Sharing Copies of Synthetic Clinical Corpora without Physical Distribution — A Case Study to Get Around IPRs and Privacy Constraints Featuring the German JSynCC Corpus

Christina Lohr **Sven Buechel** **Udo Hahn**

Jena University Language and Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena, Fürstengraben 27, D-07743 Jena, Germany
{christina.lohr|sven.buechel|udo.hahn}@uni-jena.de
http://www.julielab.de

Abstract

The legal culture in the European Union imposes almost unsurmountable hurdles to exploit copyright protected language data (in terms of intellectual property rights (IPRs) of media contents) and privacy protected medical health data (in terms of the notion of informational self-determination) as language resources for the NLP community. These juridical constraints have seriously hampered progress in resource-greedy NLP research, in particular for non-English languages in the clinical domain. In order to get around these restrictions, we introduce a novel approach for the creation and re-use of clinical corpora which is based on a two-step workflow. First, we substitute authentic clinical documents by synthetic ones, i.e., made-up reports and case studies written by medical professionals for educational purposes and published in medical e-textbooks. We thus eliminate patients' privacy concerns since no real, concrete individuals are addressed in such narratives. In a second step, we replace physical corpus distribution by sharing software for trustful re-construction of corpus copies. This is achieved by an end-to-end tool suite which extracts well-specified text fragments from e-books and assembles, on demand, identical copies of the same text corpus we defined at our lab at any other site where this software is executed. Thus, we avoid IPR violations since no physical corpus (raw text data) is distributed. As an illustrative case study which is easily portable to other languages we present JSynCC, the largest and, even more importantly, first publicly available, corpus of German clinical language.

Keywords: clinical NLP, German language corpus, legal constraints on corpus construction and distribution

1. Introduction

In both its academic and industry branches, the NLP community has established professional standards in which the open accessibility and exchange of language resources (corpora and other data sets, such as lexical resources, annotation guidelines, and software) play a dominant and fertile role. This liberal policy is one of the most important factors for the remarkable progress the field has made in the past decades. The question, however, is how NLP is going to prosper under less friendly, or even hostile, accessibility conditions strictly prohibiting the free flow of language resources.

The problems we address here are often deeply rooted in national legal systems world-wide and reflect fundamental economic as well as social concerns (Mittelstadt and Floridi, 2016). As a consequence, they are rather persist even over long periods of time and NLP research has to find ways to accommodate to the overarching legal ecosystems. *Intellectual Property Rights* (IPRs) play a key role in this discussion and are of utmost relevance for any NLP research targeting unabridged media contents (in contrast to document surrogates, such as titles, index terms, text snippets or abstracts, etc.). The stakeholders actively promoting IPRs are major publishing houses and other companies distributing media content. Commercial interests are the driving forces behind IPRs like, for example, the claims of a creative inventor of some content and enterprises distributing that work via various media channels to generate revenues for both parties. IPRs and their relations to NLP are discussed in depth by Truyens and Van Eecke (2014).

Data privacy, another crucial topic with particular relevance for biomedical or social media-focused NLP, is an ethical category deeply rooted in the civil codes of Western societies. In essence, privacy regulations are in place to preserve each citizen's right to informational self-determination (Fischer-Hübner et al., 2011). This means that each individual owns his or her personal data, which, in turn imposes the task to protect this ownership on the legal system.

The realm of medical and clinical information constitutes a typical example where privacy regulations are legally enforced to protect sensitive data associated with former and current medical statuses of individual persons and their social environment. For a survey of ethical issues to be considered by NLP research in the medical domain, cf. Šuster et al. (2017).

However, different legal cultures have emerged to balance the highly-valued societal goal of privacy protection with the sometimes competing goal of generating knowledge from biomedical research. In the Anglo-American countries, medical information is open for usage by the scientific community once clinical data are safely de-identified (as acknowledged by ethical boards) and legally binding *Data Use Agreements* (DUAs) are established between the data provider (typically, a hospital) and a data consumer (e.g., a scientist).

Within the European Union and its member states, however, much more restrictive privacy constraints are in force – making it almost impossible to access and distribute even de-identified medical information. For instance, in the German-speaking countries, access to patient information by actors outside the hospital site they originate from and by non-medical staff (e.g., computational linguists) is virtually impossible (Pommerening et al., 2014). Thus, NLP researchers in the biomedical domain (clinical NLP, in particular) are facing a lack of accessible language resources, at least when dealing with non-English languages.

