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

Background in SFB 632

- ANNIS: **ANN**otation of **I**nformation **S**tructure
  (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

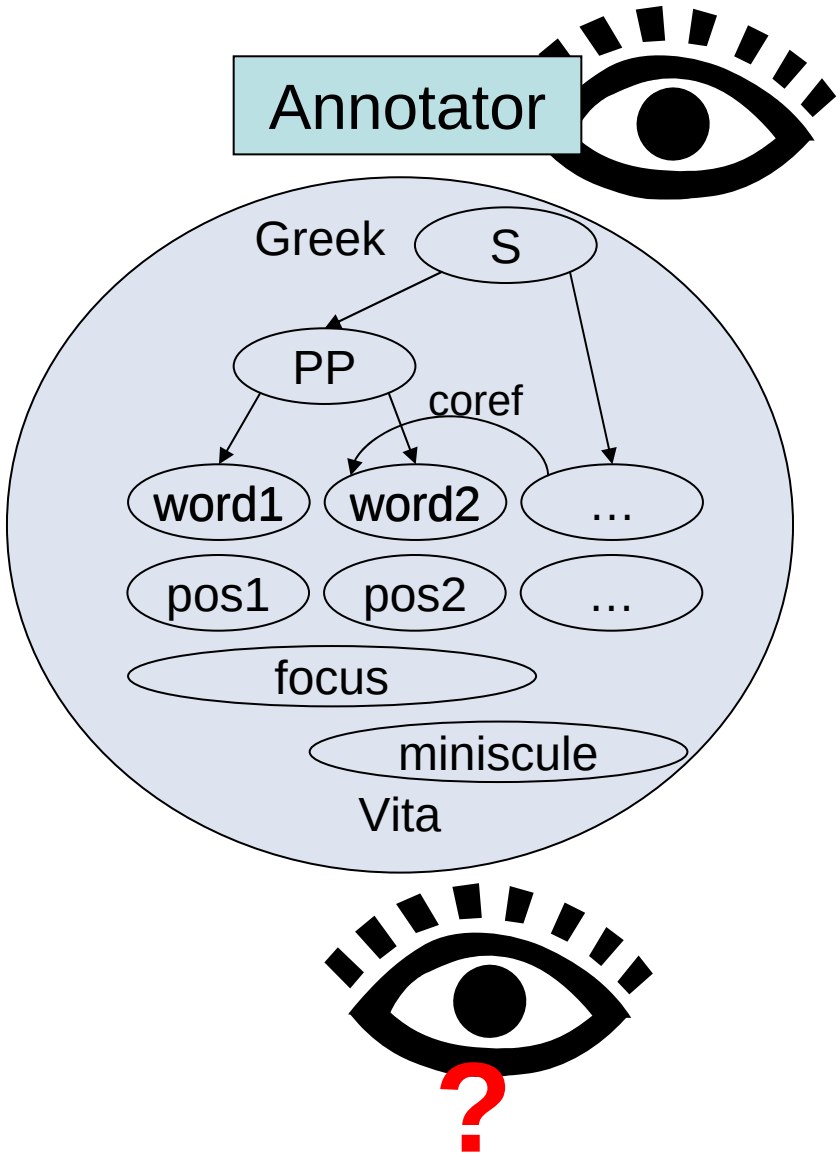
- Tree/Graph
  - Bracketing formats (e.g. Penn, Bies et al. 1995)
  - Generic Inline XML
  - Tiger XML / Negra (Synpathy, Tiger, annotate, Brants & Plaehn, 2000)
