

**Herausforderung Open Access
im DFG-Projekt
„Entwicklung einer nachhaltigen und
nutzerorientierten Speicherung und
Bereitstellung von Forschungsdaten
für die historische Linguistik“
(LAUDATIO)**

Kontext



Institut für deutsche Sprache und Linguistik [ʔɪnstitu:t fy:ɐ dɔɪʃə ʃpɛa:χə ʔunt lɪŋɡʊɪstɪk]

Professuren > Korpuslinguistik und Morphologie
ung | Abschlussarbeiten | MitarbeiterInnen | Veranstaltungen | Links



UNIVERSITÄTSBIBLIOTHEK DER HUMBOLDT-UNIVERSITÄT ZU BERLIN

Institut für Bibliotheks- und Informationswissenschaft (IBI)

BESUCHER - STUDIENINTERESSIERTE - BBK | WEBLOG - AGNES - MOODLE - DAB!



Offizielles Mitglied
des ISchools Caucus **information.**
The Power to Transform Our World.





Thema

- Open Access für Forschungsdaten
 - Teildisziplin der Linguistik: Korpuslinguistik
 - DFG-Projekt: Infrastrukturen für Forschungsdaten
 - Herausforderung für CMS/UB
 - Zukunft: Forschungsdaten an der HU Berlin



Agenda

- Forschungsdaten in der Korpuslinguistik & derzeitiger Stand des Zugangs
- Projekt LAUDATIO & Herausforderungen
- Beispielkorpus
- Zusammenfassung & Diskussion



Korpuslinguistik

"Ein **Korpus** ist eine Sammlung **schriftlicher oder gesprochener Äußerungen** in einer oder mehreren Sprachen. Die Daten des Korpus sind digitalisiert, d.h. auf Rechnern gespeichert und **maschinenlesbar**. Die Bestandteile des Korpus, die Texte oder Äußerungsfolgen, bestehen aus den **Daten** selbst sowie möglicherweise aus **Metadaten**, die diese Daten beschreiben, und aus **linguistischen Annotationen**, die diesen Daten zugeordnet sind." (Lemnitzer & Zinsmeister 2006)



Korpuslinguistik

- Aufbau, Aufbereitung und Analyse von Korpora zur Untersuchung von sprachlichen Phänomenen
- Annotation: Zuweisung von Interpretationen auf Satz-, Wort-, Lemma-, Zeichenebene (Token) zu den Rohdaten (Texte) und gemeinsame Speicherung
- Quantitative Auswertung und Vergleich



Forschungsdaten

- Primärquellen (Textkorpora, Audioaufzeichnungen+Transkriptionen)
 - Klassische Manuskripte/Handschriften bis hin zu Voll-Digitalisaten (abhängig von Forschungsfrage)
- Annotationen
- Dokumentationen (Schemata)
- Metadaten für
 - Primärquellen
 - Annotationen
 - Dokumentationen (Schemata)

ANNIS-Recherche im DDB-Korpus



ANNIS* Tutorial DeutschDiachronDigital logged in as "DeutscheDiachroneBaumbe"

Search Form

AnnisQL: cat = "S"

Query Builder: Hide <<

Result: 288

History: Query History ▾

More Corpora

Name	Texts	Tokens
☐ DDB.AHD	20	2846
☒ DDB.FNHD	5	2674
☐ DDB.MHD	4	2760

Search Export

Search Result - cat = "S" (5, 5)

Page 1 of 29 Token Annotations Show Citation URL Displaying Results 1 - 10 of 288

Result 1:

Zum vierden / Sollen wir uns und den kranken sterken und troesten mit etlichen trostspruchen aus

Dat.Sg.Neut Dat.Sg.Neut -- 1.Pl.Pres.Ind 1.Nom.Pl.* 1.Acc.Pl.* -- Acc.Sg.Masc Acc.Sg.Masc Inf -- Inf -- Dat.Pl.Masc Dat.Pl.Masc --

