

# Computerlinguistik / Natural Language Processing

---

Vorlesung *Einführung in die Digital Humanities*  
MSc Digital Humanities | Wintersemester 2018/19

Prof. Dr. Christof Schöch

---

# Ankündigungen

- Vortrag: Christian Reul (Würzburg): "OCR4all - Ein vollständiger OCR Workflow gekapselt in einem Open Source Tool" (Do, 16-18, DM 131)

# Semesterüberblick

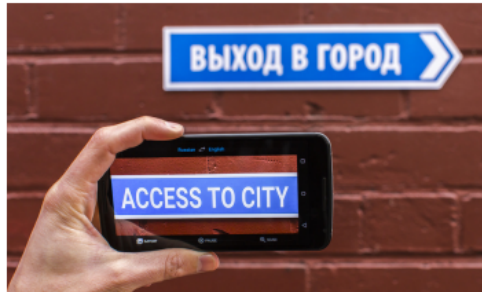
- 23.10.: Digital Humanities im Überblick
- 30.10.: Digitalisierung: Text und Bild
- 06.11.: Grundbegriffe des Programmierens
- 13.11.: Datenmodellierung 1: Modellierung
- 20.11.: Datenmodellierung 2: Datenbanken
- 27.11.: Datenmodellierung 3: Text, Markup, XML
- 04.12.: Digitale Edition
- 11.12.: Geschichte der Digital Humanities
- 18.12.: Informationsvisualisierung
- 22.12.-2.1.: *Weihnachtspause*
- **08.01.: Computerlinguistik / Natural Language Processing**
- 15.01.: Quantitative Analyse 1: Stilometrie, Topic Modeling
- 18.01.: Quantitative Analyse 2: Superv. Machine Learning
- 29.01.: Open Humanities
- 05.02.: Klausurtermin

# Sitzungsüberblick

1. Einstieg: L, CL, NLP
2. Aufbau von Text-/Sprachkorpora
3. Einzelne Annotationstechniken
4. Regeln und Wahrscheinlichkeiten
5. Beispiel: WebLicht

# 1. Einstieg: L, CL, NLP

# NLP und CL im Alltag



In der Computerlinguistik oder linguistischen Datenverarbeitung wird untersucht, wie natürliche Sprache in Form von Text- oder Sprachdaten mit Hilfe des Computers algorithmisch verarbeitet werden kann. Sie ist Schnittstelle zwischen Sprachwissenschaft und Informatik.

Chinese (Traditional)



我想訂一盤章魚湯。

Wǒ xiǎng dìng yī pán zhāngyú tāng.

English



I want to order a plate of octopus soup.

