

# A short introduction to WebLicht

Çağrı Çöltekin

University of Tübingen  
Seminar für Sprachwissenschaft

December 8, 2015

EBERHARD KARLS  
UNIVERSITÄT  
TÜBINGEN



# CLARIN: one-slide introduction

*CLARIN provides easy and sustainable access for scholars in the humanities and social sciences to digital language data and advanced tools.*



# Language resources and repositories

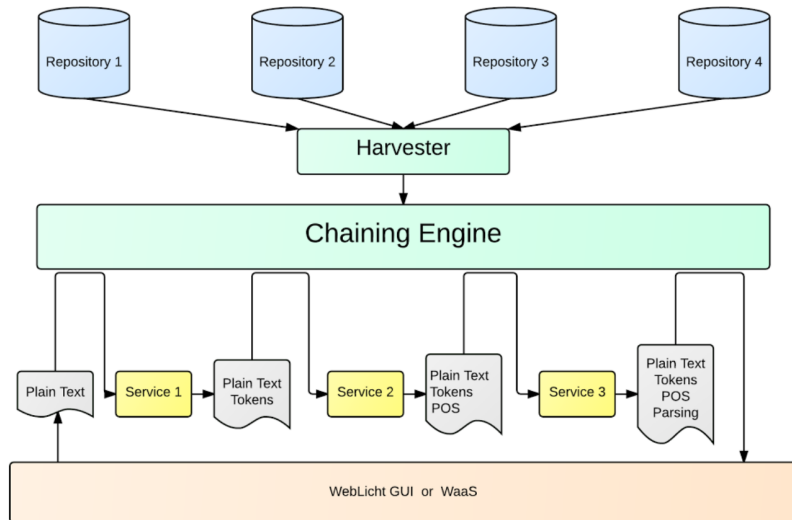
CLARIN offers depositing language resources

- ▶ Prior to deposition: help with 'data management plan'
- ▶ The data is stored in a sustainable way in a 'matching' CLARIN repository
- ▶ The resources are made available according to the depositor's terms
- ▶ All resources are assigned persistent identifiers (PIDs)
- ▶ The resources can be searched based on
  - ▶ meta data through virtual language observatory (VLO)  
<https://vlo.clarin.eu/>
  - ▶ content federated content search (FCS)  
<http://weblicht.sfs.uni-tuebingen.de/Aggregator/>

# The WebLicht infrastructure

- ▶ Centers provide RESTful services
- ▶ The metadata about the services is stored in the center's repositories
- ▶ The services use a common xml-based interchange format (TCF)
- ▶ Each service
  - ▶ receives TCF input with
  - ▶ returns TCF with a new annotation layer
- ▶ The metadata about services are harvested periodically,
- ▶ Users are authenticated through shibboleth
- ▶ The services are made available offered to the users through WebLicht or 'WebLicht as a Service' (WaaS)

# The WebLicht infrastructure: the picture



# WebLicht: an example

View Tool List



Main Page

Chain 2 ✕

+ New Chain

Show tools with status:  dev  development  production  withdrawn

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

IMS: Tokenizer

Sentences  
Tokens

SIS: Tokenizer - OpenNLP

Tokens

Berlin: Tokenizer and Sentence

Sentences  
Tokens

SIS: Tokenizer/Sentences -

newlinebounds **false** ▾  
Sentences  
Tokens



Input and Chain Selection

Run Tools

Clear Results

Download chain

**de\_Geiger** [Plain Text]

Geiger studierte ab 1902 Physik und Mathematik in Erlangen, wo er Mitglied der Burschenschaft der Bubenreuther Erlangen war und in den ersten beiden



SIS: To TCF Converter

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













# WebLicht: an example

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

Main Page **Chain 2** ✕ [+ New Chain](#)






Show tools with status:  dev  development  production  withdrawn

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

<b>IMS: Morphology</b> morphology 	<b>Berlin: Person Name</b> Named Entities: person 	<b>Berlin: Part-of-Speech Tagger</b> Part of Speech: STTS Tagset Lemmas 	<b>IMS: Constituent Parser</b> Parsing: Tiger Treebank Tagset 	<b>Berlin: CAB orthographic</b> orthography 	<b>Berlin: CAB historical text</b> Part of Speech: STTS Tagset Lemmas orthography 	<b>IMS: Stuttgart Dependency</b> Part of Speech: STTS Tagset Parsing (Dep): No Empty Tok Lemmas Parsing (Dep): None 
<b>IMS: TreeTagger</b> Part of Speech: STTS Tagset Lemmas 	<b>SIS: POS Tagger - OpenNLP</b> Part of Speech: STTS Tagset 	<b>Berlin: Tokens2Lexicon</b> Language: German Document Type: Lexicon Format TCF Version: 0.4 entities type: types 				

