

Annotating and Analysing IS in Historical Corpus Texts

13.11.2009

Universität Potsdam

ANNIS Search Form

AnnisQL: "Auto" & /*/ & /*/ & #1,* #2 & #2,* #3 | cat="S" & "das" & "Dorf" & #4 >* #5 & #5, #6

Match Count: 3

More Corpora

Name	Texts	Token
<input type="checkbox"/> b4.muspill	1	909
<input type="checkbox"/> b4.tatian	268	1502
<input type="checkbox"/> c6.hindi	1	2218
<input checked="" type="checkbox"/> d2.2samplesDEU	2	19
<input type="checkbox"/> falco.essay	133	66000
<input checked="" type="checkbox"/> pcc-11		
<input type="checkbox"/> spec		

Simple Search

Context Left: 0


Context Right: 0

Show Result

Search Result: "Auto" & /*/ & /*/ & #1,* #2 & #2,* #3 | cat="S" & "das" & "Dorf" & #4 >* #5 & #5, #6 (0, 0)

Page 1 of 1

Token Annotations * Show Citation URL



Debel hätte das Dorf jede Einnahme nötig
dabei haben der Dorf jeder Einnahme nötig
wäre v/A v/A v/A v/A v/A v/A v/A v/A v/A v/A
aida

Select Displayed Annotation Levels *

Multi-Layer Resources in ANNIS

Historical Corpora and Information Structural Annotation


Amir Zeldes, Humboldt-Universität zu Berlin
SFB 632 Information Structure, Project D1 Linguistic Database
amir.zeldes@rz.hu-berlin.de

mt-StatSeg giv-active

NPseg NP NP

TopicSeg ab

SentSeg s



Background in SFB 632

- **ANNIS: ANNotation of Information Structure**
(Dipper et al. 2004; Chiarcos et al. 2008; Zeldes et al. 2009)
<http://www.sfb632.uni-potsdam.de/d1/annis/>
- Open-source web-based corpus search tool
- Developed in project D1 of SFB 632 on *Information Structure*
- Requirements:
 - Archive data from different projects in a unified format
 - Deal with heterogeneous annotations for IS & beyond
 - Search and visualization of complex data structures

Different Research Questions

- What is the role of Information Structure for the development of German word order?
(Petrova 2006, Donhauser et al. 2006, Petrova & Solf, to appear)
- Which topic markings are possible in the Chadic languages and under what circumstances?
(Hartmann 2006, Zimmermann, to appear)
- How do universal IS categories influence prosody, morphology and syntax?
(Féry 2006, Fanselow 2007, Dipper et al. 2007)
- Can IS be inferred from morphosyntax, lexis and discourse structure?

Different Data Structures

- Token and span annotations
- syntax trees / DAGs (with / without non-terminal units)
- Labeled pointing relations
- Multimodal data, Metadata, Alignment...

Different Tools/Formats

- Automatic tools with proprietary formats (taggers, parsers, machine learning)
- Diverse dedicated manual tools (transcription, annotation, treebanking, discourse analysis)
- Asynchronously produced annotations from different sources and theories referring to the same data

Annotation Formats

- **Token Annotations**
(e.g. TreeTagger)

- POS Tagging
- Lemmatization
- Morphology
- Phonology/Phonetics

- **Tree/Graph**

- Bracketing formats
(e.g. Penn, Bies et al. 1995)
- Generic Inline XML
- Tiger XML / Negra (Synpathy, Tiger, annotate, Brants & Plaehn, 2000)
- RST tool
(rhetorical sentence trees, O'Donnel 2000)

