Erstes Tübingen-Berlin Lernerkorpustreffen

16.10.2009

ANNIS2

Suche und Visualisierung von Mehrebenenkorpora

Amir Zeldes, Korpuslinguistik / HU Berlin

amir.zeldes@rz.hu-berlin.de

Hintergrund im SFB 632

- ANNIS: **ANN**otation of **I**nformation **S**tructure
- entwickelt im Rahmen von Teilprojekt D1 des SFB 632 „Informationsstruktur“ (Dipper et al. 2004): http://www.sfb632.uni-potsdam.de/

- Anforderungen:
  - Speicherung der Daten aus unterschiedlichen Teilprojekten in einem Standard
  - Umgang mit heterogenen Annotationen
  - Durchsuchung und Visualisierung komplex strukturierter Daten

Unterschiedliche Forschungsfragen

- Wie hat Informationsstruktur diachron die Wortstellung des Deutschen beeinflusst?
  (Petrova 2006, Donhauser et al. 2006, Petrova & Solf, to appear)
- Wie und wann werden Topiks in den tschadischen Sprachen markiert?
  (Hartmann 2006, Zimmermann, to appear)
- Wie interagiert Informationsstruktur sprachübergreifend mit Prosodie, Morphologie und Syntax?
  (Féry 2006, Fanselow 2007, Dipper et al. 2007)
- Kann man Informationsstruktur anhand von Morphosyntax, Lexis und Diskursstatus automatisch erkennen?

Mehrebenenkorpora

- beliebig viele Faktoren können für ein Phänomen relevant sein (bspw. Lernerfehler)
- man weiß nicht im Voraus, was diese sind
- mehrere Kategorisierungssysteme (Tagsets)
- mehrere Werte in derselben Position (alternative Fehlerzielhypothesen, Fehlerannotationen)
- konfligierende Hierarchien
- mehrere unabhängige Annotatoren und Tools

➢ beliebig viele Ebenen, beliebige Bezeichnungen und Werte

Formate

- Bäume/Graphen
  - Klammerformate (z.B. Penn, Bies et al. 1995)
  - Inline XML
  - TigerXML / Negra (Synpathy, Tiger, annotate, Brants & Plaehn, 2000)
  - RST tool (Rhetorische Satzbäume, O'Donnel 2000)
- Diskursrelationen
  - MMAX2 (Müller & Strube 2006)
  - PALinkA (Orasan 2003)
  - Serengeti (TODO)
- Partituren
  - EXMARaLDA (Schmidt 2004)
  - ELAN (Wittenburg et al. 2006)
  - Toolbox (Stuart et al. 2007)