  - RST tool (rhetorical sentence trees, O'Donnel 2000)

- 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)

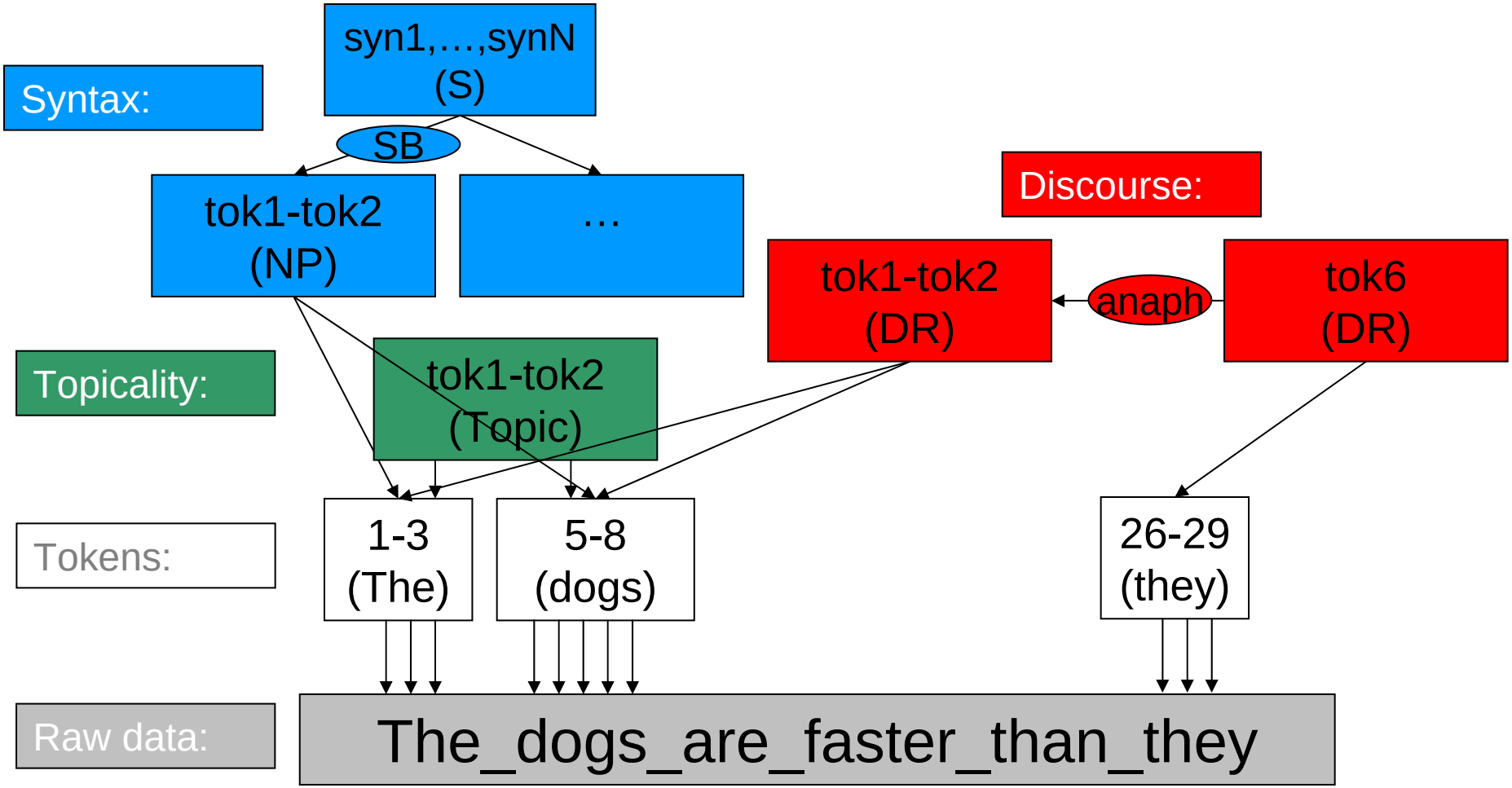
Modeling Multilevel Data

- 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)

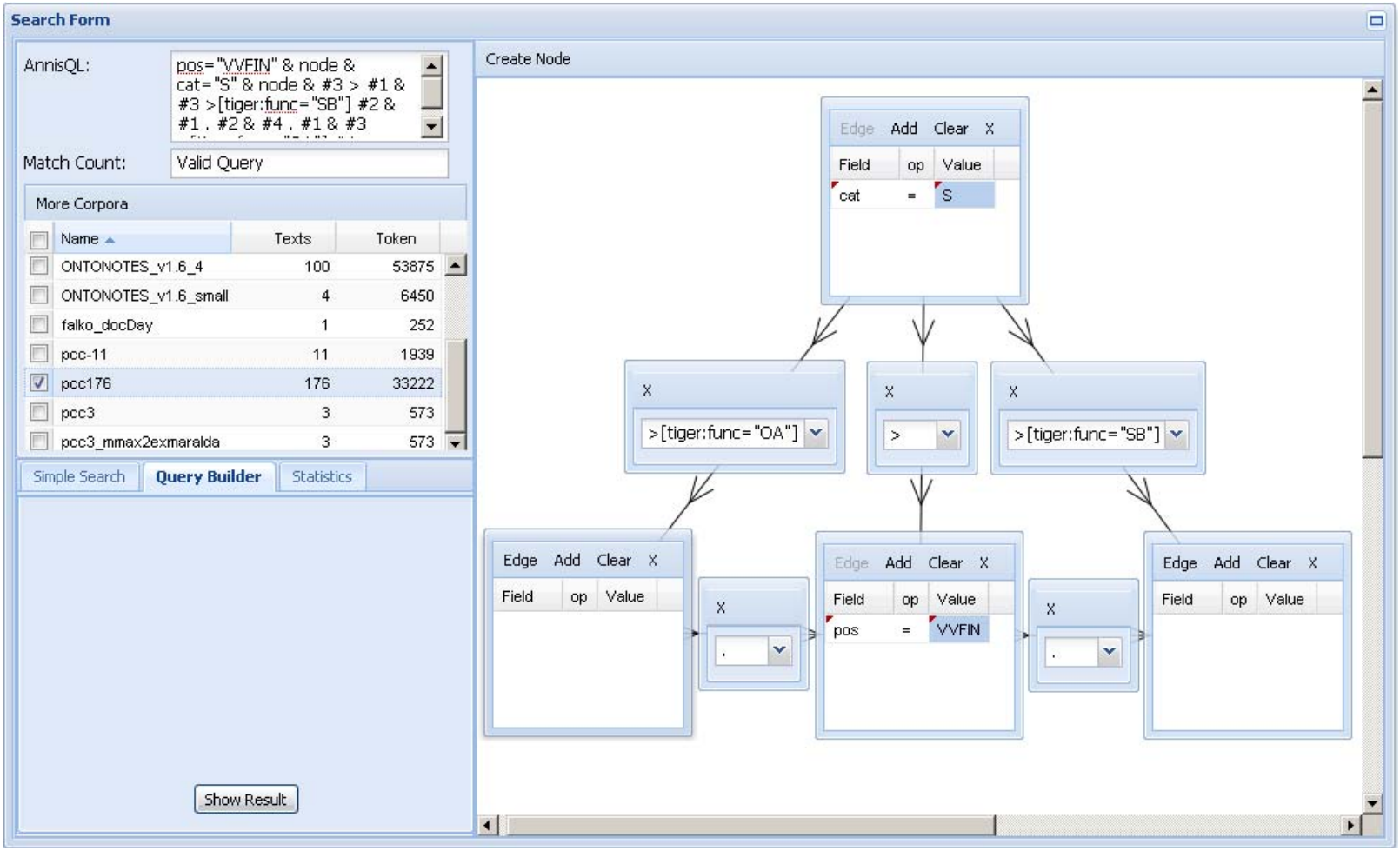


Human Readable AQL

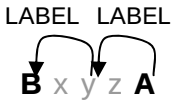
