Annotation Guidelines for German Non-standard Varieties

Marc Reznicek

empirikom

4. Arbeitstagung

8.11.2012, Aachen

Overview

Motivation & Goals

- existing resources

Data & Annotation

- test corpus
- annotation

Chat

- DCC-Chat-Protocols (Conversion, Processing)

Next steps

Motivation: linguistic corpus annotation

last 20 years:

Large effort for linguistic annotations on corpora

- large annotation projects (TiGer, TüBa)
- shared tasks (CONLL, MUC, B-CUBED, ACE, ARE, CEAF)
- SFB: Information structure

resources

- tag sets
- annotation schemes
- guidelines
- training sets

→ **designed for linguistic standard variety**

tools

- taggers
- parsers

→ **trained mostly on newspaper text**

Motivation: domain adaption in NLP

In NLP domain adaption fastly growing

- POS:
  - e.g.. Miller et al.(2006), Kübler & Baucom (2011), Hinrichs & Zastrow (2012), Rehbein et al. 2012, Rush et. al. (2012)
- syntactic analysis
  - SANCL 2012 shared task
  - e.g. Dredze et al. (07), Yoshida et al.(2007)
- coreference resolution
  - e.g.Yang et al. (2012), Uryupina, Poesio (2012)

Motivation: beyond robustness

- In NLP domain adaption is mostly seen as a matter of ***robustness*** (e.g.Balsa & Lopes 2000, Carreras & Marquez 2005)
- **assumption:**
  - there is **a right way** to annotate the data

→ **task:**

- reach at the right annotation under the condition of noisy data

! This assumption only holds for prototypical data!

Motivation: Weaknesses of existing resources

- categories of many phenomena in non-standard varieties are not defined
  - e.g. part-of-speech in **learner language**

Studenten	**sind**	in	der	Uni	**viel**	**praxisorientiert**	**.**
NN	VAFIN	APPR	ART	NN	**PIS**	**???**	
Studenten	sind	in	der	Uni	**sehr**	**praxisorientiert**	.
NN	VAFIN	APPR	ART	NN	**ADV**	**ADJD**	

Students are very practically oriented in university.

Studenten	werden	in	der	Uni	**stark**	**praxisorientiert**	.
NN	VAFIN	APPR	ART	NN	**ADJD**	**VVPP**	

Students are quite often practically oriented in university [by someone].

Motivation: Weaknesses of existing resources

- categories of many phenomena in non-standard varieties are not defined
  - e.g. part-of-speech in **historical data**

nu	**pin**	ich	groſleich	**erhocht**	**vön**		**meīem**	**chind**	vber	alles	himliſch	her
nun	**bin**	ich	sehr	**erhöht**	von		meinem	kind	über	alles	himmlische	heer
ADV	VAFIN	PPER	ADV	**ADJD**	APPR		PPOSAT	NN	APPR	PIS	ADJA	ADJA
nun	**bin**	ich	sehr	**erhöht**	von	***worden***	meinem	kind	über	alles	himmlische	heer
ADV	VAFIN	PPER	ADV	**VVPP**	APPR		PPOSAT	NN	APPR	PIS	ADJA	ADJA

Motivation: evaluation of annotation schemes

- domain adaption in a linguistic sense:
  extention of descriptive inventory of non canonical structures
- **assumption:**
  - there is a **no right way** to annotate non-canonical data with grammars that describe only canonical data

→ **task:**

- evaluate and extend annotation schemes and guidelines

Motivation: gold standard data

- **extended descriptions** of structures from more domains allows:
  - building a **gold standard** annotated data resource
    - ➢ for **evaluation** of standard tools on non-standard varieties
    - ➢ for **training**

Motivation: gold standard data

request for IBC test set in NLP

Punctuation and capitalization are often inconsistent, making it difficult to rely on features that can be predictive for newswire [the training data]. There is often a lexical shift due to increased use of slang, technical jargon or other phenomena. There is an increase in ungrammatical sentences.

…

Unfortunately, there are currently few high quality test sets available for evaluating parsers on such noisy web texts, forcing researchers to keep evaluating on a now 20 year old test (WSJ Section 23).

(SANCL2012 shared task motivation)

Motivation:
Clarin-D curation project

Clarin F-AG 7 - Curation project (KP2):
Linguistic annotation of nonstandard varieties — guidelines and "best practices"

Motivation: Goals

data:

- test corpus of German non-standard varieties

annotations:

- gold standard
  - dependencies
  - coreference
  - named entities

guidelines:

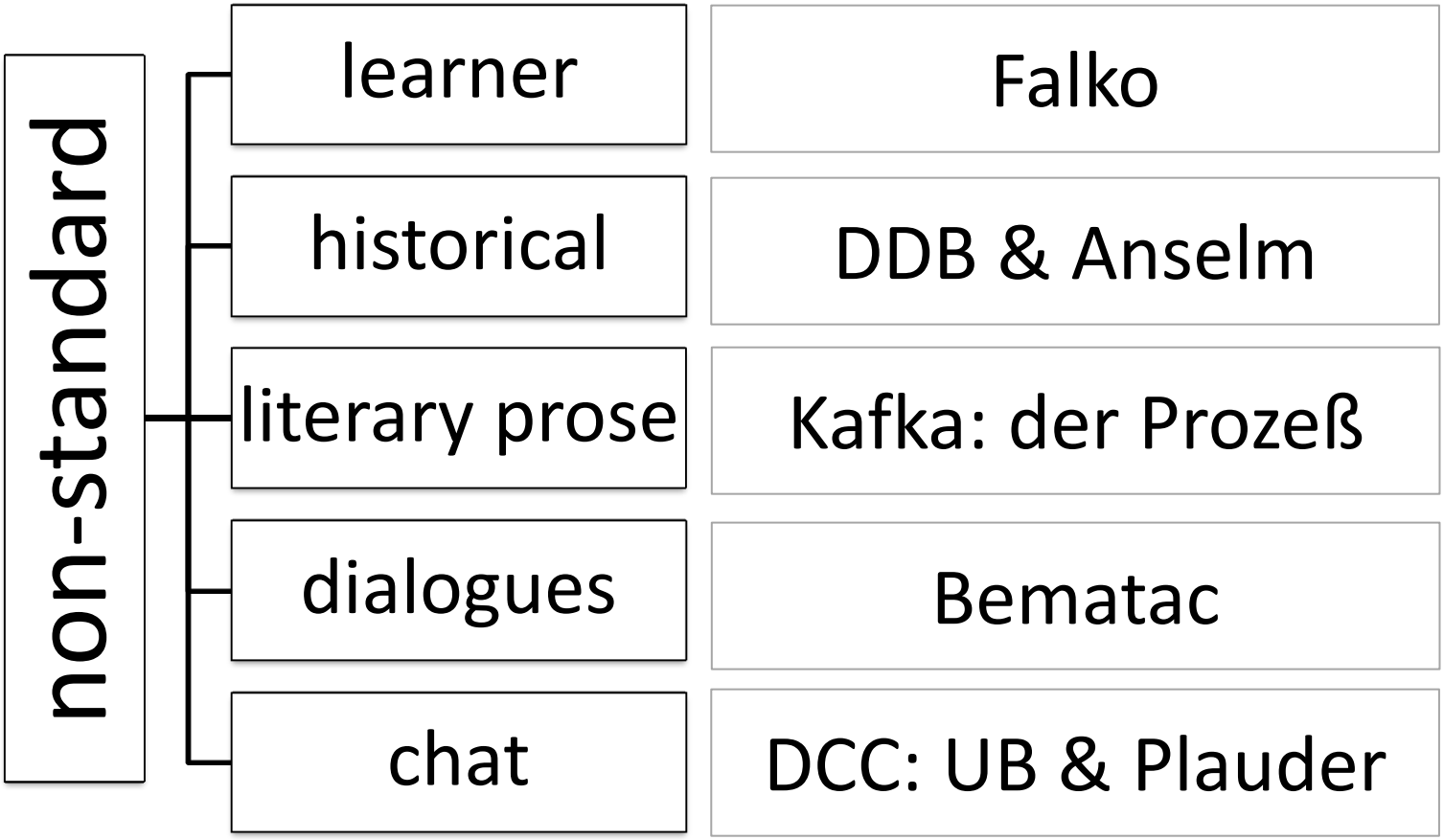
- evaluation and extension of existing guidelines
- "best practices" for lay people

tools:

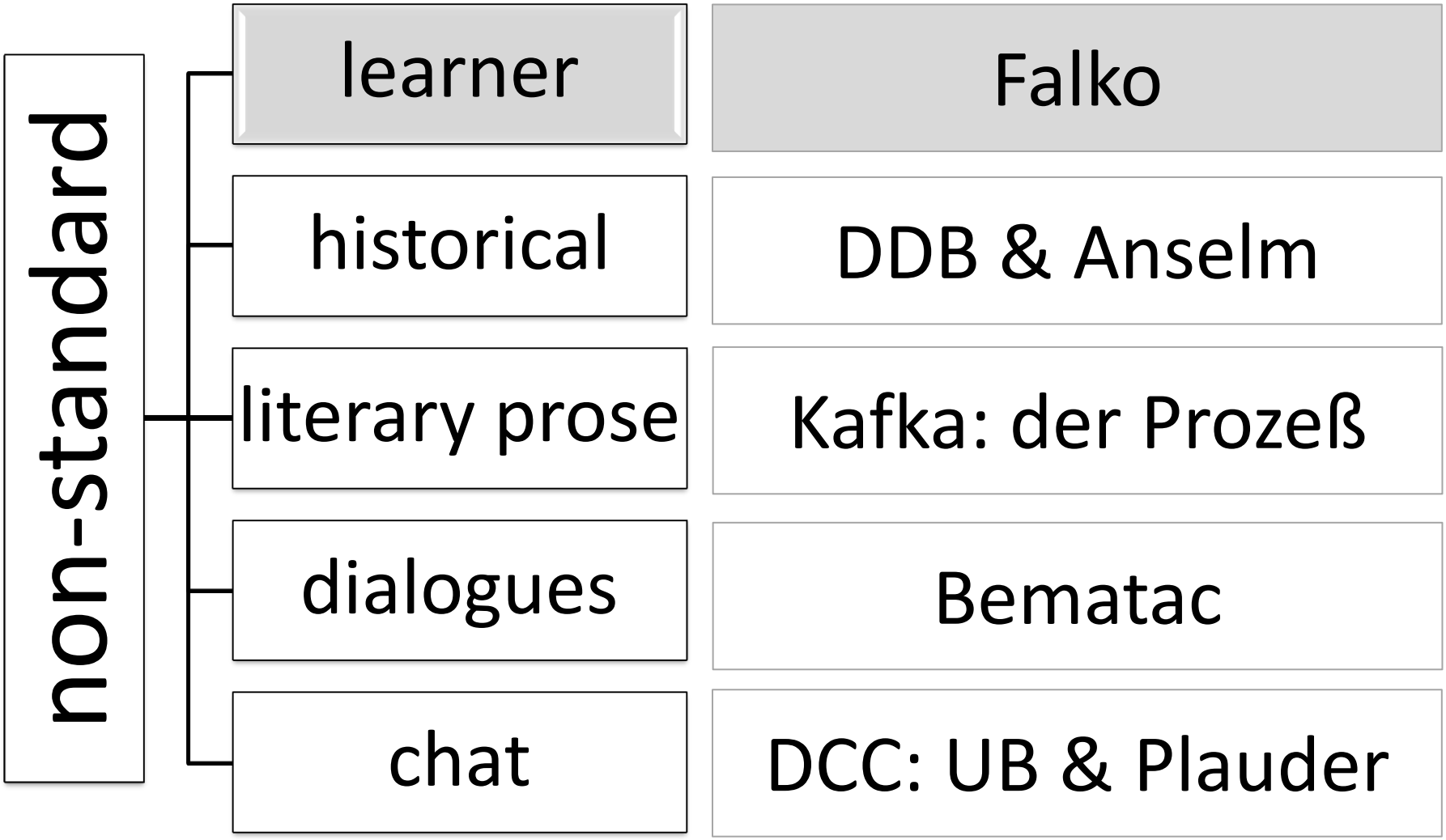
- evaluation of automatic tools for chosen annotations

Motivation: Data

- description of similarities and differences



Motivation: learner language



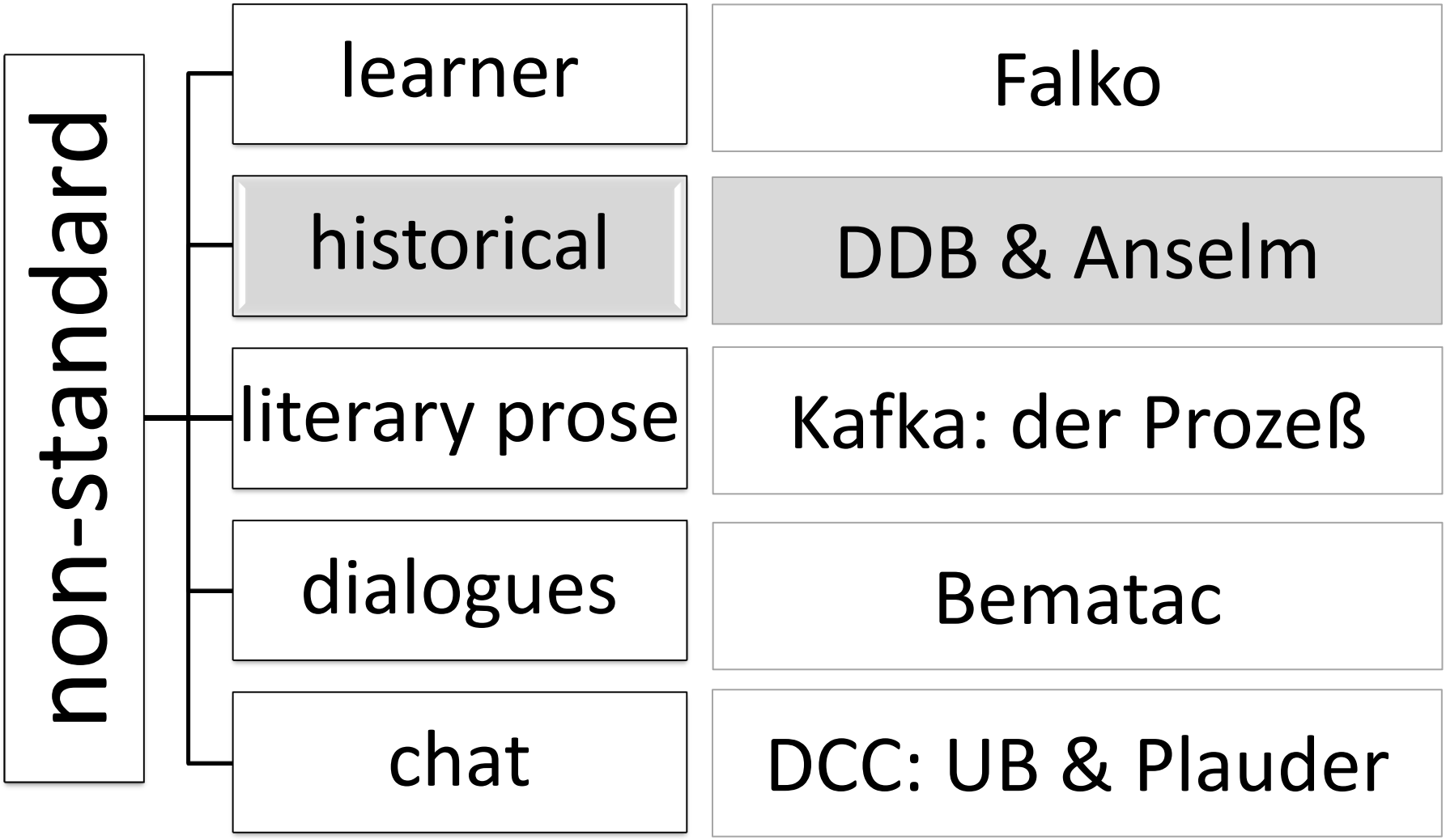
Motivation: learner language

tok	TH1a	TH1b
Meine	Meine	Meine
meinung	Meinung	Meinung
ist	ist	ist
,	,	,
dass	dass	das
	das	
examen	Examen	Examen
soll		soll
kommen		
am	am	am
Ende	Ende	Ende
des	des	des
Semester	Semesters	Semesters
	kommen	**kommen**
	soll	

