National Library as Corpus: DeLiKo-2025@DNB – A Very Large Corpus of German-language Contemporary Literature

Marc Kupietz[1], **Nils Diewald**[1], **Philippe Genêt**[2], **Andreas Witt**[1]
[1] Leibniz Institute for the German Language (IDS), [2] German National Library (DNB)
[1] Mannheim, [2] Frankfurt
[1] {kupietz, diewald, witt}@ids-mannheim.de, [2] p.genet@dnb.de

Abstract

This paper introduces DeLiKo-2025@DNB, a very large, linguistically annotated corpus of German-language contemporary literature, freely accessible via https://korap.dnb.de/. The corpus currently comprises 21 billion words from over 287,000 books published between 2005 and the present, spanning pulp and genre fiction as well as literary award-winning works. It covers the entire holdings of EPUB-format fiction ebooks deposited with the German National Library (DNB). We provide a detailed account of the corpus composition, metadata, and key features. Additionally, we explain our strategy for enabling lawful and effective access through the deployment of the open-source corpus analysis platform KorAP at the DNB, and we discuss both the transferability of our approach and work to other national libraries and our ongoing and planned extensions and enhancements.

Keywords: corpora, fiction, German, applications, legal issues, infrastructures

1. Introduction

For a long time, contemporary German-language literature was either absent or only weakly represented in major German language corpora, such as the German Reference Corpus DeReKo (Leibniz-Institut für Deutsche Sprache 2025; Kupietz et al. 2010, 2018) and the DWDS corpora (Geyken, 2007), and was therefore largely unavailable for empirically based research employing quantitative methods. The main reason for this was the seemingly insurmountable problem of copyright and licensing, and to a lesser extent, the cost of converting raw data into a standardized TEI XML format (Kupietz et al., 2014, p. 2). An improvement, and at the same time a scalable approach to addressing both the legal and technical challenges, only emerged with the release of the *Deutsches Literaturkorpus at the DNB* (DeLiKo@DNB, Deutsche Nationalbibliothek, Leibniz-Institut für Deutsche Sprache 2025; Kupietz et al., 2025) via the open-source corpus analysis platform KorAP (Bański et al., 2012). However, the first DeLiKo corpus comprised "only" a sample of 26,453 books, which nevertheless already included all digitally available German Book Prize longlist nominees since 2005. With the expanded corpus DeLiKo-2025@DNB presented here, the situation changes fundamentally: the majority of contemporary prose literature published in digital form by publishing houses is now available as a research corpus.

In the following, we first outline the legal challenges and our approach to addressing them (Section 2). We then describe the data and methods used to create the corpus (Section 3), followed by a brief overview of how to access and use DeLiKo@DNB (Section 4). Finally, we describe some of our current challenges and our approaches to tackle them, discuss the transferability of our work to other national libraries, and provide an outlook on future developments (Section 5).

2. Legal Framework and Solutions

Linguistics and literary studies both face the challenge that their research data is affected by third-party rights. Obtaining transferable, uniform licenses for fiction books is particularly costly, as typically 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 rests on two main pillars. The first relies on § 14 of the German National Library Act (DNBG), which requires that all digitally published media works be deposited with the DNB. The second pillar follows DeReKo's established strategy of addressing legal issues through infrastructure, drawing on Jim Gray's (2003) principle: *If the data cannot move, put the computation near the data* (Kupietz et al., 2022, 163ff). In our case, the full texts remain inside the DNB and are accessed through an internal instance of the corpus analysis platform KorAP (Bański et al., 2012). This arrangement allows the data to be shared more openly than permitted under the text and data

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 6528–6535
11–16 May 2026.

