RÄUME - GRENZEN - ÜBERGÄNGE:
5. Kongress der Internationalen Gesellschaft für Dialektologie des Deutschen
10. - 12. September 2015
Universität Luxemburg

GRAMMATISCHE ANNOTATION HISTORISCHER TEXTE – EIN TAGSET FÜR DAS MITTELNIEDERDEUTSCHE

DFG Deutsche Forschungsgemeinschaft

UHH Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

SARAH IHDEN
sarah.ihden@uni-hamburg.de

KATHARINA DREESSEN
katharina.dreessen@uni-hamburg.de

FABIAN BARTELD
fabian.barteld@uni-hamburg.de

INGRID SCHRÖDER
ingrid.schroeder@uni-hamburg.de

Ablauf

- ReN-Projekt
  - Rahmendaten
  - Korpusdesign und -erstellung
- Besonderheiten des Mittelniederdeutschen
- Historisches-Niederdeutsch-Tagset (HiNTS)
  - Basis
  - PoS-Tagging
  - Flexionsmorphologisches Tagging
- Resümee

Rahmendaten

Rahmendaten

Referenzkorpus Mittelniederdeutsch/ Niederrheinisch (1200-1650)

- Teil des ‚Korpus historischer Texte des Deutschen', gemeinsam mit den Referenzkorpora
  - ***Altdeutsch***
  - ***Mittelhochdeutsch***
  - ***Frühneuhochdeutsch***
- Basis sind Handschriften, Drucke und Inschriften

Ziel des Projekts

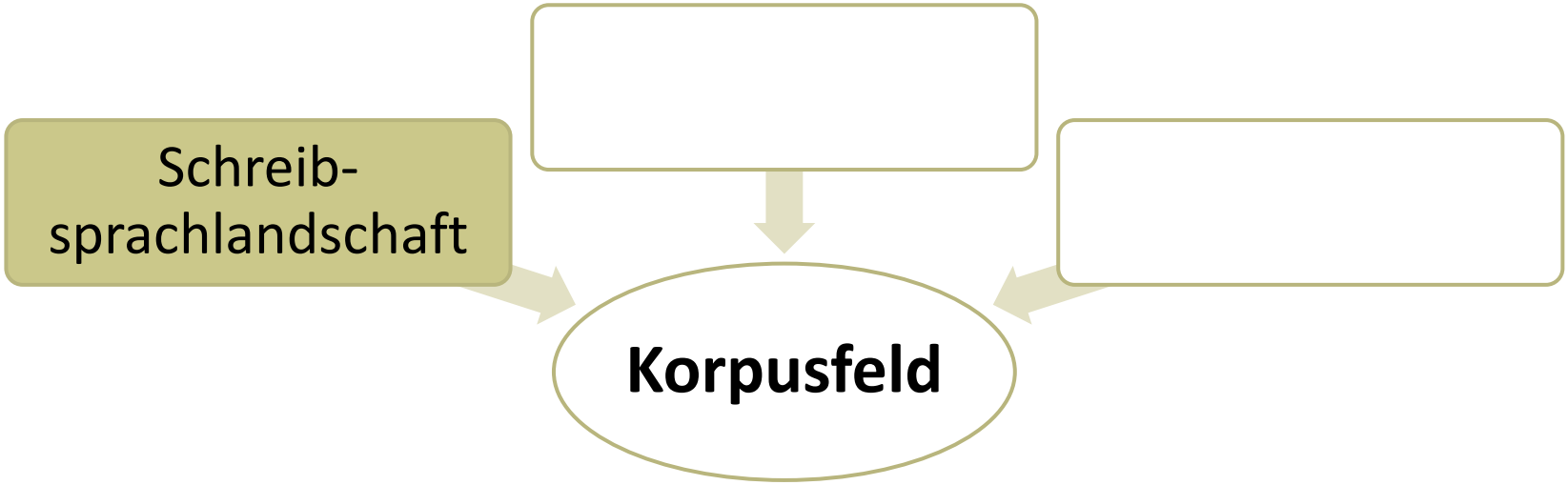
- digitale Veröffentlichung diplomatisch transkribierter, lemmatisierter und grammatisch annotierter Texte

Nutzen des Referenzkorpus

- verschafft Einblicke in die Sprach- und Textkultur des niederdeutschen und niederrheinischen Raums
- ermöglicht sprachwissenschaftliche Analysen

