



Anke Lüdeling

(Humboldt-Universität zu Berlin)

work done with

Stefanie Dipper, Lukas Faulstich, Ulf Leser, Thorwald Poschenrieder

A database for studying language change  
in German

Textual Criticism and Genetics - Confronting Methods  
Louvain-la-Neuve, Sep 2004



## outline

- bioinformatics & linguistics
- DeutschDiachronDigital
  - goal/initiative
  - architecture
- first experiments: syntactic similarity



## phylogenetic methods for linguistics

- modelling language families – similarity & distance
- phylogenetic  $\subset$  quantitative methods

## phylogenetic methods for linguistics

- comparative method problematic
    - for languages with little written record
    - not formalizable
  - testing phase ('modelling what we already know to test methods')
  - different linguistic levels (not only word lists, character-based methods):  
" ... that a classification based on word lists may not, for any number of reasons, yield the same tree as a wider ranging and more extensive comparison including data from other levels of grammar."  
(McMahon & McMahon 2003)
- 

## phylogenetic methods for linguistics: data

- we need comparative data for different languages/language stages
  - not only word lists/meaning lists but data from other levels as well
    - diachronic corpora with parallel portions
- 

## goal

- diachronic corpus of German, Old High German (800) to Modern German (≈1900) for linguistic, philological and historic research
  - current situation: a lot of digitized texts, but
    - different (mostly implicit) quality standards (source, diplomaticity)
    - different formats (WordPress, WordCruncher, XML, ...)
    - different header structures (if any)
    - different positional or structural annotation (if any)
    - unequal coverage and different corpus composition for the language stages
    - availability sometimes problematic, no common search tools
- 

## the DDD initiative

- linguists, philologists, corpus linguists, computer scientists from 15 German universities, international cooperation
  - 5 language groups + architecture group
  - pilot project for corpus architecture at Humboldt-Universität
  - grant application submitted, planned duration 7 years
  - size after 7 years
    - core corpus: 40 M words
    - extension corpus: 60 M words
-



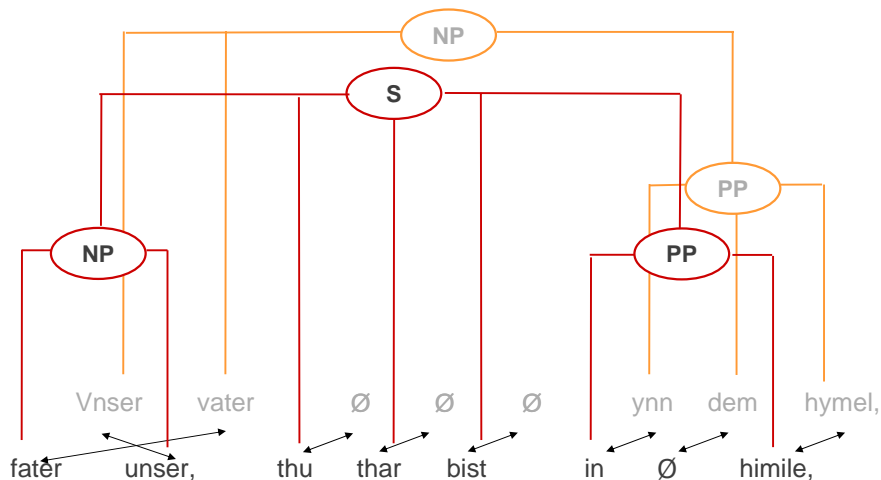
## data: the Lord's Prayer in 5 different version

- fader ist usa firio barno, thu bist an them hohen himilo rikie. Old Saxon (9th c, Heliand)
- fater unser, thu thar bist in himile, OHG (9th c, Tatian)
- Got vater unser, dâ du bist In dem himelrîche gewaltic alles des dir ist, MHG (ca. 1200, Reinmar van Zweter)
- Vnser vater ynn dem hymel. EMG (1422, Luther)
- Vater unser, der du bist im Himmel, MG

