LREC 2026

Leveraging Derived Text Formats to Unlock Copyrighted Collections for Open Science (DTF) @ LREC 2026

Workshop Proceedings

Editors
Florian Barth, Keli Du, José Calvo Tello, Philippe Genêt, Piroska Lendvai, Christof Schöch, Thorsten Trippel

12 May 2026

Proceedings of Leveraging Derived Text Formats to Unlock Copyrighted Collections for Open Science (DTF) @ LREC 2026

©ELRA Language Resources Association (ELRA), 2026
These proceedings are licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

ISBN 978-2-493814-77-7

Preface

We are pleased to present the *Book of Abstracts* for the workshop **"Leveraging Derived Text Formats to Unlock Copyrighted Collections for Open Science"**, held in conjunction with **LREC 2026** in Palma de Mallorca.

As language resources grow in scale and significance across linguistics, digital humanities, and language technology, the research community continues to grapple with the challenge of restricted-access textual data — particularly collections encumbered by copyright, licensing terms, or privacy constraints. **Derived Text Formats (DTF)**, also referred to as *extracted features*, have emerged as a promising pathway for enabling scientific inquiry and best practices of Open Science while at the same time respecting restrictions imposed by copyright law. By transforming texts into structured, interpretable, but non-reconstructible representations, DTF open important opportunities for reproducibility, comparability, and transparency in research while upholding legal and ethical obligations.

This workshop brings together researchers, legal scholars, infrastructure developers, and standardization experts to explore the multifaceted landscape of DTF. The contributions reflect active work and community experience on topics such as:

- methodologies for creating and processing Derived Text Formats
- legal and ethical considerations surrounding derived data publication
- practical use cases across digital humanities, linguistic research, corpus linguistics, and NLP
- tools, workflows, and infrastructure supporting DTF-based research
- theoretical investigations and standardization efforts

The breadth of submissions demonstrates the increasing relevance of derived data across fields that rely on sensitive or proprietary textual sources. The discussions and presentations showcased in this volume not only illuminate current practices but also point toward the development of robust community standards and sustainable infrastructures that support open science in legally complex environments.

We thank all authors for their thoughtful contributions and the members of the programme committee for their careful and constructive reviews. We are equally grateful to the participants — onsite and online — whose engagement makes this hybrid event a dynamic forum for exchange. Finally, we acknowledge the LREC 2026 Organising Committee for their support in hosting this workshop.

We hope that the papers compiled in this volume will inspire continued collaboration and innovation in leveraging Derived Text Formats to responsibly and effectively broaden access to textual resources.

The Workshop Organizers
LREC 2026
Palma de Mallorca

Organizing Committee

- Florian Barth, University of Göttingen
- Keli Du, University of Trier
- José Calvo Tello, University of Göttingen
- Philippe Genêt, German National Library
- Piroska Lendvai, Bavarian Academy of Sciences and Humanities
- Christof Schöch, University of Trier
- Thorsten Trippel, University of Tübingen and Leibniz-Institut for the German Language (IDS)

Table of Contents

Workshop Program

Tuesday May 12, 2026

14:00–15:30 **Session 1: Overview**
Room: Room 9
Chair: Philippe Genêt

14:00–14:10 ***Welcome and Introduction***

14:10–14:30 *Derived Text Formats as Strategic Transformations of In-Copyright Materials to Support Open Science: A Survey*
Christof Schöch

14:30–14:50 *A Multi-dimensional Constrained Framework for Derived Text Formats*
Keli Du and Christof Schöch

14:50–15:10 *Legal implications of Derived Text Formats - a copyright perspective*
Gianna Iacino, Pawel Kamocki and Keli Du

15:10–15:30 *Revisiting Masking After Fifteen Years: Early Approaches to Non-Reconstructable Linguistic Data in the current context*
Georg Rehm, Thorsten Trippel and Andreas Witt

15:30–16:00 ***Break***

16:00–18:00 **Session 2: Applications**
Room: Room 9
Chair: Piroska Lendvai

16:00–16:20 *Multi-Label Text Classification of Derived Text Formats with DistilBERT*
Jennifer Ecker and Roman Schneider

16:20–16:40 *Training data generation for context-dependent rubric-based short answer grading*
Pavel Šindelář, Filip Prášil, Dávid Slivka, Christopher Bouma and Ondrej Bojar

16:40–17:00 *DUO_DE A1: An Annotated Corpus of Online Learning Material for Beginning Learners of German as a Foreign Language*
Jammila Laâguidi, Vitaliia Ruban, Ronja Laarmann-Quante and Anastasia Drackert

17:00–17:20 *Why Reconstructing Scrambled Texts Fails*
Keli Du and Christof Schöch

Tuesday May 12, 2026 (continued)

17:20–17:40 *DIN 19461: A National Standard for Derived Text Formats*
Thorsten Trippel, Florian Barth, Jose Calvo Tello, Keli Du, Philippe Genêt, Daniel Kurzawe, Peter Leinen, Piroska Lendvai, Christof Schöch, Andreas Witt and Arden Zimmermann

17:40–18:00 ***Final discussion and closing***

Derived Text Formats as Strategic Transformations of In-Copyright Materials to Support Open Science: A Survey

Christof Schöch
Trier Center for Digital Humanities, Trier University
Universitätsring 15, 54296 Trier, Germany
schoech@uni-trier.de

Abstract

Derived Text Formats (DTFs) are the result of a strategic transformation of textual materials that are protected by copyright in their original form, such that the resulting data is useful for computational analyses and can be openly shared following best practices of Open Science without infringing copyright law. This paper aims to provide insights into several key aspects of this concept that is closely related to concepts such as corpus masking, non-consumptive research and extracted features. The paper establishes the motivation for using DTFs, discusses several foundational aspects of the concept and practice, describes ongoing research on issues including copyright, reconstructibility, evaluation and standardization of DTFs, and concludes with a roadmap for future work on DTFs. In this way, this paper provides a broad but concise overview of work on DTFs as a contribution to Open Science practices, with a focus on work in the Digital Humanities.

Keywords: Derived Text Formats, Non-Consumptive Research, Extracted Features, Copyright, Open Science, Digital Humanities

1. Introduction: Why do we need Derived Text Formats?

Three valuable principles contribute to shaping the regulatory framework for research today: academic freedom, Open Science, and copyright and privacy law. The legally and in some cases constitutionally-protected principle of academic freedom means, among other things, that researchers are empowered to investigate (and teach) any domain they choose to and do so in the ways they deem useful or desirable (Menand, 1996; Levy, 2026). Principles of Open Science are designed to support best practices in research, such as collaboration, transparency, accessibility, interoperability, reproducibility, and re-usability (Whyte and Pryor, 2011; Burgelman et al., 2019; Lewis, 2020). They have become increasingly central to research in the Digital Humanities against the backdrop of the 'reproducibility crisis' (Peng, 2015; Bausell, 2021; Gibney, 2022) as well as in the context of the FAIR principles for research data (Wilkinson et al., 2016). The legal framework of copyright is designed to foster innovation and creativity by granting a certain number of (moral and economic) rights to the creators of (textual or other) works (Rose, 1994; Goldstein and Hugenholtz, 2013). Finally, privacy laws are meant to enable individuals to control who is able to obtain, use, share and/or sell information regarding their identity and personal lives, and under what conditions (Bygrave, 2014; Voss, 2016), as implemented for example in the European Union's General Data Protection Regulation (GDPR, European Union, 2016).[1]

When the requirements from these three principles come together, researchers whose work relies on the computational analysis of textual or non-textual sources are forced to follow one of two strategies: Either they avoid copyright and privacy restrictions by investigating only public domain materials published before the contemporary period, that is, accept a limitation in the domains they can investigate in order to be able to practice Open Science; or they chose to investigate contemporary materials that often come with copyright and/or privacy-related restrictions, but then need to scale back their ambitions with respect to best practices of Open Science, essentially by keeping their research data locked away. DTFs are the result of a strategic transformation of in-copyright textual materials, such that these limitations are removed, thereby enabling researchers to investigate contemporary, in-copyright textual materials while both respecting copyright law and following best practices of Open Science.

This paper aims to provide a concise survey of foundational issues surrounding DTFs as well as

[1] While privacy laws are of course hugely important, this paper and most work on DTFs focuses on the copyright aspect of the issue. See, however, Greene et al. (2019) and Peloquin et al. (2020) on the implications of the GDPR for research as well as Altman et al. (2022), Joo and Kwon (2023) and Gadotti et al. (2024) on data anonymization for research in privacy-related contexts.

Proceedings of Leveraging Derived Text Formats to Unlock Copyrighted Collections for Open Science (DTF) @ LREC 2026, pages 1–15
12 May 2026. ©ELRA Language Resources Association (ELRA), 2026

recent and ongoing investigations of DTFs from the standpoints of copyright law, computational research, and infrastructure development. A particular focus is placed on the fields of Computational Literary Studies, Corpus and Computational Linguistics and Natural Language Processing in the wider context of the Digital Humanities, because researchers in these fields often work with in-copyright materials, such as large newspaper corpora or extensive collections of contemporary literary texts using computational methods of text annotation and analysis also known as text and data mining (TDM), and are directly concerned by the competing requirements of Open Science and copyright law.

In section 2 of this paper, several foundational aspects of DTFs are presented: alternative approaches to providing access to in-copyright materials (subsection 2.1), terms and concepts directly relevant to the approach represented by DTFs (subsection 2.2), prominent examples of DTFs (subsection 2.3), a process-oriented definition of DTFs (subsection 2.4) as well as several typologies of DTFs (subsection 2.5). In section 3, several areas of recent and ongoing scholarly work on DTFs are discussed: the fundamentals of copyright law as they apply to DTFs (subsection 3.1, the evaluation of various kinds of DTFs for specific research questions (subsection 3.3) as well as efforts to establish regulatory, technical and infrastructure-related standards (subsection 3.4). The paper concludes with an assessment of where we stand today and what a roadmap for scholarly work on DTFs holds for the future (section 4).[2]

2. Foundational aspects of Derived Text Formats

Having established the need not just for the provision of access to, but also for the ability of sharing materials protected by copyright, we can now turn to several foundational aspects of strategies mobilized to achieve this.

2.1. Strategies for providing access to in-copyright materials

Before turning to approaches that – conceptually, if not terminologically – fall into the class of DTFs themselves, it is worth considering several alternative approaches of providing access to materials under restrictions related to copyright and/or licensing.

Online platforms. Many databases of textual materials, such as linguistic corpora based on newspaper articles or other contemporary sources, are available online for searching and querying. The query results are usually provided in the shape of keyword-in-context views with limited context. The set of results is sometimes confined to a random sample rather than the full set. Further limitations of this approach include that they usually provide only a fixed set of query routines, do not permit merging of corpora from multiple, independent sources, do not enable custom annotation routines on the data, hinder the deployment of custom algorithms on the data, and make transparent and open sharing of results and reproducibility difficult.

This approach is useful primarily for rather simple and predictable usage scenarios of corpora. The close intertwining of platform, data (both raw text and annotations) and analytical procedures is both a conceptual weakness (because the principle of a separation of concerns is not respected) and a considerable risk in terms of sustainability (because the underlying data cannot be made available in simple, static forms independently of the platforms that are usually rather maintenance-intensive and resource-hungry. Examples of such platforms are COSMAS of the Leibniz-Institut für deutsche Sprache, providing access to the DeReKo (Deutsches Referenz-Korpus, see Kupietz et al., 2018), the DNB's korap system with access to DeLiKo-XL (Deutsches Literatur Korpus, see Jannidis et al. (2026)), or ATILF's (Analyse et Traitement Informatique de la Lange Française) Frantext database (Montémont, 2020), each providing access to large or very large corpora of historical and contemporary language materials.[3]

On-site access. In this approach, access to in-copyright materials owned or licensed by a particular institution are made available to users in a closed-room scenario on the physical premises of the institution. Usually, terminals provide access to the dataset but do not permit copying or otherwise transferring the materials. In contrast to the online querying approach, the on-site access approach enables much more sophisticated research scenarios: Access to the full data is possible, deployment of custom annotation schemes and advanced analysis scenarios is possible. Again, however, it is not usually possible to combine the datasets available on-site with other, third-party datasets. The main limitation of this approach, however, is obviously the need for researchers to be physically-present at a particular site to work with the data. The same requirement is also true for all collaborators and peer reviewers, more generally implying a strong limitation on transparency, accessibility, and reproducibility of the work performed. An example of

[2] All references are available online at https://www.zotero.org/groups/6473556/.

[3] See: https://korap.dnb.de/ and https://cosmas2.ids-mannheim.de/cosmas2-web/ as well as https://frantext.fr.

an institution offering this approach is the German National Library (DNB).

(Data / storage) capsules. The Hathi Trust Research Center (HTRC) has developed a strategy that aims to combine the advantages of the online platform approach with those of the onsite access approach, but without the need for a researcher to actually be on site, leveraging the idea of 'storage capsules' (also called 'data capsules', initially proposed by Borders et al., 2009, see also Wang et al., 2019): "The HTRC Data Capsules provides the virtual machine with two modes: a maintenance mode during which a user can access the network and install software freely, but cannot access copyrighted data; and secure mode where copyrighted texts become accessible to the user while the network access and file system access is highly constrained" (Zeng et al., 2014, 10). The advantages are obvious, but the infrastructural and technical demands remain high compared to DTFs and similar approaches where the data can be shared freely.

Sampling. Another approach is pursued by the XSamples group (Andresen et al., 2023), whose fundamental idea is to leverage the fact that copyright law allows for parts (specifically, 15%) of works to be copied, used and shared in a research and teaching context even outside provisions of the Text and Data Mining exception in European copyright law (see subsection 3.1 below). Based on this idea, the authors design an infrastructure that provides the appropriate amount of samples from a dataset to any one researcher, based on queries they can run on a suitable platform.

The approaches described so far all fundamentally aim to control how users can interact with unmodified in-copyright materials. DTFs and similar approaches, by contrast, aim to provide open accessibility to and allow unconstrained interactions with data that has been modified for this purpose.

2.2. The terminological space around Derived Text Formats

There are a number of competing terms circulating, all fundamentally describing the same idea, but with varying focus points or from differing perspectives. The most influential ones are documented and compared in this section.

Corpus masking. A pioneering and foundational proposal for using transformed texts rather than document-level metadata in order to avoid copyright restrictions is corpus masking. This proposal came from Corpus and Computational Linguistics, where researchers often work with contemporary texts that, in addition to being in-copyright, are often subject to licensing contracts. Also, such corpora often include multiple layers of linguistic or other annotations provided by different people or institutions and with varying degrees of copyright protection (Lehmberg et al., 2008). After initially experimenting with Treebank corpora (Rehm et al., 2007b), the authors went on to propose a more general framework. The term corpus masking places the focus on a masking operation, where the annotation layer(s) – such as morphological or syntactical annotations – are preserved and published, but the underlying lexical content layer – that is, the corresponding word forms – is replaced by placeholder tokens (Rehm et al., 2007a, 2026).

A related approach recently proposed by Arnold and Jäschke (2026) also separates the content and (stand-off) annotation layers, but additionally relies on a partially-masked version of the original full text that is too sparse to allow reconstruction, but informative enough to allow researchers to merge publicly available annotation layers and independently-obtained in-copyright texts layers.

(Non-consumptive / non-expressive) (use / reading / research). This is a family of terms used to underline the fact that in large-scale corpus analyses, whether platform-based or not, texts are not actually read and intellectually assimilated by any person, but algorithmic analysis processes are run on these texts at scale to identify complex features and patterns (Schreibman, 2014; Bhattacharyya et al., 2015; Kamocki, 2018; Samberg and Hennesy, 2019; Layne-Worthey, 2024; Baudry, 2023; Gruber and van Atteveldt, 2025; Zeng et al., 2014). This term sometimes encompasses or implies strategies such as the one discussed as 'online platforms' or 'data capsules' (subsection 2.1).

Extracted Features. This is an influential term mostly used in the context of the Hathi Trust Research Center's Extracted Features dataset, underlining not so much the usage but the creation of such datasets, namely by identifying and counting various kinds of features in texts – such as the number of typographical lines or the counts of all word types, per page – and making that descriptive data available (Bhattacharyya et al., 2015; Jett et al., 2020; Organisciak and Downie, 2021).

Finally, **Derived Text Formats** (in German: abgeleitete Textformate). A term introduced by Schöch et al. (2020b) that places the focus on the derivative nature of the resulting datasets, without specifying the nature of the transformation process or the kind of use being made of the data. As the most general term, this is the one preferred in this paper (see also Schöch et al., 2020a; Raue and Schöch, 2020; Genêt et al., 2025; Trippel et al., 2026; Du and Schöch, 2026a,b; Iacino et al., 2026; Ecker and Schneider, 2026.

DTFs avoid many of the limitations of the alternative approaches, in particular those connected to the need for an interactive platform. While their production and provision does have important infras-

tructural components and requirements, once DTFs have been produced and published, they are fundamentally low-maintenance, static datasets that researchers can download and work with in any way they find appropriate.

This does not mean that DTFs do not have limitations of their own. One key limitation is related to the fact that researchers do not, when working with DTFs, have access to the full, readable, original text that the DTF is based on. Currently, any DTF is designed to strike a strategic balance between the optimal obfuscation of any copyright-related features, on the one hand, and the best possible enabling of interesting and relevant research questions that can be investigated using the DTF, on the other hand. Investigating this trade-off from both a legal and a computational perspective is the object of considerable current work discussed in subsection 3.2 and subsection 3.3.

2.3. Existing Derived Text Formats

The basic idea of using not the original, full texts for research, but some proxy that stands in for them, is of course nothing new and may also be practiced for reasons other than copyright restrictions (such as privacy concerns or simply lack of full texts). A rather strong version of this strategy is using (document-level) metadata describing texts without considering their textual content at all. Examples include the use of rich qualitative and quantitative metadata (Paige, 2020), book titles in multiple languages (Patras et al., 2021) or library catalogue data (Fischer and Jäschke, 2022). In fact, one may argue that many DTFs (statistical DTFs, in particular) are simply very precise token-level metadata, in the sense that they consist of detailed information about the frequencies, distribution, co-occurrences or morpho-syntactic similarity of textual features, rather than the texts themselves.

A relatively early and very well-known example of a DTF is the **Google Ngram Viewer Dataset** (Ngram-Dataset, 2020), first published in 2009 (see also Lin et al., 2012).[4] The basic idea of the Ngram Dataset is to aggregate very large amounts of texts by year and then count the occurrences of ngrams of various sizes, allowing for an aggregated view of the rise and fall words and expressions over the years in the Ngram Viewer. Readability, recognizability and reconstruction are clearly excluded, because ngram data is aggregated by years and languages, not by individual works.[5]

While not primarily intended as a solution to copyright issues, the DLINA group's **Zwischenformat** (engl.: 'intermediary format') can also be understood as an early form of a DTF (Kampkaspar et al., 2015). It is suitable for dramatic texts encoded in XML-TEI and replaces the text contained in each scene by simple statistical information regarding the number of speeches and the number of words spoken by each character present in the scene. This information is sufficient for many analyses, in particular for network analysis, which often primarily relies on structural features.

The DTF most used in the context of Digital Humanities is probably the Hathi Trust Research Center's **Extracted Features** dataset (EF2.5, 2025) (see also Jett et al., 2020). It is also one of the more carefully-designed and well-documented such datasets, including through a JSON-LD schema.[6] The data is described as a "derived dataset consisting of metadata and data elements extracted from volumes in the HathiTrust Digital Library" (Jett et al., 2020). The data format is JSON, with metadata at the volume and page levels, and with descriptive statistics (e.g. number of lines and words) and token-level (word form and POS tag) frequency information encoded for each individual page. This dataset has been used to great effect by researchers in Digital Humanities (Underwood, 2014; Piper, 2022a; Sobchuk and Beheim, 2025), with a focus on investigating particular research problems, rather than explicitly or primarily evaluating performance compared to original full texts.

Recent work in the wider context of two projects, the consortium *Text+* in the framework of the German National Research Data Infrastructure (NFDI) and the research project *Forschen mit Derivaten* (engl.: 'Doing Research with Derivatives') have been concerned with three types of DTF used in various studies by the authors involved (Kocula, 2021; Du and Schöch, 2024; Du et al., 2025; Du and Schöch, 2026b):

- **Chunk-based term-document matrix**: Here, the original full text is tokenized, optionally receives some level of linguistic annotation, and is then split into chunks of equal (or nearly-equal) token sizes. The frequencies of all types in each chunk are then established and organized in a tabular format, that is a term-document matrix. The key parameter of this DTF is the chunk size.

[4] This resource became infamous in Digital Humanities circles not just because it is based on the Google Books Corpus of copyright settlement fame (Matulionyte, 2016; Borghi and Karapapa, 2011), but also because it was used in a highly ambitious and controversial paper using the Google Ngram Viewer to perform so-called 'Culturomics' (Michel et al., 2011).

[5] The Google Books Ngram Viewer is available at https://books.google.com/ngrams/.

[6] See Bhattacharyya et al. (2015); Jett et al. (2016); Downie (2015) and https://schemas.hathitrust.org/EF_Schema_v_3.0.

- **Chunk-based randomization of token order**: Again, the original full text is tokenized, optionally receives some level of linguistic annotation and is then split into chunks of equal (or nearly-equal) token sizes. However, instead of establishing frequencies, the token order is then randomized within each chunk, while the order of the chunks in the text remains intact.[7]
- **Selective replacement of word forms by POS**: To produce this DTF, the original full text is tokenized and annotated at least with the POS tag information. Then, a predefined proportion of randomly-selected word forms is replaced by their corresponding POS tag. The key parameter here is the replacement proportion (e.g. causing a slight irritation at less than 5% or massively interfering with readability at 40% or more).

There are many other datasets that can be understood as DTFs, including datasets that combine several types of DTF. One example of this latter type is the dataset about language use in pop culture (Songkorpus, 2022) published by Roman Schneider as accompanying data to his study on this domain (Schneider, 2022). This dataset contains word form and lemma frequencies, n-gram frequencies as well as a GloVe embedding model based on in-copyright materials. Another example is the CONLIT dataset (CONLIT, 2022) that contains a large range of derived data – from POS bigram and bookNLP supersense frequencies to – describing 2,700 contemporary books, both fiction and non-fiction (see Piper, 2022b).

What emerges primarily from this brief overview is that researchers have shown a large amount of creativity with respect to copyright-safe and useful DTFs: Corpus annotations in XML, ngram statistics or randomized texts in large CSV files, scene-level statistics in XML-TEI, extracted features in JSON, to name a few. However, while each of these formats has its justification, each project or data provider appears to have developed their own formats and strategies, with limited concern for standardization, community consensus or re-usable pipelines (a notable exception being the HTRC's Workset Ontology; see Jett et al., 2016). This is what recent and ongoing efforts in the German DH community aim to remedy.

2.4. Defining Derived Text Formats

What is common to the terms discussed in subsection 2.2 is that they all describe a strategy that enables work on in-copyright materials while respecting both the principles of Open Science and the rules of copyright law. Indeed, these approaches can be described as strategic transformations of original full texts that pursue a dual goal: on the one hand, removing or obfuscating any features of the texts that make them subject to copyright; and on the other hand, maintaining as much information as possible so that the resulting data remains useful for the investigation of one or several research questions.[8]

The authors of the emerging DIN standard on Derived Text Formats (see subsection 3.4 for context) describe this aspect of DTFs as follows: "The focus of the standard lies on identifying how enrichment and information-reduction operations produce derived formats that remain analytically useful while preventing reconstruction of the original text in ways that could infringe legal or ethical constraints" (Trippel et al., 2026). In practice, this generally means subjecting the original texts to several procedures that can be understood as both an enrichment and a reduction of information (for details, see Schöch et al., 2020b; Trippel et al., 2026).

The **enrichment** of texts means making information explicit that is implicit in the text and understood by human readers, but may not be directly accessible to machines. For example, this could mean adding information to tokens about their lemma, their part of speech, whether or not the token is a named entity, or whether or not a token is part of direct speech (such as a quote in a newspaper article or character speech in a novel). In many cases, such enrichment can only be performed on the original full texts, because the process relies on linguistic structure and contextual information.

The **reduction** of information can take the form of (selective) retention or deletion, replacement, removal or randomization. For example, a certain proportion of tokens, or a certain class of tokens, may simply be deleted. Or they may be replaced, that is either masked (i.e., replaced by a placeholder token) or replaced by information at a different (and often more abstract) level of linguistic analysis (i.e., a word form could be replaced with its corresponding POS tag). Finally, removing information can also involve randomization, e.g. removing the sequence information for tokens by randomizing their order in a document or within each of several segments of a document.

Importantly, such operations of enrichment and reduction can operate at various levels of linguis-

[7] Strictly speaking, these two DTFs contain the same information, provided that the annotations and the chunk sizes are identical. In practical terms, however, the latter format is a lot more similar to the original, annotated text.

[8] This is in line with the initial definition of DTFs in Schöch et al., 2020b: "We propose derived text formats as a solution: here, copyrighted textual materials are transformed in such a way that copyright-relevant features are removed, but that the use of various relevant methods of TDM remains possible."

tic description, in particular the character, token, n-gram, sentence, paragraph, chunk of arbitrary length, section, or work level. In addition, a given type of DTF typically has parameters, such as the proportion of tokens to be replaced by their POS tag, or the size of the chunks within which the word order is randomized.

2.5. Types of Derived Text Formats

Given the large range of theoretically possible (see subsection 2.4) and already existing (subsection 2.5) DTFs, researchers have proposed typologies of DTFs that help better understand the range of options available.

In a first approximation, Schöch et al. (2020b) have distinguished between token-based and corpus-based DTFs. In their typology, token-based DTFs rely on the manipulation of the original full text primarily at the level of the tokens (that can be enriched, deleted, replaced or see their order randomized), while the unit of production and publication is typically the individual document or work. By contrast, corpus-based DTFs identify, represent and/or count features such as ngrams (sequences of characters or words of a fixed length) or static word embeddings across multiple documents within or across a corpus.

Recently, Iacino et al. (2025) have distinguished three fundamental types of DTFs: statistical, transformative and Language-Model-based. Statistical DTFs involve extracting descriptive textual features (such as tokens, n-grams, typographical lines, or paragraphs) along with statistical metadata (e.g., frequency, length, or sequence). An example is the Google Ngram Viewer Dataset (Ngram-Dataset, 2020). Transformative DTFs introduce controlled noise to original texts by altering word order or replacing words with part-of-speech tags or placeholders, rendering the texts less readable while preserving structural and lexical information. An example is DTF600 (2025), containing texts with segment-wise randomized word order. Finally, language model-based DTFs utilize copyrighted texts to train models (such as topic models, word embeddings, or large language models), which encode textual information into algebraic vector spaces, enabling context-dependent semantic analysis or model fine-tuning for specific research tasks. Any language model can be considered a DTF in this sense.

Finally, Trippel et al. (2026) (in this volume) distinguish four types of DTFs: token-based (such as term-document-matrices or n-gram frequencies), vector-based DTFs (such as static or contextual embeddings), structured DTFs (such as formats based on shuffled segments) and multi-feature formats (combining several kinds of features, such as token statistics and embeddings).

It appears fair to say that the terminology is not yet entirely settled and that, as more such datasets are being published, there will be an opportunity to have another systematic view at the matter.

3. Recent and ongoing work on Derived Text Formats

Recent and ongoing research in particular in the Digital Humanities community has investigated legal issues around DTFs, both from a legal and an empirical perspective; has evaluated the usefulness of DTFs for specific research methods (or simply used them for research); and has concerned standardization efforts, both on a conceptual and on a technical level. This section details some of these efforts.

3.1. Derived text formats and copyright law

Chief among the legal issues is the question of how to determine whether a given DTF is actually 'safe' from the standpoint of copyright law, that is, whether the features that ground the copyrighted status of a text really have been removed or obscured, so that the data can safely be made openly and publicly available without infringing the copyright that protects the original, full-text version. In addition, the legal basis for producing DTFs in the first place is an important concern as well.

The introduction of the 'Text and Data Mining exception' for research into European copyright law in 2018/2019 (European Union, 2019) was an important adaptation of copyright to the digital age and a significant re-balancing between the interests of academics and those of copyright holders (Raue, 2017; Durantaye and Raue, 2020; Raue, 2022; Margoni and Kretschmer, 2022). In the absence of a doctrine comparable to 'fair use' in the USA and other jurisdictions, the TDM exception enables researchers in Europe to make copies of in-copyright materials without a limitation concerning the amount of the material, albeit under a certain number of conditions: their research is non-commercial in nature (note, however, that there is also a more narrow provision for commercial use cases); they have legal access to the in-copyright materials, for example through purchase or subscriptions; they intend to use these materials for the purposes of Text and Data Mining, as defined by law; and they do not openly share the full texts (data sharing is limited to the very narrow settings of direct project-based collaborations and quality checks by peers). Because of this last provision, in particular, which balances the freedoms provided to researchers against the legitimate interests of copyright holders, DTFs remain a vital strategy.

However, there is another aspect of the TDM exception that is relevant to DTFs. Because the creation of a DTF requires the creation of (albeit temporary) copies of these original texts, something which is not allowed under copyright law, it is important to ensure that there are provisions that allow the creation of DTFs in the first place. As a preparatory step of TDM, similar to cleaning, structuring and annotating texts, the creation of DTFs is covered by the TDM exception, as established by Iacino et al. (2025) and Iacino et al. (2026). However, the TDM exception requires but does not create legal access to the original full texts, so this needs to be established independently from the TDM exception.

Given that DTFs are still necessary, what legal requirements are there for DTFs in order to make sure they indeed do not infringe on the copyright holders' rights? Fundamentally, we can distinguish three aspects: readability, recognizability, and reconstructibility (Grisse, 2020; Jotzo, 2020; Iacino et al., 2025, 2026): a DTF should not allow people to read (and understand or enjoy) the text in the same way that they can read the original text;[9] a DTF should not allow people to recognize or experience the original ways in which the author of the original text used language to express their ideas and their individuality; and it should not be possible, at least not with trivial effort, to reconstruct the original full text from a DTF.

In essence, DTFs need to obfuscate or remove those aspects of the original texts in which their protection by copyright is grounded (Jotzo, 2020, 129), while making sure that the resulting DTF remains useful for research. The next section introduces various strategies that have so far been used to achieve this balance, or to identify this sweet spot.

3.2. Reconstructibility of Derived Text Formats

As shown in subsection 3.1, one of the key criteria for a suitable DTF is reconstructibility, that is, whether or not it is possible to reconstitute the original full text from one or several DTFs, and if it is possible, what effort and expertise are required (Grisse, 2020; Iacino et al., 2025). The degree of reconstructibility varies for each different type of DTF, but also with the specific parameters that were chosen when producing a particular implementation of that type of DTF.

Arguably, the deterministic reconstitution of information that has been removed (such as words forms) from the more abstract information that has been retained (such as POS tags) appears almost impossible. Similarly, randomization of word order is, in principle, an irreversible process. However, Large Language Models (LLMs) could in principle be a game-changer for this kind of task, albeit using fundamentally probabilistic, non-deterministic approaches such as vec2text or embedding inversion (Morris et al., 2023; Zhuang et al., 2024; Seputis et al., 2025). Several studies have, as a consequence, also leveraged LLMs for attempts to reconstruct various kinds of DTFs.

Work by (Kugler et al., 2024) has investigated under which conditions full texts represented as BERT-based contextualized word embeddings, a highly relevant kind of DTF that falls into the group of Language-Model-Based DTFs (in the typology by Iacino et al., 2025), can successfully be reconstructed. They conclude that the answer depends on the attack scenario: if the encoder used to produce the DTF is available, then reconstruction becomes feasible (albeit not without significant technical expertise and time); if, however, the encoder architecture is not known (and cannot itself be reverse-engineered), then any efforts of reconstructing the original texts remain futile.

Using a somewhat different approach, Du et al. (2025) have attempted to reconstruct one kind of transformative DTFs (in the sense of Iacino et al., 2025) using LLMs. Generally speaking, their findings show that while LLMs produce text based on DTFs that is somewhat more similar to the original texts than the DTF itself, in terms of an actual reconstruction of the original texts, the results are rather disappointing.

Finally, recent work by Du and Schöch (2026b) has provided a complement to performance-oriented investigations of the degree to which reconstruction of the original full text from DTFs is feasible. The authors have instead aimed to discover typical patterns of errors that occur when the reconstruction of original texts from DTFs is attempted using generative LLMs. They conclude that such reconstructed texts are typically shorter than the original texts, that LLMs tend to generate more general phrasings with missing modifiers than found in the originals, and that even when they use the same word forms, they often put them together quite differently, resulting in largely different meaning of generated texts.