- **word order**
- **argument selection**

und	und
bereiten	
die	die
Studenten	Studenten
vor	auf
die	die
wirkliche	wirkliche
Welt	Welt
	vorbereiten

Motivation: historical language



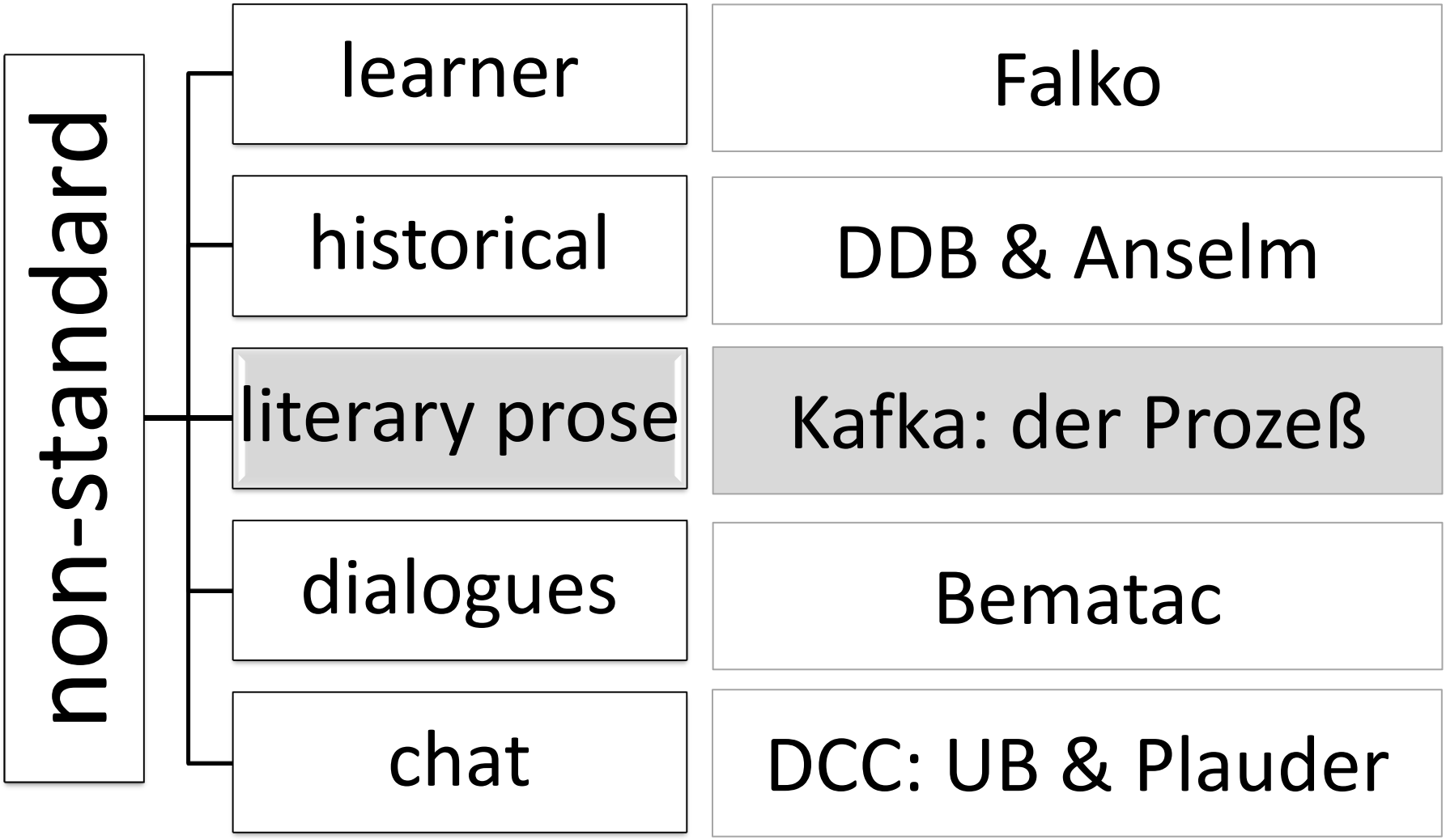
Motivation: historical language

- **sentence segmentation**

AIn höcher lerer hieſz anſhelm ***der*** ***pat vnſer frawn lang wainn̄d vnd vaſtun̄d daz*** *ſe im czu erchennen geb wie vnſer her gemartert wer vnd do er alſo nach ſeiner gewonhait vm̄b dy ſelbn̄ ſache eines males vnſer frawn pat mit ganczm̄ ernſt do erſchain im vnſer fraw vnd ſprach alſo*

der	**pat**	**vnser**	**frawn**	**lang**	**wainnnd**	**vnd**	**vastunnd**	**daz**
der	bat	unsere	Frau	lang	weinend	und	fastend	dass
REL	VVFIN	PPOSAT	NN	ADV	ADJD	KON	ADJD	KOUS
PIS	VVFIN	PPOSAT	NN	ADV	ADJD	KON	ADJD	KOUS

Motivation: literary prose



Motivation: literary prose

- **complex insertions**
- **direct speech**

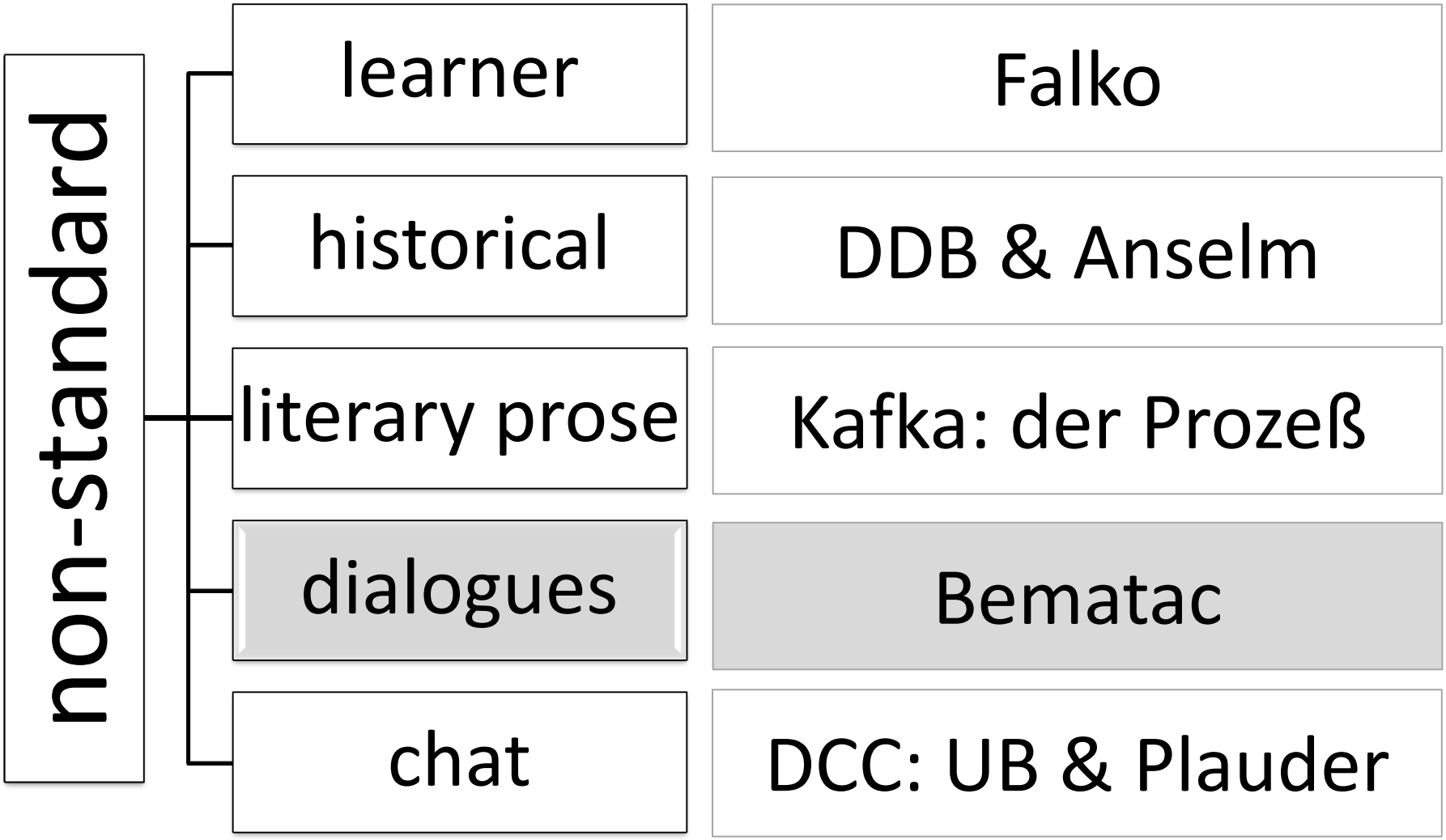
"Ja", sagte der Mann, "aber ich bin nicht mehr verpflichtet, Sie jetzt zu verhören" - ***wieder das Murren, diesmal aber mißverständlich, denn der Mann fuhr, indem er den Leuten mit der Hand abwinkte, fort, -*** *"ich will es jedoch ausnahmsweise heute noch tun. Eine solche Verspätung darf sich aber nicht mehr wiederholen. Und nun treten Sie vor!"*