- SQL queries too complex for users
- Simple query language based on nodes and relations: (cf. *NiteQL* Carletta et al. 2003, *TIGERSearch* Lezius 2002)

```
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

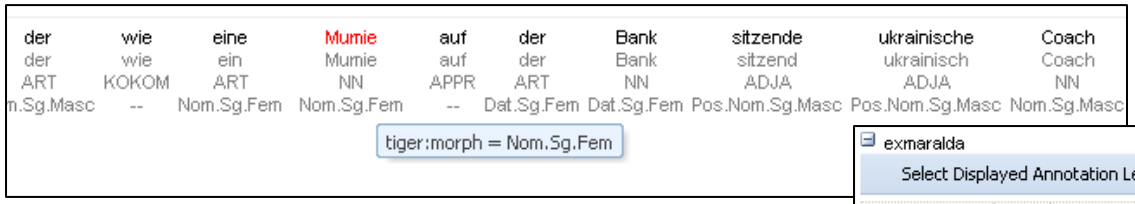


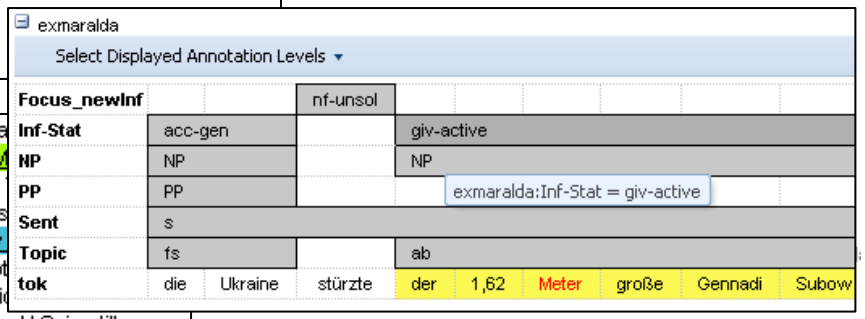
List of AQL Operators

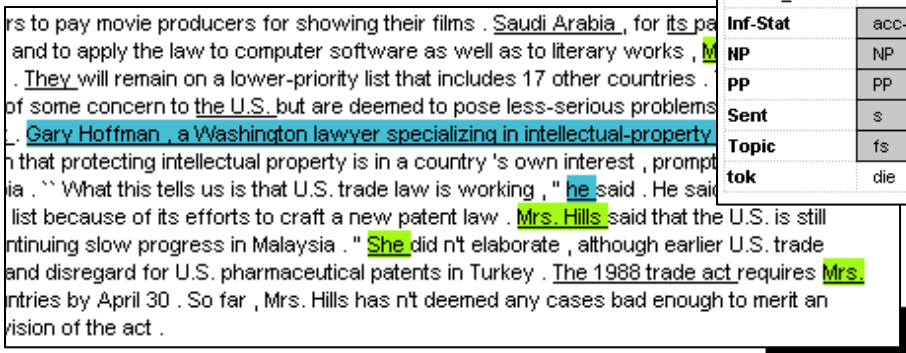
