The German Medical Text Corpus: Early 2026 Update

Justin Hofenbitzer[1], **Christina Lohr**[2], **Frank Meineke**[2], **Markus Loeffler**[2], **Martin Boeker**[1]
[1]TUM University Hospital, School of Medicine and Health, Chair of Medical Informatics, Institute for AI and Informatics in Medicine, Technical University of Munich, Germany,
[2]Institute for Medical Informatics, Statistics and Epidemiology, Leipzig University, Germany

justin.hofenbitzer@tum.de, martin.boeker@tum.de

Abstract

Clinical text resources are a central component for the study of medical language, as well as the training and evaluation of large language models, chatbots, and artificial intelligence systems supporting clinical routines. With the German Medical Text Corpus (GeMTeX), we are currently working on the largest shareable clinical document dataset in German. The multi-centric project ensures diversity across different university hospitals, clinical domains, and text sorts. After a thorough de-identification process, the clinical texts are semantically annotated using Snomed CT, a language-independent, standardized medical ontology. While the corpus is still under active development, it is accessible upon request under controlled access conditions. As of February 2026, GeMTeX comprises more than 15k documents and 20M tokens. We refer researchers interested in the resource to visit https://kiinformatik.mri.tum.de/en/gemtex or reach out to us via gemtex.mi@mh.tum.de.

1. Introduction

The German Medical Text Corpus (GeMTeX) project is a three-year (2023–2026) initiative to establish a multi-site corpus of written German clinical routine text enriched with an ontology-grounded semantic layer (Meineke et al., 2023; Faller et al., 2025). Embedded in the German *Medical Informatics Intitative* (MII) (Semler et al., 2018), the project is organized as a consortium of 18 partners, including six university hospitals contributing documents and annotations, i.e., *TUM Klinikum*, *Universitätsklinikum Leipzig*, *Universitätsklinikum Erlangen*, *Charité Berlin*, *Universitätsklinikum Carl Gustav Carus Dresden*, and *Universitätsklinikum Essen*. GeMTeX addresses two bottlenecks: The scarcity of large German clinical corpora and the limited availability of semantic layers supporting evaluation beyond surface-form matching (Névéol et al., 2018; Hahn, 2025). GeMTeX therefore aims to (i) create a standardized German clinical text corpus, (ii) enable cross-site analyses and reproducible benchmarking, and (iii) provide a Snomed CT[1]-grounded semantic annotation layer for downstream evaluation, and clinical usage scenarios.

2. Corpus Design

The GeMTeX corpus design workflow is displayed in Figure 1 and shows how raw clinical documents are taken from the local hospital information systems. Importantly, all contributing university hospitals are committed to processing only documents for which the patients actively signed the MII broad consent, i.e., they agreed to their data being used in research. Next, all documents undergo a manual

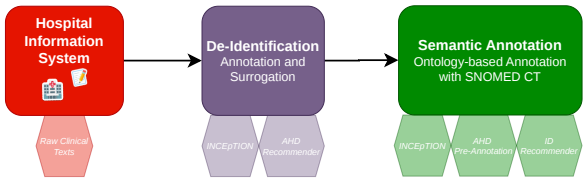

Figure 1: Schematic overview about the GeMTeX corpus design workflow. Clinical routine documents are taken from local hospital information systems, de-identified, and semantically annotated.

de-identification process, where relevant spans are annotated by two independent annotators using *INCEpTION* as annotation tool (Klie et al., 2018; Eckart De Castilho et al., 2024).[2] The annotation is supported by the industrial recommender system powered by *Averbis GmbH*[3]. A third person resolves disagreements before the annotated spans are replaced by pseudonyms[4] (Lohr et al., 2024, 2025).

Once de-identified, the documents undergo the final stage: The semantic annotation grounded in the widely shared and standardized ontology-based terminology Snomed CT in its April 2024 international version. To guide annotators through this complex task, Hofenbitzer et al. (2025) developed a comprehensive annotation guideline, which categorizes the 370k available Snomed CT concepts into three major groups and defined six annotation

[1]https://www.snomed.org/

[2]Access our de-identification guidelines and a gold standard dataset via https://doi.org/10.5281/zenodo.11502328.

[3]https://averbis.com/health-discovery/

[4]https://github.com/medizininformatik-initiative/GeMTeX/tree/main/surrogator

Proceedings of the 12th Workshop on Challenges in the Management of Large Corpora (CMLC-12) @ LREC 2026, pages 98–100. Palma, Mallorca. May 11, 2026.

maxims[5]

Each document annotation is performed by a single annotator with a medical background, supported by industrial pre-annotation from *Averbis GmbH* and a recommendation system powered by *ID Berlin*[6]. All annotators are employed as student assistants and receive fair compensation for their work. For quality assurance, 1% of the annotated documents are multiply annotated to compute site-level agreement and enable targeted adjudication. Periodic local and cross-site calibration sessions are instantiated to mitigate annotation drift.[7]

3. Current Status

The GeMTeX resource is in active implementation. As of February 2026, the project has delivered 15.4k de-identified documents (20.3M tokens, 682K de-identified spans), of which 382 documents (791.5K tokens) are semantically annotated, comprising 189.4K annotations. Besides the beforementioned annotation guidelines and gold standard examples, GeMTeX has designed a FHIR[8]-based interface for text material (Ammon et al., 2024).[9].

