Multilevel Learner Corpora

Humboldt-Universität zu Berlin

Hagen Hirschmann
hirschhx@rz.hu-berlin.de
Amir Zeldes
amir.zeldes@rz.hu-berlin.de
Anke Lüdeling
anke.luedeling@rz.hu-berlin.de

AAL/CALICO 2009

Tempe, March 10th

Overview

- Advantages of multi-level corpus architectures
- Relevance for learner corpora and learner studies
  - Error annotation & target hypotheses
  - Contrastive Interlanguage Analysis
- Outlook: Falko in Annis, a multilevel search tool

Corpus architectures

Inline

- Large (standard) corpora
- Good/fast search tools available
- Difficult to add annotation
- Difficult/impossible to represent conflicting annotation

Multilevel (standoff)

- Developed for multimodal and small/specific corpora
- Few tools available (many under development); annotation tools better than search tools
- Annotation layers can be added unrestrictedly (without changing old data)
- No problem to represent conflicting annotation

Carletta et al. 2005, Wittenburg 2008, Chiarcos et al. 2009, …

Data: the Falko corpus

- Falko (**f**ehler**a**nnotiertes **L**erner**ko**rpus), freely available multilevel learner corpus (Lüdeling et al. 2008)
- Different subcorpora
  - Summaries (Free University & Humboldt University, Berlin), L2 (many different mother tongues) & L1
  - Essays (Free University & Humboldt University, Berlin) L2 (many different mother tongues) & L1
  - Longitudinal corpus (Georgetown University), L2 (English ns)
- Automatic pos tagging and lemmatization (TreeTagger, Schmid 1994), partly manually corrected; summaries and longitudinal data topologically annotated (Doolittle 2008)

Learner corpus studies

- 2 basic approaches (Selinker 1972; Ringbom 1998; Granger et al. 2002):
  - □ Error Analysis (EA studies)
  - □ Contrastive Interlanguage Analysis (CIA)

1. Error analysis Ambiguity of errors and EA

	was		die	Novelle	oder	die	Ode	nicht	betrift
	what		the	novella	or	the	ode	not	effects
	which does not effect the novella or the ode								
A1									LEX SP
A2									SP

- Errors **are** potentially ambiguous (↑Adriane Boyd last talk tomorrow)
- How do we detect ambiguities?
- Need for transparent error analyses

1. Error analysis
Ambiguity of errors and EA

	was		die	Novelle	oder	der	Ode	nicht	betrift
	what		the	novella	or	the	ode	not	effects
	which does not effect the novella or the ode								
TH1	was	auf	die	Novelle	oder	die	Ode	nicht	zutrifft
EA1									LEX SP
TH2	was		die	Novelle	oder	die	Ode	nicht	betrifft
EA2									SP

Target hypothesis: experiment

- 5 annotations for 17 sentences (one text) (Lüdeling 2008)
- Annotation scheme identical
- Error annotations differ:

content words	function words
15	13
24	**26**
17	25
16	**12**
14	22

Conclusion target hypothesis

- Target hypothesis must be explicit/available
- It must be possible to formulate several target hypotheses for the same data
- It must be possible to formulate different analyses for the same target hypothesis (error tags)

Multiple levels → conflicts

word	He	awaited	for	his	wife
phrase	NP		PP		
targ		waited			
targ		awaited			

- Inline annotation cannot deal with these conflicting annotation spans

Multiple levels → conflicts

- Annotating errors and PP objects simultaneously in inline XML:
- <NP>He</NP> <err target="waited"> <err target="awaited"> awaited</err> <PP> for </err> his wife</PP>

He	awaited	for	his	wife
NP		PP		
	waited			
	awaited			

Summary & Conclusion

- EA and target hypothesis
  - Need for competing annotations
  - Need for conflicting annotations
- ADV case study:

2. Contrastive Interlanguage Analysis (CIA)

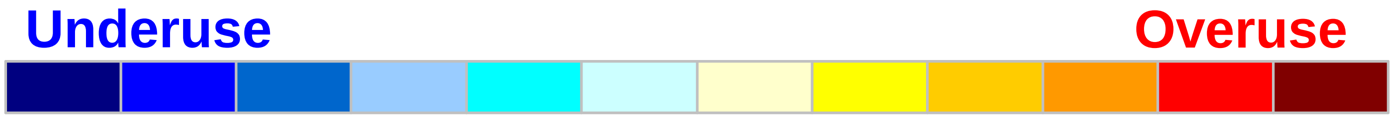
- Assumption 1: Learners have systematic interlanguage (interim language) (Selinker 1972, Corder 1981, Jordens 2003 etc.)
- Assumption 2: Interlanguage has reflexes in the observable data
- Method: Compare L2 with L1 varieties (Cobb 2003, Tono 2003, Granger 2008, Walter & Grommes 2008, Mukherjee 2008 etc.)

