

Measuring Syntactic Change

Underuse and Overuse Statistics in a Multi-Layer Diachronic Corpus of German

Hagen Hirschmann

hirschhx@hu-berlin.de

Anke Lüdeling

anke.luedeling@rz.hu-berlin.de

Julia Richling

julia@richling.de

Amir Zeldes

amir.zeldes@rz.hu-berlin.de

Humboldt-Universität zu Berlin

Plan

1. Research questions
2. Variants and Variationism
3. Underuse / overuse diagnostics
4. Multi-layer corpora
5. Two case studies:
 1. Relative clauses
 2. Periphrastic constructions

Measuring syntactic change

Historical grammars of German:

- Focus on phonological developments
- Catalogue of constructions in each language stage
- Usually no frequency information
- Relationship between the development of different constructions?

Measuring syntactic change

- We want to find out:
 - How did German syntax develop?
 - What are the most prominent changes in each period?
 - Gradual or abrupt changes?
 - How does a new construction oust an older one?
 - Concurrent use of older and newer systems?

Variationism (see e.g. Rissanen 2008)

- Variable: multiple ways of 'saying the same thing'
 - e.g. relative clauses
- Variant: one particular way
 - *The book* _A **that** / _B **which** / _C **∅** *I read*

Period 1

A
A
A
A
B
B

>

Period 2

A
B
B
B
C

Variationism

- Inherently quantitative approach
- Interest in distribution of variants in each period
- Requires identification of variables (what is 'the same thing'?)

- Diachronic corpus with uniform annotation scheme

Diachronic data

- ❑ Expensive, limited resources available
- ❑ Hard to use 'data-driven' methods (extrapolation from surface statistics)
- Use deeply annotated corpora to directly identify variables
- Annotate 'everything', since we don't know what will be interesting

The corpora

- 4 very small but deeply annotated sub-corpora, religious genres:
 - **Old High German** – Monsee Fragments (Gospel of Matthew) 2752 tokens
 - **Middle High German** – Speculum ecclesiae (Sermons) 2483 tokens
 - **Early New High German** – Sermon by Veit Nuber (written 1544) 2673 tokens
 - **New High German** – New Evangelistic Translation (Acts 1-4) 3574 tokens

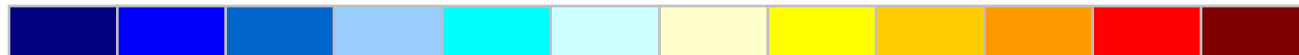
The corpora

- Uniform, deep annotation scheme:
 - POS, morphology, lemma, gramm. function
 - Hyperlemmatization of top 100 forms (unified lemma for all periods, based on NHG)
 - Syntactic annotation (constituent trees with dependency edge labels)
 - Annotated frequency, its deviation from NHG and significance for POS/POS-bigrams, word forms, hyperlemmas
 - Bibliographic information (line in edition...)
 - Normalization

Using underuse/overuse diagnostics

- Search for most significant differences between stages
- compare normalized frequencies of all annotation levels and categories
- Visualize underuse as progressively colder colors, overuse as progressively warm colors

Underuse



Overuse

Example results – POS unigrams

pos	OHG	MHG	ENHG	NHG
PDAT	0.046131	0.011679	0.007105	0.008954
PPER	0.083545	0.052759	0.075916	0.075825
ART	0	0.07934	0.065445	0.061835
VVINF	0.01126	0.015707	0.018325	0.022104
PRELS	0.009444	0.011679	0.013837	0.016788
VAFIN	0.03705	0.035038	0.047868	0.045887
VAINF	0.001453	0.001208	0.004113	0.003078
PDS	0.023974	0.026983	0.013089	0.004757

Interpreting results

- Are there simply fewer relative clauses in OHG?
- Or are sentences simply longer? (fewer sentences > fewer PRELS)
- Do all relative clauses have a PRELS?
- Need for more accurate queries, syntactic annotation

