MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI231035

Final Report on the German Clinical Reference Corpus 3000PA

Udo HAHN[a,1] Luise MODERSOHN[a,b] Jakob FALLER[a,c], and Christina LOHR[a,d]
[a] *Jena University Language & Information Engineering (JULIE) Lab, Friedrich-Schiller-Universität Jena, Jena, Germany*
[b] *Medizinische Informatik, TU München, München, Germany*
[c] *Universitätsklinikum Jena, Jena, Germany*
[d] *Institute for Medical Informatics, Statistics and Epidemiology (IMISE), Universität Leipzig, Leipzig, Germany*

Abstract. We here report on one of the outcomes of a large-scale German research program, the Medical Informatics Initiative (MII), aiming at the development of a solid data and software infrastructure for German-language clinical natural language processing. Within this framework, we have developed 3000PA, a national clinical reference corpus composed of patient records from three clinical university sites and annotated with a multitude of semantic annotation layers (including medical named entities, semantic and temporal relations between entities, as well as certainty and negation information related to entities and relations). This non-sharable corpus has been complemented by three sharable ones (JSYNCC, GGPONC, and GRASCCO). Overall, 3000PA, JSYNCC and GRASCCO feature about 2.1 million metadata points.

Keywords. Clinical text corpus, German language, annotation, clinical NLP

1. Introduction

Sharable natural language (NL) datasets (corpora) are a major prerequisite for the recent progress of natural language processing (NLP) research. Whereas this requirement can easily be fulfilled for general NL use scenarios by providing newswire or Wikipedia articles, tweets, etc., medical, even worse clinical, NLP applications suffer from a serious shortage of such corpora. This is mainly due to data privacy concerns about sensitive personal information typically contained in medical narratives.

The English-speaking medical language community, nevertheless, was truly successful in creating a variety of medical NL corpora. National legal restrictions were overcome by thorough de-identification and contractual Data Use Agreements (DUA) – the resulting corpora were (meta)data backbones for challenge competitions such as the i2b2/n2c2 task series.2 For no other NL such a wealth and diversity of datasets already exists. This holds true, in particular, for the German medical language community. Consequently, one of the goals of a large-scale national infrastructure initiative set up in Germany in 2017 (the Medical Informatics Initiative, MII)3 was to eliminate this data

[1] Corresponding Author: Udo Hahn, email: udo.hahn@uni-jena.de

[2] https://n2c2.dbmi.hms.harvard.edu/
[3] https://www.medizininformatik-initiative.de/en/start

bottleneck. In one of the four research clusters formed in MII, the SMITH Consortium,[4] the lack of resources for German clinical NLP was directly addressed by the Jena University Language & Information Engineering Lab (Professor Udo Hahn and JULIE Lab associates). That group initiated 3000PA, a German-language reference corpus compiled at three physically distributed clinical sites (the University Hospitals at Jena, Aachen and Leipzig) [1] whose fundamental quantitative characteristics are depicted in Table 1. Roughly 1,000 patient records were sampled composed of (much) more than 1,000 clinical reports about these individual patients at each site, with sentence and token numbers by far exceeding 100,000 and 1,000,000 items, respectively.

Table 1. Quantitative characteristics of the 3000PA corpus.

	Jena	**Aachen**	**Leipzig**
Patient Records	912	1 193	1 014
Clinical Reports	1 106	1 715	3 823
Sentences	187 982	147 104	715 879
Tokens	1 836 480	1 690 779	3 843 378

2. Methods

3000PA was set up in the summer of 2016 and is composed of Electronic Patient Records (EPR) from individual patients that were treated in internal medicine wards or intensive care units for at least five days between 2010 and 2015 and had died at the time the data was collected. Documents (mainly discharge summaries and transfer reports) were imported from the local clinical information system and different document formats (e.g., DOC, RTF) were transformed into UTF-8-compliant ASCII and cleansed subsequently.

The cleansed text data of 3000PA underwent a thorough annotation cycle. We, first, automatically annotated 3000PA with linguistic metadata for token and sentence boundaries using classifiers that were generated from FraMed [2], the first medical text corpus ever for the German language. In a second step, we manually annotated 3000PA over a period of five years for semantic metadata at the following five annotation layers:

- Macrostructure *segment* information of clinical documents in terms of section headings such as Family and Patient Anamnesis, Medication, Diagnosis, etc.[5] We here adhered to requirements derived from the *HL7 Clinical Document Architecture (CDA) Section Header Definitions* and adapted them to usage patterns in the three selected German hospitals [3].
- *Named entities* prevalent in clinical reports such as Medications,[6] Signs and Symptoms, Findings, Diagnoses,[7] and Protected Health Information (PHI).[8]
- *Semantic relations* between named entities, thus connecting two entity instances in terms of assertional statements (facts), e.g., <Medication> *administered-for* <Disease>, <Disease> *located-at* <Anatomical-Part>.
- *Temporal relations* between named entities or semantic relations, e.g., <Disease> *t-before* <Disease>, <Disease> *t-overlaps-with* <Medication>.

