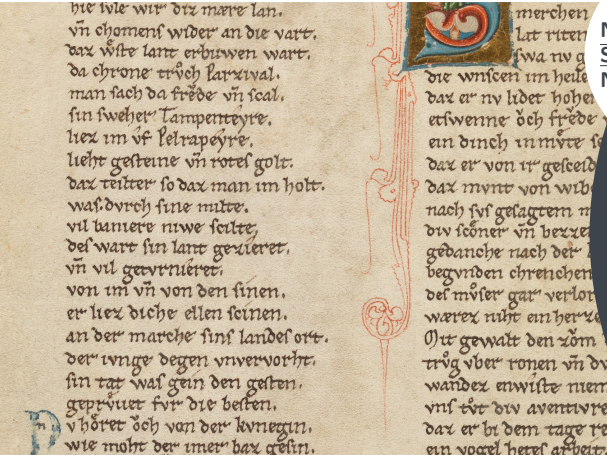
Universität Stuttgart

CRETA
CENTER FOR REFLECTED TEXT ANALYTICS



Nora Echelmeyer
Sarah Schulz
Nils Reiter

Ein PoS-Tagger für “das” Mittelhochdeutsche

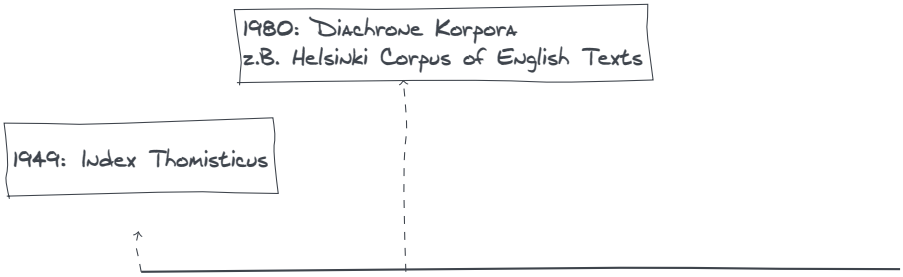
PoS-Tagging und warum's noch keiner gemacht hat

PoS-Tagging für historische Sprachen

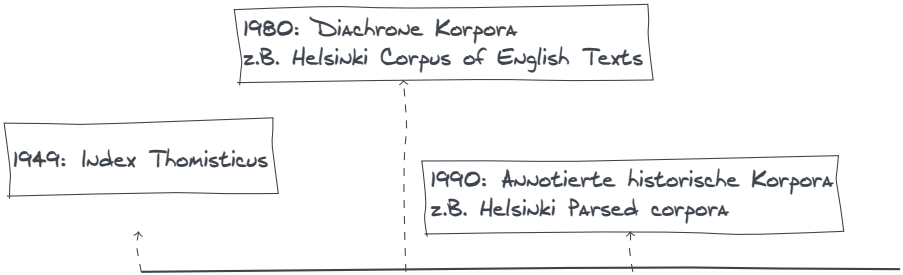
- Automatisches PoS-Tagging: auf deutschsprachigen Zeitungstexten ca. 95% Accuracy, auf Web-Texten 90-93% (Giesbrecht / Evert 2009)
- NHD-Tagger 45% Accuracy auf mittelhochdeutschen Daten
- Tagger für historische Sprachstufen des Deutschen:
  - Barteld et al. (2015): Mittelniederdeutsch
  - Dipper (2011) berichtet eine Accuracy von ca. 92% für zwei spezifische Modelle für die Dialekte Ober- und Mitteldeutsch (normalisiert)
  - Schulz / Kuhn (2016) beschreiben Ansätze zum PoS-Tagging eines spezifischen Textes

→ es fehlt ein generelles Modell für “das” MHD

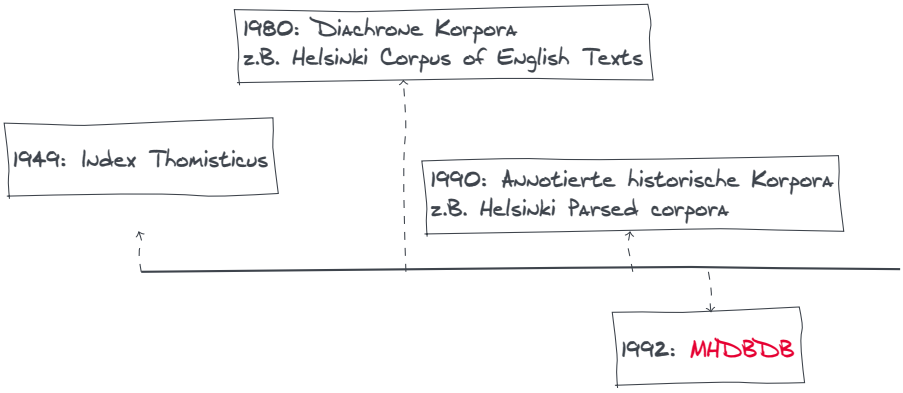
(Vgl. Dipper (2015): Annotierte Korpora für die Historische Syntaxforschung).



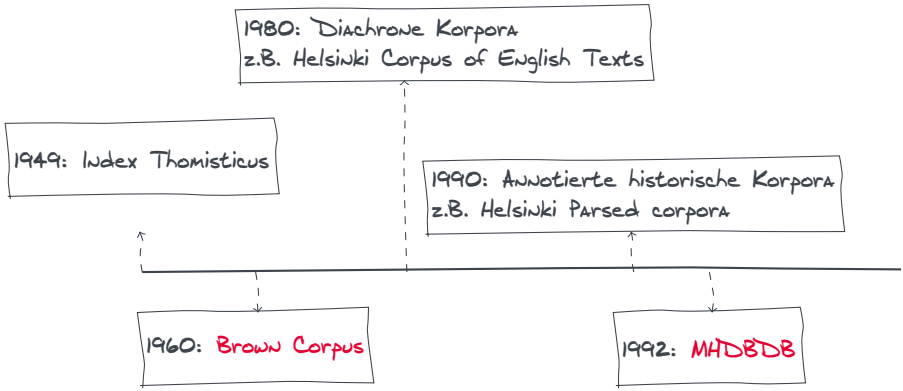
(Vgl. Dipper (2015): Annotierte Korpora für die Historische Syntaxforschung).