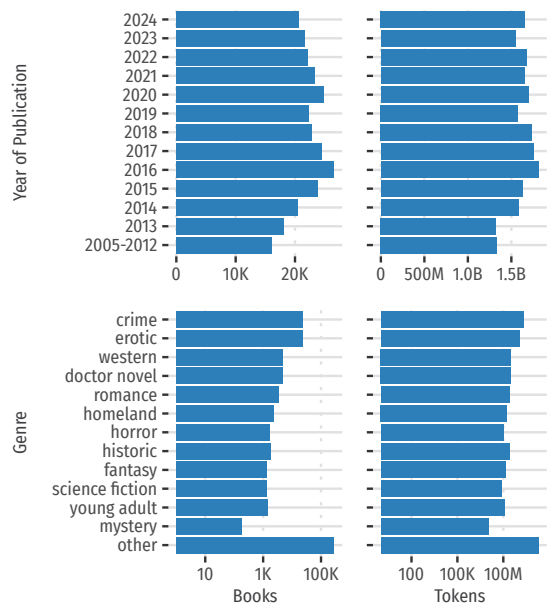

Figure 1: DeLiKo-2025@DNB composition by publication year and genre (log scaled)

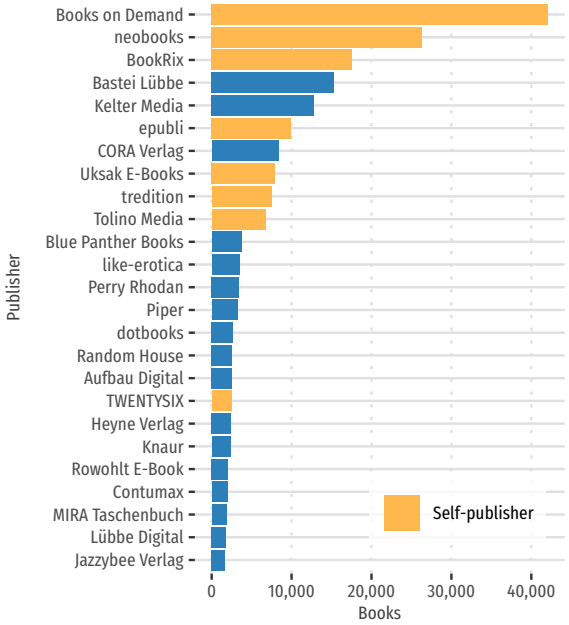

Figure 2: Top 25 publishers by number of books in DeLiKo@DNB.

mining exception,[1] while still keeping usage within the limits of copyright law. The main restriction is that the size of each search result is capped at 51 words.

3. Data and Methods

3.1. Data

The DeLiKo@DNB corpus comprises all German-language fiction ebooks that were published in EPUB format between 2012 and 2024 in Germany, summing up to almost 288,000 books containing roughly 21 billion tokens. Since the German National Library collects all publications regardless of their topics, genres or provenance, the corpus represents the full variety of published texts — from high literature to everyday prose, from science fiction to erotic dime novels — and thus reflects the complete range of contemporary German fiction (see Figure 1). This makes DeLiKo@DNB an ideal starting point for literary as well as for linguistic research.

As shown in Figure 2, the corpus contains many self-published and print-on-demand books. If needed, these can be excluded by defining appropriate virtual subcorpora (see Section 4.1.1).

[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.

3.1.1. Book Prize Winners

As a special feature, DeLiKo-2025@DNB contains all digitally available novels that have ever been longlist-nominated for the German Book Prize (*Deutscher Buchpreis*) since its establishment in 2005. This set of 364 books is searchable exclusively, by defining a virtual subcorpus (see Section 4.1.1) and has been compiled with the support of the German Publishers and Booksellers Association, organizer of the award, in celebration of the 20th anniversary of the award in 2024. This subcorpus of distinguished literature could, e.g., be analyzed for features that apparently appeal to literary criticism – the juries choosing the novels for the longlist mostly consist of literary critiques – or could be contrastively analyzed for distinctive language features.