APPRART PIS \$(VMFIN PPER PPER KON ART NN VVINF KON VVINF APPR ADJA NN APPF

tiger
token_layer
exmaralda
token_layer
tiger
paula
paula text

Result 2:

/ auff die weise / ob dich nu deine sunde / der tod oder Teufel erschrecken woll

-- -- Acc.Sg.Fem Acc.Sg.Fem -- -- 2.Akk.Sg.* -- Nom.Sg.Fem Nom.Sg.Fem -- Nom.Sg.Masc Nom.Sg.Fem -- Nom.Sg.Masc Inf 3.Pl.Pa

\$, APPR ART NN \$, KOUS PPER ADV PPOSAT NN \$, ART NN KON NN VVINF VMF

tiger
token_layer
exmaralda
token_layer
tiger
paula
paula text

Result 3:

oder Teufel erschrecken wolten / So wende dich mit allen deinen gedanken darvon / und halte dich nur

-- Nom.Sg.Masc Inf 3.Pl.Past.Ind -- -- 2.Sg 2.Akk.Sg.* -- Dat.Pl.Masc Dat.Pl.Masc Dat.Pl.Masc -- -- 2.Sg 2.Akk.Sg.* --

KON NN VVINF VMFIN -- ADV VVIMP PPER APPR PIAT PPOSAT NN PROAV \$, KON VVIMP PPER ADV

ANNIS-Recherche im DDB-Korpus



Meta Data for DDB.FNHD

meta data		available annotations		
Name	Value	node annotations		
Dokumentation_DDB	DDB Webpage	name	example (click to use query)	url
Sprachstufe	Fruehneuhochdeutsch	bib	bib="Nuber, Veit (1544) Ein kurtze und einfeltige unterweisung zum sterben nutzlich"	
Text	Ein kurtze und einfeltige unterweisung (Veit Nuber)	edition	edition="/"	
		lang	lang="Fnhd"	
		lemma	lemma="/"	
		cat	cat="S"	
		morph	morph="Nom.Sg.Masc"	
		pos	pos="NN"	
		edge types		
		edge annotations		

ANNIS-Recherche im DDB-Korpus



Home

DDB steht für "Deutsche Diachrone Baumbank", ein kleines, syntaktisch tief annotiertes Korpus, bestehend aus drei Subkorpora unterschiedlicher Sprachstufen des Deutschen:

- Subkorpora Althochdeutsch
- Subkorpora Mittelhochdeutsch
- Subkorpora Frühneuhochdeutsch

Der Aufbau des Korpus orientiert sich stark am ["TIGER-Korpus"](#), der größten frei erhältlichen Baumbank des Deutschen.

DDB ist entstanden im Rahmen des vom Berliner Senat geförderten Projekts ["Interdisziplinärer Forschungsverbund Linguistik - Bioinformatik zur Berechnung von Verwandtschaft und Abstammung"](#).

ANNIS-Recherche im DDB-Korpus



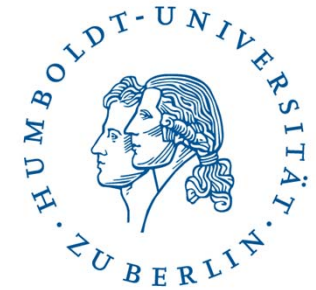
ⓘ / auff die weise / ob dich nu deine sünde / der tod oder Teufel erschrecken woll
 — — Acc.Sg.Fem Acc.Sg.Fem — — 2.Akk.Sg.* — Nom.Sg.Fem Nom.Sg.Fem — Nom.Sg.Masc Nom.Sg.Fem — Nom.Sg.Masc Inf 3.Pl.Pa
 \$, APPR ART NN \$, KOUS PPER ADV PPOSAT NN \$, ART NN KON NN VVINF VMF

