

...idance p...ar ubir p...
...toe ist h...branc ho...branc...
...ta ho...it...m...o...p...dige...



Digital Diachron Deutsch

Ein Historisches Referenzkorpus für das Deutsche



Prof. Dr. Anke Lüdeling
(Humboldt-Universität zu Berlin)

Plans for a diachronic corpus of German

NTNU, Culture-2000 Workshop
Trondheim, 19.09.2003



Outline

1. the current German historical corpus situation
2. state of the project
3. multi-modal architecture
4. annotation levels
5. example
6. aside: bioinformatics & corpus linguistics

the current German historical corpus situation

- digital text archives
 - TITUS (Thesaurus Indogermanischer Text- und Sprachmaterialien), <http://titus.uni-frankfurt.de/>: Indo-European texts, manuscripts, some texts annotated with word cruncher, publically available, diplomaticity differs
 - Mittelhochdeutsches Wörterbuch, <http://rzsun01.uni-trier.de:8084/>: texts in Middle High German (1050 – 1350), used for lexicographic purposes, some texts annotated with pos and lemma, not publically available
 - Bibliotheca Augustana, <http://www.fh-augsburg.de/~harsch/augustana.html>: collection of texts in many European languages, no linguistic annotation, standards vary
 - a number of smaller text collections



the current German historical corpus situation

- different text archives with different degrees of diplomaticity
- no uniform annotation standard
- no uniform header information
- coverage very different for different stages of the language
(best for Middle High German, worst for Early New High German)
- differing (and often bad) search and processing facilities



project plan

- language groups and architecture group work on a prototype until end 2003
- grant proposal (German science foundation or Ministry for Education and Science) due end Jan 2004
- workshop Dec 2003



language projects

responsible for

- text selection
- decisions on diplomaticity & normalization
- standardization of annotation
 - tag sets
 - annotation procedure (manually, semi-automatically, automaticall etc.)
- standardization of header information



technical requirements

- facsimile (where possible) included, aligned (by line) to digitalized text
- different linguistic and graphic annotation levels
- annotation of different units
- it must be possible to add texts
- it must be possible to add annotation levels
 - to all texts
 - only to some texts



technical requirements

- alignment of units within a text
- alignment of units across texts
- links to external resources (Middle High German lexicon)
- conformant with existing standards (TEI, perhaps CES, STTS, TIGER)
- efficient search strategies



core corpus

every text

- is normalised/standardized according to the criteria proposed by the language groups
- has obligatory header information
- can have optional header information
- has obligatory annotation levels
 - text structure (graphic word, line, verse, paragraph, ...)
 - lemma
 - inflectional morphology (derivative of STTS tag set)
- can have optional annotation levels



annotation units and levels

- sign-based: graphic information
- line-based: alignment with facsimile
- lexical word-based: lemma, pos, morphology, language, translation, ...
- stem-based: stem formation
- sentence-based: syntactic information
- ...



corpus architecture

- multi-modal corpus design
- time line \approx graphical sign
- every annotation level is in an extra file
- relational data base

→ example





aside: bioinformatics and corpus linguistics

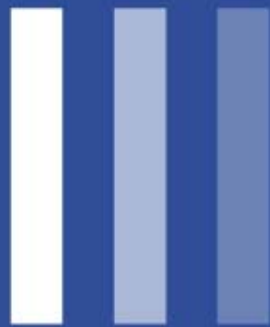
- bioinformatics methods to model relationships (similarities) between languages
- bioinformatics methods to model language change
- annotation: sequences of DNA/proteins \approx sequences of signs
 - many annotation levels
 - annotation of different units
 - efficient retrieval strategies



Thank you!

Stefanie Dipper
Karin Donhauser
Lukas Faulstich
Ulf Leser
Anke Lüdeling
Thorwald Poschenrieder

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



Digital
Diachron
Deutsch

Ein Historisches Referenzkorpus für das Deutsche



High German

Old High German

Donhauser, Gippert, Lühr

Middle High German

Gärtner, Klein, Plate, Wegera, Solms

Early New High German

Demske, Wegera, Solms

early New High German (→ 1900)

Hass-Zumkehr

Low German

Altsächsisch

Mittelniederdeutsch

Klein, Lühr, Schröder, Peters