	Name	Illustration	Options
.	direct precedence	**A B**	
.*	indirect precedence	**A** x y z **B**	.n,m
>	direct dominance	**A** | **B**	>secedge >[func="OA"]
>*	indirect dominance	**A** | ... | **B**	>n,m
->LABEL	Labeled pointing relation		
->LABEL*	Labeled pointing path		
_=_	identical coverage	**AAA**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

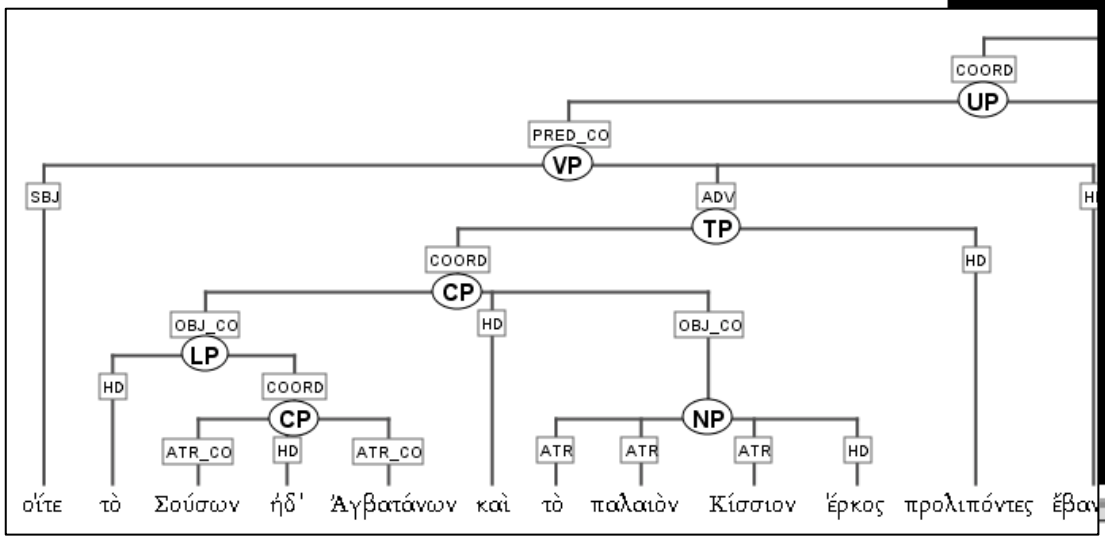
	Name	Illustration
_l_	left aligned	**AAA** **BB**
_r_	right aligned	**AA** **BBB**
>@l	left-most child	**A** / | \ **B** x y
>@r	right-most child	**A** / | \ x y **B**
$	Common parent node	x / \ **A B**
$*	Common ancestor node	x ... / \ **A B**
#x:arity=n	Arity	x / | \ **1 ... n**
#x:length=n	Length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