Querying multiple layers with ANNIS2

- ANNIS: Annotation of Information Structure
- Multi-layer corpus search architecture
- Developed in SFB 632 "Information Structure"
- Search and visualization of complex annotation graphs, spans, relations, metadata and RegEx thereof

Querying multiple layers with ANNIS2

der	wie	eine	Mumie	auf	der	Bank	sitzende	ukrain
der	wie	ein	Mumie	auf	der	Bank	sitzend	ukrain
ART	KOKOM	ART	NN	APPR	ART	NN	ADJA	AD
n.Sg.Masc	--	Nom.Sg.Fem	Nom.Sg.Fem	--	Dat.Sg.Fem	Dat.Sg.Fem	Pos.Nom.Sg.Masc	Pos.Nom.

tiger:morph = Nom.Sg.Fem

rs to pay movie producers for showing their films . Saudi Arabia , for its part , has vowed
and to apply the law to computer software as well as to literary works , Mrs. Hills said
. They will remain on a lower-priority list that includes 17 other countries . Those countr
of some concern to the U.S. but are deemed to pose less-serious problems for American
. Gary Hoffman , a Washington lawyer specializing in intellectual-property cases , said 0 the
n that protecting intellectual property is in a country 's own interest , prompted the
ia . " What this tells us is that U.S. trade law is working , " he said . He said 0 Mexico could
list because of its efforts to craft a new patent law . Mrs. Hills said that the U.S. is still
ntinuing slow progress in Malaysia . " She did nt elaborate , although earlier U.S. trade

exmaralda								
Select Displayed Annotation Levels ▾								
Focus_newInf		nf-unsol						
Inf-Stat	acc-gen							giv-active
NP	NP							NP
PP	PP							exmaralda:Inf-Stat = giv-active
Sent	s							
Topic	fs							ab
tok	die	Ukraine	stürzte	der	1,62	Meter	große	Gennadi Subov

```

graph TD
    UP((UP)) --- COORD1[COORD]
    UP --- VP((VP))
    VP --- PRED_CO[PRED_CO]
    VP --- ADV((ADV))
    VP --- TP((TP))
    VP --- HD1[HD]
    TP --- COORD2[COORD]
    TP --- HD2[HD]
    TP --- OBJ_CO[OBJ_CO]
    COORD2 --- CP1((CP))
    COORD2 --- HD3[HD]
    COORD2 --- OBJ_CO2[OBJ_CO]
    CP1 --- LP((LP))
    LP --- HD4[HD]
    LP --- COORD3[COORD]
    COORD3 --- CP2((CP))
    COORD3 --- HD5[HD]
    COORD3 --- ATR_CO1[ATR_CO]
    COORD3 --- ATR_CO2[ATR_CO]
    CP2 --- ATR1[ATR]
    CP2 --- ATR2[ATR]
    CP2 --- NP((NP))
    NP --- ATR3[ATR]
    NP --- ATR4[ATR]
    NP --- HD6[HD]
    OBJ_CO2 --- NP2((NP))
    NP2 --- ATR5[ATR]
    NP2 --- ATR6[ATR]
    NP2 --- HD7[HD]
    HD1 --- SBJ[SBJ]
    HD2 --- H[ ]
    HD3 --- CAI[ ]
    HD4 --- OITE[οίτε]
    HD5 --- TO[τὸ]
    HD6 --- SOUSISON[Σούσιων]
    HD7 --- HD8[ἠδ']
    HD8 --- AGBATANON[Ἀγβατάνων]
    HD9 --- KAI[καὶ]
    HD10 --- TO2[τὸ]
    HD11 --- PALAION[παλαιὸν]
    HD12 --- KISSION[Κίσσιον]
    HD13 --- ERKOS[ἔρκος]
    HD14 --- PROLIPONTES[προλιπόντες]
    HD15 --- EBAN[ἔβαν]
    
```

οίτε τὸ Σούσιων ἠδ' Ἀγβατάνων καὶ τὸ παλαιὸν Κίσσιον ἔρκος προλιπόντες ἔβαν

Putting it together

- Two case studies:
 1. Diagnostic: gradient underuse of PRELS
 - Increase in relative clauses?
 2. Diagnostic: underuse of VAFIN/VAINF
 - Development of periphrastic constructions?

