CLARIN-D

WebLicht Application and Workspaces

Erhard Hinrichs & Kathrin Beck
University of Tübingen

SPONSORED BY THE
Federal Ministry of Education and Research

Outline

- *Web-based Linguistic Chaining Tool* (WebLicht) for incremental annotation of language corpora

- WebLicht – Motivation
- WebLicht – Architecture
- WebLicht – Future Requirements

WebLicht – Motivation

CLARIN-D

- Many linguistic resources (corpora, dictionaries, …) and tools (tokenizer, tagger, parser, …) are available
- Most of them are implemented to run on local machines. This can be inconvenient and error-prone
- Requirements: go beyond “do-it-yourself” and “download-first” strategies
- The CLARIN solution: *Make tools and resources available as web services*

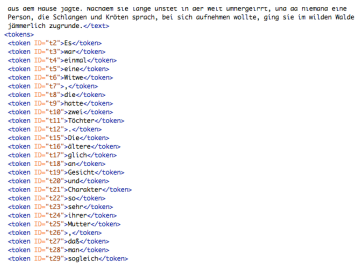
WebLicht – Architecture

- WebLicht is a *SOA* for building annotated text corpora
- Development started in October 2008
- WebLicht consists of the following components:

  - **Distributed services:** offering functionality (resources & tools) over the internet. Implemented as web services (ca. 100 currently)
  - **Repository:** stores metadata and technical information about the services
  - **Web 2.0 based user interface:** interacts with the user and combines services and information from the repository. Access also possible via scripts / programming code

WebLicht – Architecture

Stuttgart



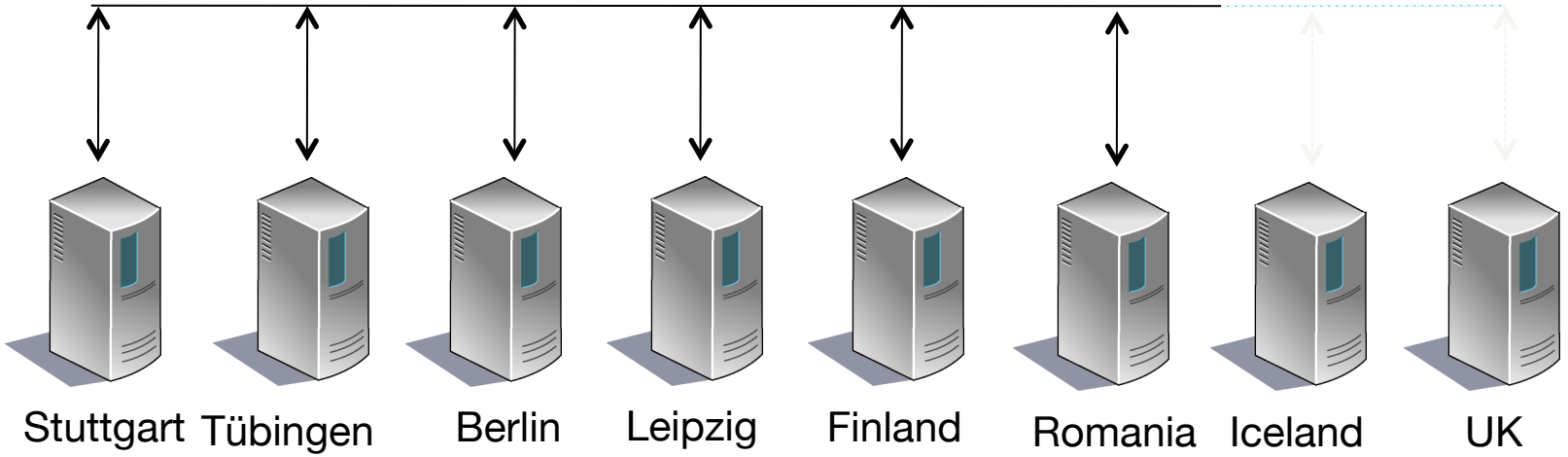
Standard-conformant
Text Corpus Encoding