2. Contrastive Interlanguage Analysis (CIA)

- Differences between varieties can be expressed in terms of frequency differences
- Over- and underuse studies
  - Esp. underuse can indicate learner difficulties
  - e.g. comparing frequencies of
    - Individual lexemes (content or function words)
    - Phrase structures
    - Pos (chains)

2. CIA

- Normalized frequencies in all Falko subcorpora (L2/L1) of pos n-grams
- Strength of over- and underuse is color-coded



Detecting structural syntactic difficulties: Pos chains

bigram	de	da	en	fr	pl	ru
$.-PPER	0.005297	0.009748	0.007963	0.006166	0.005801	0.007409
VVFIN-$,	0.006457	0.00776	0.006343	0.006937	0.006243	0.008391
PPOSAT-NN	0.008058	0.007247	0.007269	0.007066	0.006298	0.005802
ADV-ADV	**0.012858**	**0.010518**	**0.006111**	**0.006166**	**0.003094**	**0.002856**
ADV-APPR	0.009117	0.008016	0.005324	0.007837	0.004807	0.004642
PDAT-NN	0.005409	0.004233	0.005509	0.007837	0.007735	0.008837
ADV-ART	0.007629	0.006349	0.006898	0.005653	0.006133	0.004463

Consecutive adverbs are underused by all learners independent of their L1

Excel Under/Overuse Addin: http://korpling.german.hu-berlin.de/~amir/uoaddin.htm

ADV underuse case study

- ➢ ADV underuse characteristic of advanced learner variety; ADV-ADV underuse is significantly higher than underuse of single ADVs predicts
- ■ Why?
- ■ More precisely:
  - □ Are there specifically hard ADV categories and combinations of them?
  - □ Does underuse depend on complexity of ADV-ADV chains?

ADV underuse case study

- How far do we get with what we have?
- Available: surface forms, pos annotation, lemmatization
- Over-/underuse method applicable for individual lexemes:

ADV underuse case study

type		FK_Ess_L1		FK_Ess_L2	/FK_Ess_L1
immer noch *still*		2,3694		2,30485556	0,9727787
nur noch *only still*		4,4425		0,65853016	0,14823294
immer wieder *again and again*		3,2579		2,41461059	0,74116472
heute noch *today still*		1,4808		0,21951005	0,14823294
noch immer *still*		0,2962		0,21951005	0,74116472
auch noch *also still*		0,8885		0,87804021	0,98821963
immer mehr *increasingly more*		3,7021		0,43902011	0,11858636
sehr viel *very much*		0,2962		1,20730529	4,07640596

ADV underuse case study

- Measuring relative frequency of individual lexemes is easy
- Results: "Combinations with *einmal* 'once' (*xxx einmal*) are among the most underused productive L1 bigrams"
- However: No insight into syntactic structures or categories; hard to define syntactic classes
- Pos tag 'ADV' represents a heterogeneous class
  - (lexical) phrase particles (intensifiers, focus particles)
  - Verbal phrase adverbs (*bald* - *soon*)
  - Sentence adverbs (*eigentlich* - *actually*)
  - Sentence/modal particles (*wohl, doch, ja* - ??? (*well*))
- Many of the lexemes occur in more than one class

Examples from learner data (Falko)

word	und	immer	noch	kann	man	eine	unzufriedenheit	spüren
apos	KON	ADV	ADV	VMFIN	PIS	ART	unknown	VVINF
cpos	KON	ADV	ADV	VMFIN	PIS	ART	NN	VVINF
lemma	und	immer	noch	können	man	ein	unknown	spüren

(And still you can feel some dissatisfaction)

'still'

