RUHR-UNIVERSITÄT BOCHUM
HUMBOLDT-UNIVERSITÄT ZU BERLIN

ANNIS-Hist: Historische deutschsprachige Korpora in ANNIS

Marcel Bollmann[1], Stefanie Dipper[1], Mario Frank[12], Julia Krasselt[1], Florian Petran[1], Tom Ruette[2]
[1]Ruhr-Universität Bochum [2]Humboldt-Universität zu Berlin

Einführung

In den letzten Jahren gibt es mehr und mehr Initiativen mit dem Ziel, historische Sprachdaten für die Forschungsgemeinde aufzubereiten und zur Verfügung zu stellen. Dieses Poster stellt eine Reihe von neueren historischen Korpora des Deutschen vor.

Die Korpora werden alle über das Suchtool ANNIS zur Verfügung gestellt. Das ermöglicht eine Suche über alle Korpora hinweg, sodass diachrone Entwicklungen untersucht werden können. Voraussetzung dafür ist die Annotation aller Korpora mit vergleichbaren Tagsets.

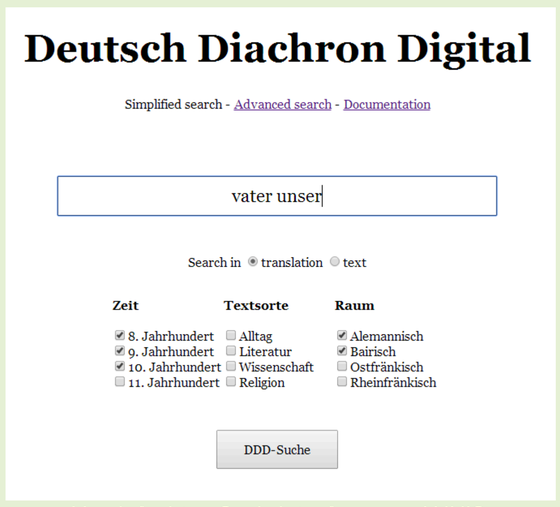
HiTS (*His*torisches *Tag*Set)

HiTS ist ein gemeinsames Tagset für die Referenzkorpora Altdeutsch, Mittel- und Frühneuhochdeutsch.

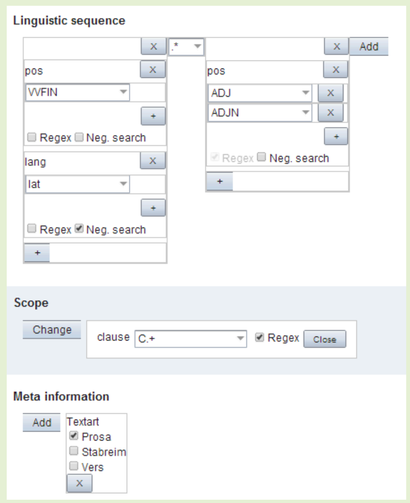
- einheitliches Tagset erlaubt diachrone Recherchen
- angelehnt an das STTS
- ergänzende Tags für historische Phänomene, z.B. ADJN für „nachgestelltes attributives Adjektiv“
- zusätzlich: Richtlinien für Tokenisierung (Wortgrenzen)
- zusätzlich: Unterscheidung zwischen Wortart des Lemmas und des Belegs

Beispiel:

- inti sie **ni / PTK > PTKNEG** quedent imo **niouuiht / PI > PNEG** „und sie nicht sagen ihm Nichts“ = „sie sagen ihm nichts“ (Ahd.)
- swer **niht / PI > PTKNEG** gloubet, der ist iu verteiltet „wer nicht glaubt, der ist schon verurteilt“ (Spätahd.)



Vereinfachtes Such-Interface von ANNIS



Beispiel-Anfrage im QueryBuilder...

ANNIS

ANNIS ist ein Korpus-Suchtool, das an der HU Berlin entwickelt wird (Zeldes et al., 2009). Das Tool eignet sich insbesondere für Daten, die auf verschiedenen Ebenen mit verschiedenen Typen von Annotationen annotiert sind. Damit unterstützt ANNIS die Idee, dass Korpora sukzessive erweitert und mit weiteren Annotationen angereichert werden können.