- **Grids**

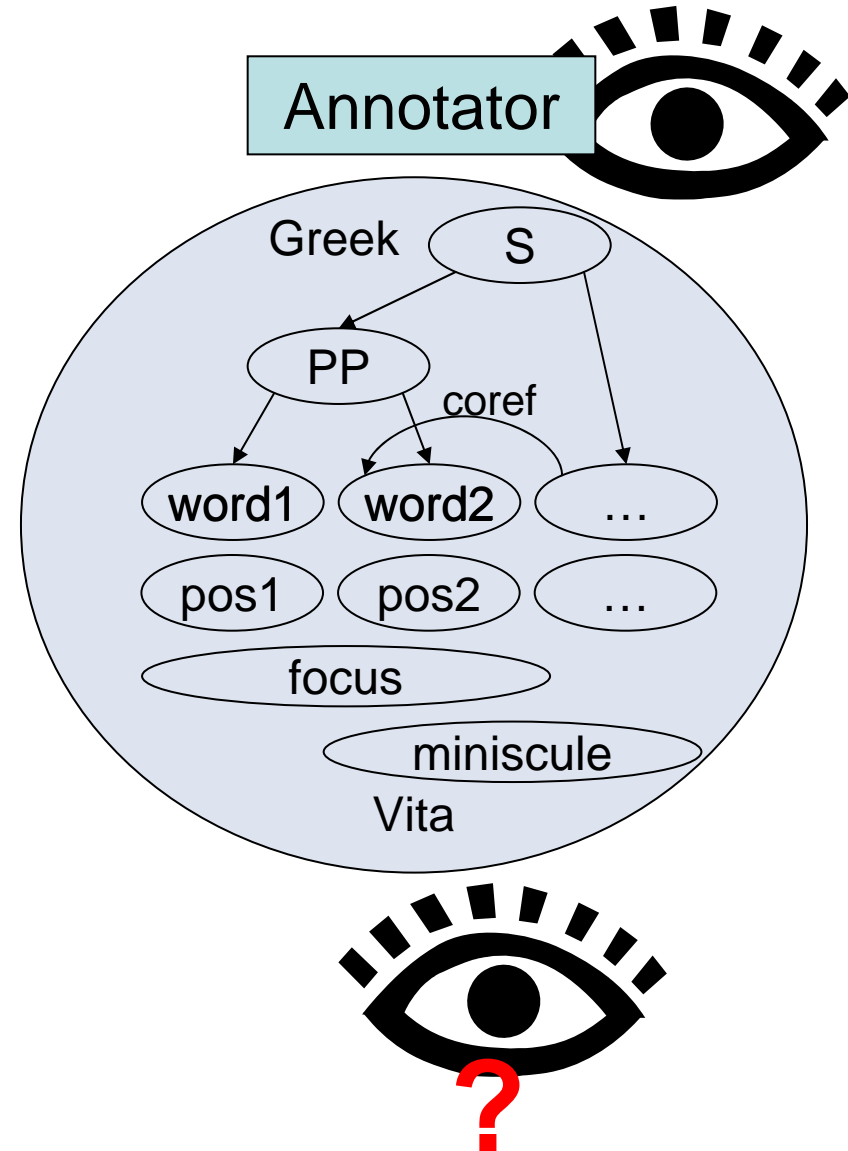
- EXMARaLDA
(Schmidt 2004)
- ELAN
(Wittenburg et al. 2006)
- Toolbox
(Stuart et al. 2007)

- **Pointing Relations**

- MMAX2
(Müller & Strube 2006)
- PALinkA
(Orasan 2003)
- Serengeti (TODO)