Motivation: literary prose

- **complex insertions**
- **direct speech**

"Ja", sagte ***der Mann****, "aber* ***ich*** *bin nicht mehr verpflichtet, Sie jetzt zu verhören"* - ***wieder das Murren, diesmal aber mißverständlich, denn der Mann fuhr, indem er den Leuten mit der Hand abwinkte, fort, -*** *"****ich*** *will es jedoch ausnahmsweise heute noch tun. Eine solche Verspätung darf sich aber nicht mehr wiederholen. Und nun treten Sie vor!"*

Motivation: spoken dialogues



Motivation : : spoken dialogues

- **filled pauses**
- **interruptions**
- **corrections**
- **repetitions**

genau **du gehst jetzt/** du gehst jetzt genau **am/** am Motorrad da unten **äh oben** lang und vorbei geradeaus du machst **n** du machst **so/** wieder so **nen** Knick

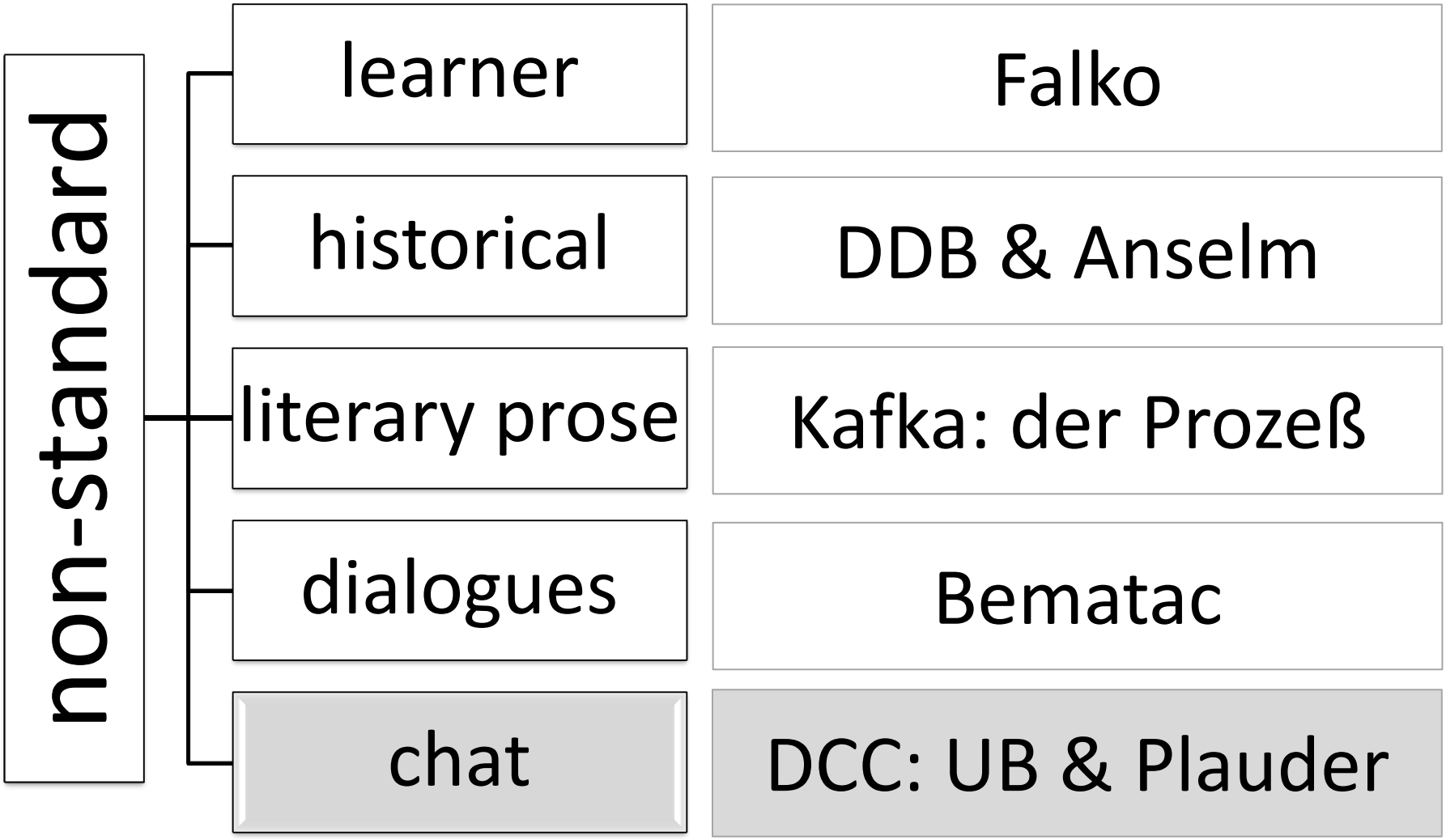
Motivation : : spoken dialogues

- **collaborative contructions**

instructor		instructee	
da	ADV		
gehst	VVFIN		
du	PPER		
rechts	ADV		
ab	PTKVZ		
		unter	APPR
		oder	KON
		über	APPR
		dem	**ART PIS**
unter	APPR		
dem	**ART PIS**		
unter	APPR		
dem	**ART**		
Bild	**NN**		

empirikom

Motivation : chat protocols



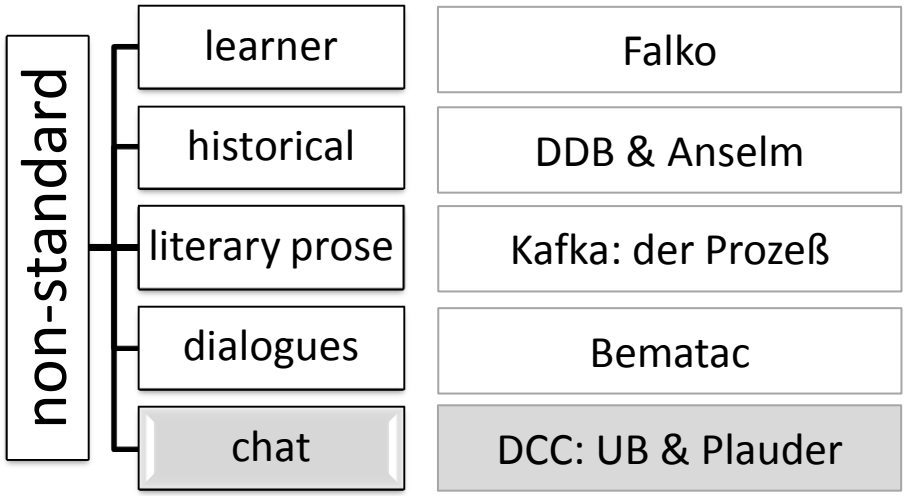
Motivation : chat protocols

- chat data especially challenging
  - hybrid variety (oral – written / close –distant)