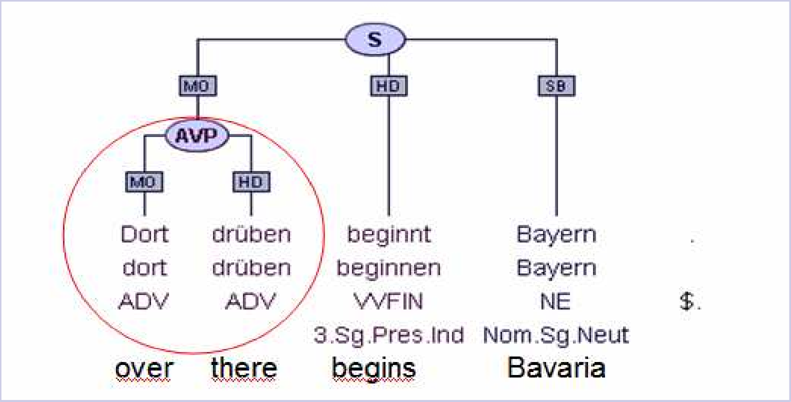
Examples from learner data (Falko)

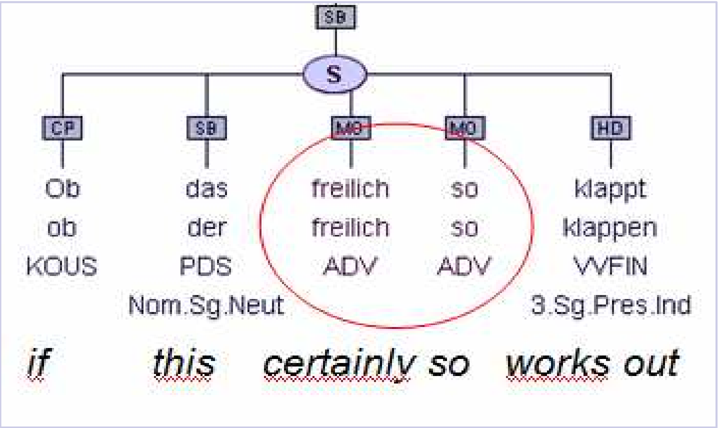
word	es	ist	doch	auch	statistisch	belegt
apos	PPER	VAFIN	ADV	ADV	ADJD	VVPP
cpos	PPER	VAFIN	ADV	ADV	ADJD	ADJD
lemma	er	sein	doch	auch	statistisch	belegen

(It is also statistically proven)

'??? also'

Types of ADV-ADV co-occurrences





Requirements

- Need for a corpus, providing
  - □ More granularity of pos annotation than the STTS tag 'ADV' offers
  - □ Phrasal annotation

ADV categories

- Syntactic classification of single ADVs
- Criteria: attachment; ±clause constituent

PT_PHR	DP/PP/AP/AdvP attached, no constituent: Phrasal particles (focus particles, intensifiers)
ADV_VP	VP attached, constituent: verbal phrase adverbs
ADV_CP	CP attached, constituent: sentence adverbs
PT_CP	CP attached, no constituent: modal particles

Annotation of ADV categories

word	und	immer	noch	kann	man	eine	unzufriedenheit	spüren
apos	KON	ADV	ADV	VMFIN	PIS	ART	unknown	VVINF
cpos	KON	ADV	ADV	VMFIN	PIS	ART	NN	VVINF
ADV_pos		PT_PHR	ADV_VP					

(And still you can feel some dissatisfaction)

Annotation of ADV categories

word	es	ist	doch	auch	statistisch	belegt
apos	PPER	VAFIN	ADV	ADV	ADJD	VVPP
cpos	PPER	VAFIN	ADV	ADV	ADJD	ADJD
ADV_pos			PT_CP	PT_CP		

(It is also statistically proven)

Phrasal annotation

word	und	immer	noch	kann	man	eine	unzufriedenheit	spüren
apos	KON	ADV	ADV	VMFIN	PIS	ART	unknown	VVINF
cpos	KON	ADV	ADV	VMFIN	PIS	ART	NN	VVINF
ADV_pos		PT_PHR	ADV_VP					
phrase		AdvP_lex			NP	NP		

(And still you can feel some dissatisfaction)

Relevance of additional layers

- Annotation of adverb types and phrase categories → Measuring ambiguity → Syntactic variability a factor for learnability?
- AdvP annotation → Acquisition of complex and lexicalized AdvPs (complex lexemes=single lexemes?)
- Falko: additional annotations in progress
- Interim results on register differences by comparing token frequencies, gathered from a German treebank (Tiger;http://www.coli.uni−sb.de/cl/projects/tiger)