Merkmale von ANNIS

1. Visualisierung
   - verschiedene Darstellungsmöglichkeiten, z.B. Tabellenansicht, Graphen, Diskursansicht
   - einzelne Ebenen können verborgen werden
   - Treffer werden farblich hervorgehoben
2. Suche in den Daten
   - AnnisQL: zugrundeliegende formale Anfragesprache
   - QueryBuilder: graphische Oberfläche
   - Vereinfachtes Such-Interface für die historischen Korpora

```
pos="VVFIN" & lang!="lat"
& pos=/(ADJ)|(ADJN)/
& clause = /C.+/
& #1_=_#2
& #2 .* #3
& #4_i_#1
& #4_i_#3
& meta::Textart = "Prosa"
```

...und übersetzt in AnnisQL

- Anfragen werden in AnnisQL übersetzt und können dort weiter verfeinert werden

Referenzkorpus Altdeutsch (750–1050)

Donhauser (HU Berlin), Gippert (Frankfurt), Lühr (Jena)

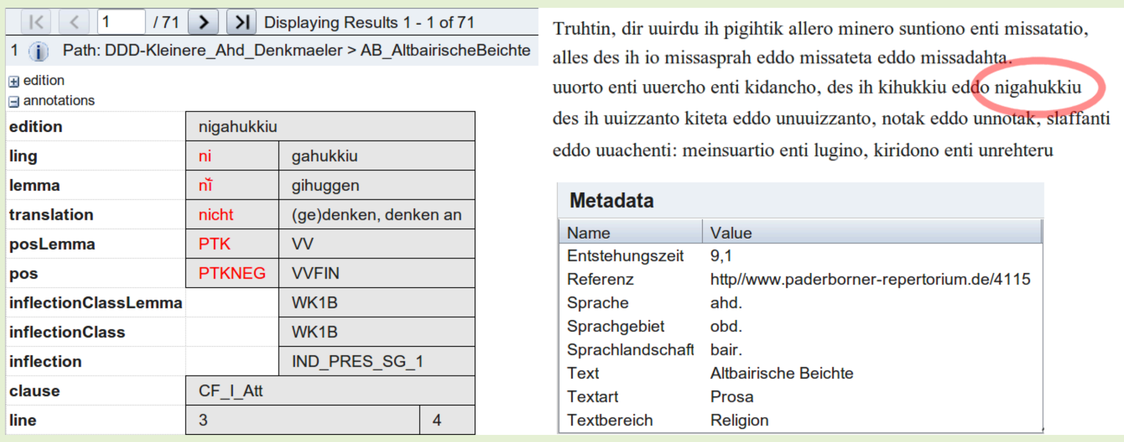
Das Referenzkorpus Altdeutsch erfasst die gesamte althochdeutsche und altniederdeutsche Textüberlieferung mit 650 000 Belegen.

Annotationsebenen:

- Wortarten-, Morphologie- und Lemma-Annotation
- Unterscheidung von Lemma- und Beleg-Wortarten (gemäß HiTS)
- Flache Annotation von Satztypen

Beispielphänomen: Negationspartikel *nī*

- typischerweise klitisch (s. Ebene „edition“)
- eigenes Token auf Annotationsebenen (s. Ebene „ling“)



Annotationsansicht (links), Editionsansicht (oben rechts) und Metadaten (unten rechts) in ANNIS

rate	ñi	bat	daz	fi	liezen	uallin	den	unrechten	willē	uon	ir
rât	ër	biten	dazz	ër	lâzen	vallen	dër	un-rëht	wille	von	ir(e)
1007	1008						1009			1010	
rate	her	bat	daz	fi	liezen	uallin	den	unrechten	willen	uon	ir
Akk_Pl	Masc_Nom_Sg_3	Ind_Past_Sg_3	--	*__Nom_Pl_3	*_Past_Pl_3		Masc_Akk_Sg	Pos_Masc_Akk_Sg_*	Akk_Sg		Neut_Dat_Sg_0
0											
p1											
NA	PPER	VVFIN	KOUS	PPER	VVFIN	VVINF	DDART	ADJA	NA	APPR	DPOSA
[ra]te	h'	bat	daz	<fi>	[liezen]	[uallin]	[den]	[unrechten]	[wi]lle-	uon	ir

Beispielergebnis aus dem Mittelhochdeutsch-Korpus bei Suche nach VVFIN gefolgt von VVINF (Suchanfrage in AnnisQL: *pos="VVFIN" & pos="VVINF" & #1 . #2*)

Referenzkorpus Mittelhochdeutsch (1050–1350)

Wegera, Dipper (Bochum); Wich-Reif, Klein (Bonn)

Das Referenzkorpus Mittelhochdeutsch enthält eine strukturierte Auswahl der Textüberlieferung mit etwa 2,1 Mio. annotierten Wortformen.

