

Webservices des Zentrums Sprache an der BBAW

Typen von Webservices

- Datenquellen (Metadaten, Texte, Belegstellen) aus DTA und DWDS
- Computerlinguistische Werkzeuge (Textanalyse und -konvertierung: Tokenizer, Tagger, Konverter) aus DTA und DWDS
- Statistische Berechnungen (syntaktische Relationen und Wortfrequenzen aus dlexDB)

Eine kleine, detailliertere Übersicht zu unseren CLARIN/WebLicht-kompatiblen Webservices findet sich in unserem CLARIN-Repository unter

<http://clarin.bbaw.de/web/>

CLARIN - Common Language Resources and Technology Infrastructure <http://clarin-d.de>

dlexDB - lexikalische Datenbank <http://dlexdb.de>

DTA - Deutsches Textarchiv <http://deutschestextarchiv.de>

DWDS - Digitales Wörterbuch der deutschen Sprache <http://dwds.de>



Webservices des Zentrums Sprache an der BBAW

Datenquellen (Belegextraktion, Metadaten und OAI-PMH-Schnittstellen)

- CLARIN-Repositorium OAI-PMH
- DTA-Repositorium OAI-PMH
- Federated Content Search/Volltextsuche DTA
- Federated Content Search/Volltextsuche Dingler
- Federated Content Search/Volltextsuche C4
- DTA/DDC OpenSearch Wrapper/Volltextsuche
- Dingler/DDC OpenSearch Wrapper/Volltextsuche
- Neue Texte auf der Webseite als Atom-Feed
- Neue Texte in DTAQ als Atom-Feed
- DTA-Autorenverzeichnis im BEACON-Format



Computerlinguistische Werkzeuge

- CAB: historische Textanalyse
- CAB: Orthographie-Kanonisierer
- Demoversion DTA::CAB
- WASTE Tokenizer
- Part-of-Speech Tagger (Wortartenerkenner)
- Eigennamenerkenner
- SynCoP Abhängigkeitsparser
- „Textcorpus nach Lexikon“-Konverter
- „Tokens nach Lexikon“-Konverter
- „Text nach TCF“-Konverter
- „Lemmas nach Lexikon“-Konverter

Webservices des Zentrums Sprache an der BBAW

Statistische Werkzeuge

- dlexDB
 - Types
 - ▶ Frequenz (absolut)
 - ▶ Frequenz (pro Million)
 - ▶ Silbifizierung
 - Lemmas
 - ▶ Frequenz (absolut)
 - ▶ Frequenz (pro Million)
- Wortverlaufskurven (Histogramme)
- Wortprofil



Virtual Language Observatory (VLO)



Die facettierte Suche des VLO ermöglicht es Nutzern, die linguistischen Ressourcen, Dienste und Werkzeuge aller CLARIN-Servicezentren zu erkunden.

Vom BBAW-Repository bereitgestellte Metadaten werden dazu durch das VLO geharvestet und automatisch in den VLO-Katalog eingebunden.

<http://catalog.clarin.eu/vlobeta>



Eine virtuelle Forschungsplattform zur automatischen Annotation von Textcorpora („Webservice Orchestrator“)

WebLicht wurde im Rahmen der CLARIN-Infrastruktur entwickelt, um einen einfachen Zugang zu Webservices zu ermöglichen (d. h. ohne Programmierkenntnisse).

Im Webfrontend werden die Webservices dann per Drag & Drop mit der Maus zu Werkzeugketten („toolchains“) zusammengesteckt.