Visualization





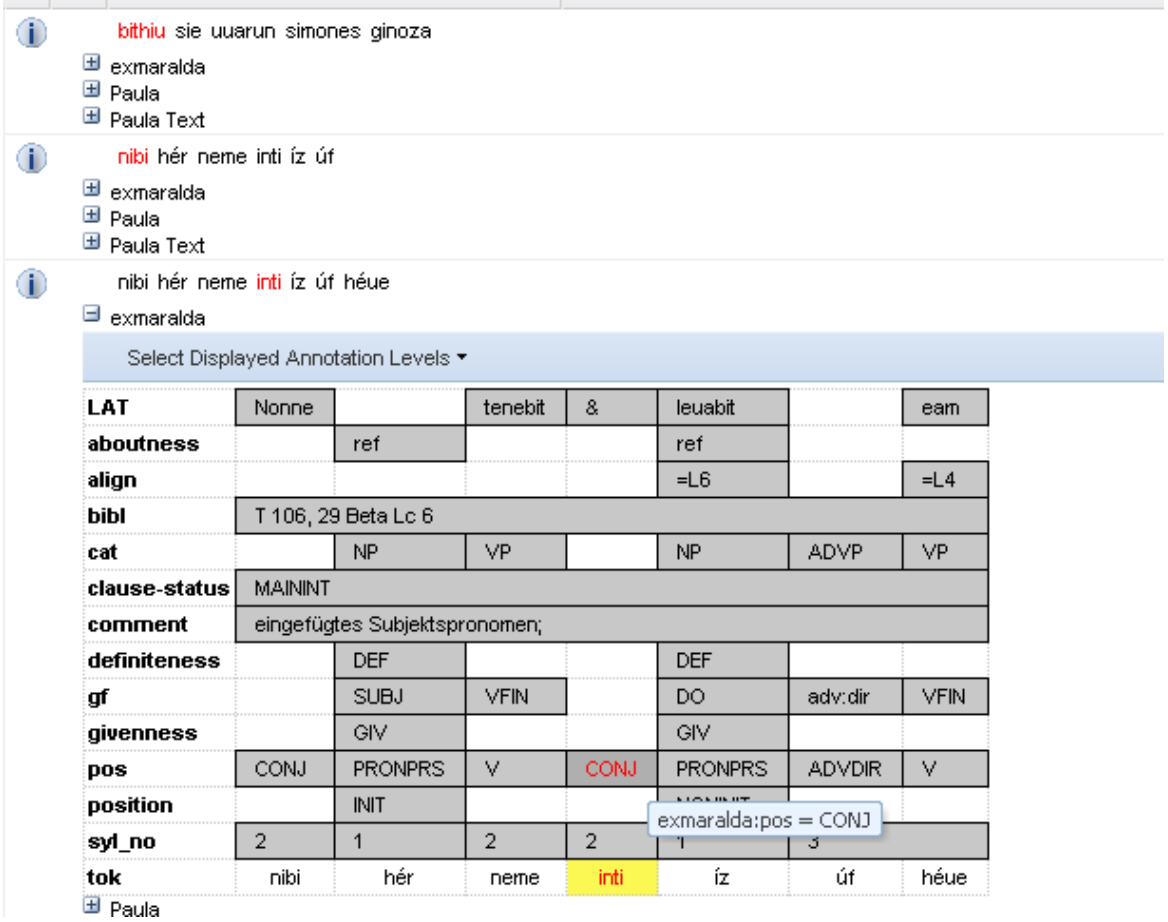




B4: T-Codex (, Muspilli, Heliand…)

(Petrova et al. 2009)

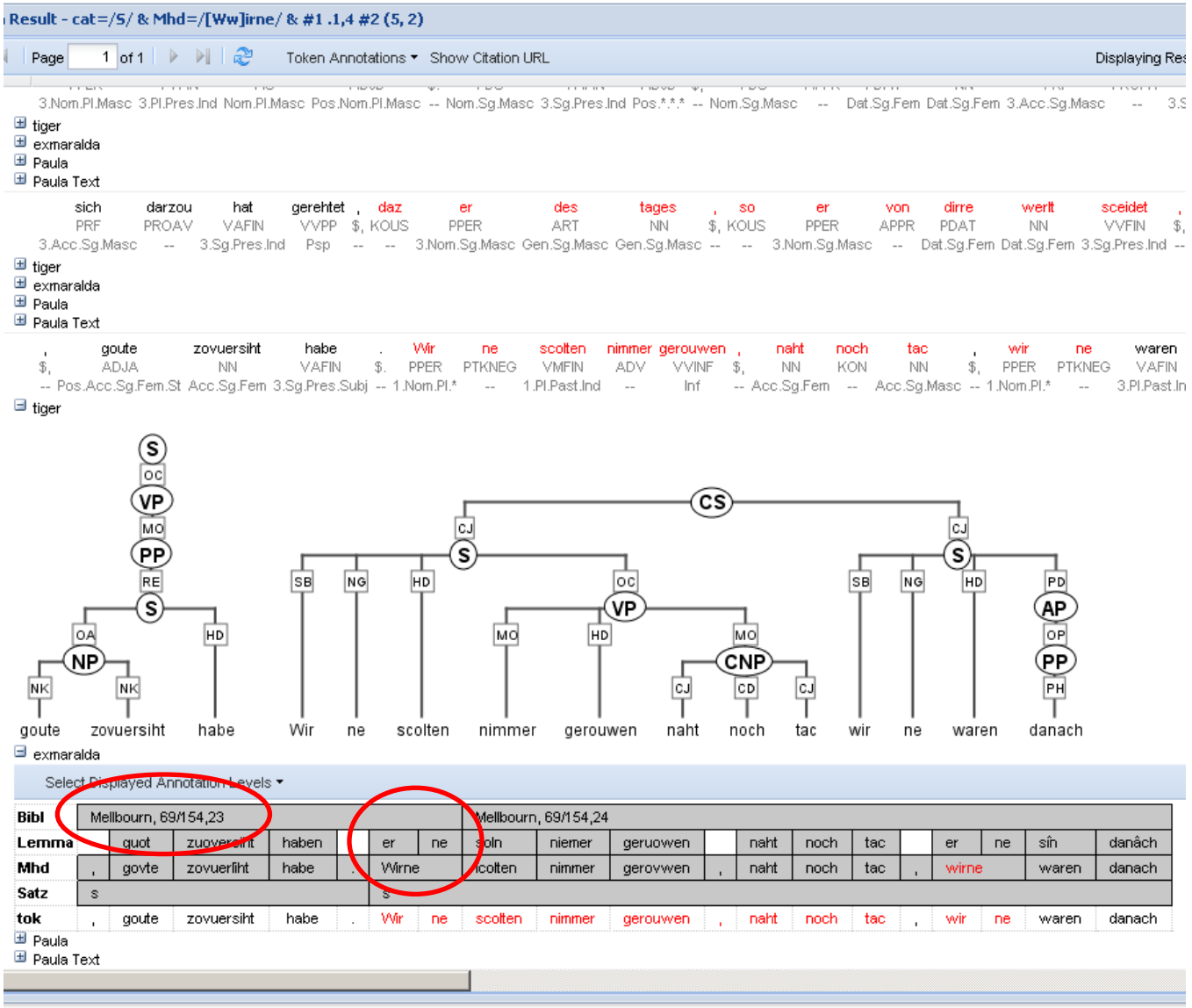
- IS (SFB guidelines)
- Morpho-syntactic annotation
- “Interlinear” Latin



MHG Speculum Ecclesiae

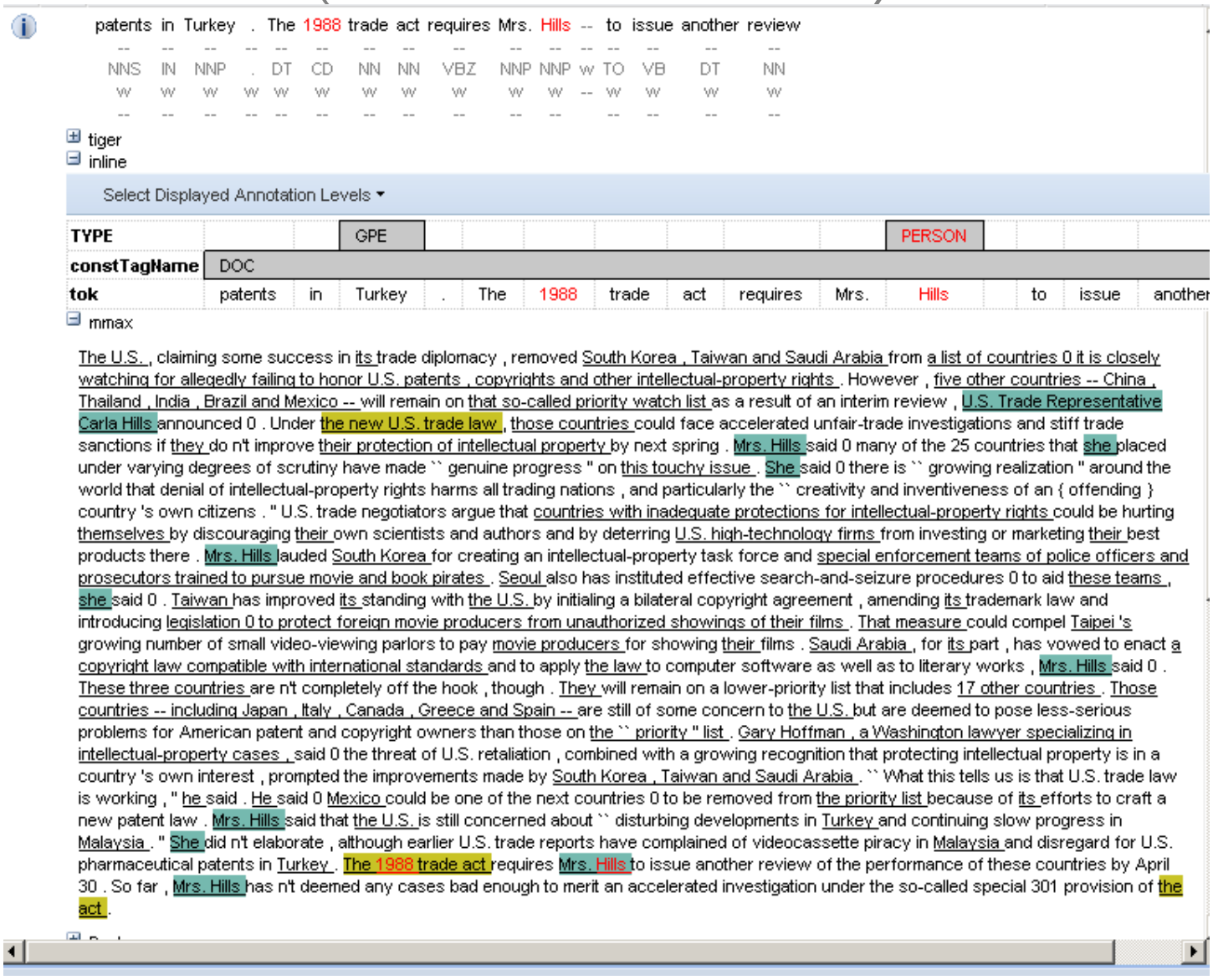
(WIP, Hagen Hirschmann/Sonja Linde et al.)

