Marc Reznicek
Marc.Reznicek@staff.hu-berlin.de
Hagen Hirschmann
hirschhx@hu-berlin.de

Parsing of Falko

Tübingen-Berlin-Meeting
University of Tübingen
12/06/2011

Overview

- Motivation
- Learner text and target hypoteses
- correction of POS-tags
- parsing the target hypothesis TH1

Motivation

- up -to -date learner corpora allow us to investigate
  - surface structure features
  - lexical patterns (lemma)
  - near surface syntactical patterns (POS)

→ e.g. under-overuse of POS n-grams:

pointers to syntactically interesting structures

example:	pre-nominal	post-nominal
modification	ART-ADJA-NN	vs. NN-APPR-NN

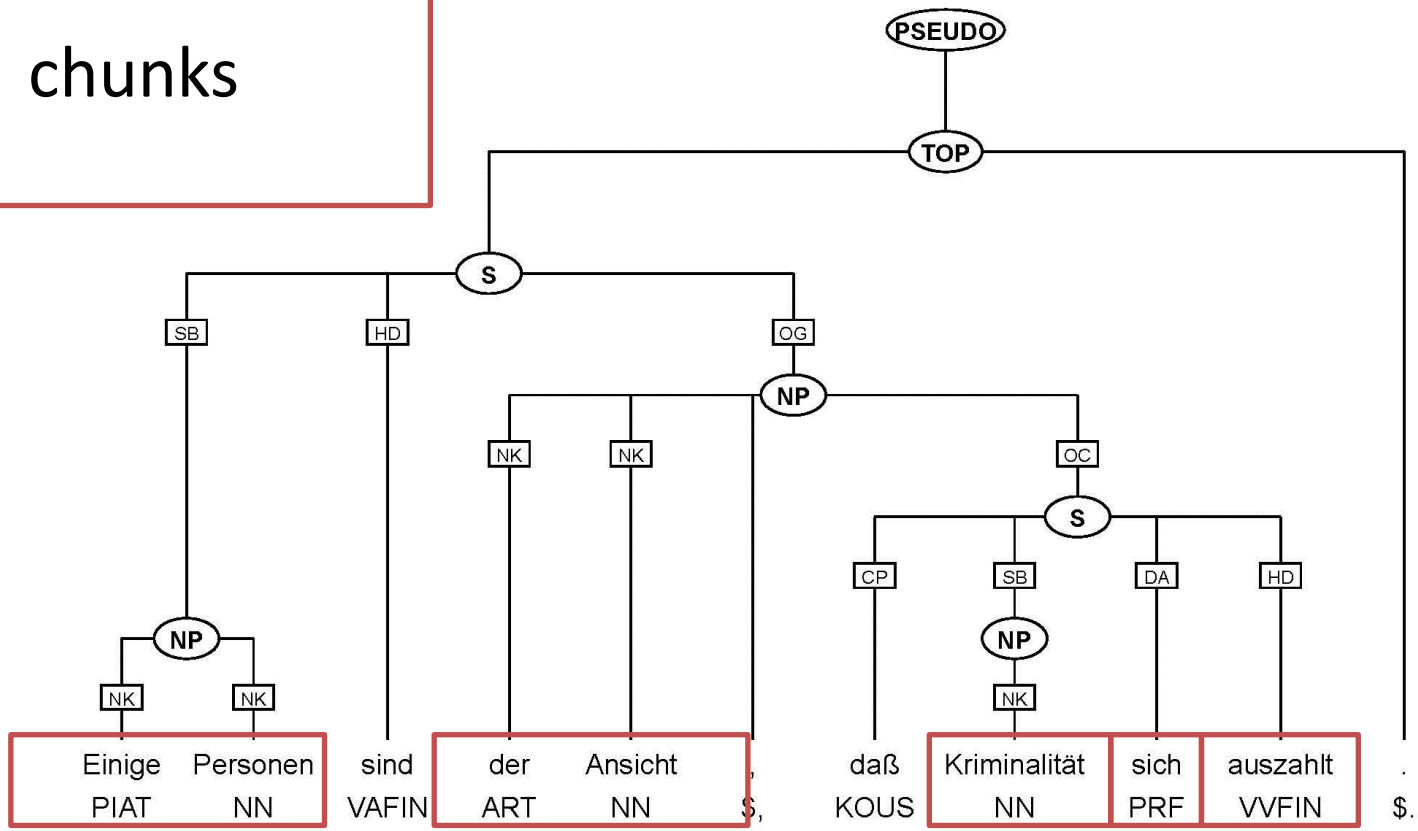
Motivation

- for further insights, we also need ...
  - more abstract linguistic representations and categories
    - topological fields
    - continuous → chunks
    - hieraquical → constituents
    - relations → dependencies
    - functions → grammatical function