3.2. Conversion and Annotation

To make the conversion into high-quality TEI-XML encoded corpora feasible, we ignored PDF ebooks and limited our focus to the books available in the XML-based EPUB format and started off with 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 we added all 364 digitally available longlisted titles from the past two decades to the corpus. As a third step, in 2025, we repeated the conversion process, with improved conversion routines for the entire set including all remaining EPUB ebooks from the DNB published until the end of 2024, resulting in a total of 287,595

books. 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 (Deutsche Nationalbibliothek, 2025) 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 et al., 2012). In subsequent steps, the TEI-XML data was converted to KorAP-XML format (Bański and Diewald, 2025), using the korapxml tool (Kupietz et al., 2026) and annotated for POS and lemma using the TreeTagger (Schmid, 1994), for POS and morphosyntactic properties using MarMoT (Müller et al., 2013), and for dependencies using MaltParser (Nivre et al., 2007). These tools were selected because of their good balance between accuracy and performance.[2] If required for specific applications, additional annotation layers can be added later, since KorAP supports any number of them.[3]

The entire conversion and annotation process was completed within 14 days on a Linux server at the DNB 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.[4] 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 section 4.1.1 and Kupietz et al., 2010; Kupietz, 2016, for a detailed account of this *primordial sample* approach).

The source code of the conversion pipeline is available on GitHub at `https://github.com/KorAP/deliko`.

[2]Annotating the whole corpus with the very accurate UDPipe 2.0 (Straka, 2018) would, for example, take 1.6 years and block two GPUs for that time. An additional POS and lemma annotation with spaCy (Honnibal et al., 2020), however, would take less than four days on the current hardware and would thus be feasible.

[3]Multiple classifications also allow for maximizing recall or precision (Belica et al., 2011).

[4]See the genre table in `https://github.com/KorAP/deliko/blob/main/xslt/epub2i5.xsl`

4. Using DeLiKo@DNB

4.1. Web UI

DeLiKo@DNB is freely accessible through the website `https://korap.dnb.de/` (see Figure 3), utilizing the full range of KorAP features, such as flexible metadata-based definitions of virtual subcorpora, multiple annotation layers, and complex annotation searches in six supported query languages, simplified by a query assistant and query-by-match functionalities.

4.1.1. Virtual Subcorpora

An essential feature of KorAP is the ability to let users define stratified subsamples, so-called virtual subcorpora, by applying any logical combination of filters to the available metadata fields (see Figure 3). For instance, one can create a virtual subcorpus of horror fiction published between 2010 and 2019, or a subcorpus of all books written by a specific author. The German Book Prize longlist nominees can be selected by adding the constraint `award=buchpreis` via the corpus composition editor (see Figure 4). There is also a help page available with pre-defined, useful virtual corpora.[5]

4.2. API and client libraries

For research applications that require programmatic access, DeLiKo@DNB can be accessed via the KorAP REST API.[6] The API supports all functionalities of the web interface, including complex queries, multiple query languages, and virtual subcorpus definitions. To make access as easy as possible, KorAP provides client libraries for R[7] and Python[8] (Kupietz et al., 2020).

The example script in Listing 1 compares the attributive adjective collocates of 'Herz' (heart) in horror fiction with those in romance fiction, using the R client library. The results are shown in Table 1 Also, the composition chart in Figure 1 was produced using the R client library.[9]

5. Challenges and Ongoing Work

We released the first version of DeLiKo@DNB publicly in October 2024 after only 6 months of development and testing, instead of working for an additional year or more to release a "fully polished

