In a nutshell

Currently available multiple parallel corpora

- there are a number of open-access parallel corpora targeting the linguistic community: Tiedemann (2012), Lison and Tiedemann (2016), Levshina (2016), and Cysouw and Wälchli (2007)
- many or offer only language pairs and/or do not feature additional linguistic or extra-linguistic annotation

MultiSub

- multiple parallel texts (subtitles of TV series) subtitles from http://www.addic7ed.com/
- linguistic & extra-linguistic annotation
- naturalistic data

Linguistic annotation (automatic)

Tool: UDPipe (Straka and Straková 2017)

Analysis: lemmatization, POS tagging, dependency parsing (s. Table 2)

Advantage: allows for direct, crosslinguistic structural comparison

Output formats

1. Multi-language parallel utterance-based segmentation

2. CoNLL-U format

3. xml format

Extra-linguistic annotation (manual)

Speaker information

- important for the study of turn taking, conversation structuring
- important for the comparison of speaker-anaphora vs. hearer-anaphora

Scene transitions

- a scene is defined as a setting without any noticeable break in the ongoing stream of time (in line with Hausler and Hanke 2016)
- scene-cuts are based on temporal progression
- important for the study of discourse, referential variation, etc.

Outlook & Applications

Next steps xml-based stand-off format, TEL-conformant, as well as reference annotation

Applications

- comparative discourse and syntactic analyses
- tab-delimited format: exploratory & qualitative studies, string queries
- CoNLL-U & xml format: comparative, quantitative studies including linguistic and extra-linguistic information

References