

## Datensammlungen und Computerprogramme

In der akademischen Forschung und Lehre werden Datensammlungen und Software-Werkzeuge seit den ersten digitalisierten Sprachdaten in den 1950er Jahren ausgetauscht und nachgenutzt. Der CLARIN-D-Zentrenverbund bietet Zugang zu vielen Referenz-Datensätzen und -Werkzeugen, die in Forschung und Lehre verwendet werden. Die folgende Übersicht stellt einige häufig genutzte Werkzeuge und Referenzdaten vor.

### Digitale Textsammlungen und Sprachaufnahmen

---

#### Deutsches Referenzkorpus (DeReKo)

DeReKo bildet mit 32,83 Milliarden Wörtern (Stand 01.10.2017) die weltweit größte linguistisch motivierte Sammlung elektronischer deutschsprachiger Korpora verschiedener Genres.

DeReKo: [www.dereko.de](http://www.dereko.de)

In DeReKo mit COSMAS2 recherchieren:

[www.ids-mannheim.de/cosmas2](http://www.ids-mannheim.de/cosmas2)

#### Deutsches Textarchiv (DTA)

Das DTA stellt einen disziplinen- und gattungsübergreifenden Grundbestand deutschsprachiger Texte aus dem Zeitraum von ca. 1600 bis 1900 aus Zeitungsartikeln, Belletristik, Gebrauchsliteratur und wissenschaftlichen Texten bereit. Das Volltextkorpus des DTA ist über das Internet frei zugänglich.

DTA: [www.deutschestextarchiv.de](http://www.deutschestextarchiv.de)

Im DTA recherchieren:

[www.deutschestextarchiv.de/doku/DDC-suche\\_hilfe](http://www.deutschestextarchiv.de/doku/DDC-suche_hilfe)

#### Sprachkorpora des BAS

Das BAS, das *Bayerische Archiv für Sprachsignale*, sammelt, standardisiert, pflegt und distribuiert digitale Sprachressourcen für gesprochenes Deutsch aus den Bereichen Telefondialoge, Gesten, Gebärdensprache, regionale Varianten des Deutschen, etc.

Online-Zugriff auf frei zugängliche Sprachkorpora für die akademische Nutzung:

[www.clarin-d.net/bas-corpora](http://www.clarin-d.net/bas-corpora)

#### PolMine

Die PolMine-Datensammlung besteht aus Texten von öffentlichen Einrichtungen, darunter Parlamentsdebatten. Die aufbereiteten Texte können z.B. mit Sprachverarbeitungswerkzeugen und statistischer Textanalyse in den Sozial- und Politikwissenschaften verwendet werden.

PolMine: [www.polmine.de](http://www.polmine.de)

## Digitale Wörterbücher

---

### DWDS

Das an der BBAW beheimatete Akademienvorhaben „Digitales Wörterbuch der Deutschen Sprache“ (DWDS) stellt Informationen über den deutschen Wortschatz in Vergangenheit und Gegenwart bereit und liefert Webservices für die Nachnutzung in CLARIN-D.

DWDS: [www.dwds.de](http://www.dwds.de)

### GermaNet

Das GermaNet ist ein lexikalisch-semantisches Netz, in dem lexikalische Einheiten nach ihren semantischen Beziehungen zueinander dargestellt werden. GermaNet kann somit als eine Art Thesaurus oder einfache Ontologie betrachtet werden.

GermaNet: [www.clarin-d.net/GermaNet](http://www.clarin-d.net/GermaNet)

### OWID

OWID ist ein Portal für wissenschaftliche, korpusbasierte Lexikografie des Deutschen und beinhaltet eine gemeinsame Suche sowie wörterbuchübergreifende und werkbezogene Suchen in den lexikologisch-lexikografischen Inhalten.

OWID: [www.owid.de](http://www.owid.de)

### Wortschatz

Mit der Sammlung zum Wortschatz steht ein Zugriff auf Verwendungsbeispiele von Wörtern für 250 verschiedene Sprachen zur Verfügung, mit deren Hilfe einzelne Wörter nachgeschlagen und in ihren Kontexten betrachtet werden können. Für das Deutsche stehen dazu über 26 Millionen Sätze zur Verfügung.

Wortschatz: [www.wortschatz.uni-leipzig.de](http://www.wortschatz.uni-leipzig.de)

## Syntaktisch annotierte Textsammlungen (Baumbanken)

---

Baumbanken sind umfangreiche syntaktisch annotierte Textsammlungen, die für die Grammatikforschung, die Computerlinguistik und den Fremdsprachenunterricht gleichermaßen Verwendung finden, darunter die Tübinger Baumbank Collection mit Daten des Deutschen, des Englischen und des Japanischen und die TIGER Baumbank. Mit Hilfe der Webanwendung Tundra lassen sich diese Baumbanken durchsuchen, die Suchergebnisse statistisch auswerten, visualisieren und als Konkordanzen aufarbeiten. Tundra unterstützt alle gängigen Datenformate für konstituenten- und dependenzbasierte Annotationen mit gegenwärtig 65 Baumbanken für 41 Sprachen.

- Tübinger Baumbank Collection: [www.clarin-d.net/tuebac](http://www.clarin-d.net/tuebac)
- TIGER: [www.clarin-d.net/TIGER](http://www.clarin-d.net/TIGER)
- Tundra: [www.clarin-d.net/tundra](http://www.clarin-d.net/tundra)

## Sprachdokumentation

---

### DOBES

Das DOBES-Archiv, das erste Archiv zur Dokumentation bedrohter Sprachen, enthält Informationen zu ca. 70 vom Aussterben bedrohter Sprachen. Vergleichbare Initiativen wie die des Hamburger Zentrum für Sprachkorpora und des Language Archive Cologne dienen der Erforschung von sprachlichen Strukturen im Bereich der Typologie und theoretischen Sprachwissenschaft. Die dort zur Verfügung gestellten audiovisuellen Daten sind aber auch für die Ethnologie, die Oral History und andere Wissenschaften von Interesse.