    (e.g.Lenke & Schmitz 1995, Beißwenger 2001, Storrer 2001)

  → shares deviations from standard with most other chosen varieties

  → chat-specific phenomena



Motivation : chat protocols

- **tokenization**

Hübsche1017_(w): Ja ich sage dir das mein Mann **zu hause** wartet. Du ich gehe jetzt. War nett mit dir. Wie **wärs** mit nächsten Dienstag so um **halb11** oder 11 Uhr?
Single9384_(w): ich verstehe mich mit jeden mann
Speedy24: Bist du Arbeiten?
Single9384_(w): das meinte ich
AufderSuche2481_(m): ja ich **probiers**
Süsser1201_(m): ja, wirklich?
AufderSuche2481_(m): fidne dich auch nett

Motivation : chat protocols

- **capitalization**

Hübsche1017_(w): Ja ich sage dir das mein Mann zu **hause** wartet. Du ich gehe jetzt. War nett mit dir. Wie wärs mit nächsten Dienstag so um halb11 oder 11 Uhr?
Single9384_(w): **ich** verstehe mich mit jeden **mann**
Speedy24: Bist du **Arbeiten**?
Single9384_(w): **das** meinte ich
AufderSuche2481_(m): **ja** ich probiers
Süsser1201_(m): **ja**, wirklich?
AufderSuche2481_(m): **fidne** dich auch nett

Motivation : chat protocols

- **punctuation**

Hübsche1017_(w): Ja [,] ich sage dir[,]das mein Mann zu hause wartet. Du [,] ich gehe jetzt. War nett mit dir. Wie wärs mit nächsten Dienstag so um halb11 oder 11 Uhr?
Single9384_(w): **ich** verstehe mich mit jeden **mann**
Speedy24: Bist du **Arbeiten**?
Single9384_(w): **das** meinte ich
AufderSuche2481_(m): **ja** [,] ich probiers
Süsser1201_(m): **ja**, wirklich?
AufderSuche2481_(m): **fidne** dich auch nett

Motivation : chat protocols

- **orthography**

Hübsche1017_(w): Ja ich sage dir **das** mein Mann zu hause wartet. Du ich gehe jetzt. War nett mit dir. Wie wärs mit nächsten Dienstag so um halb11 oder 11 Uhr?
Single9384_(w): ich verstehe mich mit jeden mann
Speedy24: Bist du Arbeiten?
Single9384_(w): das meinte ich
AufderSuche2481_(m): ja ich probiers
Süsser1201_(m): ja, wirklich?
AufderSuche2481_(m): **fidne** dich auch nett

Motivation : chat protocols

- **morpho-syntax**

Hübsche1017_(w): Ja ich sage dir das mein Mann zu hause wartet. Du ich gehe jetzt. War nett mit dir. Wie wärs **mit nächsten Dienstag** so um halb11 oder 11 Uhr?
Single9384_(w): ich verstehe mich **mit jeden mann**
Speedy24: Bist du Arbeiten?
Single9384_(w): das meinte ich
AufderSuche2481_(m): ja ich probiers
Süsser1201_(m): ja, wirklich?
AufderSuche2481_(m): fidne dich auch nett

Motivation : chat protocols

- **oral argument drop, vocatives**

Hübsche1017_(w): Ja ich sage dir das mein Mann zu hause wartet. **Du** ich gehe jetzt. **War nett mit dir**. Wie wärs mit nächsten Dienstag so um halb11 oder 11 Uhr?
Single9384_(w): ich verstehe mich mit jeden mann
Speedy24: **Bist du Arbeiten**?
Single9384_(w): das meinte ich
AufderSuche2481_(m): ja ich probiers
Süsser1201_(m): **ja, wirklich?**
AufderSuche2481_(m): **fidne dich auch nett**

Motivation : chat protocols

- **threading** (difficult for coreference)

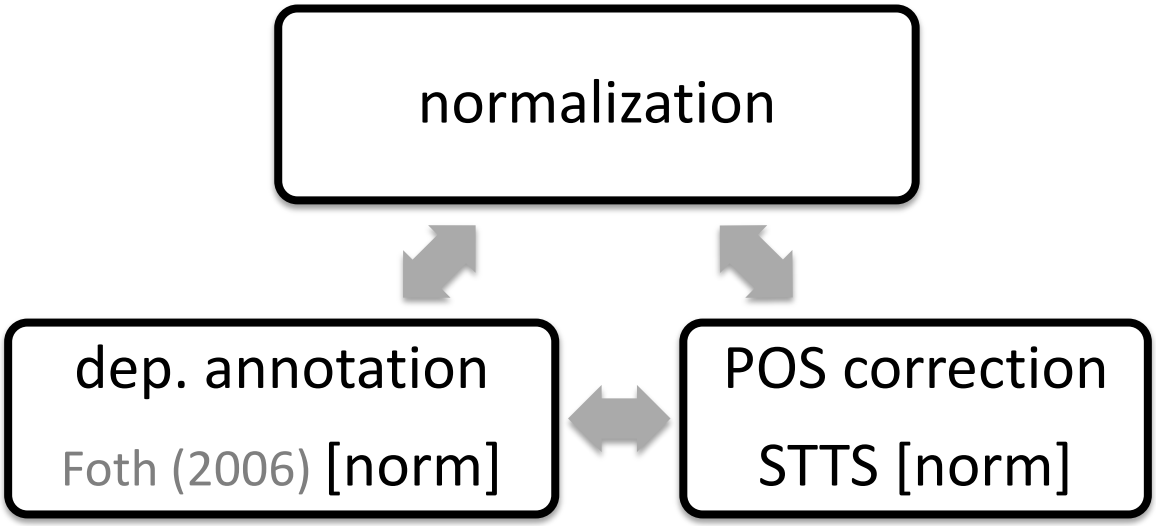
Hübsche1017_(w): Ja ich sage dir das mein Mann zu hause wartet. Du ich gehe jetzt. War nett mit dir. Wie wärs mit nächsten Dienstag so um halb11 oder 11 Uhr?
Single9384_(w): ich verstehe mich mit jeden mann
Speedy24: Bist **du** Arbeiten?
Single9384_(w): **das** meinte ich
AufderSuche2481_(m): ja ich probier**s**
Süsser1201_(m): ja, wirklich?
AufderSuche2481_(m): fidne **dich** auch nett

Motivation: chat protocols

- **verbless sentences**
- **inflectives**
- **asterisk expressions**
- **emoticons**

alles so wie immer :-)
Ich habe kürzlich gegen Riff, 6 halbe Liter Flens verloren, Thor, UND eine Überraschung.
:(((
****Roten Faden wieder in die Hand nehm****
ONKELZ !!!
Ruffputtbles betritt den Raum.
tach ruff