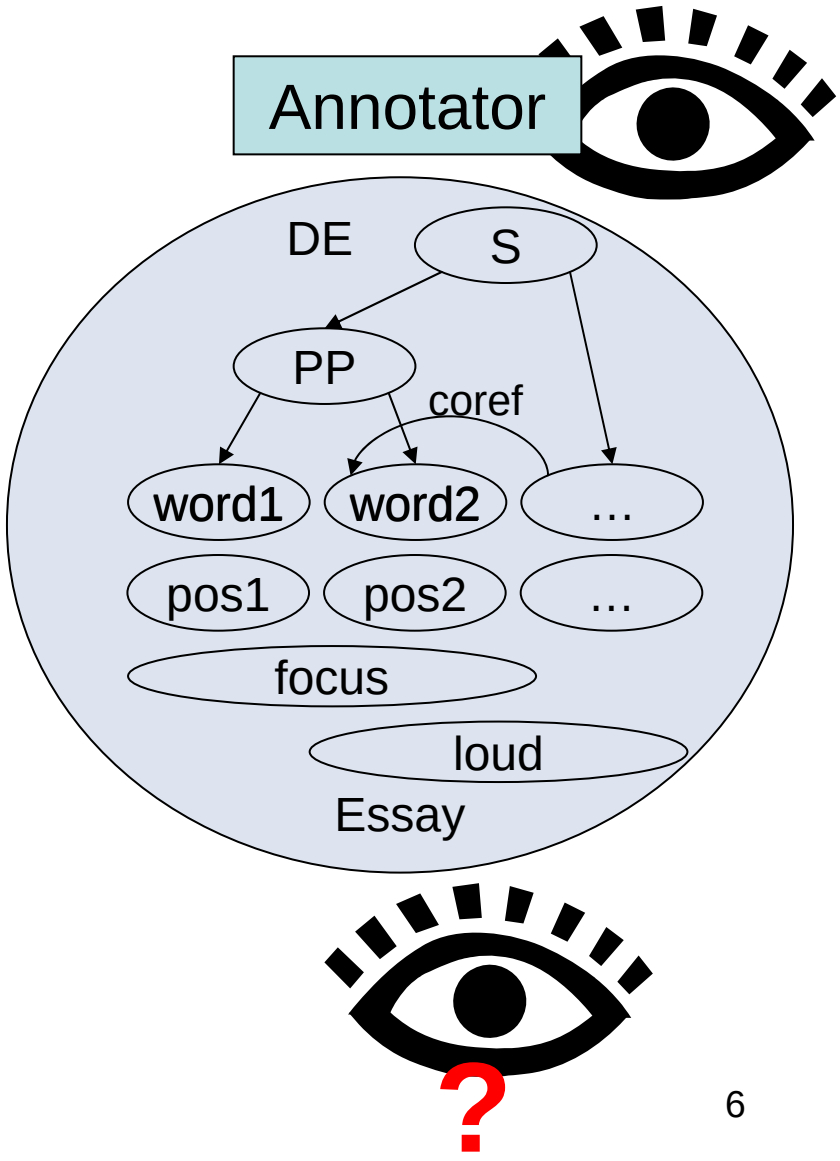
Aufgaben

- Representation heterogener evtl. konfligierender Datenstrukturen
- gleichzeitige Suche auf allen Ebenen
- angemessene Visulisierung fuer jede Datenart

  (syntaktische Konstituenten, Dependenzen, Morphologie, Phonologie, Koreferenz, multimodale Daten, …)