[4] https://www.smith.care/en/

[5] The annotation guideline for segments is available at https://doi.org/10.5281/zenodo.7707756

[6] The annotation guideline for medications is available at https://doi.org/10.5281/zenodo.7707947

[7] The annotation guideline for signs and symptoms, findings, and diagnoses is available at https://doi.org/10.5281/zenodo.7707917

[8] The annotation guideline for PHI items is available at https://doi.org/10.5281/zenodo.7707882

- *Certainty* information, including *negation*, attached to named entities or relations, such as "suspicion of" <Disease> or "no evidence for" <Disease>.[9]

Despite our efforts to get legally valid allowances and ethical votes for data sharing across local clinical walls for a de-identified (pseudonymized) version of 3000PA [4], data protection authorities denied, at that time (2018), any sort of corpus distribution.[10] Since a major goal of our project was to set up a medical/clinical text corpus that can be shared by the entire NLP research community without massive impediments, we created, in addition, three alternative corpora as proxies for 3000PA:

- JSYNCC [5] is a corpus made of fictitious case reports from medical textbooks. We took the e-book versions, scraped the reports and cleansed them. These reports are recreated, and thus *synthetic*, clinical reports that approximate real clinical language use. We cannot distribute that corpus directly due to Intellectual Property Rights (IPR) reasons, but rather offer software that automatically rebuilds a trusted copy of that corpus (plus associated metadata), if the local site at which rebuilding takes place possesses valid e-book licenses.
- GGPONC [6] constitutes a corpus of all (30) clinical practice guidelines for oncology hosted by *Deutsche Krebsgesellschaft* (DKG). Guidelines contain medical jargon but still differ from clinical language use genre-wise, and can thus only be considered as *linguistically similar* to clinical reports.
- GRASCCO [7] collects a (still small-sized) number of *synthetic* clinical reports that were derived from original ones, iteratively paraphrased by clinical experts and further augmented by clinical data noise, i.e., clinical information that was intentionally added as a camouflage for the original case.

Basic quantitative features of the three supplementary corpora are summarized in Table 2. Together with 3000PA, they form a collection of (pseudo-)clinical corpora on which the performance of clinical classifiers can subsequently be tested.

Table 2. Quantitative characteristics of three German medical corpora supplementary to 3000PA.

	JSYNCC	**GGPONC 2.0**	**GRASCCO**
Document Types	Case reports from medical textbooks	30 Clinical practice guidelines (oncology)	Discharge summaries
# Documents	399	10 193	63
# Sentences	20 860	78 090	5 430
# Tokens	199 569	1 877 100	43 667
# Annotations	343 191	448 328	177 773
Annotation Types	Named Entities: Findings, Diagnoses, Procedures, PHI	Named Entities: Findings, Substances, Procedures	Named Entities and Semantic Relations, Temporal Relations, Certainty, Negation +

3. Results

[9] The annotation guidelines for semantic and temporal relations, as well as certainty information attached to them will be made available upon publication of the classification results related to these topics at the same Zenodo site as those guidelines referred to in footnotes 5 to 8.

[10] Even training language *models* in-house and sharing them with external collaborators was interdicted. Meanwhile, this situation has changed in Germany. After many discussions among MII researchers and local, provincial and national data protection authorities, a *Broad Consent* solution is picking up more and more speed. Patients explicitly donate their EPRs after consultation and agree that their de-identified data can trustfully be shared for scientific purposes.

The results of the full annotation campaign involving 3000PA, JSynCC, and GraSCCo in terms of the number of single metadata instances per annotation layer are summarized in Table 3 (our focus is here on real and synthetic corpora with multi-layered annotations, which precludes the incorporation of metadata from GGPOnc). It depicts two important outcomes. First, cross-clinical annotations could only be achieved for named entities and, partially, for macrostructure segments (sections). This is due to different financial and manpower allocations at the three sites. The asymmetry becomes particularly obvious for Jena – this site spent more than 250,000€ extra money during the five year funding period. Second, different organizational policies for annotation team building had a big impact on productivity. Whereas Jena operated with 3 to 5 students of medicine (after their first exam) for each task, Leipzig and Aachen relied on single medical documentalists. Leipzig lately adopted the Jena policy of hiring students as annotators.

Table 3. Quantitative characteristics of the annotation layers of the corpora 3000PA, JSYNCC and GRASCCO.