Motivation

- Tiger-Scheme includes all those information

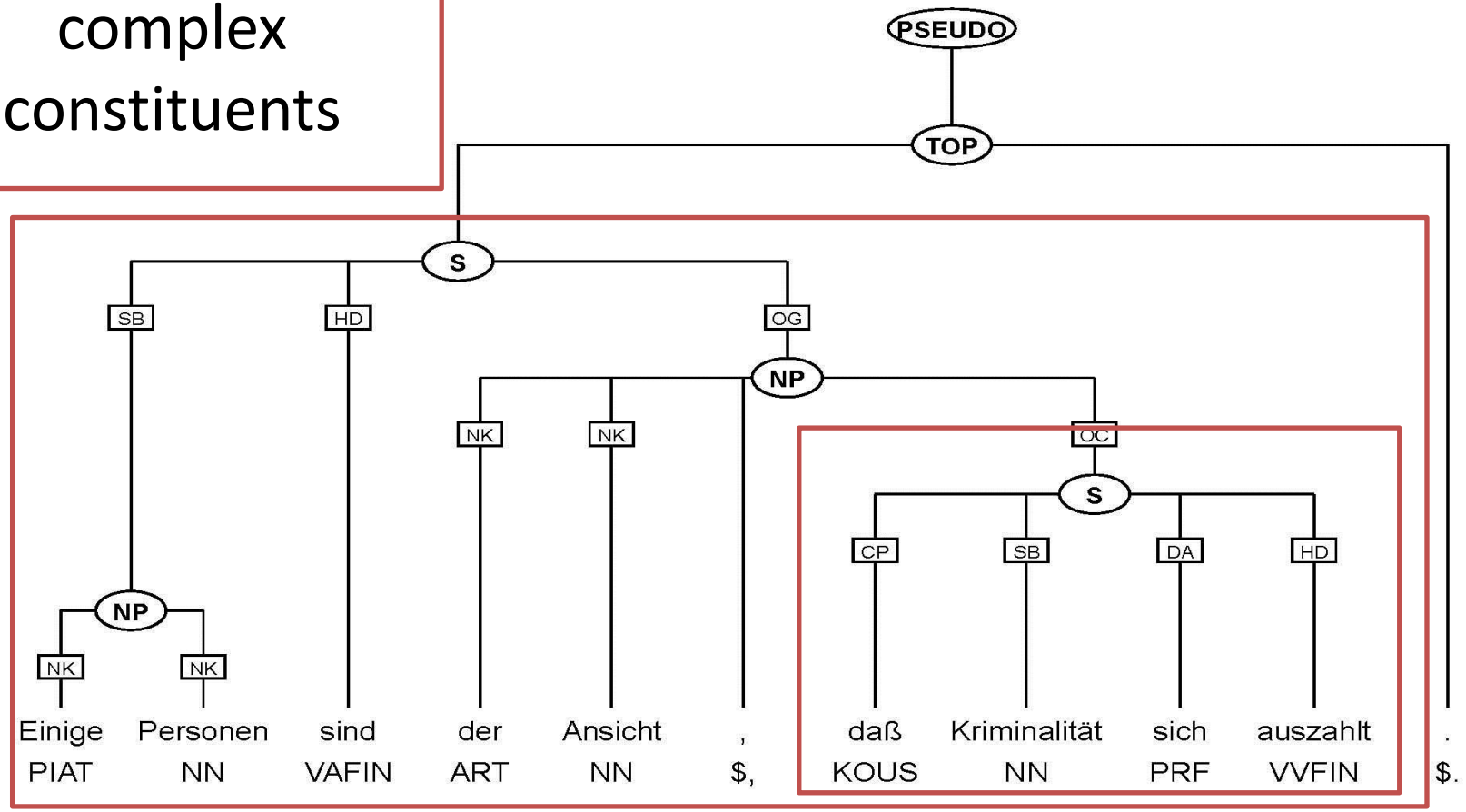
chunks



Motivation

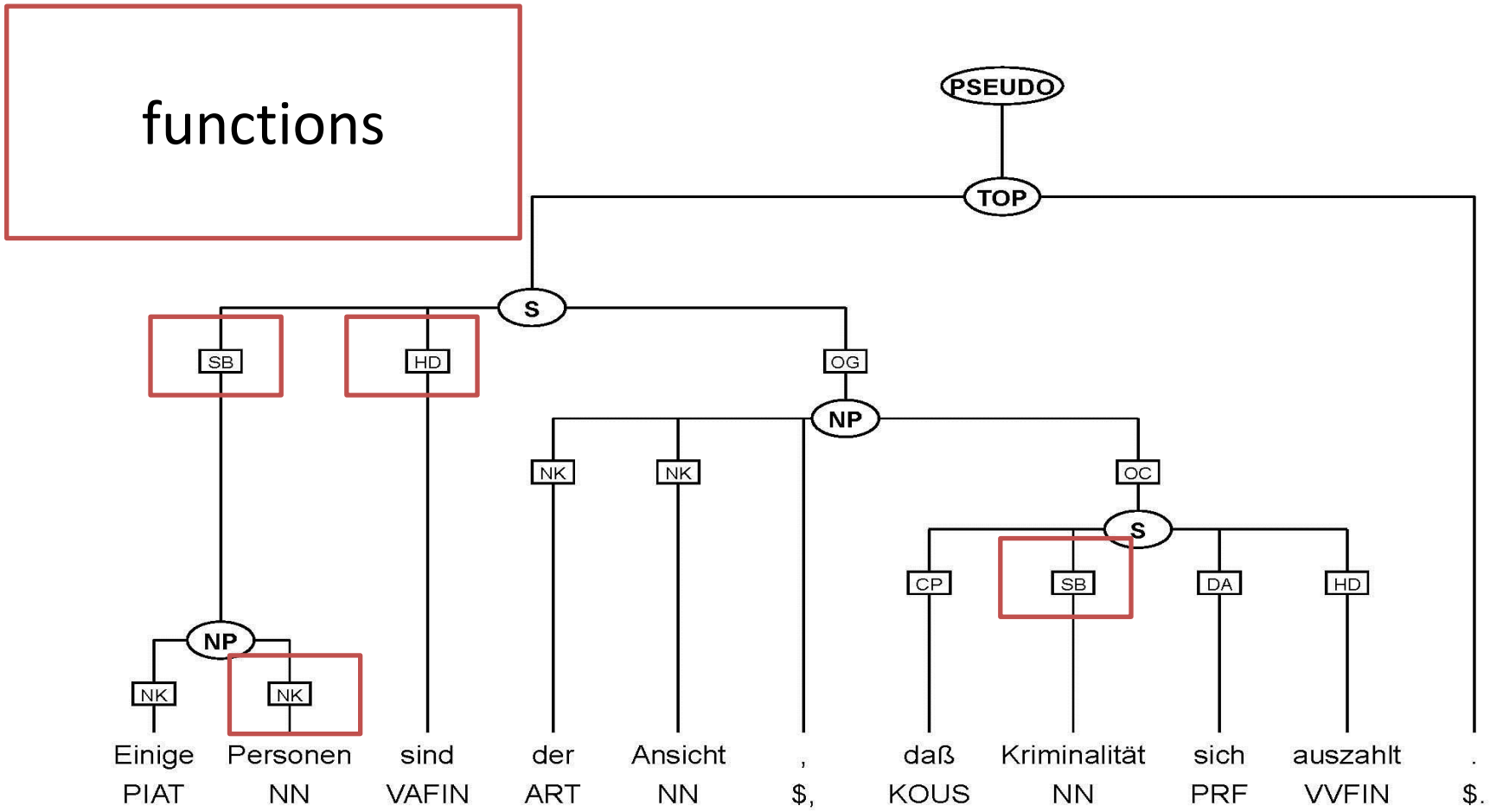
- Tiger-Scheme includes all those information

