58. Jahrestagung des Leibniz-Instituts für Deutsche Sprache vom 15. bis 17. März 2022
(als Online-Konferenz)

KORPORA IN DER GERMANISTISCHEN SPRACHWISSENSCHAFT – MÜNDLICH, SCHRIFTLICH, MULTIMEDIAL

METHODENMESSE

Mittwoch, 16. März 2022, 15:45 Uhr bis 17:45 Uhr

Das Referenzkorpus Altdeutsch.
Ein tiefenannotiertes Mehrebenenkorpus für Forschung und Lehre

Stephanie Jandt, Gohar Schnelle, Phuong Thao Tran
Humboldt-Universität zu Berlin, Institut für deutsche Sprache und Linguistik

Das Referenzkorpus Altdeutsch (ReA, aktuelle Version 1.1, Donhauser 2015) ist ein tiefenannotiertes Korpus der gesamten althochdeutschen und altsächsischen Textüberlieferung und als diachrones Sprachstufenkorpus Teil des Verbunds ‚Deutsch Diachron Digital – Referenzkorpora zur deutschen Sprachgeschichte'. Die einzelnen Textüberlieferungen sind auf Basis handschriftennaher Editionen in Form von Subkorpora oder - bei kleineren Texten - in Form von Dokumenten repräsentiert, wobei die zahlreichen lateinabhängigen Texte (bspw. Tatian, Isidor, Murbacher Hymnen) gemeinsam mit der lateinischen Vorlage aufgenommen wurden. Im ReA werden insgesamt 24 Subkorpora repräsentiert, die ungefähr 535.00 Token[1] umfassen. Die Gesamtanzahl setzt sich u.a. aus den Token der althochdeutschen Texte (ca. 241.000 Token[2)] und der lateinischen Textgrundlagen (ca. 132. 000 Token[3]) zusammen.
Als tiefenannotiertes Mehrebenenkorpus enthält das ReA eine Primärtextebene („edition") und eine Normalisierungsebene („text"), die durch eine Segmentierung der historischen Schreibung eine Grundlage für die weiteren linguistischen Annotationen bildet (siehe Abbildung 1). Den Kern der grammatischen Annotationen bilden lexikalische („lemma", „translation", „lang", „posLemma"), morphologische („inflectionClassLemma", „inflection"), syntaktische („clause", „pos") und textstrukturelle („document", „verse", „chapter", „line", „rhyme") Ebenen. Relevante Informationen der jeweiligen Texte werden in den Metadaten annotiert: zeitliche und dialektale Einordnung, Textform und inhaltliche Einordnung, Verhältnis zu einer lateinischen Vorlage u.e.m.
Eine Besonderheit des Korpus´ ist der komplementäre analytische Bezug der grammatischen Annotationsebenen: „posLemma" und „inflectionClassLemma" enthalten Informationen zum Lemma, „pos", „inflectionClass", „inflection" beziehen sich auf den Gebrauch des Lemmas im Kontext (siehe Abbildung 2). Durch die Verknüpfung lemmabezogener und auf den belegbezogenen Annotationen kann gebrauchsbasierte Variation leicht identifiziert werden, z.B. Substantivierungen. Das besondere Potential, dass diese analytischen Ebenen für die forschungsrelevante Suchanfragen liefern, möchten wir gern in unserem Videobeitrag präsentieren.

---

[1] https://korpling.org/annis3/?id=1a8aee45-fc94-4d04-8c09-d05470d9b63a
[2] https://korpling.org/annis3/?id=38c73ca5-3d49-46a7-8475-5f87ace1cd96
[3] https://korpling.org/annis3/?id=9da36a0a-cedc-431e-93a9-225337f95444

Das ReA wird als Ressource bereits vielseitig in der Forschung (u.a. Catasso 2021, Coniglio et al. 2021, Flick 2020, Jäger 2016, Luxner 2020, Mittmann 2020, Petrova 2020, Schnelle 2018, Zeige 2019) und in der sprachhistorischen universitäten Lehre (Schnelle 2019, Schnelle et al. 2021) genutzt. So stellt z.B. das ‚Lesekorpus Altdeutsch' (Mitmann & Plate 2019) eine auf philologische Textarbeit ausgerichtete Visualisierung bereit. Das seit Juli 2021 laufende Projekt „D4" (Deutsch-Diachron-Digitale Didaktik, Gohar Schnelle) erstellt auf Basis des Referenzkorpus Lern- und Forschungsmaterialien für die Universität, darunter Tutorials für die Korpusnutzung, Anleitungen für die Einbindung von Korpora in den sprachhistorischen Unterricht sowie Visualisierungen empirischer Daten (siehe Abbildung 3).
Das ReA ist auf dem Laudatio-Repositorium (Guescini & Odebrecht 2020) veröffentlicht und archiviert. Alle Referenzkorpora des DDD-Verbunds sind über die Such- und Visualisierungsplattform ANNIS (Krause 2019) weltweit kostenfrei und browserbasiert zugänglich. Annotation und Abfrage folgen gemeinsamen Standards (z.B. HiTS Dipper et al 2013). Die Plattform bietet eine Abfrage einfacher und verschränkter Suchparameter, eine transparente Visualisierung der Suchergebnisse eine Editionsansicht der Texte, sowie Analyse und Exportfunktionen.
Durch die Tiefenannotation und die umfassenden Metadaten bietet das RA einen spannenden Einblick in die Variabilität und Wandelbarkeit sprachlicher Strukturen und gemeinsam mit den Referenzkorpora des DDD-Verbunds ein Fenster in die deutsche Sprachgeschichte. Darüber hinaus macht es auch die Bearbeitung zahlreicher empirischer Forschungsfragen aus Sprach- und Literaturwissenschaft und anderen textbezogenen Wissenschaften möglich.

