

# National Library as Corpus: Introducing DeLiKo@DNB – a Large Synchronous German Fiction Corpus

## Kupietz, Marc

kupietz@ids-mannheim.de

Leibniz-Institut für Deutsche Sprache, Germany

ORCID: 0000-0001-8997-8256

## Leinen, Peter

p.leinen@dnb.de

Deutsche Nationalbibliothek, Germany

ORCID: 0000-0002-3014-000X

## Diewald, Nils

diewald@ids-mannheim.de

Leibniz-Institut für Deutsche Sprache, Germany

ORCID: 0000-0002-2993-9180

## Genêt, Philippe

p.genet@dnb.de

Deutsche Nationalbibliothek, Germany

ORCID: 0009-0001-5095-8052

## Wilm, Rebecca

wilm@ids-mannheim.de

Leibniz-Institut für Deutsche Sprache, Germany

ORCID: 0000-0002-7273-7832

## Witt, Andreas

witt@ids-mannheim.de

Leibniz-Institut für Deutsche Sprache, Germany

ORCID: 0000-0002-0299-5713

## Yaddehige, Rameela

yaddehige@ids-mannheim.de

Leibniz-Institut für Deutsche Sprache, Germany

ORCID: 0009-0008-3216-5797

## Introduction

Fiction books are weakly represented in the German Reference Corpus DeReKo (Kupietz et al. 2010, 2018). This is primarily due to the tremendously higher costs associated

with licensing and converting raw fiction data into TEI-encoded XML format, compared to newspaper articles (Kupietz et al. 2014, p. 2). In the following, we discuss how we addressed these challenges and successfully created a large extensible corpus of recent fiction books.

## The Legal Challenge

Linguistics and literary studies both face the challenge that their research data is affected by third-party rights. Obtaining transferable, uniform licenses for German fiction books is particularly costly as no licensing models for non-expressive use (previously also called “non-consumptive use”, see Kamocki 2018) of entire texts as primary research data are generally established. Moreover, individual author permissions are often required, since the use of texts as research data is not covered by standard licensing agreements between authors and publishers.

Our solution to this challenge consists of two main pillars: The first leverages § 14 of the German National Library Act (DNBG) requiring submission of all digitally published media works to the DNB. The second pillar is part of the general strategy adopted for DeReKo to address legal issues through infrastructural means, following Jim Gray's (2003) famous principle: *If the data cannot move, put the computation near the data* (Kupietz, Diewald, and Margaretha 2022, 163ff). In this application case, our approach is to leave the full texts in the DNB and make them accessible via an instance of the corpus analysis platform KorAP (Bański et al. 2012) running within DNB. This ensures that the data can be made available more openly than under the text and data mining exception,<sup>1</sup> and be used as effectively as possible for linguistic and related research purposes, but within the limits set by copyright law. The main restriction is that the context size of each search result is limited to a maximum of 50 words.

## Data and Methods

To address the second challenge, and to make conversion into high-quality TEI-XML encoded corpora feasible, we ignored PDF ebooks and limited our focus to the 273,976 books available in the XML-based EPUB format and drew a 10% random sample from these as a first step, stratified by year of publication, resulting in a sample of 26,091 ebooks. As a second step, on the occasion of the 20th anniversary of the German Book Prize in October 2024, we added all 362 digitally available longlisted titles from the past two decades to the corpus, in cooperation with the German Publishers and Booksellers Association. To convert the data to the TEI I5 format (Lüngen and Sperberg-McQueen 2012) used by DeReKo, we applied XSLT 3.0 stylesheets in three passes, via the Saxon XSLT processor and GNU Make, using the DNB SRU API to retrieve consistent metadata, a heuristic genre classifier based on this, and a MALLET (McCallum 2002) based implementation of the standard DeReKo

topic domain classifier (Weiß 2005; Klosa, Lungen, and Kupietz 2012). In subsequent steps, the TEI-XML data was converted to KorAP-XML format and annotated for POS and lemma using the TreeTagger (Schmid 1994), for POS and morphosyntactic properties using MarMoT (Mueller et al. 2013), and dependencies using MaltParser (Nivre et al. 2007). These tools were selected due to their good balance between accuracy and performance.<sup>2</sup> If required for specific applications, additional annotation layers can be added later, since KorAP supports any number of them.<sup>3</sup>

