Challenges in Building MedCorpInn - a Corpus of Unstructured German Radiology Reports

Karoline Irschara*, Claudia Posch*, Birgit Waldner*†, Stephanie Mangesius†, Leonhard Gruber†, Anna-Lena Huber†, Gerhard Rampl*, Astrid Grams†, Bernhard Glodny†

*University of Innsbruck, Department of Linguistics
Innrain 52d, 6020 Innsbruck
karoline.irschara@uibk.ac.at[1], claudia.posch@uibk.ac.at[1], gerhard.rampl@uibk.ac.at
†Medical University of Innsbruck, Department of Radiology
Anichstraße 35, 6020 Innsbruck
Birgit.Waldner@i-med.ac.at, leonhard.gruber@i-med.ac.at, anna.huber@i-med.ac.at,
stephanie.mangesius@tirol-kliniken.at, astrid.grams@i-med.ac.at, bernhard.glodny@i-med.ac.at

1. Introduction

This extended abstract reports on the ongoing project *MedCorpInn - Retrospective Intersectional Corpuslinguistic Analysis of Radiology Reports of Innsbruck Medical University*[2] and outlines the construction of the *MedCorpInn*-corpus, a large linguistically tagged corpus of medical reports. *MedCorpInn* contains 5,002,933 written reports in German (2007-2019) from the Clinic of Radiology and Neuroradiology at Medical University of Innsbruck.[3] First, we give a general description of our data and the current project status. We focus on the particular challenges the radiology reports pose for pipeline building and we suggest possible solutions for the difficulties encountered. Lastly, we give a roadmap on research intended with the corpus once completed.

2. The MedCorpInn data

Some preliminary work had been previously carried out in the smaller pilot project *KARBUN* (Irschara, Posch and Glodny, 2017), which serves as a best practice model for building the more extensive corpus of radiology reports. These texts play an essential role in the communication of physicians and serve as a legal record documenting the imaging procedures (Kahn et al., 2009). They contain information and interpretation regarding the different imaging procedures and types of examinations, e.g. computer tomography, ultrasound, magnetic resonance imaging, angiography, X-ray, fluoroscopy etc. Furthermore, they include demographic information (e.g. age, gender, nationality, type of insurance, occupational status) as well as medical metadata (e.g. mode of examination, time frame, medical indication, referral diagnosis etc.).

Concerning corpus building, there are issues at all levels of pre-processing: A first considerable problem arises when structuring the unstructured data. While the metadata occur in a rather uniform way, the report texts are different varieties of free text entries. For example, reports would often include individual headings or subheadings in a random order, depending on the date of the report, the department or/and the individual doctors. The usage of headings is however not consistent among the records. Another important task is the implementation of a de-identification strategy. The data management first of all complies with data confidentiality (§6 GDPR, current version) as declared in the approval of the ethical review committee of Medical University of Innsbruck. Most of the sensitive information (namely patient and/or doctor names and IDs) were already removed from the metadata during the process of data extraction from the clinical information system (e.g. the patient name was unticked in the extraction form and thus never appears in the metadata in the first place). Some types of metadata are generalized (e.g. date of birth, occupation). However, a limited amount of sensitive information might still appear in the free texts and must be detected and masked. For example, occasionally doctors' names appear and they are removed by using RegEx codes. This is feasible because they always are preceded by either one or more of their academic degrees and/or positions within the clinic or another form of appellation like "Frau" ('Ms.') or "Herr" ('Mr.'). The specific language used in radiology reports makes it particularly challenging to implement a text processing pipeline: Not only do the reports contain many abbreviations, short forms and ad-hoc forms, but also Latin terms or germanized

[1] corresponding authors
[2] The project *MedCorpInn* is funded by the *go!digital* 2.0 call of the Austrian Academy of Science.
[3] Radiology reports were chosen as a starting point because they were readily accessible; the use of further text types as well as data from other clinics in future projects is intended.