was ist computerlinguistik

was ist computerlinguistik

computerlinguistik was ist **das**

was **bedeutet** computerlinguistik

Google Search

I'm Feeling Lucky

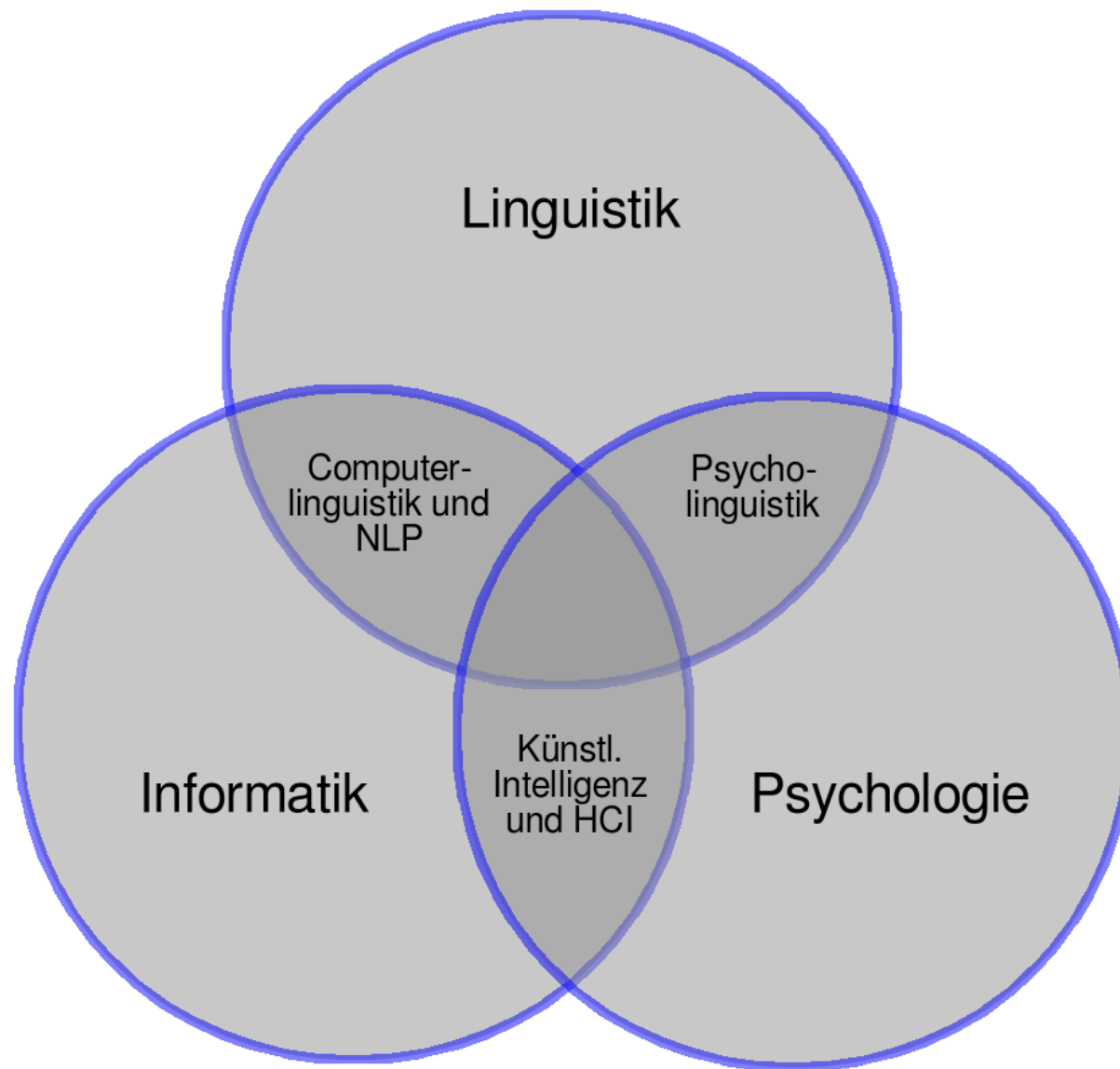
[Learn more](#)

[Report inappropriate predictions](#)



# Unterscheidungen

- Linguistik
- Computerlinguistik
- Natural Language Processing





# Linguistik

- Linguistik ist die wissenschaftliche, theoriegeleitete, synchrone und diachrone Beschreibung von Sprache
- Sprache als System mit unterschiedlichen Ebenen:
  - Phonetik / Phonologie
  - Morphologie
  - Syntax
  - Semantik
  - Pragmatik

# Computerlinguistik

- Teilgebiet der Linguistik an der Schnittstelle von Informatik und Linguistik
- Ziel: Beschreibung des Sprachsystems mit computergestützten Methoden
- Typische Arbeitsbereiche:
  - Empirische Überprüfung linguistischer Theorien
  - Automatische Annotation von Sprache auf verschiedenen Ebenen
  - Statistische Auswertung der Annotationen

# Natural Language Processing

- Teilgebiet der Informatik an der Schnittstelle von Informatik und Linguistik
- Enge Verbindungen zur Künstlichen Intelligenz und Kognitiven Psychologie
- Ziel: Anwendungsbezogener Einsatz von Techniken aus Informatik und Computerlinguistik
- Typische Arbeitsbereiche
  - Sprachverstehen (bspw. Diktiersoftware)
  - Sprachproduktion (bspw. Navigationssoftware)
  - Machine Translation (bspw. Google Translate, TM)
  - Sentiment Analyse (bspw. Marktforschung)

# Einige Aufgaben von CL und NLP

- Aufbau und Verwaltung von Korpora sprachlicher Daten
- Entwicklung von Methoden zur Modellierung / Operationalisierung sprachlicher Phänomene
- Bereitstellung von Wissen über Aspekte individueller Sprachen
- Entwicklung von Algorithmen und Methoden zur Bearbeitung von sprachlichen Äußerungen
- Entwicklung nützlicher, sprachbasierter Anwendungen
- Konzeption effektiver Evaluationsmechanismen

# State of the Art (Jurafsky)

Dan Jurafsky



## Language Technology

making good progress

mostly solved

### Spam detection

Let's go to Agra! ✓

Buy V1AGRA ... ✗

### Part-of-speech (POS) tagging

ADJ ADJ NOUN VERB ADV

Colorless green ideas sleep furiously.