complex constituents

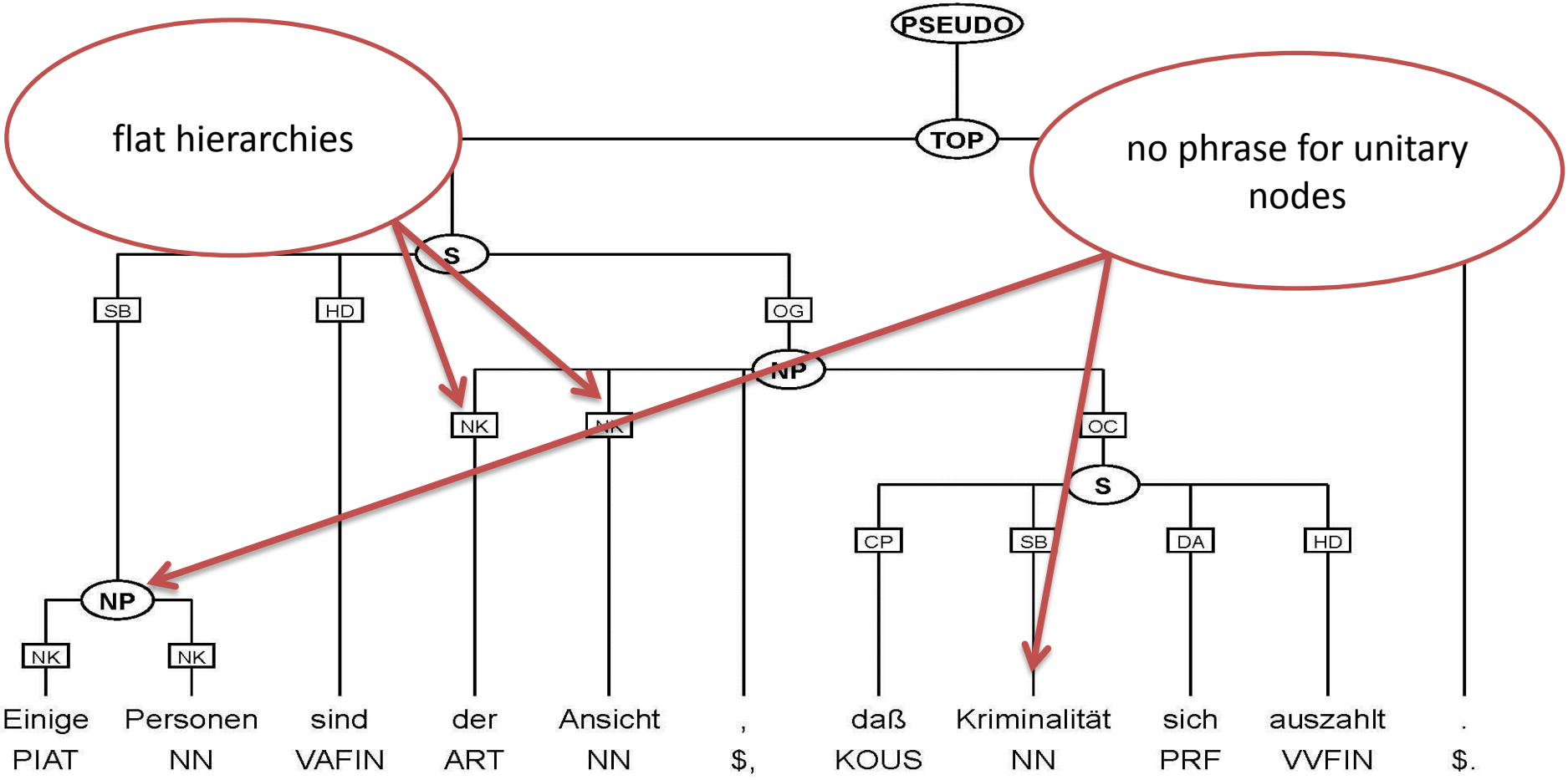


Motivation

- Tiger-Scheme includes most of this information



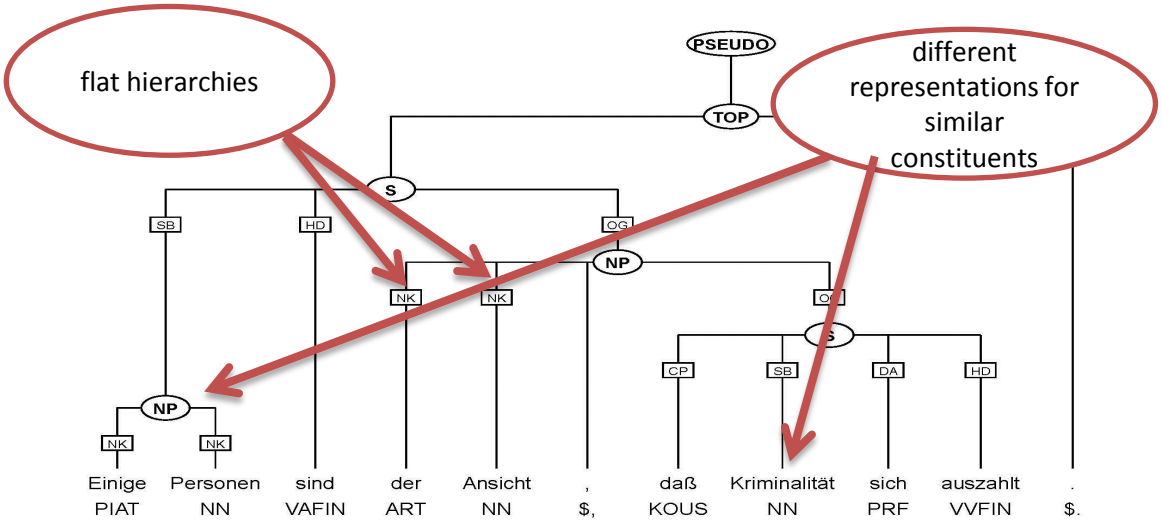
problems with Tiger-scheme



problems with Tiger-scheme

- structures are too complex
- needs a lot of time for valid annotation
- plan: annotation in teaching

→ reduction of complexity



reduction of complexity

our solution:

- separation of different information
  - topological field tagging
  - chunking
  - dependency parsing

reduction of complexity

our solution:

- separation of information of different kinda
  - topological field tagging
  - chunking
  - **dependency parsing**

