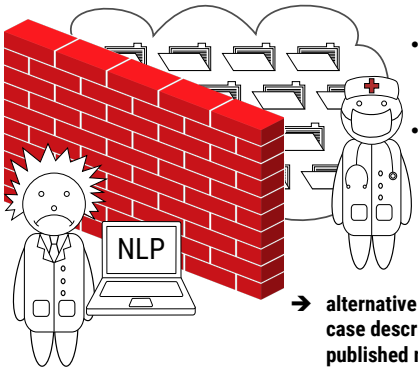
Sharing Copies of Synthetic Clinical Corpora without Physical Distribution — A Case Study to Get Around IPRs and Privacy Constraints Featuring the German JSynCC Corpus

Christina Lohr, Sven Buechel, Udo Hahn

Jena University Language & Information Engineering (JULIE) Lab, Friedrich-Schiller-Universität Jena, Germany
www.julielab.de

Clinical Data Are Always Sparse and Typically Locked



- clinical data are usually protected by national law, especially data privacy acts
- for research purposes (in the US only!): *Data Use Agreements* between data provider (hospital) and data consumer (researcher)

➔ **alternative solution: reuse fictitious clinical case descriptions and surgery reports from published medical textbooks**

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. (2013)	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	867	24,895	**32,108**	**312,784**	✓

JSynCC – Jena Synthetic Clinical Corpus

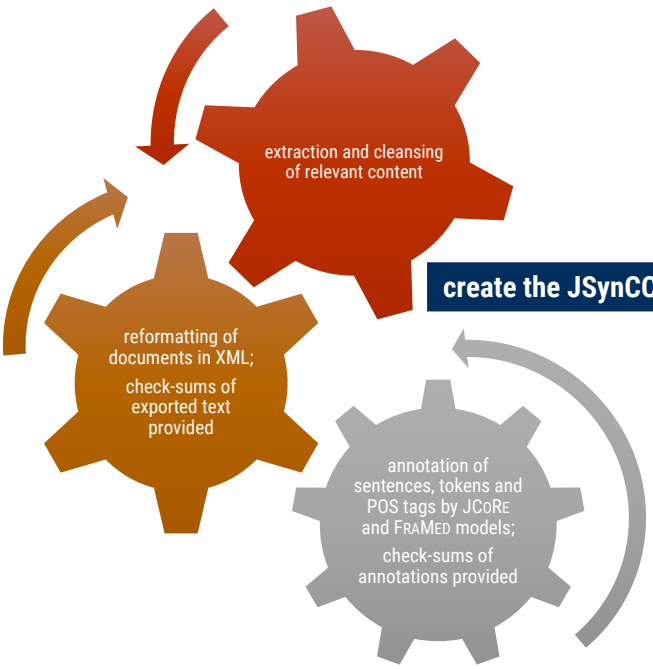
- **Java tool used to parse 10 German medical textbooks**

subscribe to 10 books

➔ **download JSynCC build software from GitHub:**

https://github.com/JULIELab/jsyncc

Source	Medical Area	Documents	Sentences	Types	Tokens
Surgery reports					
Siekmann and Irlenbusch, 2012 Siekmann and Klima, 2013 Siekmann et al., 2016	orthopedics trauma surgery	337	16,723	16,001	174,598
Hagen, 2005	general surgery	62	1,835	3,190	18,824
Case descriptions, case discussions					
Wenzel, 2015	emergency medicine	96	2,710	11,933	63,316
Case descriptions					
Eisoldt, 2017	general surgery	140	699	2,162	10,338
Hübler and Koch, 2014	anesthetics	35	934	3,783	13,919
Machado, 2013	emergency medicine	11	398	1,785	5,950
Thiel et al., 2013	ophthalmology	36	540	3,108	10,181
Hellmich, 2017	internal medicine	150	1,056	3,113	15,658
JSynCC total		**867**	**24,895**	**32,108**	**312,784**



create the JSynCC Corpus

```
<?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. Bei
      Instabilität Indikation zur Verschraubung. Der Pat. hat nach
      entsprechend umfangreicher Risikoaufklärung in die Operation
      eingewilligt. Diagnose: Geschlossene, instabile Fraktur des Dens
      axis (Anderson II, keine Neurologie) Operation: Geschlossene
      Reposition, Osteosynthese mittels zweier  kanülierte Zugschrauben
      (38 + 40 mm) Vorgehen: Unkomplizierte ITN. Cefuroxim 1,5 g  (…)
    </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>
```

Our main contributions here:

- Replace authentic clinical data (classified) by synthetic, i.e., fictitious ones (accessible on a license basis)
- Share software that allows to rebuild check-sum-identical synthetic text corpora anywhere/anytime, without distributing the source text corpora

LREC 2018 MIYAZAKI

This work was partially supported by Deutsche Forschungsgemeinschaft (DFG) under grant HA 2097/8-1 within the STAKI²B² project (Semantic Text Analysis for Quality-controlled Extraction of Clinical Phenotype Information within the Framework of Healthcare Integrated Biobanking).

DFG