

Create your own corpus – with WebLicht



SPONSORED BY THE



Federal Ministry of Education and Research

Julia Müller & Christian Mair (University of Freiburg)

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://weblight.sfs.uni-tuebingen.de/weblight/>
Log in with Shibboleth



UPLOADING A TEXT

Input Selection

Enter your text here.

1) Type or paste in your (short) text

OR

Choose a sample input:

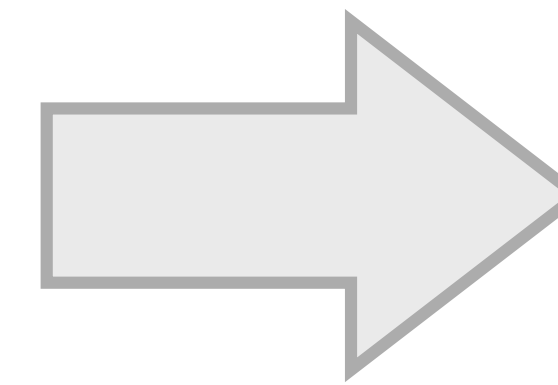
Preview of Sample Input

2) Choose a sample text to try the tool

OR

Upload a file:

3) Choose a text file from your computer and indicate the language it is written in



Mode Selection

Easy Mode (circled in red)

Use a pre-defined chain.

Advanced Mode

Build a customized chain.

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

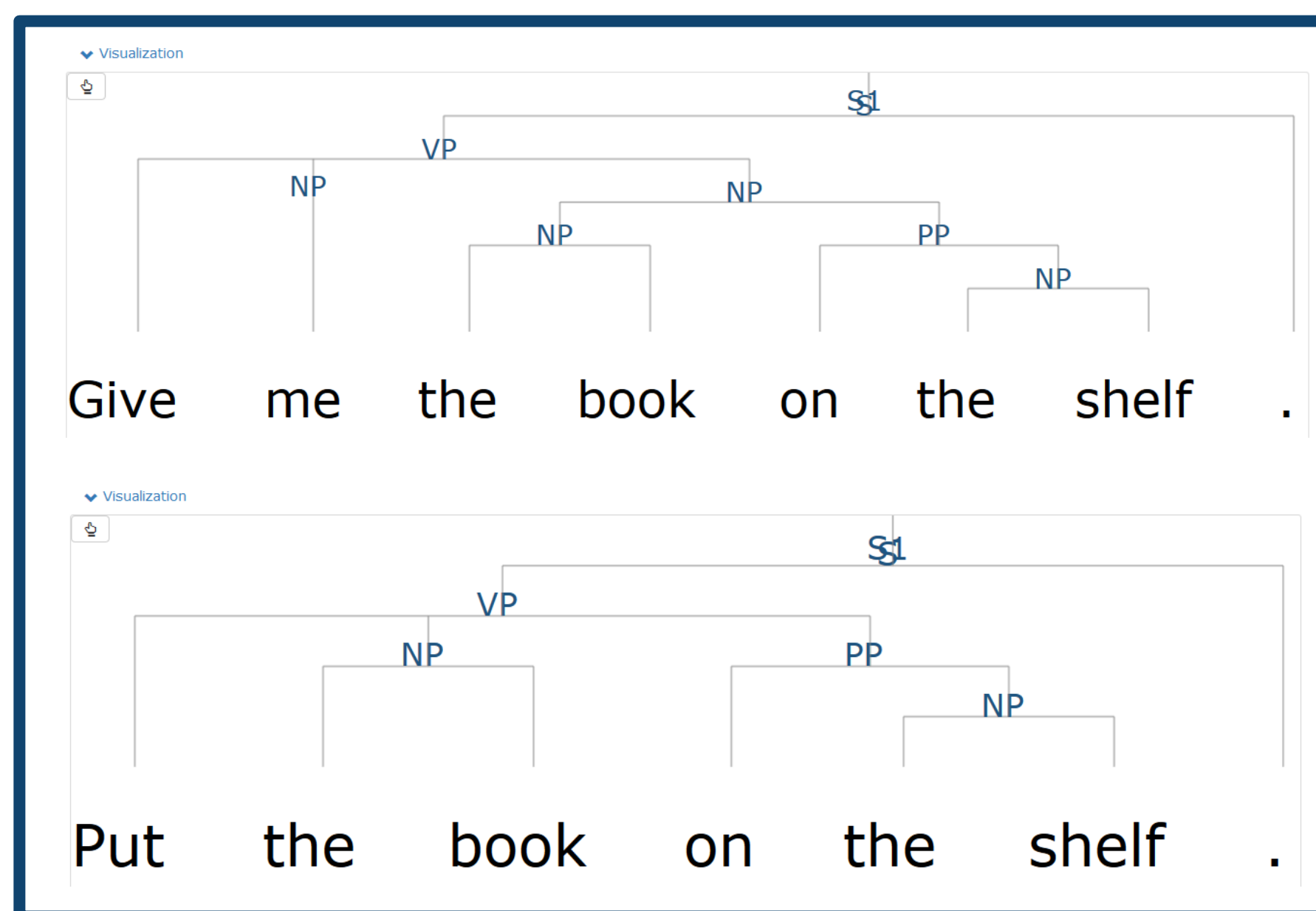
- classifies proper nouns as:
 - people
 - organizations
 - locations
 - miscellaneous
- tagset: Illinois Named Entity