One might remark that largely available resources reflecting general language use, e.g. from the newspaper domain, could be used for clinical NLP as well—with some additional efforts for domain adaptation to solve the data scarceness problem (Wermter and Hahn, 2004; Ferraro et al., 2013; Wang et al., 2015; Zhang et al., 2015). However, clinical language poses several domain-specific and rather tough challenges for NLP tools. Not only is the vocabulary abundant and highly specialized, but the language further deviates from standard usage in terms of spelling, typography and syntax: including, for example, short sentence fragments with paragrammatical structure, lack of punctuation, great volume and high degree of ambiguity of abbreviations, non-standard alphanumerical expressions, table-structured passages and mixed language use, including Greek and Latin forms (cf., e.g., Savkov et al. (2016)). Thus, adapting existing NLP tools to the clinical domain is arguably much more difficult than for many other domains and text genres. Furthermore, there is ample evidence that simply reusing standard NLP software trained on general language data (e.g., newspapers) results in severe losses of performance for biomedical applications (Tomanek et al., 2007; Ferraro et al., 2013; Hellrich et al., 2015).

The solution we are proposing here bypasses the two major obstacles for clinical NLP—IPRs and data privacy—in the following way. First, as discussed above, without any major legal changes, real, authentic clinical texts are, for the time being, inaccessible for most European languages. We therefore propose to substitute authentic clinical data by *synthetic* documents written by medical practitioners for education purposes. As a result, we are dealing with artificial, yet plausible and realistic medical scenarios, rather than identifiable social individuals and their personal legacy data and are thus able to circumvent any privacy concerns.

Nonetheless, as these synthetic documents are typically contained in the electronic versions of medical textbooks ("e-books"), the second access restriction to be overcome relates to IPRs. To do so, we employ a procedure similar to the way TWITTER corpora are often distributed (cf., e.g., Rosenthal et al. (2015)): Instead of releasing the (IPR-protected) raw data, we distribute NLP software which (given access to the original books in electronic format) reliably re-creates the same corpus we designed in our lab at any other site where this tool suite is executed. This results in a situation were a physically non-shared corpus can still be shared 'virtually' and be used for community-wide annotation as well as benchmarking efforts.

In the following, we illustrate our approach for the JSYNCC (JENA SYNTHETIC CLINICAL CORPUS), the first publicly available corpus of German clinical language ever.[1]

2. Related Work

There is a world-wide consensus on the fact that a patient's identity needs to be detached from medical data to protect sensitive personal data from any sort of misuse by non-medical or non-clinical actors. The standard way to achieve this requirement is by way of de-identification of so-called *Protected Health Information* (PHI). This is done with the help of a schema consisting of 18 categories, including data items that identify the patient in question (such as name, postal address, phone number, email address, or social security number), but also less apparent ones, such as names and locations of hospitals, their departments or clinical staff (for a complete list, cf. Stubbs and Uzuner (2015a)).

In the US, the past decade has seen a series of clinically oriented NLP shared tasks (for an extensive survey, cf. Huang and Lu (2016)). Prominent examples are the "Text Retrieval Conference" (TREC)[2] (Roberts et al., 2016) and the "Integrating Biology and the Bedside" (I2B2) initiative[3] (Chapman et al., 2011). From these activities, a wide range of de-identified and semantically annotated clinical corpora have emerged, covering the thematic foci of various competitions, such as de-identification (Stubbs and Uzuner, 2015a), medication extraction (Uzuner et al., 2010), temporal ordering of clinical events (Sun et al., 2013), or detecting risk factors for heart diseases (Stubbs and Uzuner, 2015b; Kumar et al., 2015). Each of the clinical corpora from I2B2 contains task-specific semantic metadata for slightly less than 1,000 English-language clinical reports. Those can be accessed, in a de-identified form, by simply signing a *Data Use Agreement* (DUA). Another major clinical database resource incorporating thousands of clinical reports, MIMIC III (Multiparameter Intelligent Monitoring in Intensive Care)[4] (Johnson et al., 2016), is also available via DUA contracting. Accordingly, for clinical NLP with focus on the English language, there are plenty of resources available (although not as abundant as in other non-medical fields such as newspapers or social media).

For the non-English language communities, however, less comfortable conditions apply. Only very few EU countries follow the DUA policy, such as reported for a clinical adverse drug reaction corpus for Spanish (Oronoz et al., 2015) or a comprehensive Dutch clinical corpus (Afzal et al., 2014). Some labs working on non-English languages have announced plans for releasing their resources, e.g., for French (Deléger et al., 2014), Polish (Marciniak and Mykowiecka, 2011) or Swedish (Dalianis et al., 2009). Apparently, these plans have not yet been fully realized as, to the best of our knowledge, none of these corpora is currently DUA-available for the research community.

[1] The proposed method of corpus construction and distribution aims at solving a general problem independent of specific national legislation cultures. However, for the specific case of the German legal system, very recently an interesting amendment to the national copyright law ("Urheberrechtsgesetz") has been installed by German authorities. Under certain conditions, this amendment allows for the sharing of corpora among *scientific partners* despite copyright protection, potentially mitigating some of the problems addressed in this contribution (at least for researchers located in Germany). For further information see UrhWissG (2018).

