MultiSub: a multi-parallel corpus of movie subtitles

Fahime Same[1], Laura Becker[2], Alessia Cassarà[3]
[1,3]University of Cologne, [2]University of Erlangen-Nürnberg

f.same@uni-koeln.de, gombos.becker@fau.de, alessia.cassara@uni-koeln.de

While there are a number of open-access parallel corpora targeting the linguistic community, most of the resources either only allow pairwise comparisons across languages. This poster presents MultiSub, a new corpus based on movie subtitles that can be used for comparative linguistic studies.

The largest collection of freely available parallel corpora can be found on the OPUS website (http://opus.nlpl.eu/OpenSubtitles2018.php, cf. Tiedemann 2012). There are two main issues: (i) the parallel texts are only available in language pairs; (ii) in most parts, additional morphological and syntactic information are not (yet) included. A multi-parallel corpora, the ParTy corpus (Levshina 2016), includes subtitles from 14 movies. However, this corpus only contains the aligned, primary linguistic data in a downloadable text format. Building on this solid basis, we present the MultiSub corpus which targets crosslinguistic research especially in the domain of discourse analysis. We want to provide a text collection based on open-source subtitles that can be made available online for download in text and xml-based formats, and that comes with additional annotation layers which include linguistic (lemmatization, POS tagging, syntactic annotation) and extra-linguistic information (speaker, scene cuts). Given that (at least certain types of) movies are very close to naturalistic language use, movie subtitles are a rich resource for studying discourse. We focus on movie subtitles of a smaller number of languages for which well-developed POS taggers and syntactic parsers are available. We start with gathering movie and series subtitles from a number of European languages: English, German, French, Spanish, Russian. The preprocessing of the texts involves the following steps: text standardization, tokenization, xml-file generation. Alignment is based on time stamps from the srt-files. The additional layers added are: lemma and POS tags, dependency parsing, speaker information and scene cuts (added manually for English).

References: Tiedemann, Jörg. (2012): Parallel Data, Tools and Interfaces in OPUS. *Proceedings of the Eighth International Conference on Language Resources and Evaluation.* Istanbul: ELRA. • Levshina, Natalia. 2016. Verbs of Letting in Germanic and Romance: A Quantitative Investigation Based on a Parallel Corpus of Film Subtitles. *Languages in Contrast* 16 (1): 84–117.