For the time being, then, it does not appear feasible to reliably and easily reconstruct coherent sections of original texts from some of the more common types of DTFs. However, being unsuccessful in such reconstruction attempts using technologies available today is of course not the same as proving that reconstruction will not be feasible in a few years' time or that it is, in principle, impossible.

[9] This is related to the commercial value of a copyrighted text: a DTF should not serve as a replacement to the commercial offers that exist, that is, it should not infringe on the author's capacity to derive an economic benefit from their creation.

With respect to the legal status of DTFs, however, it also needs to be reiterated that a copyright infringement based on DTFs only happens if and when someone actually reconstructs significant portions of original texts outside of a research context (where the TDM exception would very likely cover such processes), not when someone publishes a DTF that, potentially and with significant effort, would enable someone else to perform such a reconstruction.

3.3. Evaluating Derived Text Formats' usefulness for research

Evaluating the usefulness for research of a given DTF is just as important as probing its reconstructibility. The primary way of investigating this issue is by evaluating the comparative performance of specific research methods or approaches on the original full-text versions and on one or several DTFs.

Kocula (2021) has done so using a collection of 129 British novels. When comparing topic coherence for the original corpus and several different DTFs, the author demonstrated that, as expected, LDA-based Topic Modeling, as a method based on a bag-of-words representation of texts, shows comparable performance for original full texts and DTFs based on term-document matrices or randomized word order. By contrast, performance drops substantially when using a DTF based on selective replacement of word forms by POS tags. Presumably, the lexical material becomes more and more impoverished the larger the proportion of replacements becomes.

A study by Du (2023) has used a similar setup to investigate the influence of the proportion of tokens replaced by their respective POS tag on stylometric authorship attribution. Using several corpora containing German dramatic texts, French novels, and English scientific prose, respectively, the author can show that attribution accuracy drops gradually and continuously as the proportion of POS-tags replacing word forms increases. This is good news, as an appropriate balance between usefulness for research and obfuscation of copyright-related features can in this manner be identified.

A somewhat different approach is used by Du and Schöch (2024). The authors used several DTFs based on two corpora, English-language movie reviews and German-language fiction. Rather than analyzing the performance of a classifier applied to DTFs, they used the DTFs to train a DistilBERT-based sentiment classifier and then evaluated the performance of the resulting classifiers on original full texts. Interestingly, they found a sweet spot of replacing 40% of tokens with their respective POS tags, a level where readability and reconstructibility are significantly hampered, while sentiment classification performance is essentially preserved.

Finally, the paper by Ecker and Schneider (2026) contained in this volume uses similar DTFs to investigate the accuracy of text classification using DistilBERT. Using two datasets and two perturbation strategies (POS-consistent token replacement at several rates, and word-order randomization), the results show that randomization reduced accuracy by 5%, while POS replacement reduced classification accuracy by 4–9%, with performance continuously declining as perturbation intensity increases. The authors make the highly interesting observation (in line with the findings by Du and Schöch (2024) described above) that models trained on perturbed DTF data generalize better to clean text than those trained on clean data, indicating that perturbation-based training fosters more robust representations.[10]

It becomes clear from these studies that the usefulness of a given DTF may vary widely depending on the specific method chosen, and it is important to understand the systematic relationship between the kind of method and the type of DTF (see Du and Schöch, 2026a). For example, any method that relies on a bag-of-word representation of text, such as stylometric authorship attribution, topic modeling, or certain kinds of sentiment analysis, will be entirely undisturbed by token sequence randomization, at least when it is performed chunk-wise. Conversely, any method that relies on word sequence or syntactic structure will be strongly affected by DTFs that do not preserve word order and/or syntactic annotation. Also, the usefulness of a given kind of DTF will vary strongly with certain key parameters of the DTF. The key takeaway here is that many methods are surprisingly robust against the introduction of moderate amounts of noise through replacement or randomization, though at higher rates of such noise, performance will suffer more significantly.

3.4. Standardization for Derived Text Formats

An important aspect of ongoing work in the context of DTFs is their standardization. This is an aspect that, as we have seen above in subsection 2.2, has largely been neglected in the early times of the discussion on datasets avoiding the limitations of copyright or privacy law. However, standardization of formats and documentation as well as the certification of pipelines that produce DTFs are essential:

[10] This issue, however, remains an area of active research in both DH and NLP; see Eder (2013); Hill and Hengchen (2019); Aepli and Sennrich (2022); Lendvai et al. (2025).

only if researchers have good reasons to trust that the DTFs they want to use contain exactly what they are supposed to contain will they be able and willing to actually use them for research.

A working group within the German NFDI consortium Text+ has recently developed a DIN (Deutsches Institut für Normung) standard for DTFs with national scope, DIN 19461:2026-04 E. The key contribution of this standard is to define concepts and a vocabulary for describing specific DTFs. It does so by specifying, on the one hand, a certain number of operations that can be performed on the original full texts to strategically transform the information contained in a DTF relative to the original full text. And it does so by defining, on the other hand, a certain number of levels of granularity at which such operations can be applied to the original text (see subsection 2.4). The key takeaway from this work is the requirement for anyone producing a DTF to document in detail all aspects of the production process in order to foster trust and reproducibility. For further details, see Trippel et al. (2026) in this volume.

Equally important are technical pipelines to derive DTFs from full texts as well as (to a lesser degree) platforms able to handle publication of DTFs. An example of a successful integration of DTFs as the output of a text processing pipeline is MONApipe (Barth et al., 2025). This is an NLP library built on spaCy (Honnibal and Montani, 2015–2026) that is designed to process literary texts in such a way that a considerable number of textual features that are of interest to scholars in Computational Literary Studies are identified and represented (e.g. temporal expressions, speakers and speeches, or events). In its latest iteration, this pipeline is also able to produce a certain number of DTFs.

It is true that the infrastructural requirements for making sets of DTFs accessible are usually not very different from publishing regular digital corpora, given that they are usually static files intended for download and offline use, not for in-platform processing. However, some specific provisions in terms of metadata and file storage are required. An example of a repository that has been enhanced to allow publishing of several types of DTFs in XML-TEI is the TextGrid Repository (Calvo Tello et al., 2024).[11]

4. Conclusion

4.1. Where we stand today

In the past five to ten years, Derived Text Formats and related strategies have developed from relatively isolated initiatives linked to individual data collections or institutions to a well-described strategic instrument in the context of Open Science and copyright and privacy law. At the same time, this strategy has become integrated within relevant legal, technical and infrastructural contexts, chiefly in the context of the NFDI consortium Text+.

During this time, Derived Text Formats have arguably also become more necessary than ever. Practices of Open Science, for instance around the open data movement and the FAIR principles, are increasingly gaining traction in the Digital Humanities, particularly but not only in the related fields of Computational Literary Studies, Corpus and Computational Linguistics and Natural Language Processing. At the same time, concerns around copyright and research have become highly visible within the research community. Debates concerning the copyright reforms in the European Union around the years 2018/2019, the high-stakes court cases and discussions around the issue of copyright law and training corpora used for generative AI as well as the question of rights applicable to products of generative AI have received considerable media attention. In this context, solutions like DTFs are becoming more and more timely and their broad adaptation ever more urgent.

The work done so far, including the considerable work represented in this volume, certainly shows that DTFs are a robust and reliable strategy that is an important element in the Open Science toolbox supporting accessibility, transparency, reproducibility, re-usability and sustainability of datasets and results in the Digital Humanities.

DTFs have also found their way into research practices, at least to some extent: beyond the publication by institutions such as libraries of larger datasets as DTFs for research, DTFs have proven useful as a way for researchers who work on in-copyright corpora to make as much data as possible available to others, to support transparency and reproducibility. Examples include publications such as Schneider (2022); Du et al. (2022); Sperling et al. (2024). However, copyright concerns are also frequently used to justify not making any data available and there is certainly still a lot of room for increased adoption of DTFs for this kind of data sharing attached to a particular publication.

4.2. A roadmap for DTFs

The current state of affairs, as described in this paper, points to a number of challenges for the future, both with respect to specific DTFs and with respect to the legal and technical environment in the age of Artificial Intelligence.

The most fundamental challenge for the next years is to foster the publication of very large sets of DTFs, for example by institutions holding large

[11] See: https://textgridrep.org/.

amounts of materials that are protected by copyright and are of interest for researchers in the Digital Humanities, i.e. first and foremost national libraries such as the German National Library (DNB). Making such collections available will be a game changer for research on the domains covered by these materials.[12]

Related to such institutional provision of research data, but at a different level, it is also important to continue to leverage DTFs for publication of ad-hoc datasets used in individual research papers, in order to increase their transparency and reproducibility. This is a roadmap item that primarily concerns the policy and community levels. For example, journals and conferences should be encouraged and supported in their efforts to develop data deposit policies that make use of DTFs, in order to better integrate such best practices into publication habits.

An important element of the roadmap for research into DTFs concerns reconstructibility, for several reasons: new specific DTFs are likely to emerge and consolidate from ongoing experimentation; increasing attention will likely be given to vector-based DTFs on the one hand, hybrid types of DTFs (combining properties of statistical and transformative DTFs), on the other; generative LLMs will increasingly be used for reconstruction tests and are likely to become better at this task over time, including through effects of memorization as training datasets continue to increase in coverage (often with little regard for copyright, or under broad interpretations of 'fair use', see Ahmed et al., 2026); and reconstruction from multiple DTFs of different kinds but created from the same full-text originals may become relevant. All these factors point to an increased importance of research into reconstructibility of DTFs, including developing useful measures to quantify the degree of reconstructibility of different DTFs. This area of future research includes further theoretical investigations into the very concept of DTFs, following the lead of several papers in this volume (Trippel et al., 2026; Du and Schöch, 2026a).

In this context, an important avenue of further research concerns the methods of information reduction or obfuscation. There are clear limitations to approaches relying on randomization (for example by shuffling the order of tokens within chunks of text) and/or on abstractive replacement (for example replacing a certain proportion of word forms with their respective POS tags). The degree of information reduction is often quite substantial, e.g. when 50% of the word forms are missing from a text, or when the information on token sequence becomes highly imprecise. In both these approaches, the twin goals of maximal usefulness for research and minimal reconstructibility are at odds with each other, forcing researchers into a trade-off scenario. In addition (as noted above), the feasibility of reconstruction risks will likely rise in the future, with expected further improvements in generative LLMs.

An alternative approach currently being investigated against this background at Trier University relies on a specific kind of encryption called homomorphic encryption that maintains token order and token identity while making the text unreadable to humans: it is demonstrably impossible to reconstruct the full text from such encrypted data, while the process is reversible if and only if the decryption key is known. This promises to open up entirely new avenues for both computational analyses and the interpretability of the results while effectively moving beyond the trade-off between usefulness and copyright-safety.

[12] This is all the more urgent and timely as the future of one of the largest such datasets, the HTRC's Extracted Features dataset, is currently unclear; see https://web.archive.org/web/20250503202415/https://www.hathitrust.org/press-post/plans-for-hathitrust-research-center/.

5. Acknowledgements

This work has been conducted in the context of two projects: One is the National Research Data Infrastructure (NFDI) consortium *Text+* (grant number 460033370). The NFDI is funded jointly by the Federal Republic of Germany and the 16 federal states through the German Research Foundation (DFG). The other is the project *Forschen mit Derivaten* (grant number 564393508) within the DFG funding programme 'Digitalisierung und Bereitstellung (noch) rechtebewehrter Objekte'. Many thanks to all the collaborators in these two projects for the many productive discussions and activities.

6. Bibliographical References

Noëmi Aepli and Rico Sennrich. 2022. Improving Zero-Shot Cross-lingual Transfer Between Closely Related Languages by Injecting Character-Level Noise. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4074–4083, Dublin, Ireland. Association for Computational Linguistics.

Ahmed Ahmed, A. Feder Cooper, Sanmi Koyejo, and Percy Liang. 2026. Extracting books from production language models.

Micah Altman, Aloni Cohen, Francesca Falzon, Evangelia Anna Markatou, Kobbi Nissim, Michel Jose Reymond, Sidhant Saraogi, and Alexandra Wood. 2022. A principled approach to defining anonymization.

Melanie Andresen, Markus Gärtner, Sibylle Hermann, Janina Jacke, Nora Ketschik, Felicitas Kleinkopf, Jonas Kuhn, and Axel Pichler. 2023. Vorzüge von Auszügen – Urheberrechtlich geschützte Texte in den digitalen Geisteswissenschaften (nach-) nutzen. *Zeitschrift für digitale Geisteswissenschaften*, 2022(7).

Frederik Arnold and Robert Jäschke. 2026. Sharing is Caring: A Text Alignment Approach for Sharing Annotations of Copyrighted Texts. In *New Trends in Theory and Practice of Digital Libraries*, pages 135–145, Cham. Springer Natur.

Florian Barth, George Dogaru, Tillmann Dönicke, and Mathias Göbel. 2025. Infrastructures for a Community-Developed Text Processing Library. *Selected Contributions of the 5th Conference for Research Software Engineering in Germany*, 85.

Julien Baudry. 2023. Les non-consumptive research uses des ressources numériques.

R. Barker Bausell. 2021. *The Problem with Science: The Reproducibility Crisis and What to Do About It*. Oxford University Press.

Sayan Bhattacharyya, Peter Organisciak, and J. Stephen Downie. 2015. A Fragmentizing Interface to a Large Corpus of Digitized Text: (Post)humanism and Non-consumptive Reading via Features. *Interdisciplinary Science Reviews*, 40(1):61–77.

Kevin Borders, Eric Vander Weele, Billy Lau, and Atul Prakash. 2009. Protecting Confidential Data on Personal Computers with Storage Capsules. In *18th USENIX Security Symposium, Montreal, Canada, August 10-14, 2009, Proceedings*, pages 367–382. USENIX Association.

Maurizio Borghi and Stavroula Karapapa. 2011. Non-display uses of copyright works: Google Books and beyond. *Queen Mary Journal of Intellectual Property*, 1(1):21–52.

Jean-Claude Burgelman, Corina Pascu, Katarzyna Szkuta, Rene Von Schomberg, Athanasios Karalopoulos, Konstantinos Repanas, and Michel Schouppe. 2019. Open Science, Open Data, and Open Scholarship: European Policies to Make Science Fit for the Twenty-First Century. *Frontiers in Big Data*, 2.

Lee Andrew Bygrave. 2014. *Data Privacy Law: An International Perspective*. Oxford University Press.

José Calvo Tello, Mathias Göbel, Ubbo Veentjer, Stefan E. Funk, Nanette Rißler-Pipka, and Keli Du. 2024. FAIR Derived Data in TEI and its Publication in the TextGrid Repository. *Journal of the Text Encoding Initiative*, 2024(18).

DIN 19461:2026-04 E. 2026. DIN 19461:2026-04 (e). Sprachressourcen und Sprachtechnologie - Abgeleitete Textformate (ATF).

J. Stephen Downie. 2015. The HathiTrust Research Center: Providing analytic access to the HathiTrust Digital Library's 4.7 billion pages. In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '15, page 5, New York, NY, USA. Association for Computing Machinery.

Keli Du. 2023. Understanding the impact of three derived text formats on authorship classification with delta. In *Open Humanities, Open Culture: 9. Jahrestagung des Verbands Digital Humanities im deutschsprachigen Raum (DHd2023)*. Zenodo.

Keli Du, Sarah Ackerschewski, Uygar Navruz, Nazan Sınır, Julian Valline, and Christof Schöch. 2025. Reconstructing shuffled text. bad results for nlp, but good news for using in-copyright texts. *Journal of Computational Literary Studies*, 4(1).

Keli Du, Julia Dudar, and Christof Schöch. 2022. Evaluation of measures of distinctiveness: Classification of literary texts on the basis of distinctive words. *Journal of Computational Literary Studies*.

Keli Du and Christof Schöch. 2024. Shifting Sentiments? What happens to BERT-based Sentiment Classification when derived text formats are used for fine-tuning. In *Digital Humanities Conference 2024: Book of Abstracts*, Lisbon. ADHO.

Keli Du and Christof Schöch. 2026a. A multidimensional constrained framework for derived text formats. In *Leveraging Derived Text Formats to Unlock Copyrighted Collections for Open Science (Workshop at LREC 2026)*.

Keli Du and Christof Schöch. 2026b. Why reconstructing scrambled texts fails: A structural analysis of reconstruction outputs. In *Leveraging Derived Text Formats to Unlock Copyrighted Collections for Open Science (Workshop at LREC 2026)*.

Katharina De La Durantaye and Benjamin Raue. 2020. Urheberrecht und Zugang in einer digitalen Welt – Urheberrechtliche Fragestellungen des Zugangs für Gedächtnisinstitutionen und die Digital Humanities. *RuZ - Recht und Zugang*, 1(1):83–94.

Jennifer Ecker and Roman Schneider. 2026. Multilabel text classification of derived text formats with distilbert. In *Leveraging Derived Text Formats to Unlock Copyrighted Collections for Open Science (Workshop at LREC 2026)*.

Maciej Eder. 2013. Mind your corpus: Systematic errors in authorship attribution. *Literary and Linguistic Computing*, 28(4):603–614.

European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data.

European Union. 2019. Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC.

Frank Fischer and Robert Jäschke. 2022. Ein quantum literatur. empirische daten zu einer theorie des literarischen textumfangs. In Fotis Jannidis, editor, *Digitale Literaturwissenschaft: DFG-Symposion 2017*, pages 777–812. Metzler, Stuttgart.

Andrea Gadotti, Luc Rocher, Florimond Houssiau, Ana-Maria Creţu, and Yves-Alexandre de Montjoye. 2024. Anonymization: The imperfect science of using data while preserving privacy. *Science Advances*, 10(29):eadn7053.

Philippe Genêt, José Calvo Tello, Florian Barth, Peter Leinen, and Christof Schöch. 2025. Abgeleitete Textformate: Die Chance für Bibliotheken und Wissenschaft zugleich [slides].

Elizabeth Gibney. 2022. Could machine learning fuel a reproducibility crisis in science? *Nature*, 608(7922):250–251.

Paul Goldstein and Peter Bernt Hugenholtz. 2013. *International Copyright: Principles, Law, and Practice*, 3rd ed edition. Oxford University Press, Oxford.

Travis Greene, Galit Shmueli, Soumya Ray, and Jan Fell. 2019. Adjusting to the GDPR: The Impact on Data Scientists and Behavioral Researchers. *Big Data*, 7(3):140–162.

Karina Grisse. 2020. Nutzbarmachung urheberrechtlich geschützter Textbestände für die Forschung durch Dritte – Rechtliche Bedingungen und Möglichkeiten. *RuZ - Recht und Zugang*, 1(2):143–159.

Johannes B. Gruber and Wouter H. van Atteveldt. 2025. Sharing is Caring (about Research): Three Avenues for Sharing (Protected) Text Collections and the Need for Non-Consumptive Research.

Mark J Hill and Simon Hengchen. 2019. Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study. *Digital Scholarship in the Humanities*, 34(4):825–843.

Matthew Honnibal and Ines Montani. 2015–2026. spacy: Industrial-strength natural language processing in python.

Gianna Iacino, Pawel Kamocki, and Keli Du. 2026. Legal implications of derived text formats – a copyright perspective. In *Leveraging Derived Text Formats to Unlock Copyrighted Collections for Open Science (Workshop at LREC 2026)*.

Gianna Iacino, Paweł Kamocki, Keli Du, Christof Schöch, Andreas Witt, Philippe Genêt, and José Calvo Tello. 2025. Legal status of Derived Text Formats – 2nd deliverable of Text+ AG Legal and Ethical Issues –. *RuZ – Recht und Zugang*, 2025(3):149–172.

Fotis Jannidis, Philippe Genêt, Leonard Konle, Marc Kupietz, Steffen Martus, Carolin Müller-Spitzer, and Samira Ochs. 2026. Empirische Untersuchungen zur Gegenwartsliteratur. Das Literatur-Korpus DeLiKo@DNB und erste Analysen. In *Digital Humanities im deutschsprachigen Raum 2026*, Vienna. DHd-Verband.

Jacob Jett, Boris Capitanu, Deren Kudeki, Timothy Cole, Yuerong Hu, Peter Organisciak, Ted Underwood, Eleanor Dickson Koehl, Ryan Dubnicek, and J. Stephen Downie. 2020. The HathiTrust Research Center Extracted Features Dataset (2.0).

Jacob Jett, Timothy W. Cole, Christopher Maden, and J. Stephen Downie. 2016. The HathiTrust Research Center Workset Ontology: A Descriptive Framework for Non-Consumptive Research Collections. *Journal of Open Humanities Data*, 2:e1.

Moon-Ho Joo and Hun-Yeong Kwon. 2023. Comparison of personal information de-identification policies and laws within the EU, the US, Japan, and South Korea. *Government Information Quarterly*, 40(2):101805.

Florian Jotzo. 2020. Der Schutz großer Textbestände nach dem UrhG – Die Nutzbarmachung fremder Textbestände für die Forschung. *RuZ - Recht und Zugang*, 1(2):128–142.

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.

Dario Kampkaspar, Frank Fischer, and Peer Trilcke. 2015. Introducing Our 'Zwischenformat'.

Martin Kocula. 2021. Volltext vs. abgeleitetes textformat: Systematische evaluation der performanz von topic modeling bei unterschiedlichen textformaten mit python.

Kai Kugler, Simon Münker, Johannes Höhmann, and Achim Rettinger. 2024. InvBERT: Reconstructing text from contextualized word embeddings by inverting the BERT pipeline. *Journal of Computational Literary Studies*, 3(1).

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)*. LREC.

Glen Layne-Worthey. 2024. Copyright Is the Lock; Non-Expressive Fair Use Is the Key: Research with In-Copyright Texts. In *The Routledge Companion to Libraries, Archives, and the Digital Humanities*. Routledge.

Timm Lehmberg, Georg Rehm, Andreas Witt, and Felix Zimmermann. 2008. Digital Text Collections, Linguistic Research Data, and Mashups: Notes on the Legal Situation. *Library Trends*, 57(1):52–71.

Piroska Lendvai, Uwe Reichel, Anna Jouravel, Achim Rabus, and Elena Renje. 2025. Instruction Finetuning to Attribute Language Stage, Dialect, and Provenance Region to Historical Church Slavic Texts. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 654–662, Varna, Bulgaria. INCOMA Ltd.

Jacob T. Levy. 2026. Conceptualizing Academic Freedom. *Annual Review of Political Science*.

Neil A. Jr. Lewis. 2020. Open Communication Science: A Primer on Why and Some Recommendations for How. *Communication Methods and Measures*, 14(2):71–82.

Yuri Lin, Jean-Baptiste Michel, Erez Aiden Lieberman, Jon Orwant, Will Brockman, and Slav Petrov. 2012. Syntactic Annotations for the Google Books NGram Corpus. In *Proceedings of the ACL 2012 System Demonstrations*, pages 169–174. Association for Computational Linguistics.

Thomas Margoni and Martin Kretschmer. 2022. A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology. *GRUR International*, 71(8):685–701.

Rita Matulionyte. 2016. 10 years for Google Books and Europeana: Copyright law lessons that the EU could learn from the USA. *International Journal of Law and Information Technology*, 24(1):44–71.

Louis Menand, editor. 1996. *The Future of Academic Freedom*. University of Chicago Press, Chicago.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331(6014):176–182.

Véronique Montémont. 2020. De Frantext 1 à Frantext 2: La cure de jouvence d'une vieille dame. *La lexicographie informatisée: les vocabulaires nationaux dans un contexte européen*.

John Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander Rush. 2023. Text Embeddings Reveal (Almost) As Much As Text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12448–12460, Singapore. Association for Computational Linguistics.

Peter Organisciak and J. Stephen Downie. 2021. Research access to in-copyright texts in the humanities. In Koraljka Golub and Ying-Hsang Liu, editors, *Information and Knowledge Organisation in Digital Humanities*, pages 157–177. Routledge.

Nicholas D. Paige. 2020. *Technologies of the Novel. Quantitative Data and the Evolution of Literary Systems*. Cambridge University Press.

Roxana Patras, Carolin Odebrecht, Ioana Galleron, Rosario Arias, J. Berenike Herrmann, Cvetana Krstev, Katja Poniž Mihurko, and Dmytro Yesypenko. 2021. Thresholds to the 'great unread': Titling practices in eleven eltec collections. *Interférences littéraires/Literaire interferenties*, 25:163–187.

David Peloquin, Michael DiMaio, Barbara Bierer, and Mark Barnes. 2020. Disruptive and avoidable: GDPR challenges to secondary research uses of data. *European Journal of Human Genetics*, 28(6):697–705.

Roger Peng. 2015. The Reproducibility Crisis in Science: A Statistical Counterattack. *Significance*, 12(3):30–32.

Andrew Piper. 2022a. Biodiversity is not declining in fiction. *Journal of Cultural Analytics*, 7(3):768.

Andrew Piper. 2022b. The CONLIT Dataset of Contemporary Literature. *Journal of Open Humanities Data*, 8(0).

Benjamin Raue. 2017. Text und Data Mining. *Computer und Recht*, 33(10):656–662.

Benjamin Raue. 2022. Text und Data Mining in Einrichtungen des Kulturerbes – Die neuen Möglichkeiten des § 60d UrhG n.F. aus Sicht von Gedächtniseinrichtungen. *RuZ - Recht und Zugang*, 3(1):4–18.

Benjamin Raue and Christof Schöch. 2020. Zugang zu großen Textkorpora des 20. und 21. Jahrhunderts mit Hilfe abgeleiteter Textformate – Versöhnung von Urheberrecht und textbasierter Forschung. *Recht und Zugang*, 1(2):118–127.

Georg Rehm, Thorsten Trippel, and Andreas Witt. 2026. Revisiting masking after fifteen years: Early approaches to non-reconstructable linguistic data in the current context. In *Leveraging Derived Text Formats to Unlock Copyrighted Collections for Open Science (Workshop at LREC 2026)*.

Georg Rehm, Andreas Witt, Heike Zinsmeister, and Johannes Dellert. 2007a. Corpus masking: Legally bypassing licensing restrictions for the free distribution of text collections. In *Digital Humanities 2007. Conference Abstracts*, pages 166–170, Urbana-Champaign. University of Illinois.

Georg Rehm, Andreas Witt, Heike Zinsmeister, and Johannes Dellert. 2007b. Masking Treebanks for the Free Distribution of Linguistic Resources and Other Applications. In *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories*, volume 1 of *NEALT Proceedings Series*, pages 127–138.

Mark Rose. 1994. *Authors and Owners: The Invention of Copyright*, 2. print edition. Harvard Univ. Press, Cambridge, Mass.

Rachael G Samberg and Cody Hennesy. 2019. Law and Literacy in Non-Consumptive Text Mining: Guiding Researchers Through the Landscape of Computational Text Analysis. In *Copyright Conversations: Rights Literacy in a Digital World*. UC Berkeley.

Roman Schneider. 2022. Zwischen Schriftlichkeit und Mündlichkeit: Songtexte in der deskriptiven Sprachforschung. *Sprachreport*, 38(1):38–50.

Susan Schreibman. 2014. Non-Consumptive Reading. In Naomi Segal and Daniela Koleva, editors, *From Literature to Cultural Literacy*, pages 148–165. Palgrave Macmillan UK, London.

Christof Schöch, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann, and Jörg Röpke. 2020a. Abgeleitete Textformate: Prinzip und Beispiele. *Recht und Zugang*, 1(2):160–175.

Christof Schöch, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann, and Jörg Röpke. 2020b. Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen. *Zeitschrift für digitale Geisteswissenschaften*, 5.

Dominykas Seputis, Yongkang Li, Karsten Langerak, and Serghei Mihailov. 2025. Rethinking the Privacy of Text Embeddings: A Reproducibility Study of "Text Embeddings Reveal (Almost) As Much As Text". In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, RecSys '25, pages 822–831, New York, NY, USA. Association for Computing Machinery.

Oleg Sobchuk and Bret Beheim. 2025. Does literature evolve one funeral at a time? *Proceedings of the Royal Society B: Biological Sciences*, 292(2040):20242033.

Dorothy Henriette Modrall Sperling, Mike Kestemont, and Vincent Neyt. 2024. The Authorship of Stephen King's Books Written Under the Pseudonym "Richard Bachman": A Stylometric Analysis. *Journal of Computational Literary Studies*, 2(1).

Thorsten Trippel, Florian Barth, Jose Calvo Tello, Phillipe Genêt, Piroska Lendvai, and Christof Schöch. 2026. Din 19461: A national standard for derived text formats. In *Leveraging Derived Text Formats to Unlock Copyrighted Collections for Open Science (Workshop at LREC 2026)*.

Ted Underwood. 2014. Understanding Genre in a Collection of a Million Volumes. White Paper Report, University of Illinois, Urbana-Champaign.

W. Gregory Voss. 2016. European Union Data Privacy Law Reform: General Data Protection Regulation, Privacy Shield, and the Right to Delisting. *The Business Lawyer*, 72(1):221–234.

Lun Wang, Joseph P. Near, Neel Somani, Peng Gao, Andrew Low, David Dao, and Dawn Song. 2019. Data Capsule: A New Paradigm for Automatic Compliance with Data Privacy Regulations. In *Heterogeneous Data Management, Polystores, and Analytics for Healthcare*, pages 3–23, Cham. Springer International Publishing.

Angus Whyte and Graham Pryor. 2011. Open Science in Practice: Researcher Perspectives and

Participation. *International Journal of Digital Curation*, 6(1):199–213.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018.

Jiaan Zeng, Guangchen Ruan, Alexander Crowell, Atul Prakash, and Beth Plale. 2014. Cloud computing data capsules for non-consumptiveuse of texts. In *Proceedings of the 5th ACM Workshop on Scientific Cloud Computing*, pages 9–16, New York, NY, USA. Association for Computing Machinery.

Shengyao Zhuang, Bevan Koopman, Xiaoran Chu, and Guido Zuccon. 2024. Understanding and Mitigating the Threat of Vec2Text to Dense Retrieval Systems. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP 2024, pages 259–268, New York, NY, USA. Association for Computing Machinery.

7. Language Resource References

CONLIT. 2022. *CONLIT*. Edited by Andrew Piper. Figshare. PID https://doi.org/10.6084/m9.figshare.21166171.v1.

DTF600. 2025. *600 French Novels in Derived Text Format*. Edited by Christof Schöch, Keli Du, and Julia Röttgermann. Beyond Words, 1.0. PID https://github.com/Zeta-and-Company/dtf600.

EF2.5. 2025. *The HathiTrust Research Center Extracted Features Dataset (2.5)*. Edited by John A Walsh et al. HathiTrust Research Center, 2.5. PID https://doi.org/10.13012/PXP0-F135.

Ngram-Dataset. 2020. *Google Books Ngram Viewer Dataset*. Google, v3. PID https://storage.googleapis.com/books/ngrams/books/datasetsv3.html.

Songkorpus. 2022. *Abgeleitete Textformate zu popkultureller Sprache*. Edited by Roman Schneider. Leibniz-Institut für deutsche Sprache (IDS). PID https://grammis.ids-mannheim.de/download.

A Multi-dimensional Constrained Framework for Derived Text Formats

Kei Du, Christof Schöch

University of Trier
Universitätsring 15, 54296 Trier
{duk, schoech}@uni-trier.de

Abstract

Derived Text Formats (DTFs) have been proposed as a solution to enable text and data mining while avoiding copyright infringement. Building on a review of recent empirical studies of DTFs on topic modeling, authorship classification, and sentiment analysis, this paper argues that DTFs should not be treated as static formats, but as variable and task-dependent representations shaped by multiple interacting factors. In response, we propose a multi-dimensional framework that conceptualizes DTFs as configurations within a structured space defined by both internal representation parameters and external constraints. The framework includes four internal representation dimensions—feature level, degree of reduction, transformation strategy, and aggregation level—as well as two external constraining forces: legal requirements and task-specific information needs. By emphasizing the interdependence of these dimensions, the proposed framework provides a systematic way to describe, compare, and design DTFs across different analytical contexts. Therefore, this paper contributes to a more theoretically grounded understanding of DTFs and offers guidance for their responsible and effective use in text and data mining in Digital Humanities.

Keywords: derived text formats, token-based DTFs, framework

1. Introduction

The practice of transforming texts into derived representations is well established in computational text analysis. However, the notion of derived text formats (DTFs), also known as extracted features, extends this practice by embedding it within a legal and infrastructural context, thereby shifting the focus from purely methodological considerations to the interplay between analytical utility and copyright compliance (Jett et al. 2020, Schöch et al. 2020). In recent years, driven by the principles of open science, an increasing number of researchers have been making their research data publicly available alongside their publications to ensure the reproducibility and falsifiability of their work. As a result, DTFs are getting increasing attention due to the increasing needs of the storage, publication, and reuse of research data built from in-copyright texts.