[2] http://www.trec-cds.org/

[3] https://www.i2b2.org/NLP/DataSets/

[4] https://physionet.org/mimic2/

Corpus	Documents	Sentences	Types	Tokens	Available
Wermter and Hahn (2004) (FRAMED)	–	6,494	20,729	100,150	✗
Fette et al. (2012)	544	–	–	–	✗
Bretschneider et al. (2013b) Bretschneider et al. (2013a)	174	4,295	3,979	28,009	✗
Toepfer et al. (2015)	140	–	–	–	✗
Lohr and Herms (2016)	450	22,427	11,008	266,390	✗
Kreuzthaler and Schulz (2015) Kreuzthaler et al. (2016)	1,696	–	–	–	✗
Roller et al. (2016)	1,725	**27,939**	–	158,171	✗
Cotik et al. (2016)	183	2,234	–	12,895	✗
Krebs et al. (2017)	**3,000**	–	–	–	✗
Hahn et al. (2018) (3000PA)	**3,000**	–	–	–	✗
JSYNCC (this work)	867	24,895	**32,108**	**312,784**	✓

Table 1: Overview of existing corpora of German clinical language. Highest value per column in bold.

Another source of medical language resources in Europe derives from the “CLEF eHealth” initiative.[5] Established in 2013, this series of health-related challenges led to the preparation of several corpora—mostly for the English language, but also for other European languages. However, these corpora are typically very small and available only for usage directly related to the respective task, i.e., they can neither be used later on nor are they available for the research community independent of the specific CLEF task. For German-language medical corpora the situation is even worse—*all* clinical corpora are *only* accessible to the research staff within the lifetime of a project and remain inaccessible forever for the outside world. Schulz and López-García (2015) give an overview of technical, legal and organizational issues for clinical NLP in Germany (among other countries) and conclude that electronically archived patient records are typically not intended for further scientific use, e.g., text mining. Furthermore, there are no unanimously shared standards or guidelines for the storage of clinical notes and reports—resulting in a myriad of physical encodings of electronic patient data, even within a single hospital.

Nonetheless, there have been a few disconnected activities in the German NLP community to create in-project clinical corpora. In Table 1 we list, to the best of our knowledge, all existing German-language clinical corpora that have been described in scientific publications. Wermter and Hahn (2004) created FRAMED, the first German-language medical corpus ever published. It consists of a mixture of approximately 300 clinical reports, textbooks and consumer texts annotated with low-level linguistic metadata (up to the level of parts of speech). FRAMED was further used to generate in-domain machine learning models for different tasks, e.g., sentence splitting and tokenization (Faessler et al., 2014; Hellrich et al., 2015; Hahn et al., 2016). FRAMED has also become part of the multilingual extension of NEGEX, a corpus annotated for negation expressions (Chapman et al., 2013).

Bretschneider et al. (2013b) and Bretschneider et al. (2013a) introduce a corpus composed of German radiology reports and expand clinical lexical resources with information about pathology classification. Fette et al. (2012) assembled a corpus composed of 544 clinical reports from various medical domains (echocardiography, EEG, lung function, X-ray thorax, bicycle stress test) to train a CRF classifier for an information extraction (IE) task. Toepfer et al. (2015) describe a corpus made of 140 transthoracic echocardiography reports for their IE experiments. Lohr and Herms (2016) collected 450 surgery reports and used these resources to build language models adapted to metadata from two German medical thesauri. A collection of almost 1,700 de-identified clinical in- and outpatient discharge summaries were assembled from a dermatology department for an unsupervised abbreviation detection procedure (Kreuzthaler and Schulz, 2015) and supervised machine learning using an SVM for abbreviation and sentence delineation (Kreuzthaler et al., 2016). Roller et al. (2016) introduce an annotation scheme for a German corpus in the nephrology domain and a similar scheme focusing on negation phenomena is presented by Cotik et al. (2016). The latter two publications use discharge summaries and clinical notes as their document base. In the most recent publications, Krebs et al. (2017) describe a corpus of 3,000 chest X-ray reports used for term extraction (in an effort to improve IE) and Hahn et al. (2018) present 3000PA, a collection of 3,000 German discharge summaries from three different German university hospitals, currently annotated with medication information. This corpus is intended to become a national reference corpus for German clinical language based on a DUA-style agreement policy to be implemented in the future. Once again, *none* of these corpora is currently available for public use.

3. Corpus Creation

To mitigate the accessibility problems encountered for clinical corpora for most European (non-English) languages we propose an alternative workflow for the construction of a clinical corpus. This is illustrated by setting up JSYNCC, the first publicly available corpus of German clinical language. Although addressing German as an example, our approach is easily portable to other languages.