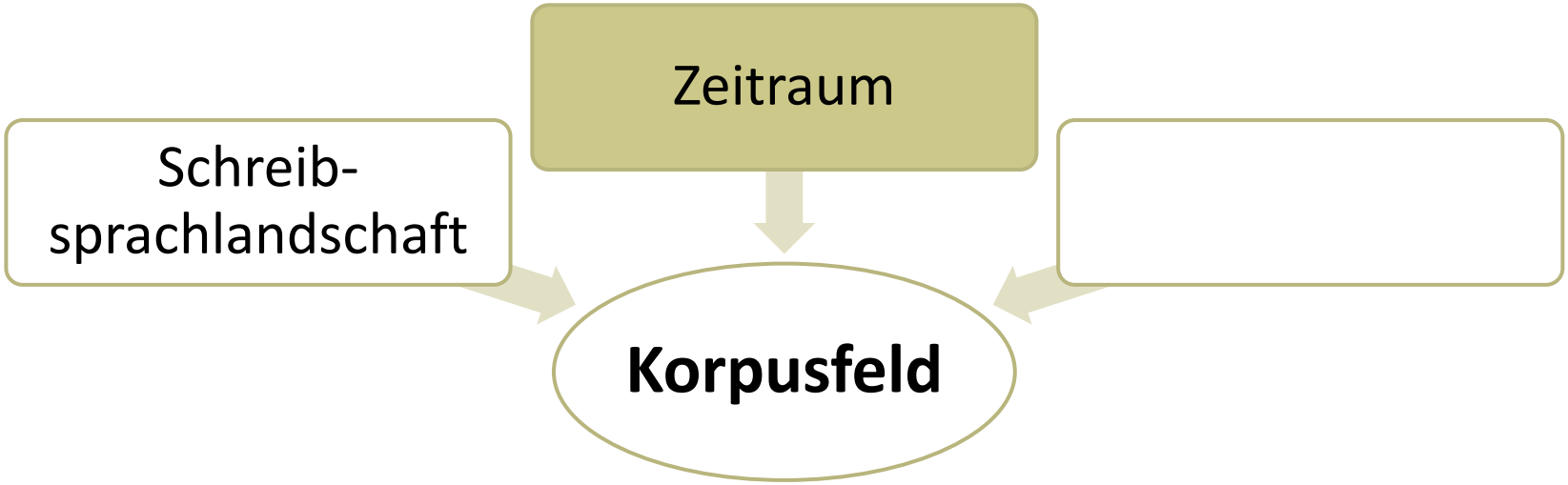
Korpusdesign

Korpusdesign



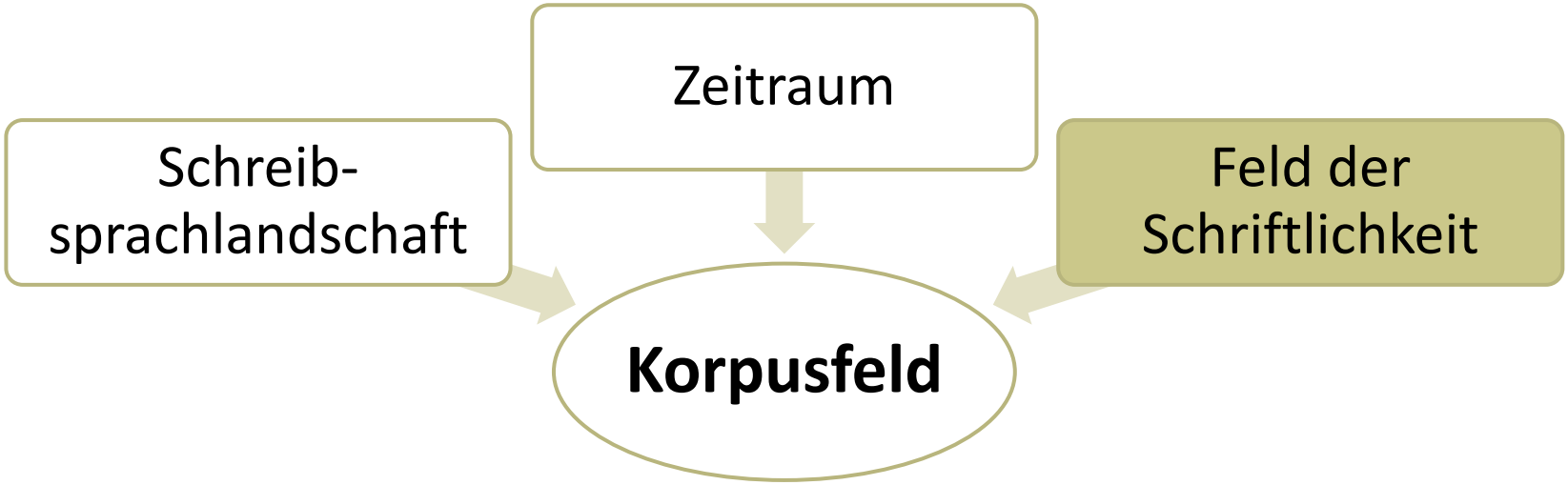
Standort Hamburg	**Standort Münster**
Nordniedersächsisch	Lübeckisch
Ostelbisch	Westfälisch
Baltisch	Ostfälisch
Südmärkisch	Elbostfälisch
	Niederrheinisch