ⓘ
 + tiger
 + token_layer
 - exmaralda

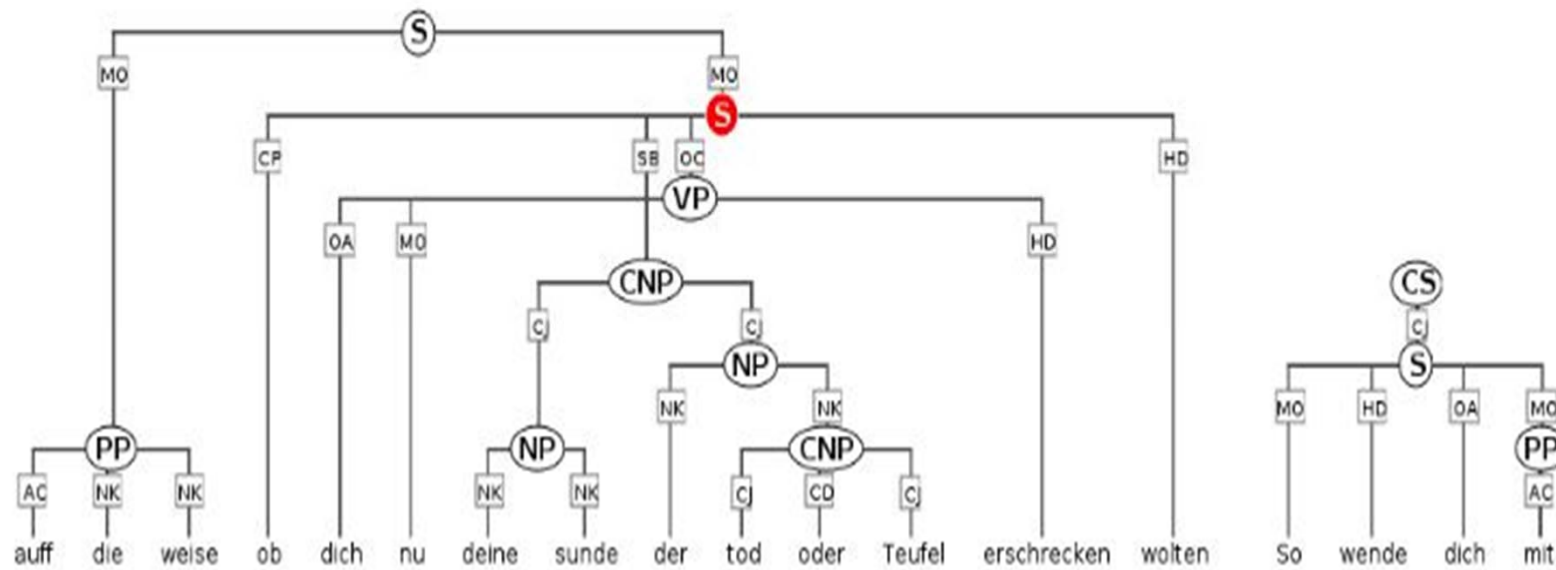
Select Displayed Annotation Levels ▼

bib	Nuber, Veit (1544) Ein kurtze und einfeltige unterweisung zum sterben nutzlich																
edition	/	auff	die	weise	/	ob	dich	nu	deine	sünde	/	der	tod	oder	Teufel	er-schrecken	wolten
lang	Fnhd																
lemma	/	auf	der	weise	/	ob	der	nun	sein	sünde	/	der	tod	oder	teufel	erschrecken	wollen
tok	/	auff	die	weise	/	ob	dich	nu	deine	sünde	/	der	tod	oder	Teufel	erschrecken	wolten

ANNIS-Recherche im DDB-Korpus



<





Forschungsdaten: Zugang

- Kein Zugang
 - unklare Rechtslage (Tondoku afrikanischer Sprachen)
 - Persönlichkeitsrechte, Schutz der Sprecher (z.B. Minderjährige)
 - Datenschutz
- Eingeschränkter Zugang
 - Nachweis von Lizenzen bspw. für Zeitungsabonnements
 - Lizenzerwerb
 - Nicht veröffentlichte Annotationen (nicht fertig, nicht ausreichend etc.)