1 Relative clauses

1. Use syntactic annotation to normalize to clauses:

2. PRELS=?RC

3. Find relative clauses in **syntactic** annotation

Sub-corpus	PRELS per 100 clauses
OHG	4.62 (sig. $p < 7.844e-06$)
MHG	10.25
ENHG	12.85
NHG	13.35

ANNIS2 browser interface

ANNIS² Corpus Search

ANNIS² Tutorial

Search Form

AnnisQL: `node & node & pos="PRELS" & ratio_word & #1 >[func="RC"] #2 & #2 > #3 & #3 _=_ #4`

Query Builder: [Show >>](#)

Result: 7

More Corpora

Name	Texts	Tokens
<input type="checkbox"/> Hebrew_Treebank_	3	1718
<input checked="" type="checkbox"/> ahd	1	2752
<input type="checkbox"/> nhd.Taten.Lukas.1-	1	3574
<input type="checkbox"/> arabic.test	1	11
<input type="checkbox"/> hildebrandtLied	1	2749
<input type="checkbox"/> hotelCorpus	208	177674
<input type="checkbox"/> pcc3v2	705	3256
<input type="checkbox"/> thukydides01	1	46
<input type="checkbox"/> tiner?	1471	888578

[Search](#) [Statistics](#)

Context Left: 5
Context Right: 5
Results per page: 10

[Show Result](#)

Search Result - node & node & pos="PRELS" & ratio_word & #1 >[func="RC"] #2 & #2 > #3 & #3 _=_ #4 (5, 5)

Page 1 of 1

Token Annotations Show Citation URL

Displaying Results 1 - 7 of 7

	furistun	dero	liuteo	in	frithoue	des	herostin	dero	euardo	der	heaz
	Nom.Pl.Masc	Gen.Pl.Masc	Gen.Pl.Masc	--	Dat.Sg.Masc	Gen.Sg.Masc	Gen.Sg.Masc	Gen.Pl.Masc	Gen.Pl.Masc	Nom.Sg.Masc	3.Sg.Past.Ind
	NN	PDAT	NN	APPR	NN	PDAT	NN	PDAT	NN	PRELS	VVFIN
		d		in		d		d		d	

exmaralda
tiger
Paula
Paula Text

	pontischin	herizohin	pilate	Duo	kasah	iudas	der	inan	dar	forreat	daz
	Pos.Dat.Sg.Masc.Vwk	Dat.Sg.Masc	Dat.Sg.Masc	--	3.Sg.Past.Ind	Nom.Sg.Masc	Nom.Sg.Masc	3.Acc.Sg.Masc	--	3.Sg.Past.Ind	--
	ADJA	NN	NE	ADV	VVFIN	NE	PRELS	PPER	ADV	VVFIN	KOUS
			da				d	sie	da		d

exmaralda

Select Displayed Annotation Levels

	pontischin	herizohin	pilate	Duo	kasah	iudas	der	inan	dar	forreat	daz
ratio_hyper_lemma						15.255360					15.255360
ratio_pos		0.679743	1.854391	0.721636	0.419785	1.854391	1.777582		0.721636	0.419785	
ratio_pos_bi	1.403514			0.298657							
ratio_word							1.911449				
tok	pontischin	herizohin	pilate	Duo	kasah	iudas	der	inan	dar	forreat	daz