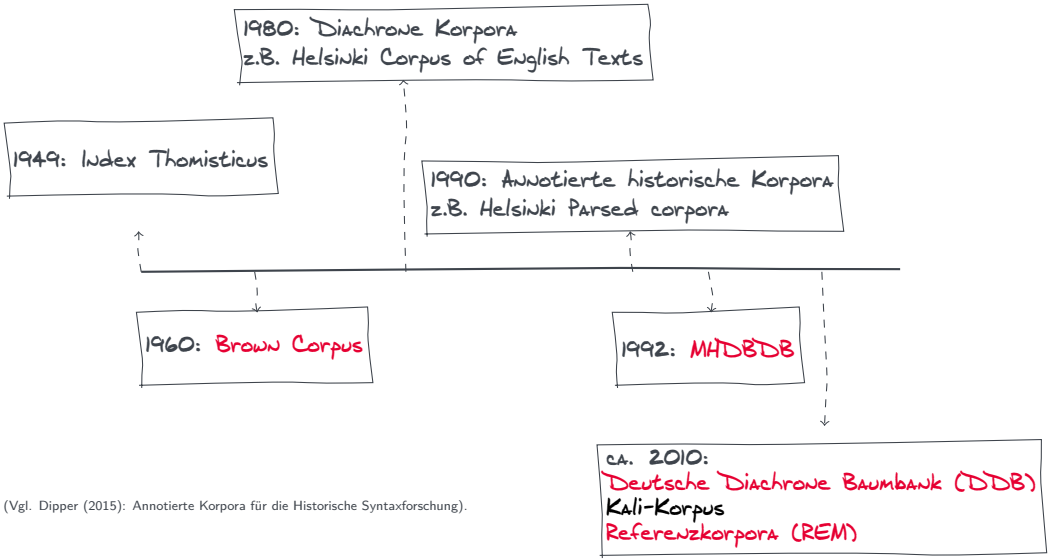
(Vgl. Dipper (2015): Annotierte Korpora für die Historische Syntaxforschung).



(Vgl. Dipper (2015): Annotierte Korpora für die Historische Syntaxforschung).



(Vgl. Dipper (2015): Annotierte Korpora für die Historische Syntaxforschung).

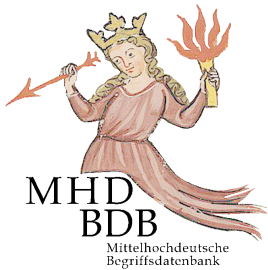


(Vgl. Dipper (2015): Annotierte Korpora für die Historische Syntaxforschung).

Daten oder: Wie man sich Daten beschafft

Mittelhochdeutsche Begriffsdatenbank (MHDBDB)

- 9,9 Millionen Tokens
- seit 1992 entwickelt
- Textkorpus: Zeitspanne von etwa vier Jahrhunderten (1100-1500)
- verschiedene dialektale Ausprägungen
- nahezu alle Gattungen und Textsorten (Heldenepik, Artusroman, Minnesang, Mären, Kochbücher u.a.)



MHDBDB

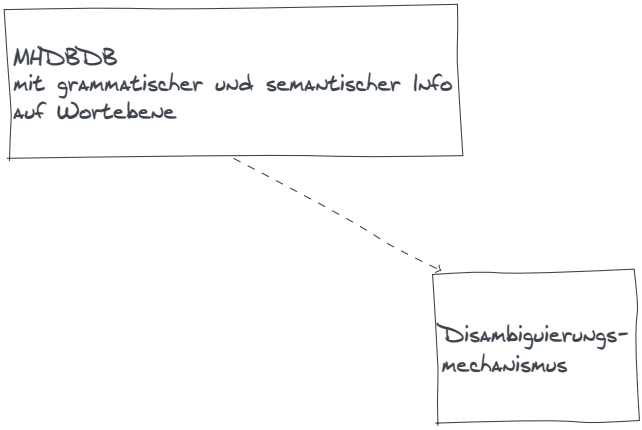
- Ziel: semantische und linguistische Fragestellung ermöglichen
- Annotation:
  - Tokenisierung
  - Lemmatisierung
  - grammatische Auszeichnungen (s. Tabelle)

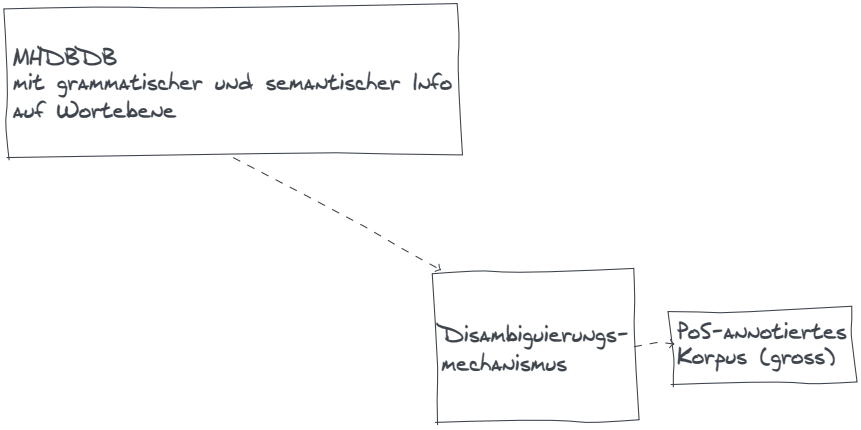
Kürzel	Name
NOM	Nomen
NAM	Name
ADJ	Adjektiv
ADV	Adverb
ART	Artikel
DET	Determinante
POS	Possessivpronomen
PRO	Pronomen
PRP	Präposition
NEG	Negation
NUM	Numeral
CNJ	Konjunktion
GRA	Gradationspartikel
IPA	Interrogativpartikel
VRB	Verb
VEX	Hilfsverb
VEM	Modalverb
INJ	Interjektion
CPA	Komparativpartikel
DIG	Zahl (Digit)

