German Medical Data Sciences 2022 - Future Medicine
R. Röhrig et al. (Eds.)

doi:10.3233/SHTI220805

GRASCCO — The First Publicly Shareable, Multiply-Alienated German Clinical Text Corpus

Luise MODERSOHN [a,c,d,e*], Stefan SCHULZ [b,1*], Christina LOHR [a,d], and Udo HAHN [a,d]
[a] *JULIE Lab, Friedrich Schiller University Jena, Germany*
[b] *Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria*
[c] *Intelligence and Informatics in Medicine, Medical Center rechts der Isar, Technical University Munich, Germany*
[d] *SMITH Consortium of the German Medical Informatics Initiative*
[e] *DIFUTURE Consortium of the German Medical Informatics Initiative*

Abstract. We describe the creation of GRASCCO, a novel German-language corpus composed of some 60 clinical documents with more than.43,000 tokens. GRASCCO is a synthetic corpus resulting from a series of alienation steps to obfuscate privacy-sensitive information contained in real clinical documents, the true origin of all GRASCCO texts. Therefore, it is publicly shareable without any legal restrictions We also explore whether this corpus still represents common clinical language use by comparison with a real (non-shareable) clinical corpus we developed as a contribution to the Medical Informatics Initiative in Germany (MII) within the SMITH consortium. We find evidence that such a claim can indeed be made.

Keywords. Clinical NLP, German Clinical Document Corpus, Case Reports

1. Introduction

Clinical natural language processing (cNLP) systematically suffers from a tremendous shortage of textual (meta-)data that can be used for training and evaluating NLP systems. This lack of data is mainly due to ethically motivated privacy concerns implemented by data protection legislation. The regulations derived therefrom interdict data/document sharing across different clinical sites and, even more so, with non-clinical, e.g., NLP, research groups – even after careful de-identification of privacy-sensitive information contained in clinical documents. This situation is particularly frustrating since sharing data and using shared data in competitively organized shared tasks are considered the main drivers of progress in the field of (biomedical) NLP [1,2].

As far as the German cNLP community is concerned, several clinical corpora have been created already, yet they are only accessible by local data management personnel on-site (for a survey, cf. [3]). Quite recently, the BRONCO150 corpus [4] has been set

[1] Corresponding author, Stefan Schulz, Medical University of Graz, Auenbruggerplatz 2, 8036 Graz, Austria, Email: Stefan.schulz@medunigraz.at
[*] These authors contributed equally to this work

up, which contains de-identified real clinical documents and is accessible upon demand via a Data Use Agreement (DUA). Clearly a milestone for German-language cNLP, this corpus also has some drawbacks: it is small in size (150 documents, 85,000 tokens) and its sentences have been shuffled randomly (to further increase data protection), which completely destroys the typical clinical document structure, e.g., in terms of sectioning. This destructive intervention not only affects medical plausibility but also dissolves any sort of inter-sentential referential relations, which is likely to negatively affect named entity recognition and relation extraction for language models trained on BRONCO.

With the exception of BRONCO150, no other German-language corpus made of *real* clinical documents is currently available for sharing. As an alternative, several research groups are considering the use of *synthetic* data resources, which simulate real clinical documents either by in-depth textual modifications of original clinical documents or by re-writing them from scratch. In the modification scenario, real clinical documents are the starting point for several rounds of alienation by experienced clinical experts, which include all kinds of paraphrasing, chopping and adding medical statements, changes of medical attributes, values, and other textual parameters relevant for re-identification attempts. All these changes, however, have to mimic the specific style and wording of the chosen report genre. The JSYNCC corpus [5] is a typical example of such a synthetic approach. It has been extracted from a wide range of introductory textbooks (e-books) for medical students. Obviously, this corpus cannot be distributed physically due to Intellectual Property Rights (IPRs), but JULIE Lab distributes the code to reliably re-create JSYNCC copies at any other physical site (including selected meta-data). As a prerequisite, all e-books incorporated in JSYNCC need to be licensed by the local institution. As another alternative, corpora have been developed, which are supposed to be *similar* in style and wording to real clinical documents. For instance, GGPONC [3] is a corpus composed of all German clinical guidelines for oncology and might be used as a proxy for real clinical data, if the degree of similarity is considered sufficient.

However, both synthetic and similar documents have to be examined how comparable they are to real clinical documents. Hence, in this paper, we not only describe the construction of a new synthetic German-language clinical corpus in Section 2, but also provide metrical evidence in Section 3 for its comparability to real (non-distributable) clinical documents. The latter are provided by the 3000PA corpus [6], a collection of more than 1,000 clinical documents each from the University Hospital of Jena (3000PAJ), Leipzig (3000PAL) and Aachen (3000PAA), respectively. Table 1 briefly summarizes major characteristics and attributes of the corpora relevant for this work.