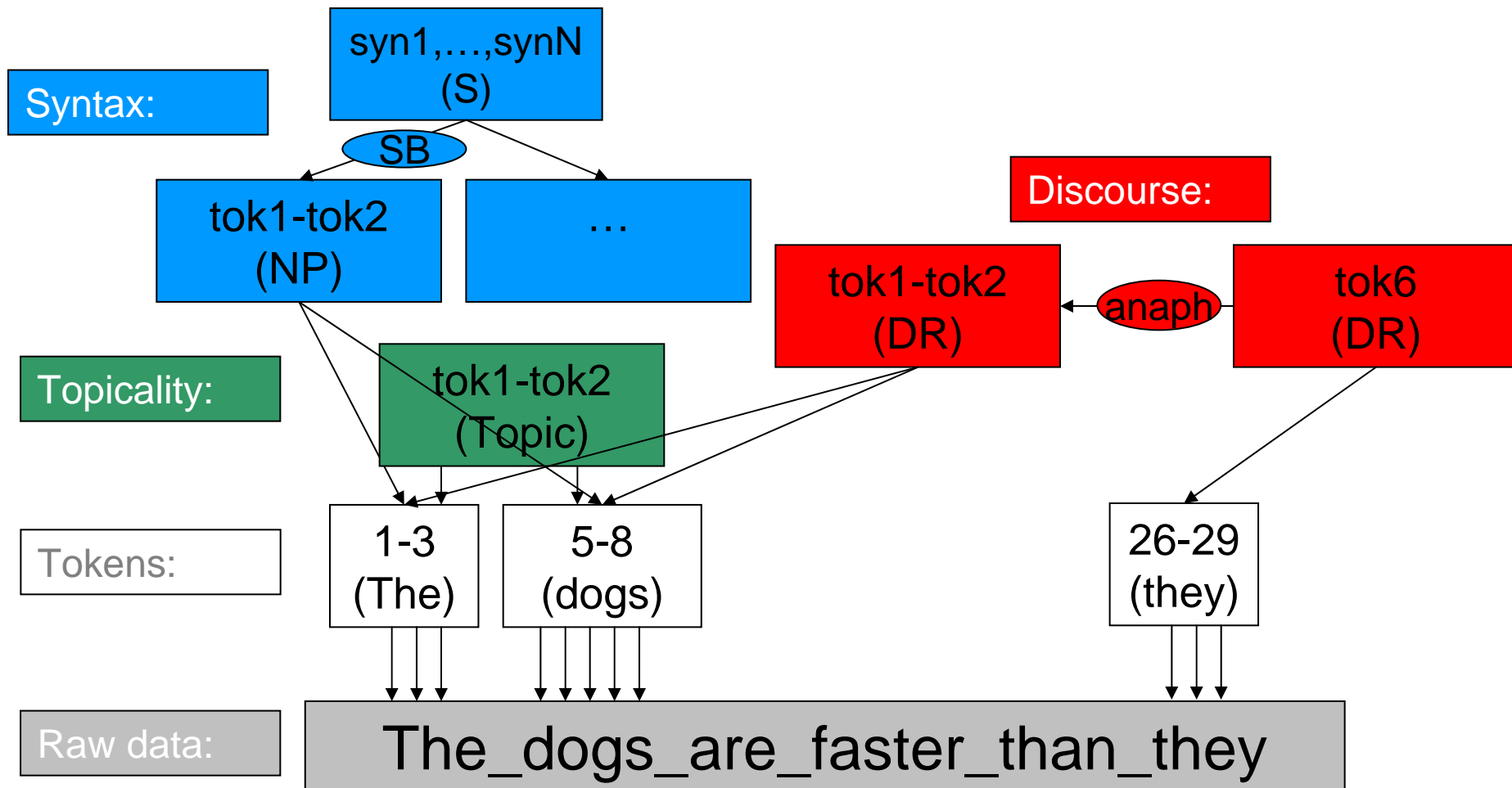
Modeling Multilevel Data

- Reduction to typed and labeled nodes and edges
- Independent layers (stand-off)
 - Retroactively insert / update / replace layers
 - Alternative interpretations (multiple pos-tags, parses, even tokenizations)



PAULA stand-off XML (simplified)

(Dipper, 2005, Dipper & Götze, 2005)



Human Readable AQL

- SQL queries too complex for users
- Simple query language based on nodes and relations: (cf. *NiteQL* Carletta et al. 2003, *TIGERSearch* Lezius 2002)

```
cat="S" & node & cat="S" & pos="PRP" &  
#1 >[func="sb"] #2 &  
#3 >[func="sb"] #4 &  
#4 ->[coref="anaphor-antecedent"] #2 &  
meta::language="en"
```

Query Builder

Search Form

AnnisQL: `pos="VVFIN" & node & cat="S" & node & #3 > #1 & #3 > [tiger:func="SB"] #2 & #1, #2 & #4, #1 & #3`

Match Count: Valid Query

More Corpora

Name	Texts	Token
<input type="checkbox"/> ONTONOTES_v1.6_4	100	53875
<input type="checkbox"/> ONTONOTES_v1.6_small	4	6450
<input type="checkbox"/> falko_docDay	1	252
<input type="checkbox"/> pcc-11	11	1939
<input checked="" type="checkbox"/> pcc176	176	33222
<input type="checkbox"/> pcc3	3	573
<input type="checkbox"/> pcc3_mmax2exmaralda	3	573