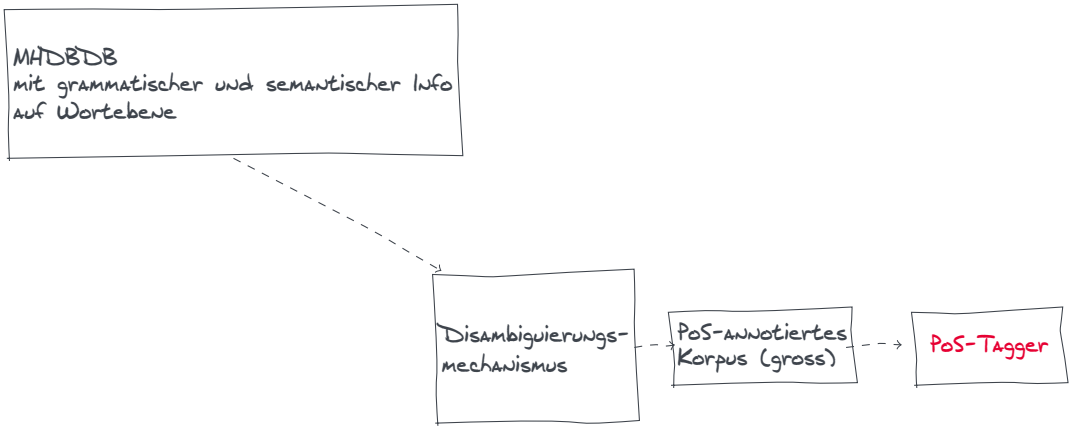
MHDBDB: Datenbank - kein annotiertes Korpus

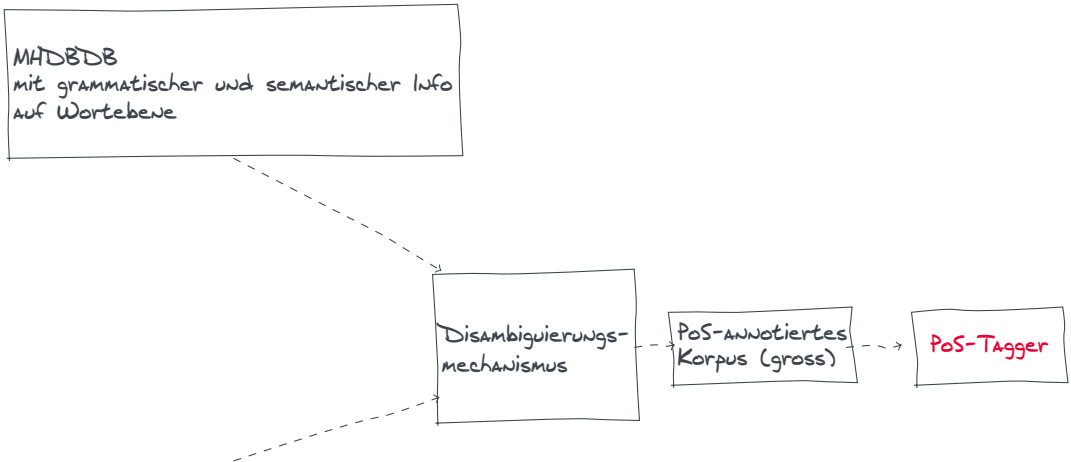
- nicht vollständig annotiert
- nicht disambiguiert: Kontext eines Wortes bleibt unberücksichtigt
- Daz: CNJ|ART
  1. Daz (ART) edel kint hât mir verjehen, daz (SCONJ) ez in troume sî geschehen. (Das edle Kind hat mir gesagt, dass es im Traum geschehen sei.)
  2. Daz (PRON) sage ich iu vür ungelogen. (Das sage ich euch wahrhaft.)
- Mehrinformationen enthalten
  - morphologische Zusammensetzung, z.B: unheil NEG|NOUN

MHDBDB
mit grammatischer und semantischer Info
auf Wortebene


MHDBDB
mit grammatischer und semantischer Info
auf Wortebene
Disambiguierungs-
mechanismus


MHDBDB
mit grammatischer und semantischer Info
auf Wortebene
Disambiguierungs-
mechanismus
PoS-annotiertes
Korpus (gross)


MHDBDB
mit grammatischer und semantischer Info
auf Wortebene
Disambiguierungs-
mechanismus
PoS-annotiertes
Korpus (gross)
PoS-Tagger


MHDBDB
mit grammatischer und semantischer Info
auf Wortebene
Disambiguierungs-
mechanismus
PoS-annotiertes
Korpus (gross)
PoS-Tagger

Universal Dependency Tagset

Kürzel	Name
ADJ	Adjektiv
ADP	Adposition
ADV	Adverb
AUX	Hilfsverb
CONJ	Konjunktion
DET	Artikel
INTJ	Interjektion
NOUN	Nomen
NUM	Nummeralie
PART	Partikel
PRON	Pronomen
PROPN	Eigenname
PUNCT	Satzzeichen
SCONJ	subordinierende Konjunktion
SYM	Symbol
VERB	Verb
X	andere

