Create your own corpus – with **WebLicht**

Julia Müller & Christian Mair (University of Freiburg)

SPONSORED BY THE
Federal Ministry of Education and Research

Introduction

What is WebLicht?

…an environment for **automatically annotating and searching text corpora**

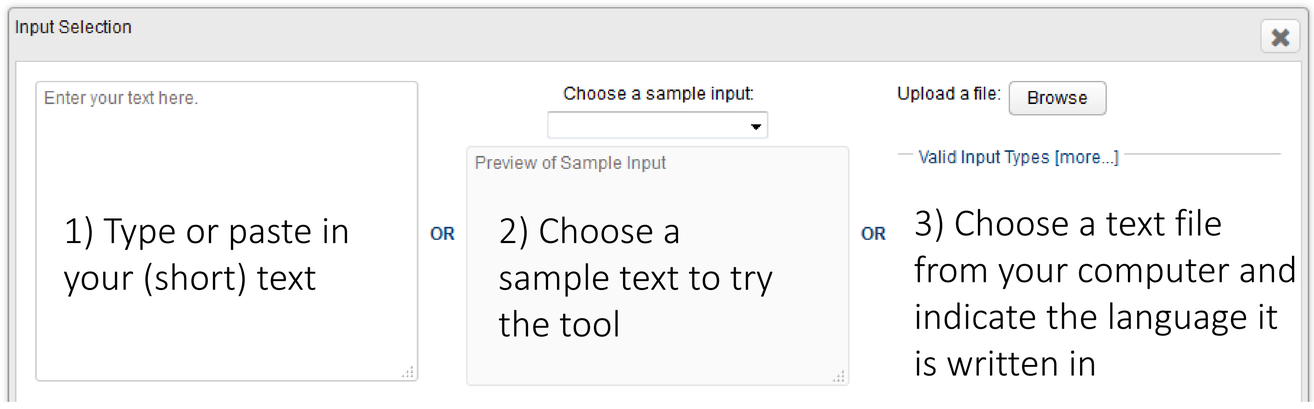
How does it work?

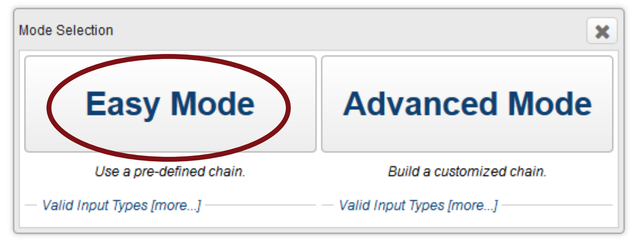
- Upload any text of your choice (format: txt)
- Choose the **annotation** you want
- **Search** the annotated corpus for words, phrases, and structures
- View and **visualize** syntactic structures

Where can I use it?

https://weblicht.sfs.uni-tuebingen.de/weblicht/
Log in with Shibboleth

Uploading a Text





WebLicht offers functional tools for English, German, French, and Italian.

Annotation Types

Pos Tags/Lemmas

- each word is associated with a:
  - Pos (= part of speech, word-class)
  - lemma (= base form)
- tagset: Penn Treebank
- search attributes: pos, lemma

Morphology

- information about syntax, tense, mood, case, person, number, degree, and negation
- tagset: NUPOS MorphAdorner
- search attributes: morph*

Constituent parsing

- divides the text into sentences, clauses, phrases, and words
- structure of each sentence is represented visually as a tree
- tagset: Penn Treebank
- search attribute: cat