(pseudo-)Latin terms. In addition, typos, missing spaces and the lack of paragraph structure make it difficult to perform paragraph and sentence splitting as well as tokenization using available NLP packages. Hence, conventional German POS-taggers have problems with the correct recognition of medical terms since they are usually oriented towards standard language texts (cf. Hellrich, Matthies, Faessler and Hahn, 2015). Nonetheless, we consider POS-tagging important for NER and lemmatization. It is also useful for word-sense disambiguation and for finding lexical or grammatical patterns in the data.

Additionally, medical term and abbreviation recognition are crucial for further corpus development, such as sentence boundary disambiguation and tokenization, which form the basis of our corpus pipeline. Regarding medical term recognition and abbreviation detection we are in the process of constructing relevant thesauri from the corpus and also partly work with the German translation of the extensive radiological ontology RadLex (RNSA, 2017).

Furthermore, the team is currently testing the possibility of transferring some of the solutions on noisiness of clinical data that exist for English (cf. Cai et al., 2016; Chapman et al., 2011; Šuster et al., 2017) onto German, even though this will not be possible for all the language specific problems.

3. MedCorpInn as a source for language and medical research

We suggest viewing the reports as in *MedCorpInn* as linguistic events, which may be affected by linguistic and social aspects, which can be analysed and researched with the corpus *MedCorpInn*. Since the 1990s, research on healthcare communication has predominantly focused on qualitative methods for investigating, for example, doctor-patient interaction (Maynard and Heritage, 2005; Atkins and Harvey, 2010; Menz, 2011). Subsequently, the focus was expanded to medical communication in a broader sense, e.g. to medical discourses or internal clinical communication (Crawford, Brown and Harvey, 2014). Recently, the fields of Corpus Linguistics and Natural Language Processing have opened new approaches to medical data which allow qualitative and quantitative methods to be combined (Taylor and Marchi, 2018; Wiegand and Mahlberg, 2019; Demjén, 2020).

For the German language, there are only few studies which apply mixed methods on clinical data: Most research of the investigations in this area focuses on developing specific NLP applications (e.g. information retrieval, de-identification etc.), but there is no substantial linguistic research on the data itself (Crawford, Brown and Harvey, 2014). Also, there is little research concerning corpus linguistic and especially discourse linguistic investigations of clinical text corpora (Demjén, 2020).

MedCorpInn therefore serves as a unique source for studies both in the fields of linguistics as well as medicine. The language used in the MedCorpInn data provides insights into the everyday communicative practices between health professionals within the fields of radiology and neuroradiology, which have only been marginally investigated from a linguistic perspective so far (Reiner, 2012).

The research questions we intend to investigate with this corpus concern (discourse) linguistic questions, which can be linked to gender medicine issues. For example, we want to analyse whether salient linguistic patterns such as keywords, n-grams or specific collocations (and collocation types) in the data are somehow connected to social categories in the metadata (e.g. age, gender, origin, type of insurance etc.) and how. Could such a connection for example indicate bias, e.g. by certain usage patterns of diminutives and amplifiers? We aim to find out how patients/groups/people are talked about, which information is made explicit on the linguistic surface and how (e.g. an examiner constantly labelling a patient as ‘asylum seeker’). As the texts are often highly standardized and schematic, variability may also be an interesting subject for study.

From a gender medicine perspective, the corpus can be queried for proposed and described medical procedures (e.g. screenings, preventive examinations) and investigate if such propositions are connected to the social and economic categories in the metadata. We also perform specific information extraction tasks on the data, e.g. the extraction of measurements of tumour diameters and the type of tumour described in the corpus. There is a number of different motivations behind researching the differences in measurement accuracies, for example to learn about mean sizes of tumours at the time of diagnosis, to look for potential gender differences or to learn about margins of resolution used in the reports in different organs, and with different modalities, such as MRI or CT and many more. We will use the results of this research to determine if there are differences regarding the accuracy of the measurements (e.g. numbers with or without decimal places) in connection with social categories such as gender or age.