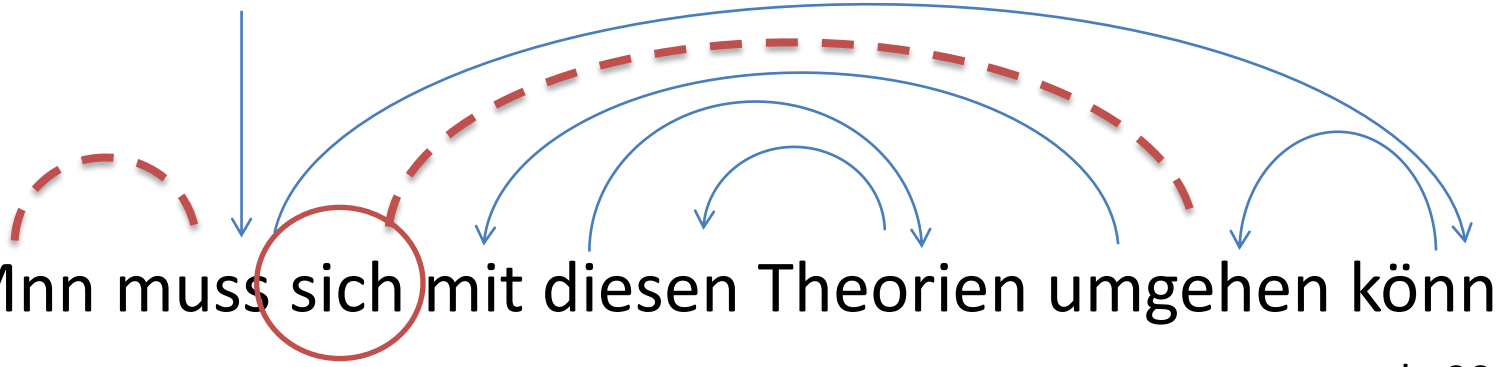
parsing the target hypothesis

Parsing the learner text directly...

... works for cannonical structures....

... runs into problems with spelling errors...

... and fails for uncanonical structures!



cbs001_2006_09

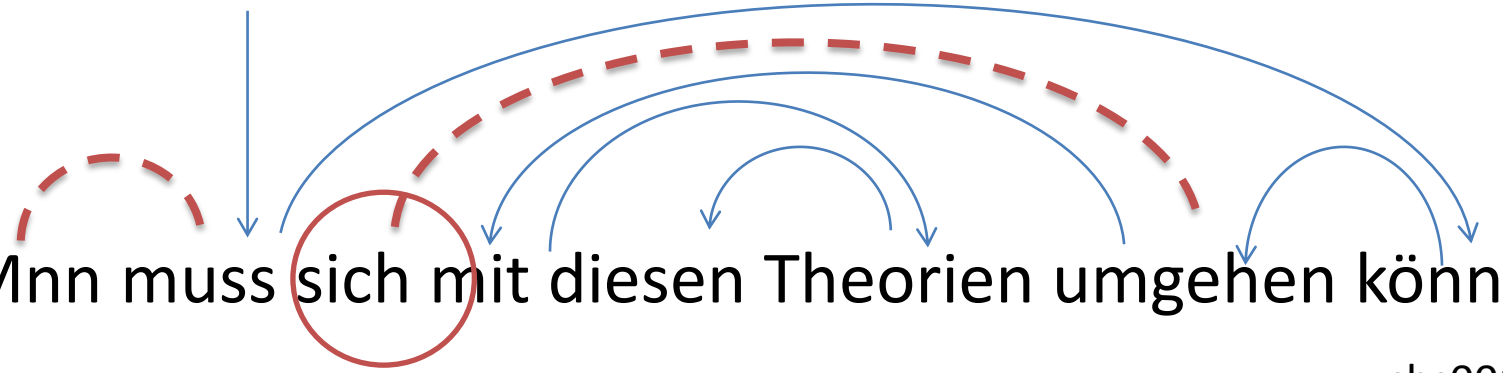
parsing the target hypothesis

We therefore construe a normalisation

→ target hypothesis (TH1)

- minimal grammatical correction
  - orthography, morphology, syntax (rule based)
  - no semantics, no pragmatics, no lexical choice, no style

TH1: Man muss [] mit diesen Theorien umgehen können ...

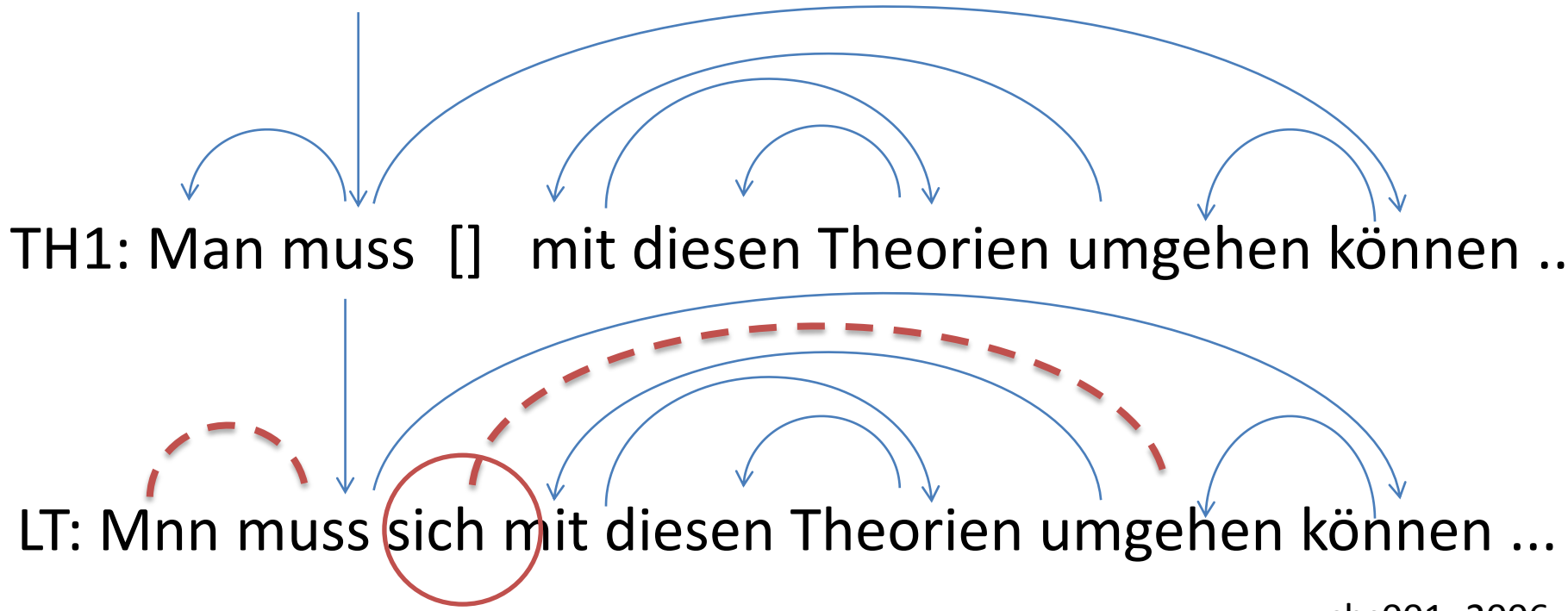


LT: Mnn muss sich mit diesen Theorien umgehen können ...

cbs001_2006_09

parsing the target hypothesis

TH1 can now be parsed according to a standard language grammar



cbs001_2006_09

parsing the target hypothesis

TH1: Man muss [] mit diesen Theorien umgehen können ...