### Named entity recognition (NER)

PERSON ORG LOC

Einstein met with UN officials in Princeton

### Sentiment analysis

Best roast chicken in San Francisco! 👍

The waiter ignored us for 20 minutes. 👎

### Coreference resolution

Carter told Mubarak he shouldn't run again.

### Word sense disambiguation (WSD)

I need new batteries for my *mouse*.

### Parsing

I can see Alcatraz from the window!

### Machine translation (MT)

第13届上海国际电影节开幕...

The 13<sup>th</sup> Shanghai International Film Festival...

### Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party  
May 27  
add

still really hard

### Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

### Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

### Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose

Economy is good

### Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

(Quelle: Dan Jurafsky, "Introduction to NLP", <https://slideplayer.com/slide/4578600/>)

## 2. Aufbau von Text-/Sprachkorpora

# Begrifflichkeiten

- Korpus, engl. corpus
- Textedition, engl. (scholarly) text edition
- Textsammlung, engl. text collection

# Korpus

- Begriff aus dem linguistischen Kontext
- kann mehr oder weniger umfangreich sein
- kann Daten in Text- oder Audioform beinhalten
- häufig: für eine bestimmte Domäne repräsentativ
- häufig: stratifiziert nach Unterdomänen; ausgeglichene Anteile; Textsamples
- häufig: mit linguistischen Annotationen versehen  
Lemma, Wortart
- in diversen Formaten: CSV, XML, XMI (standoff), TCF



# Beispiele für Korpora

- **British National Corpus** (BNC, 100 Millionen Tokens, <http://www.natcorp.ox.ac.uk/>)
- **Deutsches Referenzkorpus** (DeReKo, >42 Milliarden Tokens, <http://www1.ids-mannheim.de/kl/projekte/korpora/>)
- **Deutsches Textarchiv** (DTA, 3800 Texte, <http://www.deutschestextarchiv.de/>)
- **LAUDATIO-Korpora** (historische Korpora, <http://www.laudatio-repository.org/repository/>)

# Wiss. Textedition

- Begriff aus den Editionswissenschaften
- häufig: weniger umfangreich als Korpora
- häufig: autorzentrierte Gegenstandsdefinition
- meist sehr hohe Ansprüche an Textqualität und Transparenz der Texterstellung
- Fokus auf der (editorischen) Erschließung
- in der Regel in XML-TEI (mit Derivaten)
- Beispiele: siehe Sitzung 7

# Textsammlung

- Weniger etablierter Begriff aus den DH
- häufig: wesentlich umfangreicher als Editionen
- häufig: weiter Definition des Gegenstandsbereichs nach Sprache, Gattung, Epoche
- Formate: häufig TXT, HTML oder XML-TEI
- meistens ohne linguistische Annotation
- geringere Anforderungen an Textqualität und Transparenz

# Beispiele:

- **TextGrid's Digitale Bibliothek**, 600 Autorenwerke, <https://textgridrep.org/>
- **Théâtre classique**, 1100 Theaterstücke, <http://theatre-classique.fr/>
- **Litteraturbanken**, 2000 Texte, <http://litteraturbanken.se>
- **Papyri.info**, 50.000 Dokumente, <http://papyri.info>

# RIDE

- "A review journal for digital editions and resources"
- Reviews von Texteditionen und Textsammlungen
- Ulrike Henny und Frederike Neuber: "Criteria for Reviewing Digital Text Collections, version 1.0", 2017
- <http://ride.i-d-e.de/>

# Aspekte des Aufbaus von Textsammlungen / Korpora

- Repräsentativität (Verhältnis zur Grundgesamtheit)
- Größe (in Texten, in Tokens, in Zeichen)
- Annotationen (linguistisch, strukturell, Metadaten)
- Textselektion (Zeit, Sprache, Register, Gattungen, Autoren, etc.)
- Textform (vollständig oder Samples)