The definition of DTFs is constrained by conflicting requirements arising from the tension between legal compliance and text analysis. Legal requirements call for the reduction of textual information, while methodological needs rely on retaining such information; the tension between these two has not yet been systematically conceptualized or explained. Therefore, this paper aims to explore how to systematically describe and design DTFs within this context. In the following, we first explore the previous studies on DTFs in order to understand how different DTFs were applied in text and data mining. Then we share our observations and reflections on the use of DTFs and propose a multi-dimensional constrained framework to support the use of DTFs.

2. Previous evaluations of DTFs on text and data mining tasks

In recent years, a series of studies have evaluated DTFs' performance across various text and data mining tasks such as topic modeling, authorship classification and sentiment analysis.

Kocula (2022) systematically evaluated the performance of the Latent Dirichlet Allocation (LDA) algorithm across three DTFs: Term-Document Matrices (TDM), Segment-wise Abolished Sequence information (SAS), and Selectively Reduced Token information (TKN). This study conducted experiments using a corpus of 19th and 20th-century English novels and employed topic coherence (via the Palmetto library) and statistical significance tests to measure the quality of the generated topics. The results demonstrate that DTFs, particularly SAS and TKN, achieve topic coherence scores that are remarkably similar—and in some instances superior—to those of the original full text, suggesting that DTFs provide a robust, scientifically valid, and legally safe method for distributing research data.

Du (2023) investigated the impact of DTFs on authorship classification, with a particular focus on how information loss affects performance in stylometric analysis. The study examined three types of token-based DTFs, especially focusing on selectively replacing words in texts with their corresponding part-of-speech (POS) tags. To evaluate their effectiveness, experiments have been conducted using three corpora in three different languages. The experiment involves gradually replacing a certain proportion of the vocabulary in each text (from 0% to 100%) and measuring the resulting author classification

Proceedings of Leveraging Derived Text Formats to Unlock Copyrighted Collections for Open Science (DTF) @ LREC 2026, pages 16–19
12 May 2026. ©ELRA Language Resources Association (ELRA), 2026

performance. The results show that moderate levels of information loss (replacing or removing up to 40% of words in texts) have relatively little impact on classification accuracy. Interestingly, replacing words with POS tags does not lead to better results than simply removing them, suggesting that POS information contributes little to authorship discrimination in this context.

Du & Schöch (2024) investigated how much information loss a BERT-based model can tolerate before its sentiment classification performance significantly degrades. The authors conducted experiments using both non-literary and literary datasets and tested two DTFs: DTF-1 (randomized word order) and DTF-2 (POS Replacement). The results show that BERT is remarkably robust at picking up semantic "signals" despite the loss of syntax in scrambled text. Also, when 40% of words are replaced by their corresponding POS tags, the text becomes almost impossible for a human to read or recognize (satisfying legal safety), yet the sentiment classification accuracy remains nearly identical to the original text.

3. Reflections on the use of DTFs

By examining the decisions regarding the selection and use of DTFs in the studies mentioned above, we can identify the following characteristics of DTFs:

First, DTFs are not static; instead, they are flexible and adaptable text representations designed for different text and data mining tasks. Different tasks may require different DTFs. For example, authorship attribution only needs lexical information in text, while sequential information is necessary for training language models. When defining a DTF for a text and data mining task, a balance must be reached between, on the one hand, the legal regulations concerning copyright protection and the sharing of research data, and, on the other hand, the textual information required for the specific task. In other words, a DTF must establish trade-offs between preservation and deletion of textual information, that is between text recognizability and reconstructability, on the one hand, and analytical performance on the other.

Second, when determining a DTF, it is necessary not only to decide which types of textual information to retain or to delete, but also to determine the degree to which they should be retained or deleted. For example, we could randomly reorder 50% of the words in each sentence within a text, or replace 20% of the content words in the entire text with their corresponding POS tags. Of course, the extent to which such modifications or transformations are applied depends on the trade-offs mentioned earlier.

Third, when different DTFs are applied to the same task, if the results based on Format A are better than those based on Format B, this only indicates that Format A is more suitable for that task; it does not mean that Format A is a better DTF than Format B. This is because, during the process of transforming the same text into different DTFs, different pieces of information are filtered out of the text. For example, a document-term matrix keeps word frequency while discarding word order and syntactic relations, while a POS-based representation removes lexical content but retains aspects of grammatical structure. These transformations change the feature space on which computational models operate, as well as the statistical distribution and internal organization of the data. As a result, applying different DTFs to the same task is not simply a matter of using “more” or “less” text data for text mining; rather, it involves using different data to solve the same task. Performance differences cannot be simply attributed to methodological superiority, as these differences may reflect variations in how the underlying data is structured. Therefore, any evaluation of DTFs must consider the transformations and the information loss they introduce, rather than assuming that all derived formats are comparable simplifications of the same text.

Finally, a one-size-fits-all DTF is highly unlikely to exist, as different text and data mining tasks rely on different types of textual information. To establish a DTF capable of supporting as many tasks as possible, it is necessary to balance various types of textual information rather than optimizing solely for a single objective. This means the format should preserve diverse textual information—such as lexical, syntactic, and sequence features—while allowing for controlled information reduction. Such an attempt to address every aspect is likely to result in suboptimal performance across all tasks. Therefore, rather than pursuing a single universal format, it makes more sense to define DTF more flexibly to suit different analytical needs. However, if the same text is converted into different DTFs and all of them are made publicly available as research data, this increases the risk that the original text could be reconstructed using NLP technologies (such as large language models).

4. A multi-dimensional constrained framework of DTFs

Based on the above analysis, we believe it is necessary to establish a framework for describing DTFs so that users can take every important aspect into account when using DTFs. We consider DTFs to be the result of textual transformations shaped by mutually constraining conditions and introduce our multi-dimensional framework for DTFs.

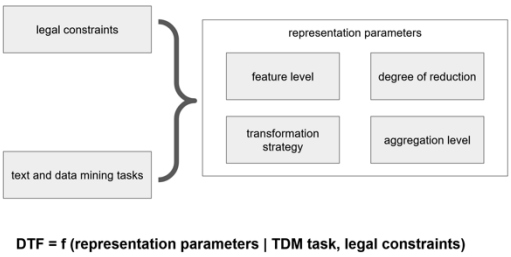

Figure 1. DTF as a multi-dimensional constrained framework

As presented in Figure 1, the graphical representation of the framework is at the top and the formula at the bottom defines DTF as a functional configuration of internal representation parameters that is dynamically shaped and restricted by two primary external forces: the legal constraints and the text and data mining tasks. The external forces serve as boundary conditions that restrict and guide the configuration of the representation parameters. The internal representation parameters have four components: feature level, degree of reduction, transformation strategy and aggregation level. Different internal parameters lead to different DTFs, each reflecting a specific balance between analytical utility and copyright compliance. Importantly, these dimensions are not independent of one another: choices made in one dimension affect the range of possible configurations in others. The following is a detailed explanation of each dimension.

- **Legal constraints**: This dimension covers the legal requirements regulating the legitimacy of DTFs, particularly those related to recognizability, reconstructability, and copyright compliance. It specifies the extent to which the original text content can be retained or reconstructed.
- **Text and data mining tasks**: This dimension refers to the specific information requirements of particular text and data mining tasks, such as authorship attribution, topic modeling, text re-use, or sentiment analysis. It determines which textual features must be kept ensuring the validity of the analysis.
- **Feature level**: This dimension refers to the selection of linguistic features kept in the DTFs, such as word forms, lemmas, POS tags, or semantic and syntactic relationships. It determines which textual information are available for analysis.
- **Degree of reduction**: This dimension quantifies the proportion of information that must be deleted or transformed in the original text. It defines the overall degree of transformation, ranging from minor modifications to total loss of information.
- **Transformation strategy**: This dimension describes methods for keeping, replacing, or deleting textual elements, such as random substitution, POS-based filtering, or building static and contextual embeddings. It determines how information loss is distributed within each text in a corpus.
- **Aggregation level**: This dimension indicates the structural level at which DTFs are constructed, ranging from token-level and sentence-level transformations to document-level or corpus-level representations. It defines how textual information is organized and interpreted.

5. Limitations

While the proposed multi-dimensional framework provides a systematic approach to defining and designing DTFs, several limitations remain.

First, although legal requirements are identified as a core constraint, the framework does not yet account for the specific statutory variations across different international jurisdictions, such as the differences between European copyright exceptions and US Fair Use.

Second, the current framework lacks quantitative metrics for measuring reconstructability and recognizability; while it defines the degree of simplification from a qualitative perspective, it does not provide a technical threshold to ensure irreversibility when confronting large language models.

Finally, the empirical evidence supporting this framework is primarily based on traditional token-based analysis, and further research is needed to validate its applicability to more abstract representations, such as high-dimensional embeddings or large-scale generative AI workflows.

6. Conclusion

This paper suggests that derived text formats (DTFs) should not be understood as static representations, but rather as flexible yet constrained configurations shaped by both text analytical and legal considerations. By examining the existing empirical research across various text and data mining tasks, we argue that differences in DTFs design fundamentally alter the structure of the underlying data. Consequently, performance differences cannot be explained in isolation from the transformation processes that generate them.

To this end, we propose a multidimensional framework that conceptualizes DTFs as contextual configurations. By distinguishing

between internal representation parameters and external constraints, this framework provides a systematic approach for defining, describing and comparing DTFs. It also emphasizes the dynamic nature of DTF construction, viewing it as a process shaped by competing factors rather than a static choice of format.

This framework provides a more robust theoretical foundation for understanding DTFs and highlights the trade-offs involved in their application. Future research could further refine this framework, explore its applicability to a broader range of text and data mining tasks, and investigate the risks regarding publishing different DTFs of the same text, particularly considering the advances in NLP technology and their potential implications for text reconstruction.

7. Acknowledgments

This work was created in the context of the work of the association German National Research Data Infrastructure (NFDI) e.V. NFDI is financed by the Federal Republic of Germany and the 16 federal states, and the consortium Text+ is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project number 460033370. The authors would like to thank for the funding and support. Furthermore, thanks also include all institutions and actors who are committed to the association and its goals.

Author contributions:

Keli Du: Conceptualization, Methodology, Investigation, Visualization, Writing - original draft, Writing - review & editing.

Christof Schöch: Funding acquisition; Supervision, Writing - review & editing.

8. Bibliographical References

Du, K. (2023). Understanding the impact of three derived text formats on authorship classification with Delta. DHd 2023 Open Humanities Open Culture. 9. Tagung des Verbands "Digital Humanities im deutschsprachigen Raum" (DHd 2023), Trier, Luxemburg. https://doi.org/10.5281/zenodo.7715299.

Du, K., & Schöch, C. (2024). Shifting Sentiments? What happens to BERT-based Sentiment Classification when derived text formats are used for fine-tuning. Digital Humanities Conference 2024 (DH2024), Washington, DC. https://doi.org/10.5281/zenodo.18161643.

Jett, Jacob, Boris Capitanu, Deren Kudeki, Timothy Cole, Yuerong Hu, Peter Organisciak, Ted Underwood, Eleanor Dickson Koehl, Ryan Dubnicek, and J. Stephen Downie (2020). The HathiTrust Research Center Extracted Features Dataset (2.0). DOI: http://doi.org/10.13012/R2TE-C227.

Kocula, M. (2021). Volltext vs. abgeleitetes Textformat: Systematische Evaluation der Performanz von Topic Modeling bei unterschiedlichen Textformaten mit Python. https://doi.org/10.5281/zenodo.5552487.

Schöch, Christof, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann, and Jörg Röpke (2020). “Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen”. In: Zeitschrift für digitale Geisteswissenschaften 5. DOI: http://doi.org/10.17175/2020_006.

Legal implications of Derived Text Formats – a copyright perspective

Ass. iur. Gianna Iacino, LL.M., Dr. iur. Pawel Kamocki, Dr. phil. Keli Du, et al.

Deutsche Nationalbibliothek, Leibniz-Institut für Deutsche Sprache, Trier Center for Digital Humanities
Adickesallee 1, 60386 Frankfurt/Germany, R 5 6-13, 68161 Mannheim, Ludwig-Weinspach Weg, 54296 Trier
g.iacino@dnb.de, kamocki@ids-mannheim.de, duk@uni-trier.de

Abstract

Text and Data Mining (TDM) methods are often used in order to analyse large amounts of text for scientific research. If the analysed text is protected by copyright, the use of such TDM methods has copyright implications. The existing copyright exceptions facilitate TDM within a narrow framework which limits the storage, publication and re-use of datasets. This paper examines the legal framework of converting the source text into a derived text format (DTF) which is no longer protected by copyright in order to allow the use of TDM without legal restrictions. First, the creation itself of a DTF is being examined: it entails copyright relevant acts which are covered by the TDM exception. In a second step the copyright status of the created DTF has to be evaluated based on three criteria: the DTF may not contain elements which are an expression of the intellectual creation of the author of the source material, the source material may not be easily reconstructable based on the DTF and the source material may not be recognizable.
Keywords: TDM, DTF, Copyright

1. Introduction

Digital Humanities research relies significantly on text corpora as foundational data. A key resarch method in this domain is Text and Data Mining (TDM). According to the legal definition in the Digital Single Market Directive (DSM Directive)[1], TDM refers to "any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations."[2]

Conducting TDM typically requires reproducing the source material, and in collaborative settings, sharing it with others or making it publicly available. However, such actions may infringe copyright if the source material is protected by copyright law. Unless a statutory exception applies, permission from rights holders is required.

Such statutory exceptions were introduced into EU law in the art. 3 and 4 of the DSM Directive, yet significant constraints persist—particularly regarding the publication, long-term storage and reuse of datasets derived from copyrighted texts.

Under the TDM exception for scientific research (art. 3 DSM Directive), source material may only be shared within a limited group of researchers for joint research or with third parties for quality assessment. Long-term storage is permitted only if the data were collected by cultural heritage institutions, research organizations, or individual researchers affiliated with such entities. These restrictions contradict the principles of open science, which are central to Digital Humanities. They hinder the reproducibility of results, limit the ability to build on prior work, and create barriers when dealing with currently copyrighted materials.

A potential mechanism to overcome these limitations are Derived Text Formats. By transforming copyrighted source material into formats that no longer contain protected content, DTFs may enable unrestricted storage, sharing and reuse of the material. This analysis focuses on the legal conditions for creating DTFs from copyrighted works under German law, as well as the criteria for determining whether DTFs themselves are protected by copyright.

2. Copyright-relevance of deriving DTFs

Certain acts are considered copyright-relevant acts by law. The 'Right of reproduction' according to art. 16 UrhG is one of those copyright relevant acts. It means the right to produce copies of the work, whether on a temporary or on a permanent basis and regardless by which means of procedure or in which quantity they are made.

[1] Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC, DSM Directive, https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32019L0790.

[2] See art. 2.2 DSM Directive.

Proceedings of Leveraging Derived Text Formats to Unlock Copyrighted Collections for Open Science (DTF) @ LREC 2026, pages 20–24
12 May 2026. ©ELRA Language Resources Association (ELRA), 2026

It is necessary to take a closer look at the creation process of a DTF: needless to say, practically the only convenient and feasible way to derive DTFs is by automated means. However, a piece of software used to derive DTFs necessarily copies (large) passages of source texts, even if only copyright-free information is extracted from the text. Even Copies that are only temporarily stored in the RAM are copyright-relevant.

The reproductions necessary to derive a DTF can therefore only be lawfully made with the permission from the right holders, unless they are covered by a statutory exception. Of course, asking for the rightholder's permission to derive a DTF is not a reasonable option, primarily for pragmatic reasons (e.g., multiple and/or unknown copyright holders). And of course: the very point of deriving a DTF in the first place is to be able to share meaningful information about the text without limitations and without having to ask for permission. It is therefore necessary to rely on an existing statutory exception for deriving a DTF.

Copyright law knows several statutory exceptions which might be applicable for the purpose of creating a DTF, due to the fact that they allow reproductions of entire works. In this context the most important exception is § 60d UrhG which allows reproductions necessary for TDM for scientific purposes.

Although first introduced into German law in 2018, § 60d UrhG, containing the exception for Text and Data Mining (TDM) for scientific research purposes, owes its current wording to Article 3 of the 2019 DSM Directive. It can be summarised as follows:

- Research organisations, cultural heritage institutions and citizien scientists[3]
- are allowed to make copies of content
- that they have lawful access to
- in order to carry out TDM
- for scientific research purposes.

The process of creating a DTF entails an automated analysis of the source material in order to generate a representation of a base text. This representation contains information about the source text. Producing a DTF is very similar in nature to applying methods of Text und Data Mining, with the exception that the process is not continued as it normally would, i.e. by summarizing the transformed data, by visualizing a selection of the data, and by drawing conclusions from it. Rather, it stops at an earlier phase. This should not be an obstacle to seeing the creation of DTFs as compatible with the legal definition of TDM and the DTF itself as a result of a TDM process.

Creating DTFs can just be a means, rather than a goal, of a TDM process. In this scenario, the information incorporated in the derived text format is not the goal of the scientific research. Rather, the derived text format serves as a means to an end: it shall now be used as a copyright-free corpus for a subsequent TDM analysis. It is therefore only an interim step in a larger process. This interim step of the TDM process already serves scientific research purposes: Scientific research generally refers to the methodical and systematic pursuit of new knowledge.[4] By already deeming the "pursuit" of new knowledge sufficient, not only the steps directly related to the acquisition of knowledge are included, rather, it is sufficient that the step in question is aimed at a (later) gain in knowledge. The creation of a dataset can be considered scientific research in this aforementioned sense. While the creation of a dataset itself may not yet be associated with knowledge gain, it is a fundamental step aimed at using the dataset for future insights.[5] The TDM regulations do not explicitly provide for a multi-step TDM process, but it is nevertheless covered by the wording as well as the intent and purpose of the regulation.

Therefore, all reproductions necessary to create DTFs are exempted under the TDM exception (§ 60 d UrhG).

3. Copyright-status of DTFs

The question that arises here is whether a DTF still contains copyright-protected content from the source material. In this case, the material continues to be protected in favor of the original author and can only be used within the limitations of copyright law. However, the goal is to create a DTF which no longer entails copyright protected content and can therefore, be used freely without

[3] The **beneficiaries** mentioned in the DSM directive are limited to research organisations (including universities and public research institutes) and cultural heritage institutions (libraries, museums, archives…). Going beyond the wording of the DSM Directive, the German transposition adds to the list "citizen scientists", i.e. researchers without academic affiliation, as long as they are acting for non-commercial purposes. This addition was possible due to a creative fusion of Art. 3 of the DSM Directive with the "general" exception for non-commercial scientific research in Art. 5(3)(a) DSM DIrective.

[4] BeckOK UrhR/Grübler, 42. Ed. 1.5.2024, UrhG § 60c Rn. 5; Dreier/Schulze/Dreier, 7. Aufl. 2022, UrhG § 60c Rn. 1.

[5] LG Hamburg, Urteil vom 27.9.2024, Az.: 310 O 227/23, Rz. 113, https://openjur.de/u/2495651.html.

any constraints of copyright law. It is therefore necessary to look at the following three criteria for the determination of copyright protection.

3.1 Are partial reproductions protected?

Under EU law, a partial reproduction is only protected if it contains elements reflecting the author's intellectual creation—i.e., meets the originality threshold. In Infopaq, the CJEU ruled that even short excerpts may be protected if they are original, but individual words are not.

In Germany, courts generally reject copyright protection for very short texts—such as slogans or brief phrases—due to insufficient originality (*Schöpfungshöhe*). The 2021 BGH ruling in *perlentaucher* confirmed that short snippets (knappe Wortfolgen) are not protected, especially when the summary is sufficiently distinct from the original. The court emphasized the need for "sufficient distance" between source and summary. While 11-word snippets were deemed potentially original in *Infopaq*, this is not a legal threshold. Shorter snippets are extremely unlikely to be protected, and their exclusion from a DTF poses negligible risk. Longer snippets, however, carry higher risk.

Thus, a DTF containing original snippets constitutes a partial reproduction—requiring either permission or a statutory exception. But for most practical purposes, the risk from very short snippets is negligible.

3.2 DTFs and the "reconstructability" criterion

The real challenge lies in ensuring that compilations of snippets do not reconstruct the original work's expressive core.

According to the CJEU snippets are to be regarded as partial reproductions if they are original or if their "cumulative effect (...) may lead to the reconstitution of lengthy fragments which are liable to reflect the originality [of the source text]"[6].

This condition, which can be referred to as "reconstructability", is instructive, but difficult to apply in the realm of language technology.

Reconstruction of source texts from DTFs may in fact be a more complicated task than it appears; in one experiment it was not possible to reconstruct even a very short source text after scrambling the word order.[7] In another, the successful reconstruction of text from a specific kind of language model-based DTF (a BERT-based contextual word embedding model) depended on the availability of the encoder used to build the DTF.[8]

Whether it is possible to do so, has to be evaluated separately for each individual DTF. Due to the constantly evolving technological possibilities, the answer to the question of reconstructibility of source texts is susceptible to changing over time. It appears that with a very large amount of effort (e.g., *ad infinitum* repetition of simple trial and error), DTFs can often be used to reconstruct source material. Therefore, when applying the criterion of reconstructability to DTFs, it appears sensible to restrict it to reconstructions possible with a "reasonable effort." However, currently copyright law does not contain such a standard, and it appears that every reconstructible copy, regardless of the effort invested in the reconstruction, remains an act of reproduction.

3.3 DTFs and the "recognisability" criterion

The third factor influencing the legal status of short textual snippets was recently established by the CJEU in the 2019 *Pelham* ruling. The Court held that the use of a very short sample from a phonogram—specifically, a 2-second rhythm sequence—in another phonogram constitutes an act of reproduction, "unless the sample is altered in a way that renders it unrecognizable to the ear."[9]

Formally, this decision applies only to Article 2(c) of the InfoSoc Directive, which concerns the reproduction rights of phonogram producers. It remains unclear whether this reasoning—particularly the "unrecognizable" threshold—can or should be extended to the broader scope of Article 2, including copyright and other related rights. The CJEU might interpret the concept of "partial reproduction" in copyright law in a manner consistent with the *Pelham-decision*. However, this does not imply that the *Pelham* test—based on recognizability—would supplant the

[6] CJEU, Infopaq, para 50.

[7] Keli Du (2024), "Rekonstruierbarkeit von abgeleiteten Textformaten", https://events.gwdg.de/event/607/contributions/1408/.

[8] Kai Kugler, Simon Münker, Johannes Höhmann, & Achim Rettinger (2023), "InvBERT: Reconstructing Text from Contextualized Word Embeddings by inverting the BERT pipeline", Journal of Computational Literary Studies 2(1), 1–18. doi: https://doi.org/10.48694/jcls.3572.

[9] CJEU, judgement of 29 July 2019, Pelham, Case C-476/17, ECLI:EU:C:2018:1002.

established criteria from *Infopaq*, namely originality and reconstructability.[10]

When applied to literary works, the concept of recognizability should not be understood as mere identification of a work or its metadata. Rather, it concerns the recognition of elements that reflect the author's creative expression or distinctive stylistic features. Thus, the use of a well-known literary character's name or the inclusion of publication details does not, by itself, satisfy the recognizability criterion.

4. Conclusions

Technically necessary acts of reproduction are required for the creation of a DTF. If the source material is protected by copyright, the creation of the DTF constitutes a copyright-relevant act. In the absence of the right holder's permission, the creation of the DTF must fall under a copyright exception in order to be lawful. DTFs can be created under the TDM exceptions (§ 60d UrhG). The DTFs can then be used as a basis for a subsequent analysis.

Whether the DTFs can also be *used freely*, made publicly available and stored for an unlimited amount of time, depends on the copyright status of these DTFs. The goal of creating a DTF which no longer contains copyright-protected content is achieved, if

- the DTF does not contain elements which are an expression of the creative individuality of the author of the source material,
- the source material cannot be reconstructed with trivial effort and
- the author's creative individuality is not recognizable.

There are many grey areas in examining the legal status of the many different forms of DTFs. In many cases, legal certainty cannot be achieved. By avoiding reproducibility and recognisability of source texts e.g. through avoidance of longer n-grams.

[10] M. Senftleben, "Flexibility Grave – Partial Reproduction Focus and Closed System Fetishism in CJEU, Pelham", IIC, 2020, 51, pp.751 – 769. See also K. Grisse, "Nutzbarmachung urheberrechtlich geschützter Textbestände für die Forschung durch Dritte – Rechtliche Bedingungen und Möglichkeiten", Recht und Zugang, 2020, 2, pp. 143–159.

5. Bibliographical References

Karina Grisse, “Nutzbarmachung urheberrechtlich geschützter Textbestände für die Forschung durch Dritte – Rechtliche Bedingungen und Möglichkeiten”, Recht und Zugang, 2020, 2, pp. 143–159.

Kai Kugler, Simon Münker, Johannes Höhmann, & Achim Rettinger (2023), “InvBERT: Reconstructing Text from Contextualized Word Embeddings by inverting the BERT pipeline”, Journal of Computational Literary Studies 2(1), 1–18. doi: https://doi.org/10.48694/jcls.3572.

BeckOK UrhR/Grübler, 42. Ed. 1.5.2024, UrhG § 60c Rn. 14.

Dreier/Schulze/Dreier, 7. Aufl. 2022, UrhG § 60c Rn. 1.

Keli Du (2024), “Rekonstruierbarkeit von abgeleiteten Textformaten”,

https://events.gwdg.de/event/607/contributions/1408/.

M. Senftleben, “Flexibility Grave – Partial Reproduction Focus and Closed System Fetishism in CJEU, Pelham”, IIC, 2020, 51, pp.751 – 769.

Revisiting Masking After Fifteen Years: Early Approaches to Non-Reconstructable Linguistic Data in the Current Context

Georg Rehm[*], **Thorsten Trippel**[**†], **Andreas Witt**[**]

[*]Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI)
Salzufer 15/16 D-10587 Berlin, Germany
Humboldt-Universität zu Berlin, Dorotheenstraße, 24, 10117 Berlin, Germany
georg.rehm@dfki.de

[†]University of Tübingen, Keplerstraße 2, D-72074 Tübingen, Germany
thorsten.trippel@uni-tuebingen.de

[**]Leibniz Institute for the German Language, R 5, 6-13, D-68161 Mannheim, Germany
{trippel, witt}@ids-mannheim.de

Abstract

This paper revisits corpus masking approaches introduced in 2007 for enabling the distribution of linguistically annotated corpora without exposing copyrighted or sensitive source texts and situates them within the contemporary framework of Derived Text Formats (DTFs). While the original work demonstrated how syntactic and morphological information could be preserved through parameterised masking, today's landscape, which is shaped by large language models, FAIR requirements, and emerging standardisation efforts, demands more formalised, robust and reproducible methods. We outline how DTFs extend early masking concepts by introducing explicit abstraction levels, reversibility classes, and machine-actionable provenance, supported by standards such as TEI, ISO linguistic annotation models, CMDI metadata, and the draft DIN DTF specification. Building on these foundations, we present a modern workflow for DTF generation, including enrichment pipelines, structural abstractions, statistical and embedding-based representations, and non-reversible transformation layers, illustrated through the MONA-pipe framework. A range of linguistic, digital-humanities and NLP use cases demonstrates the analytical utility of DTFs while maintaining legal compliance. We conclude that DTFs constitute a sustainable and infrastructure-ready solution for open, reproducible and legally secure text-based research in the decades to come.

Keywords: Derived Text Format, Masking, Linguistic Resources

1. Introduction

Research in linguistics and the digital humanities relies on empirical data. Text corpora are the backbone of many studies, enabling quantitative analyses, comparative investigations, and structural observations across a wide range of linguistic phenomena. Despite their central role in scholarly work, these corpora are often subject to legal restrictions that severely limit their distribution and reuse or even contractual restrictions that remain in force regardless of added annotation layers.

In 2007, these challenges led to early explorations of masking techniques, which aimed to separate linguistic annotations from protected textual content to enable data sharing in a legally compliant way. These initial approaches laid the conceptual groundwork later taken up and systematised within today's Derived Text Format (DTF) frameworks.

Now, almost 20 years later, the core problems remain, but the context has changed dramatically. Machine learning and large language models (LLMs) have heightened concerns about partial reconstruction of masked data, increasing the need for robust, non-reversible transformation methods. At the same time, the scientific community places greater emphasis on reproducibility, transparency, and open science, creating new incentives to standardise such transformations. Initiatives like the European Open Science Cloud (EOSC), the German National Research Data Infrastructure (NFDI) and the emergence of a DIN DTF standard (DIN 19461, currently in draft status) demonstrate that the ideas first explored in 2007 have now become part of a broader, coordinated effort to enable the lawful and sustainable use of text data in research.

This paper revisits and extends the early approaches in light of these developments. We examine how masking-based techniques can be integrated into modern DTF frameworks, how they respond to today's legal and technical challenges, and how they can support reproducible research.

2. From Masking to Derived Text Formats: A Historical Perspective

The idea of making annotated linguistic resources without distributing the underlying textual content has its origins in early discussions within compu-

Proceedings of Leveraging Derived Text Formats to Unlock Copyrighted Collections for Open Science (DTF) @ LREC 2026, pages 25–33
12 May 2026. ©ELRA Language Resources Association (ELRA), 2026

tational linguistics and the digital humanities. In 2007, two complementary approaches were proposed that would later become foundational for what is now conceptualised as DTFs.

The first, introduced at the 6th International Workshop on Treebanks and Linguistic Theories (TLT 2007), (Rehm et al., 2007b)) focused on *masking treebanks*. In the same year, a more general framework was presented at Digital Humanities 2007. This *corpus masking* (Rehm et al., 2007a) approach explored the possibility of legally bypassing licensing restrictions by systematically obfuscating the source text while preserving the annotation layers.

Both papers were motivated by the same underlying tension: while annotations constitute independently copyrightable intellectual contributions of researchers, their dissemination was limited by the rights attached to the underlying source texts. The masking approaches offered an early strategy to keep annotations usable and shareable even when the legal conditions made the distribution of the original source texts impossible.

Looking back, these works anticipated several themes that have since become central to research data infrastructures. They recognised the *importance of sustainability*, the *separation of data and annotation layers*, and the *need for flexible licensing models*. They also outlined distinctions between legal regimes and research uses that are now standard in infrastructures such as Text+ (Hinrichs and Trippel, 2024) and the National Research Data Infrastructure (NFDI) in Germany, and European initiatives such as CLARIN (Fišer and Witt, 2022), META-SHARE (Piperidis et al., 2014), European Language Grid (Rehm, 2023) or European Language Data Space (Rehm et al., 2024).

At the time, however, these approaches remained largely conceptual. The field lacked formal standards, consistent metadata vocabularies, and infrastructural support for transformations of this kind. As a result, the masking work of 2007 was forward-looking but mostly confined to experiments with specific corpora and local workflows.

From today's perspective, the significance of these early approaches has become more clear. The rise of large-scale machine learning, increasing legal scrutiny, and the emergence of formalised DTF standards have created a landscape in which the ideas of 2007 are not only relevant but foundational. The conceptual separation between text and annotation, and the possibility of distributing transformed, non-reversible representations, laid the groundwork for many of the practices and standards currently being developed.

2.1. Masked Treebanks (TLT 2007)

The work presented at TLT 2007 (Rehm et al., 2007b) introduced one of the earliest systematic attempts to address the legal barriers surrounding the distribution of annotated corpora. The masked treebank approach demonstrated that many linguistic research questions rely primarily on syntactic and morphological structure rather than on lexical content. By replacing word forms with neutral placeholders while keeping constituent hierarchies, dependency relations and grammatical features intact, the method ensured analytical usefulness without exposing protected content. Different parameter settings allowed researchers to control how much morphological or formal information was preserved. This combination of structural fidelity and legal safety made masked treebanks one of the earliest viable strategies for sharing annotated resources under restrictive conditions.

2.2. Corpus Masking (DH 2007)

The second major contribution from 2007, presented at the Digital Humanities conference (DH 2007) (Rehm et al., 2007a), generalised the idea from treebanks to any XML-annotated corpora. The proposed *corpus masking* framework applied a parameterised randomisation algorithm that replaced surface forms with automatically generated strings while preserving the annotation layers. Depending on the configuration, selected formal properties – such as token length, case patterns or affix information – could be retained to balance legal constraints with analytical usefulness. This made it possible to redistribute annotated corpora under more permissive licensing conditions and supported a range of use cases, including unlexicalised parsing, NLP evaluation, teaching scenarios and sustainability efforts where the underlying texts could not be shared.

Many of the principles introduced in 2007 – such as parameterisation, structural preservation, non-reversibility and the separation of text and annotation – anticipate the requirements of today's DTFs. Contemporary work shows that these transformations enable legally compliant text and data mining while supporting reproducible research, provided that protected content is sufficiently abstracted (Schöch et al., 2020). Empirical studies further demonstrate that even strong obfuscation methods can retain substantial analytical value while preventing reconstruction (Du et al., 2025). As a result, DTFs have developed from early experimental masking approaches into infrastructure-ready components for sustainable data stewardship and long-term reusability.