Simple Search **Query Builder** Statistics

Show Result

Create Node

```
graph TD; Root["Edge Add Clear X  
Field op Value  
cat = S"]; Node1["X  
> [tiger:func = \"OA\"]"]; Node2["X  
>"]; Node3["X  
> [tiger:func = \"SB\"]"]; Leaf1["Edge Add Clear X  
Field op Value  
pos = VVFIN"]; Leaf2["Edge Add Clear X  
Field op Value  
pos = VVFIN"]; Leaf3["Edge Add Clear X  
Field op Value  
pos = VVFIN"]; Root --> Node1; Root --> Node2; Root --> Node3; Node1 --> Leaf1; Node2 --> Leaf2; Node3 --> Leaf3;
```

List of AQL Operators

	Name	Illustration	Options
.	direct precedence	A B	
.*	indirect precedence	A x y z B	.n,m
>	direct dominance	A B	>secedge >[func="OA"]
>*	indirect dominance	A ... B	>n,m
->LABEL	Labeled pointing relation	LABEL ↙ ↘ B A	
->LABEL*	Labeled pointing path	LABEL LABEL ↙ ↘ ↙ ↘ B x y z A	
=	identical coverage	AAA BBB	
o	overlap	AAA BBB	_ol_ _or_
i	inclusion	AAA B	

	Name	Illustration
l	left aligned	AAA BB
r	right aligned	AA BBB
>@l	left-most child	A / \ B x y
>@r	right-most child	A / \ x y B
\$	Common parent node	x / \ A B
\$*	Common ancestor node	x ... / \ A B
#x:arity=n	Arity	x / \ 1 ... n
#x:length=n	Length	x ... / \ 1 ... n

B4: T-Codex (, Muspilli, Heliand...)

(Petrova et al. 2009)

- IS (SFB guidelines)
- Morpho-syntactic annotation
- “Interlinear” Latin

i **bithiu** sie uuarun simones ginoza
+ exmaralda
+ Paula
+ Paula Text

i **nibi** hér neme inti íz úf
+ exmaralda
+ Paula
+ Paula Text

i nibi hér neme **inti** íz úf héue
+ exmaralda

Select Displayed Annotation Levels ▾

LAT	Nonne		tenebit	&	leuabit		eam
aboutness		ref			ref		
align					=L6		=L4
bibl	T 106, 29 Beta Lc 6						
cat		NP	VP		NP	ADVP	VP
clause-status	MAININT						
comment	eingefügtes Subjektspronomen;						
definiteness		DEF			DEF		
gf		SUBJ	VFIN		DO	adv.dir	VFIN
givenness		GIV			GIV		
pos	CONJ	PRONPRS	V	CONJ	PRONPRS	ADVDIR	V
position		INIT			MAININT		
syl_no	2	1	2	2	1	3	
tok	nibi	hér	neme	inti	íz	úf	héue

+ Paula

exmaralda;pos = CONJ

MHG Speculum Ecclesiae

(WIP, Hagen Hirschmann/Sonja Linde et al.)

- Syntactically annotated corpus
- Normalization over multiple tokens
- Spans for edition references
- Similarly Monsee Fragments

Result - cat=/S/ & Mhd=/[Ww]irne/ & #1 .1,4 #2 (5, 2)

Page 1 of 1 Token Annotations Show Citation URL Displaying Res

3.Nom.Pl.Masc 3.Pl.Pres.Ind Nom.Pl.Masc Pos.Nom.Pl.Masc -- Nom.Sg.Masc 3.Sg.Pres.Ind Pos.*.* -- Nom.Sg.Masc -- Dat.Sg.Fem Dat.Sg.Fem 3.Acc.Sg.Masc -- 3.S

tiger
exmaralda
Paula
Paula Text

sich darzou hat gerehet , daz er des tages so er von dirre wert scheidet ,
PRF PROAV VAFIN VVPP \$, KOUS PPER ART NN \$, KOUS PPER APPR PDAT NN VVFIN \$,
3.Acc.Sg.Masc -- 3.Sg.Pres.Ind Psp -- -- 3.Nom.Sg.Masc Gen.Sg.Masc Gen.Sg.Masc -- -- 3.Nom.Sg.Masc -- Dat.Sg.Fem Dat.Sg.Fem 3.Sg.Pres.Ind --