## experiments with syntax tree distances

- syntax trees in TIGER format (Brants et al. 2002)

### Old High German and Early Modern German



## experiments with syntax tree distances

- TIGER graphs transformed into ordered trees, crossings of edges (i.e., order of terminals) ignored
- treediff software (Shasha, Wang & Zhang) used to compute distances between syntax trees of corresponding sentences in different text version
- sentence-wise distances combined to document distance:  $d(x,y)^2 = \sum_i d(x_i,y_i)^2$
- distance data visualized using Multidimensional Scaling (xgvis software): data points plotted in 2D with minimum distortion
- alternative visualization using PHYLIP phylogeny software

## preliminary results: estimation vs. Lord's Prayer

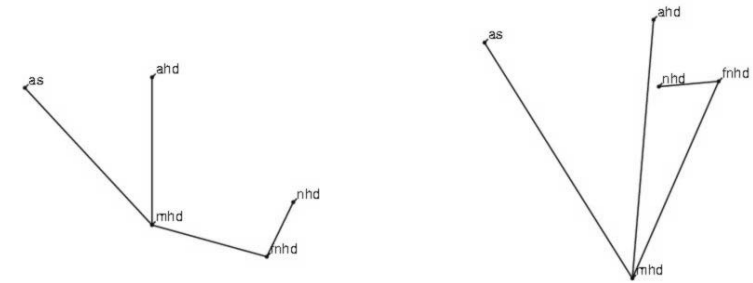
	OS	OHG	MHG	EMG	MG
OS	0	2	3	4	4
OHG	2	0	2	3	3
MHG	3	2	0	2	2
EMG	4	3	2	0	1
MG	4	3	2	1	0

	OS	OHG	MHG	EMG	MG
OS	0.00	60.58	91.74	62.95	61.26
OHG	60.58	0.00	72.11	35.19	27.03
MHG	91.74	72.11	0.00	72.98	71.81
EMG	62.95	35.19	72.98	0.00	25.09
MG	61.26	27.03	71.81	25.09	0.00

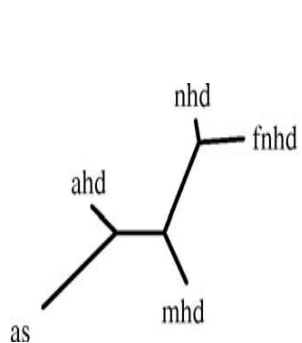
## preliminary results: Multidimensional Scaling

estimated                      Lord's Prayer

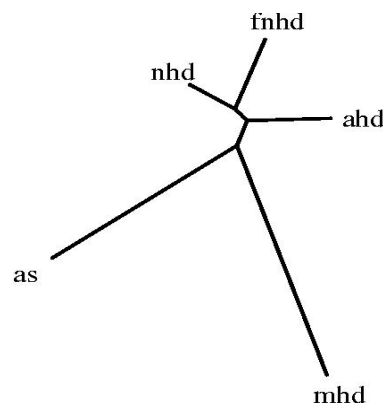


## preliminary results: phylogeny

trees  
estimated



Lord's Prayer



## problems

- problems
  - too little data
  - data is special because it is a translation and it is formulaic as a religious text
- fater unser, thu thar bist in himile, OHG (9th c, Tatian)
- Vater unser, der du bist im Himmel, MG  
vs.
- Vnsrer vater ynn dem hymel. EMG (1422, Luther)
- Unser Vater im Himmel 'real' MG



## outlook

- more data ☺
    - comparability?
  - modelling
    - different models for the linguistic levels (cladistic vs. reticulate?)
    - influence of language contact on different levels
  - for syntax: other algorithms for calculating tree similarity need to be tested
    - lexicalized vs. non-lexicalized trees
    - different weights for different edit operations
- 

<http://www.deutschdiachrondigital.de>



Digital  
Diachron  
Deutsch

A Historical Corpus of German