Korpusdesign



I: 1200-1300
II: 1301-1350
III: 1351-1400
IV: 1401-1450
V: 1451-1500
VI: 1501-1550
VII: 1551-1600
VIII: 1601-1650

Korpusdesign



Recht
Urkunden
Verwaltung
Wissensvermittlung
Religion (Kirchliches)
Literarische Texte
Private Schriftlichkeit und Korrespondenz
Inschriften

Korpuserstellung

Korpuserstellung

Phase 1: Textaufbereitung/ Transkription

Phase 2: Annotation
(PoS- und flexionsmorph. Tagging, Lemmatisierung)

Phase 3: Publikation
(Annis, TEI)

Nutzung

Sprachspezifische Besonderheiten

Sprachspezifische Besonderheiten

Wortartwechsel

to ***donde***$_{\text{VVINF}}$ *hebben*

oder

to ***donde***$_{\text{NA}}$ *hebben*

Syntaktische Ambiguität

dochter$_{\text{NA.Fem.Gen.Sg}}$ ***name***$_{\text{NA.Masc.Nom.Sg}}$

oder

dochtername$_{\text{NA.Masc.Nom.Sg}}$

Historisches-Niederdeutsch-Tagset (HiNTS)

Basis

Voraussetzungen

- low-resourced language
  - keine Tools zur automatischen Annotation vorhanden
  - keine Trainingsdaten für statistische Tools vorhanden
  - kaum elektronische Ressourcen vorhanden (z.B. Wörterbuch)

→ keine Standards für die Annotation mnd. Daten

- **POS-Tagset (inkl. Morphologie)**, Lemmainventar

Existierende Tagsets

STTS
(PoS + Morphologie)

Stuttgart-Tübingen-Tagset
(Schiller et al. 1999)

→ HiTS
(PoS)

Historisches Tagset
(Dipper et al. 2013)

Wortarten
STTS und HiTS

Nomen – N

Verb – V

Determinierer – **D** (Artikel – **ART**)

Adjektiv – ADJ

Pronomina – P (Pronominaladverb – **PAV**)

Kardinalzahlen – CARD

Adverbien – **ADV**, **AV**

Junktionen – KO

Appositionen – AP

Interjektion – ITJ

Partikel – PTK

Tag-Aufbau

- Die Hauptwortarten sind nach funktionalen und distributionellen Kriterien subklassifiziert

  Beispiel (HiTS):

  DDART

  Determinativ, definit, artikelartig

- Angabe einer Basiswortart

  Beispiel: *to **donde***$_{\text{NA < VVINF}}$ *hebben*

Anforderungen an ein MND-Tagset

Anforderungen (I)

1. Möglichst nah an existierenden Tagsets bleiben
   (POS: HiTS, Morphologie: STTS)

→ Aber:

2. Tags müssen anhand des konkreten Kontextes auswählbar sein
   (keine muttersprachliche Intuition)

- Beispiel (STTS):
  - PIAT (attribuierendes Indefinitpron., ohne Determinierer vorkommend)
    - *[etwas] Schokolade*

  vs.

  - PIDAT (attribuierendes Indefinitpron., mit Determinierer vorkommend)
    - *[solch] eine Frage*

→ dies ist erst als Ergebnis der Korpusauswertung entscheidbar

Anforderungen (II)

3. Ambiguitäten kennzeichnen
   - Beispiel:
     - KON vs. KOU nicht immer entscheidbar > HiTS hat KO*

4. Aber: so spezifisch wie möglich sein
   - STTS kennt in der Morphologie nur eine eindeutige Zuweisung oder ambig (*)

Konsequenz

→ daher weder STTS noch HiTS ohne Modifikation geeignet

→ Eigene angepasste Version von HiTS:

HiNTS (PoS + Morphologie)

Historisches-Niederdeutsch-Tagset (HiNTS)