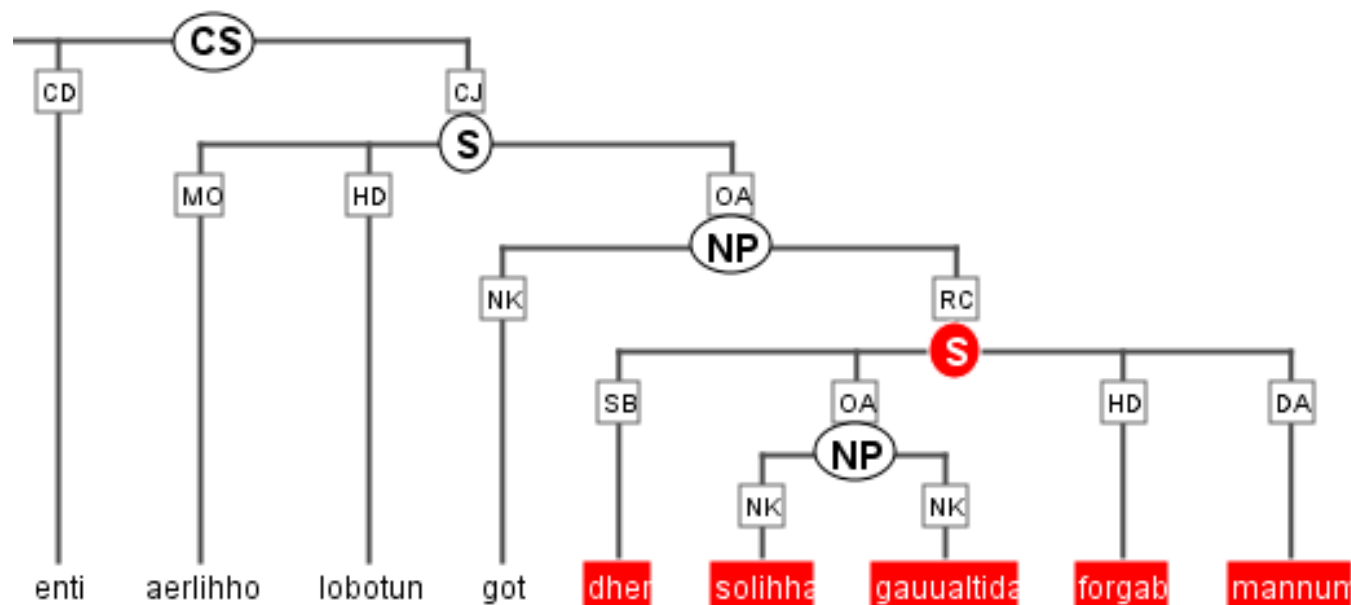
tiger

Type 1: PRELS

- and praised God, **who** gave men such power

	enti	aerlihho	lobotun	got	dher	solihha	gauuaitida	forgab	mannum
.Neut	--	--	3.Pl.Past.Ind	Acc.Sg.Masc	Nom.Sg.Masc	Acc.Sg.Fem	Acc.Sg.Fem	3.Sg.Past.Ind	Dat.Pl.Masc
:	KON	ADV	VVFIN	NN	PRELS	PIAT	NN	VVFIN	NN
	und								Mensch