Annotationsebenen:

- Wortarten-, Morphologie- und Lemma-Annotation
- Unterscheidung von Lemma- und Beleg-Wortarten (nach HiTS)

Beispielphänomen: Abfolge VVFIN/VVINF im Verbalkomplex

- z.B. *dass sie fallen ließen* vs. *dass sie ließen fallen*
- AnnisQL erlaubt Suche nach konkreter Abfolge von Wortarten-Tags

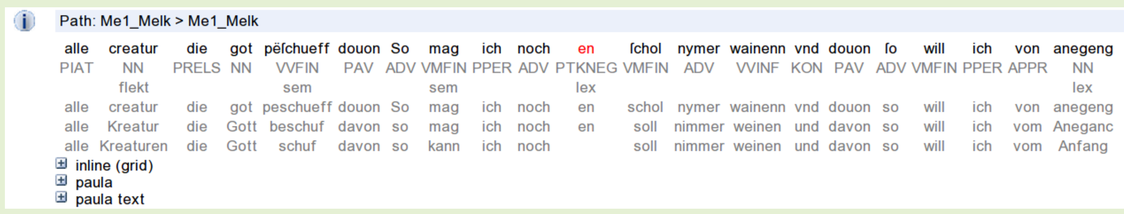
Anselm-Korpus (Frühneuhochdeutsch, 14.–16. Jh.)

Dipper, Schultz-Balluff, Wegera (Bochum)

Das Anselm-Korpus umfasst 24 frühneuhochdeutsche Überlieferungsträger des Passionstraktats „St. Anselmi Fragen an Maria“. Die Überlieferungen variieren in Länge, Textsorte (Vers vs. Prosa) und dialektalen Merkmalen.

Annotationsebenen:

- Wortarten-Annotation nach STTS
- Normalisierung: streng wortnahe Anpassung der Form an moderne Schreibung
- Modernisierung: Anpassung der Formen an moderne Grammatik und Semantik:
  1. Anpassung an moderne Flektion („flekt“)
  2. Lexikalische Anpassung extinkter Formen („lex“)
  3. Anpassung bzgl. semantischem Wandel der historischen Formen („sem“)



Ausschnitt aus dem Anselm-Korpus in ANNIS; historische Wortformen können in UTF-Darstellung oder simplifizierter Darstellung (ersetzt z.B. ſ durch s) gesucht werden

Beispielphänomen: Regelhafte graphematische Ersetzungen

- z.B. *v* → *u* (wie in *vnd* → *und*) oder *u* → *v* (wie in *douon* → *davon*)
- Normalisierungsebene erlaubt Vergleich historischer und moderner Schreibvarianten

- Referenzkorpus Altdeutsch: http://www.deutschdiachrondigital.de/
- Referenzkorpus Mittelhochdeutsch: http://referenzkorpus-mhd.uni-bonn.de/
- Anselm-Projekt: http://www.linguistics.rub.de/anselm/ sowie http://www.rub.de/wegera/sanktanselmus.html
- Amir Zeldes, Julia Ritz, Anke Lüdeling und Christian Chiarcos (2009). *ANNIS: a search tool for multi-layer annotated corpora.* In: Proceedings of Corpus Linguistics. Liverpool, UK.
- Stefanie Dipper, Karin Donhauser, Thomas Klein, Sonja Linde, Stefan Müller und Klaus-Peter Wegera (2013). HiTS: ein Tagset für historische Sprachstufen des Deutschen. *Journal for Language Technology and Computational Linguistics* (to appear).

Alle Projekte wurden gefördert von der Deutschen Forschungsgemeinschaft (DFG), Fördernummern DO 544/5-1/2, DI 1558/1-1/2 und DI 1558/4-1.