# Fokus: Repräsentativität

- Repräsentativität = auf der Grundlage der Textsammlung können verallgemeinerte Aussagen über die tatsächlichen Verhältnisse gemacht werden
- sehr hohe Anforderungen
  - Grundgesamtheit muss bekannt sein
  - es muss ein zufälliges Sample aus der Grundgesamtheit erstellt worden sein
  - reiner Umfang ist weniger entscheidend

# Alternativen

- Balancierte Sammlung (verschiedene relevante Textgruppen sind in vergleichbarem Umfang vorhanden)
- Opportunistische Sammlung (rein nach praktischer Verfügbarkeit)



# 3. Einzelne Annotationstechniken

# Annotationstechniken

- Tokenisierung
- Lemmatisierung
- POS-Tagging
- Morphologisches Tagging
- Syntaktisches Parsing
- Named Entity Recognition
- Sentiment Analysis

# Tokenisierung

- Definition: Ermittlung der Grenzen zwischen Wörtern
- Beispiel:
  - Satz: "Möge die Macht mit dir sein."
  - Tokenisiert: "Möge, die, Macht, mit, dir, sein, ."

# Herausforderungen

- Bindestriche
  - Mehrweg-Pfandflasche
  - "live off-campus"
- Mehrteilige Wörter
  - Süddeutsche Zeitung
  - New York City
- Apostrophen
  - "don't"
  - "Ella, elle l'a"

# Lemmatisierung

- Definition: Zurückführung auf die linguistische Grundform
- Beispiel:
  - Satz: “Wir wären gerne länger geblieben.”
  - Lemmatisiert: “wir, sein, gerne, lang, bleiben, .”

# Herausforderungen

- Fehler bei der Tokenisierung
  - "There, are, no, off, campus, bar, in, New, York, City"
  - "off-campus", "New York City"
- Zweideutige Tokens
  - "Die Buche steht im Wald", Lemma = "Buche"
  - "Buche bitte den Urlaub in Wallonien", Lemma = buchen

# Part-of-Speech Tagging

- Definition: Zuordnung, zu jedem Token, einer grammatischen Kategorien
- Grundlage: eine Ontologie linguistischer Klassen (=Tagset)
- Beispiel:
  - Satz: “Die Tage sind kurz.”
  - Getaggt: “Die\_ART Tage\_SUBST sind\_VERB kurz\_ADJ .\_PUNC”