VISUALISATIONS OF THE ANNOTATED CORPUS

Table view

token	pos	lemma
To	TO	To
Sherlock	VB ✗	sherlock
Holmes	NNP	holme ✗
she	PRP	she
is	VBZ	be

Automatic analysis is never flawless

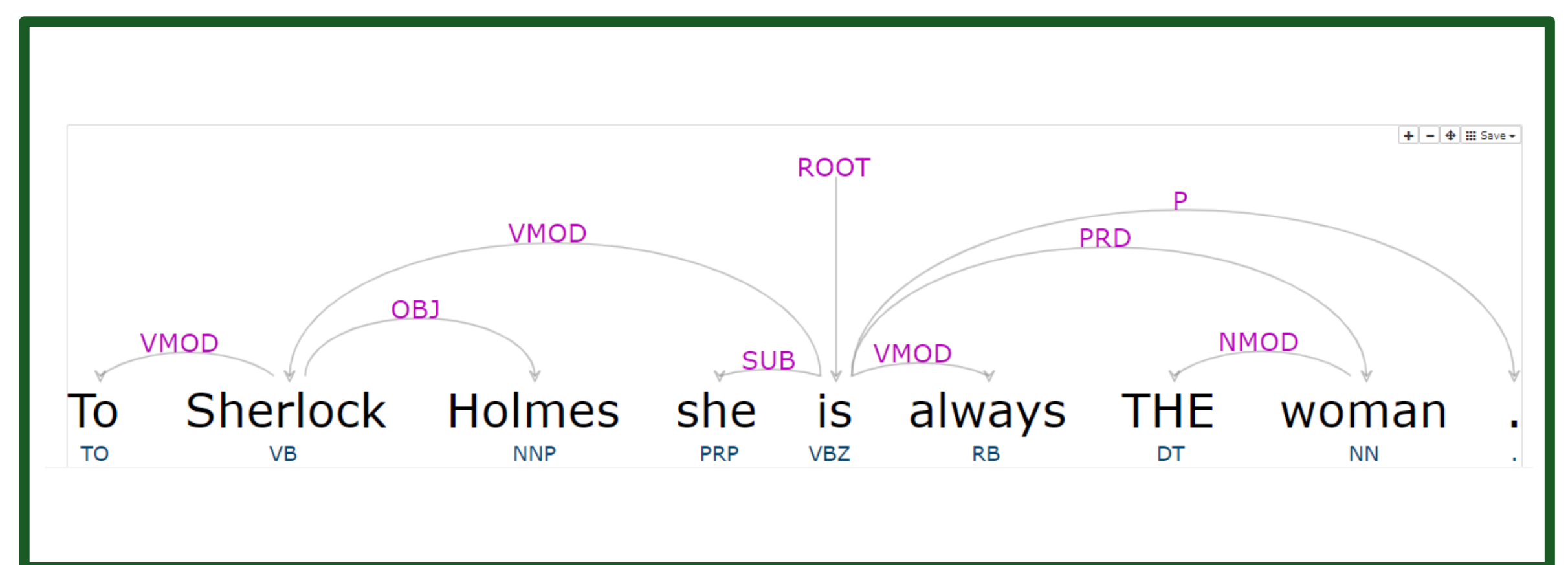


Sentence in context

#Sent.

1	To Sherlock Holmes ^{PER} she is always THE woman .
2	I have seldom heard him mention her under any other name .
3	In his eyes she eclipses and predominates the whole of her sex .
4	It was not that he felt any emotion akin to love for Irene Adler ^{PER} .
5	All emotions , and that one particularly , were abhorrent to his cold ,
6	He was , I take it , the most perfect reasoning and observing machin
7	He never spoke of the softer passions , save with a gibe and a sneer
8	They were admirable things for the observer -- excellent for drawing t
9	But for the trained reasoner to admit such intrusions into his own deli
10	Grit in a sensitive instrument , or a crack in one of his own high-pov

Token	Morphcase	Morphmajo...	Morphnumb...	Morphperson	Morphsyntax	Morph tense	Morphword...
To		adv_conj_pcl...			p		pp
Sherlock	noun	sg	np1				np
Holmes	noun	sg	np1				np
she	subj	pronoun	sg	third	pns31		pn
is	verb	sg	third	vbz	pres		v
always	adverb				av		av



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*!*/]**

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

Statistics

▼ Add/remove columns Save as CSV Save as TXT

_adjective: token	_noun: token	Occurrences	Percentage
young	lady	24	0.862
young	man	15	0.539
other	side	12	0.431
last	night	11	0.395
old	man	8	0.287

search results for **#adjective:[pos="JJ"].#noun:[pos="NN"]**

Sentence in context

Display variables in columns

#	...	adj	noun	...
1	And yet there was but one woman to him , and that woman was the	late	Irene Adler , of dubious and questionable memory .	
2	To speak plainly , the matter implicates the	great	House of Ormstein , hereditary kings of Bohemia . "	
3	" Your Majesty had not spoken before I was aware that I was addressing Wilhelm Gottsreich Sigismund von Ormstein , Grand Duke of Cassel-Felstein , and	hereditary	King of Bohemia . "	
4	That will be	next	Monday . "	
5	" Is Briony Lodge ,	Serpentine	Avenue , St. John 's Wood . "	