LT: Mnn muss sich mit diesen Theorien umgehen können ...

cbs001_2006_09

better tags give better trees?

- Parses are based on POS tags

TH1:								
	Man	muss		mit	diesen	Theorien	umgehen	können

better tags give better trees?

- Parses are based on POS tags
- assumption:
  - tagging errors lower the quality of the parses

TH1:	NN	VMFIN		APPR	PDAT	NN	VVINF	VVFIN
	Man	muss		mit	diesen	Theorien	umgehen	können

better tags give better trees?

- Parses are based on POS tags
- assumption:
  - tagging errors lower the quality of the parses
- solution:
  - semi-manual correction of TH1-POS-tags
- evaluation:
  - There is no significant difference (Rehbein et al. 2012)

→ it's enough to automatically POS-tag the target hypothesis

	NN	VMFIN		APPR	PDAT	NN	VVINF	VVINF

TH1:	NN	VMFIN		APPR	PDAT	NN	VVINF	VVFIN
	Man	muss		mit	diesen	Theorien	umgehen	können

Tools

- Parsing pos basis: RFTagger output
- Parser: Maltparser
- Learner model: libliniear
- Parsing algorithm: 2planar (2-Planar eager)
- Training data: 48,474 corpus graphs from TiGer
- (Tiger data converted to dependencies (Foth 2003, CWDG) using depsy (Daum et al. 2004))
- CoNLL data format

first evaluation results

- results for Tiger:
- UAS LAS Label acc.
- 85.56 83.43 89.10

- results for Falko L1:
- 80.32 76.97 85.27

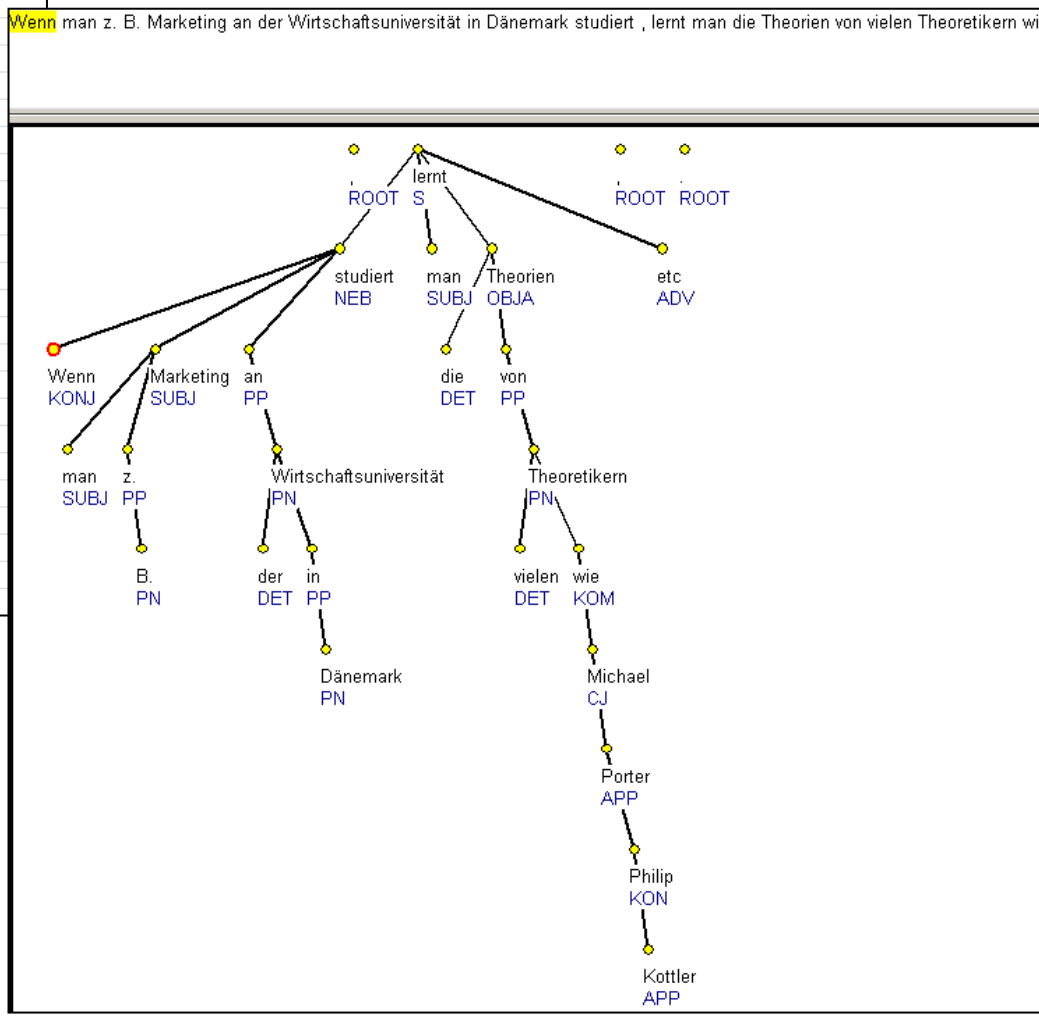
- results for Falko L2:
- 83.30 79.22 86.82

schema & list of edge labels

- constraint dependency grammar (Foth 2006)
- no phrases, exclusively dependencies: relations between terminals
- smaller set than TiGer (37) and TüBaDZ6 (48), but more detailled information than respective sets (mostly due to redundancies between edge labels and STTS tagset)

