universität innsbruck

MEDIZINISCHE UNIVERSITÄT INNSBRUCK

DHA goes digital day 25.09.2020

ÖAW ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN

dha digital humanities austria 2020

The corpus *MedCorpInn* – what we did and where we are now

Posch, Claudia; Irschara, Karoline (PIs University of Innsbruck)
Mangesius, Stephanie; Gruber, Leonhard (PIs Medical University of Innsbruck)
Huber, Anna-Lena; Waldner, Brigit (Project Staff)

The corpus MedCorpInn

- What was the project idea?
- What are the goals of the project?
- Which data is being used?
- What happened in project year one?
- Which were the challenges and obstacles?
- What‘s next?

What was the project idea?

- MedCorpInn is an interdisciplinary project between Leopold-Franzens-University and Innsbruck Medical University
- Main goal: create a (linguistically) annotated corpus of radiology reports
- Staff: Dr. Birgit Waldner (MUI & LFU: programming); Dr. Anna Lena Huber (MUI: gender medicine)
- Partners: Bernhard Glodny; Gerhard Rampl; Astrid E. Grams, Irene Gizewski
- Basic Premise: Radiology reports can be viewed as text corpora and thus can be processed and analyzed with CL and CADS methodologies.

Which data is being used?

- *MedCorpInn* currently contains 5,002,933 written reports from the *Clinic of Radiology and Neuroradiology* at *Medical University of Innsbruck*
  - 2,540,022 female patients
  - 2,440,474 male patients
- Language of reports: German
- Time period: 2007-2019
- Content: different examination methods and different medical indications
- Text structure: unstructured reports
- Metadata structure: 39 different categories, mostly structured

What happened in project year one?

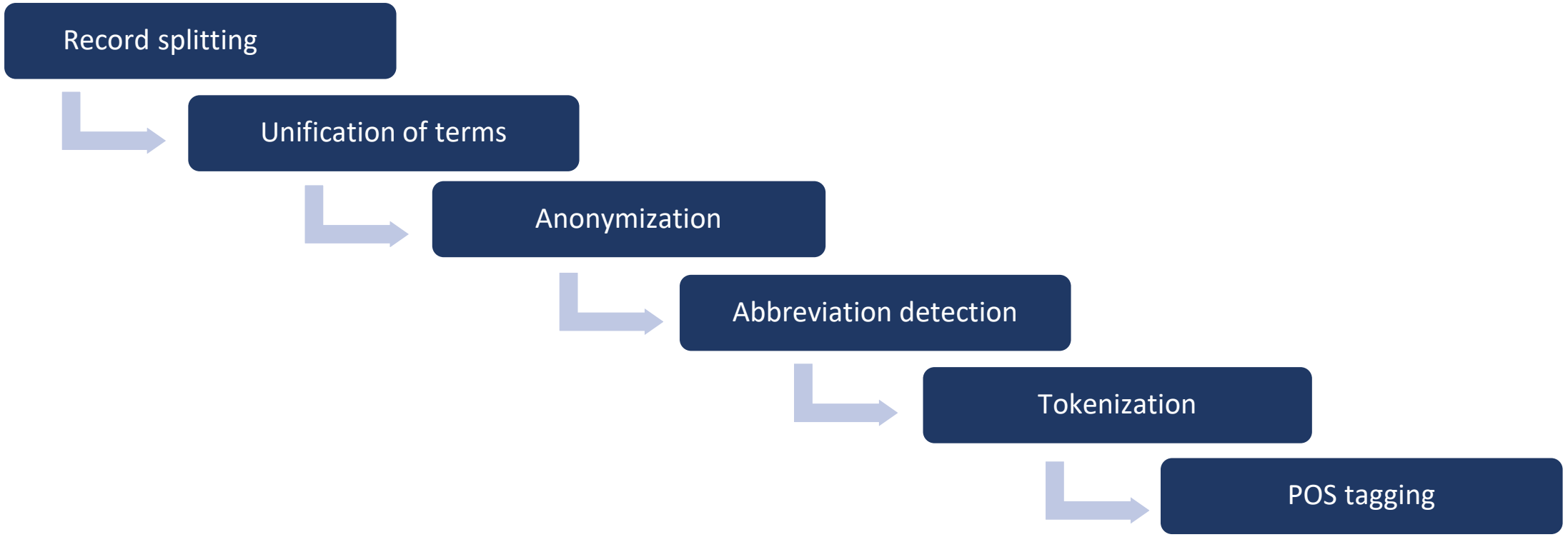
- Export, anonymisation and integration of extensive amount of new data
- Data cleaning and development of data cleaning methods
- Correction of errors in metadata; anonymisation of metadata
- Pipeline development
- Bibliography of gender medinice research; development of research questions in gender medicine;