Literatur:

Catasso, Nicholas (2021): How theoretical is your (historical) syntax? Towards a typology of Verb-Third in Early Old High German. Journal of Comparative Germanic Linguistics 24/1, S. 1–48.

Coniglio, Marco; C. De Bastiani; R. Hinterhölzl; T. Weskott (2021): In the right mood, in the right place: on mood and verb placement in Old Germanic subordinate clauses. Journal of Historical Syntax 5, S.1-27.

Dipper, Stefanie; Karin Donhauser; Sonja Linde; Stefan Müller; Klaus-Peter Wegera (2013): "HiTS: Ein Tagset Für Historische Sprachstufen Des Deutschen." *Journal for Language Technology and Computational Linguistics* 28 (1), 85–137.

Donhauser, Karin (2015): Das Referenzkorpus Altdeutsch. Das Konzept, die Realisierung und die neuen Möglichkeiten. In: Jost Gippert und Ralph Gehrke (Hg.): Historical Corpora. Challenges and Perspectives. Tübingen: Narr,S 35–49.

Donhauser, Karin; Gippert, Jost; Lühr, Rosemarie: ddd-ad. Version 1.1. Online verfügbar unter http://hdl.handle.net/11022/0000-0007-C9C7-6 .

Flick, Johanna. 2020. Die Entwicklung des Definitartikels im Althochdeutschen Eine kognitiv-linguistische Korpusuntersuchung. Language Science Press. (Empirically Oriented Theoretical Morphology and Syntax 6)

Guescini, Rolf; Odebrecht, Carolin (2020): Laudatio Repository - Long-term Access and Usage of Deeply Annotated Information: Digitale Medien, Computer- und Medienservice der Humboldt-Universität zu Berlin.

Jäger, Agnes (2016): Vergleichskasus im Althochdeutschen. In: Neri, Sergio/Schuhmann, Roland/Zeilfelder, Susanne (Hrsg.): "dat ih dir it nu bi huldi gibu" Linguistische, germanistische und indogermanistische Studien. Rosemarie Lühr gewidmet. Wiesbaden: Reichert, S. 193–208.

Krause, Thomas (2019): ANNIS: A graph-based query system for deeply annotated text corpora. Dissertation. Humboldt-Universität zu Berlin, Berlin.

Luxner, Bernhard (2020): Die althochdeutschen Adjektive auf -aht(i)/-oht(i). Eine diachron- und synchron-vergleichende Untersuchung. Dissertation. Friedrich-Schiller-Universität Jena

Mittmann, Roland (2020): Zur althochdeutschen Zeit- und Dialektgliederung. Eine computergestützte Untersuchung auf Grundlage der textlichen Überlieferung. Hamburg: Baar. (Studien zur historisch-vergleichenden Sprachwissenschaft 15)

Mittmann, Roland; Ralf Plate (2019): Das ‚Referenzkorpus Altdeutsch' als Lesekorpus. Grammatisch annotierte und mit Wörterbüchern verknüpfte Texte für Lehre und Selbststudium, in: Bleier, Roman/Fischer, Franz/Hiltmann, Torsten/Viehhauser, Gabriel/Vogeler, Georg (Hg.): Digitale Mediävistik (Das Mittelalter 24/1), 173–187. https://titus.uni-frankfurt.de/lea/

Petrova, Svetlana (2020): Embedded V2 in Old High German. A corpus study. In Rethinking Germanic V2, Rebecca Woods und Sam Wolfe (Hgg.), 555 – 573. Oxford: Oxford University Press.

Schnelle, Gohar (2019): *Corpus based teaching of Old High German. An experience report.* Digital Approaches to Teaching Historical Languages (DAtTeL), Humboldt Universität zu Berlin, Jacob-und-Wilhelm-Grimm-Zentrum – Auditorium, Berlin, March 28-29, 2019. Vortrag.

Schnelle, Gohar (2018): Funktional bedingte Variation in der Evangelienharmonie Otfrids von Weißenburg.Eine methodische Annäherung an eine variationistische korpusbasierte Registerstudie des Althochdeutschen. Masterarbeit an der Humboldt-Universität zu Berlin. https://doi.org/10.18452/19250

Schnelle, Gohar; Svetlana Petrova; Birgit Herbers (2021): Das Referenzkorpus Altdeutsch (ReA) in der universitären Lehre. Sprachgeschichte und Bildung
13. Jahrestagung, Pädagogische Hochschule Tirol, 20. bis 22. September 2021. Vortrag.