[5]reachable via Help → Useful Virtual Subcorpora

[6]documented at `https://github.com/KorAP/Kustvakt/wiki`

[7]`https://cran.r-project.org/package=RKorAPClient`

[8]`https://pypi.org/project/KorAPClient/`

[9]`https://github.com/KorAP/deliko/blob/main/scripts/DeLiKoCompositionChart.R`

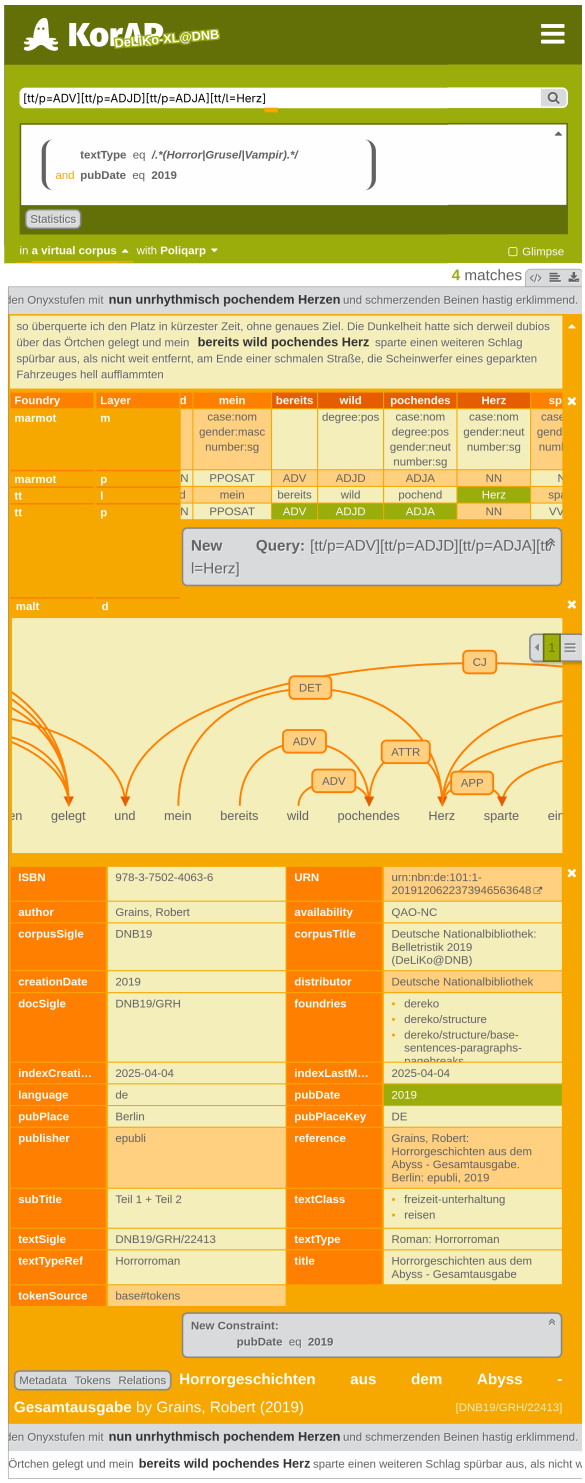

Figure 3: KorAP query in DeLiKo@DNB for the sequence of an adverb, an adverbial adjective, an attributive adjective and the lemma 'Herz' (heart), according to the TreeTagger (`tt`), in a virtual corpus restricted to genres horror, gothic, and vampire fiction, as well as a publication date in 2019. For the second query match, the token and dependency annotations as well as the book metadata are expanded. In the token annotation, query-by-match/example is shown, and in the metadata section, corpus-by-match is shown.

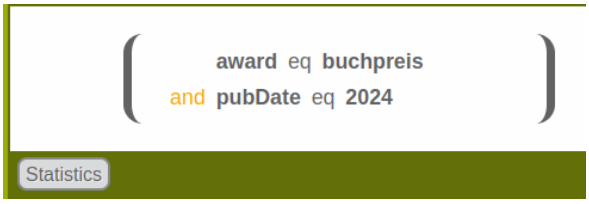

Figure 4: Definition of a virtual corpus of the German Book Prize longlist nominees, published in 2024, in the corpus composition editor of KorAP.

```
library(RKorAPClient)
library(tidyverse)
KorAPConnection("https://korap.dnb.de") |>
 collocationAnalysis(
  "focus([tt/p=ADJA] {Herz})",
  vc = c(
   'textType = /.*Liebes.*/',
   'textType = /.*(Horror|Grusel|Vampir).*/'
  ),
  leftContextSize = 1,
  rightContextSize = 0
 ) |>
 select(collocate, label, logDice) |>
 arrange(desc(logDice)) |>
 head(10)
```

Listing 1: Complete example R script comparing attributive adjective collocates of 'Herz' in horror versus romance fiction. The first argument of the collocationAnalysis function defines the query, searching for attributive adjectives according to the TreeTagger (`tt`) annotion immediately preceding 'Herz', marked by curly braces within the focus function. The vc-argument defines the two virtual corpora by applying regular expressions to the textType metadata field. The leftContextSize and rightContextSize arguments restrict the collocate search to words immediately preceding the focus. The results are sorted by logDice score (Rychlý, 2008) and limited to the top 10 results.

