

Nina Bartsch, Stefanie Dipper, Birgit Herbers,
Sarah Kwekkeboom, Klaus-Peter Wegera
Ruhr-Universität Bochum

Lars Eschke, Thomas Klein, Elke Weber
Universität Bonn

Projektziel

Das DFG-geförderte Projekt *Annotiertes Referenzkorpus Mittelhochdeutsch (1050-1350)* hat zum Ziel, die hochdeutschen Sprachdenkmäler von ca. 1050 bis 1200 weitestgehend vollständig und von ca. 1200 bis 1350 in strukturierter Auswahl für eine breite Nutzergemeinde für historiolinguistische und mediävistische Forschung verfügbar zu machen.

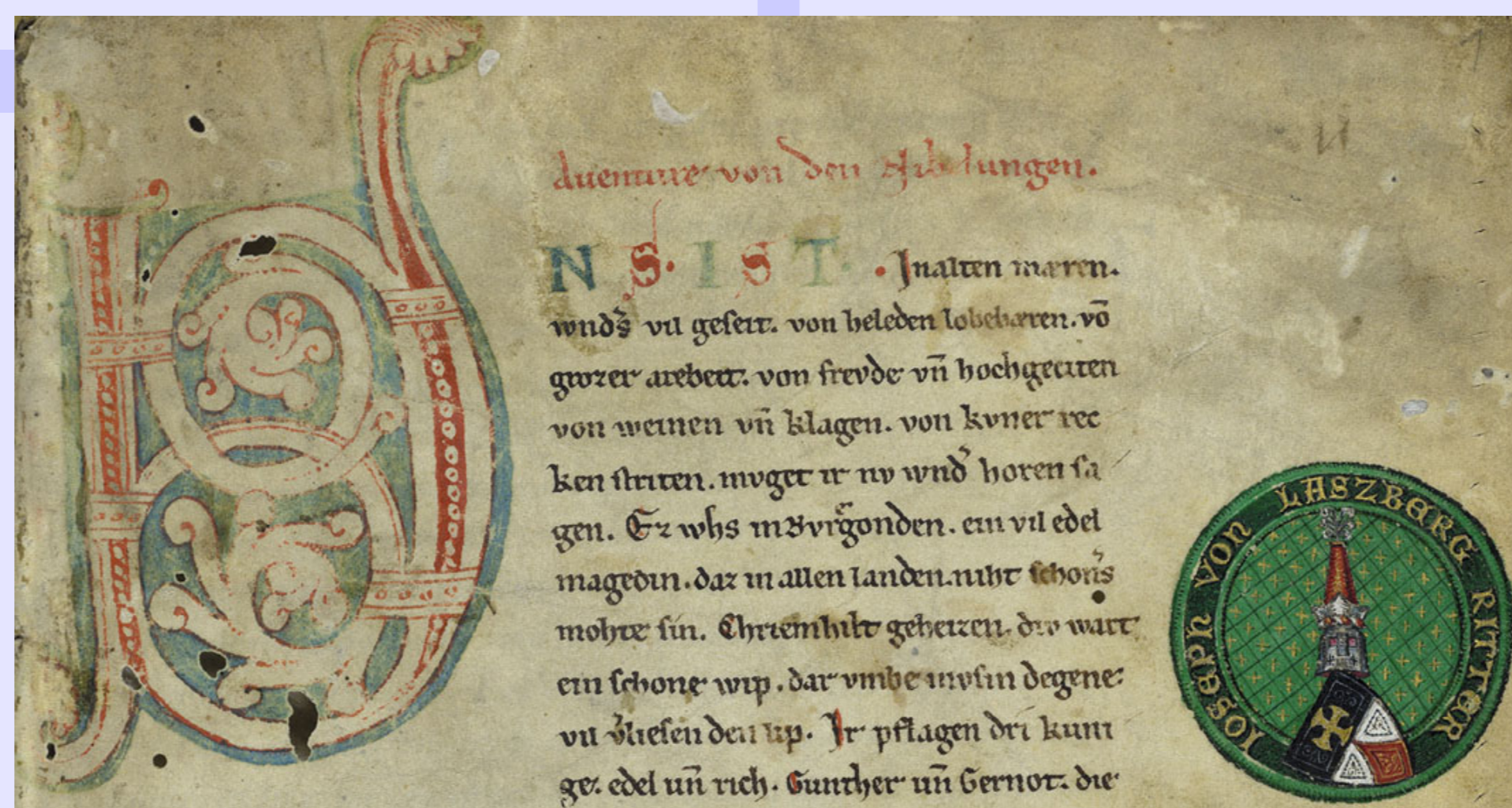
Diachrones Referenzkorpus

Ein weiteres DFG-Projekt befasst sich mit der Erstellung des *Referenzkorpus Altdeutsch*, mit Texten von ca. 750-1050 (Universitäten HU Berlin, Frankfurt, Jena).