3. New Challenges in 2026

Although the early masking approaches anticipated many of the ideas that now underpin DTFs, the

landscape in which such transformations operate has changed profoundly over the past fifteen years. Advances in machine learning – most notably in LLMs – have created new risks related to the reconstruction and inference that were not conceivable at the time. At the same time, the growing emphasis on transparent and reproducible research has increased the demand for datasets that can be shared, cited and reused without infringing on copyright or privacy laws. These developments intersect with a complex legal environment shaped by copyright, contractual restrictions and data protection regulations, all of which impose increasingly stringent requirements on the handling of textual data. Together, these factors create a new constellation of challenges that modern DTFs must address, extending far beyond the assumptions and technical constraints of 2007.

3.1. LLMs and Reconstruction Risks

Modern machine learning models are capable of inferring missing lexical or semantic material from sparse cues. Even when surface forms have been fully removed, LLMs can exploit preserved structural patterns, token-level regularities or distributional signals to reconstruct plausible fragments of the original text or to approximate its content. Du et al. (2025) point out that currently the reconstruction is far from perfect, but caution needs to reign to avoid legal implications, especially if the quality of transformation further increases.

A masked corpus that appears non-reversible to a human analyst may not be non-reversible to a state-of-the-art model trained on vast amounts of data. As a consequence, masking is no longer sufficient on its own; it must be accompanied by a *formalised assessment of non-reversibility and information leakage*. Such assessment requires evaluating the transformation against realistic attack models that reflect modern inference capabilities, rather than assuming that the removal of surface forms is intrinsically safe.

These developments highlight the need to reconsider masking not merely as a technical transformation, but as a risk-managed process grounded in formal guarantees. Contemporary DTFs must therefore include explicit definitions of abstraction levels, measurable criteria for leakage, and transformation metadata documenting which information has been removed, retained or generalised.

3.2. Reproducibility Requirements

Reproducibility has become a central requirement in contemporary machine learning and computational linguistics. This kind of research requires accessible datasets, legal frameworks often prohibit sharing the underlying content.

DTFs offer a way to reconcile these conflicting demands. The exact transformed datasets used in a workflow can be published, allowing others to replicate results with full transparency regarding preprocessing steps and data provenance.

Furthermore, DTFs allow for the creation of benchmarks that remain usable over time even when the original corpora cannot be redistributed or when licensing conditions change. By separating protected content from its derived representational layers, DTFs make it possible to document and preserve the exact data conditions under which a model was trained or evaluated.

In this sense, reproducibility is not merely a desirable feature but an operational requirement – one that DTFs are uniquely positioned to support.

3.3. Legal and Ethical Context

Legal and ethical constraints surrounding text content remain major challenges for the distribution and reuse of linguistic corpora. Copyright and contractual restrictions often prevent redistribution of source content, even when annotation layers are fully owned by researchers (Lehmberg et al., 2007). Corpora frequently comprise multiple content layers (source content, annotations, metadata, derived structures), each subject to different legal regimes, which complicates reuse across institutions.

Privacy and data protection laws add further constraints. GDPR (European Parliament and Council of the European Union, 2026) prohibits sharing corpora containing personal or identifiable information, regardless of copyright status. While masking can reduce identifiability, it must be designed carefully: insufficient abstraction or the inferential power of modern LLMs can allow partial reconstruction, similar to failures in anonymisation.

Recent legal developments relevant to machine learning include the German text-and-data-mining (TDM) exception (UrhG) (also see Text+ on TDM), which permits certain forms of model training under specific conditions, provided that resulting models do not reproduce protected content and deletion obligations are met. However, the exception does not allow the redistribution of the underlying corpora, leaving reproducibility dependent on legally shareable derived formats.

Complex licensing structures further complicate matters. Large corpora may involve multiple rights holders, annotation layers contributed by different groups, and tools or contracts imposing additional restrictions. Clear provenance documentation is essential to specify which rights apply to each layer and how they are transformed in a DTF workflow.

By separating protected source texts from non-reversible derived representations, DTFs reduce legal risks and support compliance with copyright,

contractual obligations and data protection requirements. Earlier analyses (Lehmberg et al., 2008) already highlighted the need for such differentiated data handling; in the 2026 landscape, it has become an operational requirement. DTFs thus function not only as technical artefacts but as legally and ethically robust instruments for sustainable and compliant research data practices.

4. Standardisation Landscape

The increasing legal, technical and methodological complexities surrounding textual data have intensified the need for clear, formalised frameworks that govern how DTFs are created, documented and shared. While the early masking approaches were primarily technical experiments developed in the absence of established standards, today's research data ecosystem demands structured, interoperable and machine-actionable specifications. Here, standardisation plays a crucial role: it transforms ad-hoc masking techniques into reproducible, transparent and legally robust workflows. Emerging efforts such as the DIN standard for Derived Text Formats (DIN 19461:2026-04 , E, currently a draft national standard), along with existing ISO standards for linguistic representation and metadata (such as ISO 24622-1:2015, ISO 24622-2:2019, ISO 24619:2011) provide a coherent foundation upon which modern DTF practices can be built. Below, we outline this evolving standardisation landscape and its implications for sustainable research data management.

4.1. DTF Standard DIN 19461

The emerging DTF DIN standard (DIN 19461:2026-04 , E) provides the first formalised framework for describing, generating and documenting transformed text representations in a standardised and methodologically transparent way. Unlike the early approaches, the new standard offers a systematic categorisation of transformation types, information-reduction operations and metadata obligations that apply to all formats derived from natural-language text.

At its core, the standard defines DTFs as structured text formats produced through *targeted enrichment and information reduction* of an original source text. It distinguishes clearly between enrichment steps that add analytical layers such as POS tags, lemmas, NER information or syntactic structures, and information-reduction operations designed to eliminate or generalise protected content. The explicit modelling of transformation procedures formalises what earlier work on corpus masking and masked treebanks treated implicitly.

A central requirement of the DIN standard is the assurance of *non-reconstructability*. The standard recognises that reconstruction risks may arise not only from a single transformation but also from the *combination of multiple DTFs* generated from the same source material. Section 4.4 therefore mandates a systematic evaluation of combined leakage risks – an issue anticipated in 2007, but now formulated explicitly as a normative obligation.

The standard also introduces comprehensive *documentation and metadata requirements*. All enrichment and reduction steps must be fully documented, including tools, algorithms, parameters and the granularity at which each operation was applied. Documentation should be linked to machine-readable CMDI metadata following (ISO 24622-1:2015) and (ISO 24622-2:2019), ensuring traceability and long-term sustainability. This establishes a formal reproducibility framework that earlier masking work lacked.

In addition, the standard defines canonical *categories of derived formats*, including token-based DTFs (e. g., bag-of-words and n-gram representations), vector-based DTFs (e. g., TF/IDF, Word2Vec or contextual embeddings), structured DTFs (e. g., shuffled segments) and multi-feature formats (e. g., HathiTrust Extracted Features). By codifying these patterns, the DIN standard extends concepts implicit in masking – replacing words with placeholders, removing sequence information, randomising tokens – into a general typology that reflects contemporary NLP and text-analytic needs.

While the canonical DTF categories define token-based units as the smallest standardised representational layer, subtoken-level information units – such as character n-grams, morphologically defined graphemic segments, or tokenizer-produced subword units – can also be modelled within the DTF framework. Their inclusion changes both analytical utility and leakage profiles. DIN 19461 therefore treats subtoken-level DTFs as special cases of information-reduction or enrichment steps whose granularity must be documented. Modern masking workflows often implicitly operate on such units (e. g., affix-aware masking), and evaluation must consider reconstruction pathways that exploit LLM tokenizers or subword distribution patterns.

4.2. International Standards

A crucial element of the emerging DTF ecosystem is its alignment with established standards for linguistic representation and metadata. While the DIN DTF standard defines the conceptual and operational framework for DTFs, its practical implementation relies on interoperable modelling languages, annotation formats and metadata infrastructures (DIN 19461:2026-04 , E).

The *TEI Guidelines*, in particular the module for *Feature Structures*, provide a flexible mecha-

nism for representing linguistic information independently of the underlying text. Feature structures support hierarchical, attribute–value-based annotations and are already used in real-world DTF implementations where annotation layers remain intact while textual content is masked or transformed (DIN 19461:2026-04 , E).

Beyond TEI, a suite of *ISO standards for linguistic annotation* ensures cross-infrastructure interoperability. These include models for syntactic structure (ISO 24615-1:2014) (SynAF), morpho-syntactic units (ISO 24611-1:2025) (MAF), and lexical representations (ISO 24613-1:2024) (LMF), as well as frameworks for structural, discourse and semantic annotation (ISO 24612:2012) (LAF) and the ISO 24617-x series (SemAF, see for example ISO 24617-1:2012). Together, they provide the conceptual backbone referenced throughout the DIN DTF draft (DIN 19461:2026-04 , E).

The *Component Metadata Infrastructure (CMDI)* (ISO 24622-1:2015; ISO 24622-2:2019) offers a mechanism for documenting linguistic resources. CMDI profiles can capture the enrichment and reduction steps required for DTF, ensuring transparent and machine-actionable provenance (DIN 19461:2026-04 , E).

Finally, *persistent identifiers (PIDs)* support long-term referencing and provenance tracking. Systems such as DOI, Handle or the identifier mechanisms defined in (ISO 24619:2011) allow infrastructures to systematically track all DTF variants derived from a source corpus and thus complement FAIR-aligned data stewardship practices (DIN 19461:2026-04 , E).

Together, TEI Feature Structures, ISO annotation standards, CMDI metadata infrastructure and persistent identifier systems provide the technical foundation that enables DTF to function as transparent, sustainable and interoperable components within modern research data ecosystems.

4.3. Infrastructural Implications

The introduction of a formal DIN standard for DTFs has significant implications for national and international research infrastructures. Infrastructures for language resources including sustainability and repository services depend on interoperable formats, persistent identifiers and machine-readable metadata to ensure long-term access, legal compliance and cross-project reuse. The DIN DTF standard provides a unified framework that enables derived textual data to be integrated into these ecosystems in a transparent and robust manner (DIN 19461:2026-04 , E).

By requiring the documentation of enrichment steps, information-reduction operations, granularity choices and tool parameters – captured through CMDI-compatible metadata profiles (ISO 24622-1:2015; ISO 24622-2:2019) – the standard ensures that DTFs can be archived, validated and reused across institutions. Maintaining internal and external records of all DTF variants derived from a source corpus supports long-term preservation and risk assessment, while the linkage of transformation metadata to persistent identifiers (ISO 24619:2011) allows repositories to manage DTFs as first-class digital objects (ISO 24619:2011).

Beyond technical specifications, the DIN DTF standard acts as an infrastructural enabler. By aligning DTF workflows with established standards for metadata, annotation, provenance and sustainability, it facilitates seamless integration of derived representations into the major research data infrastructures essential for open science and enduring accessibility of textual resources.

5. Revisiting and Modernising Masking Algorithms

The early masking techniques formed an important starting point for generating legally shareable representations of copyrighted or sensitive corpora. They demonstrated that substantial syntactic and morphological information can be preserved even when lexical material is removed or obfuscated. At the same time, the assumptions underlying the original methods – especially concerning reconstructability and acceptable abstraction levels – reflected the technological landscape of their time. With the advent of LLMs, machine-learning architectures and formalised DTF standards, these early approaches require systematic revision.

This section summarises the conceptual foundations of the original masking approach, outlines the requirements modern DTF must meet and presents a contemporary, multi-stage workflow for producing legally compliant and reproducible DTFs.

5.1. The Original Masking Approach

The 2007 masking strategies were built around the idea of *parameterised masking*: configurable transformation settings allowed researchers to adjust how much information was retained, modified or removed. Central to this approach was a *dictionary-based randomisation* procedure that replaced each token with an artificial string – preserving length or formal characteristics where needed – while ensuring internal coherence across the corpus. Additional *affix-aware strategies* enabled the retention of morphological cues such as case, number or tense, even when stems were fully masked. Low-risk closed-class items could be selectively retained to support syntactic modelling without introducing significant leakage risks.

These ideas were operationalised in the *CorpusMasker* tool (Dellert, 2007; Rehm et al., 2007a), which provided a GUI for experimenting with masking parameters and applying transformations to XML-annotated corpora. The resulting principles – parameterisation, affix sensitivity, selective retention and controlled randomisation – remain foundational for contemporary DTF approaches.

5.2. Updated Requirements for DTF

Transforming masking into fully fledged DTFs introduces several new requirements.

First, *robustness against reconstruction* has become essential. Modern LLMs can infer missing lexical or semantic information from subtle structural or distributional cues. Removal or randomisation of surface forms is no longer sufficient on its own; DTFs must be evaluated against realistic attack models and provide quantifiable, machine-verifiable guarantees of non-reversibility.

Second, contemporary DTFs require *explicit abstraction levels and reversibility classes*. Instead of a single continuum of obfuscation, DTF frameworks must define what types of linguistic or statistical information remain in a derived representation and what forms of reconstruction are theoretically possible. This includes classes such as token-level replacement, morphological generalisation, sequence removal or embedding-level abstraction, each with different leakage profiles.

Third, modern infrastructures require *comprehensive provenance and machine-actionable metadata*. All enrichment and reduction operations must be documented – tools, parameters, granularity choices and processing steps – so that repositories and workflows can validate, reproduce and assess the transformation. Such provenance is essential for legal compliance, long-term preservation and transparent scientific practice.

Together, these requirements transform masking from an informal technique into a principled, standardised methodology embedded within the broader DTF framework.

5.3. A Workflow for DTF Generation

Modern workflows for generating DTFs extend far beyond the ad-hoc masking procedures of 2007. They integrate enrichment, structural abstraction, statistical modelling and legally robust information reduction into reproducible multi-stage pipelines.

The process begins with the *extraction of structural representations* from the source text. Prior to any reduction, the text is enriched with annotation layers such as tokenisation, part-of-speech tags, lemmas, morphological features, named entities and syntactic dependencies. These layers form the basis for transformations that preserve analytic value while removing protected content.

Next, the workflow applies *structural abstraction*, producing feature bundles, e. g., POS or lemma classes, dependency graphs or other units defined in the DIN DTF framework. This separates the protected source text from the shareable representational layers.

In parallel, workflows may generate *statistical and embedding-based formats*, including TF/IDF features, topic-modelling inputs, token or sentence embeddings and contextualised representations. These abstractions retain analytical utility without exposing lexical material.

A dedicated *non-reversible transformation layer* ensures that the resulting DTF cannot be used to reconstruct the source text. Depending on the abstraction level and reversibility class, this layer applies operations such as deletion, replacement or randomisation. Modern pipelines must ensure that combinations of retained information do not permit reconstruction, even by advanced systems.

DTFs do not assume that the source material is written text. Any linguistic representation – handwritten, OCR-derived, ASR output, phonetic or graphemic transcription – can serve as source material. For born-derived modalities, the same principles apply: enrichment layers (POS, phonetic features, timing annotations) precede information reduction, and reversibility is assessed with respect to the original modality.

A concrete implementation of this approach is provided by *MONAPipe* (Dönicke et al., 2022; MONAPipe 2023), whose `derived_text_formatter` applies DIN-aligned operations such as `replace`, `randomize` and `keep` at multiple text levels. The use depends on the language models selected, MONApipe's default being German at present. While MONAPipe already supports the main transformation operations, it does not yet generate a complete provenance record for example in a CMDI serialisation, which is an important requirement for future development.

Taken together, these developments transform masking from a single operation into an integrated, multi-stage workflow. Contemporary DTF pipelines combine enrichment, abstraction, statistical representation and non-reversible transformation to produce legally compliant, reproducible and infrastructure-ready representations of textual data.

DTF enable a broad range of research applications, from stylometry and diachronic studies to NLP benchmarking and pedagogical use. These applications benefit from the structural preservation and non-reversibility provided by DTF workflows.

6. Discussion

The emergence of DTFs has opened new possibilities for research on copyrighted or sensitive textual data, but it also highlights several important trade-offs. Foremost among these is the tension between *utility* and *non-reversibility*: retaining too much linguistic or distributional information can increase the risk of reconstruction, while overly aggressive abstraction may reduce the analytical value of the data. Determining the appropriate level of transformation requires explicit modelling of abstraction levels, reversibility classes and information-reduction operations, as well as evaluation against potential reconstruction pathways – including those made possible by modern LLMs.

Several *limitations of current approaches* persist. Even well-designed masking or randomisation strategies cannot guarantee absolute non-reversibility, particularly when external knowledge or powerful generative models can infer missing content from subtle linguistic cues. The ability of LLMs to recover stylistic, syntactic or semantic patterns raises questions about the long-term safety of DTFs that retain traces of original structure. Furthermore, the development of DTF workflows varies across research communities, with some digital humanities domains lacking standardised pipelines. Tools such as MONA-pipe implement DIN-aligned transformation operations, but still do not automatically generate the complete provenance metadata required for reproducible and legally robust publication – an important aspect for future work.

These challenges underscore the *importance of standardisation*. The techniques introduced in 2007 established key principles – parameterisation, structured abstraction and controlled leakage – that underpin modern DTF approaches. The DIN DTF standard extends these ideas by providing formal categories, transformation operations, documentation obligations and provenance requirements.

Several *open questions* remain. Defining thresholds for irreversibility, quantifying leakage risks, automating provenance-aware metadata generation and balancing legal compliance with scientific utility will require continued interdisciplinary collaboration. Ultimately, the promise of DTFs lies not only in enabling access to restricted data but also in establishing a transparent, accountable and reproducible ecosystem for text-based research – one that remains resilient as technologies, legal frameworks and research practices continue to evolve.

Future work includes the development of metrics that quantify both divergence from the original and retained utility. Current practice relies on (i) statistical divergence measures (e. g., KL or JS divergence of POS or dependency distributions), (ii) structural similarity metrics (tree edit distances, graph similarity), (iii) task-based utility evaluations (NER, parsing, text classification performance on DTF vs. original), and (iv) leakage assessments using reconstruction experiments Du et al. (2025). While DIN 19461 introduces reversibility classes and documentation obligations, it does not prescribe specific metrics; formal benchmarks for leakage and utility are therefore an open research need.

7. Conclusion

Nearly two decades separate the first explorations of corpus masking from DTFs in 2026. During this period, the research community has progressed from prototypes toward a mature and standardised ecosystem for legally compliant text transformations. The original work anticipated many of the challenges that would later become central – most notably the need to preserve linguistic structure while preventing reconstruction – and laid the foundations on which modern approaches now build.

The updated framework presented in this paper translates the early insights of parameterised masking, affix-aware strategies and controlled leakage into a formalised model incorporating explicit abstraction levels, reversibility classes, provenance requirements and non-reversibility guarantees suited to an era shaped by LLMs. Tools such as MONAPipe, alongside emerging standards like the DIN DTF draft standard, provide practical implementations of these concepts and illustrate how masking can evolve into transparent, reproducible and extensible workflows.

DTFs also contribute directly to broader goals in research data management. By enabling the publication of legally usable, machine-actionable representations of copyrighted or sensitive corpora, they support the reproducibility of computational experiments, the creation of open benchmarks and the long-term stewardship of linguistic resources. They enhance compliance with legal and ethical requirements without diminishing the analytical potential of derived datasets.

Looking ahead, DTFs could become a sustainable and foundational component of national and international research infrastructures. Their integration into metadata standards, repository systems and NLP pipelines will help ensure that text-based research remains transparent, responsible and legally viable as technologies and legal contexts continue to evolve. Future work on automated provenance generation, leakage evaluation frameworks and interoperable DTF toolchains will further strengthen this role. DTFs represent not only a technical solution but also a structural contribution to open science: a means of enabling rigorous scholarship while respecting the rights and protections inherent in the underlying textual data.

8. Acknowledgements

Though the authors are indebted to various co-authors working on this topic for years, work on this paper was carried out within the National Research Data Infrastructure (NFDI) association. The NFDI is funded jointly by the Federal Republic of Germany and the 16 federal states, and the Text+ consortium is supported by the German Research Foundation (DFG). Georg Rehm is part of NFDI4DS – the National Research Data Infrastructure for Data Science and Artificial Intelligence, grant number 460234259; Andreas Witt and Thorsten Trippel are part of the Text+ consortium, grant number 460033370. The authors gratefully acknowledge this support, as well as the engagement of all institutions and individuals contributing to the NFDI and its goals.

The authors acknowledge the use of Large Language Models (LLMs) as writing aids in phrasing this paper, based on the authors' notes, ideas and concepts. The authors retain full responsibility for the content.

9. Bibliographical References

Johannes Dellert. 2007. Corpusmasker: A tool for parameterised masking of linguistic resources. https://www.lingexp.uni-tuebingen.de/sfb441/c2/corpus-masker-0.1.tar.gz. Version 0.1, SFB 441 "Linguistic Data Structures", University of Tübingen.

DIN 19461:2026-04 (E). 2026. Sprachressourcen und Sprachtechnologie - Abgeleitete Textformate (ATF). Technical report, Deutsches Institut für Normung, Berlin.

Tillmann Dönicke, Florian Barth, Hanna Varachkina, and Caroline Sporleder. 2022. Monapipe: Modes of narration and attribution pipeline for german computational literary studies and language analysis in spacy. In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, Potsdam, Germany.

Keli Du, Sarah Ackerschewski, Uygar Navruz, Nazan Sınır, Julian Valline, and Christof Schöch. 2025. Reconstructing shuffled text. bad results for nlp, but good news for using in-copyright texts. *Journal of Computational Literary Studies*, 4(1).

European Parliament and Council of the European Union. 2026. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation).

Darja Fišer and Andreas Witt, editors. 2022. *CLARIN: The Infrastructure for Language Resources*. De Gruyter, Berlin, Boston.

Erhard Hinrichs and Thorsten Trippel. 2024. Text+ – concept and benefits for empirical researchers. *Cybernetics and Information Technologies*, 24(4):143–163.

ISO 24611-1:2025. 2025. Language resource management — morphosyntactic annotation framework (MAF) — part 1: Core model. International Standard, International Organisation for Standardization (ISO), Geneva.

ISO 24612:2012. 2012. Language resource management — linguistic annotation framework (LAF). International Standard, International Organisation for Standardization (ISO), Geneva.

ISO 24613-1:2024. 2024. Language resource management — lexical markup framework (LMF) — part 1: Core model. International Standard, International Organisation for Standardization (ISO), Geneva.

ISO 24615-1:2014. 2014. Language resource management — syntactic annotation framework (SynAF) — part 1: Syntactic model. International Standard, International Organisation for Standardizaton (ISO), Geneva.

ISO 24617-1:2012. 2012. Language resource management — semantic annotation framework (SemAF) — part 1: Time and events (semaf-time, iso-timeml). International Standard, International Organisation for Standardization (ISO), Geneva.

ISO 24619:2011. 2011. Language resource management – Persistent identification and sustainable access (PISA). International Standard, International Organization for Standardization (ISO), Geneva.

ISO 24622-1:2015. 2015. Language resource management – Component Metadata Infrastructure (CMDI) – Part 1: The Component Metadata Model. International Standard, International Organization for Standardization (ISO), Geneva.

ISO 24622-2:2019. 2019. Language resource management – Component Metadata Infrastructure (CMDI) – Part 2: The Component Metadata Specification Language. International Standard, International Organization for Standardization (ISO), Geneva.

Timm Lehmberg, Christian Chiarcos, Georg Rehm, and Andreas Witt. 2007. Rechtsfragen bei der nutzung und weitergabe linguistischer daten. In Georg Rehm, Andreas Witt, and Lothar Lemnitzer, editors, *Datenstrukturen für linguistische Ressourcen und ihre Anwendungen*, pages 93–102. Gunter Narr Verlag, Tübingen.

Timm Lehmberg, Georg Rehm, Andreas Witt, and Felix Zimmermann. 2008. Digital text collections, linguistic research data, and mashups: Notes on the legal situation. *Library Trends*, 57(1):52 – 71.

MONAPipe 2023. 2023. Monapipe: Modes of narration and attribution pipeline. `GitLab repository`. Accessed: 2026-03-20.

Stelios Piperidis, Harris Papageorgiou, Christian Spurk, Georg Rehm, Khalid Choukri, Olivier Hamon, Nicoletta Calzolari, Riccardo del Gratta, Bernardo Magnini, and Christian Girardi. 2014. META-SHARE: One year after. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014)*, pages 1532–1538, Reykjavik, Iceland. European Language Resources Association (ELRA).

Georg Rehm, editor. 2023. *European Language Grid: A Language Technology Platform for Multilingual Europe*. Cognitive Technologies. Springer, Cham, Switzerland.

Georg Rehm, Stelios Piperidis, Khalid Choukri, Andrejs Vasiļjevs, Katrin Marheinecke, Victoria Arranz, Aivars Bērziņš, Miltos Deligiannis, Dimitrios Galanis, Maria Gavriilidou, Maria Giagkou, Katerina Gkirtzou, Dimitris Gkoumas, Annika Grützner-Zahn, Athanasia Kolovou, Penny Labropoulou, Andis Lagzdiņš, Elena Leitner, Valérie Mapelli, Hélène Mazo, Simon Ostermann, Stefania Racioppa, Mickaël Rigault, and Leon Voukoutis. 2024. Common European Language Data Space. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3579–3586, Turino, Italy. European Language Resources Association (ELRA) and International Committee on Computational Linguistics (ICCL). May 20-25, 2024.

Georg Rehm, Andreas Witt, Heike Zinsmeister, and Johannes Dellert. 2007a. Corpus masking: Legally bypassing licensing restrictions for the free distribution of text collections. In *Digital Humanities 2007. Conference Abstracts*, pages 166–170, Urbana-Champaign. University of Illinois.

Georg Rehm, Andreas Witt, Heike Zinsmeister, and Johannes Dellert. 2007b. Masking treebanks for the free distribution of linguistic resources and other applications. In *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories (TLT 2007)*, pages 127–138, Bergen, Norway.

Christof Schöch, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann, and Jörg Röpke. 2020. Abgeleitete textformate: Text und data mining mit urheberrechtlich geschützten textbeständen. *Zeitschrift für digitale Geisteswissenschaften (ZfdG)*, 5.

Text+ on TDM. 2025. Why the training of large language models falls within the scope of the text and data mining exceptions. Last modified: 23 January 2025.

UrhG. 2021. § 44b urhg – text und data mining. `https://www.gesetze-im-internet.de/urhg/__44b.html`. Zugriff am 16. Februar 2026.

Multi-Label Text Classification of Derived Text Formats with DistilBERT

Jennifer Ecker, Roman Schneider
Leibniz Institute for the German Language
R5 6-13, D-68161 Mannheim
{ecker, schneider}@ids-mannheim.de

Abstract

Derived Text Formats enable the distribution of copyrighted texts by systematically perturbing linguistic information to reduce reconstructibility. However, the extent to which such information loss affects downstream text classification remains unclear. We investigate how controlled perturbations affect learning dynamics in transformer-based classification using two datasets and two strategies: POS-consistent replacement of 30%, 40%, and 50% of tokens, and random word-order shuffling. On Wikipedia data, POS replacement increases loss by 4–9% and reduces micro-F1 by 3–8%, depending on the replacement rate, while shuffling raises loss by 5% and lowers micro-F1 by 4%. Performance degrades monotonically with higher replacement rates, and shuffling yields results between the 30% and 40% conditions, indicating that DistilBERT relies more on lexical semantics than on word order. Experiments on specialist-domain data show the same pattern, demonstrating robustness across domains. To test cross-representation generalization, we train classifiers on both clean and perturbed texts and evaluate them on the respective alternate representation. Models trained on DTF data generalize better to clean text than vice versa, suggesting that perturbation-based training promotes more robust representations. Our findings position DTF as a promising strategy for reproducible, legally compliant, and robust NLP research.

Keywords: DTF, Multi-label classification, Text perturbations, Transformer robustness, Learning dynamics

1. Introduction

Reproducibility underpins rigorous empirical NLP research, yet copyright restrictions on large text corpora preclude their public dissemination, compromising scientific transparency and replication. Derived Text Formats (DTF) address this challenge by systematically perturbing texts—through techniques such as word embeddings, Part-of-Speech (POS) tag replacement, or word-order shuffling—to reduce reconstructibility while retaining utility for downstream tasks like classification (Rehm et al., 2007). However, Kugler et al. (2024) demonstrate high reconstructibility of contextualized word embeddings produced by transformer-encoders, confirming that such DTFs enable analysis but still risk copyright violation.

Multi-label classification assigns multiple labels to instances simultaneously, prevalent in hierarchical taxonomies like Wikipedia main topic categories (Tsoumakas and Katakis, 2007; Tarekegn et al., 2021). Early approaches relied on rule-based category matching, while recent methods fine-tune BERT (Bidirectional Encoder Representations from Transformers) variants on imbalanced datasets (Sanh et al., 2019). Perturbation robustness studies reveal BERT's vulnerability to adversarial attacks (e.g., TextFooler synonym swaps) and shuffling, underscoring lexical-semantic dependence over syntax (Jin et al., 2019; Hauser et al., 2021; Reimers and Gurevych, 2019). DTF-specific work has enabled privacy-preserving analysis in computational literary studies, yet – to our knowledge – no systematic evaluation exists of transformer-based text classification performance under controlled DTF degradation.

This study aims to fill that gap. We investigate the impact of selected DTF on transformer-based language models for large-scale classification tasks. Our experiments draw on two complementary corpus sources:

1. Wikipedia articles: This text type represents general-domain language use and provides a heterogeneous and widely used benchmark corpus. Its open-data status enables the parallel publication of original and derived versions, allowing explicit quantification of the textual alterations introduced by DTF.

2. Specialized scientific texts: Although likewise publicly accessible, these texts differ substantially in register and structure. They are domain-specific, terminologically dense, and frequently include illustrative example sentences (i.e., object-language material). Their fine-grained linguistic annotations impose distinct challenges for classification models.

The combination of these two corpora seems methodologically advantageous: Together, they facilitate a systematic assessment of the robustness and transferability of multi-label classifiers across registers and levels of annotation granularity, supporting the broader goal of transferring models to

Proceedings of Leveraging Derived Text Formats to Unlock Copyrighted Collections for Open Science (DTF) @ LREC 2026, pages 34–43
12 May 2026. ©ELRA Language Resources Association (ELRA), 2026

proprietary corpora.

This paper is structured as follows: Section 2 describes the data and methodology, Section 3 presents experimental results on robustness and train–test mismatch, and Section 4 discusses implications for privacy-preserving NLP.

2. Classification task

DTFs have already been tested for author classification (Du, 2023) and sentiment classification (Du and Schöch, 2024). In this work, we extend their use to a multi-label text classification setting. It is particularly interesting to see what effects different perturbations have when training this type of text classifier. In general, difficulties arise from imbalanced class distributions, whereby some classes are overrepresented, while others contain fewer examples. Furthermore, assigning multiple labels to one text expands the problem, creating many possibilities for combining the classes.

For the Wikipedia texts, we assign categories based on the Wikipedia main topic classification taxonomy[1] as of October 22, 2024. Because this taxonomy evolves over time, categories may be merged, removed, or newly introduced. In this study, we use the following top-level classes: Academic disciplines, Business, Communication, Concepts, Culture, Economy, Education, Energy, Engineering, Entertainment, Entities, Food and drink, Geography, Government, Health, History, Human behavior, Humanities, Information, Internet, Knowledge, Language, Law, Life, Lists, Mass media, Mathematics, Military, Nature, People, Philosophy, Politics, Religion, Science, Society, Sports, Technology, Time, and Universe.

For the specialist texts, we employ fine-grained categories derived from a dedicated domain-specific ontology (Lang et al., 2018). Further details are provided in the data section below.

2.1. Data

2.1.1. Wikipedia articles

The Wikipedia text data are extracted from XML sources (Kupietz et al., 2019) containing articles included in the German Reference Corpus *DeReKo* (Kupietz et al., 2018). We build the training corpus using a rule-based string-matching method that leverages the category links stored in the XML metadata. Each article contains a reference to its corresponding Wikipedia category in the <classCode> element. Using these references, we extract all terms corresponding to the main topic classifications. A lexicon of main categories (e.g., *Geography*) is created, and each lexicon entry is matched against the category names in the XML. Categories in the XML are often deep in the Wikipedia hierarchy (e.g., *Geography of Saxony-Anhalt*), where the parent main category could theoretically be determined by traversing the hierarchy upwards. However, to maintain practical and straightforward labeling, this traversal is not performed, and categories are assigned directly from the XML tag (e.g., *Geography of Saxony-Anhalt* remains as-is).