Table 1 Overview of the German clinical text corpora

Corpus	Text Genre	# Documents	# Sentences	# Tokens	Shareability
3000PAJ [6]	Discharge summaries	1,106	146,191	1,707,019	Non-Shareable
JSYNCC OP [5]	Medical text-books	399	20,860	199,569	Code for re-creation
GGPONC 1.0 [3]	Clinical practice guidelines	12,761	77,986	1.522,588	DUA
BRONCO150 [4]	Discharge summaries	150	10,251	83,633	DUA
This work	**Alienated case reports**	**63**	**5,430**	**43.667**	**Fully Shareable**

2. Methods

The starting point for building the first version of GRASCCO (Graz Synthetic Clinical Corpus) was a heterogeneous collection of documents, to which the second first author (a medical doctor) had access for specific use in particular projects:

- Anonymized and pseudonymized discharge summaries from the University Hospital Freiburg, Germany,
- Anonymized and pseudonymized discharge summaries from KAGes, a large Austrian network of public hospitals,
- German case reports from Open Access journals,
- Discharge summaries, some of them not de-identified, published on the Web.

These documents cannot be shared as-is according to privacy regulations. In order to make them fully shareable, any references to real patients and clinical actors had to be removed. This led to a fictional re-creation of these reports (by the second first author). The transformations involved the following steps:

- Real names of patients and therapists were replaced by fictional names, often with gender assignments differing from the original documents,
- Completely fictitious place and institution names were added,
- All dates were placed in the future,
- To additionally increase the noise level for re-identification attacks, at least one factual change was introduced in each medically relevant sentence, e.g., concerning laboratory tests, test result values, patients' complaints, diagnoses, medication statements,
- Many passages were paraphrased at all linguistic levels (mostly lexically and syntactically),
- Text fragments were exchanged with other ones (flowing back and forth within the entire collection), especially when atypical medical phenomena were described.

In a second round, additional text alienations were carried out and regionally typical expressions for salutations, technical terms, abbreviations, and academic degrees were changed so that no conclusions about the true origin of the texts can be made.

The strong alienation in form and content of the synthesized documents entailed that some of these narratives became medically implausible. This is not an issue for NLP purposes, since their focus is on learning language (use) models rather than domain models. Once a document was considered safe from re-identifying all of the mentioned human individuals, according to the second first author's judgement, it was incorporated into the corpus. The entire collection was then published by the second first author as products of fiction "inspired by real clinic texts", and made publicly available under the Creative Commons license BY 4.0 (all rights attribution) at Zenodo.[2]

[2] https://zenodo.org/

3. Corpus Description and Comparison

We will now, first, give a detailed description of the synthetic GRASCCO corpus (version 1), and then render preliminary evidence for its validity as a substitute for real clinical data by measuring the linguistic closeness between both document sets.

