



Digital Diachron Deutsch

Ein Historisches Referenzkorpus für das Deutsche

Anke Lüdeling

Humboldt-Universität zu Berlin
zusammen mit
Stefanie Dipper
Lukas Faulstich
Ulf Leser
Thorwald Poschenrieder

Variationen über
DeutschDiachronDigital

SFB 631, Potsdam, 09.07.2004

DDD

- diachrones Korpus des Deutschen vom Althochdeutschen (≈ 800) bis zum Neuhochdeutschen (≈ 1900) für alle textbezogenen Forschungsrichtungen
- 13 deutsche Universitäten + internationale Kooperationspartner
- 5 Sprachstufengruppen und eine Gruppe Korpusarchitektur und -entwicklung (wir ☺)
- Laufzeit geplant 7 Jahre
- geplante Größe 60 M Wörter, davon 40 M Wörter annotiert

Ziele

- innovative Forschungsressource für breiten Anwenderkreis
 - cutting edge Methoden aus Korpuslinguistik und Informatik
 - alte Forschungsfragen können neu adressiert werden
 - neue Forschungsfelder
- philologisch gesicherter Referenzrahmen für empirische Forschung
 - Reproduzierbarkeit
 - Annotation auf vielen Ebenen
 - Entwicklung von empirischen Standards
- Sicherung eines wichtigen Teils des kulturellen Erbes



Variation & Standardisierung

- Variation
 - bei der Erstellung & Nutzung
 - in den Daten
 - in der Korpusarchitektur
 - Standardisierung
 - bei der Digitalisierung
 - bei der Annotierung
-



Variation bei der Erstellung & Nutzung

- verteiltes Arbeiten
 - unterschiedliche Traditionen
 - unterschiedliche Kompetenzen & Interessen
-



Variation & Standardisierung

- Variation
 - bei der Nutzung
 - in den Daten
 - in der Korpusarchitektur
 - Standardisierung
 - bei der Digitalisierung
 - bei der Annotierung
-



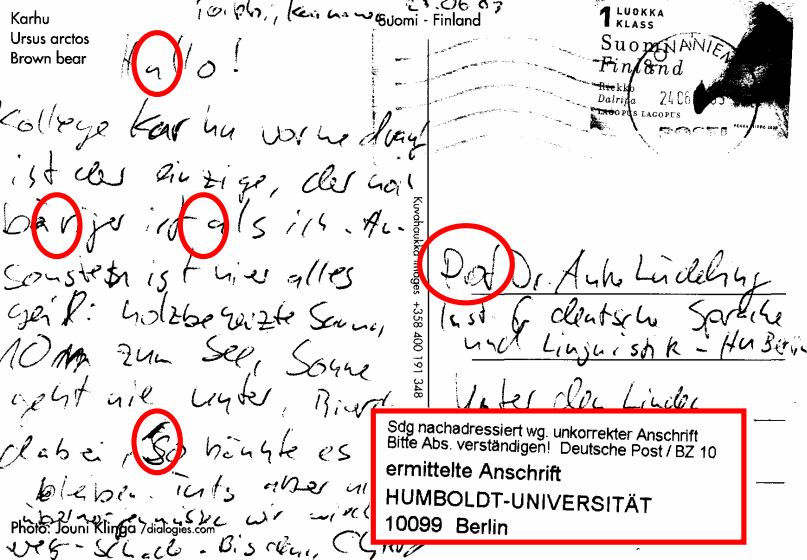
Variation & Standardisierung

- Variation
 - bei der Nutzung
 - in den Daten
 - innerhalb einer Sprachstufe
 - diachron
 - in der Korpusarchitektur
 - Standardisierung
 - bei der Digitalisierung
 - bei der Annotierung
-



Digitalisierung - Diplomazität

- Quelle – Handschrift/Originaldruck oder Edition
- Qualitätskontrolle (Kollationierung, Korrektur etc.)
- Entscheidung über Zeichenkodierung
 - Ligaturen
 - Unterschiede (f, s)
 - Spatien
 - Abkürzungen
 - Marginalien, Zusätze, Glossen
- soweit wie möglich Originale
- soweit wie möglich Alignierung mit Faksimiles
- gemeinsame Qualitätsstandards



Digitalisierung - Diplomazität

- Quelle – Handschrift/Originaldruck oder Edition
- Qualitätskontrolle (Kollationierung, Korrektur etc.)
- Entscheidung über Zeichenkodierung
 - Ligaturen
 - Unterschiede (f, s)
 - Spatien
 - Abkürzungen
 - Marginalien, Zusätze, Glossen
- hohe Diplomazität
- soweit wie möglich Originale
- soweit wie möglich Alignierung mit Faksimiles
- gemeinsame Qualitätsstandards



Swerlenrecht können wil•d~volge
dis buches lere.