[5] https://sites.google.com/site/clefehealth/

Source	Area	Register	Doc.	Sent.	Types	Tokens
(Siekmann and Irlenbusch, 2012) (Siekmann and Klima, 2013) (Siekmann et al., 2016)	orthopedics trauma surgery	surgery reports	337	16,723	16,001	174,598
(Hagen, 2005)	general surgery	surgery reports	62	1,835	3,190	18,824

(Wenzel, 2015)	emergency medicine	case descriptions
		case discussions
			48
			48
				2,710	11,933	63,316

(Eisoldt, 2017)	general surgery	case descriptions	140	699	2,162	10,338
(Hübler and Koch, 2014)	anesthetics	case descriptions	35	934	3,783	13,919
(Machado, 2013)	emergency medicine	case descriptions	11	398	1,785	5,950
(Thiel et al., 2013)	ophthalmology	case descriptions	36	540	3,108	10,181
(Hellmich, 2017)	internal medicine	case descriptions	150	1,056	3,113	15,658
JSynCC total			**867**	**24,895**	**32,108**	**312,784**

Table 2: Raw data and summary statistics of JSynCC. The top three entries contain duplicates and are therefore presented jointly.

The procedure for corpus construction first identifies publicly available (yet IPR-protected) synthetic medical language data contained in e-books as a reasonable substitute for authentic clinical data. Second, a software infrastructure is shared for constructing exact copies of the original corpus (given access to the raw data based on a purchasable license from the publisher of the e-books selected) rather than distributing the copyrighted raw data physically. This workflow can further be broken down into the following steps which we will discuss in more detail below:

- identification and selection of relevant language data,
- extraction and cleansing of relevant content,
- reformatting of the documents in XML and validation of the corpus file,
- optionally followed by annotation and/or text analysis.

3.1. Manual Selection of Raw Data

The e-books which form the basis of our corpus (see Table 2 for a detailed description) contain *synthetic* surgery reports, case descriptions as well as case discussions written by clinical professionals with the intention of training medical students in clinical documentation and decision-making. Together, they constitute the three registers of JSynCC. This way, the problem of getting clearance from third-party ethical and legal bodies to deal with clinical documents addressing *real* individuals can be avoided.

As a result of carefully reviewing available educational material for the German language, we came up with a list of ten textbooks containing suitable synthetic data: (Siekmann and Irlenbusch, 2012), (Siekmann and Klima, 2013), and (Siekmann et al., 2016) supply reports of orthopedics and trauma surgeries, including fictional administrative information, such as patient numbers, time and date information, as well as the name of the physician in charge. Since the most recent of these publications is in part a combination of the earlier two, duplicated entries were removed during extraction resulting in a total of 337 synthetic documents. Similarly, (Hagen, 2005) provides 62 synthetic reports dealing with general surgery. Altogether, these 399 documents extracted from the first four books make up the *surgery reports* register of JSynCC. Furthermore, (Wenzel, 2015) describes 48 case examples from emergency medicine including an extensive discussion for each of these cases (forming the *case discussions* register). (Eisoldt, 2017) contains 140 short surgical case descriptions and (Hübler and Koch, 2014) comprises 35 case descriptions from anesthetics. (Machado, 2013) describes 11 medical emergency situations occurring in intercultural settings. Finally, JSynCC also contains 36 case descriptions from ophthalmology by (Thiel et al., 2013) and 140 case descriptions from internal medicine by (Hellmich, 2017). Together, the documents extracted from these last six books form the *case descriptions* register of JSynCC.

3.2. Automatic Extraction of Relevant Content

The whole process of creating our corpus runs fully automatically (mainly using JAVA) and is scripted so that everyone having access to the pre-selected e-books can generate an exact copy of JSynCC without any manual intervention (other than assembling the necessary raw data). The entire software package for corpus construction is released on our lab's GITHUB site.[6]

For converting the e-books (which originally come in PDF format) to plain text, we used the JAVA-based APACHE TIKA tool[7] and the command-lines tools PDFTOTEXT[8] and PDFTOHTML[9] depending on which one generated the cleanest output for the book under scrutiny. For each book we determined unique typographic characteristics by which the starting and ending points of the relevant textual excerpts could be reliably localized. Along with each of the resulting 867 documents, we also extracted the following kinds of meta-information:

(1) the title of the book the excerpt originates from,
(2) the register of our corpus we assign it to (surgery reports, case descriptions, or case discussions),
(3) its heading as given in the textbook, as well as its topic.

The topic labels were assigned based on the chapter in which a given excerpt appeared, or similar structural in-

[6] https://github.com/JULIELab/jsyncc

[7] https://tika.apache.org/

[8] http://www.xpdfreader.com/pdftotext-man.html

[9] https://linux.die.net/man/1/pdftohtml

formation, with the heading labels being assigned based on the table of contents in the respective book.

In order to minimize the number of potential conversion errors in the final version of JSynCC, we automatically post-processed the output of the tools mentioned above in a rule-based fashion. We manually revised each of the extracted documents, iteratively improving the post-processing procedure for each individual book. In this manner, we removed artifacts originating from the conversion of the print layout to plain text such as additional line breaks (while keeping paragraph segmentation), formatting characters, as well as superfluous hyphenation.