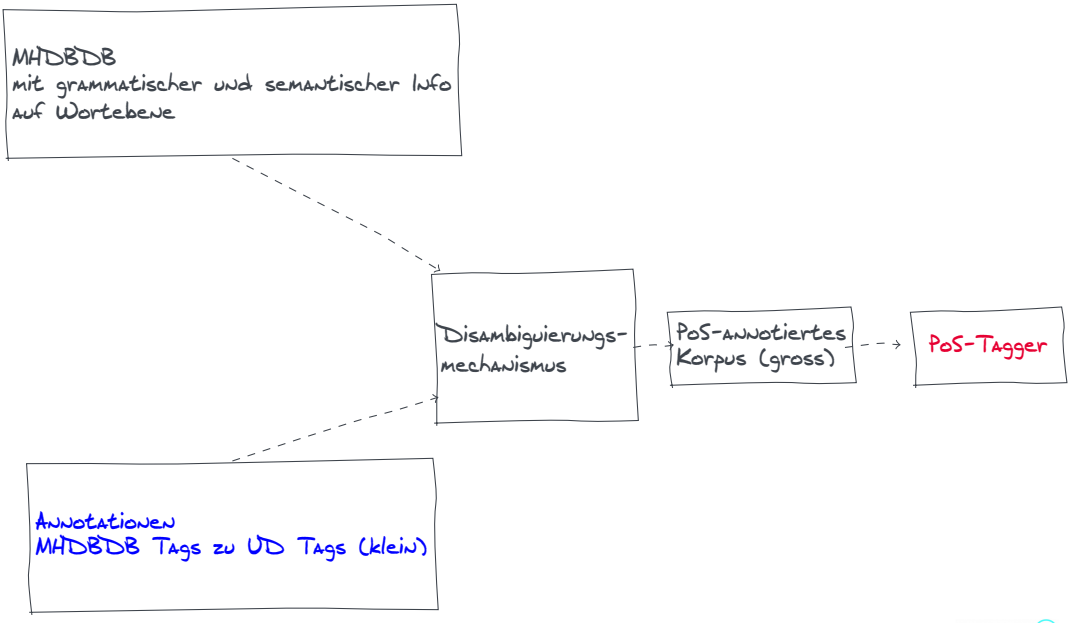
Erweitert um Kombinationstags für Wörter wie *irs* (ihr es) PRON+PRON

Automatische Disambiguierung

Token	MHDBDB Tag	Lemma
ein	NUM|ART	ein
maere	ADJ|NOM	mære
ich	PRO	ich
iu	PRO	sie
wil	VEM|VRB	wellen
niuwen	VRB	niuwen
das	CONJ|ART	daz
saget	VRB	sagen
von	PRP	von
grôzen	0	grôzen
truiwen	0	triuwen

Automatische Disambiguierung

Token	MHDBDB Tag	Lemma	UD Tag
ein	NUM|ART	ein	DET
maere	ADJ|NOM	mære	NOUN
ich	PRO	ich	PRON
iu	PRO	sie	PRON
wil	VEM|VRB	wellen	AUX
niuwen	VRB	niuwen	VERB
das	CONJ|ART	daz	PRON
saget	VRB	sagen	VERB
von	PRP	von	ADP
grôzen	0	grôzen	ADJ
truiwen	0	triuwen	NOUN


MHDBDB
mit grammatischer und semantischer Info
auf Wortebene
Disambiguierungs-
mechanismus
PoS-annotiertes
Korpus (gross)
PoS-Tagger
Annotationen
MHDBDB Tags zu UD Tags (klein)

Manuelle Annotation

3

Kontextabhängige Annotation

- DET vs. PRON:
  - DET: attribuierend, auch Possessiv-, Demonstrativpronomen: dirre âventiure, mîn bruoder
  - PRON: subsituierend: der mîne; daz was getân

Kontextabhängige Annotation

- DET vs. PRON:
  - DET: attribuierend, auch Possessiv-, Demonstrativpronomen: dirre âventiure, mîn bruoder
  - PRON: subsituierend: der mîne; daz was getân
- ADV:
  - auch adverbial gebrauchte Adjektive: guoter kneht - der im guot gedinet hat

Kontextabhängige Annotation

- DET vs. PRON:
  - DET: attribuierend, auch Possessiv-, Demonstrativpronomen: dirre âventiure, mîn bruoder
  - PRON: subsituierend: der mîne; daz was getân
- ADV:
  - auch adverbial gebrauchte Adjektive: guoter kneht - der im guot gedinet hat
- NOUN:
  - auch substantivierte Adjektive, Numeralia, Verben: die drî, die schoene, daz lachen

Problemfälle

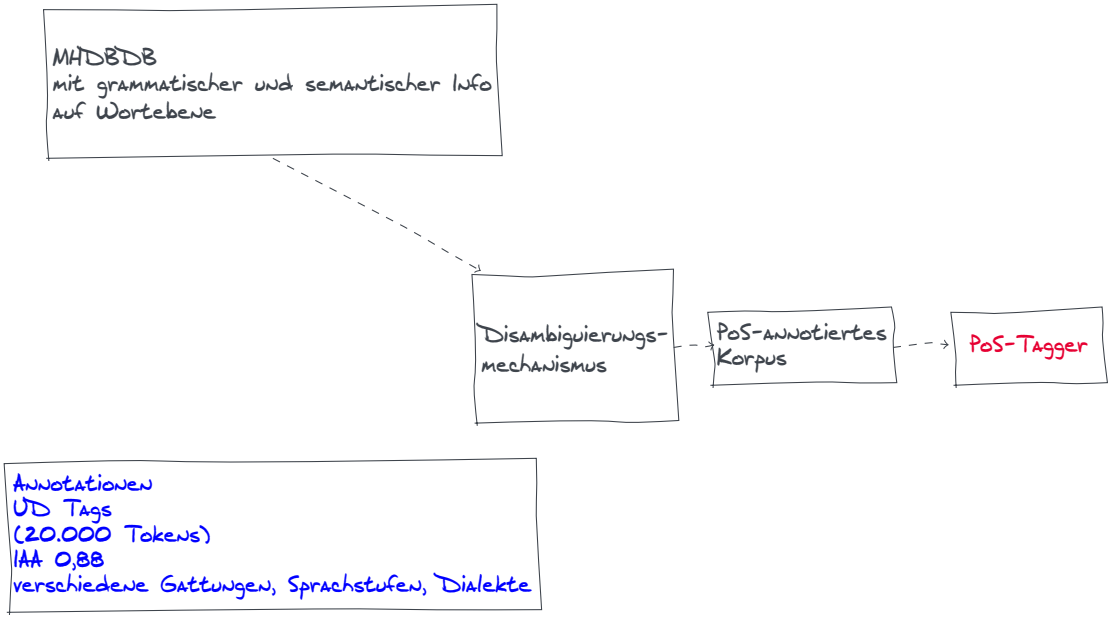
- Trennschärfe von DET und ADJ (z.B. viele, alle, einige)
  - ADJ: die vielen Frauen
  - DET: einige Frauen