The entire conversion and annotation process was completed in 48 hours on a Linux server with 96 cores and 1.5 TB of RAM. The composition of the resulting corpus, categorized by genres and publication years, is presented in Figure 1. Genre classifications were derived from the DNB metadata using string matching heuristics.<sup>4</sup> It is important to note that the relative proportions and the ‘representativeness’ or ‘balance’ of strata are not relevant in the case of DeLiKo@DNB. Instead, only the minimum absolute sizes of the strata are important, as users are invited to define their own virtual subcorpora. This allows them to create stratified, task-specific subsamples based on metadata constraints (see Kupietz et al. 2010; Kupietz 2016 for a detailed account of this *primordial sample* approach).

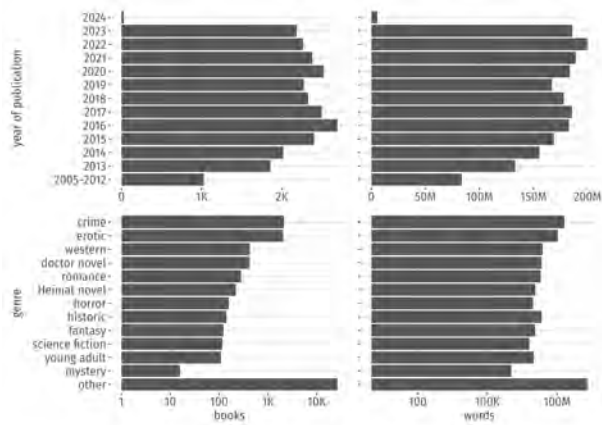


Figure 1: DeLiKo@DNB composition by publication year and genre (log scaled)

## Using DeLiKo@DNB

DeLiKo@DNB, currently comprising 2.02 billion words, is freely accessible through the website <https://korap.dnb.de> (see Figure 1), leveraging the full range of KorAP features, such as flexible metadata-based definition of virtual subcorpora, multiple annotation layers, and complex annotation search in six supported query languages, simplified by a query assistant and query-by-match functionalities. The corpus is also accessible via the KorAP API, supported by libraries for R<sup>5</sup> and Python<sup>6</sup> (Kupietz, Diewald, and Margaretha 2022).

## Outlook

We aim to regularly expand the corpus by incorporating newly published books. Additionally, we plan to enhance the search and analysis capabilities by integrating advancements from the long-term KorAP project, including updates to the user interface and client libraries. Further additions, improvements, and extensions, concerning e.g. the addition of books, text classifications, or annotation layers, will be driven by the demands of the user communities engaging with DeLiKo@DNB.

Figure 2: KorAP query in DeLiKo@DNB for the lemma 'Herz' (heart) followed by a verb and an adjective, in a virtual corpus containing only horror novels, published between 2010 and 2019

## Fußnoten

1. The German TDM implementation (§ 60d UrhG) allows only “to make the corpus available to the public for a specifically limited circle of persons for their joint scientific research“. For the DNB's holdings, this is only possible on the DNB's premises.
2. The source code of the conversion pipeline is available on GitHub at
3. For the advantages of maximising recall or precision via multiple classifiers, instead of trying to increase accuracy with just one classifier, see Belica et al. (2011).
4. See the genre table in <https://github.com/KorAP/epub2korap/blob/main/xslt/epub2i5.xsl>
5. See <https://cran.r-project.org/package=RKorAPClient>. An example R script that produces Figure 1 can be found here <https://github.com/KorAP/epub2korap/blob/main/scripts/DeLiKoCompositionChart.R>.
6. <https://pypi.org/project/KorAPClient/>