	Jena	**Aachen**	**Leipzig**	**Σ**
Macrostructure Segments	228 539	39 435	–	267 974
Medical Named Entities	859 830	296 819	286 000	1 442 649
Medical Relations	134 751	–	–	134 751
Temporal Relations	106 661	–	–	106 661
Certainty + Negation	140 727	–	–	140 727
Σ	1 470 508	336 254	286 000	2 092 762

4. Discussion

Annotating clinical data is a cognitively demanding task that is controlled by iteratively refined annotation guidelines per task (usually, 3 to 4 iteration rounds were needed in our case for each task) and continuous training and supervision of the annotators. The quality of annotations is measured by well-known metrics for inter-annotator agreement (IAA), mostly Krippendorff's α or pair-wise averaged F1 score. Our IAA results for all five annotation tasks comply with those reported for English-language clinical corpora [1,3,6].

Based on this set of medical corpora we have trained a series of classifiers for each task. Again, the vast majority of them meet the quality standards for their English counterparts. For each annotation layer, classifiers have been trained. In total, we come up with 12 classifiers.

5. Conclusions

We here reported on the final shape of 3000PA, the first national reference corpus for the German clinical language. This corpus excels with its multi-site composition (raw textual data were provided by three major university hospitals in Germany (Jena, Aachen, and Leipzig)), its multi-layer annotations (seven layers ranging from formal linguistic structure to deep semantic information), and its large number of annotation instances.

The major drawback of this real clinical corpus is its distribution status – it is locked in the walls of the local hospitals as a consequence of German data protection legislation. However, we proactively tried to break this data bottleneck by supplying three alternative

corpora: two of them (JSynCC and GraSCCo) are synthetic ones, i.e., they contain pseudo-clinical documents recreated by expert authors (medical scientists) with the intention to mimic real clinical jargon. GraSCCo is publicly accessible on Zenodo,[11] whereas JSynCC needs licenses for the e-books it is based on. GGPOnc is (only) similar to real clinical corpora in terms of medical jargon and requires signing a DUA with DKG. An open issue here remains: How comparable are the three supplementary, synthetic or similar, corpora (JSynCC, GGPOnc, GraSCCo) with a real clinical one (3000PA)?

Acknowledgments

We want to thank our colleagues at the University Hospitals in Jena (UKJ: A. Scherag, D. Ammon, K. Saleh), Aachen (UKA: I. Lutz, S. Haferkamp), and Leipzig (UKL: T. Wendt), as well as our collaborators at the IMISE Institute, University of Leipzig (M. Löffler, F. Meineke, F. Matthies). Big thanks also owe to our numerous annotators, more than 30 students at UKJ and UKL, and our documentalists at UKA (R. Kober) and UKL (U. Schönwiese, I. Strauch). Last, but not least, we wish to thank our funding agency, Bundesministerium für Bildung und Forschung (BMBF; grant numbers 01ZZ1803G, 01ZZ1803A), for supporting this work.

References

[1] Hahn U, Matthies F, Lohr C, Löffler M. 3000PA: towards a national reference corpus of German clinical language. Stud Health Technol Inform. 2018 Apr;247:26-30, doi: 10.3233/978-1-61499-852-5-26.

[2] Wermter J, Hahn U. An annotated German-language medical text corpus as language resource. In: LREC 2004 – Proceedings of the 4th International Conference on Language Resources and Evaluation; 2004 May 24-30; Lisbon, Portugal. Paris: ELRA; 2004. p.473-6.

[3] Lohr C, Luther S, Matthies F, Modersohn L, Ammon D, Saleh K, Henkel A, Kiehntopf M, Hahn U. CDA-compliant section annotation of German-language discharge summaries: guideline development, annotation campaign, section classification. In: AMIA 2018 – Proceedings of the 2018 Annual Symposium of the AMIA; 2018 Dec 5; San Francisco, CA, USA. AMIA; 2018. p.770-9.

[4] Lohr C, Eder E, Hahn U. Pseudonymization of PHI items in German Clinical Reports. Stud Health Technol Inform. 2021 May;281:273-7, doi: 10.3233/SHTI210163.

[5] Lohr C, Buechel S, Hahn U. Sharing copies of synthetic clinical corpora without physical distribution: a case study to get around IPRs and privacy constraints featuring the German JSYNCC corpus. In: LREC 2018 – Proceedings of the 11th International Conference on Language Resources and Evaluation. 2018 May 7-12; Miyazaki, Japan. Paris: ELRA; 2018. p.1259--66.

[6] Borchert F, Lohr C, Modersohn L, Witt J, Langer T, Follmann M, Gietzelt M, Arnrich B, Hahn U, Schapranow MP. GGPONC 2.0—The german clinical guideline corpus for oncology: curation workflow, annotation policy, baseline NER taggers. In: LREC 2022 – Proceedings of the 13th International Conference on Language Resources and Evaluation. 2022 Jun 20-25; Marseille, France. Paris: ELRA; 2022. p.3650-60.

[7] Modersohn L, Schulz S, Lohr C, Hahn U. GRASCCO - The First Publicly Shareable, Multiply-Alienated German Clinical Text Corpus. Stud Health Technol Inform. 2022 Aug;296:66-72, doi: 10.3233/SHTI220805.

[11] DOI: https://doi.org/10.5281/zenodo.6539131