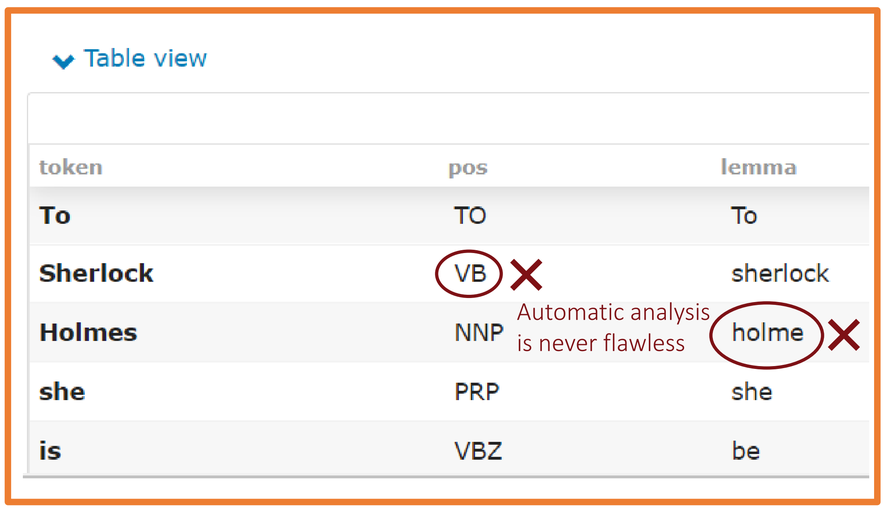
Dependency parsing

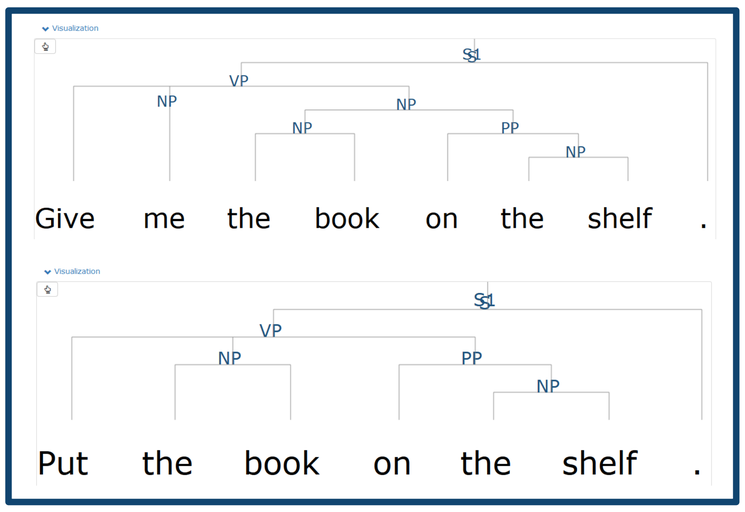
- connects words based on how they modify each other
- recognises relationships such as subject - object
- search attribute: edge

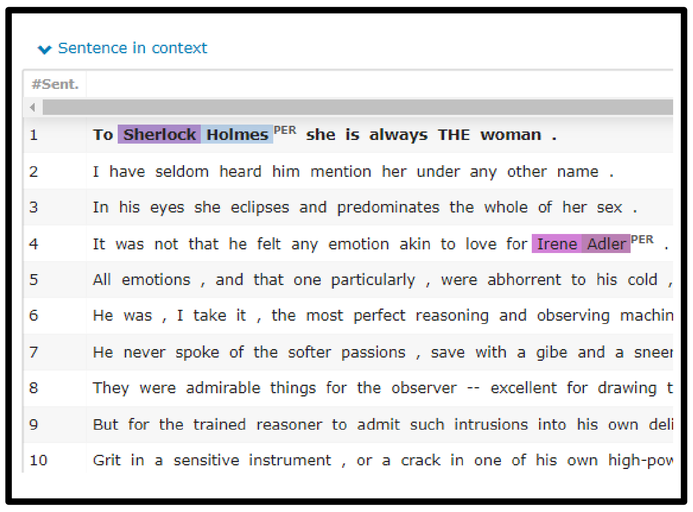
Named Entities

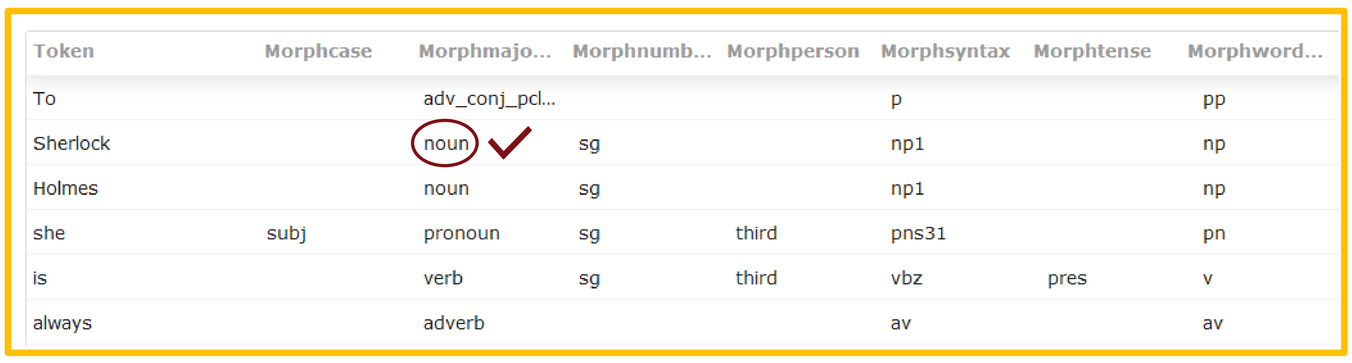
- classifies proper nouns as:
  - people
  - organizations
  - locations
  - miscellaneous
- tagset: Illinois Named Entity