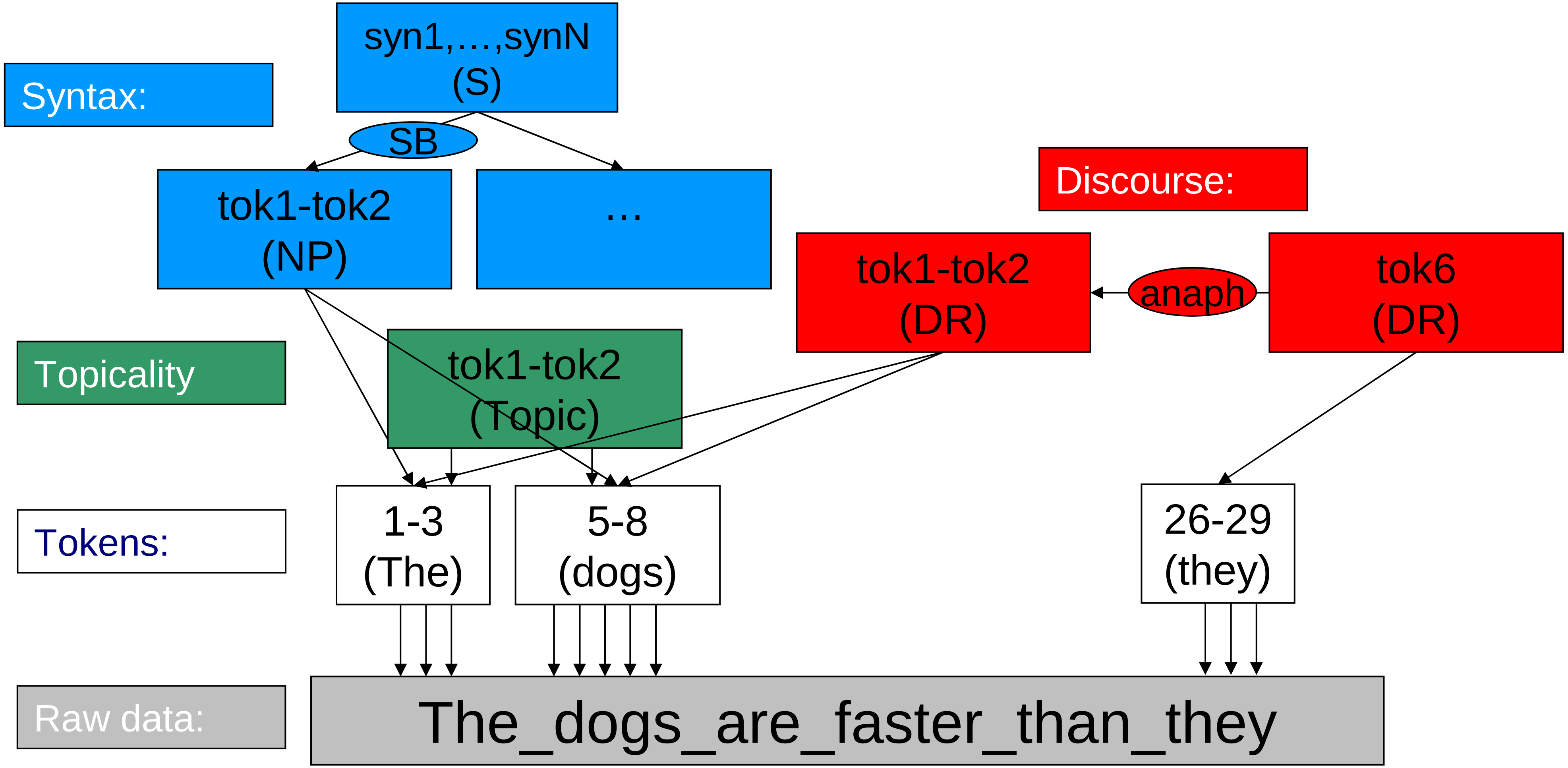
Summary & Conclusion

- EA and target hypothesis
  - Need for competing annotations
  - Need for conflicting annotations
- ADV case study
  - Need for easy addition of new (layers of) annotations (saving original annotations)

Outlook: Falko in Annis – Search in a multi-layer learner corpus

- Annis is a multilevel search architecture allowing search and visualization of:
  - □ Discontinuous surface structures
  - □ Conflicting hierarchical structures
  - □ Ambiguous annotations
- Based on the stand-off XML format PAULA (Dipper & Götze 2005, Chiarcos et al. 2009)