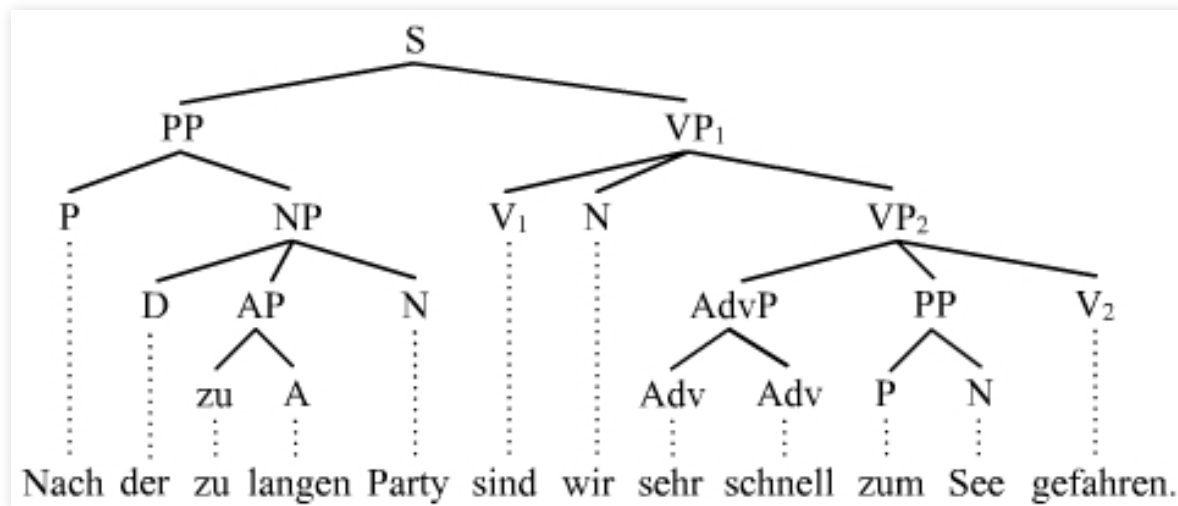
# Herausforderung

- Fehler bei der Tokenisierung
  - "There are no off-campus bars in New York City"
  - off\_PRP, campus\_N vs. off-campus\_ADJ vs.
  - New\_ADJ, York\_NP, City\_NC vs. New York City\_NP
- Ambiguitäten
  - "Klaus hat lange Nudeln gegessen"
  - a) lange\_ADJ ("lange Nudeln")
  - b) lange\_ADV ("lange ... gegessen")



# Syntaktisches Parsing

- Definition: Ermittlung der hierarchischen, grammatikalischen Struktur eines Satzes.
- Beispiel: "Nach der zu langen Party sind wir schnell zum See gefahren."



# Herausforderungen

- Syntaktische Ambiguität (siehe POS)
  - “Klaus hat lange Nudeln gegessen”
  - a) NP(ADJ+N) (“lange Nudeln”)
  - b) VP(ADV+V) (“lange ... gegessen”)

# Named Entity Recognition

- Definition: Identifikation und Klassifikation von Tokens, die sich auf Personen, Orte, Daten, Organisationen beziehen
- Beispiel: Zeitungstexte
  - Satz: “Elon Musk sagte in Paris, er wolle in der Stadt des Lichts einen großen Tesla Showroom eröffnen.”
  - Mit NER: “[Elon Musk]\_PERS sagte in [Paris]\_PLACE, er wolle in der Stadt des Lichts einen großen [Tesla]\_ORG Showroom eröffnen.

# NER mit "Coreference Resolution"

- Definition: Identifikation der Referenten von anaphorischen Ausdrücken
- Beispiel: Zeitungstexte
  - Satz: "Elon Musk sagte in Paris, er wolle in der Stadt des Lichts einen großen Tesla Showroom eröffnen."
  - NER mit CR: "[Elon Musk]\_1 sagte in [Paris]\_2, [er]\_1 wolle in der [Stadt des Lichts]\_2 einen großen [Tesla]\_3 Showroom eröffnen."

# Sentiment Analysis

- Definition: Einen Satz (oder eine Phrase) nach ihrer Wertung (positiv, negativ, neutral) klassifizieren
- Beispiel Produktbewertungen
  - “Die Kamera macht richtig tolle Bilder”, 0.8
  - “Der Zoombereich ist nicht nützlich”, -0.4
  - “Der Autofokus funktioniert meistens nicht”, -0.1

# Herausforderungen

- Negation:
  - "Ich finde, dass die Kamera tolle Bilder macht!"
  - vs. „Ich finde nicht, dass die Kamera tolle Bilder macht!“
  - vs. "Ich sage ja nicht, dass die Kamera nicht tolle Bilder macht!"
- Ironie, Sarkasmus, Implizites
  - "Die Kamera ist sicher für Katzenbilder gut geeignet."

# 4. Verschiedene Annotationsstrategien

# Beispiel: POS-Tagging

- "Der Baum ist grün"
- Der\_DET Baum\_NC ist\_V grün\_ADJ
  
- "Klaus isst lange Nudeln"
- Klaus\_NP isst\_V lange\_ADJ|ADV Nudeln\_NC



# Verschiedene Annotationsstrategien

- Prinzipien
  - wörterbuchbasiert
  - regelbasiert
  - Machine Learning
- Eigenschaften
  - Kontextfrei oder sequenzinformation
  - deterministisch vs. probabilistisch

# wörterbuchbasiert: Prinzip

- Wörterliste mit der Zuordnung von POS ("Vollformenlexikon")
- Algorithmus "schlägt nach"
- Einfach, kontextfrei, deterministisch

# wörterbuchbasiert: Grundlage

- Der = DET
- Baum = NP
- ist = V
- grün = ADJ
- isst = V
- lange = ADJ
- Nudeln = NP
- Klaus = ?