3.3. XML Conversion and Validation

The extracted and post-processed documents are stored in a single XML file together with their accompanying metadata. Figure 1 illustrates the structure of this file, as well as an exemplary section of the raw text in JSynCC. To make sure that the corpus creation script works as intended for anyone wanting to re-build JSynCC, we also provide checksums in a separate XML file for each individual document on which a validation script is based. By this, we can trustfully guarantee that each corpus copy created using our technical set-up is a valid copy (that is, it always produces the same corpus given the same selection of input documents).

3.4. Automatic Annotation and Analysis

In order to illustrate the potential of JSynCC as a future benchmark corpus for German clinical NLP, we also provide automatically derived annotations on token-, sentence- and part-of-speech level. We used the Uima-based tool suite JCoRe[10] (Hahn et al., 2016) employing the publicly available models trained on the confidential clinical FraMed corpus (Faessler et al., 2014). The resulting annotations are available in a stand-off XML format.

Based on these metadata, we computed the number of types, tokens, and sentences for each of the source textbooks (see Table 2) as well as those of the JSynCC corpus as a whole. As can be seen from Table 1, besides being the first publicly available data set of German clinical language, JSynCC is also the largest corpus ever published (containing over 300k tokens and 30k types).

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<corpus>
  <document>
    <id>1</id>
    <text>Vorgeschichte/Indikation: Sturz auf den
      Schädel unter Alkoholeinfluss. Anschl.
      HWS-Schmerzen. Konventionell radiologisch sowie
      im CT Nachweis der u.g. Fraktur. (...)</text>
    <type>operation report</type>
    <heading>Densfraktur - Verschraubung</heading>
    <topic>Orthopädie</topic>
    <topic>Unfallchirurgie</topic>
    <source>Siekmann, H., Irlenbusch, L., and Klima, S.
      (2016). Operationsberichte Orthopädie und
      Unfallchirurgie. Springer-Verlag.</source>
  </document>  (...)
</corpus>
```

Figure 1: Illustration of the automatically extracted and post-processed corpus in XML format.

4. Conclusion

Almost unsurmountable legal problems encountered when dealing with clinical documents in Germany and many other European countries have led us to consider using made-up *synthetic* rather than real *authentic* language data. Such substitutes can easily be extracted from electronically published educational medical textbooks. Thus, privacy protection concerns do not arise since artificial actors rather than real-life individuals are in focus.

Based on this design decision, we here outline a methodology for corpus development which leads to the creation of copy-identical (as guaranteed by checksums) corpora which can be trustfully built on demand at any physical lab site using the same software and selections of textual raw data. Since no textual data are physically distributed, legal IPR issues are avoided as well. Thus, we share corpus building software without touching sensitive legal ground related to IPR-protected raw data. Still, the NLP community is able to work with these corpora without any restriction and loss in raw data quality based on an easily manageable technical bypass.

We illustrated this highly portable approach introducing JSynCC, the largest and—even more importantly—first *publicly available* corpus of German clinical language. Hence, for the first time ever, research on German clinical NLP (and other language communities on which strict legal protection constraints are imposed) can benefit from community-wide annotation efforts which may transform JSynCC (and potential follow-up data sets) into a benchmark corpus for various tasks in future work.

The next obvious problem that needs to be tackled relates to the main assumption underlying our approach, i.e., assessing the similarity of authentic and synthetic clinical documents and thus estimating their substitutability. Accordingly, a stylistic sublanguage comparison study will be carried out in the future.

5. Acknowledgements

This work was partially supported by *Deutsche Forschungsgemeinschaft* (DFG) under grant HA 2097/8-1 within the STAKI^{2}B^2 project (Semantic Text Analysis for Quality-controlled Extraction of Clinical Phenotype Information within the Framework of Healthcare Integrated Biobanking).