Zeige, Lars Erik (2019): Die Präpositionalgruppe im ältesten Deutsch. Korpus, Grammatik, Darstellung. 2. Bd. Habilitationsschrift, Humboldt-Universität zu Berlin.

edition	nu	noh	uuerchum	deononte	uueraltí	kelaubuliugan		cote
text	nu	noh	uuerchum	deononte	uueraltí	kelaubu	liugant	cote
lemma	nu	noh	werk	dionon	weralt	gilouba	liogan	got
posLemma	ADV	ADV	NA	VV	NA	NA	VV	NA
pos	ADV	ADV	NA	VVPS	NA	NA	VVFIN	NE
inflectionClassLemma			A_NEUT	WK2	I_FEM	O_FEM	ST2A	A_MASC
inflectionClass			A_NEUT	WK2	I_FEM	O_FEM	ST2A	A_MASC
inflection			PL_DAT	P_POS_MASC_PL_NOM_ST	SG_DAT	SG_DAT	IND_PRES_PL_3	SG_DAT
lang	goh	goh	goh	goh	goh	goh	goh	goh
clause	CP					CF_U_M		
document	(Althochdeutsche) Benediktinerregel							
chapter	1							
translation	nun	noch	Werk, Tat	(be)dienen, aufwarten	Zeit; Zeitalter, Welt	Glaube(nsbekenntnis)	lügen	Gott

Abbildung 1: Getrenntschreibung auf text- Ebene von nicht-lemmakonformer Zusammenschreibung auf **edition**- Ebene. Die Ebene **posLemma** und **inflectionClassLemma** enthalten Information zu den Lemmata *gilouba* und *liogan*. Die Annotation auf der Ebene **pos, inflectionClass** und inflection enthalten Analysen zum Beleg im Gebrauchskontext.

edition	thorfti	,	sô	thar	êr	ênig	thegno
text	thorfti	,	sô	thar	êr	ênig	thegno
lemma	thurban		so	thar	er	enig	thegan
posLemma	VM	$,	KO	ADV	ADV	ADJ	NA
pos	VMFIN	$,	KOUS	ADV	ADV	NA	NA
inflectionClassLemma	PRPR					A,O	A,I_MASC
inflectionClass	PRPR					A,O	A_MASC
inflection	SUBJ_PAST_SG_3					POS_MASC_SG_NOM_0	PL_GEN
lang	osx		osx	osx	osx	osx	osx
clause	CF_I_O		CF_I_Adv_Mod				
document	Heliand						
chapter	III						
verse	178						
rhyme	S1b						S2
translation	Veranlassung haben, brauchen, dürfen		wie	da, dort	vor, eher, früher	irgendein(er)	der Mann des Gefolges, Dienstmann

Abbildung 2: Auf der **posLemma**-Ebene wurde das Lemma *enig* Adjektiv annotiert. Diese Information bezieht sich auf das zugewiesene Lemma. Aus dem gebrauchbasierten Kontext auf der **text**-Ebene hat *enig* auf der **pos**-Ebene den Wert eines Nomen Appelativum.

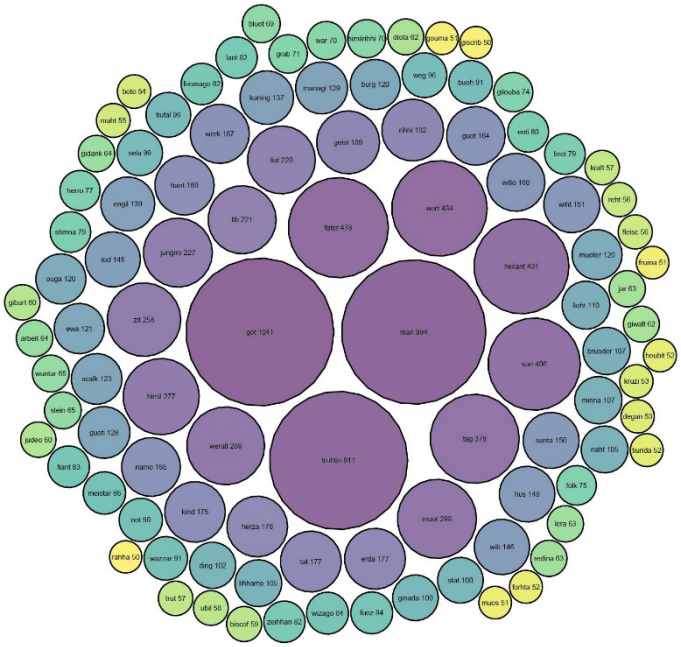

Abbildung 3Wordbubbles der 100 häufigsten Nomen und deren Frequenz im ReA-Korpus.