Medical Informatics Europe:
MIE 2018

3000PA—Towards a National Reference Corpus of German Clinical Language

Udo Hahn[a], Franz Matthies[a], Christina Lohr[a], Markus Löffler[b]

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

[b]Institute for Medical Informatics, Statistics and Epidemiology (IMISE)
Universität Leipzig, Germany
www.imise.uni-leipzig.de

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

SMITH
Smart Medical Information Technology for Healthcare

www.smith.care

UNIVERSITÄT LEIPZIG

Towards a German Clinical Reference Corpus

German-language clinical corpora

- FRAMED: hybrid mix of several (non-)clinical text genres (non-distributable; LREC 2004)
- many other small- and medium-sized corpora from single clinical sites and single genres; non-distributable (German data privacy law)
- vision of a national reference corpus: cross-hospital, cross-genre collection of clinical reports (distributable under DUAs)

3000PA Corpus

- part of SMITH (one of four funded consortia (40 Mio. €) within a major national German initiative to foster medical informatics research (BMBF))
- first national reference corpus for German clinical documents
- ≈ 1000 electronic patient records from three German university hospitals (Aachen, Jena and Leipzig)
- 2010-2015; internistic or ICU units; patients deceased

Why a (German) text corpus?

- **collections of (machine-readable) text, either used for**
  - training NLP systems in a (semi-)supervised way
  - evaluating the performance of (NLP) systems (benchmark data sets)

- **(German) clinical text data**
  - medical jargon constitutes a sublanguage on its own
  - differ across hospitals, clinical departments and text genres
  - evidence from (distributable) English clinical corpora is not transferable to German

3000PA

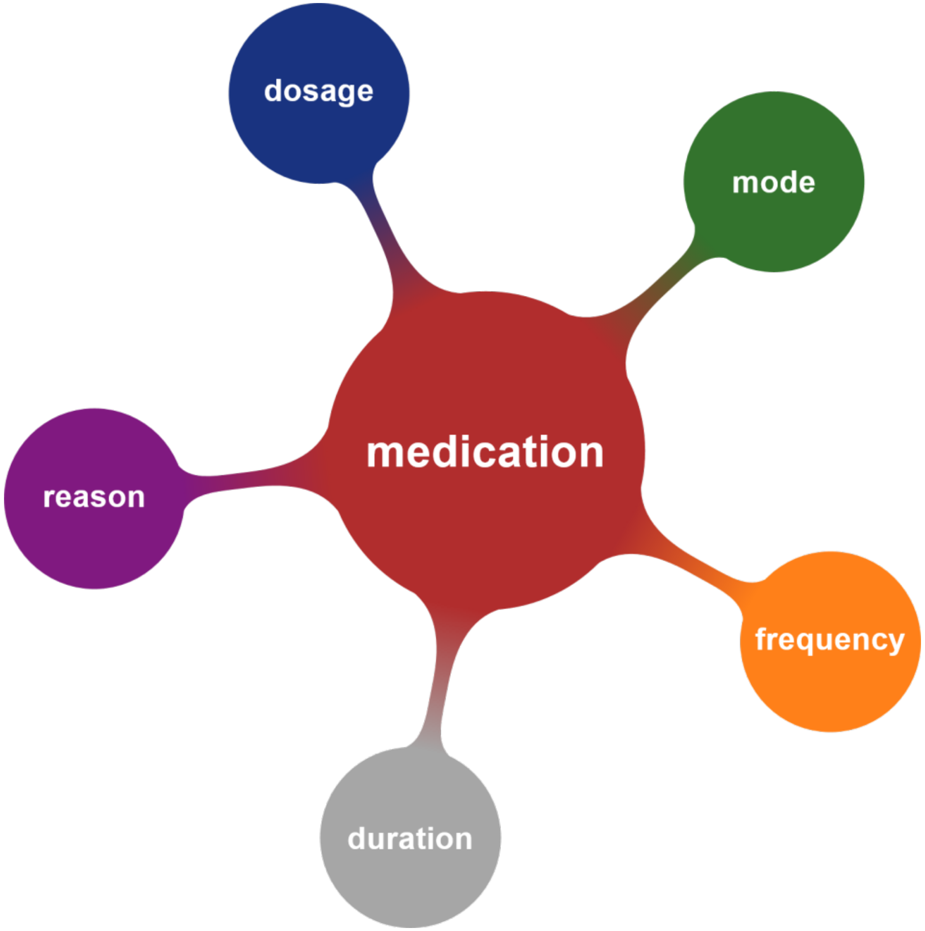
Jena slice of 3000PA

- 1107 text documents
  - 620 discharge summaries
  - 487 transfer letters
- 1.75 Mio tokens
- 180 K sentences

- Leipzig and Aachen slices exhibit similar numbers
  - altogether, roughly > 5 Mio tokens and 500 K sentences
  - meanwhile, five clinical sites have joined the SMITH consortium and will, additionally, contribute around 1000 EPRs each

Case study – Medication

- pilot study: corpus can be used by clinic-external staff (JULIE Lab)
- replication of a task similar to the 3rd i2b2 challenge on medication extraction (JAMIA 2010)
- adaptation of English i2b2 guidelines to German clinical language using 3000PA



Metadata relevant for medication extraction

• medication	*The patient received* ***aspirin****.*
• dosage	*Oxycodon* ***5-10 mg***
• mode	***intravenous*** *prednisolone therapy*
• frequency	***twice a day***
• reason	*Due to the* ***dyspnoea symptoms*** *...*

Two studies related to medication extraction

manual annotation campaign

- documents annotated with medication information
- BRAT tool
- five students of medicine
- 52 documents double-annotated for measuring the agreement (IAA)

automatic medication extraction

- adaptation of the (English) MEDXN system (JAMIA 2014) to German: JUMEX
- based on regular expressions and German dictionaries (Rote Liste)
- rapid prototype only, not tuned for competitions
- processing based on the full Jena slice of 3000PA

Evaluation study

- performance of human annotation (based on inter-annotator agreement – IAA)
- performance of automatic annotation (based on JUMEX)
- all performance data vary dependent on the choice of string overlap criteria (centroids; LREC 2012)

	IAA	JUMEX
frequency	0.91 - 0.98	0.81 - 0.83
dosage	0.81 - 0.83	0.81 - 0.83
medication	0.90 - 0.96	0.49 - 0.50
duration	0.66 - 0.78	0.30 - 0.34
mode	0.69 - 0.85	0.19 - 0.22
reason	0.27 - 0.69	–

Conclusion

- 3000PA: first prototype of a German national reference corpus of clinical documents
  - cross-hospital (3+5), cross-genre (2+x)
  - currently, around 5 Mio. tokens, and 500k sentences
  - annotations available for sentences, tokens, section headings, medications (diseases, symptoms, and therapies soon to come)
- pilot study testing its usability for manual and automatic annotation
- replication of the 3rd i2b2 challenge task for German: medication extraction
- first published German corpus on medication metadata and automatic medication extraction

Medical Informatics Europe:
MIE 2018

3000PA—Towards a National Reference Corpus of German Clinical Language

Udo Hahn[a], Franz Matthies[a], Christina Lohr[a], Markus Löffler[b]

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

[b]Institute for Medical Informatics, Statistics and Epidemiology (IMISE)
Universität Leipzig, Germany
www.imise.uni-leipzig.de

FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

SMITH
Smart Medical Information
Technology for Healthcare

UNIVERSITÄT
LEIPZIG