6. Bibliographical References

Afzal, Z., Pons, E., Kang, N., Sturkenboom, M. C. J. M., Schuemie, M. J., and Kors, J. A. (2014). ContextD: An algorithm to identify contextual properties of medical terms in a Dutch clinical corpus. *BMC Bioinformatics*, 15:373.

Bretschneider, C., Zillner, S., and Hammon, M. (2013a). Grammar-based lexicon extension for aligning German radiology text and images. In *RANLP 2013 — Proceedings of the 9th International Conference on Recent Advances in Natural Language Processing. Hissar, Bulgaria, September 7-13, 2013*, pages 105–112.

Bretschneider, C., Zillner, S., and Hammon, M. (2013b). Identifying pathological findings in German radiology

[10] http://julielab.github.io/

reports using a syntacto-semantic parsing approach. In *BioNLP 2013 — Proceedings of the 2013 Workshop on Biomedical Natural Language Processing @ ACL 2013. Sofia, Bulgaria, August 8, 2013*, pages 27–35.

Chapman, W. W., Nadkarni, P. M., Hirschman, L., D'Avolio, L. W., Savova, G. K., and Uzuner, O. (2011). Overcoming barriers to NLP for clinical text: The role of shared tasks and the need for additional creative solutions. *Journal of the American Medical Informatics Association*, 18(5):540–543.

Chapman, W. W., Hillert, D., Velupillai, S., Kvist, M., Skeppstedt, M., Chapman, B. E., Conway, M., Tharp, M., Mowery, D. L., and Deléger, L. (2013). Extending the NEGEX lexicon for multiple languages. In *MEDINFO 2013 — Proceedings of the 14th World Congress on Medical and Health Informatics. Copenhagen, Denmark, 20-23 August 2013*, pages 677–681.

Cotik, V., Roller, R., Xu, F., Uszkoreit, H., Budde, K., and Schmidt, D. (2016). Negation detection in clinical reports written in German. In *BioTxtM 2016 — Proceedings of the 5th Workshop on Workshop on Building and Evaluating Resources for Biomedical Text Mining @ COLING 2016. Osaka, Japan, December 12, 2016*, pages 115–124.

Dalianis, H., Hassel, M., and Velupillai, S. (2009). The Stockholm EPR Corpus: Characteristics and some initial findings. In *ISHIMR 2009 — Evaluation and Implementation of e-Health and Health Information Initiatives: International Perspectives. Proceedings of the 14th International Symposium for Health Information Management Research. Kalmar, Sweden, October 14-16, 2009*, pages 243–249.

Deléger, L., Ligozat, A.-L., Grouin, C., Zweigenbaum, P., and Névéol, A. (2014). Annotation of specialized corpora using a comprehensive entity and relation scheme. In *LREC 2014 — Proceedings of the 9th International Conference on Language Resources and Evaluation. Reykjavik, Iceland, May 26-31, 2014*, pages 1267–1274.

Faessler, E., Hellrich, J., and Hahn, U. (2014). Disclose models, hide the data: How to make use of confidential corpora without seeing sensitive raw data. In *LREC 2014 — Proceedings of the 9th International Conference on Language Resources and Evaluation. Reykjavik, Iceland, May 26-31, 2014*, pages 4230–4237.

Ferraro, J. P., Daumé, H., DuVall, S. L., Chapman, W. W., Harkema, H. H., and Haug, P. J. (2013). Improving performance of natural language processing part-of-speech tagging on clinical narratives through domain adaptation. *Journal of the American Medical Informatics Association*, 20(5):931–939.

Fette, G., Ertl, M., Wörner, A., Klügl, P., Störk, S., and Puppe, F. (2012). Information extraction from unstructured electronic health records and integration into a data warehouse. In *INFORMATIK 2012: Was bewegt uns in der/die Zukunft? Proceedings der 42. Jahrestagung der Gesellschaft für Informatik e.V. (GI). Braunschweig, Germany, September 16-21, 2012*, number P-208 in GI-Edition - Lecture Notes in Informatics (LNI), pages 1237–1251. Gesellschaft für Informatik e.V. (GI), Bonner Köllen Verlag.

Fischer-Hübner, S., Hoofnagle, C., Krontiris, I., Rannenberg, K., and Waidner, M. (2011). Manifesto from Dagstuhl Perspectives Workshop 11061 "Online Privacy: Towards Informational Self-Determination on the Internet". Technical report, Schloss Dagstuhl, Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany.

Hahn, U., Matthies, F., Faessler, E., and Hellrich, J. (2016). UIMA-based JCORE 2.0 goes GITHUB and MAVEN CENTRAL: State-of-the-art software resource engineering and distribution of NLP pipelines. In *LREC 2016 — Proceedings of the 10th International Conference on Language Resources and Evaluation. Portorož, Slovenia, May 23-28, 2016*, pages 2502–2509.

Hahn, U., Matthies, F., Lohr, C., and Löffler, M. (2018). 3000PA: Towards a national reference corpus of German clinical language. In *MIE 2018 — Proceedings of the 29th Medical Informatics in Europe Conference. Gothenburg, Sweden, April 23-25, 2018*.

