EmpiriST 2.0: Adding normalization, lemmatization and semantic tags to a German web and social media corpus

Thomas Proisl, Natalie Dykes, Philipp Heinrich, Besim Kabashi, Stefan Evert

FAU Erlangen-Nürnberg, Lehrstuhl für Korpus- und Computerlinguistik

{thomas.proisl, natalie.mary.dykes, philipp.heinrich, besim.kabashi, stefan.evert}@fau.de

The EmpiriST corpus is a manually annotated corpus consisting of almost 23,000 tokens of German web pages and German computer-mediated communication (CMC), i.e. written discourse. Examples for CMC genres are monologic and dialogic tweets, social and professional chats, threads from Wikipedia talk pages, WhatsApp interactions and blog comments.

The dataset was originally created for the EmpiriST 2015 shared task (Beißwenger et al., 2016) and featured manual tokenization and part-of-speech tagging. Subsequently, Rehbein et al. (2018) incorporated the dataset into their harmonised testsuite for POS tagging of German social media data (https://www.cl.uni-heidelberg.de/~rehbein/tweeDe.mhtml), manually added sentence boundaries and automatically mapped the part-of-speech tags to UD POS tags. In our own annotation efforts, we manually normalized and lemmatized the data and converted the corpus into a "vertical" format suitable for importing it into the Open Corpus Workbench, CQPweb, SketchEngine, or similar corpus tools.

During normalization, we corrected, for example, obvious spelling errors, e.g. *hinstelt*, and normalized non-standard variants to their canonical form, e.g. *uuuh* to *uh* or *hab* to *habe*. Then, we produced two lemma variants: Surface-oriented lemmata that are mainly based on the inflectional suffixes of the token and retain, as far as possible, any non-standard orthographical features of the token (the surface-oriented lemma for *hinstelt* would be *hinstelen*) and normalized lemmata that are based on the normalized token (e.g. *hinstellen*). The corpus was annotated by four student annotators, with agreement scores between 92.7 and 98.2 (Cohen's κ). The corpus and the annotation guidelines are available online under a Creative Commons license (https://github.com/fau-klue/empirist-corpus).

We will also report on our ongoing efforts to annotate the corpus with the semantic tagset used by the multilingual UCREL Semantic Analysis System (USAS; Piao et al., 2016) which consists of 232 category labels grouped into 21 major discourse fields (Archer et al., 2002).

References: Archer, D., Wilson, A., & Rayson, P. (2002). Introduction to the USAS category system. http://ucrel.lancs.ac.uk/usas/usas_guide.pdf • Beißwenger, M., Bartsch, S., Evert, S., & Würzner, K.-M. (2016). EmpiriST 2015: A shared task on the automatic linguistic annotation of computer-mediated communication and web corpora. Proceedings of the 10th web as corpus workshop (WAC-X) and the EmpiriST shared task, 44–56. • Piao, S., Rayson, P., Archer, D., Bianchi, F., Dayrell, C., El-Haj, M., Jiménez, R.-M., Knight, D., Křen, M., Löfberg, L., Nawab, R. M. A., Shafi, J., Teh, P. L., Mudraya, O. (2016). Lexical coverage evaluation of large-scale multilingual semantic lexicons for twelve languages. Proceedings of the tenth international conference on language resources and evaluation (LREC 2016), 2614–2619. • Rehbein, I., Ruppenhofer, J., & Zimmermann, V. (2018). A harmonised testsuite for pos tagging of German social media data. Proceedings of the 14th conference on natural language processing (KONVENS 2018), 18–28.