Part of Speech (PoS) Flexionsmorphologie

Historisches-Niederdeutsch-Tagset (HiNTS)

Part of Speech (PoS) Flexionsmorphologie

Abweichungen von HiTS

HiTS (Dipper et al. 2013: 36f.) unterscheidet zwischen:

- Tags mit **D...** (Determinativa) → 2 Unterkategorien: Typ + Position, z.B.:
  - *[dise]*$_{\text{DDA}}$ *rede*

    (Determinativ, definit/demonstrativ, attributiv)
  - *[dizze]*$_{\text{DDS}}$ *ist ein anphanclich zít*

    (Determinativ, definit/demonstrativ, substituierend)
- Tags mit **P...** (Pronomen): stets substituierend → nur eine Unterkategorie: Typ, z.B.:
  - *man*$_{\text{PI}}$ (Pronomen, indefinit)

→ Problem:

vorab als Pronomen klassifiziertes Lexem in anderer Distribution (nicht substituierend), z.B. *man* vor einem Substantiv i.S.v. ‚irgendein'$_{\text{(Bsp. konstruiert)}}$

Abweichungen von HiTS

→ lexembezogene Vorannahmen in HiNTS vermeiden → 2 Gruppen:

- Tags mit **D...** → attributiv, z.B.:
  - *[dyt]*$_{\text{DDA}}$ *ghut kanstu allene nyth ghe wynnen*

    (Determinativ, definit/demonstrativ, attributiv, vorangestellt)
- Tags mit **DP...** → substituierend, z.B.:
  - *dat my [nemant]*$_{\text{DPNEGS}}$ *kunne lyken*

    (Determinativ/Pronomen, negativ, substituierend)

↓

Nutzen: Ermittlung von Lemmata, die nur substituierend vorkommen und daher tatsächlich Pronomen sind

Qualitätssicherungsverfahren

Inter-Annotator-Agreement

- Anwendung des MAMA-Zyklus (vgl. Pustejovsky/ Stubbs 2012)
- 2 Inter-Annotator-Agreements mit je 2 Annotatoren
  - IAA I:
    - PoS-Taggings und flexionsmorphologisches Tagging
    - keine Lemmatisierung
    - ausschließlich manuelles Tagging
  - IAA II:
    - PoS-Taggings und flexionsmorphologisches Tagging
    - Lemmatisierung mittels Lemmaliste
    - halbautomatisches Tagging

Qualitätssicherungsverfahren

Inter-Annotator-Agreement PoS

Art der Abweichung	IAA I	IAA II	IAA III
Abweichungen aufgrund fehlender Regeln	8,9%	3,7%	↘
Abweichungen trotz bestehender Regeln	3,1%	5,7%	↘
unterschiedliches Textverständnis	1,0%	1,9%	→
gesamt	13,0%	11,3%	↘

Abweichungen trotz bestehender Regeln:

- Regelverstoß
- Folgefehler
- Annotation vergessen/ Aufmerksamkeit

Historisches-Niederdeutsch-Tagset (HiNTS)

Part of Speech (PoS) **Flexionsmorphologie**

Flexionsmorphologisches Tagset

- Die Art und der Umfang der Tags sind abhängig von PoS

Beispiele:

*Her marcus meyger myn leue **frunth** klagent **mack** my **nycht** baten [...]* (Brief v. A. Willeken, 1535)

frunth$_{\text{NA.Masc.Nom.Sg}}$ **mack**$_{\text{VMFIN.3.Sg.Pres.Ind}}$ **nycht**$_{\text{PTKNEG}}$

- STTS nutzt das Sternchen (*) für Ambiguitäten

→ Nachteile:

- keine Angabe der konkreten möglichen Werte (z.B. Dat. und Akk., aber nicht Gen.)
- Tendenz, eine Entscheidung herbeizuführen, vgl. TIGER : „[...] Nur wenn es nicht gelingt, im gegebenen Kontext dem Attribut einen eindeutigen Wert zuzuweisen, soll der Wert * zugewiesen werden.“ (TIGER-Morphologie-Annotationsschema 2015: 5)

→ für Nhd. möglich, für historische Sprachstufen problematisch, da zu interpretatorisch

Genusambiguität

Beispiel:

Dit is der sassen speyghel (Oldb. Ssp., Überschrift)

Token	PoS	Morphologie
Dit	DPDS	
is	VVFIN	
der	DDARTA	
sassen	NA	
speyghel	**NA**	**Masc-Neut.Nom.Sg**

Vgl. Barteld et al. (2014)

Qualitätssicherungsverfahren

Inter-Annotator-Agreement Morph.