Problem Tokenisierung

- Unterscheidung zwischen graphischem Wort und 'lexikalischem' Wort
- Kodierung von Wortteilen (Stammbildung)
- Abkürzungen

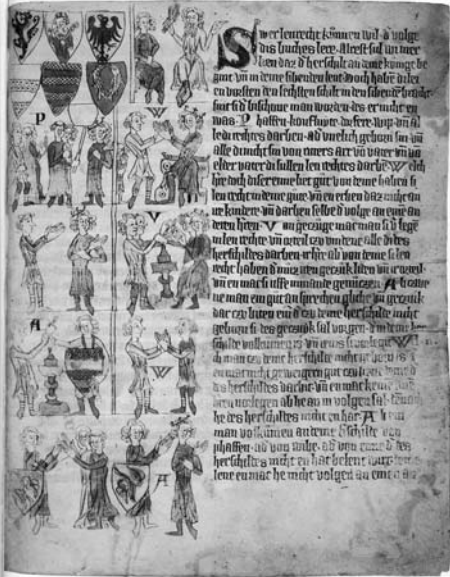
- stand off Architektur
- Zeichen als kleinste Einheit

Normalisierung - Lemmatisierung

- Problem: Normalisierung
mhd, fnhd *und*:
undi, vndi, unndi, vnndi, unti, unnti, vnti, vnnti, unde, unnde, unte, unnte, vnde, vnnde, vnnte, vnnte, ...
- innerhalb einer Sprachstufe
- über die Sprachstufen hinweg: Hyperlemmatisierung

Tagsets & Annotationsrichtlinien

- mindestens Wortart, Flexionsmorphologie, Syntax
- Probleme
 - unterschiedliche Fokussierung in traditioneller Grammatik (Wortarten eher flexionsmorphologisch definiert) und gängigen Tagsets wie STTS (Wortarten eher syntaktisch definiert)
traditionell: Flexionsklassen
STTS: substituierend - attribuierend
 - Veränderung von Kategorien über die Zeit
Dual im Althochdeutschen
 - Latein
- Ausgangspunkt STTS
- Unterspezifizierung/Ambiguitätserhaltung



Swerlenrecht können wil•d~volge
dis buches lere.

S	w	e	r	l	e	n	r	e	c	h	t	k	ü	n	n	e	n	w	i	l	•	d	~	v	o	l	g	e
sw	e	r	l	e	n	r	e	c	h	t		k	ü	n	n	e	n	w	i	l	•	d	~	v	o	l	g	e
wer																												
Pronomen																												
I																												

Verfahren

- keine rein statistischen Annotationsverfahren möglich, da zu wenig Text, zu wenig Standardisierung
- in vielen Fällen zu aufwendig, regelgeleitete Verfahren zu schreiben
- keine/kaum digitale Ressourcen (Lexika) verfügbar
- hohe Genauigkeit (das Wissen ist ja oft schon da! - ?)
- genaue Kenntnis der Sprachstufe, der Überlieferungssituation etc.

Umgang mit Variationen & Ambiguitäten

- typische moderne Textkorpora:
Ambiguitäten werden aufgelöst,
Variationen werden normalisiert,
Heuristiken
- alte Sprachstufen:
Variationen wichtig – müssen dokumentiert werden
Unterspezifikation/Ambiguitätserhaltung
Standardisierung auf zusätzlichen Ebenen

inhaltliche Anforderungen - Standardisierung

- gemeinsamer Qualitätsstandard
 - Quelle
 - Diplomtizität
- gemeinsame Headerinformationen (TEI)
 - Raum
 - Textsorte
 - Paläographie/Kodikologie
- gemeinsame strukturelle Annotation
 - graphisch
 - logisch
 - konfligierende Hierarchien

inhaltliche Anforderungen - Standardisierung

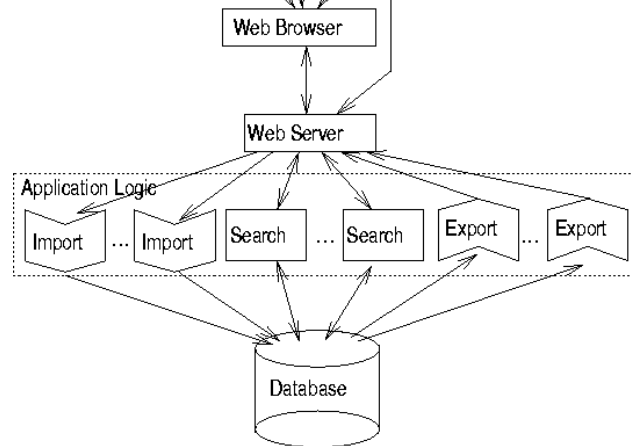
- gemeinsame positionelle Annotation
 - Ebenen
 - Tagsets und Richtlinien
- Lemmatisierung
 - innerhalb einer Sprachstufengruppe
 - über die Sprachstufengruppen hinweg - Hyperlemma
 - Multilingualität