Hellrich, J., Matthies, F., Faessler, E., and Hahn, U. (2015). Sharing models and tools for processing German clinical texts. In *MIE 2015 — Proceedings of the 26th Medical Informatics in Europe Conference. Madrid, Spain, May 27-29, 2015*, pages 734–738.

Huang, C.-C. and Lu, Z. (2016). Community challenges in biomedical text mining over 10 years: Success, failure and the future. *Briefings in Bioinformatics*, 17(1):132–144.

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M. M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:#160035.

Krebs, J., Corovic, H., Dietrich, G., Ertl, M., Fette, G., Kaspar, M., Krug, M., Stoerk, S., and Puppe, F. (2017). Semi-automatic terminology generation for information extraction from German chest X-ray reports. In Rainer Röhrig, et al., editors, *German Medical Data Sciences: Visions and Bridges. Proceedings of the 62nd Annual Meeting of the German Association of Medical Informatics, Biometry and Epidemiology (gmds e.V.). Oldenburg (Oldenburg), Germany, 17-21 September 2017 — GMDS 2017*, number 243 in Studies in Health Technology and Informatics, pages 80–84. IOS Press.

Kreuzthaler, M. and Schulz, S. (2015). Detection of sentence boundaries and abbreviations in clinical narratives. *BMC Medical Informatics and Decision Making*, 15(Suppl 2):S4.

Kreuzthaler, M., Oleynik, M., Avian, A., and Schulz, S. (2016). Unsupervised abbreviation detection in clinical narratives. In *ClinicalNLP 2016 — Proceedings of the Clinical Natural Language Processing Workshop @ COLING 2016. Osaka, Japan, December 11, 2016*, pages 91–98.

Kumar, V., Stubbs, A., Shaw, S. Y., and Uzuner, O. (2015). Creation of a new longitudinal corpus of clinical narratives. *Journal of Biomedical Informatics*, 58(Supplement: Proceedings of the 2014 i2b2/UTHealth Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data)):S6–S10.

Lohr, C. and Herms, R. (2016). A corpus of German clinical reports for ICD and OPS-based language modeling. In *CLAW 2016 — Proceedings of the 6th Workshop on Controlled Language Applications @ LREC 2016. Portorož, Slovenia, May 28, 2016*, pages 20–23.

Marciniak, M. and Mykowiecka, A. (2011). Towards morphologically annotated corpus of hospital discharge reports in Polish. In *BioNLP 2011 — Proceedings of the Workshop on Biomedical Natural Language Processing @ ACL-HLT 2011. Portland, Oregon, USA, June 23-24, 2011*, pages 92–100.

Mittelstadt, B. D. and Floridi, L. (2016). The ethics of Big Data: Current and foreseeable issues in biomedical contexts. *Science and Engineering Ethics*, 22(2):303–341.

Oronoz, M., Gojenola, K., Pérez, A., Diaz de Ilarraza, A., and Casillas, A. (2015). On the creation of a clinical gold standard corpus in Spanish: Mining adverse drug reactions. *Journal of Biomedical Informatics*, 56:318–332.

Pommerening, K., Drepper, J., Helbing, K., and Ganslandt, T. (2014). *Leitfaden zum Datenschutz in medizinischen Forschungsprojekten. Generische Lösungen der TMF 2.0*. Number 11 in Schriftenreihe: Projektergebnisse aus der TMF und Referenzwerke zur Gesundheitstelematik. TMF — Technologie- und Methodenplattform für die vernetzte medizinische Forschung e.V., Berlin.

Roberts, K., Simpson, M. S., Demner-Fushman, D., Voorhees, E. M., and Hersh, W. R. (2016). State-of-the-art in biomedical literature retrieval for clinical cases: A survey of the TREC 2014 CDS Track. *Information Retrieval Journal*, 19(1-2):113–148.

Roller, R., Uszkoreit, H., Xu, F., Seiffe, L., Mikhailov, M., Staeck, O., Budde, K., Halleck, F., and Schmidt, D. (2016). A fine-grained corpus annotation schema of German nephrology records. In *ClinicalNLP 2016 — Proceedings of the Clinical Natural Language Processing Workshop @ COLING 2016. Osaka, Japan, December 11, 2016*, pages 69–77.

Rosenthal, S., Nakov, P. I., Kiritchenko, S., Mohammad, S. M., Ritter, A., and Stoyanov, V. (2015). SemEval-2015 Task 10: Sentiment analysis in TWITTER. In *SemEval 2015 — Proceedings of the 9th International Workshop on Semantic Evaluation @ NAACL-HLT 2015. Denver, Colorado, USA, June 4-5, 2015*, pages 451–463.

Savkov, A., Carroll, J. A., Koeling, R., and Cassell, J. A. (2016). Annotating patient clinical records with syntactic chunks and named entities: The HARVEY Corpus. *Language Resources and Evaluation*, 50(3):523–548.

Schulz, S. and López-García, P. (2015). Big Data, medizinische Sprache und biomedizinische Ordnungssysteme. *Bundesgesundheitsblatt – Gesundheitsforschung – Gesundheitsschutz*, 58(8):844–852.