collocate	text type / genre regex	logDice
schlagendes	/.*(Horror|Grusel|Vampir).*/	10.34
pochendes	/.*(Horror|Grusel|Vampir).*/	10.15
gutes	/.*Liebes.*/	9.76
ganzes	/.*Liebes.*/	9.64
klopfendes	/.*(Horror|Grusel|Vampir).*/	9.57
gebrochenes	/.*Liebes.*/	9.55
tiefste	/.*Liebes.*/	9.48
weiches	/.*Liebes.*/	9.42
schwaches	/.*(Horror|Grusel|Vampir).*/	9.18
klopfendes	/.*Liebes.*/	9.05

Table 1: Result of the R script in Listing 1: Top 10 attributive adjective collocates of 'Herz' (heart) in horror, gothic and vampire versus in romance fiction, according to their logDice scores. Note that the collocates are linked to corresponding KorAP queries in the corresponding virtual corpora, so that users can directly check the underlying concordances.

final version". The reasons for this were threefold: First, we didn't want to withhold this valuable resource for any longer than necessary; second, we plan to extend DeLiKo regularly, anyway; and third, we were convinced that in this case, it makes more sense to treat the corpus, including the software to access it, which is essential in this case, as a living resource that needs to be continuously improved and maintained, rather than as a static snapshot that is "finished" once and for all. This is, of course, a commonplace in software development, but still not so much in the field of corpora and language resources. Accordingly, the challenges described in the following sections — together with our solution approaches and current work in progress — should be seen as integral parts of an ongoing development process that is inseparable from DeLiKo itself, and one we invite wider user and contributor communities to discuss with us.

5.1. Metadata

The metadata for books in the corpus were obtained via the DNB SRU API (Deutsche Nationalbibliothek, 2025) using the ISBNs encoded in the EPUB files. Since this metadata is manually curated by the DNB, it exhibits a relatively consistent structure (Deutsche Nationalbibliothek, 2021). However, it is not highly detailed, as it is based on publisher-provided information and is encoded in simple Open Archive Initiative (OAI) Dublin Core format (OpenAIRE, 2012). Consequently, certain elements such as translators are only encoded through conventions within the element text rather than through formalized markup (Dublin Core Metadata Initiative, 2005). Other information, such as the year of first publication or original publication date, is entirely absent.

This gap is manageable for close-reading applications, where researchers can consult external catalogues as needed. It is likewise less critical for corpus-linguistic studies that treat translations as independent texts — apart from translation effects such as shining through (Teich, 2003) — provided the translation status is explicitly recorded. For comparative literary studies and other distant-reading approaches, however, reliable translation and original publication year metadata are essential, because mislabelled translations and misleading publication years can quietly skew analysis results.

A separate problem concerns original publication dates. Many ebooks represent second or later editions of works first released years earlier; without metadata that distinguishes these cases, diachronic analyses can be distorted. To mitigate both shortcomings, we plan to enrich the curated records with estimated translation labels and original publication years (or year spans) by submitting the existing metadata to an ensemble of large language models. In a pilot study involving 1,000 titles that could not be resolved through Wikidata queries, DeepSeek Chat, Gemini 2.5 Flash-Lite, Claude 3 and 4.5 Haiku, and GPT-4o Mini were used to generate classifications. The results indicate that applying LLM-based inference can substantially improve the completeness and accuracy of the metadata.

As shown in Table 2, among the 500 books marked as translations in the DNB metadata via the name of the translator, 94.0% (470) were assigned an original publication year by at least one LLM, and 64.2% (321) received at least two agreeing years. We will continue to develop this approach, as it shows strong potential to enhance the overall metadata quality and expand the application scope of DeLiKo@DNB.

In addition, we intend to improve the genre classification, which currently relies only on string matching over the metadata available from the DNB. Achieving a richer and more reliable genre inventory will, however, require classifiers that inspect longer text passages or book-level representations, making the task substantially more resource-intensive — both computationally and in terms of the legal safeguards.