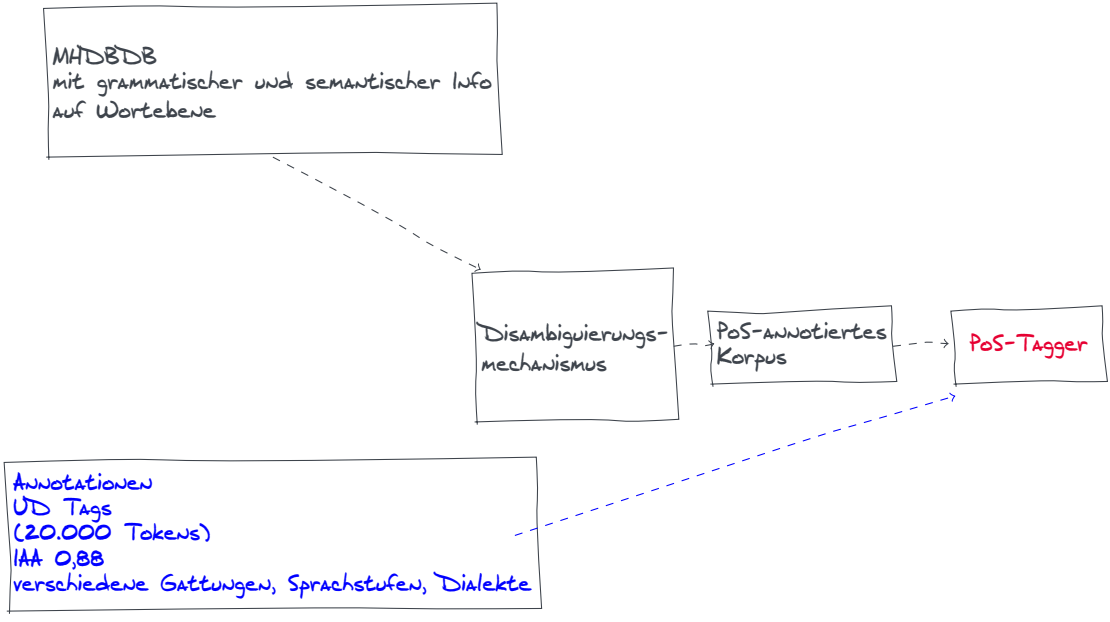
Problemfälle

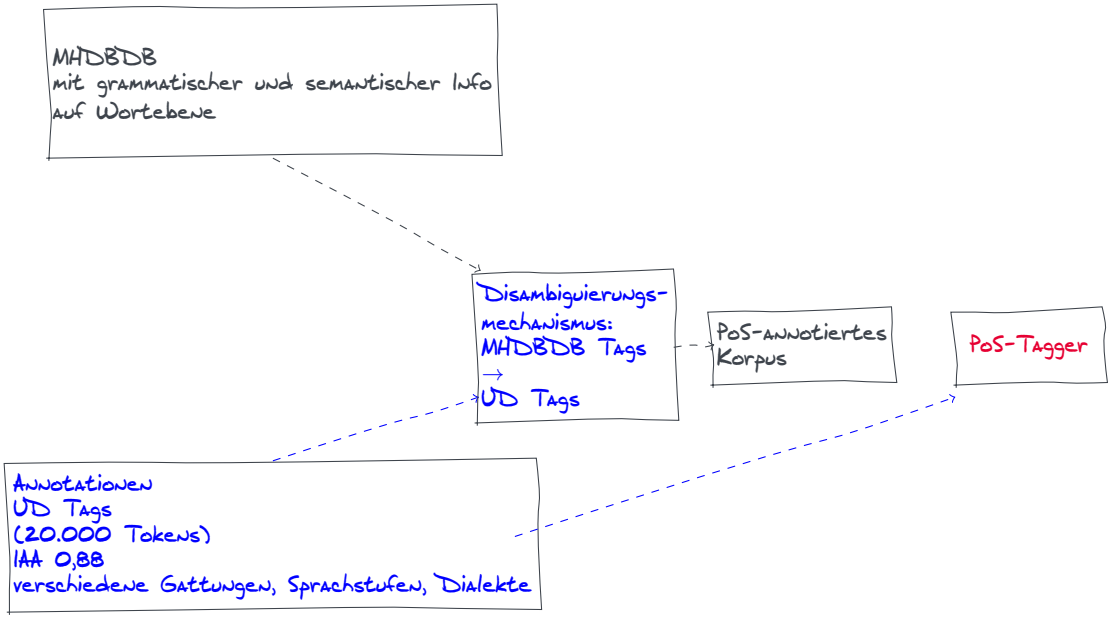
- Trennschärfe von DET und ADJ (z.B. viele, alle, einige)
  - ADJ: die vielen Frauen
  - DET: einige Frauen
- Im Mhd. können mehrere DET zusammentreten
  - der bruoder mîn