PAULA - Stand-Off XML (simplified)



Syntax:

syn1,…,synN (S)

SB

tok1-tok2 (NP)

…

Discourse:

tok1-tok2 (DR)

anaph

tok6 (DR)

Topicality

tok1-tok2 (Topic)

Tokens:

1-3 (The)

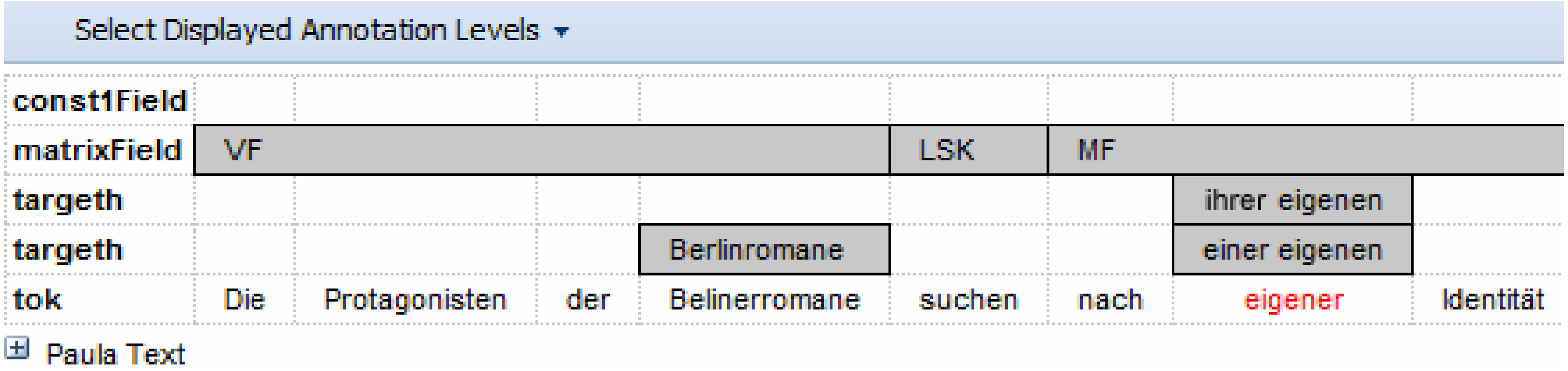
5-8 (dogs)

26-29 (they)

Raw data:

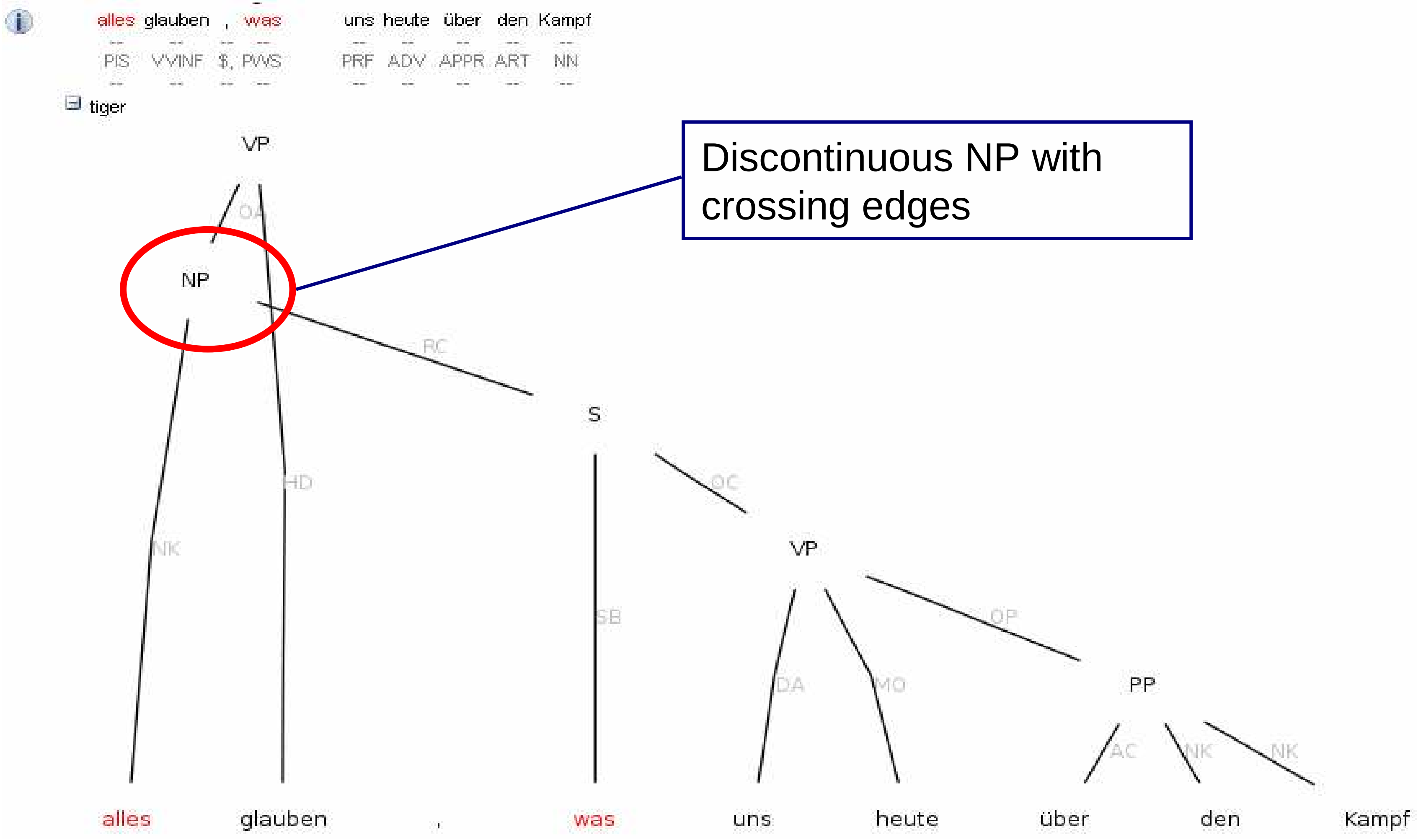
The_dogs_are_faster_than_they

Annis: Search span annotations



const1Field								
matrixField	VF				LSK	MF		
targeth							ihrer eigenen	
targeth				Berlinromane			einer eigenen	
tok	Die	Protagonisten	der	Belinerromane	suchen	nach	eigener	Identität

Annis: Search treebanks



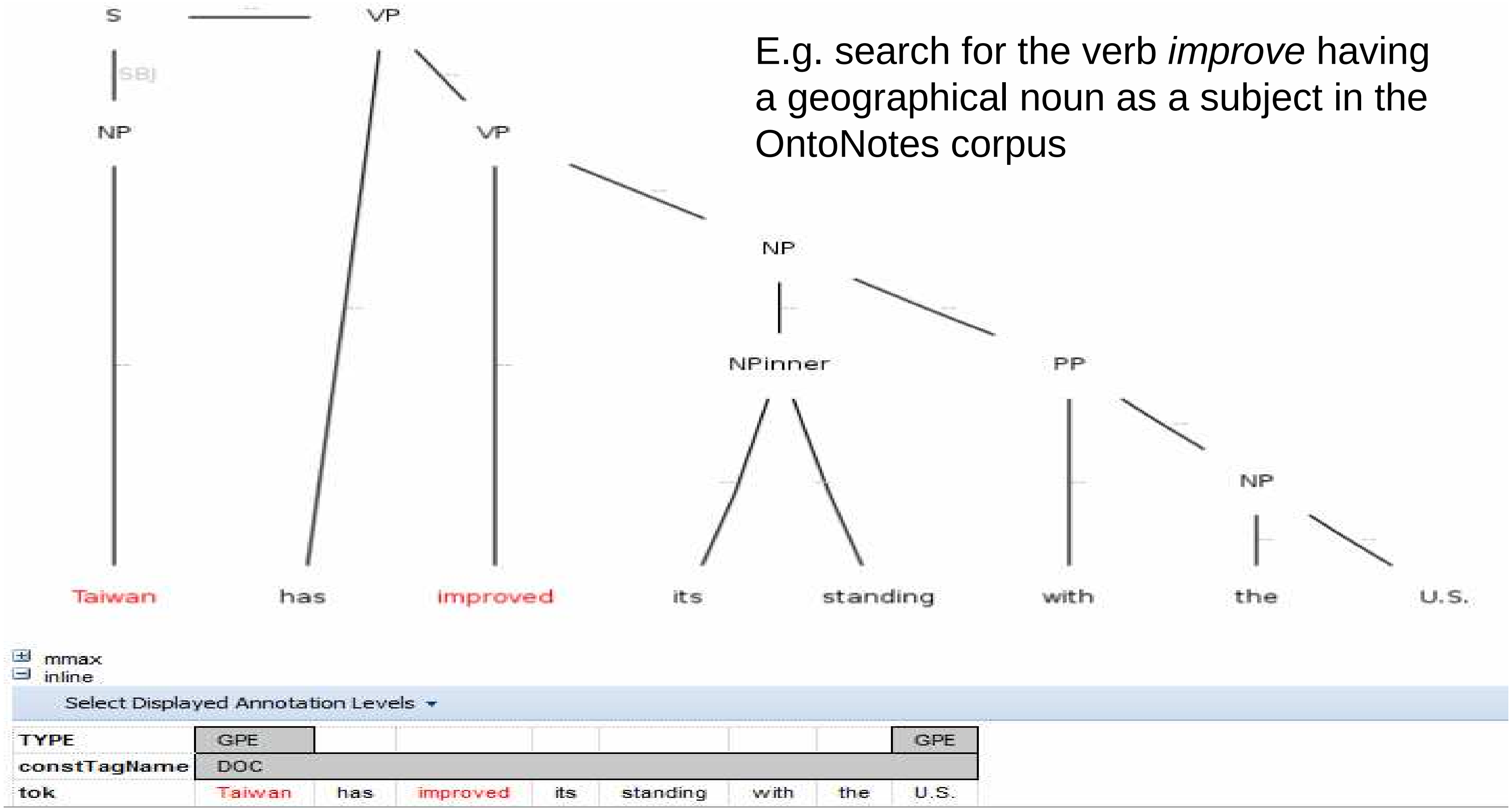
alles	glauben	,	was	uns	heute	über	den	Kampf
--	--	--	--	--	--	--	--	--
PIS	VVINF	$,	PWS	PRF	ADV	APPR	ART	NN
--	--	--	--	--	--	--	--	--