Zusammen mit einem geplanten Projekt zum Frühneuhochdeutschen ergibt sich somit ein einheitlich erfasstes diachrones Korpus des Deutschen.

Transkription

- ♦ Diplomatisch
- ♦ Sonderzeichen, z.B.
 - ♦ "langes s": ſ
 - ♦ Diakritika u.ä.: y, û
 - ♦ Abkürzungen: ʒ ("er")
- ♦ Layout (Spalten u.ä.)
- ♦ Spätere Ergänzungen
- ♦ Editionsählung



Nibelungenlied (C), 1.Hälfte 13.Jh., obd.
(Karlsruhe, Landesbibl., Cod. Donaueschingen 63, Bl. 1r)

Annotation

- ♦ Wort- und Satzgrenzen
- ♦ Lemma, Wortart, Morphologie
- ♦ Normalisierung
- ♦ semi-automatisch, interaktiv
- ♦ Wort- und Satzgrenzen (manuell)

1. Transkription

OTTO (Online Transcription Tool...)

Text Editor

Transcription

Zeilenzählungs-Editor

Header - Editor

Lokale Transkriptions-Regeln

2. semi-automatische Annotation

```
<DDMhd>
<header sigle="Nib"/>
<follo id="f 1" count="1" side="r">
<line id="l 1" count="1">
  <token id="t 1" count="1">
    <ref_second page="0" line="1"/>
    <form dipl="auentiere" trans="auentiere" norm="auentiere"/>
    <lemma gen="auentiere" inst="auentiere"/>
    <pos gen="NA" inst="NA"/>
    <infl class="Fem" inst="Fem"/>
    <infl val="Nom_Sg"/>
    <comment val="/R"/>
  </token>
  <token id="t 2" count="2">
    <ref_second page="0" line="1"/>
    <form dipl="von" trans="von" nor/>
    <lemma gen="von" inst="von"/>
    <pos gen="AP" inst="APPR"/>
    <comment val="/R"/>
  </token>
  <token id="t 3" count="3">
    <ref_second page="0" line="1"/>
    <form dipl="den" trans="den" nor/>
    <lemma gen="den" inst="den"/>
    <pos gen="DD" inst="DDART"/>
  </token>
  </line>
</follo>
```

1:er	pron	NSm	GP	GSF	DSF				
2:der	art	NSm	GP	GSF	DSF				
3:hêre	m	NS							
1:be-vêlher	stv3b	3SVI	1SVI						
2:sin	anv	3SGK	1SGK	I	1PG				
3:sich	pron	A							
1:got	m	DS	NP	AP	GP				
2:gotte	f	NS	AS	GS	DS	NP	AP		
3:goten	swv	1SG	3SGK	1PG	2SGB	I			
1:mit	prap								
2:mit(e)	adv								
3:ich	pron	DS							
1:inner	adj	DSn	DSm						
1:ge-bêt	n	DS	GP						
2:ge-bête	f	NS	AS	GS	DS	NP	AP		
3:ge-biten	stv5b	3SVK	1SVK	2SVI	1PVK				
1:Benjamin	PN	AS	NS	DS					

3. Korpusnutzung in ANNIS

- ← Diplomatische Transkription (Unicode)
- ← Position im Satz
- ← Lemma
- ← Diplomatische Transkription (ASCII)
- ← Wortart
- ← normalisierte Wortform
- ← Morphologie

Textauswahl

- ♦ Frühmhd. Zeit (1070-1200): Alle überlieferten Texte werden komplett erfasst.
- ♦ Zweite Hälfte der mhd. Zeit (1200-1350): Textauswahl, die zeitlich und räumlich gleichmäßig verteilt ist und auch die Überlieferungsformen Vers, Prosa und Urkunde gleichermaßen repräsentiert.
- ♦ Das Gesamtkorpus wird aus rund 300 Texten mit knapp 2 Mio. Wortformen bestehen.

URLs

- ♦ Projektseite: <http://www.linguistics.ruhr-uni-bochum.de/mhd>
- ♦ Korpus der Mittelhochdeutschen Grammatik: <http://www.degruyter.de/files/pdf/9783484110038Quellenkorpus%20Uebersicht.pdf>
- ♦ Tools der Mittelhochdeutschen Grammatik: <http://www.mittelhochdeutsche-grammatik.de>
- ♦ ANNIS: <http://www.sfb632.uni-potsdam.de/annis>