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

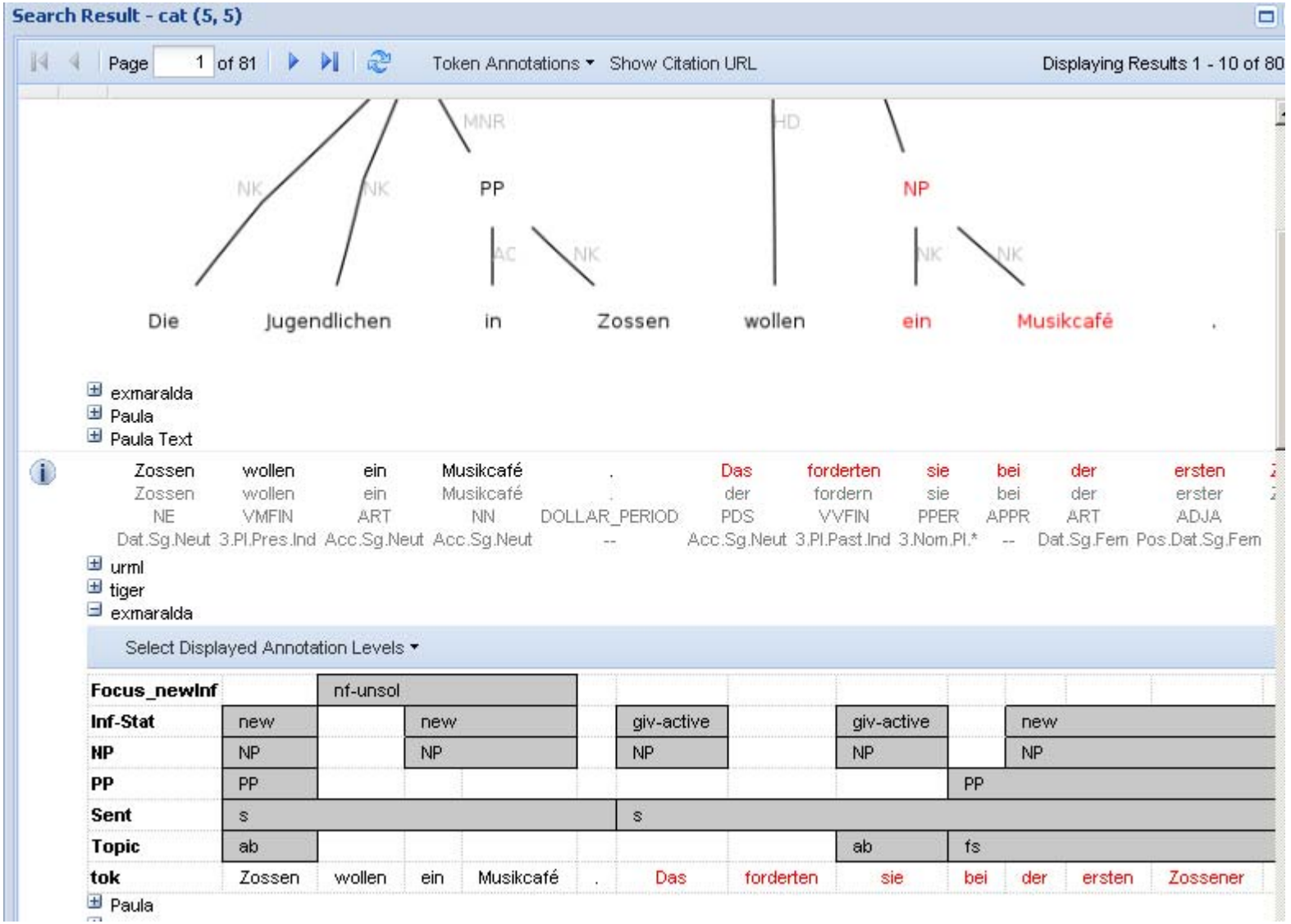


OntoNotes: Coreference/Syntax/NER

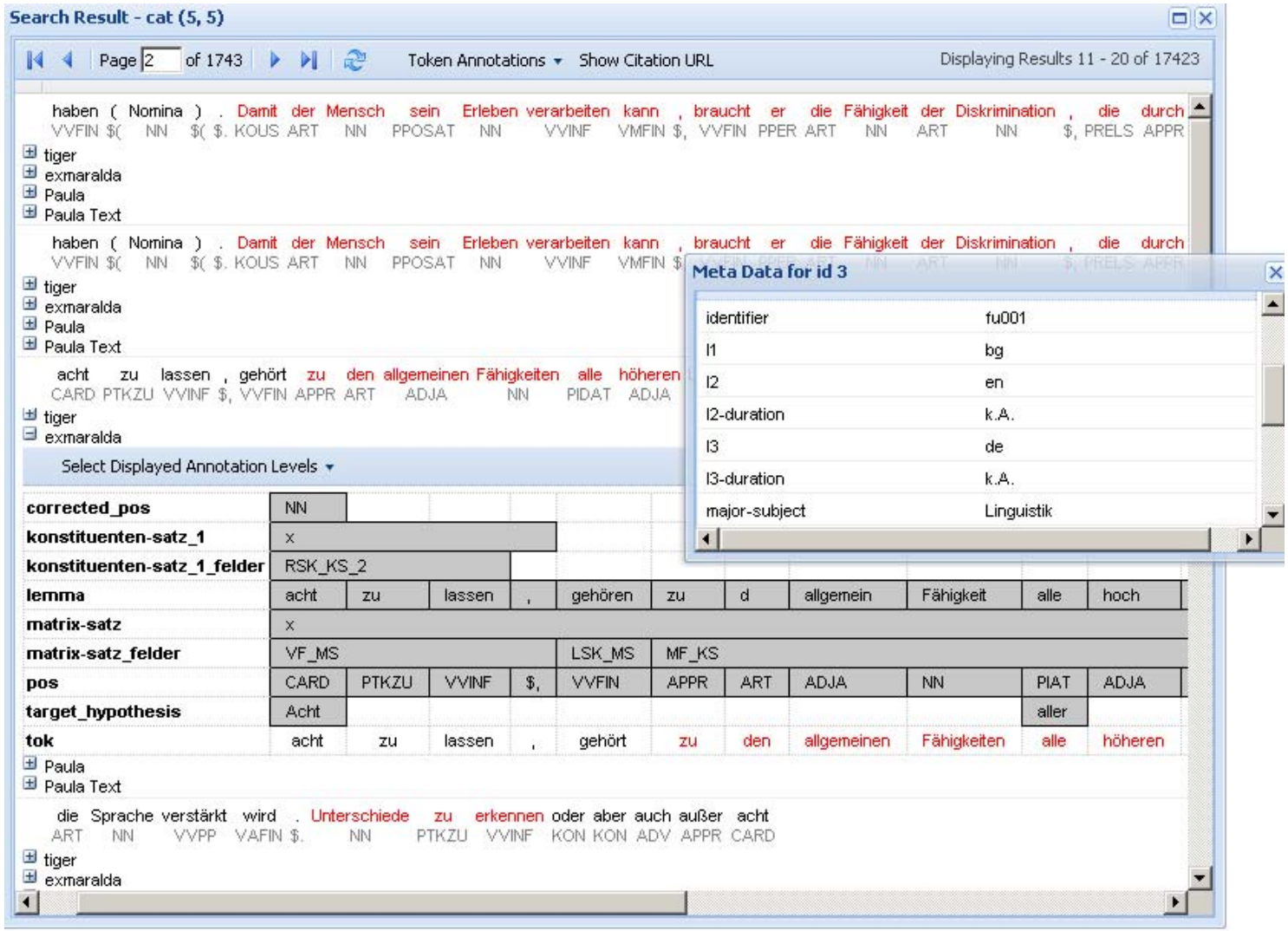
(Weischedel et al. 2008)



PCC: RST, IS (Stede 2004)



Falko: Topological Fields, Errors