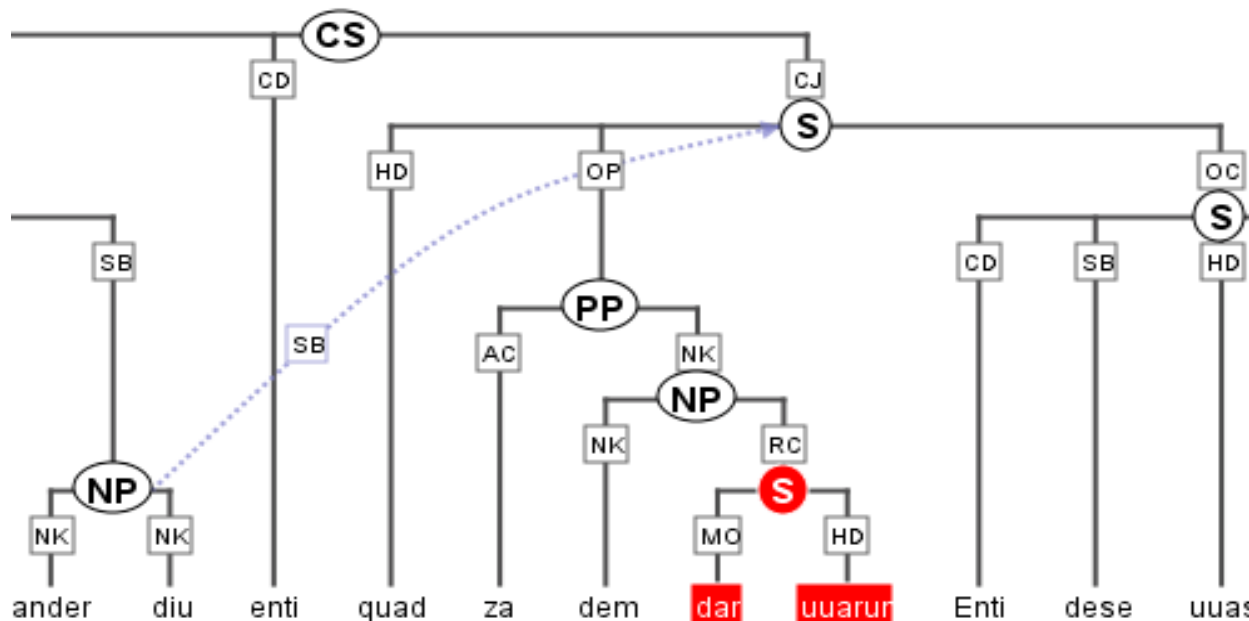
Ja



Type 2: Asyndetic

□ and said to them **were there**

inan	ander	diu	enti	quad	za	dem	dar	uuarun
3.Acc.Sg.Masc	Pos.Nom.Sg.Fem.St	Nom.Sg.Fem	--	3.Sg.Past.Ind	--	Dat.Pl.*	--	3.Pl.Past.Ind
PPER	ADJA	NN	KON	VVFIN	APPR	PDS	ADV	VAFIN
sie		d	und	sagen	zu	d	da	sein