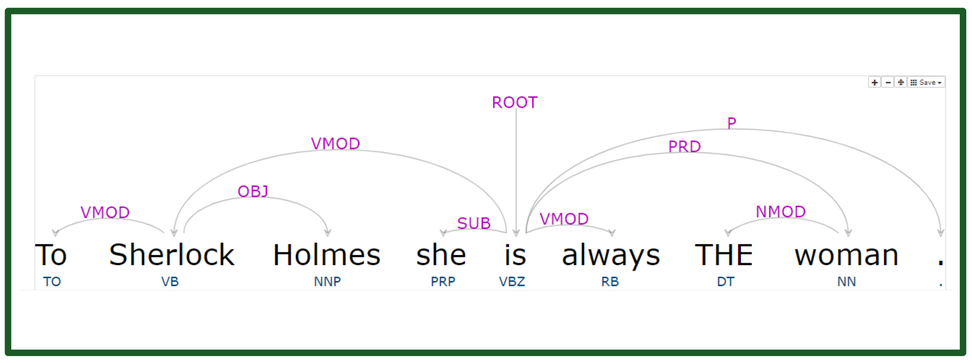
Visualisations of the Annotated Corpus











Searching the Annotated Corpus

Basics

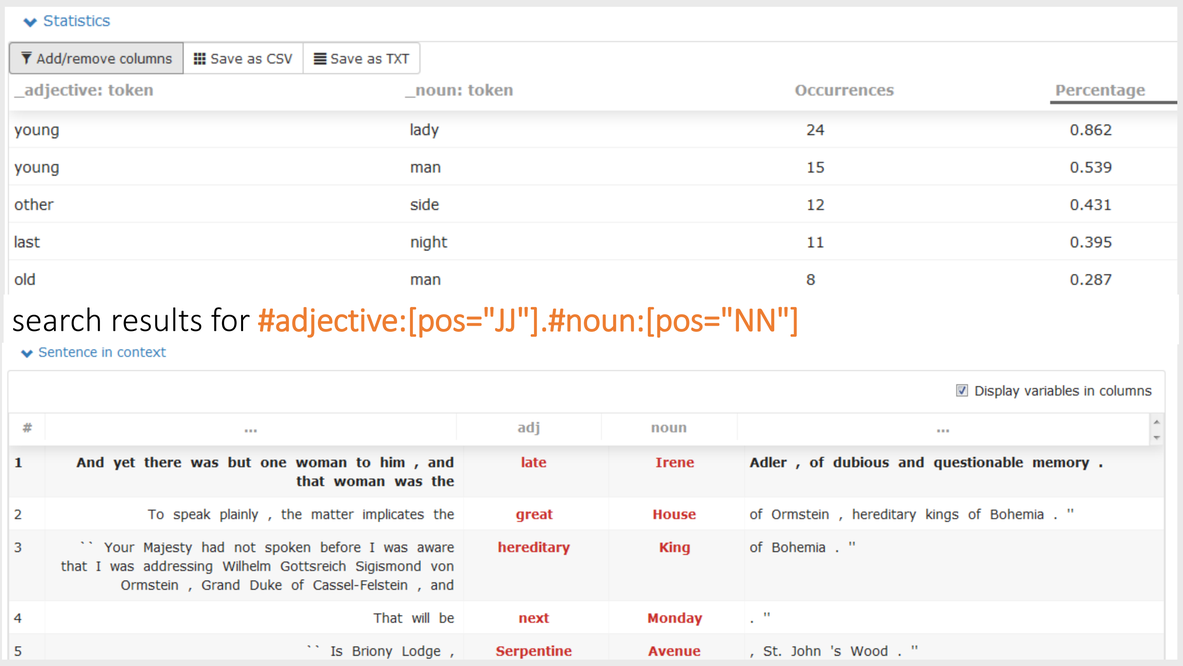
- search with this model: [attribute="value"]
- search can be named (e.g. #query1:)
- possible in all annotations:
  #q1:[word="going"]
  → instances of 'going'
  #q2:[token="has"]
  → instances of 'has'
- Regular expressions in forward slashes, e.g. [word=/lo*l*/]

Combining criteria

Combine criteria using:
& 'and' | 'or' ! 'not'
[morphwordclass="vm"&word!="will"]
→ all modals, excluding 'will'

Sequences

full stop between queries to search for sequences
[lemma="student"].#q2:[pos="VBP"|pos="VBZ"]
→ all forms of 'student' followed by singular or plural verbs



For questions, comments, or more detailed information on how to use WebLicht, please contact julia.mueller@anglistik.uni-freiburg.de