Tübingen

Web 2.0 Application for
Tool Chaining
and Execution

Leipzig

Repository



Available Services

CLARIN-D

Transformations

- Extraction:
  PDF, Word, ...
- Import/Export:
  MAF, TEI, Negra, Paula,...
- Integration of ISOcat

Resources

- Access to text corpora
- Upload of user data
- Integration of query languages

Annotation

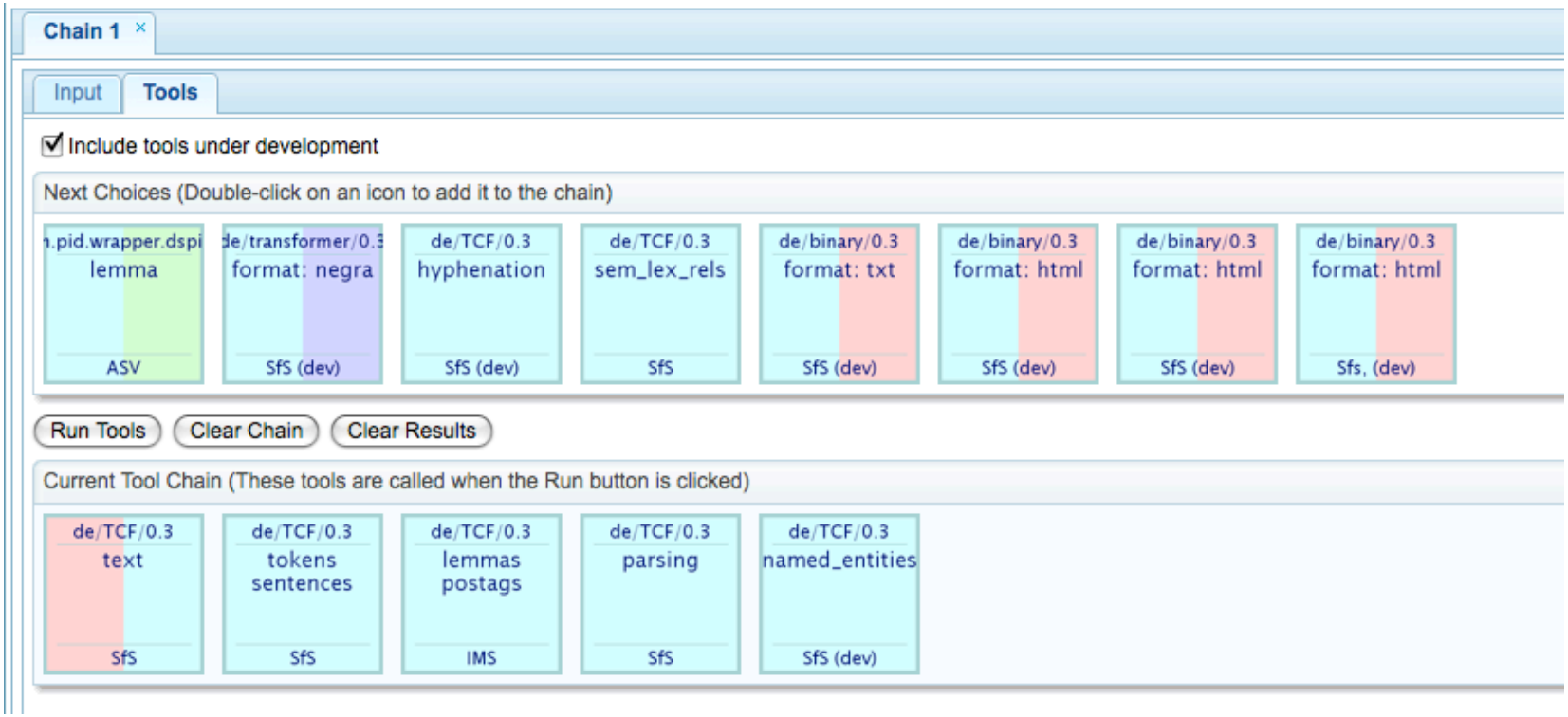
- *8 languages*
- *ca. 100 web services*

- Tokens
- POS Tags
- Constituency Parses
- Semantic Annotation
- NER
- ...