5.2. Applications that Require Full Text Access

We are aware that full text access is sometimes essential, particularly for language technology applications. However, legal and ethical constraints prevent us from making the complete texts of books in DeLiKo@DNB available for download. To maximize the usability of DeLiKo@DNB while respecting the rights of copyright holders, we implement a graduated variant of the "put the computation near the data" approach, following the model established by DeReKo (Kupietz et al., 2022). This approach provides multiple levels of access: through the user interface, via the API, through extensions and plugins for the query engine, and ultimately, in the case of DeLiKo@DNB, through supervised full text access during organized on-site hackathons and workshops.

5.3. Technical Precautions Against Abuse

When making legally protected data available for use in a scientific context, preventing misuse of this data (for personal or commercial purposes) represents a particular challenge. When making copyright-protected texts available, this means preventing the reconstruction of complete texts in circumvention of the legal limitations the system allows. The goal when implementing mechanisms to ensure that, is always to strike a balance between:

translated	total	≥2 agreements	≥2 guesses	≥1 guesses
yes	500	64.2% (321)	84.0% (420)	94.0% (470)
not known	500	29.2% (146)	88.8% (444)	97.4% (487)
all	1000	46.7% (467)	86.4% (864)	95.7% (957)

Table 2: LLM ratings of and agreement on the original publication year of a random sample of 1,000 books in DeLiKo@DNB, stratified by translation status according to DNB metadata.

- *minimizing* the attack surface for abusive usage, while also
- *maximizing* the resource's utility for researchers.

This can be achieved by restricting access to identifiable users, where misuse cannot be ruled out but can be detected and sanctioned (as is the approach for, e.g., DeReKo). For DeLiKo, however, a low-threshold solution without the need for user authentication was preferred, which is why various mechanisms were considered and implemented that would at least require considerable effort and time to reconstruct the original texts. These approaches include, for example, gradually slowing down successive queries and signing matches to prevent direct access of successive text passages based on match positions. Like security vulnerabilities, these methods are continuously reviewed and improved.

5.4. DeLiKo Expansions

We aim to expand the corpus by incorporating newly published books, including the German Book Prize longlist nominees, on an annual basis. As the conversion and annotation pipeline is fully automated (see Section 3.2), this process can be executed with minimal manual effort. Persistence and reproducibility issues can be avoided, because we only add newly published books that can be pinned each to its publication date and thus, if needed, excluded through virtual-corpus restrictions (like `pubDate<2025`).

5.5. Transferability of the Approach to Other National Libraries

The National Library as Corpus approach is underpinned by two complementary legal bases in Germany: the Text and Data Mining exception, which is based on an EU directive but implemented differently across member states (Kamocki et al., 2018), and Digital Legal Deposit, which stems from national legislation. Digital legal deposit regulations vary considerably between countries, and even where such regulations exist, they are sometimes only voluntary, as is the case in the Netherlands and Italy (Roudik et al., 2018). Therefore, transferring the approach presented here to other national libraries is not readily possible. Nevertheless, the challenges that national libraries face in making their digital collections available for research are similar, and researchers' interests in accessing national libraries, too. We therefore hope that our approach and experiences can serve as inspiration for other national libraries to initiate similar projects, which would ultimately also enable valuable cross-linguistic research opportunities within initiatives like EuReCo (Kupietz et al., 2024). In addition, our work also provides a tested full-stack technical foundation, from the raw EPUB files to a search and analysis user interface, that other national libraries and language resource providers can re-use and adapt to their specific legal and organizational settings, without much effort.

6. Summary & Conclusions

With DeLiKo@DNB, we have created an unprecedented research resource: a freely accessible, linguistically annotated corpus of over 287,000 contemporary German-language fiction books containing nearly 21 billion words. This corpus fills a long-standing gap in language resources by making available the complete spectrum of contemporary German fiction – from literary award winners to everyday prose – enabling both corpus-linguistic and digital humanities research on scale.