Process:
path to corpus and guidelines



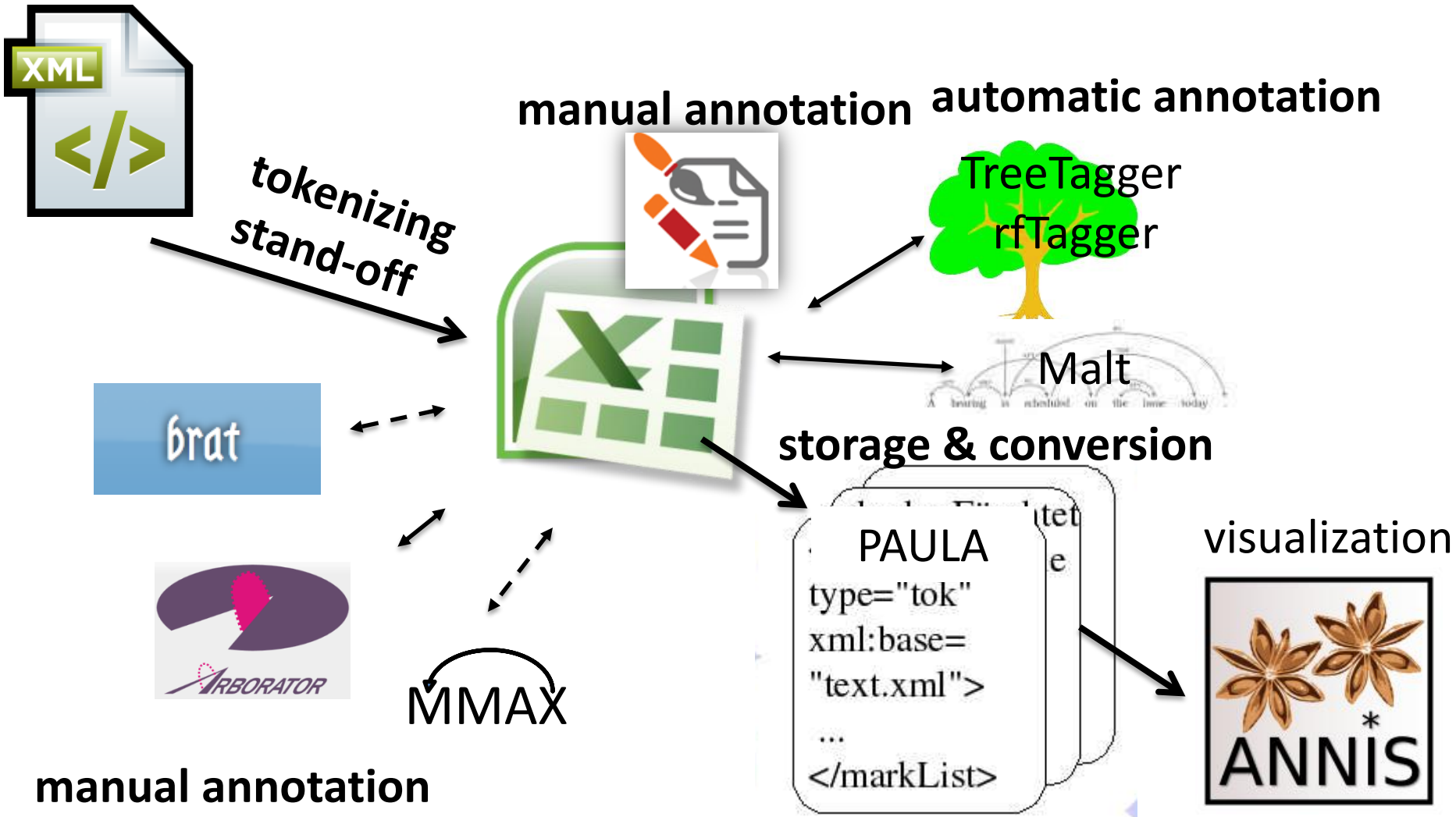
correction of Malt (Nivre 2006) parses (on normalization)

4) description of deviation between norm. and orig. data

5) proposal for changes in annotation schemes

6) evaluation of tools

Annotations tool chain



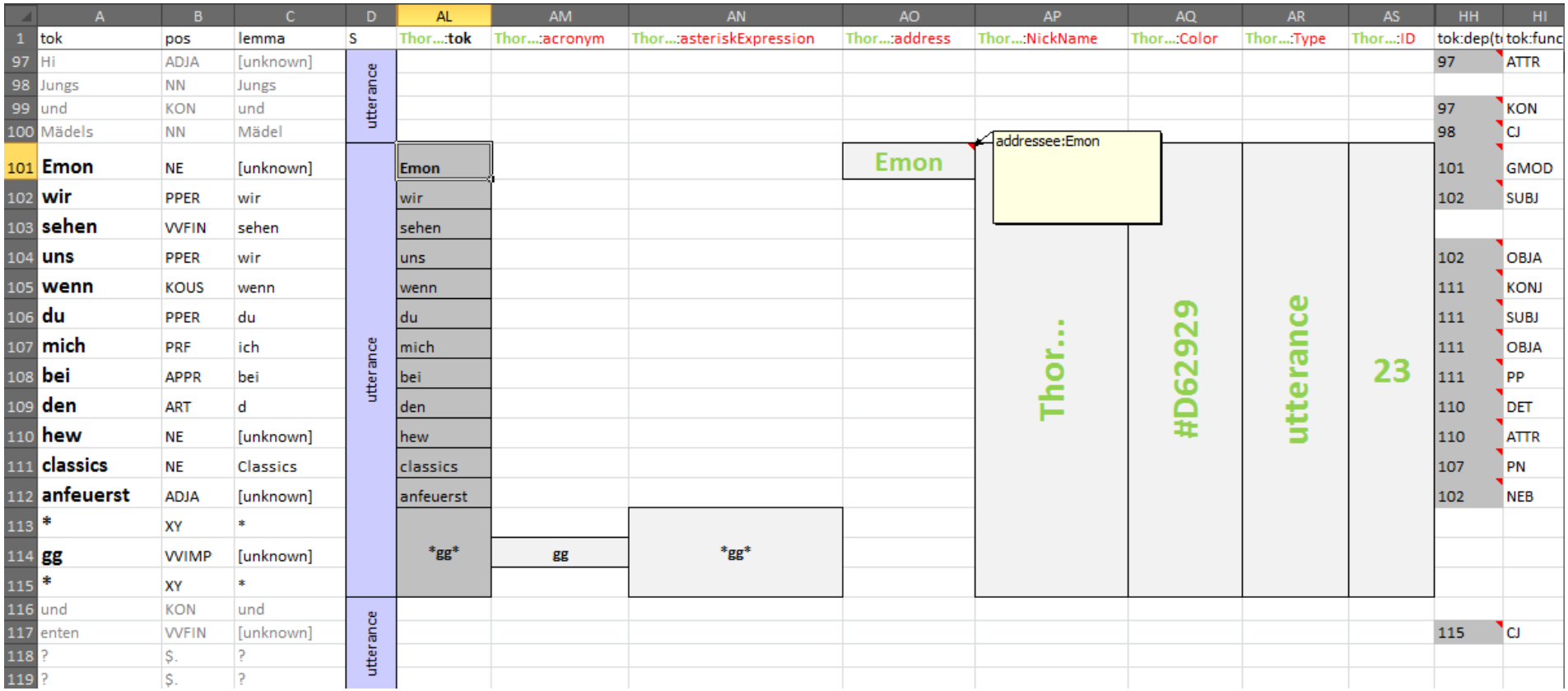
Annotations chat xml

```
<message id="22" type="utterance" creator="süpi" color="#D29552">
        <messageHead>
                <nickname>süpi</nickname>
        </messageHead>
        <messageBody>Hi Jungs und Mädels</messageBody>
</message>

<message id="23" type="utterance" creator="Thor..." color="#D62929">
        <messageHead>
                <nickname>Thor...</nickname>
        </messageHead>
        <messageBody>
                <address addressee="Emon">Emon</address> wir sehen uns wenn du mich bei den
                hew classics  anfeuerst <asteriskExpression>*<acronym>gg</acronym>*
                </asteriskExpression>
        </messageBody>
        </message>

<message id="24" type="utterance" creator="quaki" color="#D62994">
        <messageHead>
                <nickname>quaki</nickname>
        </messageHead>
        <messageBody>und enten??</messageBody>
</message>
```