tiger
exmaralda
Paula
Paula Text

goute zovuersiht habe , Wir ne scoltten nimmer gerouwen , naht noch tac , wir ne waren
\$, ADJA NN VAFIN \$, PPER PTKNEG VMFIN ADV VVINF \$, NN KON NN \$, PPER PTKNEG VAFIN
-- Pos.Acc.Sg.Fem.St Acc.Sg.Fem 3.Sg.Pres.Subj -- 1.Nom.Pl.* -- 1.Pl.Past.Ind -- Inf -- Acc.Sg.Fem -- Acc.Sg.Masc -- 1.Nom.Pl.* -- 3.Pl.Past.Ind

tiger

goute zovuersiht habe Wir ne scoltten nimmer gerouwen naht noch tac wir ne waren danach

exmaralda

Select Displayed Annotation Levels

Bibl	Mellbourn, 69/154,23		Mellbourn, 69/154,24																
Lemna	gout	zovuersiht	haben	er	ne	scoltten	nimmer	gerouwen		naht	noch	tac		er	ne	sih	danach		
Mhd	,	goute	zovuersiht	habe	,	Wirne	scoltten	nimmer	gerouwen	,	naht	noch	tac	,	wirne	waren	danach		
Satz	s			s															
tok	,	goute	zovuersiht	habe	,	Wir	ne	scoltten	nimmer	gerouwen	,	naht	noch	tac	,	wir	ne	waren	danach

Paula
Paula Text

PCC: RST, IS (Stede 2004)

Search Result - cat (5, 5) Displaying Results 1 - 10 of 80

Page 1 of 81 Token Annotations Show Citation URL

NK
NK
MNR
HD

PP
NP

AC
NK
NK

Die
Jugendlichen
in
Zossen
wollen
ein
Musikcafé

- exmaralda
- Paula
- Paula Text

i Zossen wollen ein Musikcafé . Das forderten sie bei der ersten
 Zossen wollen ein Musikcafé . der fordern sie bei der erster
 NE VMFIN ART NN DOLLAR_PERIOD PDS VVFIN PPER APPR ART ADJA
 Dat.Sg.Neut 3.PI.Pres.Ind Acc.Sg.Neut Acc.Sg.Neut -- Acc.Sg.Neut 3.PI.Past.Ind 3.Nom.Pl.* -- Dat.Sg.Fem Pos.Dat.Sg.Fem

- urml
- tiger
- exmaralda

Select Displayed Annotation Levels ▾

Focus_newInf		nf-unsol										
Inf-Stat	new		new		giv-active		giv-active		new			
NP	NP		NP		NP		NP		NP			
PP	PP								PP			
Sent	s			s								
Topic	ab						ab		fs			
tok	Zossen	wollen	ein	Musikcafé	.	Das	forderten	sie	bei	der	ersten	Zossen