Our approach demonstrates a scalable solution to the persistent legal and technical barriers that have historically prevented national libraries from sharing their digital collections as research infrastructure. By combining the legal foundation of digital legal deposit with the computational principle of "put the computation near the data," we have created a model that respects copyright while maximizing research access through the open-source KorAP platform deployed at the DNB.

The corpus is accessible through multiple pathways suited to different research needs: an intuitive web interface with powerful query capabilities, a comprehensive REST API, and client libraries for R and Python that enable programmatic analysis at scale. Users can flexibly define virtual subcorpora based on rich metadata, enabling stratified, task-specific research on precisely tailored datasets.

We have also identified the key challenges ahead – including metadata enrichment through LLM-based inference, improved genre classification, and technical safeguards against misuse – and outlined concrete approaches to address them. Importantly, we treat DeLiKo@DNB not as a finished product but as a living resource, continuously enriched with newly published books and subject to ongoing technical and methodological improvements.

Finally, while the specific legal framework enabling this approach is not universally transferable, our full-stack technical foundation and documented experiences provide a tested blueprint that other national libraries and language resource providers can adapt to their own legal and organizational contexts, potentially enabling valuable cross-linguistic research collaborations.

7. Bibliographical References

Piotr Bański and Nils Diewald. 2025. Dealing with multiple annotations. In Piotr Bański, Ulrich Heid, and Laura Herzberg, editors, *Harmonizing language data. Standards for linguistic resources*, volume 4 of *Digital Linguistics*, pages 169–200. De Gruyter.

Piotr Bański, 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)*, pages 2905–2911, Istanbul, Turkey. European Language Resources Association (ELRA).

Cyril Belica, Marc Kupietz, Harald Lüngen, and Andreas Witt. 2011. The Morphosyntactic Annotation of DeReKo: Interpretation, Opportunities and Pitfalls. In *Grammar & Corpora. Third international conference, Mannheim, Sept., 22-24 2009*, pages 451–469, Tübingen. Narr.

Deutsche Nationalbibliothek. 2021. Delivering metadata to the German National Library: Metadata core set for all types of publication. Technical report, Deutsche Nationalbibliothek, Leipzig, Frankfurt am Main.

Deutsche Nationalbibliothek. 2025. SRU Interface - Deutsche Nationalbibliothek.

Dublin Core Metadata Initiative. 2005. Guidelines for Encoding Bibliographic Citation Information in Dublin Core Metadata.

Alexander Geyken. 2007. The DWDS corpus: A reference corpus for the German language of the twentieth century. In Christiane Fellbaum, editor, *Idioms and collocations: Corpus-based linguistic and lexicographic studies*. Continuum, London.

Jim Gray. 2003. Distributed Computing Economics. Technical Report MSR-TR-2003-24, Microsoft Research.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Pawel Kamocki, Erik Ketzan, Julia Wildgans, and Andreas Witt. 2018. New exceptions for Text and Data Mining and their possible impact on the CLARIN infrastructure. In *CLARIN Annual Conference 2018, Proceedings. 8-10 October 2018, Pisa, Italy*, pages 39 – 42, Utrecht. CLARIN.

Paweł Kamocki. 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*, pages 237–258. Edward Elgar Publishing.

Annette Klosa, Marc Kupietz, and Harald Lüngen. 2012. Zum Nutzen von Korpusauszeichnungen für die Lexikographie. *Lexicographica*, 28:71–97. Place: Berlin.

Marc Kupietz. 2016. Constructing a Corpus. In Philip Durkin, editor, *The Oxford Handbook of Lexicography*, pages 62–75. OUP, Oxford.

Marc Kupietz, Piotr Bański, Nils Diewald, Beata Trawiński, and Andreas Witt. 2024. EuReCo: Not building and yet using federated comparable corpora for cross-linguistic research. In *Proceedings of the BUCC 2024: The 17th workshop on building and using comparable corpora*, pages 94–103, Torino, Italia. ELRA and ICCL.

Marc Kupietz, 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).

