATION DATES

Start: 2021-01, End: 2023-02

DATASET CREATORS

Moniek Kuijpers, University of Basel

Priska Hadayani Rüegg, University of Basel

Jana Lüdtke, Freie Universität Berlin

Marina Lehmann, Johannes Gutenberg-Universität Mainz

Anne Heumann, Johannes Gutenberg-Universität Mainz

LANGUAGE

Data: German

Metadata: English

LICENCE

Poem-level metadata, token-level metadata, word-frequency table, TTR data and poemtool.py: CC 0

REPOSITORY NAME

Zenodo

PUBLICATION DATE

Version 2.0: 2023–05–15

(4) REUSE POTENTIAL

Although there is much research available on German poetry, both on corpora (e.g. Haider & Eger, 2019) and computational assessments (e.g. Reinig & Rehbein, 2019), these works never focus on German poetry for children alone. Thus, our data offers new research scenarios for anyone interested in poetry for children, such as empirical scholars, researchers in didactics or digital humanists. In experimental studies the texts can be used as stimulus material to investigate children's emotional involvement when reading poetry. Elaborate metadata allows for a precise poem selection along specific criteria, including rhyme, sonority or onomatopoeia. However, the corpus cannot provide all information which might be useful for empirical studies, including publication dates for individual poems or an evaluation of age appropriateness.

In the context of digital humanities, especially computational literary studies, our data allows for investigations of different poetic features and their correlations. There are plenty of possible approaches from the field of Natural Language Processing which can be performed on the data and might yield new insights on the study of German poetry for children. These include linguistic corpus analysis, sentiment analysis (for an example for children's books see Jacobs et al., 2020), text similarity assessment, topic modelling, named entity recognition or explorative approaches through visualisations.

Overall, the childPoeDE corpus lays the foundations for a wide range of research scenarios while being extensible at the same time. The data could be enriched with additional metadata (e.g. sentiment values, reading age or text complexity measures), linked to other data sets through the authors' GND ids or used for comparisons with corpora from other genres (i.e. childLex (Schroeder et al., 2015)).

ACKNOWLEDGEMENTS

Mesian Tilmatine, Freie Universität Berlin.

Nico Kestel, Technische Universität Berlin.

FUNDING INFORMATION

This data was collected within the project "Advanced sentiment analysis for understanding affective-aesthetic responses to literary texts: A computational and experimental psychology approach to children's literature" (424250469) funded by the DFG grant "SPP 2207: Computational Literary Studies" (402743989).

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

- Marina Lehmann: writing – final draft and revision, data curation, data publication
- Anne Heumann: writing – final draft and revision, data curation, data quality check
- Moniek M. Kuijpers: dataset creation, sample selection, data curation, writing – first draft and revision
- Gerhard Lauer: funding acquisition, supervision, writing – revision
- Jana Lüdtke: funding acquisition, data curation, supervision, writing – revision

AUTHOR AFFILIATIONS

Marina Lehmann orcid.org/0000-0002-6818-6169
Department of Book Studies, Johannes Gutenberg-Universität Mainz, Mainz, Germany

Anne Heumann orcid.org/0009-0000-5791-6982
Department of Book Studies, Johannes Gutenberg-Universität Mainz, Mainz, Germany

Moniek M. Kuijpers orcid.org/0000-0002-3676-5879
Digital Humanities Lab, University of Basel, Basel, Switzerland

Gerhard Lauer orcid.org/0000-0003-0230-2574
Department of Book Studies, Johannes Gutenberg-Universität Mainz, Mainz, Germany

Jana Lüdtke orcid.org/0000-0002-1581-6120
Department of Education and Psychology, Freie Universität Berlin, Berlin, Germany

REFERENCES

Andresen, U. (Ed.). (1991). *Im Mondlicht wächst das Gras: Gedichte für Kinder und alle im Haus*. Ravensburg: Otto Maier.

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & **Matsuo, A.** (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774. https://quanteda.io. DOI: https://doi.org/10.21105/joss.00774.