Data so far

1	Wenn	wenn	KOUS	KOUS	_	11	KONJ
2	man	man	PIS	PIS	_	5	SUBJ
3	z.	z.	APPRART	APPRART	_	5	PP
4	B.	B.	NN	NN	_	3	PN
5	Marketing	Marketing	NN	NN	_	11	SUBJ
6	an	an	APPR	APPR	_	11	PP
7	der	d	ART	ART	_	8	DET
8	Wirtschaftsu	Wirtschaftsu	NN	NN	_	6	PN
9	in	in	APPR	APPR	_	8	PP
10	Dänemark	Dänemark	NE	NE	_	9	PN
11	studiert	studieren	VVFIN	VVFIN	_	13	NEB
12	,	,	$,	$,	_	0	ROOT
13	lernt	lernen	VVFIN	VVFIN	_	0	S
14	man	man	PIS	PIS	_	13	SUBJ
15	die	d	ART	ART	_	16	DET
16	Theorien	Theorie	NN	NN	_	13	OBJA
17	von	von	APPR	APPR	_	16	PP
18	vielen	viel	PIAT	PIAT	_	19	DET
19	Theoretikern	Theoretiker	NN	NN	_	17	PN
20	wie	wie	KOKOM	KOKOM	_	19	KOM
21	Michael	Michael	NE	NE	_	20	CJ
22	Porter	unknown	NE	NE	_	21	APP
23	,	,	$,	$,	_	0	ROOT
24	Philip	Philip	NE	NE	_	22	KON
25	Kottler	unknown	NE	NE	_	24	APP
26	etc.	etc.	ADV	ADV	_	13	ADV



Data correction

- manual corrections in seminars and by experienced annotators
- Editor: Vākyārtha

➢ http://arborator.ilpga.fr/vakyartha/

(Kim Gerdes; http://arborator.ilpga.fr/wiki/Vakyartha)

Edit tags in Falko

tok	pos	lemma
Mnn	NN	[unknown]
muss	VMFIN	müssen
sich	PRF	er|es|sie|Sie
mit	APPR	mit
diesen	PDAT	dies
Theorien	NN	Theorie
umgehen	VVINF	umgehen
können	VMINF	können
um	APPR	um
die	ART	d
Klausuren	NN	Klausur
zu	PTKZU	zu
bestehen	VVINF	bestehen
,	$,	,
aber	ADV	aber
sind	VAFIN	sein
eigentlich	ADV	eigentlich
sie	PPER	sie
nicht	PTKNEG	nicht
praxisorientiert	ADJD	praxisorientiert
.	$.	.

Edit tags in Falko

tok	pos	lemma	ZH1	ZH1gpos	ZH1lemma
Mnn	NN	[unknown]	Man	PIS	man
muss	VMFIN	müssen	muss	VMFIN	müssen
sich	PRF	er|es|sie|Sie			
mit	APPR	mit	mit	APPR	mit
diesen	PDAT	dies	diesen	PDAT	dies
Theorien	NN	Theorie	Theorien	NN	Theorie
umgehen	VVINF	umgehen	umgehen	VVINF	umgehen
können	VMINF	können	können	VMINF	können
			,	$,	,
um	APPR	um	um	KOUI	um
die	ART	d	die	ART	d
Klausuren	NN	Klausur	Klausuren	NN	Klausur
zu	PTKZU	zu	zu	PTKZU	zu
bestehen	VVINF	bestehen	bestehen	VVINF	bestehen
,	$,	,	,	$,	,
aber	ADV	aber	aber	KON	aber
			eigentlich	ADV	eigentlich
sind	VAFIN	sein	sind	VAFIN	sein
eigentlich	ADV	eigentlich			
sie	PPER	sie	sie	PPER	sie
nicht	PTKNEG	nicht	nicht	PTKNEG	nicht
praxisorientiert	ADJD	praxisorientiert	praxisorientiert	ADJD	praxisorientiert
.	$.	.	.	$.	.

Edit tags in Falko

tok	pos	lemma	ZH1	~Diff	ZH1gpos	~Diff	ZH1lemma	~Diff	ZH1S
Mnn	NN	[unknown]	Man	CHA	PIS	CHA	man	CHA	
muss	VMFIN	müssen	muss		VMFIN		müssen		
sich	PRF	er|es|sie|Sie		DEL		DEL		DEL	
mit	APPR	mit	mit		APPR		mit		
diesen	PDAT	dies	diesen		PDAT		dies		
Theorien	NN	Theorie	Theorien		NN		Theorie		
umgehen	VVINF	umgehen	umgehen		VVINF		umgehen		
können	VMINF	können	können		VMINF		können		
			,	INS	$,	INS	,	INS	
um	APPR	um	um		KOUI	CHA	um		
die	ART	d	die		ART		d		
Klausuren	NN	Klausur	Klausuren		NN		Klausur		s9
zu	PTKZU	zu	zu		PTKZU		zu		
bestehen	VVINF	bestehen	bestehen		VVINF		bestehen		
,	$,	,	,		$,		,		
aber	ADV	aber	aber		KON	CHA	aber		
			eigentlich	MOVT	ADV		eigentlich		
sind	VAFIN	sein	sind		VAFIN		sein		
eigentlich	ADV	eigentlich		MOVS					
sie	PPER	sie	sie		PPER		sie		
nicht	PTKNEG	nicht	nicht		PTKNEG		nicht		
praxisorientiert	ADJD	praxisorientiert	praxisorientiert		ADJD		praxisorientiert		
.	$.	.	.		$.		.		

Merge parses

tok	ZH1	~Diff	ZH1gpos	ZH1Knoten	ZH1Kante	ZH1Funktion	ZH1S
Mnn	Man	CHA	PIS	1	2	SUB	
muss	muss		VMFIN	2	0	ROOT	
sich		DEL					
mit	mit		APPR	3	6	PP	
diesen	diesen		PDAT	4	5	DET	
Theorien	Theorien		NN	5	3	PN	
umgehen	umgehen		VVINF	6	7	AUX	
können	können		VMINF	7	2	AUX	
	,	INS	$,	8	0	ROOT	
um	um		KOUI	9	13	KONJ	
die	die		ART	10	11	DET	
Klausuren	Klausuren		NN	11	13	OBJA	s9
zu	zu		PTKZU	12	13	PART	
bestehen	bestehen		VVINF	13	2	NEB	
,	,		$,	4	0	ROOT	
aber	aber		KON	15	2	KON	
	eigentlich	MOVT	ADV	16	17	ADV	
sind	sind		VAFIN	17	15	CJ	
eigentlich		MOVS					
sie	sie		PPER	18	17	SUBJ	
nicht	nicht		PTKNEG	19	17	ADV	
praxisorientiert	praxisorientiert		ADJD	20	17	PRED	
.	.		$.	21	0	ROOT	

ANNIS2 merging information

Path: FalkoEssayL2v2.0 > cbs001_2006_09_L2v2.0

Mann	muss	sich	mit	diesen	Theorien	umgehen	können	um	die	Klausuren	zu	bestehen	,	aber	eigentlich	sind	sie	nicht	praxisorientiert	.
NN	VMFIN	PRF	APPR	PDAT	NN	VVINF	VMINF	APPR	ART	NN	PTKZU	VVINF	$,	ADV	ADV	VAFIN	PPER	PTKNEG	ADJD	$.
Mann	müssen	er|es|sie|Sie	mit	dies	Theorie	umgehen	können	um	d	Klausur	zu	bestehen	,	aber	eigentlich	sein	sie	nicht	praxisorientiert	.