Metadata categories

» demographic and medical metadata

meta_list = ['Versicherungstyp', 'Datum', 'Sterbedatum', 'Maßnahme', 'Signierer', 'PLZ', 'Zustand', 'Ort',
'Modalität', 'Mitarbeiter', 'Institut', 'Überweisertyp', 'Bereich', 'Geschlecht',
'Geburtsdatum', 'Freigeber', 'Durchführender', 'Diktierer', 'Beruf', 'Religion', 'Befundender',
'Aufnahmeart', 'Arbeitsplatz', 'Aufnahmedatum', 'Abrechnungsstelle', 'Abbruchgrund',
'Maßnahmestatus', 'Größe', 'Gewicht', 'Dienstart', 'Patienten Fremd ID', 'Fall Fremd ID',
'Fallzusatz ID', 'Patient Zusatz ID', 'Anforderungs ID', 'Befund ID', 'Anforderungs Gruppen
ID', 'Interne Patienten ID', 'Überweiser', 'Alter']

Processing Pipeline scheme



Data structure

» structured metadata, but completely unstructured reports
» headings and subheadings are not used consistently

```
Privatkasse;*;*;*;;SKSPL;EKA;6067;;*;;US;RAD-
I;Station;RAUN;weiblich;*;EKA;WOC;EKA;Pension;RK;;Ambulant;UNFF;01.01.2007;RAUN;;bef;;;;201041922
3;35178020;200712438709;1337724;;12309535;;1337724;UNAA;59;
```

Untersuchung: Sprunggelenk und hohe Fibula links vom *:**Untersuchung:** Sprunggelenk und hohe Fibula links vom *:**BEFUND:**Mehrfragmentfraktur im Bereich des distalen Fibulaschaftes Typ Weber C. Das obere Sprunggelenk regelrecht artikulierend..1.1.,0.36h:Die distale Fibulafraktur in Gipsverband in annähernd anantomischer Stellung stabilisiert.Dr. *. */mCT/MR-Terminvergabe: MO-FR von 08.00 Uhr - 16.00 Uhr, Tel.-Nr. 25655 (Fax: 22779)MR-Notfalluntersuchungen bitte unter Tel.-Nr. 25277 anmeldenCT/MR/PET/SPECT - Bildfusion(SIP-Labor)-Terminvergabe: MO-FR von 09.00 Uhr - 13.00 Uhr, Tel.-Nr. 80831 (Fax: 28992)

Data structure

» missing spaces, typos, missing or erroneous punctuation
» end line in various forms

```
Privatkasse;*;*;*;;SKSPL;EKA;6067;;*;;US;RAD-
I;Station;RAUN;weiblich;*;EKA;WOC;EKA;Pension;RK;;Ambulant;UNFF;01.01.2007;RAUN;;bef;;;;2010419
223;35178020;200712438709;1337724;;12309535;;1337724;UNAA;59;
```

```
Untersuchung: Sprunggelenk und hohe Fibula links vom *:Untersuchung: Sprunggelenk und hohe
Fibula links vom *:BEFUND:Mehrfragmentfraktur im Bereich des distalen Fibulaschaftes Typ Weber
C. Das obere Sprunggelenk regelrecht artikulierend..1.1.,0.36h:Die distale Fibulafraktur in
Gipsverband in annähernd anantomischer Stellung stabilisiert.Dr. *. */mCT/MR-Terminvergabe: MO-
FR von 08.00 Uhr - 16.00 Uhr, Tel.-Nr. 25655 (Fax: 22779)MR-Notfalluntersuchungen bitte unter
Tel.-Nr. 25277 anmeldenCT/MR/PET/SPECT - Bildfusion(SIP-Labor)-Terminvergabe: MO-FR von 09.00
Uhr - 13.00 Uhr, Tel.-Nr. 80831 (Fax: 28992)
```