Discontinuous NP with crossing edges

everything believe which us today about the war

Annis: Combined search

E.g. search for the verb *improve* having a geographical noun as a subject in the OntoNotes corpus



TYPE	GPE							GPE
constTagName	DOC							
tok	Taiwan	has	improved	its	standing	with	the	U.S.

Thank you! Danke!

Contact:
Hagen Hirschmann
hirschhx@hu-berlin.de

Information and contact Annis:
http://www.sfb632.uni-potsdam.de/~d1/annis/
Our learner corpus Falko is freely available at
http://korpling.german.hu-berlin.de/falko/

References

- Carletta, Jean; Evert, Stefan; Heid, Ulrich; Kilgour, Jonathan; Chen, Yiya (2005) The NITE XML Toolkit: data model and query. *Language Resources and Evaluation Journal (LREJ)*, 39(4), 313-334.
- Chiarcos, Christian; Dipper, Stefanie; Götze, Michael; Leser, Ulf; Lüdeling, Anke; Ritz, Julia; Stede, Manfred (2009) A flexible framework for integrating annotations from different tools and tagsets. In: Traitement Automatique des Langues.
- Doolittle, Seanna (2008) Entwicklung und Evaluierung eines auf dem Stellungsfeldermodell basierenden syntaktischen Annotationsverfahrens für Lernerkorpora innerhalb einer Mehrebenenarchitektur mit Schwerpunkt auf schriftlichen Texten fortgeschrittener Deutschlerner. Magisterthesis, Humboldt-University.
- Lüdeling, Anke; Doolittle, Seanna; Hirschmann, Hagen; Schmidt, Karin; Walter, Maik (2008) Das Lernerkorpus Falko. In: *Deutsch als Fremdsprache 2(2008)*, 67-73.
- Lüdeling, Anke (2008) Mehrdeutigkeiten und Kategorisierung. Probleme bei der Annotation von Lernerkorpora. In: Maik Walter & Patrick Grommes (Hrsg.) *Fortgeschrittene Lernervarietäten*. Tübingen: Niemeyer, 119-140.
- Schmid, Helmut (1994) Probabilistic Part-of-Speech Tagging Using Decision Trees. In: *Proceedings of the International Conference on New Methods in Language Processing.* Manchester, 44-49. [extended version available at http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf].
- Wittenburg Peter (2008) Preprocessing multimodal corpora. In: A. Lüdeling & M. Kytö (eds) *Corpus Linguistics. An International Handbook*, Mouton de Gruyter, Berlin.