Art der Abweichung	IAA I	IAA II	IAA III
Abweichungen aufgrund fehlender Regeln	8,2%	1,7%	↘
Abweichungen trotz bestehender Regeln	**11,6%**	**12,6%**	↘
unterschiedliches Textverständnis	0%	0,5%	→
gesamt	19,8%	14,8%	↘

Abweichungen trotz bestehender Regeln:

- Regelverstoß
- Folgefehler
- Annotation vergessen/ Aufmerksamkeit

Resümee

Resümee

- Entwicklung des **HiNTS** aufgrund sprachspezifischer Besonderheiten des Mittelniederdeutschen
- **Qualitätssicherungsverfahren** sind von hoher Wichtigkeit
- Inter-Annotator-Agreements zeigten:
  - HiNTS ist erfolgreich anwendbar
  - mit HiNTS sinkt der Grad der Interpretation
  - Abweichungen zwischen den Annotatoren haben unterschiedliche Ursachen
    - systematische Abweichungen lassen sich reduzieren
    - Routinierte Anwendung wird aufmerksamkeitsbedingte Fehler und Regelverstöße herabsetzen

Herzlichen Dank für Ihre Aufmerksamkeit!

Literatur

- Barteld, Fabian/ Ihden, Sarah/ Schröder, Ingrid/ Zinsmeister, Heike (2014): „Annotating descriptively incomplete language phenomena". In: *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*, August 2014, Dublin, S. 99-104. Online verfügbar unter: http://www.aclweb.org/anthology/W14-4915.
- 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*, Special Issue, 28(1), 85-137.
- Pustejovsky, James/ Stubbs, Amber (2012): *Natural Language Annotation for Machine Learning*. A Guide to Corpus-Building for Applications. Beijing [u.a.].
- Rehbein, Ines/ Hirschmann, Hagen/ Lüdeling, Anke/ Reznicek, Marc (2012): "Better tags give better trees – or do they?". In: *Linguistic Issues in Language Technology* (LILT). Volume 7, S. 1-18.
- Rehbein, Ines/ Schalowski, Sören (2013): „STTS goes Kiez – Experiments on Annotating and Tagging Urban Youth Language". In: *Journal for Language Technology and Computational Linguistics* (JLCL), Special Issue, 28(1), S. 199-227.
- Schiller, Anne/ Teufel, Simone/ Stöckert, Christine: Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset). Stuttgart, Tübingen 1999: Univ. Stuttgart, Univ. Tübingen
- TIGER Morphologie-Annotationsschema (2015). Auf: http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/TIGERCorpus/annotation/tiger_scheme-morph.pdf. [Zuletzt gesehen am 02.09.15]

Anhang

Qualitätssicherungsverfahren

Inter-Annotator-Agreement PoS

Art der Abweichung	IAA I	IAA II	IAA III
Abweichungen aufgrund fehlender Regeln	8,9%	3,7%	
Abweichungen trotz bestehender Regeln	3,1%	5,7%	
unterschiedliches Textverständnis	1,0%		
gesamt	13,0%		

IAA PoS
Abweichungen trotz bestehender Regeln

Art der Abweichung	IAA I	IAA II
Regelverstoß	78,6%	62,5%
Folgefehler	21,4%	23,2%
Annotation vergessen, Aufmerksamkeit	0%	14,3%
gesamt	100%	100%

Qualitätssicherungsverfahren

Inter-Annotator-Agreement Morph.

Art der Abweichung	IAA I	IAA II	IAA III
Abweichungen aufgrund fehlender Regeln	8,2%	1,7%	
Abweichungen trotz bestehender Regeln	11,6%	12,6%	
unterschiedliches Textverständnis	0%		
gesamt	19,8%		

IAA Morphologie

Abweichungen trotz bestehender Regeln

Art der Abweichung	IAA I	IAA II
Regelverstoß	35,6%	47,3%
Folgefehler	57,8%	25,5%
Annotation vergessen, Aufmerksamkeit	5,6%	17,3%
tagger-reproduzierte Fehler	-	10,0%
verschiedene	1,1%	0%
gesamt	100%	100%

IAA in anderen Korpora
(PoS)

- Rehbein et al. (2012, 8):
  - Annotation von Lernersprache mit dem STTS
  - Annotatorenübereinstimmung von 97,9 %
- Rehbein/ Schalowski (2013: 208)
  - Annotation des Kiezdeutsch-Korpus mit dem dafür erweiterten Tagset
  - Annotatorenübereinstimmung von 96,5 %