## wörterbuchbasiert: Ergebnis

- Der\_DET Baum\_NC ist\_V grün\_ADJ ✓
- Klaus\_? isst\_V lange\_ADJ Nudeln\_NC ✗

# wörterbuchbasiert: Varianten

- a) nur mit einem (dem „normalen“) POS: lange\_ADJ
- b) mit allen möglichen POS: lange\_ADJ|ADV
- c) mit allen möglichen POS und  
Auftrittswahrscheinlichkeit: lange\_ADJ-80%|ADV-20%

# regelbasiert: Prinzip

- Es werden deterministische Regeln definiert
- der Algorithmus prüft und wendet die Regeln an
- (fast) ohne Wörterbuch können so Zuordnungen vorgenommen werden
- Regeln können kontextfrei sein oder Sequenzinformation nutzen;
- Regeln können mit Wahrscheinlichkeiten ausgestattet sein

# regelbasiert: Grundlage

- Wenn das Wort "(D|d)(er|ie|as)" lautet, ist es ein DET
- Wenn das Wort mit Großbuchstaben anfängt, oder wenn vor dem Wort ein DET steht, ist es ein N
- Wenn vor dem Wort ein N steht, ist es ein V
- Wenn vor dem Wort ein V steht, ist es ein ADJ

# regelbasiert: Ergebnis

1. Der Baum ist grün
2. Der\_DET Baum ist grün
3. Der\_DET Baum\_N ist\_V grün
4. Der\_DET Baum\_N ist\_V grün\_ADJ



# regelbasiert: Varianten

- Alternativen, bspw.: "wenn auf das Wort ein N folgt, ist es ein DET oder ein ADJ" (der Baum, grüner Baum)
- Alternativen mit Wahrscheinlichkeiten: "wenn auf das Wort ein N folgt, ist es mit 80% W. ein DET, mit 20% ein ADJ"
- Regeln mit Wahrscheinlichkeiten: "wenn das Wort mit Großbuchstaben anfängt, ist es mit 60% Wahrscheinlichkeit ein N"; wenn zudem ein DET davor steht, ist es sogar mit 80% Wahrscheinlichkeit ein N"

# Machine Learning: Prinzip

- Weder Wörterbücher noch deterministische Regeln
- Lernalgorithmus bekommt viele Beispiele präsentiert
- Bekommt Informationen über jedes Wort und korrektes Label
- Kontextfreie und Sequenzinformationen
- Algorithmus erlernt selbst Regeln, mit Alternativen und Wahrscheinlichkeiten
- Diese (oft hochkomplexen Regeln) wendet er dann auf neue Texte an
- probabilistisch; eher intransparent

# Beispiel: Informationen

- "Der Baum ist grün."
  - Wortlänge: 4 Buchstaben
  - Erster Buchstabe: "b", Majuskel
  - Zweiter Buchstabe: "a"
  - Dritter Buchstabe: "u"
  - Vierter Buchstabe: "m"
  - Wort davor: "der"
  - Wort danach: "ist"
  - Wort davor: 80% DET
  - etc.

# 5. Beispiel: WebLicht

# WebLicht (CLARIN-D)

- Ein Webservice für linguistische Annotationen
- Bereitgestellt von CLARIN-D (Common Language Resources and Technology Infrastructure)
- <https://weblicht.sfs.uni-tuebingen.de/weblicht/>

# Beispieltext

*"Der Zauberberg ist ein 1924 erschienener Bildungsroman von Thomas Mann. Der berühmte Roman spielt in Davos."*

Main Page

Chain 1 ✕

+ New Chain

# WebLicht



## Welcome to WebLicht

WebLicht consists of a collection of web-based linguistic annotation tools, distributed repositories for storing and retrieving information about the tools, and this web application, which allows you to easily create and execute tool chains without downloading or installing any software on your local computer.

This application and its associated tools are continually being updated and improved.

For more information, visit our websites at [WebLicht](#), [CLARIN-D](#), and [CLARIN](#).

## What's New

There are 2 modes for building tool chains:

- Easy Mode lets you choose pre-defined processing chains
- Advanced Mode allows you to build customized tool chains.

In this version, the input selection/upload process was made more intuitive.