As mentioned above, *MedCorpInn* is a corpus project still in progress. It is the first large corpus of its kind for the German language and will be used to research (discourse) linguistic as well as gender medicine research questions. Additionally, the specific type of data will also help furthering NLP methods for German.

4. References

Sarah Atkins and Kevin Harvey. 2010. How to use corpus linguistics in the study of health communication. In: Anne O'Keeffe and Michael McCarthy, ed., *The Routledge Handbook of Corpus Linguistics*. Routledge, New York.

Tianrun Cai, Andreas A. Giannopoulos, Sheng Yu, Tatiana Kelil, Beth Ripley, Kanako K. Kumamaru, Kanako K. Kumamaru, Frank J. Rybicki, and Dimitrios Mitsouras. 2016. Natural Language Processing Technologies in Radiology Research and Clinical Applications. In: *Radiographics: a review publication of the Radiological Society of North America*, Inc 36 (1), S. 176–191. DOI: 10.1148/rg.2016150080.

Paul Crawford, Brian Brown, and Kevin Harvey. 2014. Corpus linguistics and evidence-based health communication. In: Chou Wen-ying, Sylvia Hamilton, Heidi Ehernberger, ed., *The Routledge Handbook of Language and Health Communication*, pages 75–90. Routledge, London, New York.

Wendy W. Chapman, Prakash M. Nadkarni, Lynette Hirschman, Leonard W. D'Avolio, Guergana K. Savova, and Ozlem Uzuner. 2011. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. In: *Journal of the American Medical Informatics Association*: JAMIA 18 (5): S. 540–543. DOI: 10.1136/amiajnl-2011-000465.

Zsófia Demjén. 2020. *Applying linguistics in illness and healthcare contexts*. Bloomsbury Academic, London, New York.

Johannes Hellrich, Franz Matthies, Erik Faessler, and Udo Hahn. 2015. Sharing models and tools for processing German clinical texts. In: *Studies in health technology and informatics*, 210: 734–738.

Karoline Irschara, Claudia Posch, and Bernhard Glodny. 2017. *KARBUN. Korpus radiologischer Befunde der Universitätskliniken für Neuroradiologie und Radiologie.* Leopold-Franzens-Universität Innsbruck.

Charles E. Kahn, Curtis P. Langlotz, Elizabeth S. Burnside, John A. Carrino, David S. Channin, David M. Hovsepian, and Daniel L. Rubin. 2009. Toward best practices in radiology reporting. In: *Radiology* 252 (3): S. 852–856. DOI: 10.1148/radiol.2523081992.

Douglas Maynard and John Heritage. 2005. Conversation analysis, doctor-patient interaction and medical communication. In: *Medical education* 39 (4): 428–435.

Florian Menz. 2011. Doctor-Patient-Communication. In: Ruth Wodak, Barbara Johnstone, Paul Kerswill, ed., *The Sage Handbook of Sociolinguistics*, pages 330–344. Sage, Los Angeles, Calif.

Bruce I. Reiner. 2012. Using Analysis of Speech and Linguistics to Characterize Uncertainty in Radiology Reporting. In: *J Digit Imaging*, 25: 703–707. https://doi.org/10.1007/s10278-012-9535-x.

RNSA. 2017. RadLex Term Browser, available from http://radlex.org/ (cited 2020 Sept 7).

Simon Šuster, Stéphan Tulkens, and Walter Daelemans. 2017. A Short Review of Ethical Challenges in Clinical Natural Language Processing. In: *arXiv preprint* arXiv:1703.10090.

Charlotte Taylor and Anna Marchi, ed. 2018. *Corpus Approaches to Discourse*. Routledge, New York.

TEI Consortium. 2020. TEI P5: Guidelines for Electronic Text Encoding and Interchange. Release 4.0.0.

Viola Wiegand and Michaela Mahlberg. 2019. *Corpus Linguistics, Context and Culture*. De Gruyter, Berlin, Boston.