Marc Kupietz, Nils Diewald, and Eliza Margaretha. 2020. RKorAPClient: An R Package for Accessing the German Reference Corpus DeReKo via KorAP. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7015–7021, Marseille, France. European Language Resources Association (ELRA).

Marc Kupietz, Nils Diewald, and Eliza Margaretha. 2022. Building paths to corpus data: A multi-level

least effort and maximum return approach. In Darja Fišer and Andreas Witt, editors, *CLARIN. The Infrastructure for Language Resources.* deGruyter, Berlin.

Marc Kupietz, Peter Leinen, Nils Diewald, Philippe Genêt, Rebecca Wilm, Andreas Witt, and Rameela Yaddehige. 2025. National Library as Corpus: Introducing DeLiKo@DNB – a Large Synchronous German Fiction Corpus. In *DHd 2025 Book of Abstracts*. Zenodo.

Marc Kupietz, Harald Lüngen, Piotr Bański, and Cyril Belica. 2014. Maximizing the potential of very large corpora. In Marc Kupietz, Hanno Biber, Harald Lüngen, Piotr Bański, Evelyn Breiteneder, Karlheinz Mörth, Andreas Witt, and Jani Takhsha, editors, *Proceedings of the LREC-2014-Workshop Challenges in the Management of Large Corpora (CMLC2)*, pages 1–6. European Language Resources Association (ELRA), Reykjavik, Iceland.

Marc Kupietz, Harald Lüngen, Nils Diewald, Helge Stallkamp, Uyen-Nhu Tran, and Rameela Yaddehige. 2026. EuReCo, KorAP and DeReKo: Updates on Ingestion and Annotation Pipelines, Backend, Interfaces, Operation, and Corpora. In Piotr Bański, Dawn Knight, Marc Kupietz, Andreas Witt, and Alina Wróblewska, editors, *Proceedings of the 12th Workshop on the Challenges in the Management of Large Corpora (CMLC-12 2026)*. European Language Resources Association (ELRA), Palma de Mallorca, Spain.

Marc Kupietz, Harald Lüngen, Paweł 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).

Harald Lüngen and Christopher M. Sperberg-McQueen. 2012. A TEI P5 Document Grammar for the IDS Text Model. *Journal of the Text Encoding Initiative*, 3:1–18.

Andrew Kachites McCallum. 2002. MALLET: A Machine Learning for Language Toolkit.

Thomas Müller, 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*, pages 322–332, Seattle, Washington, USA. Association for Computational Linguistics.

Joakim Nivre, 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*, 13:95–135.

OpenAIRE. 2012. Use of OAI-DC.

Peter Roudik, Kelly Buchanan, Tariq Ahmad, Laney Zhang, Nerses Isajanyan, Nicolas Boring, Jenny Gesley, Ruth Levush, Dante Figueroa, Sayuri Umeda, Elin Hofverberg, Graciela Rodriguez-Ferrand, Clare Feikert-Ahalt, and Law Library of Congress (U.S.), editors. 2018. *Digital legal deposit in selected jurisdictions*. The Law Library of Congress, Global Legal Research Directorate, Washington, D.C.

Pavel Rychlý. 2008. A Lexicographer-Friendly Association Score. *Proc. 2nd Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN*, 2:6–9.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Milan Straka. 2018. UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Elke Teich. 2003. *Cross-Linguistic Variation in System and Text: A Methodology for the Investigation of Translations and Comparable Texts*. Mouton de Gruyter, Berlin.

Christian Weiß. 2005. Die thematische Erschließung von Sprachkorpora. *OPAL - Online publizierte Arbeiten zur Linguistik*, 2005(1).

8. Language Resource References

Deutsche Nationalbibliothek, Leibniz-Institut für Deutsche Sprache. 2025. *DeLiKo-2025@DNB*. Deutsche Nationalbibliothek, Leibniz-Institut für Deutsche Sprache, Deutsches Literaturkorpus DeLiKo, DeLiKo-2025@DNB.

Leibniz-Institut für Deutsche Sprache. 2025. *DeReKo-2025-I*. Leibniz-Institut für Deutsche Sprache, German Reference Corpus DeReKo, DeReKo-2025-I.