Forschungsdaten: Zugang

- Zugang mit kleinen Hürden
 - Verfügbar mit Angabe von Nutzerdaten
 - Sicherstellung der rein wissenschaftlichen Nutzung
 - Angebot per Mail (Falko-Korpus)
- Freier Zugang
 - Zugang über bspw. ANNIS2
 - Urheberrechtsfreie Werke
 - Freie Verfügbarkeit der Texte & Annotationen in verschiedenen Formaten



Projektkontext LAUDATIO

- Long term Access and Usage of Deeply Annotated Information
- Projektziel: Entwicklung einer Infrastruktur für die historische Linguistik
 - unter Einbeziehung eines Repository
 - für die Erfassung, persistente Speicherung und jederzeit zugreifbare Bereitstellung fachspezifischer Primärdaten in Form tief annotierter und komplex strukturierter Textkorpora



Projektkontext LAUDATIO

- Förderungsbereich: DFG Wissenschaftliche Literaturversorgungs- und Informationssysteme (LIS)
 - Ausschreibung: Informationsinfrastrukturen für Forschungsdaten (1. Phase: 2011-2014)
 - Kategorie „Disziplinspezifische Organisationsformen“
 - Thema: Entwicklung eines nutzerorientierten Repository zur nachhaltigen Speicherung und Bereitstellung von Forschungsdaten für die historische Linguistik



Projektkontext LAUDATIO

- Projektpartner an der HU Berlin
 - Institut für deutsche Sprache und Linguistik
 - Prof. Dr. Anke Lüdeling
 - Prof. Dr. Karin Donhauser
 - Laurent Romary, PhD (+INRIA)
 - Computer- und Medienservice
 - Prof. Dr. Peter Schirmbacher



Projektkontext LAUDATIO

- Datenquellen: Historische Textkorpora (zunehmend digital)
 - Fragestellungen: Text und Autor, einzelsprachliche Phänomene (Lexikal., Sprachstruktur), sprachübergreifende Phänomene, Sprachwandel
- Zugriff auf Textkorpora, Annotationen
 - über das Annotations- und Suchwerkzeug ANNIS / Webschnittstelle
 - Referenzierung, persistente Verlinkung
 - Download, Import- & Exportmöglichkeiten
 - Aufbauende Dienste über API
 - Verteiltes Arbeiten an Korpora



Herausforderungen

- Urheberrechtsschutz bzw. Copyright
- Wann ist eine Version fertig?
- Projektkontexte, Communities
- Versionierung von Korpora (Architektur, Texte) – Management für Versionen erforderlich
- Editionen: für bestimmte Fragestellungen müssen u.U. entsprechende Abkommen mit den Lizenzinhabern geschlossen werden
- etc.



Herausforderungen

- LAUDATIO – Zuschnitt auf historische Linguistik:
 - Lizenzmodelle für Eigenkorpora (die an der HU verfügbar sind wie "DeutschDiachronDigital")
 - Lizenzmodelle für annotierte Fremdkorpora (die wir zum Ende der ersten Projektphase erwarten)
 - Verfügbarkeit
 - Vorgaben aus den Partnerprojekten

Data Sharing: Ziele für LAUDATIO



- Ziel für Community: urheberrechtsfreie Versionen
- Regeln / „Data-Sharing-Kultur“ angepasst an die Community & Umgebungen
 - Lizenzen für Korpora: CC-BY?
 - Lizenzen für Dokumentationen/Annotationsschemata: CC-BY?
 - Lizenzen für Metadaten: CC-BY?