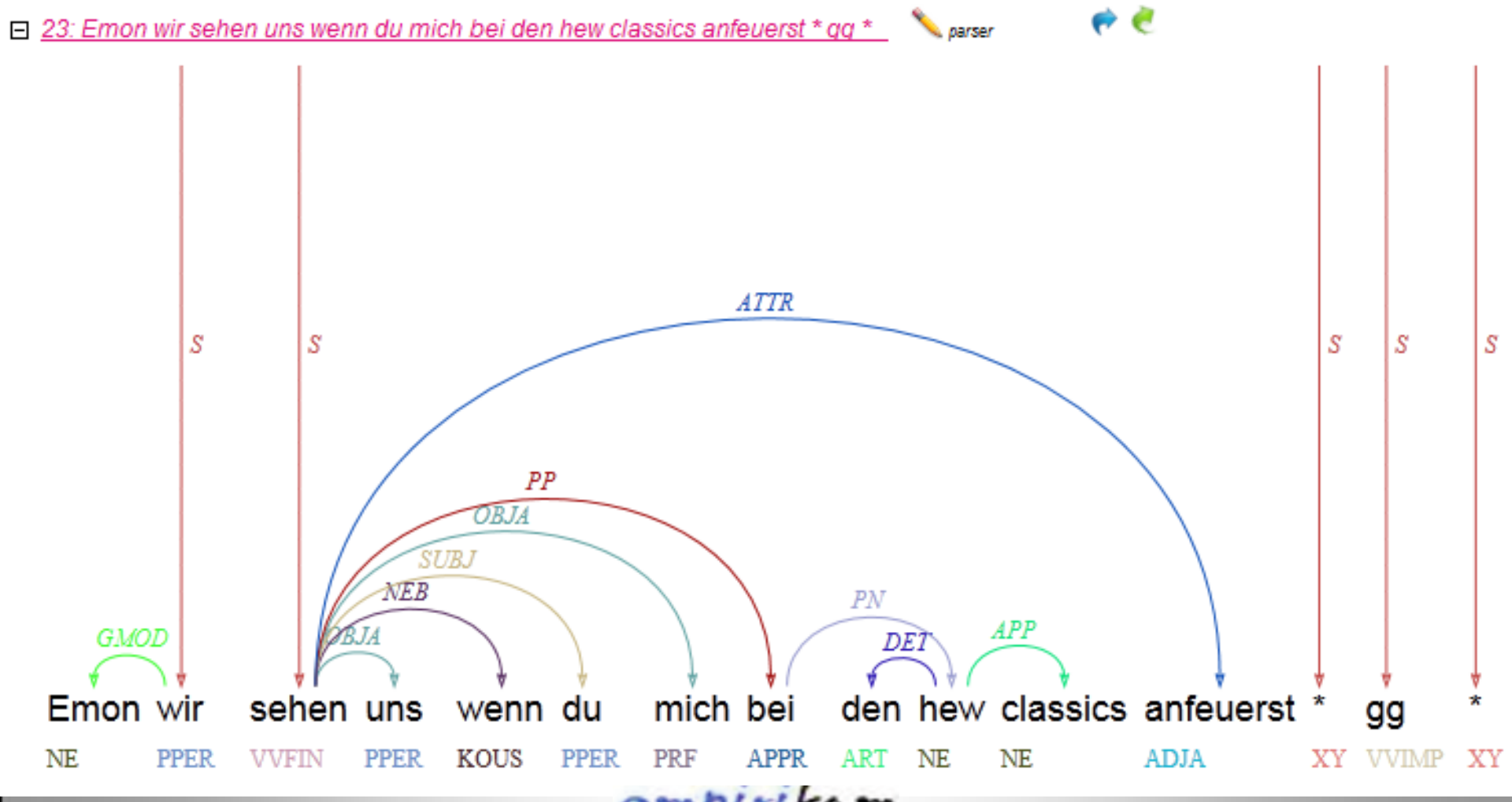
Annotations stand-off



	A	B	C	D	AL	AM	AN	AO	AP	AQ	AR	AS	HH	HI
1	tok	pos	lemma	S	Thor...:tok	Thor...:acronym	Thor...:asteriskExpression	Thor...:address	Thor...:NickName	Thor...:Color	Thor...:Type	Thor...:ID	tok:dep(t	tok:func
97	Hi	ADJA	[unknown]										97	ATTR
98	Jungs	NN	Jungs											
99	und	KON	und										97	KON
100	Mädels	NN	Mädel										98	CJ
101	**Emon**	NE	[unknown]		Emon								101	GMOD
102	**wir**	PPER	wir		wir								102	SUBJ
103	**sehen**	VVFIN	sehen		sehen									
104	**uns**	PPER	wir		uns								102	OBJA
105	**wenn**	KOUS	wenn		wenn								111	KONJ
106	**du**	PPER	du		du								111	SUBJ
107	**mich**	PRF	ich		mich								111	OBJA
108	**bei**	APPR	bei		bei								111	PP
109	**den**	ART	d		den								110	DET
110	**hew**	NE	[unknown]		hew								110	ATTR
111	**classics**	NE	Classics		classics								107	PN
112	**anfeuerst**	ADJA	[unknown]		anfeuerst								102	NEB
113	*	XY	*											
114	**gg**	VVIMP	[unknown]		*gg*	gg	*gg*							
115	*	XY	*											
116	und	KON	und											
117	enten	VVFIN	[unknown]										115	CJ
118	?	$.	?											
119	?	$.	?											

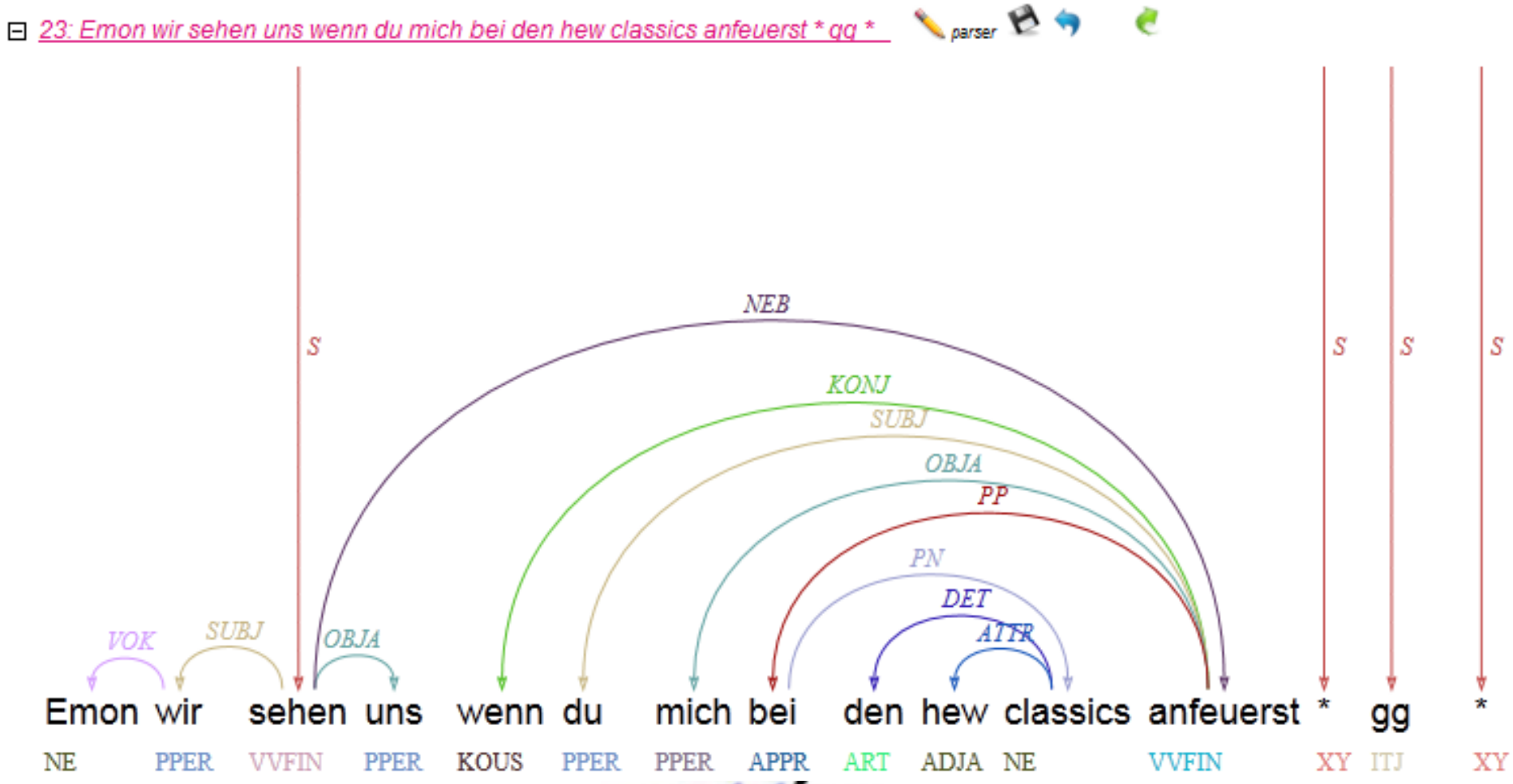
Annotation dependencies: Arborator (soon brat)