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



Auswahl einer vordefinierten Toolchain

Available
Annotations for:
German Plain Text

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

So sieht die vordefinierte POS-Toolchain aus:

The screenshot displays the WebLicht POS-Toolchain interface with four panels:

- Karin.txt Plain Text:** Contains the text "Karin fliegt nach New York. Sie will dort Urlaub machen." and a double-headed arrow icon at the bottom.
- SfS To TCF Converter:** Shows "Document Type" as "Plain Text", "Language: German", "Document Type: TCF", and "TCF Version: 0.4". It includes a "Text" input field and a "Convert" button.
- IMS Tokenizer:** Displays "Sentences" and "Tokens" and has an information icon at the bottom.
- IMS TreeTagger:** Displays "Part of Speech: STTS Tagset" and "Lemmas" and has an information icon at the bottom.

Ergebnis der Toolchain-Analyse

TreeTagger

Annotation Layers: language = de [Download as Excel sheet](#) [Download as CSV](#)

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

token ID	tokens	POSTags	lemmas
t1	Karin	NE	Karin
t2	fliegt	VFIN	fliegen
t3	nach	APPR	nach
t4	New	NE	New
t5	York	NE	York
t6	.	\$.	.
t7	Sie	PPER	Sie sie sie
t8	will	VMFIN	wollen
t9	dort	ADV	dort
t10	Urlaub	NN	Urlaub
t11	machen	WINF	machen
t12	.	\$.	.

[Download TCF](#)