Using this approach, the resulting Wikipedia dataset comprises 430,767 texts totaling 421,659,916 tokens for classifier training. The assigned categories serve as thematic classes, with each text associated with an average of 1.24 class labels. Measured in characters, document length ranges from 1 to 466,945 characters, with a mean of 4,341 (median: 2,420; standard deviation: 7,765). When measured in words, texts range from 1 to 66,034 words, with a mean of 595 (median: 335; standard deviation: 1060). The observed minimum document length of a single character can be attributed to artifacts introduced during the automated data extraction process (e.g., formatting remnants or parsing inconsistencies). These extremely short texts represent rare outliers, as reflected by the substantially higher median (2,420 characters) and mean (4,341 characters) document lengths. Due to their negligible semantic content, such instances are unlikely to have a meaningful impact on classifier training.

2.1.2. Specialist texts

The specialist texts are linguistics research texts written by expert linguists and published in the online grammar portal *grammis* (Schneider and Lang, 2022). We use 1,649 documents from this source, which are part of COrLiCo (Corpus for the Oral–Literate Continuum) (Schneider, 2026). Each document has been manually annotated with 593 fine-grained thematic categories, averaging 2.03 labels per text. The category distribution is highly skewed, with some frequent labels (e.g., *Wortstellung, Wortbildung, Komposition*) and around 300 singleton labels (e.g., *Adjunktorgruppe, Diktumsgraduierung, Gattungsname*). For comparability with the Wikipedia sub-corpus, we retain only the 39 most frequent linguistic categories for classification.

In terms of length, the *grammis* texts exhibit substantial variation. Measured in characters, document length ranges from 100 to 36,194 characters, with a mean of 2,826 (median: 1,995; standard deviation: 2,956). When measured in words, texts range from 10 to 4,875 words, with a mean of 358 (median: 244; standard deviation: 392). This indicates a highly heterogeneous dataset with a long-tailed distribution of document lengths.

[1] https://en.wikipedia.org/wiki/Category:Main_topic_classifications

We expand this specialist text dataset using controlled paraphrastic augmentation to increase the amount of training data. Each of the 1,649 original texts is paraphrased ten times with Cohere's Command A, a 111B-parameter large language model optimized for multilingual processing of long documents (Cohere, 2025), which we access locally via an Ollama runtime. The model is prompted as a German linguistics expert and rewrites each text to preserve the original content and specialized terminology while varying the surface form and phrasing. To ensure diversity and quality, we discard near-duplicate paraphrases that exceed a cosine similarity threshold of 0.80. Additionally, we manually inspect a random sample of 100 outputs to verify that the paraphrases are grammatical, semantically faithful, and suitable as additional training instances.

The generated texts differ slightly in their length distribution from the original *grammis* texts. In characters, they range from 82 to 159,724, with a mean of 1,996 (median: 1,807; standard deviation: 1,869). In terms of words, lengths vary between 9 and 14,918 words, with a mean of 245 (median: 222; standard deviation: 209). Compared to the original texts, the paraphrases are on average shorter and less variable, although a small number of extreme outliers remain.

This process generates up to 16,490 paraphrase candidates, resulting in a final specialist dataset of 18,139 linguistic texts, totalling 4,628,838 tokens.

A critical point concerns the extensive use of paraphrastic augmentation. While this procedure substantially increases the amount of available training data, it also introduces a strong dependency on model-generated text: roughly 90% of the resulting specialist dataset consists of LLM-produced paraphrases rather than naturally occurring expert texts. This raises the question to what extent observed patterns reflect genuine properties of the underlying corpus as opposed to artefacts of the generative model.

The paraphrases are generated without explicitly specified decoding parameters (e.g., temperature or nucleus sampling), meaning that the degree of variation is not systematically controlled and may depend on implicit model defaults. At the same time, large language models tend to regularize stylistic variation, smooth rare constructions, and favor preferred lexical and syntactic patterns. Even in the absence of deterministic decoding, this can induce subtle distributional shifts, for instance a bias toward more prototypical or "canonical" formulations. Although we filter near-duplicates and manually verify a subset of outputs, such measures cannot fully eliminate these effects.

Consequently, results obtained on this dataset should be interpreted with caution. Performance gains may partly reflect a model's ability to learn and exploit the stylistic and structural regularities of the generating LLM, rather than the full diversity of authentic specialist writing. Future work could address this limitation by (i) explicitly reporting and controlling decoding parameters to improve reproducibility, (ii) comparing against non-augmented baselines, and (iii) systematically analyzing differences between human- and model-generated texts, for example with respect to lexical diversity and syntactic variation.

2.1.3. Generating Derived Text Formats

For both Wikipedia and linguistic specialist texts, we generated multiple DTF versions using the derived text formatter provided by MONAPipe (Dönicke et al., 2022). We systematically generate four DTF variants to probe distinct information bottlenecks:

- **Word replacement**: Replace X% of content words per sentence with their POS tags (X $\in$ {30, 40, 50}), preserving syntactic structure while eliminating lexical-semantic information for substituted tokens.
- **Word-order randomization**: Fully shuffle tokens at document level, eliminating sequential and syntactic relations while preserving the full lexicon.

These perturbations create orthogonal degradation axes: word replacement ablates *semantics* at controlled rates, while word-order randomization ablates *syntax/sequence*, enabling dose-response analysis of DistilBERT's representational reliance on each signal type. Specifically, the graduated POS replacement rates quantify lexical contribution incrementally – each 10% increase in preserved content words should yield measurable F1-micro gains if semantics dominate. Conversely, randomization tests positional encoding utility, hypothesizing lesser degradation since transformer self-attention is permutation-invariant in theory (though task-specific patterns may emerge). This controlled experimental design isolates DistilBERT's feature learning priorities – lexical vs. sequential – while mimicking real-world DTF privacy constraints, providing actionable guidance for copyright-compliant corpus sharing.

2.2. Classifier

We fine-tune a pre-trained DistilBERT model[2] from HuggingFace to perform multi-label classification using PyTorch Lightning. DistilBERT (Sanh et al.,

[2] https://huggingface.co/distilbert/distilbert-base-german-cased

2019) was chosen for its compact size and fast processing speed. It is a knowledge-distilled version of BERT, a transformer-based language model that produces contextualized word representations by considering both left and right context. Through distillation, a smaller "student" model learns to approximate the behavior of a larger pre-trained BERT, retaining much of its semantic and syntactic understanding while being computationally more efficient. This makes DistilBERT particularly well suited for large-scale and resource-sensitive classification tasks.

All models were trained for up to 25 epochs for the Wikipedia data and for up to 40 epochs for the specialist data. For the Wikipedia data, the baseline model and one DTF model were trained five times to quantify residual stochasticity. All experiments used identical data splits and random seeds for NumPy, PyTorch, and CUDA. Although the hyper-parameters had to be tuned separately for each data set to achieve optimal performance, the resulting configurations remained stable across all DTF versions within a given data set. All code will be released in a public repository for the replication of the classification models.

For training, we use Binary Cross-Entropy with Logits Loss, computed independently for each class and averaged across all classes. Each label is thus treated as a separate binary classification task. The loss measures the discrepancy between predicted probabilities and true labels and serves as the differentiable objective optimized during training. We report both training and validation loss to monitor optimization dynamics and generalization. While decreasing training loss indicates successful fitting, divergence between training and validation loss can reveal overfitting or poor generalization.

In multi-label settings with class imbalance, the loss can be dominated by the large number of negative label predictions per instance (Lin et al., 2017). Because it is averaged across all classes, correctly predicting negatives can substantially reduce the overall loss. Consequently, low loss values do not necessarily imply strong performance on positive or minority labels. As noted by Terven et al. (2025), the loss should therefore primarily be interpreted as an optimization objective rather than a comprehensive performance metric.

To evaluate predictive quality, we report the F1-micro score. Unlike the loss, the F1-score operates on discrete predictions and combines precision and recall into a single measure. The micro-averaged variant aggregates decisions across all labels, making it well-suited for imbalanced multi-label scenarios and aligned with our focus on overall predictive behavior. We report both training and validation F1-micro to assess classification quality on seen and unseen data. In contrast, F1-macro assigns equal weight to each class and would mainly capture performance changes in minority labels rather than reflecting overall model behavior.

3. Experiments and Results

We divide our study into two complementary experiments: Experiment A evaluates classifier robustness under controlled text perturbations, using both the Wikipedia and specialist texts corpora to assess domain-general degradation patterns across general-domain and terminologically dense registers. Experiment B investigates representational transfer under train-test mismatch by cross-evaluating models trained on clean vs. DTF-perturbed texts; this analysis is restricted to the Wikipedia corpus, as its larger size and balanced class distribution enable more reliable estimation of generalization gaps, whereas the smaller specialist corpus risks inflated variance from stochastic effects in fewer runs.

3.1. Experiment A: Robustness to Controlled Text Perturbations

3.1.1. Comprehensive Evaluation on Wikipedia Dataset

These experiments were conducted with two different types of data manipulation: (i) replacing X% of the words in the training sentences with their corresponding part-of-speech (POS) tags and (ii) completely randomizing the word order in the training texts. For both manipulations, models with identical hyper-parameters and initializations (seeds) were trained to ensure a fair comparison.

Figure 1 presents the training and validation loss over 20 epochs for the baseline model and two DTF models of the Wikipedia data. The baseline model achieved the best performance, with training loss converging to approximately 0.30 and validation loss stabilizing at the same level, indicating effective learning without significant overfitting. The POS replacement strategy (DTF: POS 50%) resulted in substantially degraded performance, with both training and validation losses plateauing around 0.39. Notably, the training and validation curves remained closely aligned throughout training, suggesting that while overfitting was avoided, the model's capacity to learn discriminative features was severely limited by the reduced lexical information. The randomization strategy (DTF: Randomize) yielded intermediate results, with validation loss converging to approximately 0.34. Although this represents a significant performance degradation compared to the baseline, the model performed better than the POS replacement condition, likely because the original lexical items remained accessible despite the disrupted word order.

To further investigate the impact of lexical information loss, we conducted an ablation study varying the POS replacement rate at 50%, 40%, and 30%. Figure 2 depicts the loss curves for these three configurations. The results demonstrate a clear dose-response relationship between POS replacement rate and model performance. The 30% replacement condition achieved the best performance among the three variants, with validation loss converging to approximately 0.34. The 40% replacement condition showed intermediate performance at around 0.36 validation loss, while the 50% replacement condition exhibited the poorest performance, plateauing at approximately 0.39.

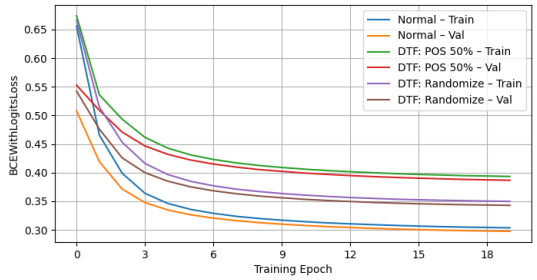

Figure 1: Training and validation loss across epochs for Wikipedia baseline and types of DTF.

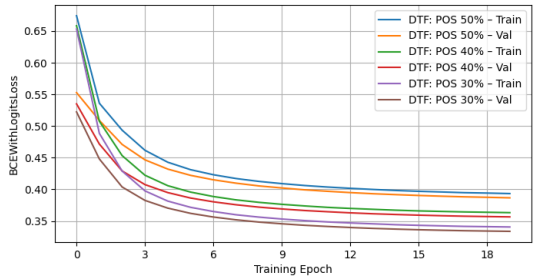

Figure 2: Training and validation loss across epochs for Wikipedia DTF POS types.

Figure 3 presents the F1-micro scores over 20 epochs for the baseline model and two DTF models. The baseline model achieved the best performance, with F1-micro scores converging to approximately 0.48 on the validation set and 0.47 on the training set by epoch 20, demonstrating robust classification capability. While this absolute performance may appear modest, it reflects the inherent difficulty of the task, which involves multi-label classification over 39 categories. Importantly, the focus of this study is on the relative impact of perturbation strategies rather than absolute performance, and the baseline therefore serves as a consistent reference point for comparison. The POS replacement strategy (DTF: POS 50%) resulted in substantially degraded performance, with F1-micro scores plateauing at approximately 0.40 for validation and 0.39 for training. The closely aligned training and validation curves throughout the learning process suggest that the model reached its learning capacity early, limited by the reduced semantic information available when half of the tokens were replaced with generic part-of-speech tags. The randomization strategy (DTF: Randomize) yielded intermediate results, with F1-micro scores converging to approximately 0.43-0.44 on both training and validation sets. This represents a moderate performance degradation of roughly 4-5% compared to the baseline. Notably, the randomization condition outperformed the POS replacement strategy by 3-4%, confirming that preserving lexical semantics even when word order is disrupted provides more discriminative information for multi-label classification than maintaining syntactic structure alone.

Figure 4 depicts the F1-micro score curves for POS replacement rate at 50%, 40%, 30%. The 30% replacement condition achieved the best performance among the three variants, with F1-micro scores converging to approximately 0.45 on validation and 0.44 on training. The 40% replacement condition showed intermediate performance, reaching around 0.43 on validation and 0.42 on training, while the 50% replacement condition exhibited the poorest performance, plateauing at approximately 0.40 for validation and 0.39 for training. Notably, all three conditions showed remarkably similar learning dynamics, with rapid improvement during the first 3-5 epochs followed by gradual convergence. The consistent gap of approximately 2-4% F1-micro points between successive replacement rates suggests that each additional 10% of lexical information preserved contributes meaningfully to classification performance.

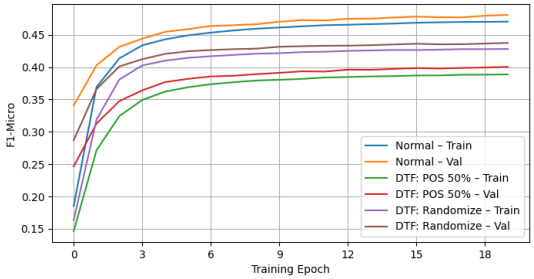

Figure 3: Training and validation F1-micro score across epochs for Wikipedia baseline and types of DTF.

The difference between the loss and the F1-score for the different DTF models from the baseline model is shown in Table 1 displaying the exact differences to the baseline model indicated above.

Model	Δ Train-Loss	Δ Val-Loss	Δ Train-F1	Δ Val-F1
Baseline	0.0000	0.0000	0.0000	0.0000
DTF: POS 50%	0.0870	0.0866	−0.0803	−0.0788
DTF: POS 40%	0.0583	0.0581	−0.0507	−0.0505
DTF: POS 30%	0.0356	0.0353	−0.0296	−0.0293
DTF: Randomize	0.0450	0.0444	−0.0416	−0.0430

Table 1: Difference in loss and F1-score from the baseline model of the Wikipedia data for all DTF models.

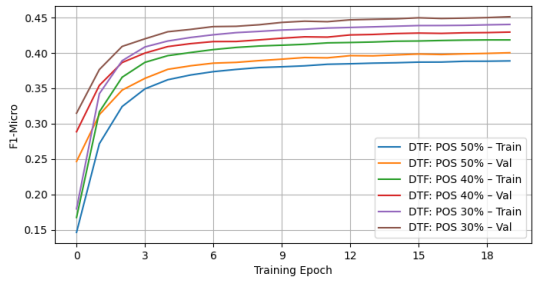

Figure 4: Training and validation F1-micro score across epochs for Wikipedia DTF POS types.

To determine whether the differences arise from the training data or the model itself, we trained the baseline configuration and one DTF configuration (DTF: POS 50%) five times with different random seeds. Due to a limitation of GPU computing resources only these two models were trained more than once. The resulting variance and standard deviation across runs was negligible (see Table 2) indicating that the chosen hyper-parameters are stable and that the model architecture is robust.

3.1.2. Targeted Evaluation on Specialist Dataset

We extend the robustness analysis to the specialist corpus, quantifying degradation in loss and F1-micro relative to the baseline across DTF conditions. The 39-class model achieves a validation F1-micro score of 0.53 (validation loss: 0.28), indicating solid multi-label performance given the fine-grained categories and terminological density. Figures 5 and 6 reveal patterns analogous to those observed for the Wikipedia data.

Qualitative comparison against expert-assigned ground-truth keywords reveals linguistically plausible predictions that align reasonably well with the texts' core themes:

- Text "Wie flektieren entlehnte Adjektive?" (expert: *Adjektiv, Deklination, Flexion, Lehnwort*) predicts *Adjektiv* (0.6428) directly alongside related categories *Satzmodus* (0.5555) and *Satzadverbiale* (0.5567), plausibly extending the adjectival inflection focus.
- Text "Subjektkomplement im Vorfeld" (expert: *Vorfeld, Subjekt, Komplement*) assigns high probabilities to *Passiv* (0.9545), *Vorfeld* (0.7806), *Satzadverbiale* (0.7222), and *Supplement* (0.5436), capturing key aspects of clause-initial argument structure and word order.
- Text "Satz-Nomen- und Phrase-Nomen-Komposita" (expert: *Komposition, Nominalphrase, Phrase, Satz*) predicts *Wortstellung* (0.7111), *Mittelfeld* (0.6096), and *Wortart* (0.5850), reasonably associating compounding with phrasal and sentence-level syntactic phenomena.

These examples demonstrate face-validity: the model's top predictions are interpretable and thematically coherent with expert labels, focusing on grammatical core concepts (inflection, word order, argument structure) rather than random or implausible categories. This supports DistilBERT's capacity for meaningful thematic classification in specialized registers, though exact keyword recovery remains partial as expected in a reduced 39-class taxonomy. Moreover, text length and the usage of specialized terminology also play an important role in the predictions, as less terminology, shorter text length, and an overload of many example sentences illustrating grammatical phenomena lead to a more inaccurate assessment.

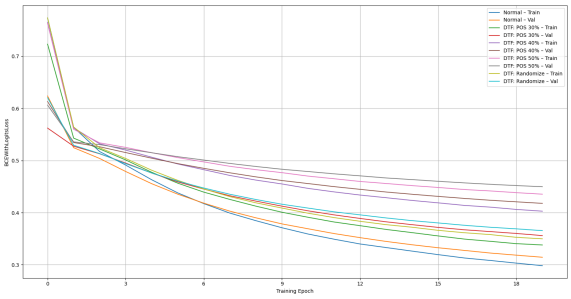

Figure 5: Training and validation loss across epochs for specialist dataset baseline and types of DTF.

Furthermore, Table 3 confirms the same consistent ranking observed for Wikipedia: POS replace-

Metric	Baseline		DTF: POS 50%	
	Var	Std-Dev	Var	Std-Dev
Train loss	1.1×10^{-5}	3.38×10^{-3}	2.5×10^{-6}	1.59×10^{-3}
Val loss	1.2×10^{-5}	3.50×10^{-3}	2.4×10^{-6}	1.54×10^{-3}
Train F1-Micro	4.0×10^{-6}	2.10×10^{-3}	7.0×10^{-7}	8.35×10^{-4}
Val F1-Micro	8.0×10^{-6}	2.82×10^{-3}	2.0×10^{-6}	1.43×10^{-3}

Table 2: Variance and standard deviation across five runs with different random seeds for the baseline and DTF (POS 50%) models.

Model	Δ Train-Loss	Δ Val-Loss	Δ Train-F1	Δ Val-F1
Baseline	0.0000	0.0000	0.0000	0.0000
DTF: POS 50%	0.1482	0.1457	−0.1904	−0.2136
DTF: POS 40%	0.1109	0.1092	−0.1312	−0.1418
DTF: POS 30%	0.0432	0.0469	−0.0651	−0.0745
DTF: Randomize	0.0570	0.0569	−0.0863	−0.0975

Table 3: Difference in loss and F1-score from the baseline model of the specialist data for all DTF models.

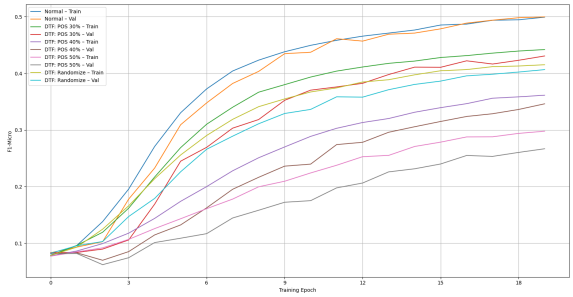

Figure 6: Training and validation F1-micro score across epochs for specialist dataset baseline and types of DTF.

ment at 50% causes the largest drops (e.g., validation F1-micro worsens by 0.2136 relative to baseline), with performance improving monotonically as replacement rates decrease to 40% (Δ=0.1418) and 30% (Δ=0.0745); word-order randomization falls between POS-40 and POS-30 (Δ=0.0975). Absolute degradation is markedly steeper, roughly 2–3× worse across metrics.

The results likely reflects the specialist corpus's extreme terminological density, where domain-specific lexical items — such as technical terms denoting fine-grained grammatical phenomena — carry disproportionate discriminative weight for thematic classification, rendering POS substitution particularly catastrophic compared to Wikipedia's more general-domain prose with broader semantic redundancy. In such narrow, technical registers, DistilBERT's reliance on precise lexical-semantic cues dominates even more markedly than positional or syntactic signals, amplifying overall sensitivity to DTF perturbations. Syntax preservation alone proves insufficient to sustain classification performance when irreplaceable terminology is systematically ablated. These findings underscore a register-dependent vulnerability: while general-domain models tolerate moderate degradation via contextual inference, specialist tasks demand verbatim lexical fidelity, posing steeper challenges for privacy-preserving text transformations in domain-specific NLP applications.

3.2. Experiment B: Train/Test Mismatch

The aim of this experiment is to answer the question: how strongly is the learned representation tied to a specific text form? Two models were trained on the Wikipedia data. The first model (B1) was trained on clean text (no alterations) and evaluated using DTF (50% POS replacement). The second model (B2) was trained on DTF with 50% POS replacement and evaluated on clean text. We chose the 50% POS replacement, because it produces the largest statistically reliable performance drop of approximately 8% micro-F1 loss on the validation set, while still allowing training to converge. This provides a balanced level of difficulty, enabling meaningful analysis of the influence of text form. For training both models, we used early stopping to avoid computational overload.

Figure 7 shows the training and validation loss under representational train–test mismatch. Model B1 (Train clean→Val DTF (50% POS)) fails to transfer from lexically grounded representations to representations where lexical semantics is partially removed and replaced by syntactic category information. The training loss (clean) decreases sharply and continuously and the validation loss (DTF) only decreases at the beginning and plateaus early at a significantly higher level. Overall, training on clean text leads to representations that rely heavily on lexical-semantic cues and exhibit

limited transfer to POS-based DTF. For model B2 (Train DTF (50%)→Val clean), the training loss (DTF) decreases more slowly and remains higher. The validation loss (clean) decreases continuously and reaches lower values than in model B1 (clean→DTF). Training on POS-based DTF induces representations that generalize better to clean text despite reduced optimization efficiency.

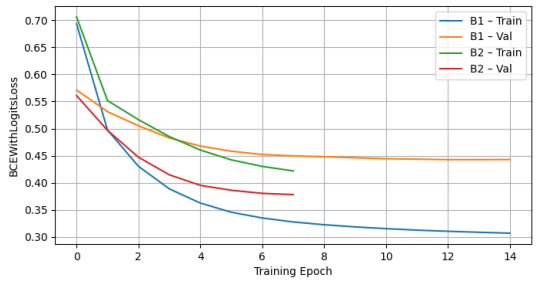

Figure 7: Training and validation loss under representational train–test mismatch. Models are trained either on clean text or on DTF (50% POS replacement) and evaluated on the opposite representation.

Figure 8 presents the F1-micro scores. The training F1-micro score of model B1 rises quickly and high (to 0.47). The validation F1-micro score rises significantly more slowly, reaching saturation at 0.33 and remaining clearly below training F1-micro score. The model learns very well on clean text. However, much of this knowledge cannot be transferred to POS-DTF due to missing lexical-semantic clues. High in-domain performance does not translate to robustness under representational degradation. The training F1-micro score of model B2 rises more slowly and remains below the clean training F1-micro score of model B1 (0.37). The validation F1-micro score rises steadily to reach 0.35, which is above the validation F1-micro score of model B1. According to the lower training F1-micro score, DTF training is more difficult. The model learns more robust features that are less dependent on lexical content. Overall, the generalization of DTF training to clean text is better than the reverse. F1-micro scores reveal an asymmetric transfer behavior: models trained on clean text achieve high in-domain performance but struggle to generalize to POS-based DTF, whereas models trained under POS-induced information limitations exhibit lower training performance, but improved generalization to clean text.

4. Conclusion

Using public-domain data from both Wikipedia and specialist linguistics corpora, this study demonstrates DTF's viability for transformer-based thematic classification by quantifying perturbation effects on learning dynamics and generalization.

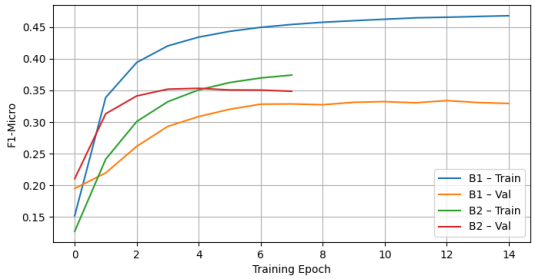

Figure 8: Training and validation F1-micro score under representational train–test mismatch. Models are trained either on clean text or on DTF (50% POS replacement) and evaluated on the opposite representation.

Experiment A yields several key insights: the baseline's superior performance confirms the joint importance of lexical-semantic and sequential information; POS replacement degrades performance more severely than randomization — especially in specialist texts due to terminological density — revealing lexical semantics' greater discriminative power, with performance scaling linearly such that each additional 10% of preserved tokens yields measurable gains. Randomization, in turn, demonstrates syntax's secondary but meaningful contribution.

Experiment B shows that models trained with DTF generalize more effectively to clean text than vice versa. While DTF training leads to slower optimization, it improves cross-representation generalization. In contrast, models trained exclusively on clean text exhibit pronounced sensitivity to representational shifts.

The results indicate that DistilBERT relies primarily on lexical-semantic information, as evidenced by the stronger performance degradation under POS replacement compared to randomization. The asymmetric transfer performance – DTF-trained models generalizing better to clean text than vice versa – demonstrates that training under controlled information loss induces more robust representations, less prone to overfitting on specific lexical cues. This aligns with findings in adversarial robustness studies, where BERT-like models trained on perturbed inputs develop features resilient to distributional shifts, akin to regularization effects observed in TextFooler attacks and adversarial training setups (Jin et al., 2019; Hauser et al., 2021). For DTF applications in computational literary studies, moderate POS replacement rates offer an optimal balance, given the linear scaling of performance

with preserved lexical content, while prioritizing semantics over syntax.

A central issue not addressed in this study is the reconstructibility of the generated DTF texts. Prior to publication, this should be assessed to mitigate potential data protection risks. In our case, however, this concern is moot, as all data are publicly available.

5. Acknowledgements

The work for this paper has been carried out within the SATEK project, funded by the German Research Foundation, grant number 531750631. One of the authors further acknowledges involvement in the Text+ project (grant number 460033370), which contributed to the development of this work.

The authors acknowledge the use of a large-language model (LLM) to draft the descriptive text for the figures in Chapter 3.1.1. The drafts were subsequently revised and edited by the authors, who retain full responsibility for the content.

6. Bibliographical References

Cohere. 2025. Command a: An enterprise-ready large language model. *10.48550/arXiv.2504.00698*.

Tillmann Dönicke, Florian Barth, Hanna Varachkina, and Caroline Sporleder. 2022. MONAPipe: Modes of narration and attribution pipeline for German computational literary studies and language analysis in spaCy. In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS)*, pages 8–15, Potsdam.

Keli Du. 2023. Understanding the impact of three derived text formats on authorship classification with delta. In *DHd*, page 309.

Keli Du and Christof Schöch. 2024. Shifting sentiments? what happens to bert-based sentiment classification when derived text formats are used for fine-tuning.

Jonas Hauser, Zhao Meng, Damián Pascual, and Roger Wattenhofer. 2021. Bert is robust! a case against synonym-based adversarial examples in text classification. *arXiv preprint arXiv:2109.07403*.

Di Jin, Zhijing Shin, Junjie Kim, Jiaqi Duan, Xiangyu Tang, Tongche McCoy, Michael Ramezani, Hannaneh Levine, and Byron C Wallace. 2019. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. *arXiv preprint arXiv:1907.11932*.

Kai Kugler, Simon Münker, Johannes Höhmann, and Achim Rettinger. 2024. Invbert: Reconstructing text from contextualized word embeddings by inverting the bert pipeline. *Journal of Computational Literary Studies*, 2:1–18.

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).

Christian Lang, Roman Schneider, and Karolina Suchowolec. 2018. Extracting specialized terminology from linguistic corpora. In Eric Fuß, Marek Konopka, Beata Trawiński, and Ulrich H. Waßner, editors, *Grammar and Corpora 2016*, pages 425–434. Heidelberg University Publishing, Heidelberg.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Georg Rehm, Andreas Witt, Heike Zinsmeister, and Johannes Dellert. 2007. Corpus masking: Legally bypassing licensing restrictions for the free distribution of text collections. In *Digital Humanities 2007*, pages 166–170.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Roman Schneider and Christian Lang. 2022. Das grammatische Informationssystem grammis – Inhalte, Anwendungen und Perspektiven. *Zeitschrift für germanistische Linguistik*, 50(2):407–427.

Adane Nega Tarekegn, Mario Giacobini, and Krzysztof Michalak. 2021. A review of methods for imbalanced multi-label classification. *Pattern Recognition*, 118:107965.

Juan Terven, Diana-Margarita Cordova-Esparza, Julio-Alejandro Romero-González, Alfonso Ramírez-Pedraza, and Edgar A Chavez-Urbiola. 2025. A comprehensive survey of loss functions and metrics in deep learning. *Artificial Intelligence Review*, 58(7):195.

Grigorios Tsoumakas and Ioannis Katakis. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13.

7. Language Resource References

Kupietz, Marc and others. 2019. *DeReKo - Deutsches Referenzkorpus: wpd19*. Institute for the German Language (IDS), Mannheim.

Schneider, Roman. 2026. *Abgeleitete Textformate zu gesprochener und geschriebener Sprache im Nähe-Distanz-Kontinuum*. Institute for the German Language (IDS), Mannheim.

Training data generation for context-dependent rubric-based short answer grading

Pavel Šindelář, Filip Prášil, Dávid Slivka, Christopher Bouma, Ondřej Bojar
Faculty of Mathematics and Physics at Charles University
sindelar@ufal.mff.cuni.cz, filip.prasil176@student.cuni.cz, david.slivka745@student.cuni.cz,
christopher.bouma496@student.cuni.cz, bojar@ufal.mff.cuni.cz

Abstract

Every four years, the PISA test is administered by the OECD to test the knowledge of teenage students worldwide and allow for comparisons of educational systems. However, having to avoid language differences and annotator bias makes the grading of student answers challenging. For these reasons, it would be interesting to consider methods of automatic student answer grading. To train some of these methods, which require machine learning, or to compute parameters or select hyperparameters for those that do not, a large amount of domain-specific data is needed. In this work, we explore a small number of methods for creating a large-scale training dataset using only a relatively small confidential dataset as a reference, leveraging a set of very simple derived text formats to preserve confidentiality. Using the proposed methods, we successfully created three surrogate datasets that are, at the very least, superficially more similar to the reference dataset than a straightforward result of prompt-based generation. Early experiments suggest one of these approaches might also lead to improved training of automatic answer grading models.

Keywords: Short Answer Grading, Optimal Subset Selection, Unlocking Copyrighted Collections

1. Introduction

Access to high-quality domain-specific data is commonly one of the biggest obstacles in natural language research, especially for under-explored tasks.

Data for many domains is publicly available in large quantities. However, certain potentially highly scientifically valuable tasks and domains are heavily restricted due to privacy regulations and intellectual property protections. The lack of access to these data restricts the research on it to specific institutions, and also limits the reproducibility of such research.

One example of such highly restricted but interesting domains is the evaluation of short open-ended answers. Training a robust evaluator that can consistently grade open-domain, open-ended answers requires a substantial amount of data from a variety of domains, with diverse question types, and containing both positive and negative examples. It should also include grading rubrics that explain the grading procedures for each question to reduce grading ambiguity.

Data collection in this domain is often limited by the right to privacy of students. Most openly available datasets fail to meet our criteria. They are typically restricted to a single domain, contain only correct answers, or lack grading rubrics. With the goal of creating the training data for a subtask within the Sensemaking 2026[1] task at the ELOQUENT lab[2], our research team was granted privileged access to a closed-source dataset containing the relevant information from previous iterations of the Programme for International Student Assessment (PISA) test by OECD. However, because access to this dataset is bound by a non-disclosure agreement, it cannot be published or shared directly.