4. Accessibility

GeMTeX follows an access model under MII governance. Raw, de-identified, and semantically annotated texts remain at contributing sites and are not unconditionally redistributed. Access to the corpus or subsets is managed via the *German Portal for Medical Research Data* (FDPG)[10] under MII governance. Requests require a study protocol, ethics approval, and a data use application. The most recent information on accessibility options can be found at https://kiinformatik.mri.tum.de/en/gemtex, and interested researchers may reach out via gemtex.mi@mh.tum.de.

5. Outlook

The final project phase of GeMTeX focuses on completing annotations and consolidating guidelines, as well as releasing tooling. GeMTeX is intended to support benchmarking for named entity recognition, Snomed CT entity linking, as well as cross-site linguistic or medical analyses. By combining multi-site coverage, an ontology-grounded semantic layer, and controlled access, GeMTeX aims to provide a sustainable resource for clinical text-based research. In addition, structured clinical data, e.g., diagnoses, treatments, or laboratory values, are available for included patients and can be analyzed jointly with text data upon request.

6. Acknowledgments

We owe special thanks to all GeMTeX consortium members, associated partners, and contributing clinics at the university hospitals. This work is funded by the Federal German Ministry of Research, Technology, and Space under the grants 01ZZ2314A and 01ZZ2314B.

References

Danny Ammon, Maximilian Kurscheidt, Karoline Buckow, Toralf Kirsten, Matthias Löbe, Frank Meineke, Fabian Prasser, Julian Saß, Ulrich Sax, Sebastian Stäubert, Sylvia Thun, Reto Wettstein, Joshua P Wiedekopf, Judith A H Wodke, Martin Boeker, and Thomas Ganslandt. 2024. Arbeitsgruppe interoperabilität: Kerndatensatz und informationssysteme für integration und austausch von daten in der Medizininformatik-Initiative. 67(6):656–667.

Richard Eckart De Castilho, Jan-Christoph Klie, and Iryna Gurevych. 2024. Integrating INCEpTION into larger annotation processes. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 110–121, Miami, Florida, USA. Association for Computational Linguistics.

Jakob Faller, Christina Lohr, Martin Boeker, and Frank Meineke. 2025. Building the Infrastructure for the German Medical Text Corpus Project (GeMTeX). *Studies in Health Technology and Informatics*, 327:894–895.

Udo Hahn. 2025. Clinical document corpora—real ones, translated and synthetic substitutes, and assorted domain proxies: a survey of diversity in corpus design, with focus on German text data. *JAMIA Open*, 8(3):ooaf024.

Justin Hofenbitzer, Stefan Schulz, Martin Boeker, Peter Klügl, Sarah Riepenhausen, Christina Lohr, Jacqueline Lammert, Andrea Riedel, and Luise Modersohn. 2025. Introducing Medical Semantic Annotation Guidelines for German Clinical Documentation with SNOMED CT.

[5] Access the semantic annotation guidelines via https://doi.org/10.5281/zenodo.15689930.

[6] https://www.id-berlin.de

[7] Access a single, synthetic semantic gold standard document via https://doi.org/10.5281/zenodo.18861607.

[8] https://www.hl7.org/fhir/

[9] https://www.medizininformatik-initiative.de/Kerndatensatz/KDS_Dokument/MIIIGModulDokument.html

[10] https://forschen-fuer-gesundheit.de/en/

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. Association for Computational Linguistics. Veranstaltungstitel: The 27th International Conference on Computational Linguistics (COLING 2018).

Christina Lohr, Jakob Faller, Andrea Riedel, Hung Manh Nguyen, Markus Wolfien, Justin Hofenbitzer, Luise Modersohn, Jutta Romberg, Fabian Prasser, Jazia Omeirat, Yutong Wen, Oksana Galusch, Udo Hahn, Marvin Seiferling, Christoph Dieterich, Peter Klügl, Franz Matthies, Janina Kind, Martin Boeker, Markus Löffler, and Frank Meineke. 2025. GeMTeX's De-Identification in Action: Lessons Learned & Devil's Details. In *German Medical Data Sciences 2025: GMDS Illuminates Health*, pages 274–282. IOS Press.

Christina Lohr, Franz Matthies, Jakob Faller, Luise Modersohn, Andrea Riedel, Udo Hahn, Rebekka Kiser, Martin Boeker, and Frank Meineke. 2024. *De-Identifying GRASCCO - A Pilot Study for the De-Identification of the German Medical Text Project (GeMTeX) Corpus*, volume 317, pages 171–179. IOS Press.

Frank Meineke, Luise Modersohn, Markus Loeffler, and Martin Boeker. 2023. Announcement of the German Medical Text Corpus Project (GeMTeX).

Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana K. Savova, and Pierre Zweigenbaum. 2018. Clinical natural language processing in languages other than English: opportunities and challenges. *Journal of Biomedical Semantics*, 9:1–13, article no. 12.

Sebastian C Semler, Frank Wissing, and Ralf Heyder. 2018. German medical informatics initiative. *Methods of Information in Medicine*, 57(S 01):e50–e56.