- Paula

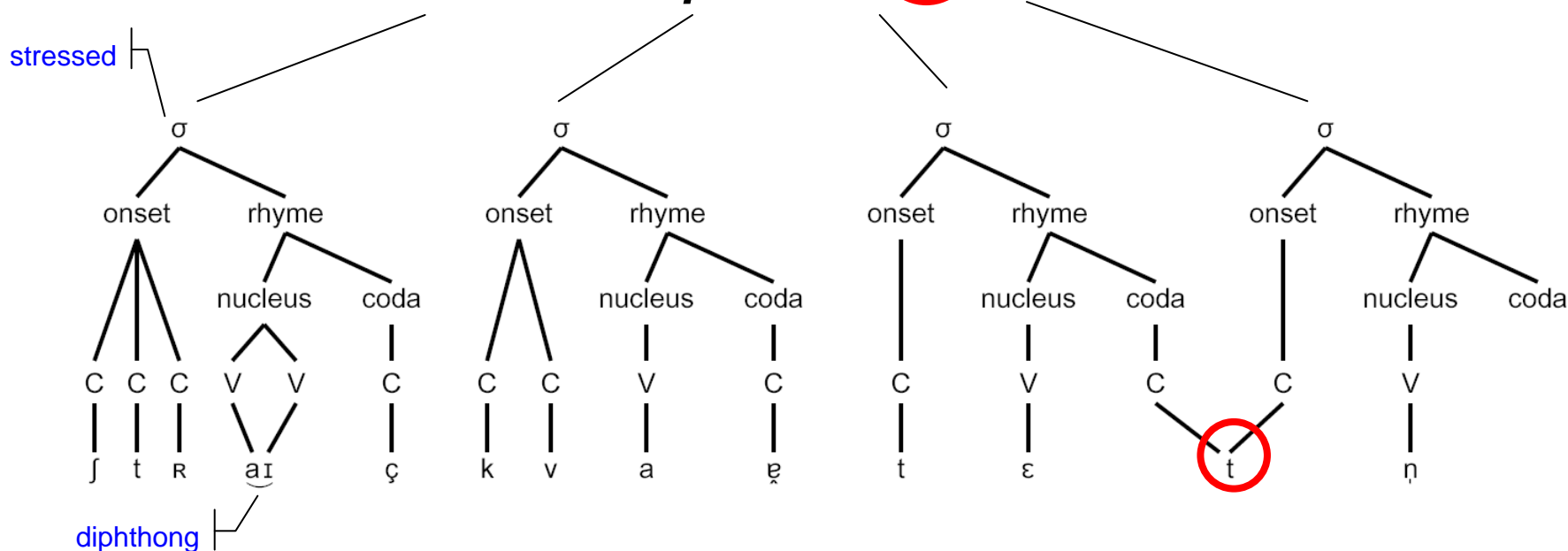
Future Directions / Open Issues

- Subtokenization (cf. MAF, Clément & de la Clergerie 2005)
 - Structural challenge to token based concept:
 - "Physical" tokens (atoms) = smallest units
 - "Reference" tokens (e.g. word forms) = unit of reference for search operators (*“within 5 tokens”*) and context visualization (*“10 tokens left and right, sort by second token to the right”*)
 - Users unaware if desired results are subtokenized!

Future Directions / Open Issues

- Subtoken DAGs, features \rightarrow operators?
- Ambiguous subtoken-token alignment

Streich·quar·tet·ten (“string quartets”)



Parallel Corpora

- Representation of aligned document sets
- Edges to align multiple levels paragraph/
sentence/word... (TEI, Romary & Bonhomme 2000)
- Handling of mismatching / circular /
incomplete alignment
- Search and visualization

Summary

- ANNIS is an open-source search and visualization architecture for multi-layer corpora
- Merge annotations from multiple tools
- Annotate IS, syntax, morphology, discourse structure...
- Modular visualizations can be written for further types of data – we're always looking for partners!

Thanks!

Christian Chiarcos, Thomas Krause, Anke
Lüdeling, Julia Richling, Julia Ritz, Viktor
Rosenfeld, Manfred Stede, Amir Zeldes and
Florian Zipser

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

References

- Bies, A./Ferguson, M./Katz, K./MacIntyre, R. (1995) *Bracketing Guidelines for Treebank II Style Penn Treebank Project*. Technical report, University of Pennsylvania.
- Brants T./Plaehn, O. (2000) Interactive Corpus Annotation. In: *Proc. LREC 2000*, Athens.
- Carletta, J./Evert, S./Heid, U./Kilgour, J./Robertson, J./Voormann, H. (2003) The NITE XML Toolkit: Flexible Annotation for Multi-modal Language Data. *Behavior Research Methods, Instruments, and Computers* 35(3), 353-363.
- Chiarcos, C./Dipper, S./Götze, M./Leser, U./Lüdeling, A./Ritz, J./Stede, M. (2008) A Flexible Framework for Integrating Annotations from Different Tools and Tag Sets. *TAL* 49(2), 271-293.
- Clément, L./de la Clergerie, É. (2005) MAF: A Morphosyntactic Annotation Framework. In: *Proc. of the Language and Technology Conference, Poznan, Poland*, 90-94.
- Dipper, S. (2005) XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation. In: *Proceedings of Berliner XML Tage 2005 (BXML 2005)*, Berlin, Germany, 39-50.
- Dipper, S. & Götze, M. (2005) Accessing Heterogeneous Linguistic Data – Generic XML-based Representation and Flexible Visualization. In: *Proc. 2nd Language & Technology Conference*. Poznan, Poland, 206-210.
- Dipper, S./Götze, M./Skopeteas, S. (eds.) (2007) Information Structure in Cross-Linguistic Corpora: Annotation Guidelines for Phonology, Morphology, Syntax, Semantics, and Information Structure. In: *ISIS* 9. Potsdam: Universitätsverlag Potsdam.
- Donhauser, K./Solf, M./Zeige, L. (2006) Informationsstruktur und Diskursrelationen im Vergleich Althochdeutsch – Altisländisch. In: Hornscheidt, A. et al. (eds.), *Grenzgänger. Festschrift zum 65. Geburtstag von Jurij Kusmenko*. Berlin: Nordeuropa-Institut, 73-90.
- Fanselow, G. (2007) The Restricted Access of Information Structure to Syntax - A Minority Report. *ISIS* 6.
- Féry, C. (2006) The Fallacy of Invariant Phonological Correlates of Information Structural Notions. *ISIS* 6.
- Grust, T./Keulen, M. V./Teubner, J. (2004) Accelerating XPath Evaluation in any RDBMS. *ACM Trans. Database Syst.* 29 (1), 91-131.
- Hartmann, K. (2006). Focus Constructions in Hausa. In: Molnár, V./Winkler, S. (eds.), *The Architecture of Focus. Studies in Generative Grammar*. Berlin: Mouton de Gruyter, 579-607.
- Lezius, W. (2002) *Ein Suchwerkzeug für syntaktisch annotierte Textkorpora*. Ph.D. thesis. (Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS) 8(4).) Stuttgart: IMS, University of Stuttgart.