To address these issues and stay within the bounds of our non-disclosure agreement, we decided to use the confidential data to generate an open proxy dataset to be used instead. However, generating this data is not as straightforward as it might first seem, because using the original data in the generation process directly is difficult and, depending on the method used, could constitute a breach under the NDA. Unfortunately, generating entirely synthetic data would likely produce data that are not very similar to the reference. To mitigate this, we leveraged Derived Text Formats (DTFs). Specifically, we computed a variety of linguistic and semantic features for the reference dataset and the synthetic samples. We then used those features to select the most similar synthetic samples to create the final dataset. However, since these features are almost entirely focused on the relations between different pieces of text defining a datapoint, it is difficult to compare our method with most other works using DTFs. For example, practically no information about the original datapoints can be recovered from our DTFs without access to parts of the original datapoints.

The code used for subsampling and the full gen-

[1] https://ufal.mff.cuni.cz/see/sensemaking-2026

[2] https://eloquent-lab.github.io/

Proceedings of Leveraging Derived Text Formats to Unlock Copyrighted Collections for Open Science (DTF) @ LREC 2026, pages 44–50
12 May 2026. ©ELRA Language Resources Association (ELRA), 2026

erated dataset can be found in a GitHub repository located at https://github.com/sindelarp/training-data-generation-for-context-dependent-rubric-based-short-answer-grading.

2. Related works

There have been several other works on optimal subset selection, often referred to as dataset condensation, distribution matching, or dataset alignment.

GRAD-MATCH (Killamsetty et al., 2021) selects a subset by approximately minimizing the gradient difference with the validation set for the specific model state at the chosen training step.

LESS (Xia et al., 2024) uses a procedure similar to Tracin (Pruthi et al., 2020) to estimate sample influence on minimizing model loss on a reference dataset by looking at multiple checkpoints during one training run trained on a static dataset. It then selects the samples with the most influence to create a new dataset.

While potentially less powerful, our method requires significantly less computing power than most of these methods by leveraging the fact that our data has a large number of dependent fields. An added advantage of our method is that the resulting features are potentially shareable. Thus, the generation process can be trivially extended to other datasets we wish to subsample.

The paper Guo et al. (2022) gives an overview of a few methods that are more similar in computational requirements to our method. These, however, deal mostly with monolithic texts and focus on selecting the best subsample without the benefit of a reference dataset, and so models trained on them often have worse results on the test dataset than models trained on the full dataset.

3. Methodology

3.1. Reference Dataset

The target distribution for our data generation was derived from the dataset of selected questions and answers from the previous iteration of the PISA test, provided to us by the OECD under a non-disclosure agreement. The dataset contained the following fields:

- **Context:** A text containing all of the relevant context for the question.
- **Question:** A question regarding the context.
- **Grading Rubric:** A grading rubric defining the criteria an answer should meet to be given full, partial, or no credit.
- **Answer:** An answer to the question.
- **Rating:** A ternary rating of the answer, deciding whether it deserves full credit (FC/numerical label 2), partial credit (PC/numerical label 1), or no credit (NC/numerical label 0).

Due to the non-disclosure agreement, these items could not effectively be used for anything beyond a closed evaluation. To solve this issue, we decided to treat the dataset as a collection of linguistic and semantic features, which we would then target to mimic in our generated dataset, which would not have such limitations. For the present experiment, we decided to focus only on working with the English subset of the data.

3.2. Synthetic Candidate Generation

The initial synthetic dataset had to be large enough to allow for effective sample selection. This was achieved through a multi-stage pipeline that used openly available sources from the Internet and Large Language Model (LLM)[3] prompting.

3.2.1. Context Extraction

As the base data for the creation of the contexts, we extracted plain text from several openly accessible websites. These texts were then segmented into sentences using regular expressions. Consecutive sentences were concatenated until they reached a target word count randomly selected between 150 and 800 words. The last sequence for each source was kept only if it contained at least 150 words and was discarded otherwise.

The resulting text chunks were then given to an LLM with the instruction to clean the text of any relics introduced during parsing or transcription. These cleaned texts were then used as contexts.

3.2.2. Question Generation

The questions were generated using a 2-stage process. An LLM was given a context and was instructed to generate 1-5 open-ended questions asking about the information from it. It was also instructed to generate questions only satisfying these criteria:

- The question is answerable using only the information in the context.
- The question is not a subset, non-standalone follow-up question, or rephrasing of any of the previously generated questions.
- The question tests the logical understanding of the text, not just factual recall.

[3]The open-weights gpt-oss-120b model was used for all parts of the synthetic candidate generation.

- The question does not ask for anything subjective. There must always exist a single or only a few correct answers to it.

Another LLM was then given all of the generated questions and was tasked with deciding whether each of the given questions follows these criteria. The questions that were identified as compliant were kept; the rest were discarded.

3.2.3. Grading Rubric Generation

The grading rubrics were generated in three parts: full credit, partial credit, and no credit. Similarly to question generation, the generation process had two stages.

In the first stage, an LLM was given a context and a question and was tasked with generating all of the three grading rubric parts so that they follow these criteria:

- The full credit section describes a fully correct answer to the question.
- The partial credit section describes an answer that is not fully correct (unambiguously), but shows at least partially correct reasoning and understanding of the context and the question.
- The no credit section describes an answer that is fully incorrect or shows little to no correct reasoning and understanding of the context and the question.
- The individual sections do not have any overlap, which would cause ambiguity when using them to grade.

As an additional criterion, we instructed the LLM to make the sections similar in quality and detail to three examples of grading rubrics openly available on the OECD website.

In the second stage, another LLM (the judge) was given the context, question, and the generated grading rubrics and was tasked with deciding whether it follows the generation criteria or not. In the case the text did not follow some criteria, the judge was instructed to generate a critique explaining which criteria it does not follow and why. This explanatory text was then provided along with the instructions from the first stage, and this process was repeated until the judge LLM decided that all of the criteria were satisfied. A limit of 10 iterations was set, but it was never reached.

3.2.4. Answer Generation and Grading

The answers were again generated using a 2-stage process. In the first stage, the LLM was instructed to generate five candidate answers for a given context and question, following a given grading rubric section and additional instructions. This was done for each correctness level (full/partial/no credit) separately, with some common and some level-specific instructions. The common instructions were:

- The answer follows the given grading criteria.
- The answer is not a copy of any of the previously generated answers. It can contain the same information, but it must be worded differently or include/exclude additional information that does not change the correctness of the answer.
- Some of the answers may contain common typos and grammatical issues, such as their/they're errors[4] or slightly awkward phrasing.
- The answers should be written as a high school student might write them under time pressure.

During the full-credit-level generation, the model was not given any additional specific instructions.

During the partial-credit-level generation, the specific instruction was:

- The incorrectness in the answers comes from a plausible misinterpretation of the text or from simply excluding information required to be fully correct.

During the no-credit-level generation, the specific instruction was:

- The incorrectness in the answer comes from a plausible misinterpretation of the text, from simply excluding necessary correct information, or trying to "talk around" the question without really answering it.

Similarly to the first stage, the second stage was also split by the correctness levels. For each correctness level, an LLM judge was given the context, the question, the grading rubric section, the instructions that the generation process of the answers was supposed to follow, and the answers themselves. Its task was to decide whether each of the answers was generated by correctly following the generation instructions. The answers for which this was not the case were discarded.

3.3. Feature Extraction

To better understand the generated dataset, we extracted the following features for each datapoint:

- **BAAI/bge-m3[5] context question cosine similarity** is the cosine similarity between dense

[4] We opted to illustrate a spelling error to increase the chances that the model will actually generate some.

[5] https://huggingface.co/BAAI/bge-m3

embeddings of the context and the question, where embeddings are produced by the **BAAI/bge-m3** encoder.

- **BAAI/bge-m3 context answer cosine similarity** is the cosine similarity between **BAAI/bge-m3** embeddings of the context and the answer.
- **BAAI/bge-m3 rubrics/FC answer cosine similarity** is the cosine similarity between **BAAI/bge-m3** embeddings of the full-credit rubric text and the student answer.
- **recall 2gram FC answer** is the number of shared bigrams between the full-credit rubric and the answer divided by the length of the answer. The value **recall 2gram NC answer** is similar, but uses the no-credit rubric instead of the full-credit rubric.
- **answer length** is the length of the answer measured as the number of tokens.
- **answer lexical density** is the proportion of answer tokens tagged as content words (nouns, verbs, adjectives, adverbs) by NLTK POS tagging.
- **Jaccard 1gram question answer** and **Jaccard 1gram context answer** are the Jaccard similarity between the unigram sets of the question and the answer and between the unigram sets of the context and the answer, respectively.
- **recall 2gram question answer** and **recall 2gram context answer** are the fraction of answer bigrams that also appear in the question and answer bigrams that also appear in the context, respectively.
- **recall 2gram context overlap with answer minus question** is a recall measure. Let S be the set of n-grams in the answer with n-grams from the question removed (set difference). This feature is the fraction of n-grams in S that also occur in the context.
- **tfidf cosine question answer** and **tfidf cosine context answer** are the cosine similarity between TF–IDF vector representations of the question and the answer and of the context and the answer, respectively.
- **precision 1gram question answer** is the fraction of question unigrams that also appear in the answer.
- **recall 1gram question answer** and **recall 1gram context answer** are the fraction of answer unigrams that also appear in the question and answer unigrams that also appear in the context, respectively.
- **recall 1gram context overlap with answer minus question** is a recall measure similar to **recall 2gram context overlap with answer minus question**.

3.4. Data selection

To make the data resemble the OECD data more closely, we computed the features mentioned above for each datapoint in the synthetic candidate dataset and the OECD dataset. We then compared three feature-based matching methods.

First selection method In the **first selection method**, we use a very small dataset and simply take the 5% of samples that are the most similar to the mean of the OECD features of their label (i.e., we only compare FC with FC, PC with PC, and NC with NC) under the L2 metric.

Second selection method As shown in the section below, the **first selection method** did not yield any noticeable improvement in **E1** and **E2**. We identified two potential issues that we tried to alleviate in the **second selection method**.

1. The lack of domain coverage after subsampling. – We take the best 5% in every domain separately to try to maintain data diversity. For the few-shot examples, we sample 2 datapoints from the subsampled data for each domain.
2. The potential of the single mean being too reductive. – Instead of looking at the similarity to a single mean over all OECD data, we decided to look at the similarity to the closest of eight different representatives of the OECD data feature space. These representatives were selected using k-means with k-means++ initialization.

Third selection method In the **third selection method**, we tried to further relax the conditions. Instead of demanding that the features not reveal information about individual samples, we used access to the unreduced features to select the training dataset. We decided to remove the direct reliance on distance comparisons and instead look at the rank of the first five closest datapoints from the reference dataset. To get the negative score of the sample a, we sort the reference dataset and the generated dataset together by their ascending L2 distance to a, and average the indices of the first five samples belonging to the reference dataset in this ordering.

4. Results

We decided to compare the usefulness of the subsampled dataset with the full dataset using multiple comparison methods.

E1 For a fast preliminary experiment, we examined how different the respective dataset utility is by sampling few-shot examples from both the full dataset and the dataset sub-sampled via the selection method and comparing the accuracy of *Qwen3-8B* using these two different parameters on the confidential data. We used a simple prompt generated using DSPy (Khattab et al., 2022).

E2 Here we check how good our subsampled subset would be for model selection. If the results of metrics on the selected subset are closer than on the whole dataset, it could be expected that if the subset was used to select between different model architectures, different models, training hyperparameters, or different model checkpoints, it would lead to better results on the confidential dataset than selecting based on the full dataset.

E3 For an additional comparison, we trained a small model on the dataset resulting from the application of the selection method. To make the small size of the dataset work with the model and make the training more stable, we chose to use the "tasksource/ModernBERT-base-nli" as a base for our training. This model is based on the pretrained model ModernBERT and further finetuned on a mix of natural language inference datasets. As a sort of regularization, we added a subset of the 20% samples with the longest tokenized length from the MNLI dataset. We used a learning rate of 2e-5 and did gradient accumulation over 32 samples in each step. These were selected by training on an unrelated training dataset and scoring on a manually verified development dataset, where the model achieved a mean accuracy of 0.722 during the second half of training. We decided to treat the evaluation on the confidential dataset during training as a set of paired experiments. We evaluated the model trained on a randomly subsampled dataset and a model trained on a dataset subsampled using the method after every 187 steps, and then calculated the Wilcoxon signed-rank test statistic (Wilcoxon, 1945). We note that while consecutive accuracies of the model on the reference set during training are not independent, they are approximately independent enough for an initial estimate. Ideally, we would do many independent runs and analyze accuracy at each number of steps separately.

4.1. First selection method

The feature similarity results of **first selection method** can be seen in Table 1. The improvement was particularly apparent for answer length, which is, of course, a trivial feature.

On the other hand, some features, such as the bigram recall between the no credit rubric and the answer, now differ more. For this particular feature example, it is most likely caused by the fact that the no-credit conditions were often described abstractly.

	original	selected
BAAI/bge-m3 context question cosine similarity	0.009	0.026
BAAI/bge-m3 context answer cosine similarity	0.047	0.009
BAAI/bge-m3 rubrics/FC answer cosine similarity	0.074	0.053
recall 2gram FC answer	0.045	0.087
recall 2gram NC answer	0.005	0.028
answer length	30.691	16.987
answer lexical density	0.030	0.014
jaccard 1gram question answer	0.054	0.024
jaccard 1gram context answer	0.015	0.001
recall 2gram question answer	0.009	0.017
recall 2gram context answer	0.143	0.130
recall 2gram context overlap with answer minus question	0.157	0.140
tfidf cosine question answer	0.101	0.046
tfidf cosine context answer	0.025	0.014
precision 1gram question answer	0.197	0.110
recall 1gram question answer	0.028	0.055
recall 1gram context answer	0.141	0.093
recall 1gram context overlap with answer minus question	0.142	0.083

Table 1: Absolute differences between the overall feature means between confidential data and datasets.

As we can see in Table 2, our method did not improve model performance in comparison method **E1**. This could be because **first selection method** is too coarse for the selected model to benefit from the selected few-shot data more.

We can see in Table 2 that when we used method **E2**, the results on the selected data seem to be overall better than on both the entire dataset and on the confidential data, indicating it would likely not be a good dataset for model selection.

This indicates that **first selection method** is not strong enough to balance out the specific biases this dataset elicits in our model.

fewshot source	eval target	Accuracy	Quadratic Weighted Kappa
All	Confidential	0.000	0.000
Selected	Confidential	-0.040	-0.010
All	Selected	+0.250	+0.380
All	All	+0.120	+0.410

Table 2: Comparison of performance measures. Note that the accuracy was calculated on a dataset resampled to be balanced. Quadratic Weighted Kappa is computed using the respective numerical labels.

4.2. Second selection method

We later tried to evaluate the **second selection method** on a larger dataset. The results are very similar to the results of **first selection method**, and can be seen in Table 3.

Again, there seems to be a small decrease in accuracy, 0.58 vs 0.61, in **E1**. This could be because our method does not maintain diversity. When comparing how useful the data from **second selection method** is for model selection. We can see in the results of **E2** in Table 3 that the accuracy on the selected subset, 94 vs 92, is very similar to the accuracy on the whole dataset.

fewshot source	evaluation target	Accuracy	Quadratic Weighted Kappa
All	Confidential	0.000	0.000
Selected	Confidential	-0.033	-0.019
All	Selected	+0.333	+0.469
All	All	+0.313	+0.456

Table 3: Comparison of performance measures on the second method, see Table 2.

As an initial estimate using **E3**, we got the high p of 0.715, significantly above the threshold of 0.05. This suggests further experiments would likely be fruitless.

4.3. Third selection method

Because we found it the most rigorous, we only evaluated the third selection using **E3**, the differences used for the test can be seen in Figure 1. The result was a p of 0.00091, below the threshold of 0.05. Over the training, the model on the subselected data achieved an average gain of 0.0206 accuracy.

Afterwards, we reran the experiments twice more for each dataset with a different order of the training set. Since the weights of all layers were kept from the base model, this was the only way of randomizing the training run we attempted.

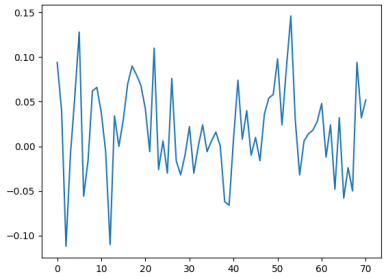

Figure 1: Accuracy of the model training in the subsampled dataset minus the accuracy of the model training on a randomly subsampled dataset.

Here, we did a Wilcoxon signed-rank test by selecting a pair of training runs to compare separately at each step of evaluation, using each accuracy value exactly once. The result was a p of 0.00102, below the threshold of 0.05. Over the training, the models on the subselected data achieved an average gain of 0.0116 accuracy.

5. Conclusion

We presented a method for deriving a dataset of contexts, questions, scoring rubrics, and scored answers from a closed seed set, namely the OECD PISA test items. We generate these items using an open-weight LLM and compare different methods of optimal subset selection.

Based on our experiments, only the most relaxed method (**third selection method**) has achieved significant results. This direction of research seems promising, but determining whether such methods truly select useful data for the purposes of training in general would require more thorough experiments on many different models.

In the future, we would like to iterate on the subset selection methods that use reduced features. It seems sets of simple, averaged features are often not expressive enough, or they need to be weighted using a complex set of hyperparameters, which would need to undergo a resource-intensive selection process. Likely, a more expressive set of features will be necessary. These methods might work better on more heavily structured datasets.

6. Limitations

Our evaluations generally work with only a single run, and a more rigorous repetition of the experiments on many different datasets and models would be necessary to test the potential of the method.

7. Declaration on Generative AI

LLMs were used for the purposes of a preliminary search for style and grammar errors. This work contains no LLM-generated text.

8. Acknowledgements

The authors acknowledge the funding from the Project OP JAK Mezisektorová spolupráce Nr. CZ.02.01.01/00/23_020/0008518 named "Jazykověda, umělá inteligence a jazykové a řečové technologie: od výzkumu k aplikacím", the support of the National Recovery Plan funded project MPO 60273/24/21300/21000 CEDMO 2.0 NPO and the support by EC Digital Europe Programme (DIGITAL) grant number 101195233 (OpenEuroLLM).

Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

9. Bibliographical References

Chengcheng Guo, Bo Zhao, and Yanbing Bai. 2022. Deepcore: A comprehensive library for coreset selection in deep learning. In *International Conference on Database and Expert Systems Applications*, pages 181–195. Springer.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive NLP. *arXiv preprint arXiv:2212.14024*.

Krishnateja Killamsetty, Sivasubramanian Durga, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. 2021. Grad-match: Gradient matching based data subset selection for efficient deep model training. In *International Conference on Machine Learning*, pages 5464–5474. PMLR.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. 2020. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33:19920–19930.

Frank. Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics*, 1:196–202.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. Less: Selecting influential data for targeted instruction tuning. *arXiv preprint arXiv:2402.04333*.

DUO_DE A1: An Annotated Corpus of Online Learning Material for Beginning Learners of German as a Foreign Language

Jammila Laâguidi[1], Vitaliia Ruban[1], Ronja Laarmann-Quante[1], Anastasia Drackert[2,3]
Ruhr University Bochum, Germany
[1]Faculty of Philology, Department of Linguistics
[2]Faculty of Philology, Institute for German Language and Literature
[3]g.a.s.t. (Society for Academic Study Preparation and Test Development)

Abstract

This paper describes the creation of DUO_DE A1, a corpus based on A1-level learning material from the Deutsch-Uni Online (DUO) language courses for German as a foreign language. We split the material into small segments and manually annotated each with fine-grained information such as the type of segment (e.g. task description, description of grammar), the medium (e.g. text, table, audio), the text units it contains (e.g. words, phrases, sentences) and other special features (e.g. marking cloze texts). Furthermore, we automatically tokenized, POS tagged and lemmatized the corpus and compared the performance of three models on these steps for different kinds of segments. We publish the created corpus in a manner that respects copyright, releasing all structural features, metadata and POS tags.

Keywords: German as a foreign language, corpus creation, learning material, beginning learners, Common European Framework of Reference

1. Introduction

When learning a foreign language, structured material like traditional textbooks or online learning material typically plays an important role. Especially in the beginning of the learning process, the learning material to a large degree determines what vocabulary and grammatical structures the learners are exposed to. Analyzing the learning material is thus of great interest to foreign language acquisition research, see e.g. the comprehensive overview of studies on the language in English as a foreign language (EFL) textbooks in Le Foll (2023). However, such studies can typically not be replicated because code and data are not made available, often for copyright reasons (Le Foll, 2023, p.66).

The goal of the present paper is to contribute to transparency in corpus linguistic research on language learning material. We present DUO_DE A1, a corpus compiled from an online course for learning German as a foreign language (GFL) on level A1 of the Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001). While the texts themselves cannot be published for copyright reasons, we release all structural features, metadata and linguistic annotations (part of speech tags).

Preparing the learning material for corpus linguistic research requires a careful consideration of several aspects. For example, the material does not only consist of coherent standard German text. There are also e.g. word lists or tables or tasks that present incomplete sentences for didactic purposes (like cloze texts) or that require the learner to correct the word order. Such elements need to be identifiable in the corpus, as they might distort certain analyses e.g. regarding collocations or syntactic dependencies. Furthermore, the material does not only contain the primary learning content but also e.g. task instructions or metalinguistc information about grammar. These elements should be clearly distinguishable as they may play different roles depending on the research question.

In this paper, we describe how we addressed such aspects. We split up the learning material into small segments and manually annotated them with various information about the type of text (e.g. task instruction or grammar description), the medium (e.g. text, table or audio), the text units it contains (e.g. words, phrases or sentences) and other special features like marking cloze texts (Sec. 3). Furthermore, we automatically tokenized the texts and added part of speech (POS) tags

Proceedings of Leveraging Derived Text Formats to Unlock Copyrighted Collections for Open Science (DTF) @ LREC 2026, pages 51–62
12 May 2026. ©ELRA Language Resources Association (ELRA), 2026

and lemma information. We compare the performance of three models on these tasks for different types of segments (Sec. 4). We hypothesize that segments containing coherent sentences would be processed more accurately than those consisting of isolated words or phrases (see also Volodina et al., 2014). Finally, we present an analysis of the composition of the whole corpus (Sec. 5) and describe the JSON format in which it is stored (Sec. 6). For copyright reasons, the word forms and lemmas can only be made available for research purposes upon request. All other parts of the corpus including structural information, metadata and POS tags are made publicly available under a CC BY-NC-SA 4.0 license. The whole corpus titled *DUO_DE A1: An Annotated Corpus of A1-Level Learning Material from the Deutsch-Uni Online Courses for German as a Foreign Language* can be accessed via the following link: https://doi.org/10.5281/zenodo.19113347 (Laâguidi et al., 2026).

2. Related Work

Corpus-based studies on GFL learning material have typically worked with corpora compiled for the purpose of the particular study without sharing the data or annotations. The data is often pruned according to the specific research questions as the following examples show. Furthermore, a detailed description of the creation of the corpus is rarely the focus of the researchers.

Bautista Zambrana (2018) studied phraseological units in learning material. Their corpus consisted of one GFL textbook and workbook on A1 level, which they divided into three subcorpora (written, oral and exercises). They excluded single word forms, morphological units, task instructions, grammar reference sections and vocabulary lists. The remaining corpus comprised 20,806 tokens.

Behnke (2023) investigated how GFL textbooks deal with language change phenomena. They compiled a corpus from five GFL textbook series across CEFR levels A1 to C1. They excluded task instructions and tasks that explicitly targeted one of the phenomena under investigation. The paper does not provide information about the number of tokens.

A different example is the DAFLex project (François et al., 2021). They used a corpus of GFL textbook reading activities and simplified readers to build a CEFR-graded lexicon of receptive vocabulary. The texts were lemmatized and POS tagged, resulting in a lexicon of 41,646 lemma-POS pairs. While the corpus is not public, there is an online tool that allows users to check the frequency of a given word according to the CEFR levels. Furthermore, it can analyse a text and assign a CEFR level to each word (François et al., 2021).

For other languages, there are a few well-prepared and documented textbook corpora. The corpus of Textbook Material (TeMa, Meunier and Gouverneur, 2009) consists of 724,174 words from 32 volumes of English for general purposes coursebooks. It is stored in XML format and comes with a detailed annotation scheme focusing on vocabulary exercises. The corpus itself is not published but sections of it can be accessed for research purposes upon request.[1] The Textbook English Corpus (TEC, Le Foll, 2023) was compiled from 43 EFL coursebooks and comprises 3,023,958 words. It is stored in XML format and all material is annotated for register. While the texts are not available for copyright reasons, all metadata, annotations and code for processing the corpus are published.

A textbook corpus with extensive annotations on various levels is the Corpus of CEFR-based Textbooks as Input for Learner Levels' modelling (COCTAILL, Volodina et al., 2014). It was compiled from twelve Swedish as a foreign language coursebooks and comprises 708,589 tokens. Text passages are annotated for topics and genres and all other material has annotations about target skills, target competences, activity types, activity formats and linguistic units. Furthermore, the data was automatically POS tagged, lemmatized and annotated for syntactic dependencies. While for copyright reasons the corpus is not freely available, for research purposes it can be browsed and parts of the corpus can be downloaded as a bag of sentences upon request.

[1] https://www.uclouvain.be/en/research-institutes/ilc/cecl/tema

3. Data

The DUO_DE A1 corpus consists of data from *Deutsch-Uni Online (DUO)* ('German-University online')[2], a language learning platform for learning German on CEFR levels A1 through C1. Its content focuses on German within a university context. For the DUO_DE A1 corpus, we used all learning materials from the A1 level.[3]

3.1. Structure of the Material

In the following, we describe the internal structure of the learning material. The A1 level consists of two **courses** A1.1 and A1.2. Each course comprises six chapters that each represent a **scenario** such as *Ein Kochabend mit Freunden* 'An evening of cooking with friends' (A1.1. ch. 3). Each chapter is divided into six **phases**. Each of these phases consists of multiple **learning activities** that deal with an overarching topic, e.g. how to use the verb *(to) like* while talking about the seasons (A1.1, ch. 5, phase 2, learning activity 2). Within each learning activity, students are given multiple **tasks** that correspond to the respective topic and have a different focus (grammar, vocabulary, etc.). The kinds of tasks vary greatly and include, for example, cloze texts, listening exercises, or free writing tasks. Each task can be further subdivided into different **segments**, containing different kinds of texts. For example, there can be a task description, further input, e.g. an audio with a transcription, and an editing field for the learner's answer.

3.2. Data Extraction

The data was extracted manually by the first two authors of this paper. We annotated the data based on **segments** of tasks as described above, i.e. the segments serve as our *unit of observation* (Le Foll, 2023, p.76). For the most part, this meant copy-pasting any data that can be read by the learners or transcriptions of texts that were presented as audio. For images containing text that could not be copy-pasted, we typed the text as accurately as possible. For tables, we structured the content into paragraphs for titles and table content.

3.3. Segment-Level Metadata

Each segment is annotated with structural information and some additional information pertaining to the segment as a whole. In this paper and in the published JSON format (see Sec. 6), we refer to this as metadata, which must not be confused with corpus-level metadata such as the language and creators of the corpus, which can be obtained directly from the data repository (Laâguidi et al., 2026). The annotated information comprise the following:

- **course** A1.1 or A1.2
- **scenario** 1-6
- **phase** 1-6
- **learning activity** 1-6 (varies)
- **task** 1-14 (varies),
- **segment**, type of segment, one of
  - title
  - task type
  - situation and instructions
  - task description
  - media
  - editing field
  - further information
  - model solution
  - additional information
  - description of grammar
  - description of learning content
  - list of new expressions
- **medium**, one of
  - text
  - table
  - audio
  - video
- **text unit**, one of
  - words
  - phrases
  - sentences
  - dialogue
  - other (e.g. suffixes)
- **comments**, optional, one or multiple of
  - multiple choice cloze text
  - incomplete cloze text

[2] https://www.deutsch-uni.com/de/

[3] We extracted the data from files from which the content had been entered into the online platform. There may be minor differences between these files and the final online version.

  - intentionally incorrect grammar
  - wrong word order
  - unsegmented
  - model solution
  - partly crossed out
  - duplicate numbering
  - without text
- **Other**, optional, any other comment (free text)

The annotations of the **course**, **scenario**, **phase**, **learning activity** and **task** can be used to locate the segment in the course, e.g. to analyze progression in the material.

The annotation of the type of **segment** follows the inherent structure of the material and allows to distinguish between different kinds of text that a learner encounters. For example, the *editing field* and *media* typically contain the primary learning content of level A1 about a certain topic, for example a cloze text about groceries. The *instructions* or *task description* in turn may contain vocabulary or grammatical constructions that are not targeted with the current task but which are needed to convey the task. Other categories like *description of grammar* or *description of learning content* consist of even more abstract language, e.g. about grammatical categories like *unbestimmter Artikel im Nominativ* ('indefinite article in nominative case').

The **medium** informs about the mode of presentation (*text* vs. *audio* vs. *video*). If text is presented as a table, its layout is not preserved in our corpus but it is annotated with the category *table* so that this information can be considered.

Under **text unit** it is stored whether the segment consists of whole sentences or a dialogue, which can be analyzed syntactically, or if it is only a list of words or phrases, where e.g. dependency parsing may not be meaningful.

We defined a set of annotations stored under **comments** which can be taken into account when further processing the corpus. For example, an *incomplete cloze text* (with gaps such as *Wir ___ viel Spaß!* 'We ___ a lot of fun!') or a multiple choice cloze text (such as *Ich arbeite seit drei Jahren / vor einem Jahr als Ingenieur* 'I have been working as an engineer for three years / one year ago') may present challenges for parsing as the sentences are incomplete or contain superfluous words. There are also sorting tasks which present learners with a wrong word order. Such sentences should not be taken into account when e.g. analyzing co-occurrences of words. When there is no transcription for audio material within a listening task, we include this segment in the corpus to retain the course structure. In this case, it does not contain any tokens but only metadata, in particular the *without text* comment.

4. Automatic Linguistic Annotation

4.1. Gold Standard Annotation

We created an evaluation set with manual gold standard annotations for lemmas and POS tags based on the STTS tagset (Schiller et al., 1999) using the annotation platform INCEpTION (Klie et al., 2018). We aimed at making the evaluation set as diverse as possible, covering all categories for *segment*, *medium*, and *text unit*. We included a wide range of different kinds of learning material from different chapters, such as cloze texts with gaps, word lists both with and without articles, words annotated with grammatical categories or suffixes like *Sg.* (Singular) or *-e*, phrases or isolated words without context as well as dialogues, both extended ones and those consisting of only two sentences. Our evaluation set consists of 843 tokens. The first two authors of this paper independently annotated the texts and subsequently discussed diverging cases for reaching a gold standard. Most cases were clear, only a few tags were debatable, for example *Arbeiten im Semester?*, where *Arbeiten* could be read as an infinitive verb ('Working during the semester') or as a plural noun ('Work during the semester').

4.2. Models

We compare three different models for automatic POS tagging and lemmatization, the **small** and **medium** German models of **spaCy** (Honnibal et al., 2020)[4] and **Stanza** (Qi et al., 2020)[5]. In order to avoid alignment issues due

[4] spaCy v3.8.4, models de_core_news_sm v.3.8.0 and de_core_news_md v.3.8.0

[5] Stanza version 1.10.1

Model	Accuracy
spaCy small	0.886
spaCy medium	0.902
Stanza	0.896
Stanza + APPRART	0.910

Table 1: Overall accuracy for POS tagging.

to differences in tokenization, we use the tokenization from the INCEpTION platform for the evaluation set for all models. Differences in tokenization arise, for instance, because Stanza, trained on Universal Dependencies treebanks, tokenizes contractions of prepositions and definite articles such as *im* (= *in dem* 'in the') as two tokens *in* and *dem*, while spaCy treats them as a single token. For the evaluation, we measure accuracy, precision, and recall on the evaluation set using *scikit-learn* (Pedregosa et al., 2011).

4.3. Evaluation of POS Tagging

The upper part of Table 1 shows the overall accuracy of the three models across all POS tags. Their performance just around .90 is only slightly worse than what Ortmann et al. (2019) report for other registers. While numerically, spaCy (medium) and Stanza perform almost on par, looking at the tagging errors that each model makes reveals some important qualitative differences. We restrict the following discussion to Stanza and the spaCy medium model, which performed slightly better than the small model.