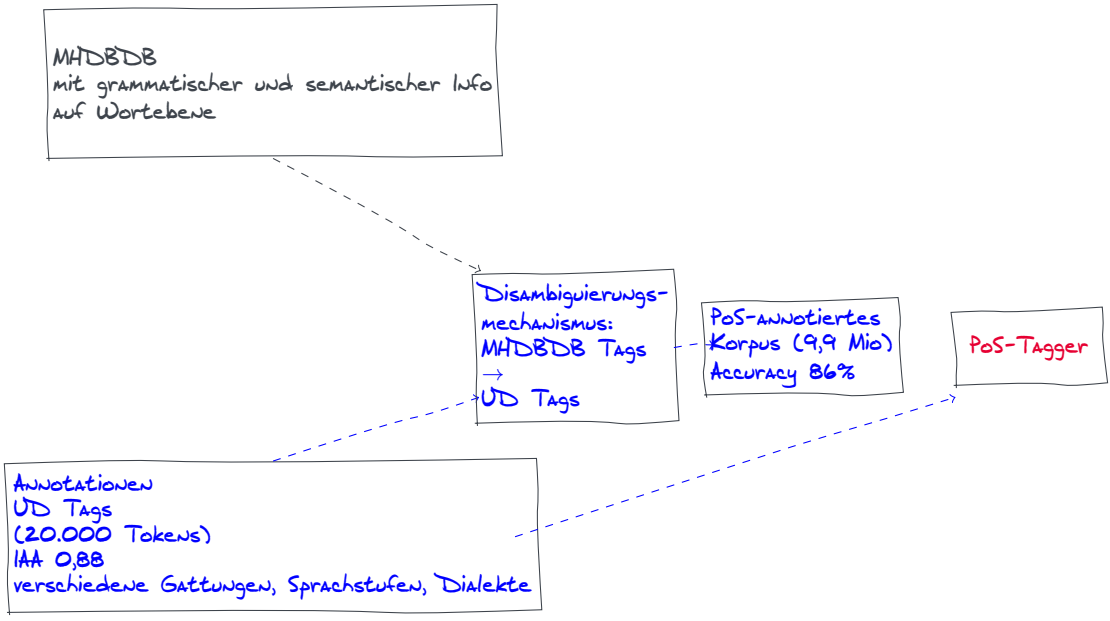
Problemfälle

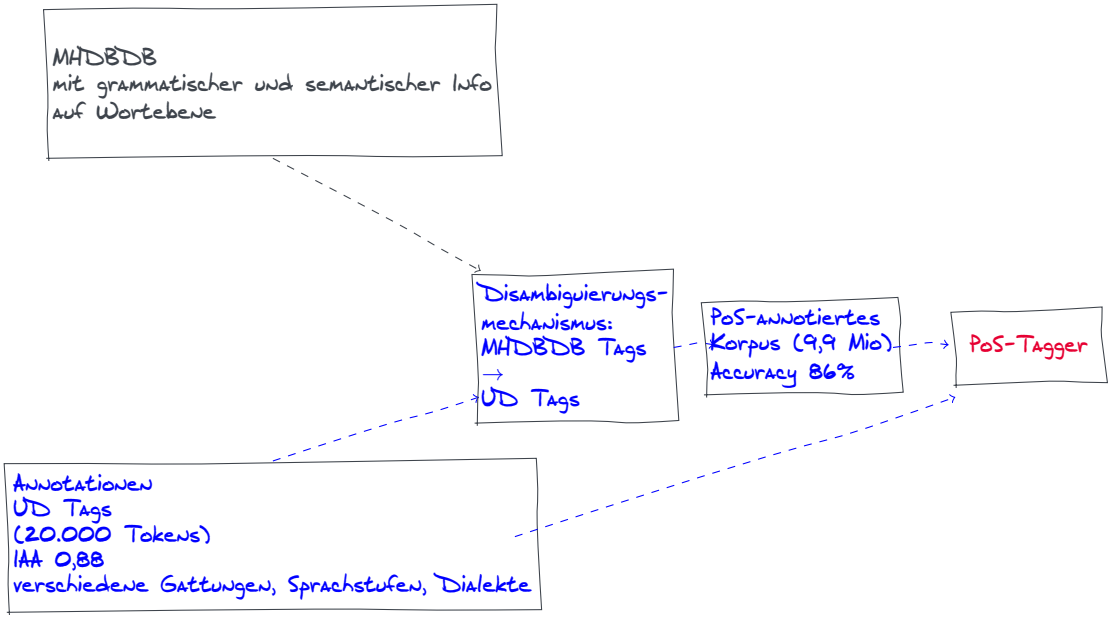
- Trennschärfe von DET und ADJ (z.B. viele, alle, einige)
  - ADJ: die vielen Frauen
  - DET: einige Frauen
- Im Mhd. können mehrere DET zusammentreten
  - der bruoder mîn
- Noch nicht lexikalisierte o. grammatikalisiert Formen:
  - sît daz = ADP und PRON
  - aber: seitdem = SCONJ


MHDBDB
mit grammatischer und semantischer Info
auf Wortebene
Disambiguierungs-
mechanismus
PoS-annotiertes
Korpus
PoS-Tagger
Annotationen
UD Tags
(20.000 Tokens)
IAA 0,88
verschiedene Gattungen, Sprachstufen, Dialekte


MHDBDB
mit grammatischer und semantischer Info
auf Wortebene
Disambiguierungs-
mechanismus
PoS-annotiertes
Korpus
PoS-Tagger
Annotationen
UD Tags
(20.000 Tokens)
IAA 0,88
verschiedene Gattungen, Sprachstufen, Dialekte


MHDBDB
mit grammatischer und semantischer Info
auf Wortebene
Disambiguierungs-
mechanismus:
MHDBDB Tags
→
UD Tags
PoS-annotiertes
Korpus
PoS-Tagger
Annotationen
UD Tags
(20.000 Tokens)
IAA 0,88
verschiedene Gattungen, Sprachstufen, Dialekte