References II

- Müller, C./Strube, M. (2006), Multi-Level Annotation of Linguistic Data with MMAX2. In: Braun, S./Kohn, K./Mukherjee, J. (eds.), *Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods*. Frankfurt: Peter Lang, 197–214.
- O'Donnell, M. (2000) RSTTool 2.4 – A Markup Tool for Rhetorical Structure Theory. In: *Proc. of the International Natural Language Generation Conference (INLG'2000), 13-16 June 2000*, Mitzpe Ramon, Israel, 253–256.
- Orasan, C. (2003), Palinka: A Highly Customisable Tool for Discourse Annotation. In: *Proc. of the 4th SIGdial Workshop on Discourse and Dialogue*, Sapporo.
- Petrova, S. (2006) A discourse-based approach to verb placement in early West-Germanic. *ISIS* 5, 153-182.
- Petrova, S./Solf, M. (to appear) Syntaktischer Wandel und Satzmodus. Beobachtungen zur Wortstellung in direkten Fragesätzen des Althochdeutschen. In: *Beiträge zur Geschichte der deutschen Sprache und Literatur*.
- Schmidt, T. (2004) Transcribing and Annotating Spoken Language with Exmaralda. In: *Proc. of the LREC-workshop on XML Based Richly Annotated Corpora, Lisbon 2004*. Paris: ELRA.
- Stede, M. (2004) The Potsdam Commentary Corpus. In: *Proc. ACL Workshop on Discourse Annotation*. Barcelona, 96–102.
- Stuart, R./Aumann, G./Bird, S. (2007) Managing Fieldwork Data with Toolbox and the Natural Language Toolkit. *Language Documentation & Conservation* 1(1), 44–57.
- Weischedel, R./Pradhan, S./Ramshaw, L./Palmer, M./Xue, N./Marcus, M./Taylor, A./Greenberg, C./Hovy, E./Belvin, R./Houston, A. (2008) *OntoNotes Release 2.0*. Philadelphia: LDC.
- Wittenburg, P./Brugman, H./Russel, A./Klassmann, A./Sloetjes, H. (2006) ELAN: a Professional Framework for Multimodality Research. In: *Proceedings of LREC 2006*.
- Zeldes, A./Ritz, J./Lüdeling, A./Chiarcos, C. (2009) ANNIS: A Search Tool for Multi-Layer Annotated Corpora. In: *Proceedings of Corpus Linguistics 2009, July 20-23*, Liverpool, UK.
- Zimmermann, M. (to appear) Focus in Western Chadic: A Unified OT Account. In: Davis, C./Deal, A.-R./Zabbal, Y. (eds.), *Proceedings of NELS 36*. Amsterdam: Benjamins.

RelAnnis (simplified)

(WIP, Viktor Rosenfeld & Florian Zipser, cf. Grust et al. 2004)

