



Digital Diachron Deutsch

Ein Historisches Referenzkorpus für das Deutsche



Prof. Dr. Karin Donhauser
(Humboldt-Universität zu Berlin)

Projektvorstellung

Deutsch Diachron Digital
Ein Referenzkorpus für das Deutsche

Akademie der Wissenschaften
Berlin, den 29.10.2003



Die Initiative

- DDD ist eine fach- und institutionenübergreifende Initiative, in der sich Sprach- und Literaturhistoriker des Deutschen aus verschiedenen Universitäten zusammengeschlossen haben, um gemeinsam ein digitales Referenzkorpus für das Deutsche zu entwickeln, das verschiedenen Benutzeranforderungen Rechnung trägt.
- Mitglieder der Initiative:
Ulrike Demske, (Saarbrücken), Karin Donhauser (HU Berlin), Kurt Gärtner (Trier), Jost Gippert (Frankfurt), Ulrike Hass-Zumkehr (IDS Mannheim), Thomas Klein (Bonn), Rosemarie Lühr (Jena), Marga Reis (Tübingen), Hans Joachim Solms (Halle), Klaus Peter Wegera (Bochum)
Anke Lüdeling (HU Berlin), Ulf Leser (HU Berlin)
- Sprecher der Initiative
Karin Donhauser (HU Berlin) und Kurt Gärtner (Trier)



Die Ausgangssituation

- Ein spezielles Defizit: Es fehlt für das Deutsche ein historisches Referenzkorpus vergleichbar dem Penn-Helsinki-Korpus für das Englische
- Statt dessen: Zahlreiche Einzelkorpora mit unterschiedlichem Umfang und von unterschiedlicher Qualität
- Generelle Defizite:
 - kein (einheitlicher) Annotationsstandard
 - keine (einheitliche) Header-Information
 - unterschiedliche Grade an Diplomtizität
 - nur zum Teil öffentlich zugänglich
 - schlechte Such- und Recherchemöglichkeiten
 - unterschiedliche Abdeckung der verschiedenen Sprachstufen des Deutschen



Die Zielsetzung

Mit dem Aufbau eines Historischen Referenzkorpus des Deutschen verfolgt die Initiative im Wesentlichen drei Zielsetzungen:

- Sicherung und Zugänglichmachen eines wichtigen Teil des kulturelles Erbes Deutschlands und Europas
- Bereitstellung von zuverlässigen Textgrundlagen für sprach-, literatur- und kulturwissenschaftliche Forschungen
- Entwicklung eines modernen Forschungstools für eine Vielzahl sprach- und literaturwissenschaftlicher Fragestellungen



Die Leitlinien

- **Ökonomie:**

Kein vollständiger Neueinsatz: bereits vorhandene Ressourcen werden - soweit als möglich - integriert.

- **Qualität:**

Qualität verlangt Expertise. Um Qualitätsstandards zu gewährleisten müssen für alle Aufgabenbereiche Experten gewonnen werden.

- **Flexibilität:**

Das Korpus muss ausbaufähig und korrekturfähig sein, kumulatives Arbeiten muss möglich werden.



Die Leitlinien

- **Zugänglichkeit**
Das Korpus muss öffentlich zugänglich sein - für Wissenschaftler, aber auch für interessierte Laien.
- **Benutzerkreis**
Die Benutzerkreis sollte nicht zu eng gefasst werden. Kein Benutzerinteresse sollte exklusiv verfolgt werden.



Das technische Konzept

- Übernahme technischer Standards bei der Kodierung und Annotation der Texte (XML, TEI/CES, STTS; TIGER)
- Folgende Merkmale dagegen stellen Neuerungen gegenüber bestehenden historischen Textkorpora dar:
 - Kleinste Einheit
Zeitlinie: graphisches Zeichen
 - Standoff-Architektur
Jede Annotationsebene befindet sich in einer eigenen Datei
 - Multimodale Datenbank
Möglichkeit zur Verbindung von unterschiedlichen Datentypen
(z.B. Text, Bild, Ton)



Das technische Konzept

Verantwortlich für das technische Konzept:

- Anke Lüdeling / Korpuslinguistik (HU Berlin)
- Ulf Leser / Bioinformatik (HU Berlin)

Pilotprojekt unterstützt von der Senatverwaltung für
Wissenschaft und Forschung des Landes Berlin



Das inhaltliche Konzept

- Zeitliche Abdeckung

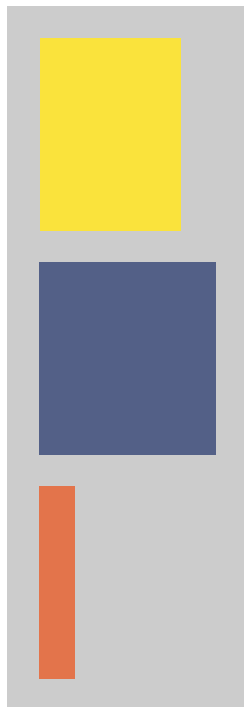
- Von den Anfängen der deutschen Überlieferung bis zur Gegenwart.
- Schwerpunkt: die älteren Sprachstufen des Deutschen (Althochdeutsch, Mittelhochdeutsch und Frühneuhochdeutsch)
- Für den Bereich des Neuhochdeutschen Kooperation mit dem DTA und dem IDS Mannheim

- Räumliche Abdeckung

Hochdeutsch, Niederdeutsch, [Jiddisch]

Das inhaltliche Konzept

- Korpusstruktur



Das Kernkorpus

Repräsentativ

Obligatorisch: Headerinformation; Standardannotation

Das Erweiterungskorpus

Nicht repräsentativ

Obligatorisch: Headerinformation; Basisannotation,

Das Präsentationskorpus

Exemplarisch

Faksimile; Tonf



Die Sprach(stufen)-Projekte

Althochdeutsch	Karin Donhauser (HU Berlin), Jost Gippert (Frankfurt), Rosemarie Lühr (Jena)
Mittelhochdeutsch	Kurt Gärtner (Trier), Thomas Klein (Bonn), Hans Joachim Solms (Halle)
Frühneuhochdeutsch	Ulrike Demske (Saarbrücken), Klaus Peter Wegera (Bochum)
Neuhochdeutsch	Ulrike Hass-Zumkehr (IDS Mannheim) > DTA
Niederdeutsch	Robert Peters (Münster), Ingrid Schröder (Hamburg), Heinrich Tiefenbach (Regensburg)
[Jiddisch	Tony Kroch/Beatrice Santorini (Pittsburg)]

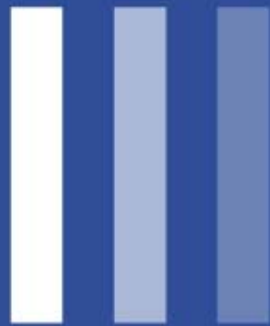


Zeitplan und Bearbeitungsstand

- Gründung der Initiative:
Juni 2002 (Tübingen)
- Erarbeitung des Konzeptes:
Juni 2002-November 2003
- Antragskonferenz:
10/11.12.2003 in Berlin

Sie finden uns im Internet unter:

<http://korpling.german.hu-berlin.de/ddd/>



Digital
Diachron
Deutsch

Ein Historisches Referenzkorpus für das Deutsche