Beispiel „Ridges Corpus“

- Register in Diachronic German Science
 - Untersuchung der deutschen Wissenschaftssprache ab dem 16. Jahrhundert, Kräuterkunde
 - Projektseminar am IdSL HU Berlin
 - Gefördert durch Google Digital Humanities Award
 - Korpus aus Digitalisaten hauptsächlich von Google Books (urheberrechtsfrei)



Our commitment to the digital humanities

7/14/2010 03:45:00 AM

(Cross-posted on the [Google Research Blog](#))





Metadaten und Schemata

W >>

DDB.AHD 20

DDB.FNHD 5

History

Texts

e

Meta Data for ridges.herbology

meta data		available annotations		
Name	Value	node annotations		
annotators	Alexander Meyer, Anna Tóth, Antonia Dittberner, Cora Döhn, Imke Driemel, Ina Riesler, Johannes Mursell, Judith Esser, Katharina Reinig, Kornél Kovács, Kristina Stephan, Laura Lueders, Malte Belz, Réka Tóth, Samuele Maniscalco, Silke Andresen, et al.	name	example (click to use query)	url
projectDesc	RIDGES Project (Register In Diachronic GERman Science), funded by the Google Digital Humanities Research Award	atLeast	atLeast="1"	↔
respStmt	Compiled by Amir Zeldes, Anke Lüdeling	atMost	atMost="3"	↔
sponsor	Google Inc.	author_ref	author_ref="pron1sg"	↔
version	1.0	clean	clean="/"	↔
		definition	definition="trans_to_german"	↔
		del	del="del"	↔
		div1	div1="div1"	↔
		div1_type	div1_type="book"	↔
		div2	div2="div2"	↔
		div2_n	div2_n="5"	↔
		edge types		+
		edge annotations		+



Downloadmöglichkeiten

Home

Dokumentation

Zugang

Download

Korpuslinguistik

 English

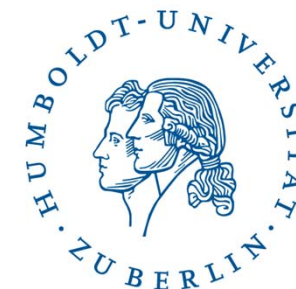
Ridges

Download

ridges.herbology-Korpus (V 1.0)

Das gesamte ridges.herbology-Korpus kann in den folgenden Formaten heruntergeladen werden:

Korpus	PAULA XML	relAnnis
ridges.herbology		



Downloadmöglichkeiten

Dokument	Faksimile	TEI XML	TreeTagger/ CWB SGML	Excel
All documents (Zipped)				
New Kreüterbüch / Leonhart Fuchs (Basel, 1543) [S. 2-e4 (25 Seiten), 5169 Token]				
Alchemistische Praktik / Andreas Libavius (Frankfurt, 1603) [S. 4-26, 5282 Token]				
Pflanz-Gart (Vorrede) / Daniel Rhagor (Bern, 1639) [S. 1-10, 2225 Token]				
Pflanz-Gart (Capitel 4) / Daniel Rhagor (Bern, 1639) [S. 33-45, 2756 Token]				
Curioser Botanicus oder sonderbares Kräuterbuch (1675) [S. 1-11, 2230 Token]				



Zusammenfassung & Diskussion

- Urheberrechtsfreie Werke
- Empfehlungen für CC:BY-Lizenzen?
- Rolle von Forschergruppen / Communities
- Rolle von Umgebungen (eResearch, VRE)
- „Abgeschlossenheit“ von Forschung
- Rolle der Serviceeinrichtungen wie Bibliotheken und Rechenzentren



OA-Publikationen als Beitrag zu freien Korpora!



Beispiele Korpora

- http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/links/korpora_links
- CLARIN resources inventory: http://www.clarin.eu/view_resources
- „Deutsch Diachron Digital“: <http://korpling.german.hu-berlin.de/ddd/search.html>
- Ridges Herbiology Corpus: http://korpling.german.hu-berlin.de/ridges/access_de.html
- ANNIS-Demos: <http://korpling.german.hu-berlin.de/trac/annis/>