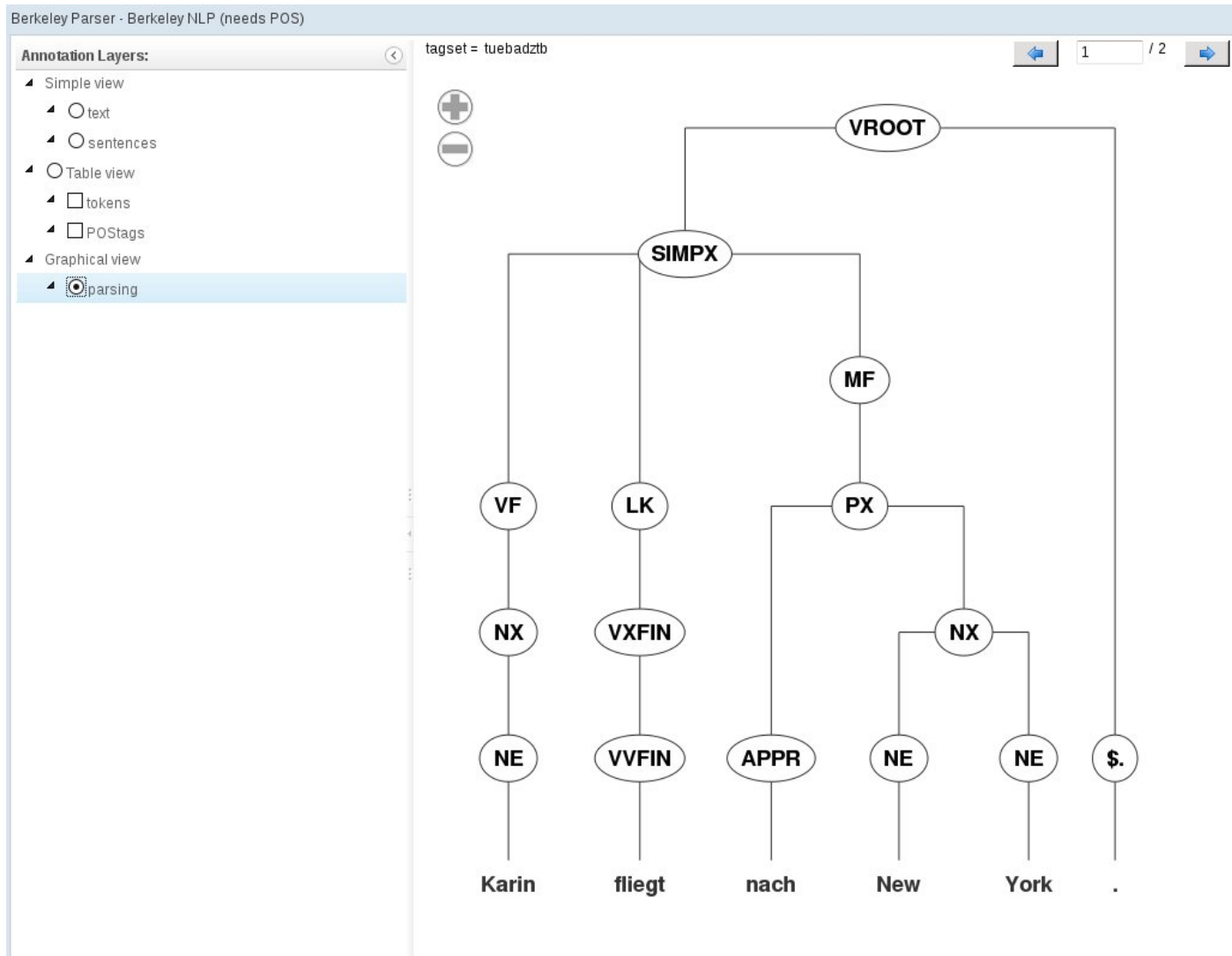


Erweiterung der Toolchain um einen Parser

The screenshot displays the WebLicht web application interface. At the top, there's a navigation bar with 'Main Page' and 'Chain 1 x +'. Below this, a filter section allows users to show tools with specific statuses: ☐ dev, ☐ development, ☒ production, and ☐ withdrawn. A section titled 'Next Choices (Double-click on an icon to add it to the chain)' contains seven tool cards. Each card includes a title, a subtitle, and a description. The tools are: 1. IMS Morphology (morphology), 2. SfS Convert to Negra (Document Type: NEGRA Forr), 3. Berlin-Brandenburg Person N (Named Entities: person), 4. Berlin-Brandenburg Lemmas2 (Language: German, Document Type: Lexicon Forr, TCF Version: 0.4, entries.type: lemmas), 5. Berlin-Brandenburg Tokens2L (Language: German, Document Type: Lexicon Forr, TCF Version: 0.4, entries.type: types), 6. Berlin-Brandenburg CAB orthc (orthography), and 7. SfS German Named Entity Model (Model: conll2003). Below this, an 'Input and Chain Selection' section shows five tool cards: 1. Karin.txt Plain Text (Karin fliegt nach New York. Sie will dort Urlaub machen.), 2. SfS To TCF Converter (Language: German, Document Type: TCF, TCF Version: 0.4, Text), 3. IMS Sentences Tokens (Tokenizer), 4. IMS Part of Speech: STTS Tagset Lemmas (TreeTagger), and 5. SfS Berkeley Parser - Berk Parsing: tuebadztb. Each tool card has a blue bar at the bottom with an 'i' icon and a red 'x' icon, indicating its status or availability.



Ansicht des Parsebaums



DTA::CAB (“*Cascaded Analysis Broker*”)*

* ... sometimes also “*confusing and broken*”

Purpose: linguistic analysis of (historical) German text data

Features:

- TEI-XML serialization (*e.g. in situ <note>, @prev/@next chains, ...*)
- Word- and sentence-tokenization (*abbreviations, ordinals, apostrophes, ...*)
- Heuristic language recognition (*sentence-wise, esp. DE, LA*)
- **Orthographic normalization** (*e.g. Thür $\mapsto$ Tür*)
- Morphological analysis (*TAGH; Haus $\mapsto$ {haus+en/V.imp.sg, Haus/N.neut.sg.nom, ...}*)
- Part-of-Speech tagging (*STTS; Tür $\mapsto$ NN*)
- Lemmatization (*Ärzte $\mapsto$ Arzt*)
- Lexical expansion (*e.g. Haus $\mapsto$ {Haus, Häuser, Haws, Häußren, ...}*)

... and more!