4.3.1. Confusion of POS Tags

For spaCy medium, there are 83 tagging errors in total. The most frequent confusion (12 times) is to tag a normal noun (NN) as a proper noun (NE). Stanza has 88 tagging errors and the confusion of NE for NN only happened seven times. This indicates that Stanza has a broader lexicon because normal nouns such as *HNO* ('ENT specialist'), *Nachhilfelehrer* ('tutor'), *Babysitter* ('babysitter') and *Animateur* ('entertainer') were treated as proper nouns by spaCy, a frequent fallback tag for unknown words, but recognized correctly by Stanza.

The most frequent confusion of Stanza (19 times) is to tag non-words (XY) as punctuation marks ($()). Firstly, this concerns emojis and secondly, we used the tag XY to tag lines indicating gaps in a cloze text. It is very plausible to tag such cases as punctuation marks instead and while it affects Stanza's overall accuracy in this evaluation, in practical applications, these word classes should only play a minor role. spaCy, on the other hand, tags them partly as foreign words (FM) and partly as adjectives (ADJA). This could impact analyses of the learning material in an undesirable way.

4.3.2. Creation of a Combined Model

Based on the previous observations, we decided for Stanza to be more appropriate than spaCy for tagging the DUO_DE A1 corpus. However, as addressed in Sec. 4.2, Stanza treats contracted prepositions and articles such as *im* as two tokens. We find this undesirable because there are differences in meaning between the contracted forms and the split forms and they cannot be used interchangeably (see e.g. Cieschinger, 2016), hence keeping the original form is important. Contracted prepositions and articles always get the POS tag APPRART in the gold standard, which Stanza never assigns because it was not present in its training data. Instead, Stanza would assign the tags for article (ART, 11 times) or preposition (APPR, 1 time) in our evaluation set. Therefore, we decided to create a combined POS tagging model which uses Stanza's POS tags except for tokens where spaCy assigns APPRART, which then overwrites the Stanza tag. This combined model achieves a higher accuracy (0.91) than the individual models, see Table 1.

Figure 1 shows the confusion matrix of its remaining tagging errors and Table 2 the precision, recall and F1 score per POS tag. We can see that besides NE, XY and $(that were discussed above, most of the tags with an F1 score $<$ .90 occurred less than 10 times: ITJ (5), PDAT (1), PDS (1), PIAT (8), PIS (1), PTKANT (2), PWAV (6), VVIMP (1). Adverbs (ADV) only have an F1 score of .81 partly because Stanza tags answer particles (PTKANT) as adverbs, and infinitive verbs (VVINF) are challenging (F1 = .86) because they are of-

XPOS	precision	recall	F1	# toks.
$(	.82	.94	.88	94
$,	1	1	1	24
$.	.97	1	.98	85
ADJA	1	.85	.92	13
ADJD	.92	1	.96	12
ADV	.73	.92	.81	12
APPR	1	.97	.99	37
APPRART	1	1	1	12
ART	1	.95	.97	60
CARD	1	1	1	34
ITJ	0	0	0	5
KON	.94	.94	.94	17
NE	.72	.91	.80	43
NN	.94	.94	.94	160
PAV	1	1	1	1
PDAT	0	0	0	1
PDS	.33	1	.50	1
PIAT	.71	.62	.67	8
PIS	.20	1	.33	1
PPER	.95	1	.97	55
PPOSAT	1	.85	.92	20
PTKANT	0	0	0	2
PTKNEG	1	1	1	2
PTKVZ	1	1	1	1
PWAT	1	1	1	1
PWAV	1	.33	.5	6
PWS	1	1	1	7
VAFIN	1	.87	.93	31
VMFIN	1	1	1	9
VMINF	1	1	1	1
VVFIN	.88	.97	.92	38
VVIMP	0	0	0	1
VVINF	.94	.79	.86	19
VVPP	.88	1	.93	7
XY	1	.09	.16	23
macro avg	.76	.76	.73	843
weighted avg	.92	.91	.90	843
accuracy		.91		843

Table 2: By-tag performance of the combined model of Stanza + spaCy's APPRART.

ten confused with finite verbs which share the same word form.

4.3.3. Impact of Type of Text

Table 3 shows the performance of the models across the different categories for *text unit*, *medium* and *segment*. We hypothesized that the taggers would struggle most with fragmented text segments such as tables or word lists where no complete sentence context is given. When comparing the results for the different *text units*, we see that numerically, all models indeed perform better on sentences/ dialogues and phrases than on isolated words. However, the worse performance for words vs. sentences is only statistically significant for spaCy ($\chi^2(1) = 7.50, p < .01$, one-sided) but not for Stanza ($\chi^2(1) = 0.91, p > .17$, one-sided) nor the combined model ($\chi^2(1) = 2.18, p > .06$, one-sided).

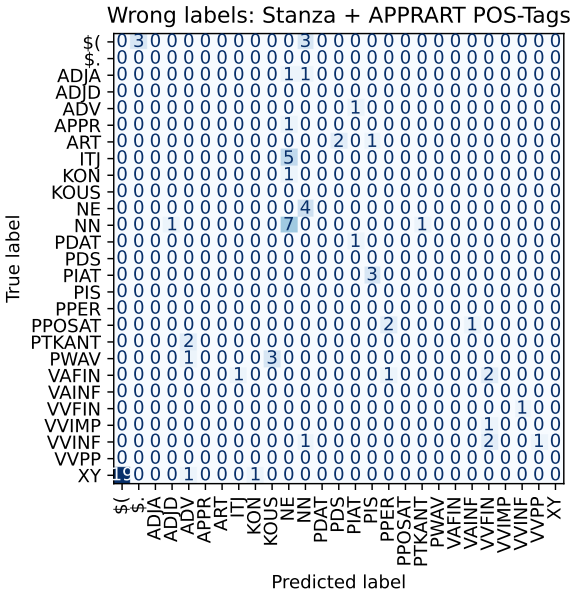

Figure 1: Confusion matrix of POS tags for the combined model of Stanza + spaCy's APPRART. Only wrong assignments are counted.

When comparing the different types of text segments, we see no large differences between them for either of the models, except for a worse performance on the category *further information*. A closer inspection revealed that this was largely caused by presenting isolated word forms in an inflectional paradigm, which are ambiguous without sentence context.

4.4. Evaluation of Lemmatization

For the evaluation of the lemmatization, we focus only on nouns, verbs and adjectives (39.6% of the evaluation set). Other word classes either do not inflect or, e.g. in case of pronouns or determiners, their lemma is usually not of primary interest but rather their grammatical properties. Table 4 shows the accuracy of the three models for nouns, verbs and adjectives.

text unit				
	spaCy	**Stanza**	**Mix**	**#toks**
dialogue	.93	.90	.93	73
sentences	.92	.90	.92	382
phrases	.90	.91	.91	284
words	.82	.86	.87	104
medium				
	spaCy	**Stanza**	**Mix**	**#toks**
audio	.87	.91	.91	23
table	.81	.86	.86	43
text	.91	.90	.91	777
segment				
	spaCy	**Stanza**	**Mix**	**#toks**
title	1.0	.90	.90	10
situation & instructions	.95	.95	1.0	22
task description	.94	.97	.97	35
media	.96	.87	.96	46
editing field	.89	.89	.90	375
hint	.68	.75	.76	76
model solution	.98	.99	.99	102
descr. of grammar	1.0	.95	.95	22
descr. of learning content	.95	.91	.93	94
list of new expressions	.92	.87	.87	61

Table 3: Accuracy of POS tagging per text unit, medium and segment type for spaCy (md), Stanza and the combined model (*Mix*).

POS	**spaCy sm**	**spaCy md**	**Stanza**	**#toks**
Noun	.92	.93	.98	203
Verb	.89	.90	.96	106
Adj.	.84	.92	.88	25
total	.90	.92	.96	334

Table 4: Lemmatization accuracy for nouns, verbs, adjectives and all three word classes.

We can see that the overall lemmatization accuracy is very high, especially for Stanza, where the accuracy of 96% corresponds to only 12 wrongly assigned lemmas. This includes cases where Stanza uses the old spelling (*Schloß* for *Schloss* 'lock') and cases where it indicates an ambiguous lemma (e.g. *Dosis/Dose* 'dose/can' for *Dosen*). spaCy, in contrast, often fails to provide a lemmatized form, sticking with the inflected word form (e.g. *warst* '(you) were' or *Würste* 'sausages') or it seems to apply rules that lead to non-existent word forms, such as **isen* instead of *essen* for the (irregularly inflected) verb form *isst* '(you) eat' or **Wetterlag* for *Wetterlage* 'weather conditions'.

5. Corpus Analysis

Following the results on the evaluation set, the whole corpus was tokenized and split into sentences with spaCy (medium) because it keeps contractions of prepositions and articles as one token. We lemmatized the corpus with Stanza and POS tagged it with the combined model based on Stanza with spaCy's APPRART tags. In total, the DUO_DE A1 corpus consists of 126,142 tokens across 4,084 segments.[6] This number includes punctuation marks and whitespace tokens such as tabs and newlines, which are preserved for some layout information. The following analyses are based on pure words, which we define as tokens that consist only of alphabetic characters and potentially hyphens. The corpus contains a total of 85,680 words. The ten most frequent lemmas are shown in the first column of Table 5.

Table 6 shows how many words belong to each kind of *text unit*, *medium* and *segment*. We can see that coherent text, i.e. sentences and dialogue only make up about 80% of the corpus. While phrases and isolated words showed high POS tagging accuracy as well, this part of the corpus may not be suitable for subsequent syntactic analyses like dependency parsing.

Using the *segment* annotation, we can approximate a distinction of text containing the primary learning content vs. instructional or metalinguistic content, which in the following we refer to as meta language. Meta language, when defined as comprising the segments *task description*, *task type*, *situation and instructions*, *description of learning content* and *description of grammar*, makes up about 35% of the words in the course. The language of the primary learning content (= all other segments) differs from the meta language as shown in Table 5. The table contains the ten most frequent noun, verb and adjective lemmas (according to the automatic annotation) for learning content vs. meta language. For this analysis, we leave out the *task type* because it consists of rather standardized task descriptions such as

[6] Six of these segments do not contain any tokens but were only included for structural reasons, see the *without text* annotation in Sec. 3.3.

overall		nouns				verbs				adjectives			
		content		meta		content		meta		content		meta	
der	7,157	Uhr	315	Übung	212	sein	1,702	lesen	379	gut	426	anderer	84
the		*clock*		*exercise/task*		*be*		*read*		*good*		*other*	
sie	4,539	Zeit	139	Tip	151	haben	844	sein	377	neu	98	gut	72
she/you		*time*		*hint*		*have*		*be*		*new*		*good*	
ich	2,258	Tag	134	Dialog	139	können	508	hören	258	alt	85	neu	45
ich		*day*		*dialogue*		*can*		*listen*		*old*		*new*	
sein	2,179	Jahr	120	Frage	93	gehen	404	passen	186	schön	66	richtig	35
be		*year*		*question*		*go*		*pass*		*pretty*		*right*	
und	2,031	Frau	114	Verb	75	machen	344	machen	182	schwarz	54	passend	20
and		*woman/Ms.*		*verb*		*make*		*make*		*black*		*fitting*	
in	1,668	Hose	101	Person	75	kommen	267	haben	157	anderer	45	verschieden	16
in		*pants*		*person*		*come*		*have*		*other*		*different*	
ein	1,665	Haus	100	Studierende	72	müssen	209	sehen	153	super	43	wichtig	14
a		*house*		*student*		*must*		*see*		*super*		*important*	
haben	1,002	Freund	86	Bild	72	wollen	197	sagen	152	klein	42	falsch	14
have		*friend*		*picture*		*want*		*say*		*small*		*wrong*	
was	998	Zimmer	83	Satz	61	finden	191	lernen	134	kalt	41	international	12
what		*room*		*sentence*		*find*		*learn*		*cold*		*international*	
mit	816	Wochenende	79	Gespräch	50	mögen	183	können	127	warm	41	spät	11
with		*weekend*		*conversation*		*like*		*can*		*warm*		*late*	

Table 5: Most frequent lemmata for primary learning content (*content*) and instructional and metalinguistic content (*meta*).

text unit		
	#words	**%**
sentences	83,146	65.9
dialogue	20,569	16.3
phrases	17,820	14.1
words	4,450	3.5
other	157	0.1
medium		
	#words	**%**
text	71,104	83.0
table	10,137	11.8
audio	4,355	5.1
video	84	0.1
segment		
	#words	**%**
editing field	32,715	38.2
media	14,126	16.5
task description	13,087	15.3
task type	8,528	10.0
situation and instructions	6,605	7.7
hint	4,044	4.7
list of new expressions	2,265	2.6
model solution	1,944	2.3
description of learning content	1,768	2.1
title	361	0.4
description of grammar	179	0.2
additional information	58	0.1

Table 6: Distribution of categories for *text unit*, *medium* and *segment* based on words in the whole corpus.

Verbinden Sie die Elemente 'Connect the elements'. While the *task type* segments consist of 8,528 words, one can find only 131 different lemma types and these would skew the analysis.

6. JSON Format

The corpus is stored in Tabular JSON format (Roussel, 2024). Each segment is represented as one JSON object. Each file represents one chapter/scenario with an array of all JSON objects belonging to this chapter. On the top level, each JSON object consists of an `id`, a `metadata` object, as well as a `tokens` and `sentences` array. The `metadata` object includes the information described in Sec. 3.3, e.g. what course and scenario the text is from or what units of text it contains. Additionally, it indicates how the data was tokenized and annotated with POS and lemmas. The `tokens` array consists of one object per token, which contains the token's id, word form, lemma, and POS tag. The `sentences` array describes the span of each sentence, i.e. from which token to which token each sentence within the current segment spans. The JSON format is made publicly available with the exception of the word forms and lemmas. In this derived text format, copyright is respected as the original

texts cannot be reconstructed from our annotations. A complete example for one segment can be found in Appendix A.

7. Conclusion and Future Work

We presented the creation of DUO_DE A1, a corpus compiled from A1-level learning material from the Deutsch-Uni Online GFL online course. The corpus is enriched with fine-grained metadata characterizing different kinds of text segments and was tokenized, POS tagged and lemmatized with high accuracy. While the texts are copyrighted, we publish all structural features, metadata and POS tags in a structured JSON format in order to contribute to transparency in corpus linguistic research on GFL learning material. The corpus can be used, for example, to analyse the structural composition of the learning material or the progression of POS distributions. In future work, we want to enrich the corpus with further linguistic annotations like syntactic dependency parsing and morphological analysis. We are also planning to process material from other CEFR levels from the Deutsch-Uni Online course in a similar way.

8. Acknowledgments

We gratefully acknowledge the Society for Academic Study Preparation and Test Development (g.a.s.t. e.V.) for providing access to the learning materials from German University Online (DUO) and for supporting this research. Furthermore, we thank the anonymous reviewers for their helpful comments.

9. Bibliographical References

Maria Rosario Bautista Zambrana. 2018. Corpus analysis of phraseology in an A1 level textbook of German as a foreign language. *Quaderns de Filologia - Estudis Lingüístics*, 22:13–32.

Lars Behnke. 2023. Korpuslinguistische Betrachtungen zum grammatischen Wandel in DaF-Lehrwerken. Zwischen Authentizität und Lernbarkeit. *AUC PHILOLOGICA*, 2022(3):11–39.

Maria Cieschinger. 2016. *The contraction of preposition and definite article in German. Semantic and pragmatic constraints*. PhD Thesis, University of Osnabrück.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge University Press.

Thomas François, Patricia Kerres, Damien De Meyere, and Ferran Suñer Muñoz. 2021. DAFLex: A CEFR-graded lexical resource for German as a foreign language. Presentation at the first Workshop on Building CEFR-graded resources for second and foreign language learning (GR4L2), Louvain-la-Neuve, Belgium.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in Python.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, USA. Association for Computational Linguistics.

Elen Le Foll. 2023. *Textbook English: A corpus-based analysis of the language of EFL textbooks used in secondary schools in France, Germany and Spain*. PhD Thesis, University of Osnabrück. Publisher: Universität Osnabrück.

Fanny Meunier and Céline Gouverneur. 2009. New types of corpora for new educational challenges: Collecting, annotating and exploiting a corpus of textbook material. In Karin Aijmer, editor, *Studies in Corpus Linguistics*, volume 33, pages 179–201. John Benjamins Publishing Company, Amsterdam.

Katrin Ortmann, Adam Roussel, and Stefanie Dipper. 2019. Evaluating off-the-shelf NLP tools for German. In *Proceedings of the Conference on Natural Language Processing (KONVENS)*, pages 212–222, Erlangen, Germany.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake. Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Adam Roussel. 2024. Tabular JSON: A proposal for a pragmatic linguistic data format. In *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*, pages 166–172, Vienna, Austria. Association for Computational Linguistics.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1999. *Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset)*. Universität Stuttgart, Universität Tübingen.

Elena Volodina, Ildikó Pilán, Stian Rødven Eide, and Hannes Heidarsson. 2014. You get what you annotate: A pedagogically annotated corpus of coursebooks for Swedish as a second language. In *Proceedings of the third workshop on NLP for computer-assisted language learning*, pages 128–144, Uppsala, Sweden. LiU Electronic Press.

10. Language Resource References

Thomas François and Patricia Kerres and Damien De Meyere and Ferran Suñer Muñoz. 2021. *DAFlex: A CEFR-graded lexical resource for German as a foreign language*. Centre de Traitement automatique du Langage (CENTAL). PID https://cental.uclouvain.be/cefrlex/daflex/.

Laâguidi, Jammila and Ruban, Vitaliia and Laarmann-Quante, Ronja and Drackert, Anastasia. 2026. *DUO_DE A1: An Annotated Corpus of A1-Level Learning Material from the Deutsch-Uni Online Courses for German as a Foreign Language*. Zenodo.

A. Appendix

The following shows a complete example of an annotated segment in JSON format. The segment reads *Günstig oder teuer? Wie ist Ihre Meinung?* ('Cheap or expensive? What is your opinion?'). In the published corpus, the lemma and word form (marked in red) are removed for copyright reasons.


```
{
    "id": "seg324",
    "metadata": {
        "course": "A1.1",
        "scenario": "3",
        "phase": "6",
        "learning_activity": "1",
        "task": "4",
        "medium": "text",
        "segment": "task description",
        "text_unit": "sentences",
        "comments": null,
        "other": null,
        "annotations": {
            "pos": {
                "use": "pos_xpos"
            },
            "token": {
                "type": "property",
                "model": "de_core_news_md
                    ",
                "source": "spaCy"
            },
            "pos_xpos": {
                "type": "property",
                "source": "stanza"
            },
            "lemma": {
                "type": "property",
                "model": "de",
                "source": "stanza"
            }
        }
    },
    "tokens": [
        {
            "id": "t1",
            "form": "Günstig",
            "lemma": "günstig",
            "pos_xpos": "ADJD"
        },
        {
            "id": "t2",
            "form": "oder",
            "lemma": "oder",
            "pos_xpos": "KON"
        },
        {
            "id": "t3",
            "form": "teuer",
            "lemma": "teuer",
            "pos_xpos": "ADJD"
        },
        {
            "id": "t4",
            "form": "?",
            "lemma": "?",
            "pos_xpos": "$."
        },
        {
            "id": "t5",
            "form": "Wie",
            "lemma": "wie",
            "pos_xpos": "PWAV"
        },
        {
            "id": "t6",
            "form": "ist",
            "lemma": "sein",
            "pos_xpos": "VAFIN"
        },
        {
            "id": "t7",
            "form": "Ihre",
            "lemma": "ihr",
            "pos_xpos": "PPOSAT"
        },
        {
            "id": "t8",
            "form": "Meinung",
            "lemma": "Meinung",
            "pos_xpos": "NN"
        },
        {
            "id": "t9",
            "form": "?",
            "lemma": "?",
            "pos_xpos": "$."
        }
    ],
```

```
    "sentences": [
        {
            "id": 1,
            "begin": 1,
            "end": 4
        },
        {
            "id": 2,
            "begin": 5,
            "end": 9
        }
    ]
}
```

Why Reconstructing Scrambled Texts Fails: A Structural Analysis of Reconstruction Outputs

Kei Du, Christof Schöch

University of Trier
Universitätsring 15, 54296 Trier
{duk, schoech}@uni-trier.de

Abstract

This paper explores the limitations of reconstructing scrambled text within the context of Derived Text Formats (DTFs). While previous research has treated reconstruction as a technical challenge, this study shifts the focus to investigating the causes of reconstruction failure. Through a detailed analysis of outputs generated by language models on non-literary (IMDb reviews) and literary (Gutenberg texts) datasets, several systematic patterns were identified. First, reconstructed texts are generally shorter than the originals, indicating that the generated results are often incomplete. Second, models simplify expressions by omitting specific modifiers, thereby producing more general outputs. Third, high similarity at the string level does not guarantee semantic equivalence, revealing fidelity-related issues in text reconstruction. In literary texts, chunk-based segmentation poses additional challenges; this approach disrupts syntactic and contextual coherence, leading to sentences that are structurally correct but semantically distorted. These findings suggest that reconstruction difficulty is not merely a matter of model performance but also reflects the importance of higher-level textual organization. This study highlights the fundamental limitations of current language models and reframes reconstruction failure as an analytical perspective for understanding how meaning is constructed in text.

Keywords: derived text formats, scrambled text, reconstructibility

1. Introduction

In Digital Humanities, derived text formats (DTFs), also sometimes called extracted features, have been proposed for the storage, publication, and reuse of datasets built from in-copyright texts (Jett et al. 2020, Schöch et al. 2020). One approach is to scramble the order of the words in the text so that it becomes unreadable to humans but can still be used for text and data mining. An important precondition for applying this DTF, of course, is that the original text cannot be reconstructed. This question has attracted increasing attention. For example, Du et al. 2025 largely framed this issue as a problem of performance: to what extent can large language models (LLMs) successfully restore an original sequence of a text once its linear structure has been disrupted. While such an approach yields valuable empirical insights, it also risks narrowing the scope of inquiry by treating reconstruction primarily as a technical challenge.

This paper adopts a different perspective. Rather than asking whether disordered texts can be reconstructed, it focuses on why reconstruction often fails. It reveals the extent to which textual meaning depends on higher-level structural organization — such as logical sequencing and thematic coherence — which cannot be easily recovered once disrupted. By reinterpreting reconstruction failure as an analytical resource, this study seeks to reposition the problem within a broader theoretical framework. In doing so, it aims to demonstrate that reconstructability is not merely a measure of technical capability, but a reflection of the structural constraints that determine how texts generate and maintain meaning.

2. Previous work

Research on reconstructing scrambled texts draws on several interconnected strands of work in natural language processing and computational literary studies. In Du et al. 2025, reconstruction has been framed as a problem of recovering linguistic structure from transformed representations. Experiments have been conducted in which a language model (T5-base) was fine-tuned to take scrambled texts as input and generate outputs that resemble the original text. Two datasets including IMDb reviews (non-literary texts) and Gutenberg novels (literary texts) were used.

For the experiments on IMDb reviews, each review was treated as a single data point and was converted into DTF format by shuffling the word order within sentences while keeping sentence order unchanged. Three datasets (25k, 50k, and 75k reviews) were used for training the T5-model, with separate validation and test sets of 5,000 unseen reviews each. For the experiments using Gutenberg novels, texts from four genres were randomly selected: detective, historical, romance, and science fiction. Two datasets were built (12 and 60 novels) to study the effect of size of training data, with texts split into chunks where words in each chunk were shuffled but chunk order in each novel preserved. These chunks (50, 100, and 500 words) were used as data points and divided into training, validation, and test sets in an 80/10/10 ratio.

Proceedings of Leveraging Derived Text Formats to Unlock Copyrighted Collections for Open Science (DTF) @ LREC 2026, pages 63–66
12 May 2026. ©ELRA Language Resources Association (ELRA), 2026

The reconstruction quality was evaluated using string similarity metrics (word error rate, rouge scores and sacreBLEU) between the reconstructed and original texts. The results showed that reconstruction performance is generally poor: similarity scores remain low across most cases, with only a few outliers achieving higher similarity. Longer and more complex literary texts are especially difficult to reconstruct. While models trained with more data offer slight improvements, they do not significantly change the overall outcome. The models tend to recover only fragments of vocabulary rather than accurate sentence structure or full semantic content. Overall, the study concludes that reconstructing original texts from scrambled texts is still a challenging task.

3. Analysis of reconstructed texts

While the previous study evaluated reconstructed texts by comparing the string similarity between the original and reconstructed texts, the present study reports on a detailed analysis of the reconstructed (non-literary and literary) texts and provides an in-depth examination of the reasons for reconstruction failure. The primary objective is to identify the differences between the reconstructed text and the original text across various levels, and to explore whether the quality of the reconstruction can be improved.

3.1 Non-literary texts (IMDb-reviews)

Based on the evaluation of the reconstruction results in Du et al. 2025, the model trained using 75,000 reviews produced better results than other models. Therefore, we have carefully compared the differences between the reconstructed text using this model and the original text and can make the following three observations.

First, the text length of the original and reconstructed IMDb-reviews was compared. The difference in length between the original and the reconstructed texts was calculated as follows:

$$difference = \frac{len(orig) - len(recon)}{len(orig)}$$

In Figure 1, the Y-axis represents the percentage difference in length, and the X-axis shows three models trained on 25,000, 50,000, and 75,000 texts, respectively. As can be seen, most of the values are larger than 0, indicating that most of the original texts are longer than their reconstruction. It is particularly noteworthy that, regardless of which model was used for text reconstruction, more than 75% of the reconstructed texts are at least 20% shorter than the original text. In contrast, only a very small number of reconstructed texts are as long as or longer than their original counterparts. This reflects a systemic issue with the model when reconstructing text: it tends to generate outputs that are shorter than the original text. To address this issue, a possible solution is to require that the length of the reconstructed text by the model must match that of the original text.

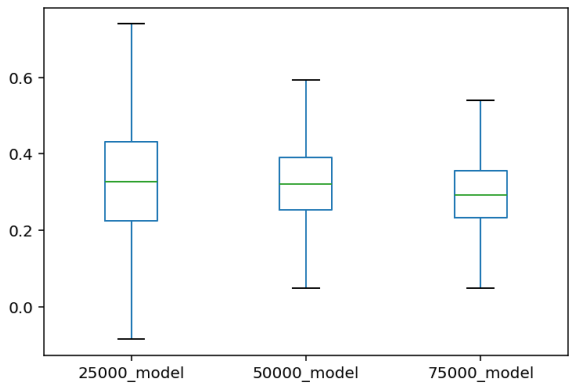

Figure 1: Distribution of the difference in length between the original and the reconstructed IMDb-texts (outliers are not visualized).

Second, the model tends to simplify sentences in reconstruction. Here are some examples:

- “A good solid piece” was reconstructed as “A solid piece”. (review no. 23)
- “Typical American movie” was reconstructed as “Typical movie”. (review no. 1)
- “This is one of the worst Stephen King movies I've ever seen” was reconstructed as “This is one of the worst movies I've ever seen”. (review no. 2073)

As we can see, the main subject (e.g. “piece”, “movie”) in these examples remains unchanged, and the overall sentiment or evaluation is preserved. However, specific modifiers such as adjectives, authorship, or origin are missing after the reconstruction. This is a common phenomenon in NLP text generation, and it is mainly caused by probability-based generation favoring high-frequency tokens, the loss of long-tail information, and decoding strategies that bias toward safer outputs. Prior works such as Guo et al. 2024 have shown that standard language model training tends to overemphasize high-frequency, low-information tokens, leading to fluent but generic outputs that lack linguistic diversity. This issue can be mitigated through controllable generation techniques such as incorporating weighting mechanism conditioned on token frequency (Jiang et al. 2019).

Third, string similarity does not fully indicate whether the reconstruction was successful, because similar texts may have completely different meanings. Here are some examples:

- “I hope everyone had a good time making this mess” was reconstructed as “I had a good time making this mess”. (review no. 4050)

- “After the first 5 minutes there is nothing worth watching in this film” was reconstructed as “After watching this film there is nothing worth watching”. (review no. 4690)
- “Predictable soaps, you've seen every soppy story full of sad” was reconstructed as “Predictable soaps, you've seen every soppy story full of sad”. (review no. 4884)

In fact, this issue is known as the “faithfulness” problem in text generation tasks and has been discussed in e.g. Maynez et al. 2020. In this work, textual entailment measures were suggested to address the need of developing evaluation and training methods that go beyond lexical overlap and can accurately model semantic faithfulness.

3.2 Literary texts (Gutenberg texts)

Compared to reconstructing non-literary texts, reconstructing literary texts presents both similar and additional challenges.

First of all, the comparison of the lengths of the reconstructed text and the original text reveals a different pattern to that observed before. Only when the Gutenberg texts are divided into longer 500-words-chunks, the original chunks are often 40% longer than the reconstructed text, or even more (see Figure 2). In contrast, the reconstructed text of 50-words-chunks and 100-words-chunks are similar in length to the originals or (in some cases) even longer than the originals.

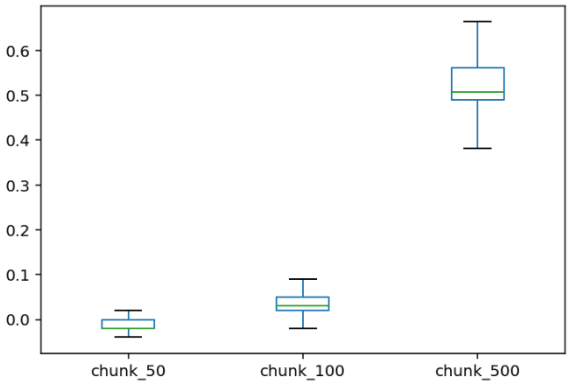

Figure 2: Distribution of the difference in length between the original and the reconstructed Gutenberg chunks (outliers are not visualized).

The evaluation results in Du et al. 2025 showed that reconstructing Gutenberg texts was more challenging than IMDb reviews. When we examined the reconstructed texts, we uncovered a key factor that led to this outcome: the different segmentation methods: The IMDb reviews were split into sentences and the word order within each sentence was shuffled. In comparison, the Gutenberg texts were split into chunks with exact number of words (50, 100 or 500). As a result, many chunks begin and end with incomplete sentences. Reconstructing such language chunks is more challenging than reconstructing a shuffled sentence, as it requires the model to maintain syntactic and semantic correctness while taking context into account and selectively discarding some information. We randomly selected a number of 50-words-chunks and carefully reviewed their reconstructed versions, and found that the beginnings and endings of nearly all the reconstructed chunks differed from those of the original text. There are three examples in Table 1. An interesting observation is that the first token in all three reconstructed examples is a punctuation mark, which is very common in reconstruction results. The reasons behind this phenomenon clearly require further analysis and research.

original text	reconstruction
room door . The diagram of this portion of the hotel will give you an idea of these connecting rooms . There are three of them , as you will see , all reception - rooms . Mr. Ransom had passed through them all in looking for his wife .	. There are three rooms in this hotel , all of them connecting . Mr. Ransom had an idea of his wife , as you will see . The reception - room will give you an idea of all these rooms , as you pass through the door .
to talk to Sir Andrew , if only for a moment . He felt lonely and desperately anxious . He had hoped to tire out his nerves as well as his body , but in this he had not succeeded . As soon as he had given up his tools	. As soon as he had succeeded in talking to Sir Andrew , he felt as if he had given up his nerves . He had not only given up his tools , but he had desperately hoped to tire out his body in this lonely moment , as well as
no other man could have come to him in that place ; and his whole body was wrung with torturing pains , and he was in the very article of death . And so it was , my prudence leading me to speak few and simple words , and my	; and he was so simple in his words that he could have no other article to speak of ; and in torturing him , and in leading me to my death , my whole body was wrung with pains and prudence . And so it was with my man ,

Table 1. Three examples of Gutenberg text chunks and their reconstruction

Another phenomenon worth noting is that, since each chunk may contain more than one sentence, and the words in each chunk were shuffled, the reconstructed text — while grammatically correct — often combines words of phrases from different sentences in the original chunk into a single sentence. This results in reconstructed text that may be very similar to the original at the word or n-gram level, but whose content differs significantly from the original. This phenomenon

is very common even in the reconstruction of short chunks containing just 50 words, let alone longer chunks containing 100 or 500 words. To avoid such errors, the model likely needs sufficient contextual information to accurately determine which subject corresponds to which verb and object within a sentence. However, since the order of the words within each chunk has been scrambled, the precise contextual information found in the original text is no longer present in the scrambled text.

4. Conclusion

In this study, we have examined the results of reconstructing scrambled text in detail. The results indicate that significant further improvements would be necessary for the successful reconstruction of DTFs, with areas of improvement ranging from the methods used to reconstruct the text to the evaluation of the quality of the reconstructed results. Given the current state of the art, reconstructing DTFs remains a highly challenging task. This also implies that reconstructability is not currently a factor that hinders the widespread use of DTFs for the publication of in-copyrighted text as research data.