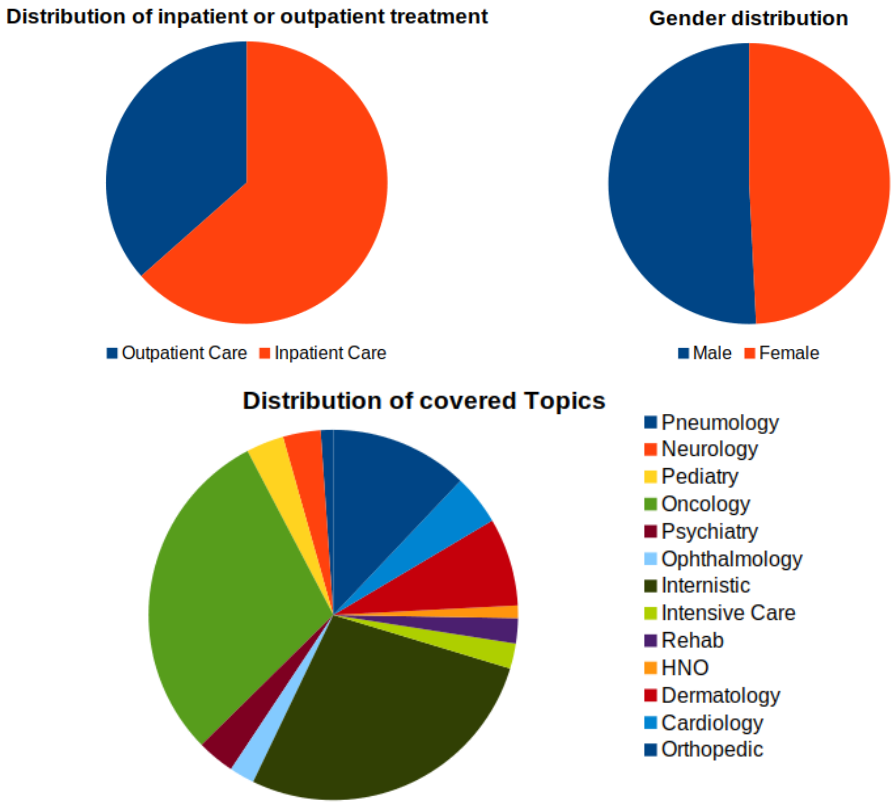

Figure 1. Document types, gender and topic distribution

GRASCCO v.1 consists of 63 documents with about 5,000 sentences and 43,000 tokens. An average document comprises 93 sentences, with about 740 tokens. A more detailed quantitative comparison with alternative German medical datasets is depicted in Table 1. Our corpus is almost perfectly gender-balanced, covers two-thirds of all patients as hospitalized in-patients, and also incorporates a large variety of medical topics, such as ophthalmology, oncology, or orthopedics as visualized in Figure 1.

In order to judge whether the documents from GRASCCO v.1 are truly linguistically close to real clinical documents, we took syntactic and semantic criteria into account and compared synthetic and real clinical corpora with non-clinical ones on a larger scale (for similar diagnostics of clinical reports, cf. also [7,8]).

Figure 2. Clustering results of k-means (left) and t-SNE (right)

For syntactic measurements, we used SPACY[3] as a pipeline with the general German language model[4] to automatically count the number of sentences, tokens, stop words, nouns, verbs, etc. of each document (for a broader set of linguistic features, cf. [9]). As far as semantic criteria are concerned, we used a SPACY-based medical named entity extraction pipeline, with concepts from the Unified Medical Language System (UMLS) [10], such as `Anatomy`, `Disorders` or `Living Being`, and medication names from the ROTE LISTE. [5] The resulting normalized counts of sentences, tokens, and occurrences of named entities were used as features for the subsequent clustering step.

We used t-SNE (T-distributed Stochastic Neighbor Embedding) [11] and k-means to cluster the document features of our corpus and compared these aggregated data with those from other real clinical (the Jena part of 3000PA [6], BRONCO150 [4]), synthetic clinical (JSYNCC [5]), similar-to-clinical (GGPONC [3], German PUBMED case reports),[6] and non-clinical, i.e., Wikipedia- or newspaper-rooted German corpora (WIKIWARSDE [12] and KRAUTS [13]). This test was accomplished using the Python library *scikit-learn*[7] and its implementations of `t-SNE` and `k-means`.

We only used up to 1000 documents per dataset to ensure a more balanced and fair comparison. As can be seen from Figure 2, `k-means` clustering shows that all datasets are clearly distinguishable, albeit not perfectly separable. Thus, the selection of features seems to be appropriate to describe our datasets for comparison. Both the `k-means` and the `t-SNE` plot show a clear separation between the clinical 3000PA documents and the other ones. Interestingly, documents from GRASCCO v.1 can be found both in the

[3] `https://spacy.io/`
[4] *de_core_news_sm*: `http://spacy.io/models/de`
[5] `https://www.rote-liste.de/produkte`
[6] Query: `Case Reports[Publication Type] AND GER[LA]`
[7] `https://scikit-learn.org/stable/index.html`

3000PA and BRONCO150 cluster and near the JSYNCC cluster. This yields preliminary evidence that our synthetic corpus is, at least partially, comparable with real German discharge summaries.

4. Discussion

We introduced a new clinical German-language corpus, GRASCCO, which can be publicly distributed within the (c)NLP community without any legal or contractual constraints. We also assessed the linguistic closeness of this synthetic corpus to other clinical and non-clinical corpora. The results give first hints that GRASCCO might be a reasonable substitute for non-shareable real clinical corpora. With the unconstrained shareability and open usability of GRASCCO, we intend to contribute to better comparability of cNLP systems for the German language when GRASCCO is used as an experimental frame of reference.