DTA::CAB / Implementation

HTTP/REST Service <http://deutschestextarchiv.de/demo/cab>

- **Requests:** HTTP form GET, POST, hybrid, or XML-RPC (deprecated)
- **Manual Interfaces:** Live browser demo (HTML+JavaScript), file upload
- **Program Interfaces:** command-line, curl, Perl, JavaScript, C, C++, ...
- **Parameters:** **q**/**qd** (query string/document), **a** (analyzer), **fmt** (format), ...
- **Analyzers:** default (robust analysis), expand (lexical expansion), ...
- **I/O Formats:** JSON, TCF, TEI-XML, Text, TT, YAML, XML-RPC, ...

Example Query

<http://deutschestextarchiv.de/demo/cab/query?q=Elephant&fmt=json>



DTA::CAB / Example Output (JSON)

Input:

*EJn zamer Elephant gilt ohngefähr zweihundert Thaler.
Ceterum censeo Carthaginem esse delendam.*

Output:

```
{ "body": [
  { "lang": "de", "tokens": [
    { "...", "word": "Ein", "tag": "ART", "lemma": "eine", "...", },
    { "...", "word": "zahmer", "tag": "ADJA", "lemma": "zahn", "...", },
    { "...", "word": "Elefant", "tag": "NN", "lemma": "Elefant", "...", },
    { "...", "word": "gilt", "tag": "VVFIN", "lemma": "gelten", "...", },
    { "...", "word": "ungefähr", "tag": "ADJD", "lemma": "ungefähr", "...", },
    { "...", "word": "zweihundert", "tag": "CARD", "lemma": "zweihundert", "...", },
    { "...", "word": "Thaler", "tag": "NN", "lemma": "Thaler", "...", },
    { "...", "word": ".", "tag": "$.", "lemma": ".", "...", }
  ] },
  { "lang": "la", "tokens": [
    { "...", "word": "Ceterum", "tag": "FM.la", "...", },
    { "...", "word": "censeo", "tag": "FM.la", "...", },
    { "...", "word": "Carthaginem", "tag": "FM.la", "...", },
    { "...", "word": "esse", "tag": "FM.la", "...", },
    { "...", "word": "delendam", "tag": "FM.la", "...", },
    { "...", "word": ".", "tag": "$.", "lemma": ".", "...", }
  ] }
]
```



DTA::CAB / User Interface

DTA::CAB Demo



Query:

Analyzer:

Format:

Flags: ☐ pretty ☒ clean ☒ exlex ☒ tokenize

Options:

URL: <http://kaskade.dwds.de/cab/query?a=default&fmt=csv&clean=1&q=EJn%20zamer%20Elephant%20gilt%20200%20Thaler.>

```
%% $s:lang=de
EJn    Ejn    Ein    ART    eine
zamer  zamer  zahmer ADJA   zahm
Elephant      Elephant  Elefant NN    Elefant
gilt    gilt  gilt  VVFIN  gelten
200     200   200   CARD   200
Thaler  Thaler Taler  NN     Taler
.        .     .     $.     .
```

[File Demo](#) | [Analyzers](#) | [Formats](#) | [Status](#) | [Documentation](#)



DDC Corpus Search (*“Dialing/DWDS Concordance”*)[†]

[†] ... occasionally also “*diabolically defective cruft*”

Purpose: fast flexible search over large, richly annotated text corpora

Features:

- Token-level attributes (*e.g. \$Token (\$w), \$Pos (\$p), \$Lemma (\$l)*)
- Lemma-based search by default (*via CAB*)
- Fast regular expression queries (*(PCRE)*)
- Meta-data and context-based filters & sorting (*e.g. date, left-context*)
- Exact hit counts (*(by sentence or token)*)
- Result-set aggregation (*(histograms)*)
- External term expansion via HTTP (*e.g. thesauri, semantic similarity*)
- Distributed server architecture (*(parallel processing & scalability)*)

... and more!