MHDBDB
mit grammatischer und semantischer Info
auf Wortebene
Disambiguierungs-
mechanismus:
MHDBDB Tags
→
UD Tags
PoS-annotiertes
Korpus (9,9 Mio)
Accuracy 86%
PoS-Tagger
Annotationen
UD Tags
(20.000 Tokens)
IAA 0,88
verschiedene Gattungen, Sprachstufen, Dialekte


MHDBDB
mit grammatischer und semantischer Info
auf Wortebene
Disambiguierungs-
mechanismus:
MHDBDB Tags
→
UD Tags
PoS-annotiertes
Korpus (9,9 Mio)
Accuracy 86%
PoS-Tagger
Annotationen
UD Tags
(20.000 Tokens)
IAA 0,88
verschiedene Gattungen, Sprachstufen, Dialekte

Let’s train that thing

4

Experimente

- Ziele
  a. Vergleich manuell annotierte Daten vs. automatisch annotierte Daten
  b. Vergleich wenige hochqualitative Daten vs. viele Daten
- Baseline: neuhochdeutsches TreeTagger-Modell
- Modell 1: TreeTagger trainiert mit manuell annotierten Daten
  16k Tokens
- Modell 2: TreeTagger trainiert mit automatisch disambiguierten Daten
  16K Tokens
- Modell 3: TreeTagger trainiert mit automatisch disambiguierten Daten
  9,9M Tokens

Ergebnisse

	Precision	Recall	F-Score	Accuracy
Baseline	40,3	35,4	37,7	45,4
Modell 1 (kleines Trainingsset, manuell annotiert)	86,0	80,3	83,0	87,0
Modell 2 (kleines Trainingsset, autom. disambiguiert)	84,8	68,8	76,0	84,7
Modell 3 (großes Trainingsset, autom. disambiguiert)	91,2	79,6	85,0	90,9

Fehleranalyse

PoS-Klassen, die mit $> 90\%$ richtig getaggt werden:

- geschlossene Klassen
  - CONJ 4,5%
  - DET 5,1%
  - ADP 6,2%
- Inhaltswörter, offene Klassen
  - NOUN 6,6%
  - VERB 8,6%
  - PROPN 9,7%

Fehleranalyse

Häufige Fehlerquellen

- Kombinierte Tags 53,3%
- NUM 42,4%
  - DET/ADJ
- PART 33,7%
  - ADV/ADP
- ADJ 27,1%
  - ADV/NOUN/VERB

Fazit

5

PoS-Tagger für “das” MHD?

- PoS-Tagger für “das” Mittelhochdeutsche,
  → nahezu universell einsetzbar (Genres, Sprachstufen, Dialekte)
- Bereitstellung des Modells (Modell 3) als TreeTagger-Modell
  - `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`
- Bereitstellung des Modells (Modell 3) über eine Online-Benutzeroberfläche
  - `http://clarin05.ims.uni-stuttgart.de/mhdtt/index.html`

CRETA
CENTER FOR REFLECTED TEXT ANALYTICS

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Quellen

- Barteld, Fabian / Schröder, Ingrid / Zinsmeister, Heike (2015): Unsupervised regularization of historical texts for POS tagging, in: Proceedings of the 4th Workshop on Corpus-based Research in the Humanities (CRH) 312 www.slm.uni-hamburg.de/germanistik/personen/zinsmeister/downloads/barteld-etai-2015.pdf [letzter Zugriff 23.08.2016].
- Cohen, Jacob (1960): A coefficient of agreement for nominal scales, in: Educational and Psychological Measurement 20: 3746.
- Dipper, Stefanie (2015): Annotierte Korpora für die Historische Syntaxforschung. Anwendungsbeispiele anhand des Referenzkorpus Mittelhochdeutsch, in: Zeitschrift für Germanistische Linguistik 43: 516563.
- Dipper, Stefanie (2011): Morphological and Part-of-Speech Tagging of Historical Language Data: A Comparison, in: Journal for Language Technology and Computational Linguistics 26: 2537 (= Proceedings of the TLT-Workshop on Annotation of Corpora for Research in the Humanities 2012) www.jlcl.org/2011_Heft2/2.pdf [letzter Zugriff 23.08.2016].
- Dipper, Stefanie / Donhauser, Karin / Klein, Thomas / Linde, Sonja / Müller, Stefan / Wegera, Klaus-Peter (2013): HiTS: ein Tagset für historische Sprachstufen des Deutschen, in: Journal for Language Technology and Computational Linguistics 28: 85137 www.jlcl.org/2013_Heft1/5Dipper.pdf [letzter Zugriff 23.08.2016].

Quellen

- Giesbrecht, Eugenie / Evert, Stefan (2009): Is Part-of-speech tagging a Solved Task? An Evaluation of POS Taggers for the German Web as Corpus, in: Proceedings of the 5th Web as Corpus Workshop (WAC5) `www.stefan-evert.de/PUB/GiesbrechtEvert2009_Tagging.pdf` [letzter Zugriff 23.08.2016].
- Nivre, Joakim / de Marneffe, Marie-Catherine / Ginter, Filip / Goldberg, Yoav / Haji, Jan / Manning, Christopher D. / Mc Donald, Ryan / Petrov, Slav / Pyysalo, Sampo / Silveira, Natalia / Tsarfaty, Reut / Zeman, Daniel (2016): Universal Dependencies v1: A Multilingual Treebank Collection, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016) 16591666 `www.petrovi.de/data/lrec16.pdf` [letzter Zugriff 23.08.2016].
- Schiller, Anne / Teufel, Simone / Stöckert, Christine / Thielen, Christine (1999): Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset), Universität Stuttgart / Tübingen `www.sfs.uni-tuebingen.de/resources/stts-1999.pdf` [letzter Zugriff 23.08.2016].
- Schmid, Helmut (1995): Improvements in Part-of-Speech Tagging with an Application to German, in: Proceedings of the ACL SIGDAT-Workshop 4750 http://www.cis.uni-muenchen.de/schmid/tools/TreeTagger/data/tree-tagger2.pdf [letzter Zugriff 23.08.2016].
- Schulz, Sarah / Kuhn, Jonas (2016): Learning from Within? Comparing PoS Tagging Approaches for Historical Text, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016) 43164322 `www.lrec-conf.org/proceedings/lrec2016/pdf/1237_Paper.pdf` [letzter Zugriff 23.08.2016].