Analysis & Visualization

- Frequency analysis
- Geographic localization
- Reports:
  Machine learning
  Diagrams, ...

WebLicht – Processing Chains



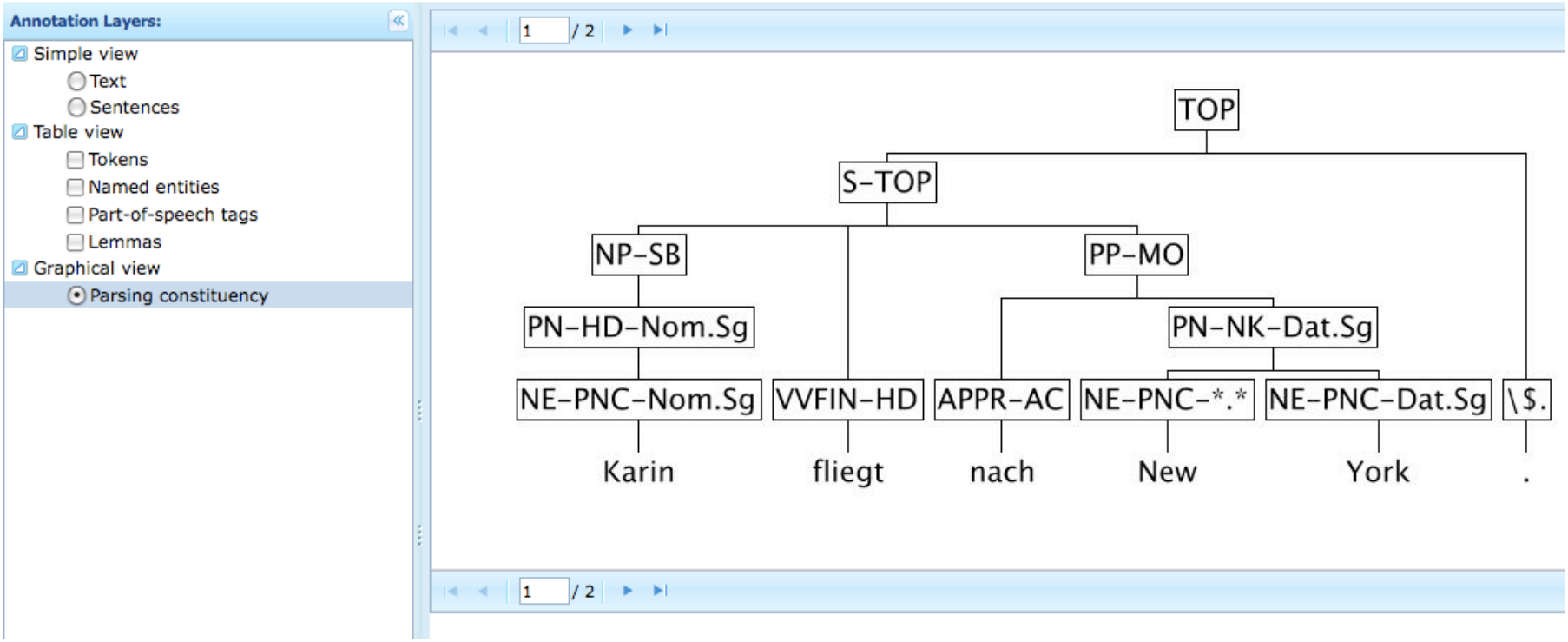
WebLicht – Results

CLARIN-D

```
<tns:parsing tagset="TigerTB">
  <tns:parse>
    <tns:constituent cat="TOP">
      <tns:constituent cat="S-TOP">
        <tns:constituent cat="NP-SB">
          <tns:constituent cat="PN-HD-Nom.Sg">
            <tns:constituent cat="NE-PNC-Nom.Sg">
              <tns:tokenRef tokID="t_0"/>
            </tns:constituent>
          </tns:constituent>
        </tns:constituent>
        <tns:constituent cat="VVFIN-HD">
          <tns:tokenRef tokID="t_1"/>
        </tns:constituent>
        <tns:constituent cat="PP-MO">
          <tns:constituent cat="APPR-AC">
            <tns:tokenRef tokID="t_2"/>
          </tns:constituent>
          <tns:constituent cat="PN-NK-Dat.Sg">
            <tns:constituent cat="NE-PNC-*.*">
              <tns:tokenRef tokID="t_3"/>
            </tns:constituent>
            <tns:constituent cat="NE-PNC-Dat.Sg">
              <tns:tokenRef tokID="t_4"/>
            </tns:constituent>
          </tns:constituent>
        </tns:constituent>
      </tns:constituent>
      <tns:constituent cat="\$.">
        <tns:tokenRef tokID="t_5"/>
      </tns:constituent>
    </tns:constituent>
  </tns:parse>
```