Data structure

- short forms
- ad-hoc forms
- Latin terms or (pseudo-)Latin terms with German morphology
- English terms
- (non-standard) abbreviations
- multi word expressions in English or Latin, e.g. *medulla oblongata*
- Missing or erroneous punctuation

→ effects tokenization, sentence splitting, POS-tagging

Anonymization and pseudonymization

- patients' and doctors' names were removed from metadata
- in the unstructured reports doctors' names appear occasionally
- IDs are pseudonymized (PAT_number), patient follow up is possible through pseudonymized number

```
termlist = ["Mag\. med\. \w. \w+", "FA Priv\.Doz\.Dr\. \w\. \w", "FA Priv\. Doz\. Dr\. \w. \w+", "OA Dr\.* \w\. \w+ \w+-\w+\/\w+","OA \w+", "OA Dr\. \w+",
    "OA Dr\. \w. \w*", "O Dr\. \w\. \w", "OA Dr\. \w\. \w+", "Dr\. \w+\. \w+\/\w+", "Dr\. \w+ \w+\/\w+", "Dr\. \w+\/\w+", "Dr\.* \w\. \w+","OA Dr. \w\. \w+",
    "Dr\. \w\.\w+", "Dr\.* \w+ *%", "Dr \w+", "Prof\. \w+-\w+", "Prof\. \w+", "Prof \w+", "Dr\. \w+“, "Prof\.Dr.\w+", "Dr.\w+", "OA Priv\.-Doz\.* \w\. \w+ \w+-
    \w+\/\w+", "Priv\.-Doz\.* \w\. \w+ \w+-\w+\/\w+", "OA Ass\.-Prof. Priv\.-Doz\.* \w\. \w+ \w+-\w+\/\w+", "ao\. Univß.-Prof\.* \w\. \w+ \w+-\w+\/\w+",
    "o\. Univ\.-Prof\.* \w\. \w+ \w+-\w+\/\w+", "o\. Univ.-Prof. \w\. \w+", "Univ\.-Doz\.* \w\. \w+ \w+-\w+\/\w+", "o\.Univ\.-Prof\.* \w\. \w+ \w+-
    \w+\/\w+", "ao\.Univ\.-Prof\.* \w\. \w+ \w+-\w+\/\w+", "OA Doz\.* \w\. \w+ \w+-\w+\/\w+", "Doz\.* \w\. \w+ \w+-\w+\/\w+", "Hr\. \w\. \w+", "Fr\. \w.
    \w+", "Hr\. \w+", "Fr\. \w+", "Ass\. Prof\.* \w\. \w+ \w+-\w+\/\w+","Ass\. Prof\. \w\. \w+", "ao\. Univ\.-Prof\. \Dr. \w. \w+", "o\. Univ\. Prof\. Dr\. \w.
    \w+", "o\. Univ\. -Prof\.", "o\. Univ\.-Prof\.", "o\. Univ\.-Prof", "Hr\.        Ass\. Prof\. \w+", "OA PD Dr\. * \w\. \w+ \w+-\w+\/\w+", "OÄ Dr\.* \w\. \w+
    \w+-\w+\/\w+", "OA Assoz\.-Prof\. Priv\.-Doz\. Dr\. \w\. \w+", "\. \w\. \w+", "Fecit: \w+", "Fecit: \w\. \w+","Ass\. \w+", "Kopie an: \w", "Kopie an: \w\.
    \w+","Abschrift an: \w+","Abschrift an: \w\. \w+"]
```

MedCorpInn as a source for language and medical research

» Corpus Assisted Discourse Studies, Gender Linguistics & Gender Medicine
» reports as discursive, linguistic events

Research questions:

» Are salient linguistic patterns (keywords, n-grams, collocations) somehow connected to social categories in the metadata?
» How are groups/people/patients talked about?
» Are proposed / described medical procedures connected to social categories in the metadata?
» Are there differences regarding the accuracy of measurements of tumours in connection with social categories?
» …