DOBES: [dobes.mpi.nl](http://dobes.mpi.nl)

## Virtuelle Forschungsumgebungen zur Analyse von Sprachdaten

### EXMARaLDA

EXMARaLDA besteht aus einem Transkriptions- und Annotationseditor, einem Tool zum Verwalten von Korpora und einem Such- und Analysewerkzeug. EXMARaLDA erlaubt die zeitalignierte Transkription von Audio- oder Videodaten mit frei wählbaren Analysekatoren.

EXMARaLDA: [www.exmaralda.org](http://www.exmaralda.org)

### Saarbrücken Corpus Collection and Query Portal

Das Saarbrücker Corpus Collection and Query Portal bietet Zugriff auf derzeit 93 verschiedene Textsammlungen, die nach Textgattung, Zeit, Sprache und Sammlungszweck unterschieden werden und die mittels der Korpusanfragesprache CQP in einer Webumgebung analysiert werden können.

SaCoQP: [www.clarin-d.net/SaCoQP](http://www.clarin-d.net/SaCoQP)

### WebAnno

WebAnno ist ein webbasiertes Werkzeug zur Annotation in verschiedenen Ebenen von, z.B. linguistischen Annotationen mit morphologischen, syntaktischen und semantischen Ebenen. Eigene Annotationsebenen können definiert werden, z.B. für literaturwissenschaftliche Informationen.

WebAnno: [www.clarin-d.net/WebAnno](http://www.clarin-d.net/WebAnno)

### WebLicht

WebLicht ist ein Annotationswerkzeug zur Kombination von über 100 Webservices (z.B. Tokenisierer, Lemmatisierer, Wortarten- und Eigennamenerkennung, Parser) für die automatisierte Annotation von Texten für Forschende aus z.B. Literatur-, Sprach-, Politik- und Geschichtswissenschaften.

WebLicht: [www.clarin-d.net/WebLicht](http://www.clarin-d.net/WebLicht)

### BAS Speech Tools

Der Transkriptionseditor Octra, die automatische multilinguale Segmentation und Aligierung mit MAUS sowie die Emu WebApp zur Auswertung empirischer Sprachdaten unterstützen den bisher extrem zeitaufwendigen und weitgehend manuellen Workflow bei der Aufbereitung gesprochener Sprache.

[www.clarin-d.net/BASWebServices](http://www.clarin-d.net/BASWebServices)

Weitere Datensammlungen und Computerprogramme sind über das Virtual Language Observatory (VLO) auffindbar: [vlo.clarin.eu](http://vlo.clarin.eu)



# Serviceangebote

---

## Datenmanagement

Sprachbasierte Forschungsprojekte können ihre von Drittmittelgebern geforderten Datenmanagementpläne mit einem CLARIN-Zentrum erstellen, dadurch eigene Daten nachhaltig archivieren und zugänglich machen, sowie Daten anderer Projekte und Referenzdaten nutzen.

Datenmanagementplan entwickeln: [www.clarin-d.net/dmp](http://www.clarin-d.net/dmp)

## Handreichungen

Das CLARIN-D Benutzerhandbuch ist ein Leitfaden für die Anpassung und Integration existierender Sprachressourcen mit grundlegenden Themen und Modellierungsprinzipien, die auf alle Arten von linguistischen Ressourcen und Werkzeugen anwendbar sind (Datenkategorien, Metadaten, Annotationen, rechtliche Aspekte, Qualitätsmanagement).

Benutzerhandbuch: [www.clarin-d.net/benutzerhandbuch](http://www.clarin-d.net/benutzerhandbuch)

## Helpdesk

Der CLARIN Helpdesk stellt individuelle Unterstützung für konkrete Fragen zur Nutzung von CLARIN-Ressourcen und -Programmen sowie zum Serviceangebot zur Verfügung. Anhand von konkreten Fragen und Beispielen oder auch im direkten Kontakt mit erfahrenen Experten erhalten Geistes- und Sozialwissenschaftler Unterstützung bei ihren Forschungsfragen.

Helpdesk: [www.clarin-d.net/hilfe](http://www.clarin-d.net/hilfe)

## Rechtliche und ethische Richtlinien und Informationen

Sprachbasierte Daten in den Geistes- und Sozialwissenschaften berühren häufig die Rechte Dritter, angefangen von Verlagen bis zu Autoren oder Sprechenden. Eine Sammlung von allgemeinen Informationen zu juristischen und ethischen Fragestellungen in den Geisteswissenschaften gibt es im Legal Helpdesk.

Legal Helpdesk: [www.clarin-d.net/recht](http://www.clarin-d.net/recht)



[facebook.com/clarindeutschland](https://facebook.com/clarindeutschland)



[plus.google.com/+CLARIND](https://plus.google.com/+CLARIND)



[twitter.com/clarin\\_d](https://twitter.com/clarin_d)

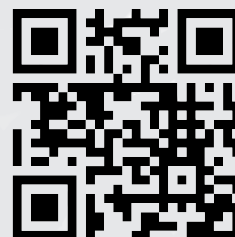


[youtube.com/user/CLARINGermany](https://youtube.com/user/CLARINGermany)

GEFÖRDERT VOM



Bundesministerium  
für Bildung  
und Forschung



[www.clarin-d.net](http://www.clarin-d.net)