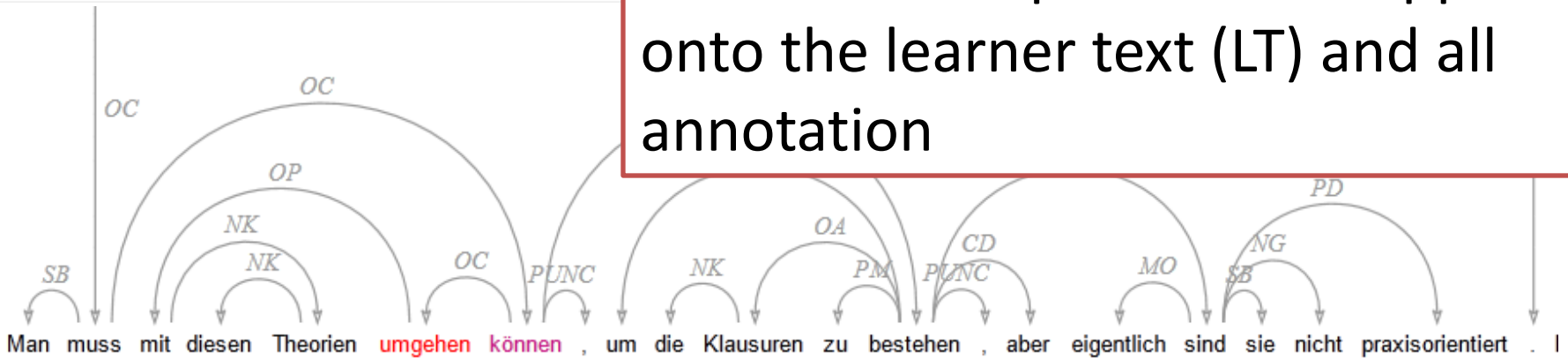
ZH2 (grid)
learner (grid)
falko (grid)
ZHverb (grid)
ZH1 (grid)

Select Displayed Annotation Levels

ZH1	Man	muss		mit	diesen	Theorien	umgehen	können	,	um	die	Klausuren	zu	bestehen	,
ZH1Diff	CHA		DEL						INS						
ZH1lemma	man	müssen		mit	dies	Theorie	umgehen	können	,	um	d	Klausur	zu	bestehen	,
ZH1pos	PIS	VMFIN		APPR	PDAT	NN	VVINF	VMINF	$,	KOUI	ART	NN	PTKZU	VVINF	$,
tok	Mann	muss	sich	mit	diesen	Theorien	umgehen	können		um	die	Klausuren	zu	bestehen	,

errors (grid)
text (discourse)

In ANNIS2 the parses are mapped onto the learner text (LT) and all annotation



ANNIS2 merging information

Path: FalkoEssayL2v2.0 > cbs001_2006_09_L2v2.0

Mann	muss	sich	mit	diesen	Theorien	umgehen	können	um	die	Klausuren	zu	bestehen	,	aber	eigentlich	sind	sie	nicht	praxisorientiert	.
NN	VMFIN	PRF	APPR	PDAT	NN	VVINF	VMINF	APPR	ART	NN	PTKZU	VVINF	$,	ADV	ADV	VAFIN	PPER	PTKNEG	ADJD	$.
Mann	müssen	er|es|sie|Sie	mit	dies	Theorie	umgehen	können	um	d	Klausur	zu	bestehen	,	aber	eigentlich	sein	sie	nicht	praxisorientiert	.

ZH2 (grid)
learner (grid)
falko (grid)
ZHverb (grid)
ZH1 (grid)

Select Displayed Annotation Levels

ZH1	Man	muss		mit	diesen	Theorien	umgehen	können	,	um	die	Klausuren	zu	bestehen	,	
ZH1Diff	CHA		DEL						INS							
ZH1lemma	man	müssen		mit	dies	Theorie	umgehen	können	,	um	d	Klausur	zu	bestehen	,	
ZH1pos	PIS	VMFIN		APPR	PDAT	NN	VVINF	VMINF	$,	KOUI	ART	NN	PTKZU	VVINF	$,	
tok	Mann	muss	sich	mit	diesen	Theorien	umgehen	können		um	die	Klausuren	zu	bestehen	,	

errors (grid)
text (discourse)

DEL → LT-tokens have no parses

ANNIS2 merging information

Path: FalkoEssayL2v2.0 > cbs001_2006_09_L2v2.0

Mann	muss	sich	mit	diesen	Theorien	umgehen	können	um	die	Klausuren	zu	bestehen	,	aber	eigentlich	sind	sie	nicht	praxisorientiert	.
NN	VMFIN	PRF	APPR	PDAT	NN	VVINF	VMINF	APPR	ART	NN	PTKZU	VVINF	$,	ADV	ADV	VAFIN	PPER	PTKNEG	ADJD	$.
Mann	müssen	er|es|sie|Sie	mit	dies	Theorie	umgehen	können	um	d	Klausur	zu	bestehen	,	aber	eigentlich	sein	sie	nicht	praxisorientiert	.