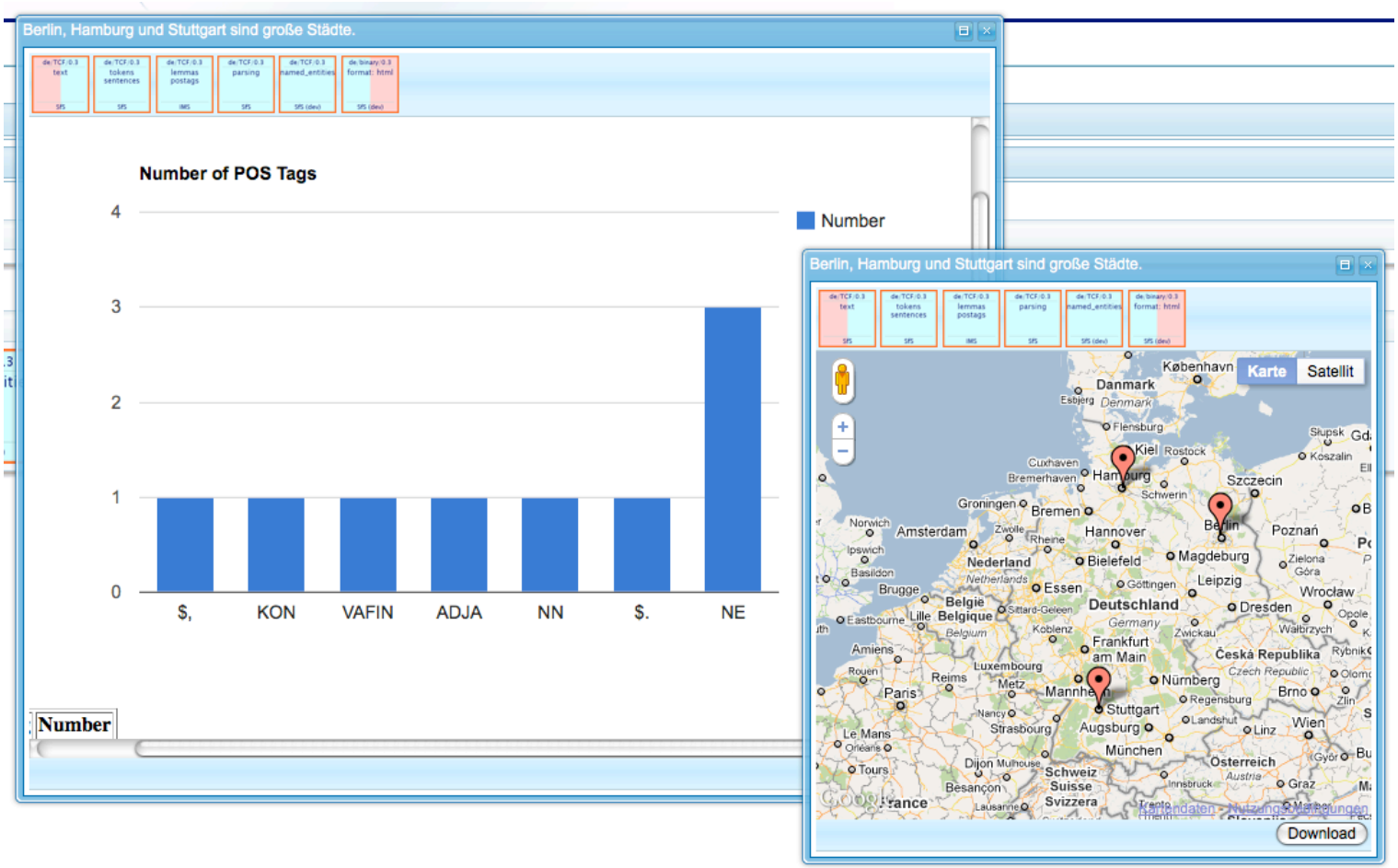
WebLicht – Visualizations

Integrated web application for visualization of TCF data

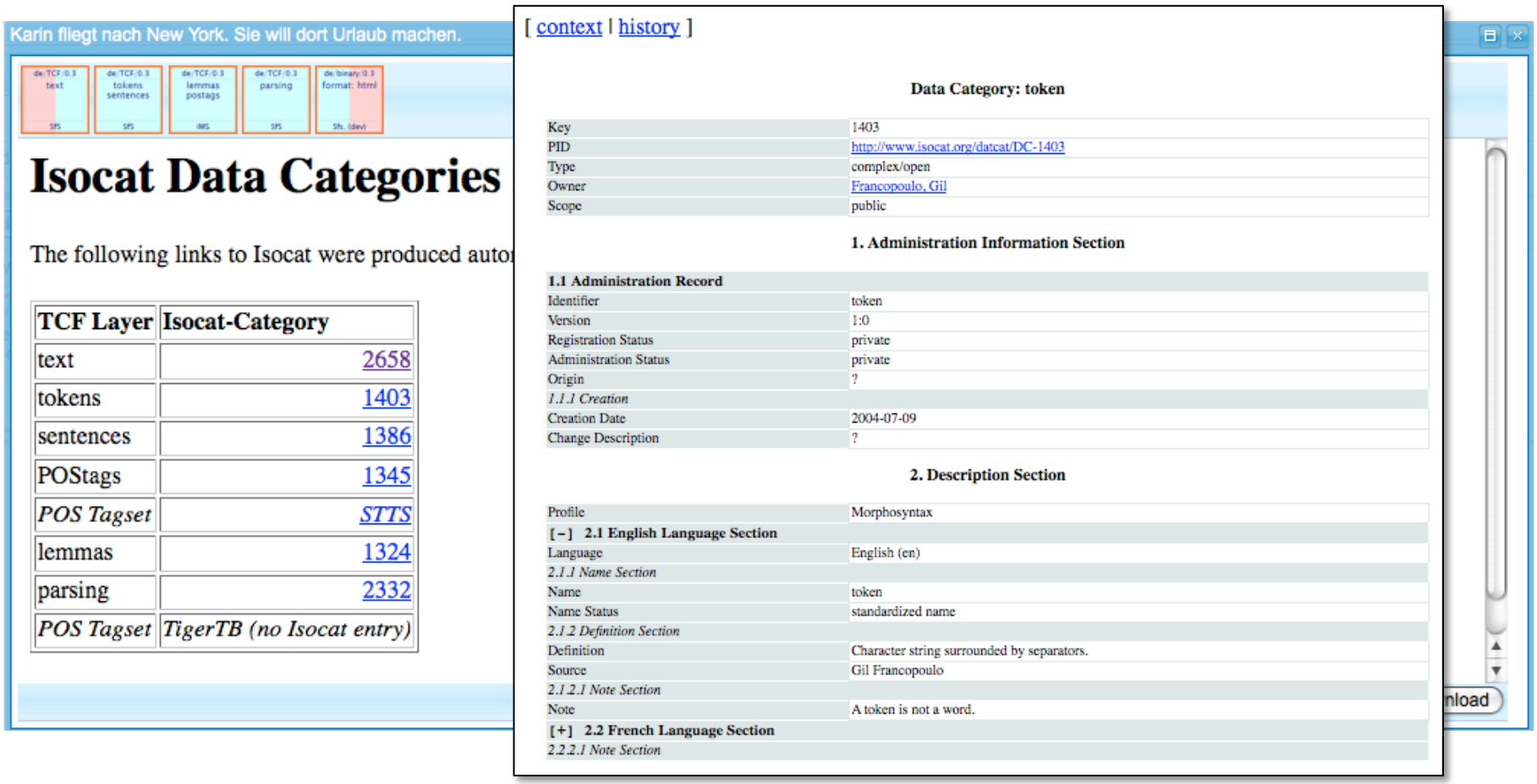


WebLicht – Visualizations

Web service for showing locations on a map that occur in a text



Web service for displaying ISOcat data categories

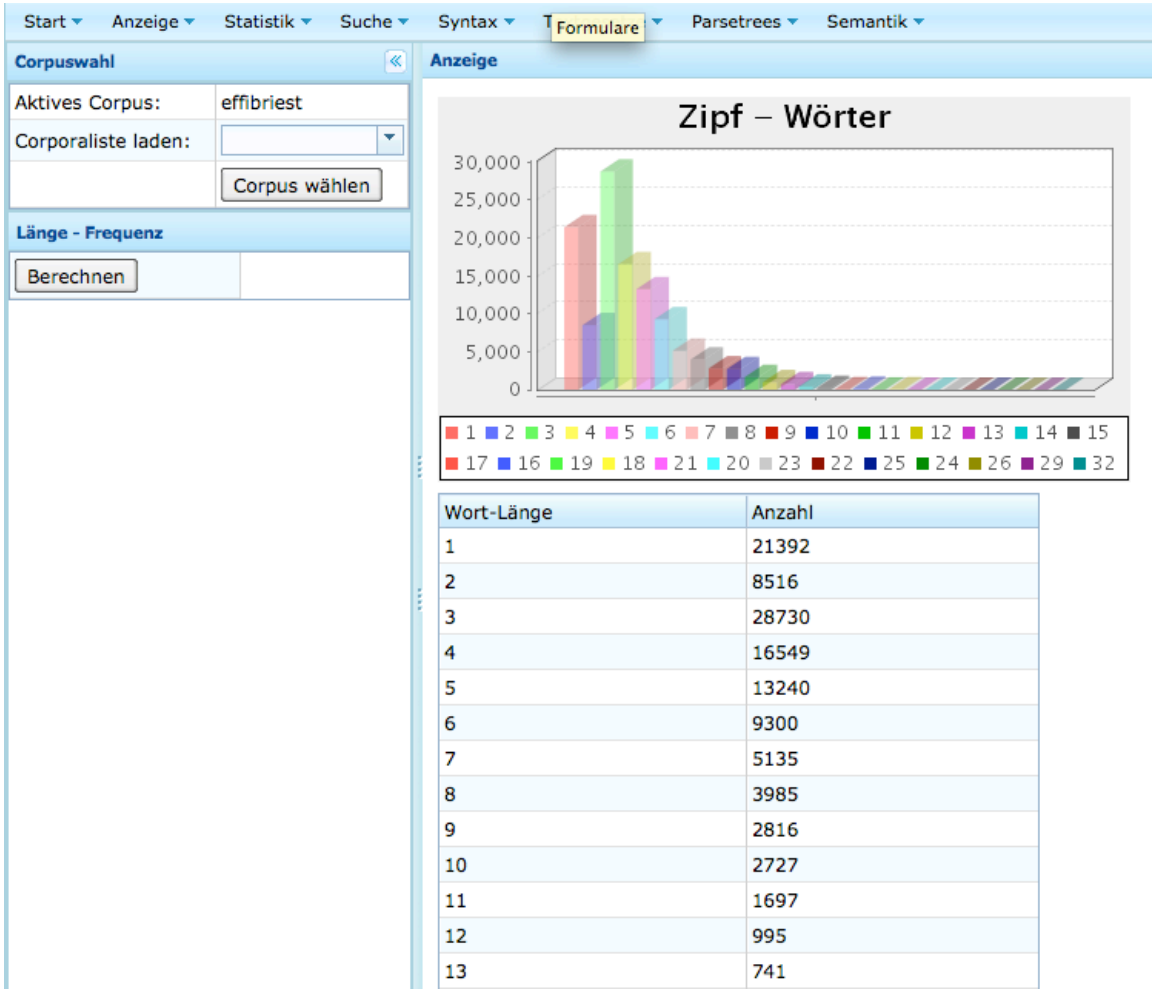


WebLicht - Integration of External Web Applications

CLARIN-D

Dynamic Corpus Analyzer (DCA)

- Web application for visualization and statistical analysis of linguistically annotated text corpora
- Can use WebLicht's TCF format as input



Wort-Länge	Anzahl
1	21392
2	8516
3	28730
4	16549
5	13240
6	9300
7	5135
8	3985
9	2816
10	2727
11	1697
12	995
13	741

WebLicht – Features

- With REST-style web services, it is easy to implement a web service for WebLicht (4-page tutorial)