Deutsche Forschungsgemeinschaft. (2023). *CHYLSA – Advanced sentiment analysis for understanding affective-aesthetic responses to literary texts: A computational and experimental psychology approach to children's literature*. Last accessed 24 May 2023. https://gepris.dfg.de/gepris/projekt/424250469?language=en.

Covington, M. A., & **McFall, J. D.** (2010). Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR). *Journal of Quantitative Linguistics*, *17*(2), 94–100. DOI: https://doi.org/10.1080/09296171003643098.

Fried, A., & Hein, S. (Ed.) (2006). *Ich liebe dich wie Apfelmus: Die schönsten Gedichte für Kleine und Große*. Munich: cbj.

Gelberg, H. J. (Ed.) (2000). *Großer Ozean: Gedichte für alle. Bilder, Fotos, Illustrationen*. Weinheim: Beltz und Gelberg.

Gutzschhahn, U. M. (2018). *Sieben Ziegen fliegen durch die Nacht: hundert neue Kindergedichte*. Munich: Dtv.

Haider, T. N., & **Eger, S.** (2019). Semantic Change and Emerging Tropes In a Large Corpus of New High German Poetry. *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*. Florence, Italy, August 2, 2019, 216–222. https://aclanthology.org/W19-4727.pdf. DOI: https://doi.org/10.18653/v1/W19-4727.

Jacobs, A. M. (2017). Quantifying the Beauty of Words: A Neurocognitive Poetics Perspective. *Frontiers in Human Neuroscience*, *11*, Article 622. DOI: https://doi.org/10.3389/fnhum.2017.00622.

Jacobs, A. M., Herrmann, B., Lauer, G., Lüdtke, J., & **Schroeder, S.** (2020). Sentiment Analysis of Children and Youth Literature: Is There a Pollyanna Effect? *Frontiers in Psychology*, *11*. DOI: https://doi.org/10.3389/fpsyg.2020.574746.

Kruse, M. (Ed.). (2003). *Die schönsten Kindergedichte*. Berlin: Aufbau.

Krüss, J. (Ed.). (1998). *So viele Tage wie das Jahr hat: 365 Gedichte für Kinder und Kenner* (6th ed.). Gütersloh: Bertelsmann.

Plecháč, P. (2018). A Collocation-Driven Method of Discovering Rhymes (in Czech, English, and French Poetry). In M. Fidler & V. Cvrček (Eds.), *Taming the Corpus: From Inflection and Lexis to Interpretation* (pp. 79–95). New York: Springer. DOI: https://doi.org/10.1007/978-3-319-98017-1_5.

Reinig, I., & **Rehbein, I.** (2019). Metaphor detection for German Poetry. *Preliminary proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019). Erlangen-Nürnberg, Germany, October 9–11, 2019*, 149–160. https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/9316/file/Reinig_Rehbein_Metaphor_detection_2019.pdf.

Sailer, S. (Ed.). (2010). *Sieben kecke Schnirkelschnecken: Lustige Kindergedichte und Reimspaß zum Lachen*. Boston: Arena.

Schmid, H. (1995). Improvements in Part-of-Speech Tagging with an Application to German [Conference presentation]. *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland. https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger2.pdf.

Schroeder, S., Würzner, K.-M., Heister, J., Geyken, A., & **Kliegl, R.** (2015). childLex: A lexical database of German read by children. *Behavior Research Methods*, *47*, 1085–1094. DOI: https://doi.org/10.3758/s13428-014-0528-1.

Stenneken, P., Bastiaanse, R., Huber, W., & **Jacobs, A. M.** (2005). Syllable structure and sonority in language inventory and aphasic neologisms. *Brain and Language*, *95*(2), 280–292. DOI: https://doi.org/10.1016/j.bandl.2005.01.013.

TO CITE THIS ARTICLE:
Lehmann, M., Heumann, A., Kuijpers, M. M., Lauer, G., & Lüdtke, J. (2023). The ChildPoeDE Corpus: 1082 German Children's Poems for Computational and Experimental Studies on Poetry Reception. *Journal of Open Humanities Data*, 9: 6, pp. 1–6. DOI: https://doi.org/10.5334/johd.102

Published: 02 June 2023

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