- **Malt parses**



Annotation dependencies: Arborator (soon brat)

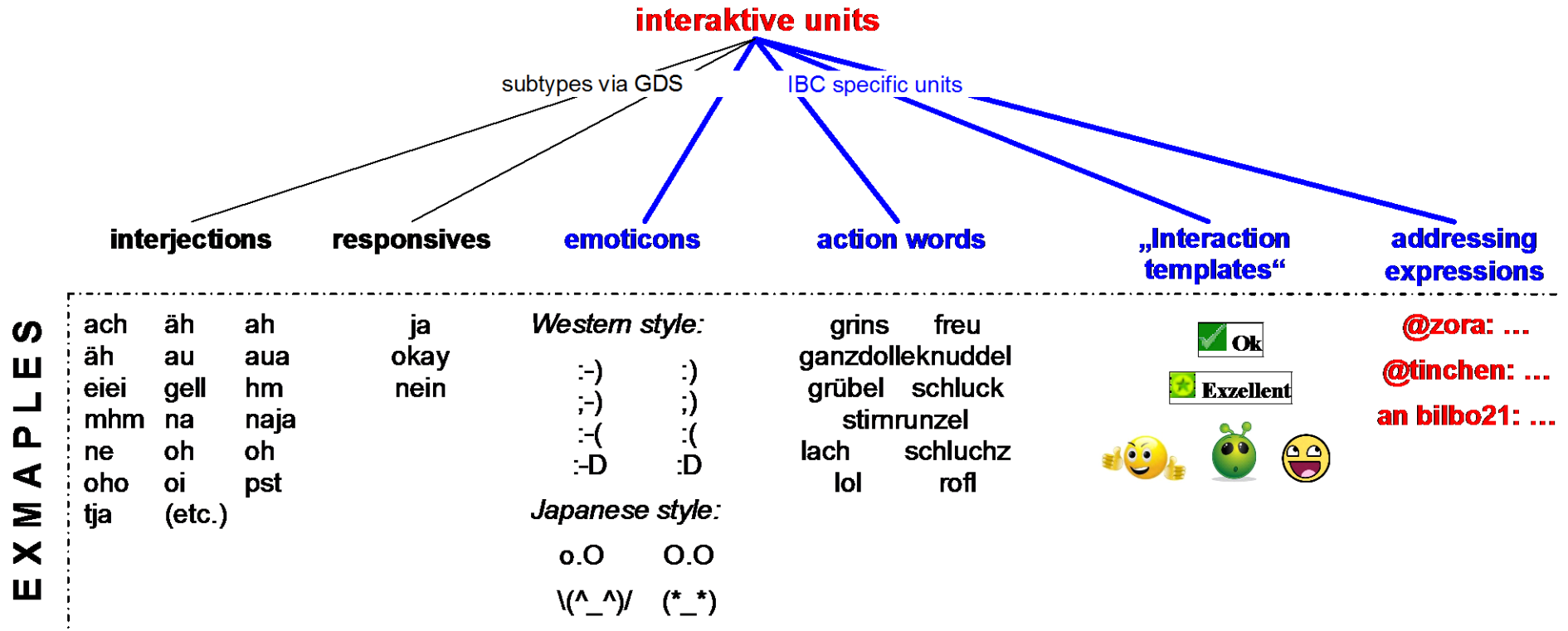
- **corrected parses**



Annotation extension of POS

- **proposal for enlargement of STTS für chat data**

Michael Beißwenger (STTS-Workshop 2012)



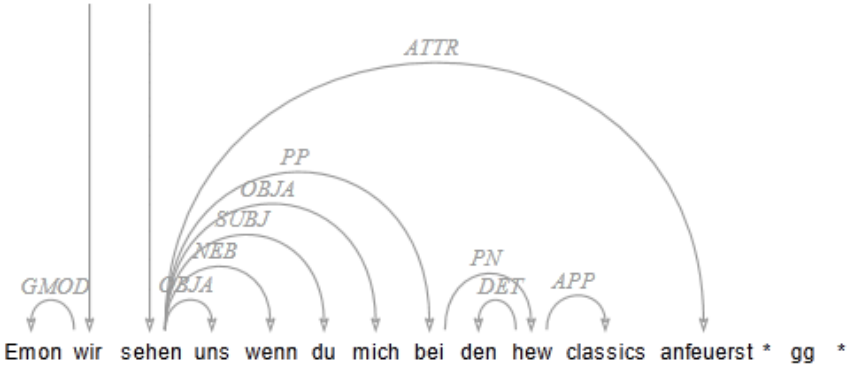
IBK: Data

Path: chat > 2221007_unicum_21-02-2003_2

Emon	wir	sehen	uns	wenn	du	mich	bei	den	hew	classics	anfeuerst	*	gg	*
NE	PPER	VVFIN	PPER	KOUS	PPER	PRF	APPR	ART	NE	NE	ADJA	XY	VVIMP	XY
[unknown]	wir	sehen	wir	wenn	du	ich	bei	d	[unknown]	Classics	[unknown]	*	[unknown]	*

chat (grid)

chat (dependencies)



Malt parses

Search Form

AnnisQL: S="utterance"

Show Result | Query Builder | History

Result: 413

Thor (grid)

Select Displayed Annotation Levels

Color	#D62929													
ID	23,000000													
NickName	Thor...													
Type	utterance													
acronym													gg	
address	Emon													
asteriskExpression												*gg*		
tok	Emon	wir	sehen	uns	wenn	du	mich	bei	den	hew	classics	anfeuerst	*gg*	
tok	Emon	wir	sehen	uns	wenn	du	mich	bei	den	hew	classics	anfeuerst	*	gg	*

empirikom

Visualization: ANNIS2 (soon ANNIS3)

Path: chat > 2221007_unicum_21-02-2003_2

Emon	wir	sehen	uns	wenn	du	mich	bei	den	hew	classics	anfeuerst	*	gg	*	
NE	PPER	VVFIN	PPER	KOUS	PPER	PPER	APPR	ART	NE	NE	VVFIN	XY	ITJ	XY	
[unknown]	wir	sehen	wir	wenn	du	ich	bei	d	[unknown]	Classics	[unknown]	*	[unknown]	*	

chat (grid)

chat (dependencies)

corrected parses

Thor (grid)

Select Displayed Annotation Levels

Color	#D62929														
ID	23,000000														
NickName	Thor...														
Type	utterance														
acronym													gg		
address	Emon														
asteriskExpression													*gg*		
tok	Emon	wir	sehen	uns	wenn	du	mich	bei	den	hew	classics	anfeuerst	*gg*		
tok	Emon	wir	sehen	uns	wenn	du	mich	bei	den	hew	classics	anfeuerst	*	gg	*

Next steps

- Coreference
- Named entities

Thanks to

Stefanie Dipper
Anke Lüdeling
Michael Beißwenger

Data:
Test corpus

Korpus	Textsorte	Norm	POS	NE	Syntax		Koref
					DepRel	GramFkt	
TüBa-D/Z Link	Zeitungs-texte	n.a.	STTS	*organisation person location, geo-political entity*	TüBa	TüBa	Anaph.& Koref. nom. & pron. Antezendenz
DDB – Deutsche Diachrone Baumbank Link	**Historische Texte Fnhd.**	ja	STTS	nein	TIGER	DDB-Liste	nein
Anselm		ja	nein	nein	nein	nein	nein
Bematac	Gesprochene Maptasks	nein	STTS	nein	nein	nein	nein
Dortmund Chat-Korpus Link	Chat-Protokolle	nein	nein	nein	nein	nein	nein
Falko Link	Lerner-aufsätze	mehrere Ziel-hypothesen	STTS	nein	ja	ja	nein
Franz Kafka: Der Prozeß	literarische Prosa	nein	nein	nein	nein	nein	nein