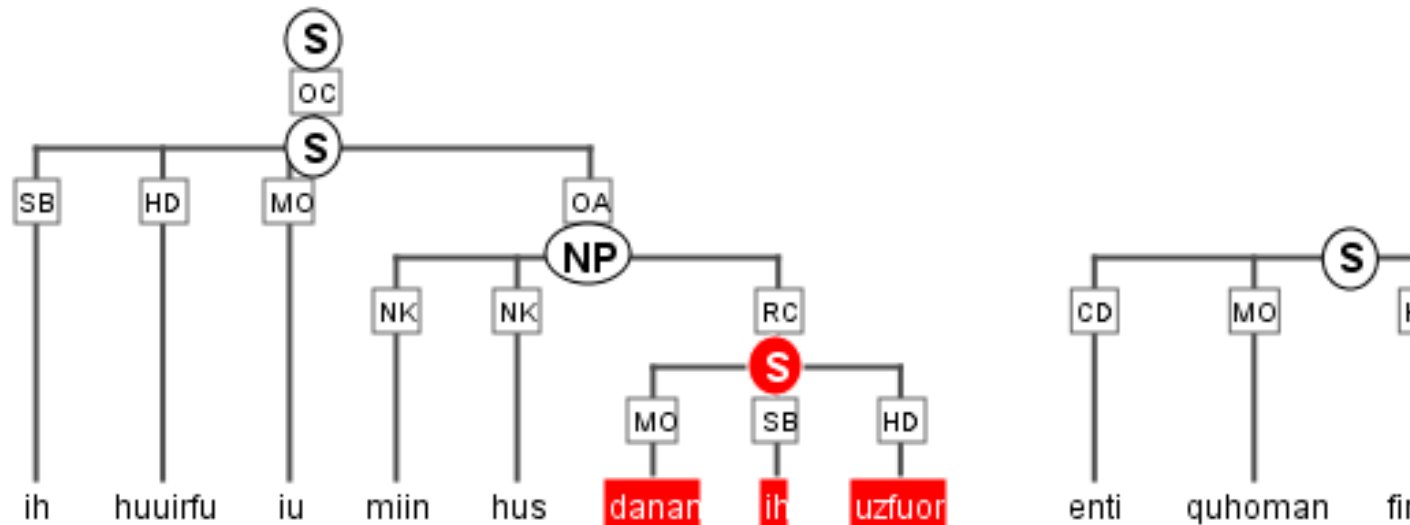


Type 3: PWAV

- I return now to my house, **whence** I departed

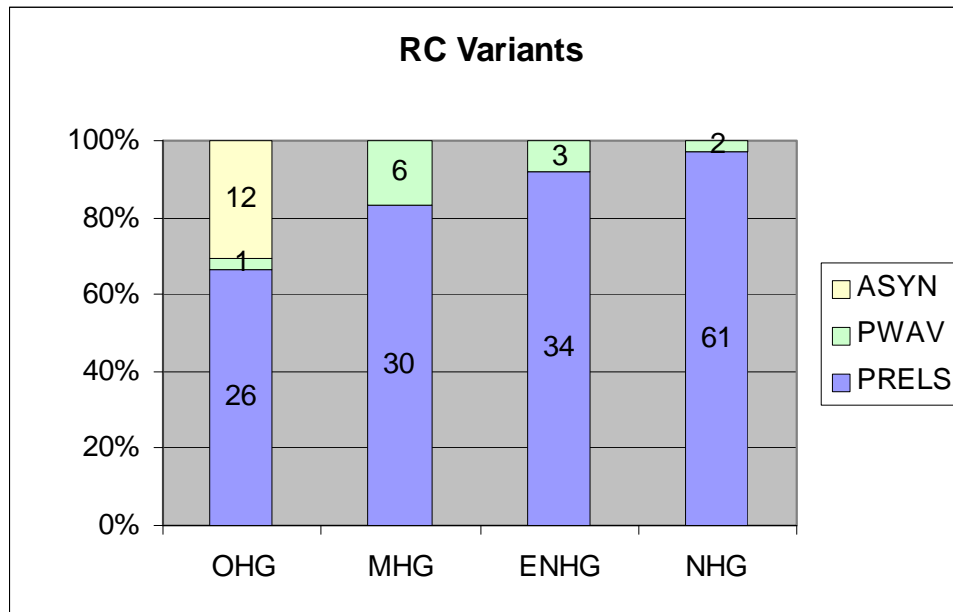
ih	huuirfu	iu	miin	hus	danan	ih	uzfuor
1.Nom.Sg.Masc	1.Sg.Pres.Ind	--	Pos.Acc.Sg.Neut	Acc.Sg.Neut	--	1.Nom.Sg.Masc	1.Sg.Past.Ind
PPER	VVFIN	ADV	PPOSAT	NN	PWAY	PPER	VVFIN
ich		ihr	mein	Haus	fort	ich	

tiger



Relative clauses - Overview

	with PRELS	with PWAV	Asyndetic	total RC
OHG	26 (4.62)	1 (0.18)	12 (2.13)	39 (6.93)
MHG	30 (10.60)	6 (2.12)	0	36 (12.72)
ENHG	34 (11.81)	3 (1.04)	0	37 (12.85)
NHG	61 (13.35)	2 (0.44)	0	63 (13.79)




(contrast significant at p -value $< 2.2e-16$ for all parameters, and within MHG-NHG for PRELS/PWAV)

Summary

- Variable: RC
 - PRELS dominant in all periods
 - Asyndetics present in OHG but completely absent later
 - PWAV in all periods, but decreasing
 - Progressively limited to adverbial usage though also known dialectally (not discussed)