Quellen

- Spoustová, Drahomira / Haji, Jan / Raab, Jan / Spousta, Miroslav (2009): Semi-supervised training for the averaged perceptron POS tagger, in: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009) 763771 www.aclweb.org/anthology/E09-1087 [letzter Zugriff 23.08.2016].
- Zeldes, Amir / Ritz, Julia / Lüdeling, Anke / Chiarcos, Christian (2009): ANNIS: a search tool for multi-layer annotated corpora, in: Proceedings of the Corpus Linguistics Conference (CL 2009) 2023 `http://edoc.hu-berlin.de/oa/conferences/reS4Xo05sncZc/PDF/29i8VTIzYfT3M.pdf` [letzter Zugriff 23.08.2016].

Internetquellen

- Mittelhochdeutsche Begriffdatenbank: `http://mhdbdb.sbg.ac.at/`
- Referenzkorpus Mittelhochdeutsch: `http://referenzkorpus-mhd.uni-bonn.de`
- Wörterbuchnetz (Uni Trier): `http://woerterbuchnetz.de/`
- TreeTagger (Uni München): `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`
- Mittelhochdeutscher PoS-Tagger (Uni Stuttgart): `http://clarin05.ims.uni-stuttgart.de/mhdtt/index.html`; `http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/PoS_Tag_MHG.html`

Anhang I

PoS-Klassen, die Fehlerquellen sind:

- Kombinierte Tags: 53,3%
  - z.B. enkunde VERB (statt PART+VERB); siz PRON (statt PRON+PRON)
- NUM: 42,4%
  - DET (57%) oder ADJ (14%): zêhen vrouwen clâre, die drî ritter
- PART: 33,7%
  - ADV (42%) oder ADP (18%): z.T. gleiche Form (uf gên - dar ûf - uf die burc)
- ADJ: 27,1%
  - VERB (24%) : v.a. Partizipien (z.B. mîn herze ist versêrt)
  - ADV (29%): gleiche Form (z.B. guote knehte - guot gedinet)
  - NOUN (25%): gleiche Form (z.B. die schoene - daz schoene wîp)

Anhang II

http://clarin05.ims.uni-stuttgart.de/mhdtt/index.html

Uns ist in alten mæren wunders vil geseit
von helden lobebæren, von grôzer arebeit,
von freuden, hôchgezîten, von weinen und von klagen,
von küener recken strîten muget ír nu wunder hœren
sagen.

Uns	PRON
ist	VERB
in	ADP
alten	ADJ
mæren	NOUN
wunders	NOUN
vil	ADV
geseit	VERB
von	ADP
helden	NOUN
lobebæren	ADJ
,	PUNCT
von	ADP
grôzer	ADJ
arebeit	NOUN

Anhang III

Universal Dependency Tagset

Tag		Anmerkungen	Beispiele
ADJ	adjective	vorangestellt, nachgestellt, Partizipien	der ritter guot; daz elder kint; roemisch lant; der ander man
ADP	adposition	Prä-, Post- und Zirkumposition	mit dem swerte; gein Nantes; âne ir schulde
ADV	adverb	auch adverbial gebrauchte Adjektive und relativischer Gebrauch	der ritter lîdenlîche leit; so sprach der künec; rehte liebe im nie geschach; hôret, swie er ze strîte quam
AUX	auxiliary verb	Hilfs- und Modalverben	ich muoz ir dienen; die sint enterbet; ir habet ez von mir gehôrt
CONJ	coordinating conjunction	nebenordnend	ritter unde diep; zwei teil oder mêr; denne ich welle jehen
DET	determiner	Artikel (bestimmt und unbestimmt); attribuierende Demonstrativ-, Possessiv- und Relativpronomen	ein maere; der ritter guot; diz bîspel; dirre âventiure; ir triuwe; mîn bruoder; dehein man; in welhem lande
INTJ	interjection		ouwê; ach!
NOUN	noun	auch substantivierte Adjektive, Verben, Numeralia	diu vrouwe; duc Orilus; rîche und arme; die drî; daz singen
NUM	numeral	nur Kardinalzahlen	die drî ritter
PART	particle	Negationspartikel; abgetrennter Verbzusatz; zu (mit Inf.); Vergleichspartikel; Abtönungspartikel	daz enweiz ich niht; lâzet allez trüren abe; daz ist swere ze halten; snêwîz als ein harm
PRON	pronoun	Personal-, Relativ-, Reflexivpronomen; substituierende Possessiv-, Indefinit-, Demonstrativ-, Interrogativpronomen	er lac tôt; Isenhärten, der den lîp verlôs; die kuenen heten sich berâten; der sîne sprach; da was nieman; allez, daz ich habe; man saget; diz was dô getân; swaz er gebôt
PROPN	proper noun	Eigennamen (auch mehrteilig)	Parzivâl; Orilus de Lalander; Nantes
SPUNCT	punctuation	Satz-beendende Zeichen	. : ! ?
PUNCT	punctuation	alle sonstigen Satzzeichen	, ; < > „ / ( ) usw.
SCONJ	subordinating conjunction	unterordnend	Er sagete daz Isenhart küneclîch bestatet wart; sît er an mir ist sus verzagt; ob mich gelücke wil bewarn
SYM	symbol		
VERB	verb	alle Vollverben	Er lac tôt; wir suln kurzwîl phlegen; er hat ein grôz her; ir reht was vernomen
X	other		