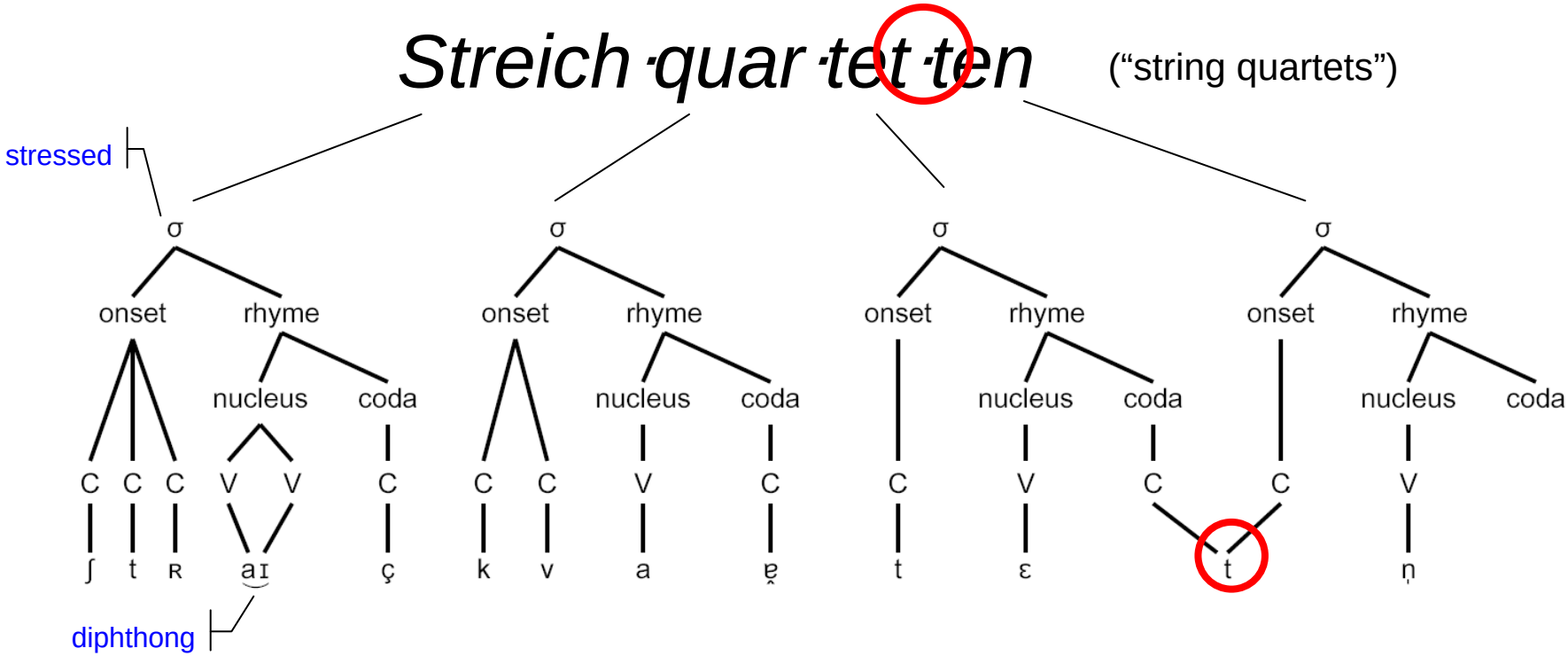


Future Directions / Open Issues

- 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 → operators?
- Ambiguous subtoken-token alignment



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)