5. Acknowledgments

This work was created in the context of the work of the association German National Research Data Infrastructure (NFDI) e.V. NFDI is financed by the Federal Republic of Germany and the 16 federal states, and the consortium Text+ is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project number 460033370. The authors would like to thank for the funding and support. Furthermore, thanks also include all institutions and actors who are committed to the association and its goals.

Author contributions:

Keli Du: Conceptualization, Methodology, Investigation, Visualization, Writing - original draft, Writing - review & editing

Christof Schöch: Funding acquisition and Supervision, Writing - review & editing

6. Bibliographical References

Du, K., Ackerschewski, S., Navruz, U., Sınır, N., Valline, J. & Schöch, C., (2025) “Reconstructing Shuffled Text. Bad Results for NLP, but Good News for Using In-Copyright Texts”, *Journal of Computational Literary Studies* 4(1). DOI: https://doi.org/10.48694/jcls.4163.

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2024. The Curious Decline of Linguistic Diversity: Training Language Models on Synthetic Text. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3589–3604, Mexico City, Mexico. Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/2024.findings-naacl.228.

Jett, Jacob, Boris Capitanu, Deren Kudeki, Timothy Cole, Yuerong Hu, Peter Organisciak, Ted Underwood, Eleanor Dickson Koehl, Ryan Dubnicek, and J. Stephen Downie (2020). *The HathiTrust Research Center Extracted Features Dataset (2.0)*. DOI: http://doi.org/10.13012/R2TE-C227.

Shaojie Jiang, Pengjie Ren, Christof Monz, and Maarten de Rijke. 2019. Improving Neural Response Diversity with Frequency-Aware Cross-Entropy Loss. In The World Wide Web Conference (WWW '19). Association for Computing Machinery, New York, NY, USA, 2879–2885. DOI: https://doi.org/10.1145/3308558.3313415.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/2020.acl-main.173.

Schöch, Christof, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann, and Jörg Röpke (2020). “Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen”. In: *Zeitschrift für digitale Geisteswissenschaften* 5. DOI: http://doi.org/10.17175/2020_006.

DIN 19461: A National Standard for Derived Text Formats

Thorsten Trippel[|||] [**], **Florian Barth**[*], **Jose Calvo Tello**[*],
Keli Du[§], **Philippe Genêt**[‡], **Daniel Kurzawe**[*],
Peter Leinen[+], **Piroska Lendvai**[¶], **Christof Schöch**[§]
Andreas Witt[**], **and Arden Zimmermann**[‡]

[|||]University of Tübingen
Keplerstraße 2, D-72074 Tübingen
thorsten.trippel@uni-tuebingen.de

[**]Leibniz Institute for the German Language
R 5, 6-13, D-68161 Mannheim
{trippel, witt}@ids-mannheim.de

[*]University of Göttingen
Papendiek 14, D-37073 Göttingen
florian.barth@uni-goettingen.de, { kurzawe,calvotello}@sub.uni-goettingen.de

[§]University of Trier
Universitätsring 15, 54296 Trier
{duk,schoech}@uni-trier.de

[‡]German National Library
Adickesallee 1, D-60322 Frankfurt am Main
{P.Genet, Ar.Zimmermann}@dnb.de

[+]Technical University of Darmstadt
Karolinenplatz 5, 64289 Darmstadt
peter.leinen@tu-darmstadt.de

[¶]Bavarian Academy of Sciences and Humanities
Alfons-Goppel-Str. 11, 80539 München
piroska.lendvai@badw.de

Abstract

We present DIN 19461:2026-06 (E), a German draft national standard that defines categories, terminology, and process requirements for Derived Text Formats (DTFs) created from text documents in natural language. The standard specifies enrichment and information reduction operations, requirements for combining multiple DTFs, and documentation obligations for publication, archiving, and reuse. Its aim is to enable legally compliant sharing and analysis of texts–especially where copyright or data protection prevents distributing originals–while maintaining scientific utility and reproducibility through explicit process and parameter recording. We outline the scope, the key concepts, the four core reduction operations (retain, delete, replace, randomise), together with examples across token-, structure-, and vector-based DTFs, and implications for infrastructures (e.g., ISO 24622-based metadata). Finally, we discuss limitations, open questions (e.g., reconstruction risks with modern ML models), and next steps for adoption and maintenance.

1. Introduction

Access to large text collections and their analysis via Text and Data Mining (TDM) is foundational for many research domains, yet legal constraints often restrict sharing copyrighted texts in their original form. The standard DIN 19461:2026-06 (E), which is a standard developed within the national organization of standardisation in Germany (DIN), addresses this by standardising how Derived Text Formats (DTFs) are defined, produced, and documented with the goal that they can be shared without enabling trivial reconstruction of the source text.

This paper contributs: (i) a concise overview of DIN 19461's scope and rationale; (ii) a presentation of its conceptual framework, requirements, and operations; (iii) worked examples spanning token-, structure-, and vector-based DTFs; and (iv) implications for infrastructures and reproducibility.

Proceedings of Leveraging Derived Text Formats to Unlock Copyrighted Collections for Open Science (DTF) @ LREC 2026, pages 67–75
12 May 2026. ©ELRA Language Resources Association (ELRA), 2026

2. Background and Motivation

In the context of the German National Research Data Infrastructure (NFDI), it is essential to provide researchers with reliable and legally compliant access to a wide range of research data. NFDI (Kraft et al., 2021) is a public initiative in Germany to build a research infrastructure for all research disciplines build on the FAIR principles (Wilkinson et al., 2016) with a clear focus on research data management established by the federal government and the German states (GWK – Gemeinsame Wissenschaftskonferenz, 2018). The consortium Text+ (see e.g. Hinrichs and Trippel, 2024) addresses language and text based research data. For language- and text-based data—central in linguistics, digital humanities, and natural language processing— the requirement of providing reliable and legal access to data is particularly challenging. Such data are frequently subject to legal and ethical constraints that restrict their distribution, while at the same time posing technical challenges related to processing, documentation, and reproducibility. This section outlines these two major dimensions: on the one hand, the legal and ethical considerations that shape the conditions under which text data may be shared, and on the other hand, the technical prerequisites needed to create formats that remain useful for research without compromising applicable restrictions, such as limited reusability due to licenses not granting this, or possible privacy related data.

2.1. Legal and ethical constraints

Language and text data are frequently subject to restrictions arising from both copyright and privacy considerations. In many jurisdictions, copyright law limits what can be done with contemporary texts and governs how authors' economic rights are protected. These rights ensure that creators can control the use and dissemination of their works and can benefit financially from them. As a result, research projects cannot always redistribute original texts, even when these texts are needed for scientific analysis. In addition to such copyright-related constraints, ethical obligations also arise from the nature of the data. Text collections may contain sensitive or personal information—for example, learner corpora that document individual writing performance. Even if such corpora do not always reach the threshold of originality required for copyright protection, they may still raise privacy concerns (see European Parliament and Council of the European Union (2026)) if the individuals who produced the text samples could be identified or exposed, especially in contexts where the content may be perceived as personal or potentially embarrassing. In Germany, the rights of authors to creative works are established by the ¨Urheberrechtsgesetz¨(UrhG).

DIN 19461 addresses these issues by positioning DTFs as a mechanism to reduce and abstract content in ways that preserve the ability to answer research questions while preventing reconstruction of the original text or making such reconstruction require disproportionate effort. The determination of whether a specific workflow meets legal obligations remains the responsibility of implementers.

2.2. Technical Challenges and Gaps

Technical challenges in working with language and text data arise at several levels. Prior approaches to "masking"(see for example Rehm et al., 2007 and Lehmberg et al., 2008) or abstracting text differ substantially across projects and infrastructures, and until recently no common terminology or requirements catalogue has existed. As a consequence, practices for reducing or altering text to meet legal and ethical constraints have been inconsistent and often difficult to compare. DIN 19461 responds to this inconsistent practices by introducing a unified vocabulary–for example the notions of DTFs (see also Schöch et al., 2020), information reduction, transformation, and generalisation–and by specifying that all operations and parameters must be documented at clearly defined levels of granularity, such as token, sentence, paragraph, document, or collection.

Beyond these conceptual gaps, practical technical issues must also be considered. On the one hand, many research workflows rely on large quantities of text data that are already available in digital form. These may include formats such as EPUB publications or HTML-based sources. Once such digital formats are available, the central question becomes how they may be used and what technical means exist to transform them automatically into representations that continue to support research needs while reducing legal and ethical risks.

Automated procedures play a key role in this transformation. When large-scale collections of digital text are involved, it is often impractical to rely on manual curation. Instead, workflows must support efficient, reproducible, and scalable operations that alter or abstract content without requiring excessive computational effort or complex manual intervention. Developing such workflows–capable of handling diverse formats, documenting each processing step, and ensuring that legal and ethical constraints remain respected–constitutes a significant technical challenge.

3. Standardisation Process

The development of DIN 19461 followed established procedures for national standardisation.This section provides an overview of the working group, its methodology, and the alignment of the resulting standard with related terminology and models from existing standards. For more information on the standardisation process see also (Preissner and Heid, 2025).

Standards developed within formal processes such as those of national standards, in Germany by the standardisation organisation DIN, or ISO, the International Organization of Standardisation, follow highly structured and rigorously regulated procedures. The documents themselves typically conform to established templates and organisational principles, which we do not discuss in detail here, as they are shared across many standards. Nevertheless, it is useful to provide a brief overview of the types of provisions contained in DIN 19461 in order to clarify how the standard supports the production, documentation, and publication of DTFs.

3.1. Working Group and Methodology

As in other formal standardisation processes, the creation of DIN 19461 began with the submission of a work-item proposal intended to assess whether a standard on DTFs was necessary and feasible. The responsible standards committees reviewed the proposal and concluded that, although only a limited number of data-holding institutions were at that time providing DTFs, it was nonetheless important to establish a structured and consistent basis for evaluating the legal situation surrounding derived formats. A standard would allow such evaluations to follow reproducible criteria rather than ad-hoc institutional decisions.

Based on this assessment, a working group was established within the German Standards Organization (DIN) working group NA 105-00-06 AA “Sprachressourcen und Sprachtechnologie”, drawing on expertise from the national research data infrastructure. The group developed a draft standard that defines terminology, describes the relevant operations and procedures, and specifies requirements for the creation and documentation of DTFs.

The question of why an international standard was not pursued from the outset was considered in the early stages. Although other jurisdictions, particularly within the broader European legal context, may eventually recognise the usefulness of such a standard, the immediate use case was rooted in national requirements and the needs within national research data providers. DIN 19461 therefore focuses first on addressing needs emerging from research data infrastructures in Germany, while leaving open the possibility that its concepts and procedures may inform international efforts in the future.

3.2. Alignment and References

DIN 19461 does not exist in isolation. It draws on several concepts and models already established in existing standards for linguistic resources. Within ISO/TC 37/SC 4, numerous standards are relevant to the creation, documentation, and referencing of language resources, and these form part of the conceptual background against which DIN 19461 was developed.

For example, persistent identifiers (PIDs) as defined in (ISO 24619:2011) are essential for reproducibility. When information-reduction operations are applied and a DTF no longer contains the original text in recognisable form, PIDs ensure that the underlying source can still be referenced reliably. Similarly, ISO 24622 (CMDI, see ISO 24622-1:2015 and ISO 24622-2, 2019) provides a component-based framework for metadata and serves as a foundation for modelling the metadata required for DTFs. Using CMDI components allows the documentation of enrichment steps, reduction operations, provenance, and processing parameters in a structured and interoperable manner.

Beyond these standards, several other models are relevant as conceptual or technical precursors. These include frameworks such as the Linguistic Annotation Framework (LAF, ISO 24612:2012) and feature-structure-based models (ISO 24610:2008) used for representing linguistic information, which can underpin enrichment steps and serve as starting points for generating DTFs based on stand-off annotations. Such standards and frameworks do not prescribe specific procedures for creating DTFs, but they provide established terminology, structural patterns, and annotation practices that can be employed within the workflows described in DIN 19461.

4. Overview of DIN 19461

For researchers, it is not only important to understand that DTFs provide a way to make text data usable under legal and ethical constraints, but also that a standardised approach exists for creating and documenting such formats. Data-holding institutions carry responsibility for complying with legal requirements while still enabling collaboration, transparency, and reliable long-term access. A common standard helps ensure that decisions about which data can be shared, and under what conditions, are based on consistent criteria rather than local interpretations.

Against this background, the following section

introduces the scope, conceptual framework, and requirements defined in the standard, and explains how these elements assist institutions in creating DTFs that are both legally robust and technically transparent.

4.1. Scope and Purpose

DIN 19461 applies to the classification and uniform description of methods and procedures used for creating DTFs from natural-language text documents. It covers both semi-structured data, such as XML or TEI representations, and unstructured sources, such as plain text, provided that these materials encode language at the character level. The focus of the standard lies on identifying how enrichment and information-reduction operations produce derived formats that remain analytically useful while preventing reconstruction of the original text in ways that could infringe legal or ethical constraints.

The scope of DIN 19461 explicitly excludes representations of text as images; only once such materials have been transformed into machine-readable digital text–for example through OCR–do they fall within the domain of the standard. Moreover, the standard does not make legal determinations about the status of any particular DTF. Instead, it provides the conceptual and procedural framework that allows practitioners to produce, document, and evaluate DTFs in a consistent, transparent, and assessable manner.

Within this scope, DIN 19461 defines terminology, units of granularity, categories of operations, and requirements that govern the creation of DTFs. It also establishes the documentation obligations necessary to ensure that such formats can be interpreted, compared, and reused across institutions and projects. The standard thereby provides a stable foundation for producing derived formats that simultaneously support research goals and respect the legal and ethical boundaries of the underlying source material.

4.2. Conceptual Framework

The conceptual framework of DIN 19461 introduces the fundamental notions required to describe, generate, and evaluate DTFs.

The standard distinguishes several key concepts. *Information reduction* refers to operations that remove, alter, or generalise textual content in a controlled manner. *Transformation* captures the conversion of text segments into new representational forms, such as linguistic categories or numerical vectors. *Generalisation* describes abstraction steps that replace specific linguistic content with higher-level descriptors. To ground these operations, DIN 19461 formalises the units on which they may apply, including text, tokens, and sequence information such as sentences, paragraphs, or larger structures.

The conceptual framework also incorporates terminology from linguistic and computational methods, including part-of-speech categories, lemmas, named entities, syntactic relations, and embeddings. Such concepts enable the enrichment of source material with linguistic annotations that may subsequently form the basis for reduction or transformation into a DTF.

Finally, the standard enumerates units of granularity and their associated types of annotation. These granularities–ranging from individual tokens to entire document collections–provide the structural reference points for both enrichment and reduction. By defining concepts and units systematically, DIN 19461 establishes a coherent vocabulary for describing how DTFs are produced and how their properties can be evaluated across projects and institutions.

4.3. Requirements Structure

DIN 19461 provides the requirement structure setting out the central requirements that govern the creation, documentation, and publication of DTFs. These requirements address the entire workflow from initial enrichment to the evaluation and dissemination of the resulting formats. First, the standard defines how enrichment procedures must be described, including the linguistic or structural annotations applied to the source material and the tools, models, and parameters used. Second, it specifies the requirements for information reduction, which is implemented through four well-defined operations: selective retention, deletion, replacement, and randomisation. Each operation must be applied at an explicitly stated level of granularity, and its effects must be documented in a way that enables transparent assessment of the resulting DTF.

Beyond individual operations, the standard also provides requirements for evaluating combinations of DTFs that originate from the same source material. Such combinations may increase the risk of reconstructing the original text, and DIN 19461 therefore mandates that their joint effects be considered when assessing whether a set of DTFs remains compliant with legal and ethical constraints. Finally, the standard outlines the prerequisites for publication, including mandatory metadata describing methods, tools, granularity levels, parameters, and any additional contextual information required to ensure reproducibility and to support the evaluation of reconstruction risks (see for example (Du et al., 2025)). Collectively, these requirements create a framework that enables consistent production and responsible sharing of DTFs across institutions and projects.

5. Core Operations for DTFs

In the following subsections, we describe the four core operations defined in the standard: enrichment, selective retention, deletion, and randomisation.

5.1. Enrichment

Enrichment is the first step in producing a DTF and serves as the foundation for all subsequent information-reduction operations. In DIN 19461, enrichment refers to the addition of linguistic, structural, or statistical information to the source text before any transformation, deletion, or abstraction takes place. At the same time, the standard recognises that enrichment may also consist of *not* adding any additional annotation. In such cases, the text is used exactly in the form in which it is available, without further linguistic or structural augmentation. This minimal form of enrichment remains a valid option whenever no additional annotation is required for the intended reduction steps.

If enrichment *is* performed, it must precede all information-reduction operations. This ensures that any subsequent transformations rely on consistent, traceable, and high-quality input data. Enrichment can include a wide range of annotations, depending on the research context and the level of granularity relevant to the intended DTF. Examples include the assignment of part-of-speech categories, lemmas, named entities, syntactic or dependency relations, and other forms of linguistic analysis. It may also involve information obtained through computational methods, such as vector-based representations or statistical measures extracted from the source material.

5.2. Selective Retention

Selective retention refers to the controlled preservation of certain pieces of information from the source text, provided that these elements are relevant for subsequent analytical tasks and can be retained at a given level of granularity without, on their own, revealing the original textual content. In DIN 19461, selective retention does *not* aim to produce a legally unproblematic DTF by itself. Instead, it constitutes an initial step that ensures that the information required for later processing remains available while preparing the ground for further information-reduction operations.

The key requirement is that selective retention should preserve only those elements that remain compatible with the intended reduction workflow. At the chosen granularity level—whether token, sentence, paragraph, or document—retained information must support the analytical purpose without obstructing the subsequent deletion, replacement, or randomisation that may be necessary to ensure that the final DTF cannot be trivially reconstructed. Selective retention therefore contributes to shaping the representational basis on which later reduction steps operate, but it does not by itself guarantee non-reconstructibility or legal compliance.

In practice, selective retention may involve keeping structural delimiters, metadata, positional information, segment boundaries, or statistical properties that support later analytical methods. Where linguistic annotations such as lemmas, part-of-speech categories, or named-entity labels are retained, this must be done with the understanding that further reduction operations may still be needed to prevent reverse mapping to the original lexical items. DIN 19461 therefore requires transparent documentation explaining why the retained information is relevant for the intended analysis, how it relates to the chosen granularity level, and how it fits into the broader sequence of reduction steps. Through this staged approach, selective retention helps ensure that the resulting DTF can be transformed into a legally robust format by applying the subsequent operations defined in the standard.

5.3. Deletion

Deletion constitutes one of the fundamental operations in the derivation of a DTF, focusing on the removal of elements from the source material that are not required for the planned analytical tasks. As defined in DIN 19461, deletion reduces the textual detail contained in the intermediate representation and thereby contributes to limiting the potential for reconstructing the original text. However, deletion alone does not produce a format that is legally or ethically unproblematic; rather, it functions as one coordinated step within a broader sequence of reduction operations.

In this operation, specific units of the text–such as individual tokens, multi-word expressions, sentences, or larger structural segments–are removed according to defined criteria. These criteria may be rule-based, algorithmic, or derived from annotations produced during the enrichment phase. The granularity selected for deletion must be compatible with later reduction operations, ensuring that the workflow as a whole leads toward a representation that can fulfil the legal and methodological aims of the DTF.

DIN 19461 requires that every deletion step be documented precisely. This includes a description of which elements are removed, the procedures or heuristics used to identify them, and the granularity level at which the deletion occurs. Such documentation clarifies how deletion supports the transformation of the data and how it interacts with the other operations–replacement and randomisation–that may still be necessary to ensure that the final DTF

cannot be interpreted or reverse-engineered as the original text.

5.4. Replacement

Replacement is an information-reduction operation in which selected elements of the source text are substituted with abstract or categorical representations. In DIN 19461, this operation serves to transform concrete linguistic material into forms that retain analytical value while reducing the possibility of reconstructing the original wording. As with other reduction operations, replacement does not by itself ensure that the resulting Derived Text Format (DTF) complies with legal or ethical requirements; rather, it forms part of a coordinated sequence of steps that collectively lead toward a non-reconstructible representation.

In this operation, textual units–such as tokens, multi-word expressions, or larger linguistic structures–are replaced according to explicitly defined rules. These replacements may take the form of linguistic categories (e.g., part-of-speech labels, lemma identifiers, named-entity types), structural abstractions, or numerical or vector-based representations. Replacement may operate at various levels of granularity, and the chosen level must be compatible with the intended analytical purpose as well as with subsequent reduction steps such as deletion or randomisation.

DIN 19461 requires that each replacement step be documented thoroughly. This documentation includes the description of the transformation rules applied, the models or algorithms used (for instance, tagging models or embedding frameworks), version information, and any parameter settings relevant to the operation. Such transparency ensures that the replacement is interpretable, reproducible, and assessable within the broader reduction workflow. Through this mechanism, replacement helps preserve analytical features of the text while progressively distancing the derived representation from the original content.

5.5. Randomisation

Randomisation is an information-reduction operation in which the order or internal structure of textual units is deliberately altered according to defined randomness parameters. In DIN 19461, randomisation contributes to distancing the resulting DTF from the original text by disrupting sequential patterns that could otherwise support reconstruction. As with the other operations, randomisation is not sufficient on its own to guarantee a legally or ethically unproblematic DTF; instead, it operates as one coordinated element within a broader reduction workflow.

Randomisation can be applied at various levels of granularity. At the token level, shuffling or reordering disrupts the syntax and surface structure of the text, effectively eliminating the sequence information needed to reconstruct meaningful sentences. At the sentence or paragraph level, randomising segment order breaks larger structural relationships, while still maintaining the internal consistency of each segment if required for analysis. At the document level, randomising whole-document order affects only the arrangement of documents within a collection and does not, by itself, prevent reconstruction of the content of individual documents. The effects of randomisation therefore depend strongly on the chosen granularity, and this choice must be aligned with the intended analytical purpose and with the overall reduction strategy.

DIN 19461 requires that all randomisation procedures be documented precisely. This includes specifying the units subject to randomisation, the algorithms or methods used, parameters such as random seeds or shuffle constraints, and any rules governing how randomisation interacts with preserved or enriched information. Such documentation is essential for reproducibility and for assessing how randomisation contributes to reducing reconstructive potential when combined with other reduction operations such as deletion or replacement. By systematically altering structural order at defined levels of granularity, randomisation supports the creation of DTFs that maintain analytical utility while further limiting the possibility of recovering the original text.

6. DTF Generation Workflow

The workflow is designed so that each step builds on the previous one, with enrichment establishing the informational basis and reduction operations progressively removing or abstracting content.

6.1. Enrichment Phase

Before any information-reduction steps are applied, enrichment may be used to add linguistic, structural, or statistical information to the source text. This can include annotations such as lemmas, part-of-speech categories, named-entity labels with authority links, syntactic or dependency structures, coreference chains, or disambiguation results. In addition to linguistic annotations, enrichment may also involve the extraction of statistical measures or vector-based features. These annotations may be represented in formats such as XML, TEI, or JSON, depending on the needs of the workflow and the characteristics of the data.

DIN 19461 does not prescribe specific tools, models, or annotation schemas, but it requires that all

enrichment steps be documented clearly and in detail. Such documentation must include the methods and tools used, version and parameter information, the level of application (e.g., token or sentence), and the file formats in which annotations are stored. By providing this information up front, the enrichment phase establishes a consistent and interpretable starting point for the subsequent reduction steps.

6.2. Reduction Phase and Granularity

Once enrichment is complete (or if no enrichment is required), one or more of the four core reduction operations—selective retention, deletion, replacement, and randomisation—are applied. These operations may be used individually or in combination, depending on the analytical purpose of the DTF and the legal or ethical constraints associated with the underlying material.

One aspect defined within DIN 19461 is that each reduction operation must be applied at a clearly specified level of granularity. Possible levels include individual tokens, sentences, paragraphs, documents, or entire collections. The standard provides illustrative examples showing how the choice of granularity affects both the usefulness of the resulting DTF and its resistance to reconstruction. For instance, segment-wise randomisation disrupts intra-segment structure while preserving overall grouping; selective replacement of tokens with part-of-speech categories abstracts away from lexical form while retaining syntactic patterns. These examples highlight that granularity is not an arbitrary choice but a decisive factor in the structure, utility, and safety of the resulting DTF.

6.3. Reproducibility and Metadata

To support scientific reuse, evaluation of non-reconstructibility, and long-term preservation, every DTF must be accompanied by comprehensive metadata describing the full generation workflow. This includes all processing steps, the tools and models used, segmentation decisions, random seeds (where applicable), parameters for enrichment and reduction operations, as well as software versions. Such documentation makes it possible to reproduce the DTF creation process and to assess whether the combination of reduction operations sufficiently prevents reconstruction of the source text.

The standard recommends using structured, interoperable metadata representations, such as CMDI components defined in ISO 24622, to ensure that enrichment and reduction steps are linked to the final DTF in a machine-readable manner. This linkage enables consistent archiving, facilitates dissemination across infrastructures, and allows data-holding institutions to provide transparent accounts of how a given DTF was produced.

7. Challenges in the Development of DIN 19461

The development of DIN 19461 involved navigating several challenges arising from the intersection of legal, linguistic, and technical requirements. One central difficulty was balancing these perspectives in a way that would allow the standard to be broadly applicable across institutions while remaining sufficiently precise to support reliable evaluation of derived text formats (DTFs). Legal considerations emphasised the need to avoid creating formats from which the original text could be reconstructed, whereas linguistic and technical considerations focused on preserving analytical utility and ensuring that workflows could be implemented with existing tools and methods.

Another challenge concerned the level of detail required to document enrichment and reduction operations. The standard needed to be specific enough to support reproducibility and assessment—particularly regarding parameters, granularity choices, and processing steps—while also remaining tool-agnostic so that institutions could use different software environments without deviating from the standard's requirements. Achieving this balance required careful formulation of definitions, documentation rules, and workflow descriptions.

A further challenge was clarifying how combinations of DTFs derived from the same source material should be evaluated. While individual DTFs may meet the requirements for non-reconstructibility, their combination can increase the potential for recovering information from the original text. This issue becomes particularly relevant in light of modern machine-learning capabilities, which may detect patterns or correlations across multiple representations that are not apparent in any single DTF. The standard therefore emphasises the need to assess reconstruction risks not only at the level of individual DTFs but also for sets of derived formats taken together.

8. Discussion

Benefits. DIN 19461 provides a shared terminology and a unified set of requirements for the creation and description of DTFs. By defining enrichment and information-reduction operations, specifying levels of granularity, and requiring explicit documentation of tools, parameters, and workflow decisions, the standard enables reproducible and transparent production of derived formats. This contributes not only to methodological clarity but

also to the lawful and privacy-respecting dissemination of text-based data, as institutions are supported in assessing what information may be preserved, transformed, or removed within a controlled process. The standard therefore serves as a foundational reference for data-holding institutions and research infrastructures seeking to balance analytical utility with legal and ethical responsibilities.

Limitations. While DIN 19461 establishes a structured framework for deriving text formats, it explicitly refrains from making case-by-case legal determinations. Implementers must independently assess the applicable copyright and data-protection requirements for their specific use cases. Moreover, the concept of non-reconstructibility is not an absolute property but depends on context, available auxiliary information, and the evolving capabilities of analytical tools.

As computational methods continue to advance, especially in machine learning, the risk that certain derived formats could be partially reconstructed may increase. For this reason, the standard recommends conservative combinations of reduction operations and thorough documentation to support informed assessment of residual reconstruction risks.

Open Questions. Several aspects lie beyond the current scope of the standard and remain open for future work. These include formal models for assessing reconstruction risks–particularly in the presence of modern language models—benchmark tasks or evaluation suites for validating non-reconstructibility, and the development of CMDI profiles specifically tailored to DTFs. In addition, mappings to repository schemas and guidance for integrating DTF workflows into existing archival infrastructures require further elaboration.

9. Conclusion

DIN 19461 systematises how DTFs are defined, produced, combined, and documented. By establishing a unified terminology, specifying enrichment and information-reduction operations, and outlining clear requirements for granularity, metadata, and workflow transparency, the standard provides a structured foundation for creating derived formats that can be shared lawfully and used reliably in research contexts where original texts cannot be distributed. It thereby supports data-holding institutions and research infrastructures in enabling analytical work while respecting legal and ethical constraints.

The standard also aims to foster community engagement. We encourage infrastructures, projects, and research communities to pilot DTF workflows, provide practical feedback on their applicability, and contribute to the continued development and refinement of the standard. Such collaboration will help ensure that future revisions reflect emerging needs, evolving technologies, and potentially open pathways toward broader—possibly international—alignment.

10. Acknowledgements

Though the authors are indebted to various co-authors working on this topic for years, work on this paper was carried out within the National Research Data Infrastructure (NFDI) association. The NFDI is funded jointly by the Federal Republic of Germany and the 16 federal states, and the Text+ consortium is supported by the German Research Foundation (DFG). The authors are affiliated with the Text+ consortium, grant number 460033370. The authors gratefully acknowledge this support, as well as the engagement of all institutions and individuals contributing to the NFDI and its goals. We acknowledge the work of the DIN committee NA 105-00-06 AA "Sprachressourcen und Sprachtechnologie" and contributing institutions, who were providing initial feedback and discussion, resulting finally in the draft standard, that is expected to be finalized as a national standard in the course of 2026.

The authors acknowledge the use of Large Language Models (LLMs) as writing aids in phrasing this paper, based on the authors' notes, ideas and concepts, including notes that were created during the development of the national standard. The authors retain full responsibility for the content.

11. Bibliographical References

DIN 19461:2026-06 (E). 2026. Sprachressourcen und Sprachtechnologie - Abgeleitete Textformate (ATF). National Standard, Deutsches Institut für Normung (DIN), Berlin.

Keli Du, Sarah Ackerschewski, Uygar Navruz, Nazan Sınır, Julian Valline, and Christof Schöch. 2025. Reconstructing shuffled text. bad results for nlp, but good news for using in-copyright texts. *Journal of Computational Literary Studies*, 4(1).

European Parliament and Council of the European Union. 2026. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

GWK – Gemeinsame Wissenschaftskonferenz. 2018. Bund-Länder-Vereinbarung zu Aufbau und Förderung einer Nationalen Forschungsdateninfrastruktur (NFDI) vom 26. November 2018. Accessed: 2026-03-24.

Erhard Hinrichs and Thorsten Trippel. 2024. Text+ – concept and benefits for empirical researchers. *Cybernetics and Information Technologies*, 24(4):143–163.

ISO 24610:2008. 2008. Iso 24610-1:2008 – language resource management – feature structures – part 1: Feature structure representation. Technical report, International Organisation for Standardization (ISO), Geneva, Switzerland.

ISO 24612:2012. 2012. Language resource management — linguistic annotation framework (LAF). International Standard, International Organisation for Standardization (ISO), Geneva.

ISO 24619:2011. 2011. Language resource management – Persistent identification and sustainable access (PISA). International Standard, International Organization for Standardization (ISO), Geneva.

ISO 24622-1:2015. 2015. Language resource management – Component Metadata Infrastructure (CMDI) – Part 1: The Component Metadata Model. International Standard, International Organization for Standardization (ISO), Geneva.

ISO 24622-2. 2019. Language resource management – Component Metadata Infrastructure (CMDI) – Part 2: The Component Metadata Specification Language. International Standard, International Organization for Standardization (ISO), Geneva.

Sophie Kraft, Angela Schmalen, Hendrik Seitz-Moskaliuk, York Sure-Vetter, Jennifer Knebes, Eva Lübke, and Elena Wössner. 2021. Nationale Forschungsdateninfrastruktur (NFDI) e. V.: Aufbau und Ziele. *Bausteine Forschungsdatenmanagement*, (2):1–9.

Timm Lehmberg, Georg Rehm, Andreas Witt, and Felix Zimmermann. 2008. Digital text collections, linguistic research data, and mashups: Notes on the legal situation. *Library Trends*, 57(1):52 – 71.

Annette Preissner and Ulrich Heid. 2025. *The life of an ISO standard*, pages 427–446. De Gruyter.

Georg Rehm, Andreas Witt, Heike Zinsmeister, and Johannes Dellert. 2007. Masking treebanks for the free distribution of linguistic resources and other applications. In *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories (TLT 2007)*, pages 127–138, Bergen, Norway.

Christof Schöch, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann, and Jörg Röpke. 2020. Abgeleitete textformate: Text und data mining mit urheberrechtlich geschützten textbeständen. *Zeitschrift für digitale Geisteswissenschaften (ZfdG)*, 5.

UrhG. 2021. Gesetz über Urheberrecht und verwandte Schutzrechte. https://www.gesetze-im-internet.de/urhg/. Accessed 24 March 2026.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018.

Author Index