## Getting Started

Click on the "+ New Chain" tab at the top left of this page or click on the "Start" button below:

**Start**

[CLARIN-D](#)

<https://weblicht.sfs.uni-tuebingen.de/WebLicht-4/>
Search
View Tool List
HELPDESK

Main Page
Chain 3 x

### Available Annotations for: German Plain Text

- ☐ POS Tags/Lemmas
- ☐ Morphology
- ☒ Constituent Parses
- ☐ Dependency Parses
- ☐ Named Entities

#### Berkeley Parser - Berkeley NLP POS

Annotation Layers:

- Simple view
  - ☐ text
  - ☐ sentences
  - ☒ Table view
    - ☒ tokens
    - ☒ POSTags
  - Graphical view
    - ☐ parsing

language = de

Download as Excel sheet   Download as CSV

[ 1 - 30 / 52 ]

token ID	tokens	POSTags
t1	The	NE
t2	Wire	NE
t3	ist	VAFIN
t4	eine	ART
t5	US-amerikanische	ADJA
t6	Fernsehserie	NN
t7	,	\$,
t8	die	PRELS
t9	von	APPR
t10	2002	CARD
t11	bis	APPR

Download TCF

#### Input and Chain Selection

Run Tools   Clear Results   Download chain

My Input
Plain Text
The Wire ist eine US-amerikanische Fernsehserie, die von 2002 bis 2008 in Baltimore (Maryland) gedreht wurde. Autor ist der ehemalige

Sfs To TCF Converter
Document Type
Language: German
Document Type: TCF
TCF Version: 0.4
Text

IMS Tokenizer
Sentences
Tokens

Sfs Berkeley Parser - Berkele
Part of Speech: STTS Tagset
Parsing: tuebadzib

Done running tools.



<https://weblicht.sfs.uni-tuebingen.de/WebLicht-4/>

[View Tool List](#)
[HELPDESK](#)

[Main Page](#)
[Chain 3 x](#)

### Available Annotations for: German Plain Text

- ☐ POS Tags/Lemmas
- ☐ Morphology
- ☐ Constituent Parses
- ☐ Dependency Parses
- ☒ Named Entities

#### German Named Entity Recognizer

**Annotation Layers:**

- Simple view
  - ☐ text
  - ☐ sentences
- Table view
  - ☐ tokens
  - ☐ POSTags
  - ☐ lemmas
  - ☐ namedEntities
- Highlighted view
  - ☒ namedEntities

tagset = tuebadz8 ☒ Show all sentences

☒ GPE

☒ PER

☒ ORG

The Wire ist eine US-amerikanische Fernsehserie , die von 2002 bis 2008 in **Baltimore** ( **Maryland** ) gedreht wurde .

Autor ist der ehemalige Polizeireporter **David Simon** , der schon die Vorlage zur Krimiserie **Homicide** schrieb .

Gegenstand der Serie ist die Realität der postindustriellen amerikanischen Stadt und des amerikanischen Gemeinwesens .

[Download TCF](#)

#### Input and Chain Selection

[Run Tools](#)
[Clear Results](#)
[Download chain](#)

**My Input** Plain Text
 

The Wire ist eine US-amerikanische Fernsehserie, die von 2002 bis 2008 in Baltimore (Maryland) gedreht wurde. Autor ist der ehemalige

**Sfs To TCF Converter**

Document Type

Language: German  
Document Type: TCF  
TCF Version: 0.4  
Text

**IMS Tokenizer**

Sentences  
Tokens

**IMS TreeTagger**

Part of Speech: STTS Tagset  
Lemmas

**Sfs German Named Entity R**

Named Entities: tuebadz8

| Done running tools.

<https://weblicht.sfs.uni-tuebingen.de/WebLicht-4/>
Search
View Tool List
HELPDESK

Main Page
Chain 3 x +

### Available Annotations for: German Plain Text

- ☐ POS Tags/Lemmas
- ☐ Morphology
- ☒ Constituent Parses
- ☐ Dependency Parses
- ☐ Named Entities

#### Berkeley Parser - Berkeley NLP POS

tagset = tuebadztb

1 / 3

Annotation Layers:

- Simple view
  - ☐ text
  - ☐ sentences
  - ☐ Table view
    - ☐ tokens
    - ☐ POSTags
  - Graphical view
    - ☒ parsing

Download TCF

#### Input and Chain Selection

My Input
Plain Text
The Wire ist eine US-amerikanische Fernsehserie, die von 2002 bis 2008 in Baltimore (Maryland) gedreht wurde. Autor ist der ehemalige

Sfs
To TCF Converter
Document Type
Language: German
Document Type: TCF
TCF Version: 0.4
Text

IMS
Tokenizer
Sentences
Tokens

Sfs
Berkeley Parser - Berke
Part of Speech: STTS Tagset
Parsino: tuebadztb

Run Tools
Clear Results
Download chain

| Done running tools.