- The SOA infrastructure is independent of programming languages or operating systems
- The loose-coupling of web services is failsafe
- The chaining algorithm is independent of the data format used
- Web services are located in the institute where they were created

WebLicht – Security

- AAI federation addresses user authentication
  - Open issues:
    - decentralized authorization
    - delegation of user credentials
- Monitoring:
  - The availability of individual web services is monitored and benchmarked continuously

WebLicht – License and IPR

- Currently, only open-source tools are integrated into WebLicht
- Tools and resources may have different license agreements
- Open issues:
  - Harmonization of licenses used in WebLicht
  - Integration of copyright-protected resources and tools

WebLicht – Future Requirements

- Web services are *synchronous*: some linguistic annotation processes are very time consuming
  - an asynchronous behavior of these service would be desirable
- The processing power is limited by local computing resources
  - a centralized approach with High Performance Computing (HPC)
- The current architecture is not sufficiently parallelized and therefore does not scale up:
  - accommodate a large number of simultaneous users
  - parallelization of processes

WebLicht – Future Requirements

CLARIN-D

- Currently, users must store both input data and results on their local machines
  - Online storage in the form of personal workspaces with reliable backup solutions
- Linguistic tools are typically developed in a variety of heterogeneous software environments and programming languages (Java, Perl, Python, C/C++, Prolog, Lisp, …)
  - Encapsulation of individual services with common APIs for interoperability
- Currently, WebLicht services are limited to processing text corpora
  - Incorporation of web services that operate on spoken language and multi-modal datasets

WebLicht – Related Work

- Desktop-based annotation pipelines:
  - Gate
  - UIMA
- Web service environments:
  - Taverna-based workflows (CLARIN Spain)
  - Static pipeline execution (CLARIN-NL)
- Related Projects:
  - EUDAT
  - DARIAH

Links etc.

CLARIN-D Homepage:
http://www.clarin-d.de

WebLicht (login via DFN AAI):
https://weblicht.sfs.uni-tuebingen.de/

Erhard Hinrichs, Kathrin Beck
Seminar für Sprachwissenschaft
Universität Tübingen

Wilhelmstr. 19
D-72074 Tübingen

kathrin.beck@uni-tuebingen.de
erhard.hinrichs@uni-tuebingen.de