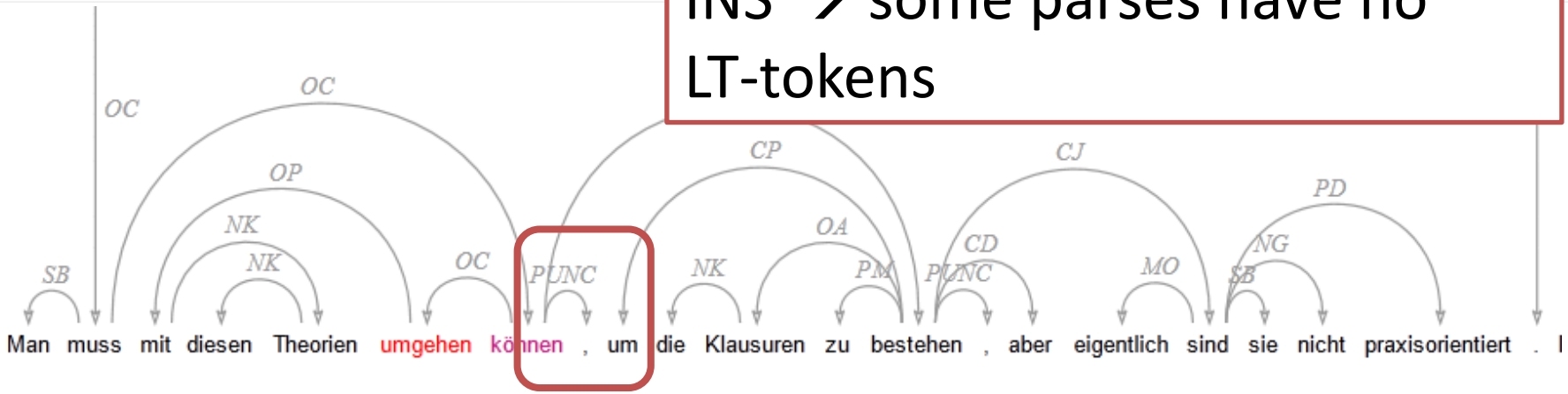
ZH2 (grid)
learner (grid)
falko (grid)
ZHverb (grid)
ZH1 (grid)

Select Displayed Annotation Levels

ZH1	Man	muss		mit	diesen	Theorien	umgehen	können	,	um	die	Klausuren	zu	bestehen	,	
ZH1Diff	CHA		DEL						INS							
ZH1lemma	man	müssen		mit	dies	Theorie	umgehen	können	,	um	d	Klausur	zu	bestehen	,	
ZH1pos	PIS	VMFIN		APPR	PDAT	NN	VVINF	VMINF	$,	KOUI	ART	NN	PTKZU	VVINF	$,	
tok	Mann	muss	sich	mit	diesen	Theorien	umgehen	können		um	die	Klausuren	zu	bestehen	,	

errors (grid)
text (discourse)

INS → some parses have no LT-tokens



ANNIS2 merging information

Path: FalkoEssayL2v2.0 > cbs001_2006_09_L2v2.0

Mann	muss	sich	mit	diesen	Theorien	umgehen	können	um	die	Klausuren	zu	bestehen	,	aber	eigentlich	sind	sie	nicht	praxisorientiert	.
NN	VMFIN	PRF	APPR	PDAT	NN	VVINF	VMINF	APPR	ART	NN	PTKZU	VVINF	$,	ADV	ADV	VAFIN	PPER	PTKNEG	ADJD	$.
Mann	müssen	er|es|sie|Sie	mit	dies	Theorie	umgehen	können	um	d	Klausur	zu	bestehen	,	aber	eigentlich	sein	sie	nicht	praxisorientiert	.

ZH2 (grid)
learner (grid)
falko (grid)
ZHverb (grid)
ZH1 (grid)

Select Displayed Annotation Levels

ZH1	Man	muss		mit	diesen	Theorien	umgehen	können	,	um	die	Klausuren	zu	bestehen	,
ZH1Diff	CHA		DEL						INS						
ZH1lemma	man	müssen		mit	dies	Theorie	umgehen	können	,	um	d	Klausur	zu	bestehen	,
ZH1pos	PIS	VMFIN		APPR	PDAT	NN	VVINF	VMINF	$,	KOUI	ART	NN	PTKZU	VVINF	$,
tok	Mann	muss	sich	mit	diesen	Theorien	umgehen	können		um	die	Klausuren	zu	bestehen	,

errors (grid)
text (discourse)

INS → some parses have no LT-tokens

Open questions

- Alter schema?
- ➢ Retraining of parser possible?
- How to deal with empty slots in Falko ctok data?

Thank You!

Parses condcuted by Ines Rehbein
Additional assistence Yannick Versley

email: hirschhx@hu-berlin.de

Literatur

- Díaz-Negrillo, Ana; Meurers, Detmar; Valera, Salvador & Wunsch, Holger (2010) Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. In: *Language Forum* 36(1-2). Special Issue on New Trends in Language Teaching, edited by Carmen Pérez Basanta.
- Granger, Sylviane (2008) Learner corpora. In: Anke Lüdeling & Merja Kytö (eds) *Corpus Linguistics. An International Handbook*. Vol 1. Mouton de Gruyter, Berlin, 259-275.
- Hirschmann, H.; Doolittle, S. & Lüdeling, A. (2007) Syntactic annotation of non-canonical linguistic structures. In: Proceedings of Corpus Linguistics 2007, Birmingham.
- Lorenz, G. (1998). Overstatement in advanced learners' writing: Stylistic aspects of adjective intensification. In S. Granger (Ed.), *Learner English on computer* (pp. 53– 66). London: Longman.
- Lüdeling, Anke; Doolittle, Seanna; Hirschmann, Hagen; Schmidt, Karin & Walter, Maik (2008) Das Lernerkorpus Falko. In: Deutsch als Fremdsprache 2, 67-73.
- Möllering, Martina (2004) *The acquisition of German modal particles*. Lang, Bern.
- Tenfjord, Kari; Hagen, Jon Erik; Johansen, Hilde (2006) The hows and whys of coding categories in a learner corpus(or How and why an error-tagged learner corpus is not ipso facto one big comparative fallacy). In: *Rivista di Psicolinguistica Applicata* (RiPLA) VI.(3), 93-108
- Rosén, Victoria & de Smedt, Koenraad (to appear) Syntactic Annotation of Learner Corpora. In: Hilde Johansen, Anne Golden, Jon Erik Hagen and Ann-Kristin Helland (eds.), Systematisk, variert, men ikke tilfeldig. Antologi om norsk som andrespråk i anledning Kari Tenfjords 60-årsdag (Systematic, varied, but not arbitrary. Anthology about Norwegian as a second language on the occasion of Kari Tenfjord's 60th birthday), Novus forlag, Oslo, 2010.
- Vyatkina, Nina (2007) Development of second language pragmatic competence: The data-driven teaching of German modal particles based on a learner corpus. Dissertation, Pennsylvania State University.
- Maden-Weinberger. (2009) *Modality in Learner German: A corpus-based study investigating modal expressions in argumentative texts by British learners of German*. Doctoral Thesis. Lancaster University: Department of Linguistics and English Language.