<https://weblicht.sfs.uni-tuebingen.de/WebLicht-4/>

[View Tool List](#)
[HELPDESK](#)

[Main Page](#)
[Chain 8 x +](#)

Show tools with status:
 ☐ dev
 ☐ development
 ☒ production
 ☐ withdrawn

Next Choices (Double-click on an icon to add it to the chain)

<b>IMS</b> Morphology morphology 	<b>Berlin-Brandenburg</b> Person Na Named Entities: person 	<b>Berlin-Brandenburg</b> Part-of-Sp Part of Speech: STTS Tagset Lemmas 	<b>IMS</b> Constituent Parser Parsing: Tiger Treebank Tagset 	<b>Berlin-Brandenburg</b> CAB ortho orthography 	<b>Berlin-Brandenburg</b> CAB histor Part of Speech: STTS Tagset Lemmas orthography 
<b>IMS</b> Stuttgart Dependency F Part of Speech: STTS Tagset Parsing (Dep): No Empty Token: Lemmas Parsing (Dep): tiger Parsing (Dep): false 	<b>IMS</b> TreeTagger Part of Speech: STTS Tagset Lemmas 	<b>SfS</b> Berkeley Parser - Berke Part of Speech: STTS Tagset Parsing: tuebadztb 	<b>Berlin-Brandenburg</b> Tokens2Le Language: German Document Type: Lexicon Formal TCF Version: 0.4 entries.type: types 	<b>SfS</b> POS Tagger - OpenNLP Part of Speech: STTS Tagset 	

Input and Chain Selection

[Run Tools](#)
[Clear Results](#)
[Download chain](#)

<b>TheWire.txt</b> Plain Text The Wire ist eine US-amerikanische Fernsehserie, die von 2002 bis 2008 in Baltimore (Maryland) gedreht wurde. Autor ist der ehemalige 	<b>SfS</b> To TCF Converter Language: German Document Type: TCF TCF Version: 0.4 Text 	<b>SfS</b> Tokenizer/Sentences - C newlinebounds false Sentences Tokens 
---	---	--

Done running tools.

# Abschluss

Fragen?

# Lektürehinweise

## Referenztext

- Kai-Uwe Carstensen, Susanne Jekat und Ralf Klabunde (Hrsg.). "Computerlinguistik – Was ist das?", in: *Computerlinguistik und Sprachtechnologie. Eine Einführung*. Hrsg. von Ralf Klabunde et al. Heidelberg: Spektrum, 2009. <https://www.linguistics.rub.de/CLBuch/>

## Weitere Empfehlungen

- Christopher Manning und Hinrich Schütze, *Foundations of Statistical Natural Language Processing*, Boston: MIT Press, 1999. (Oder neuere Auflagen)
- Dan Jurafsky und James A. Martin, *Speech and Language Processing*, Englewood Cliffs: Prentice-Hall, 1999. (Oder neuere Auflagen)

## Darüber hinaus

- Stefan Müller, *Einführung in die Computerlinguistik*. FU Berlin, Fachbereich Philosophie und Geisteswissenschaften, 2013. <https://hpsg.hu-berlin.de/~stefan/PS/cl-slides.pdf> (Foliensatz)
- Christof Schöch, "Aufbau von Datensammlungen", in: *Digital Humanities: Eine Einführung*, hrsg. von Fotis Jannidis, Hubertus Kohle und Malte Rehbein. Stuttgart: Metzler, 2017, S. 223-233.

# Nächste Sitzungen

- 15.1.2019: "Quantitative Analyse 1: Stilometrie, Topic Modeling"
- 18.2.2019: "Quantitative Analyse 2: Supervised Machine Learning"
- Vorbereitung: "Quantitative Analyse", *Digital Humanities: Eine Einführung* (in StudIP)



Christof Schöch, 2019  
<http://www.christof-schoech.de>

---

Lizenz: Creative Commons Attribution 4.0