GRASCCO is composed of synthetic documents, i.e., original, yet subsequently anonymized and content-wise altered, real German-language clinical documents. They have undergone several rounds of alienation so that the re-identification of individual patients can be ruled out to the best of our beliefs. A preliminary comparison with real German-language clinical documents reveals that they approximate these gold data at both the syntactic and semantic level of comparison. Thus, we may recommend GRASCCO as a reasonable, hopefully valid, substitute for real clinical documents, since the latter are out of reach for free distribution even after lots of additional curation efforts (e.g., trusted and certified de-identification). However, our comparison with real clinical data is currently limited in many ways. Both the syntactic and semantic features chosen for comparison are quite simplistic, the syntactic ones, in particular. We might easily complement simple sentence and token counts by more expressive features involving n-gram statistics or the syntactic complexity mirrored in parse trees (see [9] for additional, more sophisticated features). In a similar way, the comparison of overlapping medical terminology could also be complemented by incorporating lexical semantic relations, such as synonymy or taxonomies. However, these investigations are not the focus of this work but are under way using a stylistic workbench to be published elsewhere. Furthermore, providing semantic metadata (e.g., annotations for named entities and semantic relations holding between them) could be the starting point for establishing a commonly shared German-language clinical gold standard for training and evaluation in cNLP.

Follow-up versions of GRASCCO will contain more documents and further alienation steps to even more heavily perturb potential adverse attacks on these data. The corpus can be found under the following link:

DOI: https://doi.org/10.5281/zenodo.6539131

Declarations

Ethical vote: The usage of 3000PA (Jena) is based on the approval by the local ethics committee (4639-12/15) and the data protection officer of the Jena University Hospital.
Study-Registry: not applicable
Conflict of Interest: The authors declare that there is no conflict of interest.
Author contributions: StS: corpus design and data creation; LM: study design, LM, CL: data analysis and interpretation; LM, CL, StS, UH: paper writing, UH: manuscript coordination. All authors approved the manuscript in the submitted version and take responsibility for the scientific integrity of the work.
Acknowledgement: This work was supported by BMBF within the SMITH and DIFUTURE projects under grant 01ZZ1803G and 01ZZ2009, respectively. We thank André Scherag, Danny Ammon, and all members of the Data Integration Center of the Jena University Hospital for their continuous support.

References

[1] Huang CC, Lu Z. Community challenges in biomedical text mining over 10 years: success, failure and the future. Briefings in Bioinformatics. 2015;17(1):132-44. Available from: https://doi.org/10.1093/bib/bbv024.
[2] Nissim M, Abzianidze L, Evang K, van der Goot R, Haagsma H, Plank B, et al. Sharing is caring: the future of shared tasks. Computational Linguistics. 2017;43(4):897-904.
[3] Borchert F, Lohr C, Modersohn L, Langer T, Follmann M, Sachs JP, et al. GGPONC: a corpus of German medical text with rich metadata based on clinical practice guidelines. In: LOUHI 2020 – Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis; 2020, pp. 38-48.
[4] Kittner M, Lamping M, Rieke DT, Götze J, Bajwa B, Jelas I, et al. Annotation and initial evaluation of a large annotated German oncological corpus. JAMIA Open. 2021;4(2). Ooab025. Available from: https://doi.org/10.1093/jamiaopen/ooab025.
[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, pp. 1259-1266.
[6] Hahn U, Matthies F, Lohr C, Löffler M. 3000PA: towards a national reference corpus of German clinical language. In: Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth. vol. 247 of Studies in Health Technology and Informatics. IOS Press; 2018, pp. 26-30. Available from: https://doi.org/10.3233/978-1-61499-852-5-26.
[7] Campbell DA, Johnson SB. Comparing syntactic complexity in medical and non-medical corpora. In: AMIA 2001 – Proceedings of the 2001 Annual Symposium of the American Medical Informatics Association. A Medical Informatics Odyssey: Visions of the Future and Lessons from the Past; 2001, pp. 90-94.
[8] Lysanets Y, Morokhovets H, Bieliaieva, O. Stylistic features of case reports as a genre of medical discourse. Journal of Medical Case Reports. 2017; 11, #83.
[9] Neal T, Sundararajan K, Fatima A, Yan Y, Xiang Y, Woodard D. Surveying stylometry techniques and applications. ACM Computing Surveys. 2017;50(6):#86.
[10] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research. 2004;32(suppl 1):D267-70. Available from: https://doi.org/10.1093/nar/gkh061.
[11] van der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research. 2008;9(86):2579-605.
[12] Strötgen J, Gertz M. WikiWarsDE: a German corpus of narratives annotated with temporal expressions. In: Multilingual Resources and Multilingual Applications. GSCL 2011 – Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology; 2011, pp. 129–134.
[13] Strötgen J, Minard AL, Lange L, Speranza M, Magnini B. KRAUTS: a German temporally annotated news corpus. In: LREC 2018 – Proceedings of the 11th International Conference on Language Resources and Evaluation. 2018, pp. 536-540.