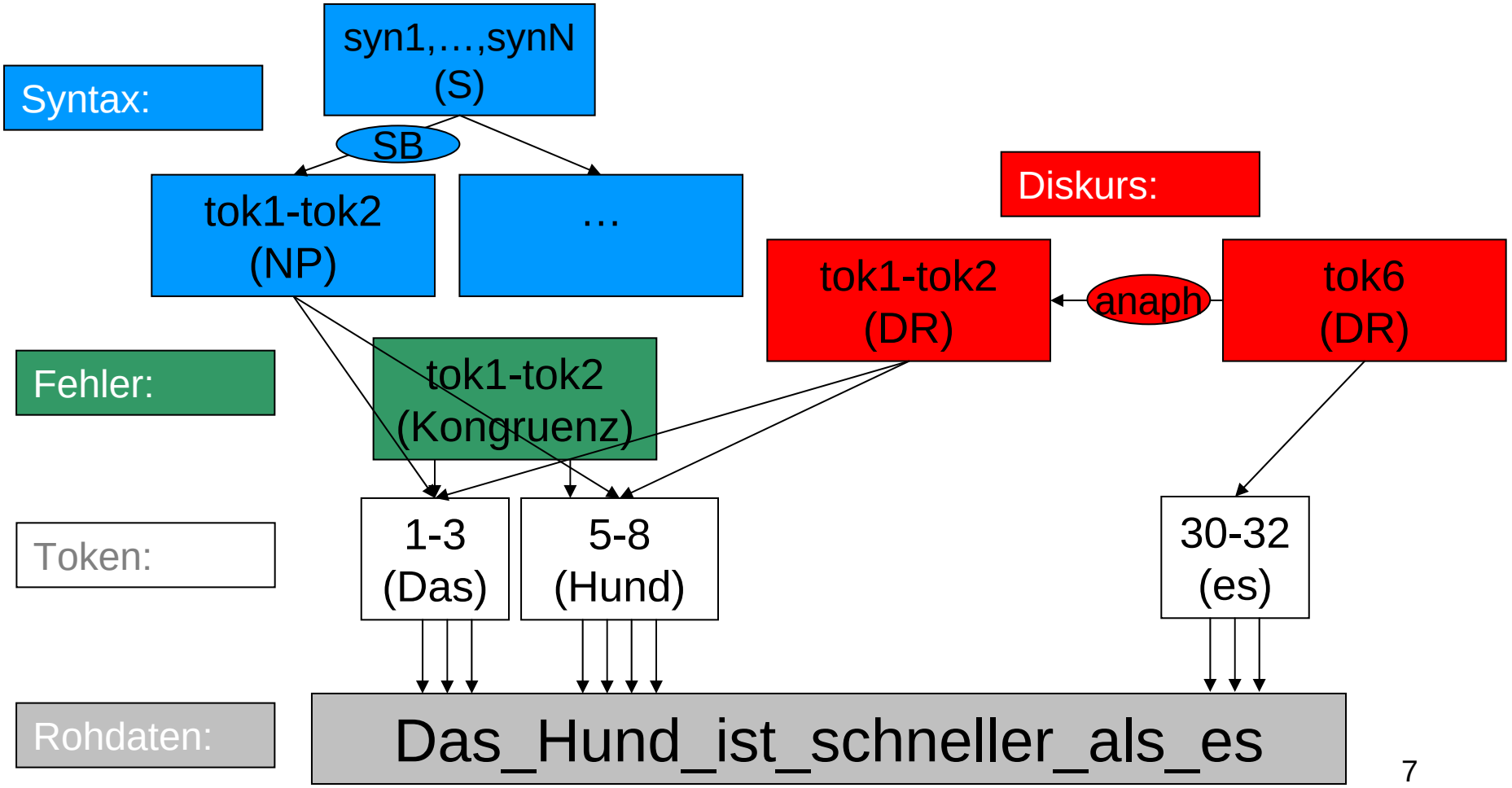
Mehrebenendaten

- Daten auf annotierte Knoten und Kanten reduzieren
- Unabhängige Ebenen (standoff)
  - Darstellung einer Graphenstruktur mit PAULA XML
  - Man kann Ebenen im Nachhinein verändern, ersetzen, hinzufügen
  - alternative Annotationen (mehrere Ebenen derselben Art)



PAULA standoff XML (vereinfacht)

(Dipper, 2005, Dipper & Götze, 2005)

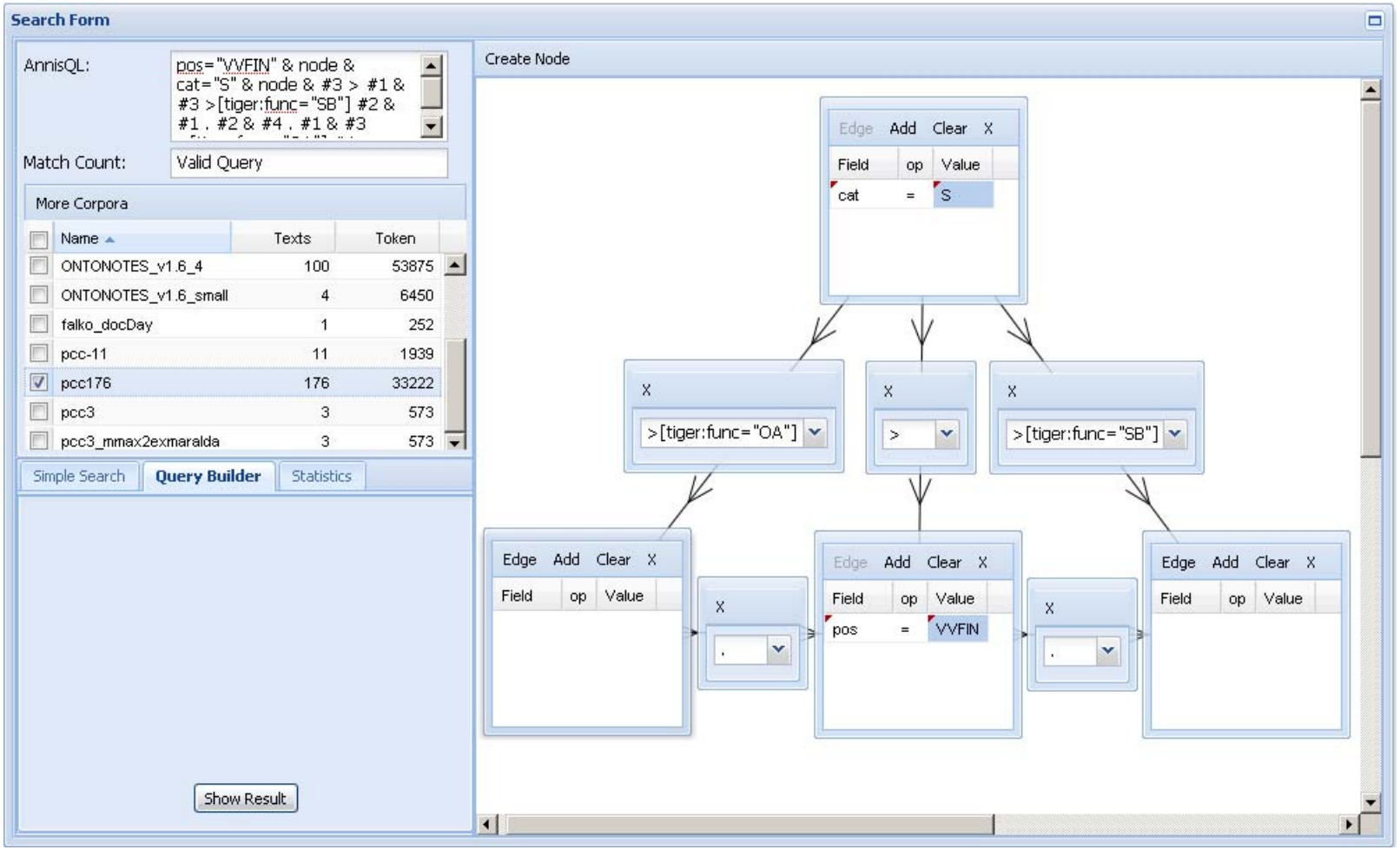


Suche mit AQL

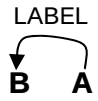
- SQL-Anfragen zu komplex
- Einfache Anfragesprache auf der Basis von gesuchten Elementen und Beziehungen:
  (vgl. *NiteQL* Carletta et al. 2003, *TIGERSearch* Lezius 2002)

```
cat="S" & node & cat="S" & pos="PPER" &
#1 >[func="SB"] #2 &
#3 >[func="SB"] #4 &
#4 ->anaph #2 &
meta::language="en"
```