HTTP/REST Service (DTA)

<http://kaskade.dwds.de/dtaos>

- **Requests:** HTTP form GET or POST wrapping legacy protocol
- **Interfaces:** HTML, OpenSearch, Perl, Java, or generic HTTP form data
- **Parameters:** **q** (query string), **start** (paging offset), **limit** (page size)
- **Output Formats:** HTML, KWIC, Text, JSON, YAML, ATOM, RSS, ...
- **Corpora:** [Deutsches Textarchiv \(DTA\)](#), [DTA+DWDS](#), [Dingler](#), [ZEIT](#), ...

Example Query

<http://kaskade.dwds.de/dtaos/query?q=Elefant+zahm>



DDC / Example Output (DDC-JSON)

Query: **Elefant && zahm**

```
{ "nhits_":38, "hits_": [
  {
    "meta_":{
      "dtadir":"thunberg_reisen02_1794",
      "author":"Thunberg, Carl Peter (#119036495)",
      "title":"Reisen durch einen Theil von Europa, Afrika und Asien ...",
      "date_":"1794",
      "indices_":["w","p","l",...],
      /* ... additional metadata attributes go here ... */
    },
    "ctx_": [[], [
      [0, "Im", "APPRART", "im", "..."],
      [0, "Durchschnitt", "NN", "Durchschnitt", "..."],
      [0, "gilt", "VVFİN", "gelten", "..."],
      [0, "ein", "ART", "eine", "..."],
      [1, "zahmer", "ADJA", "zahn", "..."],
      [1, "Elephant", "NN", "Elefant", "..."],
      [0, "ohngefähr", "ADJD", "ungefähr", "..."],
      [0, "zweyhundert", "CARD", "zweihundert", "..."],
      [0, "Thaler", "NN", "Taler", "..."],
      [0, ".", "$.", "."],
    ], []]
  },
  /* ... additional hits go here ... */
]
```



DDC+CAB / Query Lizard



DTA Query Lizard

Query Term(s): "Elefant"

[New Search](#)

[Query Lizard](#)

[Help](#)

Base Query

Query Term(s):

Expander:

Expansions:

☒ [Elefannten](#)

☒ [Elefantin](#)

☒ [Elephant](#)

☒ [elefant](#)

☒ [Elefant](#)

☒ [Elefantine](#)

☒ [Elephanten](#)

☒ [elephanten](#)

☒ [Elefanten](#)

☒ [Elephandten](#)

☒ [Elesanten](#)

[New Search](#)

[Query Lizard](#)

[Help](#)

jurish@bbaw.de

DTA Query Lizard version 0.03

0.109725 sec



berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

DDC+CAB / DTA Search

D**T****A**

zahn Elefant #has[author,/Thunber]
☐ in den Titeldaten ☒ im Korpus ☐ in der Dokumentation

Suche im Deutschen Textarchiv (Kernkorpus)

Treffer 1 - 9 von 9

[KWIC](#) · [Neue Suche](#) · [Vorherige](#) · [Nächste](#) · [Hilfe](#)

zahn Elefant #has[author,/Thunberg/] #random[89]

10 Treffer pro Seite Sortierung: Datum aufsteigend/absteigend · zufällig



1: Thunberg, Carl Peter (Übers. Christian Heinrich Groskurd): Reisen durch einen Theil von Europa, Afrika und Asien [...] in den Jahren 1770 bis 1779. Bd. 2. Berlin, 1794. #474 [mehr]

Auf diese Art gefangne männliche **Elephanten** sah ich verschiedenemal an einen dicken Baum gebunden, und in wenig Tagen **zahn** gemacht.



2: Thunberg, Carl Peter (Übers. Christian Heinrich Groskurd): Reisen durch einen Theil von Europa, Afrika und Asien [...] in den Jahren 1770 bis 1779. Bd. 2. Berlin, 1794. #472 [mehr]

Im Durchschnitt gilt ein **zahmer Elephant** ohngefähr zweyhundert Thaler.