Stubbs, A. and Uzuner, O. (2015a). Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth Corpus. *Journal of Biomedical Informatics*, 58(Supplement):S20–S29.

Stubbs, A. and Uzuner, O. (2015b). Annotating risk factors for heart disease in clinical narratives for diabetic patients. *Journal of Biomedical Informatics*, 58(Supplement):S78–S91.

Sun, W., Rumshisky, A., and Uzuner, O. (2013). Annotating temporal information in clinical narratives. *Journal of Biomedical Informatics*, 46(Supplement: Proceedings of the 2012 i2b2 NLP Challenge on Temporal Relations in Clinical Data):S5–S12.

Šuster, S., Tulkens, S., and Daelemans, W. (2017). A short review of ethical challenges in clinical natural language processing. In *Proceedings of the 1st ACL Workshop on Ethics in Natural Language Processing @ EACL 2017. Valencia, Spain, April 4, 2017*, pages 80–87.

Toepfer, M., Corovic, H., Fette, G., Klügl, P., Störk, S., and Puppe, F. (2015). Fine-grained information extraction from German transthoracic echocardiography reports. *BMC Medical Informatics and Decision Making*, 15:#91.

Tomanek, K., Wermter, J., and Hahn, U. (2007). A reappraisal of sentence and token splitting for life sciences documents. In Klaus A. Kuhn, et al., editors, *MedInfo 2007 — Proceedings of the 12th World Congress on Health (Medical) Informatics. Building Sustainable Health Systems. Brisbane, Australia, August 20-24, 2007*, number 129 in Studies in Health Technology and Informatics, pages 524–528, Amsterdam. IOS Press.

Truyens, M. and Van Eecke, P. (2014). Legal aspects of text mining. In Nicoletta Calzolari, et al., editors, *LREC 2014 — Proceedings of the 9th International Conference on Language Resources and Evaluation. Reykjavik, Iceland, May 26-31, 2014*, pages 2182–2186.

UrhWissG. (2018). Gesetz zur Angleichung des Urheberrechts an die aktuellen Erfordernisse der Wissensgesellschaft (Urheberrechts-Wissensgesellschafts-Gesetz – UrhWissG) vom 1. September 2017 (BGBl. Teil I Nr. 61, Bonn, September 7th 2017), came into force on March 1, 2018. In *BGBl. I*, pages 3346–3351. Available: `https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/BGBl-UrhWissG.pdf`.

Uzuner, O., Solti, I., Xia, F., and Cadag, E. (2010). Community annotation experiment for ground truth generation for the i2b2 medication challenge. *Journal of the American Medical Informatics Association*, 17(5):519–523.

Wang, Y., Pakhomov, S. V. S., Ryan, J. O., and Melton, G. B. (2015). Domain adaption of parsing for operative notes. *Journal of Biomedical Informatics*, 54:1–9.

Wermter, J. and Hahn, U. (2004). Really, is medical sublanguage that different? Experimental counter-evidence from tagging medical and newspaper corpora. In Marius Fieschi, et al., editors, *MEDINFO 2004 — Proceedings of the 11th World Congress on Medical Informatics. San Francisco, California, USA, September 7-11, 2004*, volume 1 of *Studies in Health Technology and Informatics*, pages 560–564, Amsterdam. IOS Press.

Zhang, Y., Tang, B., Jiang, M., Wang, J., and Xu, H. (2015). Domain adaptation for semantic role labeling of clinical text. *Journal of the American Medical Informatics Association*, 22(5):967–979.

7. Language Resource References

Eisoldt, Stefan. (2017). *Fallbuch Chirurgie: 140 Fälle aktiv bearbeiten*. Georg Thieme Verlag, 5th edition.

Hagen, Monika. (2005). *Operationsberichte für Einsteiger-Chirurgie: Operation vorbereiten—Bericht diktieren*. Georg Thieme Verlag.

Hellmich, Bernhard. (2017). *Fallbuch Innere Medizin*. Georg Thieme Verlag, 5th edition.

Hübler, Matthias and Koch, Thea. (2014). *Komplikationen in der Anästhesie*. Springer-Verlag, 3th edition.

Machado, Carl. (2013). *Patienten aus fremden Kulturen im Notarzt- und Rettungsdienst: Fallbeispiele und Praxistipps*. Springer-Verlag.

Siekmann, Holger and Irlenbusch, Lars. (2012). *Operationsberichte Unfallchirurgie*. Springer-Verlag.

Siekmann, Holger and Klima, Stefan. (2013). *Operationsberichte Orthopädie: mit speziellen unfallchirurgisch-orthopädischen Eingriffen*. Springer-Verlag.

Siekmann, Holger and Irlenbusch, Lars and Klima, Stefan. (2016). *Operationsberichte Orthopädie und Unfallchirurgie*. Springer-Verlag.

Thiel, Michael A and Bernauer, Wolfgang and Schüpfer, Marlis Zürcher and Schmid, Martin K. (2013). *Fallbeispiele Augenheilkunde*. Springer-Verlag.

Wenzel, Volker. (2015). *Fallbeispiele Notfallmedizin: Einprägsam—spannend—mit Lerneffekt*. Springer-Verlag.