2 Periphrastic constructions

pos	OHG	MHG	ENHG	NHG
PDAT	0.046131	0.011679	0.007105	0.008954
PPER	0.083545	0.052759	0.075916	0.075825
ART	0	0.07934	0.065445	0.061835
VVINF	0.01126	0.015707	0.018325	0.022104
PRELS	0.009444	0.011679	0.013837	0.016788
VAFIN	0.03705	0.035038	0.047868	0.045887
VAINF	0.001453	0.001208	0.004113	0.003078
PDS	0.023974	0.026983	0.013089	0.004757



2 Periphrastic constructions

- Categories VAFIN/VAINF too coarse
- Include all forms of *haben* 'have', *sein* 'be', *werden* 'become'
- Use hyper lemmatization to distinguish:

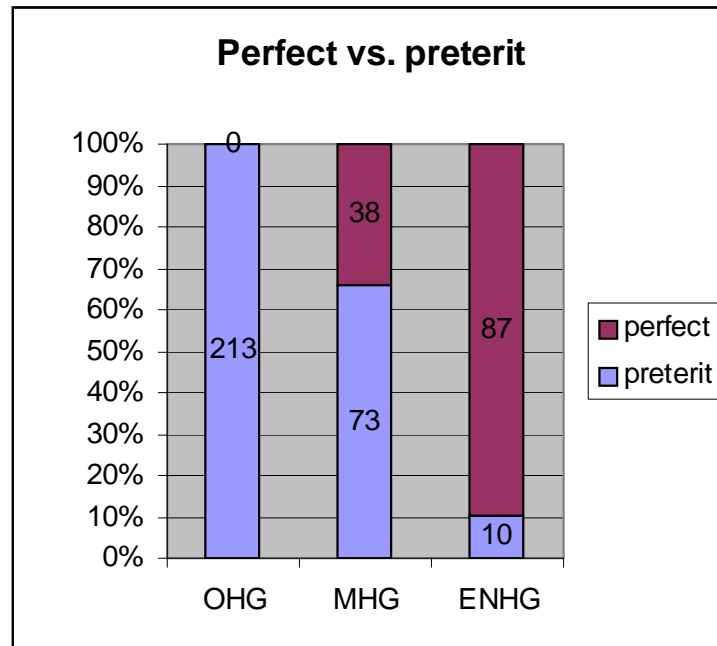
	OHG	MHG	ENHG	NHG
<i>haben</i>	0.0021802	0.00362465	0.016455	0.020145
<i>sein</i>	0.0308166	0.02698349	0.028048	0.024902
<i>werden</i>	0.0043143	0.00161095	0.013089	0.01455

2 Periphrastic constructions

- 4 Uses for VAFIN/VAINF:
 1. Periphrastic perfect
 2. Periphrastic passive
 3. Copula
 4. Full lexical verb
- Concentrate on cases 1 & 2
- Compare to synthetic preterit (morphological annotation)

2 Periphrastic constructions

	preterit	perfect	passive
OHG	213 (37.83)	0 (0.00)	33 (5.86)
MHG	73 (25.80)	38 (13.42)	15 (5.30)
ENHG	10 (3.47)	87 (30.21)	17 (5.90)



Summary

- Variable: Periphrasis
 - No periphrastic tenses in OHG
 - Perfect encroaches on preterit gradually, parallel development of future periphrasis
 - Passive use develops independently, represents different variable

Conclusion

- Annotation scheme determines variables and variants we find
- Decision 'what is the same thing' crucial in diachronic corpora
- Need to code and examine multiple (independent) annotation levels

Conclusion

- Uniform scheme across periods
- Underuse / overuse statistics as a diagnostic for interesting phenomena
- Closer look at all annotation levels for more fine-grained categorization
- Quantitative interplay of variants shapes diachronic syntax

Thank you for your attention!

- ANNIS:

<http://www.sfb632.uni-potsdam.de/d1/annis/>

- Underuse/overuse add-in for Excel:

<http://korpling.german.hu-berlin.de/~amir/uoadin.htm>