GRK 2839 Winter School: Corpus & Computational Linguistics

Linguistic annotation: tools and pipelines

Andreas Blombach, Philipp Heinrich
Lehrstuhl für Korpus- und Computerlinguistik
https://www.linguistik.phil.fau.de

CL FAU
Friedrich-Alexander-Universität
Philosophische Fakultät und
Fachbereich Theologie

Tools for manual annotation

- WebAnno / INCEpTION (focus on linguistics):
  - https://webanno.github.io/webanno/documentation/
  - https://www.youtube.com/user/webanno
  - https://inception-project.github.io
  - https://youtube.com/playlist?list=PL5Hz5pttaj96SlXHGRZf8KzlYvpVHIoL-
- prodigy (focus on linguistics):
  - https://prodi.gy
- CATMA (focus on literary science)
  - e.g. annotation of quoted and indirect speech
  - https://fortext.net/routinen/lerneinheiten/manuelle-annotation-mit-catma (in German)

Automatic annotation: complete pipelines (1)

- Stanford CoreNLP (https://stanfordnlp.github.io/CoreNLP/)
    - Long-running project, Java
    - Tokenisation, part-of-speech tagging, lemmatisation, named entity recognition, syntactic parsing, coreference resolution, sentiment analysis, …
- Stanza (https://stanfordnlp.github.io/stanza/)
    - Python, deep learning (+ interface to CoreNLP, e.g., for coreference resolution)
    - Tokenisation, POS tagging, lemmatisation, NER, dependency parsing, sentiment analysis
- spaCy – „fastest in the world“ (https://spacy.io)
    - Python, deep learning (transformer-based pipelines available)
    - Tokenisation, POS tagging, lemmatisation, NER, dependency parsing

Automatic annotation: complete pipelines (2)

NB: deep-learning-based tools generally require a decent GPU!

- Trankit (https://github.com/nlp-uoregon/trankit)
  - Python, deep learning (transformer-based)
  - Multilingual annotation possible
  - Tokenisation, POS tagging, lemmatisation, NER, dependency parsing

- Apache OpenNLP (https://opennlp.apache.org/)
  - Java
  - Tokenisation, POS tagging, lemmatisation, NER, dependency parsing, coreference resolution

- UDPipe (http://ufal.mff.cuni.cz/udpipe)
  - C++/Python, available as a library for multiple programming languages
  - Tokenisation, POS tagging, lemmatisation, dependency parsing

Automatic annotation: tokenisation and tagging

- Dedicated tokenisers
  - Python: SoMaJo (DE, EN)
  - generic tokeniser: Unitok
  - NLTK's tokeniser is mediocre at best
  - tokeniser must be compatible with POS tagger etc.!
- Part-of-speech taggers (often including their own tokeniser)
  - TreeTagger (fast, easy to use, support for many languages, incl. lemmatisation)
  - RNNTagger (deep learning successor of TreeTagger; Python; incl. lemm.)
  - SoMeWeTa (Python; DE, EN, FR)
  - Twitter data (EN): TweetNLP
  - … and many specialised tokenisers / taggers for other languages
- Select individual tools to create your own pipeline in your browser: https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page

Other NLP tools and resources

- Python is currently the language of choice for most NLP stuff
- Curated list: https://github.com/keon/awesome-nlp

- Topic modelling: gensim
- Word embeddings: fasttext
- Sentence embeddings: SBERT, SimCSE

- Transformer-based architectures and pre-trained language models: transformers (https://github.com/huggingface/transformers, https://huggingface.co/docs/transformers/index)