Input and Chain Selection

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

<b>de_Geiger</b> [Plain Text] Geiger studierte ab 1902 Physik und Mathematik in Erlangen, wo er Mitglied der Burschenschaft der Bubenreuther Erlangen war 	<b>SIS: To TCF Converter</b> Language: German Document Type: TCF TCF Version: 0.4 Text  	<b>SIS: Tokenizer/Sentences -</b> newlinebounds: false Sentences Tokens  
---	---	---

# WebLicht: an example

Main Page Chain 2 x + New Chain

View Tool List HELPDESK

Show tools with status:  dev  development  production  withdrawn

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

<b>IMS: Morphology</b> morphology	<b>Berlin: Lemmas/Lexicon</b> Language: German Document Type: Lexicon Format TCF Version: 0.4 entries.type: lemmas	<b>Berlin: CAB orthographic</b> orthography	<b>SIS: Convert to Negra</b> Document Type: NEGRA Format	<b>Berlin: Tokens/Lexicon</b> Language: German Document Type: Lexicon Format TCF Version: 0.4 entries.type: types
--------------------------------------	--	--	---	---

Input and Chain Selection

Run Tools Clear Results Download chain

<b>de_Geiger (Plain Text)</b> Geiger studierte ab 1902 Physik und Mathematik in Erlangen, wo er Mitglied der Burschenschaft der Rubenreuther Erlanger war	<b>SIS: To TCF Converter</b> Language: German Document Type: TCF TCF Version: 0.4 Text	<b>SIS: Tokenizer/Sentences -</b> newlinebounds Sentences Tokens	<b>IMS: TreeTagger</b> Part of Speech: STTS Tagset Lemmas	<b>SIS: Berkeley Parser - Berkeley</b> Parsing: tuebatzdb	<b>SIS: MaltParser</b> Parsing (Dep): No Empty Tokens Parsing (Dep): With Multi Govs Parsing (Dep): tuebatz	<b>Berlin: Person Name Recognizer</b> Named Entities: person
--	--	---	---	--	--	---

Calling MaltParser ...



# WebLicht: an example

TCF-Dep | File | Navigate | Help

**Search**  
Enter Query Here. Use double quotes around strings or use TüNDRA query syntax.

Run Stats Stop Clear ?

**Tree**

Browse Treebank

Tree No.: 23 / 30

#23: Hans Geiger wurde auf dem Neuen Friedhof Potsdam beigesetzt.

Diagram illustrating the dependency tree for the sentence: "Hans Geiger wurde auf dem Neuen Friedhof Potsdam beigesetzt."

The tree structure shows the following nodes and their dependencies:

- START** (Root) connects to **ROOT**.
- ROOT** connects to **SUBJ** (Hans Geiger) and **ALIX** (wurde).
- SUBJ** connects to **APP** (Hans) and **SUBJ** (Geiger).
- ALIX** connects to **PR** (wurde) and **RP** (auf dem Neuen Friedhof Potsdam).
- RP** connects to **PR** (auf), **DET** (dem), **ATR** (Neuen), **APP** (Friedhof), and **APP** (Potsdam).
- PR** (auf) connects to **PR** (wurde) and **PUNCT.** (beigesetzt).
- DET** (dem) connects to **ATR** (Neuen) and **APP** (Friedhof).
- ATR** (Neuen) connects to **APP** (Friedhof) and **APP** (Potsdam).
- APP** (Friedhof) connects to **APP** (Potsdam) and **PUNCT.** (beigesetzt).
- APP** (Potsdam) connects to **PUNCT.** (beigesetzt).
- PUNCT.** (beigesetzt) connects to **PUNCT.** (beigesetzt) and **PUNCT.** (.)

# WebLicht: an example

**Search**  
[lemma = "haben"] >AUX#verb

**Tree**  
Query Match: 4 / 5  
Sent match: 4 / 5

#24: [ Sein Grab hat<sup>verb</sup>[ sich erhalten<sup>verb</sup>. ],verb ]

START  
ROOT  
DEF SUBJ AUX OBJA PUNCT.  
Sein Grab hat sich erhalten .  
Lemma sein PoS PPOSAT  
Lemma Grab PoS NN  
Lemma haben PoS VAFIN  
Lemma erhalte PoS PRF  
Lemma erhalten PoS VVPP  
Lemma . PoS \$


# WebLicht: an example

**Search**

[Run](#) [Stats](#) [Stop](#) [Clear](#) [?](#)

**Tree**

Matches: 5  
Sent match: 5

[Download statistics](#)  [Browse Treebank](#)

Variable:  Attribute:

Value	Frequency	Percentage
erhalten	2	40.00
zurückziehen	1	20.00
intensivieren	1	20.00
habilitieren	1	20.00

# Summary

- ▶ CLARIN offers easy-to-use and sustainable ways to deposit language data
- ▶ WebLicht infrastructure makes common computational linguistics tools accessible through easy-to-use interfaces
- ▶ WebLicht is in constant development:
  - ▶ new visualization/search/statistics tool
  - ▶ more languages
  - ▶ more, diverse tools as services