LeiKo: A corpus of easy-to-read German

Sarah Jablotschkin, Heike Zinsmeister

Universität Hamburg

sarah.jablotschkin@uni-hamburg.de , heike.zinsmeister@uni-hamburg.de

Donnerstag,
05.03.2020
10:30-11:15
ESA1 W Foyer

'Easy to read' is a general concept to describe variants of natural languages that are systematically reduced in vocabulary and syntax to be more easily accessible for adults with low reading competence. For German, there is 'Leichte Sprache', which adheres to specific linguistic and typographical rules, and the less restricted 'einfache Sprache'. Both variants have been the subject of growing attention in the academic and non-academic discourse - not least because in 2009 Germany ratified the UN Convention on the Rights of Persons with Disabilities (CRPD), which demands accessibility for all people not only to physical environments but also to information and media (cf. United Nations 2006, Art. 21). Implementing the Convention in German law resulted in rule sets for Leichte Sprache (BMJV 2011; BMAS 2014) followed by linguistic descriptions (Bredel & Maaß 2016).

The empirical validations of how much the simplified structures contribute to facilitated text comprehension are still in their beginnings. A particular desideratum is how well the rules for syntactic simplification allow to convey discourse structure-related information, e.g. the causal structure underlying (1) (cf. Jablotschkin 2017).

(1) Der Mensch bekommt Rente. Weil er krank ist.
'The man gets a pension. Because he is ill.' (Dt. Rentenversicherung Bund 2017)

To contribute to this kind of research, we introduce LeiKo, a comparable corpus of German easy-to-read news texts. This freely available resource is systematically compiled and linguistically annotated for linguistic and computational linguistic research.

LeiKo consists of 216 news and newspaper texts (approx. 50.000 tokens) and their meta data structured in four subcorpora according to the websites they were published on. All texts are tokenized, lemmatized, part-of-speech tagged and dependency parsed and can be queried in ANNIS (Krause/Zeldes 2016). A pilot corpus of 40 texts is manually corrected.

The poster will present challenges of the manual and automatic annotation. Due to the specific use of punctuation marks in easy-to-read German, one such challenge was to determine sentence borders. Segments that seem to be separate syntactic units from the punctuation point of view tend to be related by syntactic dependencies. Automatic annotation quality could be improved by training the sentence segmentation on easy-to-read data. The poster also will describe the corpus by statistics based on the subcorpora and comparable standard German data with a focus on characteristic syntactic constructions.

References: Bredel, U. & C. Maaß (2016). Leichte Sprache. Mannheim: Duden-Verlag. BMJV (2011): Verordnung zur Schaffung barrierefreier Informationstechnik nach dem Behindertengleichstellungsgesetz (Barrierefreie-Informationstechnik-Verordnung – BITV 2.0). Last modified: 25.11.2016. BMAS (2014) (ed.): Leichte Sprache. Ein Ratgeber. Deutsche Rentenversicherung Bund (ed.) (2017). Die Renten-Versicherung. Jablotschkin, S. (2017). Kausalrelationen in Leichter Sprache. Master's thesis, Universität Hamburg. Krause, T. & A. Zeldes (2016). ANNIS3: A new architecture for generic corpus query and visualization. Digital Scholarship in the Humanities 2016 (31). United Nations (2006): Convention on the Rights of Persons with Disabilities.