## Bibliographie

- Bański, Piotr, Peter M. Fischer, Elena Frick, Erik Ketzan, Marc Kupietz, Carsten Schnober, Oliver Schonefeld, and Andreas Witt. 2012. “The New IDS Corpus Analysis Platform: Challenges and Prospects.” In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 2905–11. Istanbul, Turkey: European Language Resources Association (ELRA). [http://www.lrec-conf.org/proceedings/lrec2012/pdf/789\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2012/pdf/789_Paper.pdf).
- Belica, Cyril, Marc Kupietz, Harald Lungen, and Andreas Witt. 2011. “The Morphosyntactic Annotation of DeReKo: Interpretation, Opportunities and Pitfalls.” In *Grammar & Corpora. Third International Conference*, Mannheim, Sept., 22–24 2009, 451–69. Tübingen: Narr. <https://nbn-resolving.org/urn:nbn:de:bsz:mh39-44890>.
- Gray, Jim. 2003. “Distributed Computing Economics.” MSR-TR-2003-24. Microsoft Research.
- Kamocki, Pawel. 2018. “The Argument for ‘Non-Consumptive Use’ in the EU: How Copyright Could Be Redefined to Allow Text and Data Mining.” In *Intellectual Property Perspectives on the Regulation of New Technologies*, 237–58. Edward Elgar Publishing. <https://doi.org/10.4337/9781786436382.00016>.
- Klosa, Annette, Marc Kupietz, and Harald Lungen. 2012. “Zum Nutzen von Korpusauszeichnungen für die Lexikographie.” *Lexicographica* 28:71–97.
- Kupietz, Marc. 2016. “Constructing a Corpus.” In *The Oxford Handbook of Lexicography*, edited by Philip Durkin, 62–75. Oxford: OUP. <https://doi.org/10.1093/oxfordhb/9780199691630.013.5>
- Kupietz, Marc, Cyril Belica, Holger Keibel, and Andreas Witt. 2010. “The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research.” In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA). [http://www.lrec-conf.org/proceedings/lrec2010/pdf/414\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2010/pdf/414_Paper.pdf).
- Kupietz, Marc, Nils Diewald, and Eliza Margaretha. 2022. “Building Paths to Corpus Data: A Multi-Level Least Effort and Maximum Return Approach.” In *CLARIN. The Infrastructure for Language Resources.*, edited by Darja Fišer and Andreas Witt. Berlin: deGruyter. <https://doi.org/10.1515/9783110767377-007>.
- Kupietz, Marc, Harald Lungen, Piotr Bański, and Cyril Belica. 2014. “Maximizing the Potential of Very Large Corpora.” In *Proceedings of the LREC-2014-Workshop Challenges in the Management of Large Corpora (CMLC2)*, edited by Marc Kupietz, Hanno Biber, Harald Lungen, Piotr Bański, Evelyn Breiteneder, Karlheinz Mörrth, Andreas Witt, and Jani Takhsha, ELRA, 1–6. Reykjavik, Iceland. <https://nbn-resolving.org/urn:nbn:de:bsz:mh39-31634>
- Kupietz, Marc, Harald Lungen, Pawel Kamocki, and Andreas Witt. 2018. “The German Reference Corpus DeReKo: New Developments – New Opportunities.” In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). <https://www.aclweb.org/anthology/L18-1689>.
- Lungen, Harald, and Christopher M. Sperberg-McQueen. 2012. “A TEI P5 Document Grammar for the IDS Text Model.” *Journal of the Text Encoding Initiative*, no. 3, 1–18.
- McCallum, Andrew Kachites. 2002. “MALLET: A Machine Learning for Language Toolkit.” <http://mallet.cs.umass.edu>.
- Mueller, Thomas, Helmut Schmid, and Hinrich Schütze. 2013. “Efficient Higher-Order CRFs for Morphological Tagging.” In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, edited by David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard, 322–32. Seattle, Washington, USA: Association for Computational Linguistics. <https://aclanthology.org/D13-1032>.
- Nivre, Joakim, Johan Hall, Jens Nilsson, Atanas Chanev, Gülşen Eryiğit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. “MaltParser: A Language-Independent System for Data-Driven Dependency Parsing.” *Natural Language Engineering* 13 (June): 95–135. <https://doi.org/10.1017/S1351324906004505>.
- Schmid, Helmut. 1994. “Probabilistic Part-of-Speech Tagging Using Decision Trees.” In *Proceedings of International Conference on New Methods in Language Processing*, 44–49. Manchester, UK. <https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf>.
- Weiß, Christian. 2005. “Die Thematische Erschließung von Sprachkorpora.” *OPAL - Online publizierte Arbeiten*

*zur Linguistik*, OPAL - Online publizierte Arbeiten zur Linguistik - 2005,1, 2005 (1). <http://pub.ids-mannheim.de/laufend/opal/pdf/opal2005-1.pdf>.