Anforderungen - Flexibilität

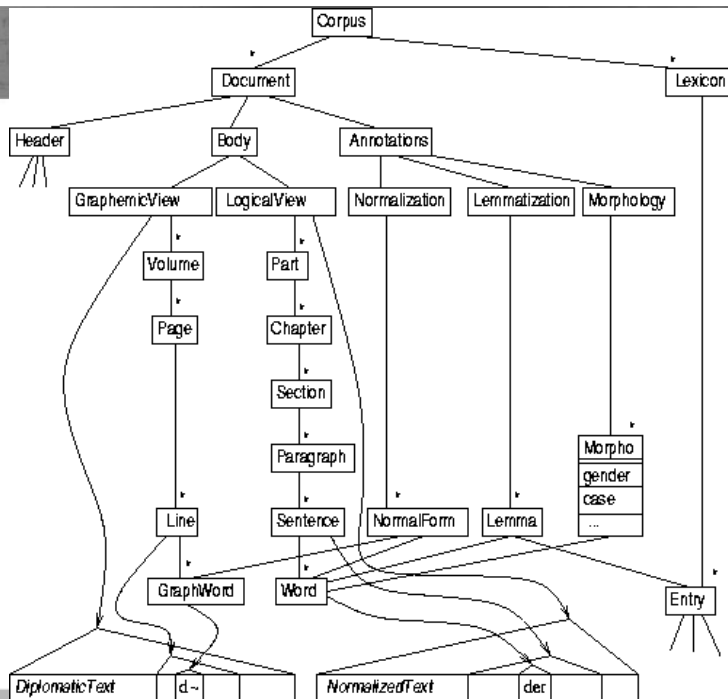
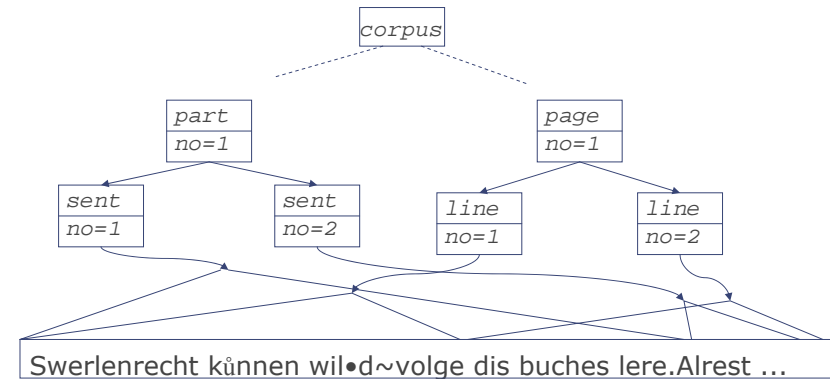
- verschiedene Texte können verschiedene Annotationsebenen haben
 - jeder Text hat Headerinformation und minimale strukturelle Annotation - Extensionskorpus
 - zusätzlich Wortart & Lemma - Kernkorpus
 - aligniert mit Faksimile oder Sounddatei - Präsentationskorpus
 - Multimodalität
 - zusätzlich: jeder Text kann weitere Annotationsebenen haben (Syntax, Informationsstruktur, ...)
- Texte und Annotationsebenen können jederzeit hinzugefügt werden

Systemarchitektur

- webbasierte Client-Server-Architektur
- Korpus ist in einer relationalen Datenbank gespeichert
- zusätzliche Client-Werzeuge zur Annotation, Suche, Präsentation etc.



das Datenmodell – konfligierende Hierarchien



Import & Export

- Export aus der Datenbank in XML für Präsentation
- Export aus der Datenbank in XML für externe Annotationstools
- Import von dem von Annotationstools erzeugten XML-Dateien in die Datenbank
- dh – verschiedene Formate für ganz unterschiedliche Anwendungen werden unterstützt



Zusammenfassung

- Ziel: diachrones Korpus des Deutschen
- höchste Konsistenz und gleichzeitig hohe Flexibilität
 - gemeinsame Qualitätsstandards,
gemeinsame Headerstrukturen,
gemeinsame Annotationsebenen mit Tagsets & Richtlinien
- Implementation
 - webbasierte Client-Server-Architektur, Relationale Datenbanken
 - Import & Export (XML)

Sie finden uns im Internet unter:
<http://korpling.german.hu-berlin.de/ddd/>



Digital
Diachron
Deutsch

Ein Historisches Referenzkorpus für das Deutsche