Query Builder

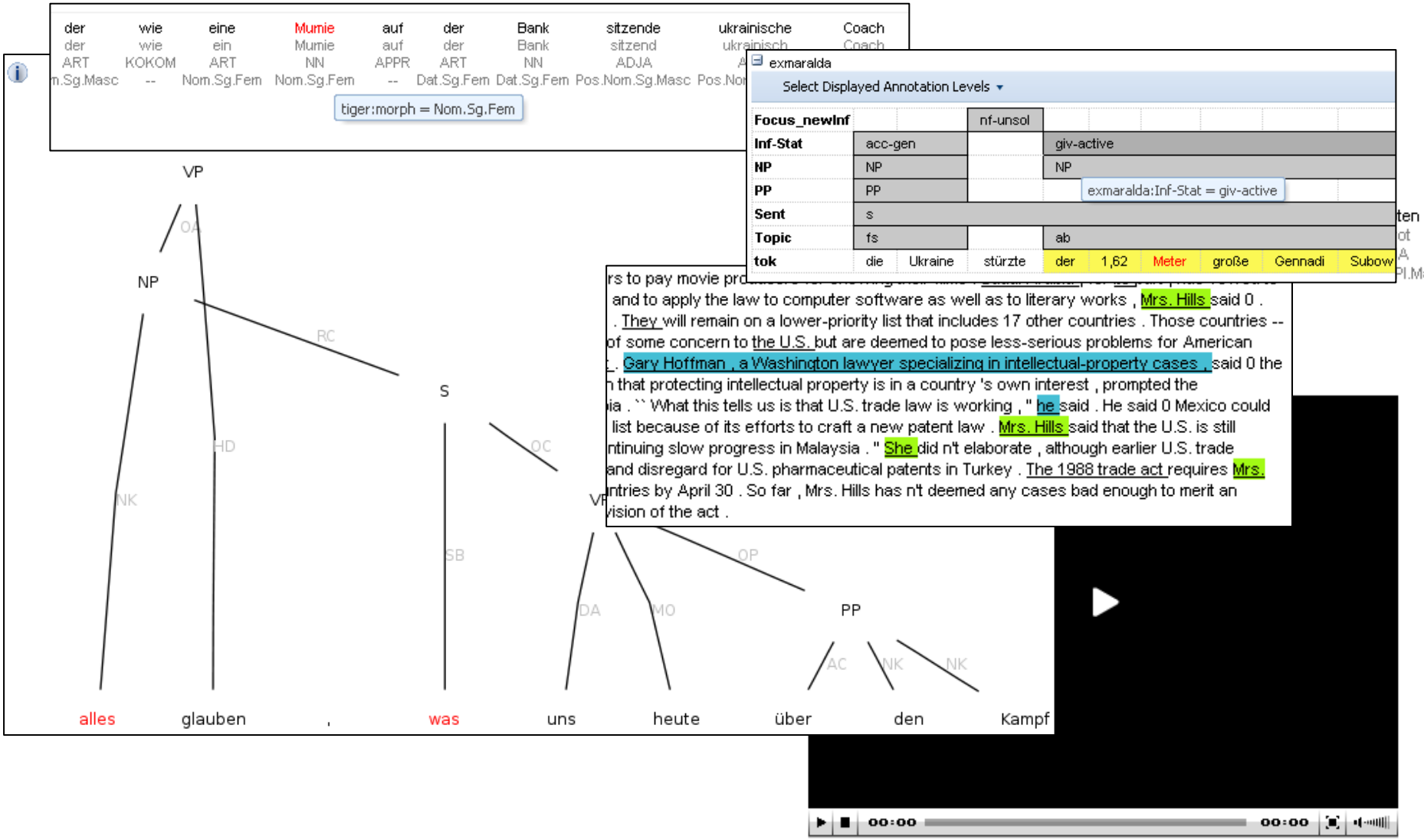


AQL Operatoren

	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		
_o_	overlap	**AAA** **BBB**	_ol_ _or_

	Name	Illustration
_=_	identical coverage	**A** **B**
_i_	inclusion	**AAA** **B**
_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**
$.*	Precedent + common parent node → Common ancestor node	x ... / \ **A B**

Visualisierung



Ausblick - Datenverarbeitung

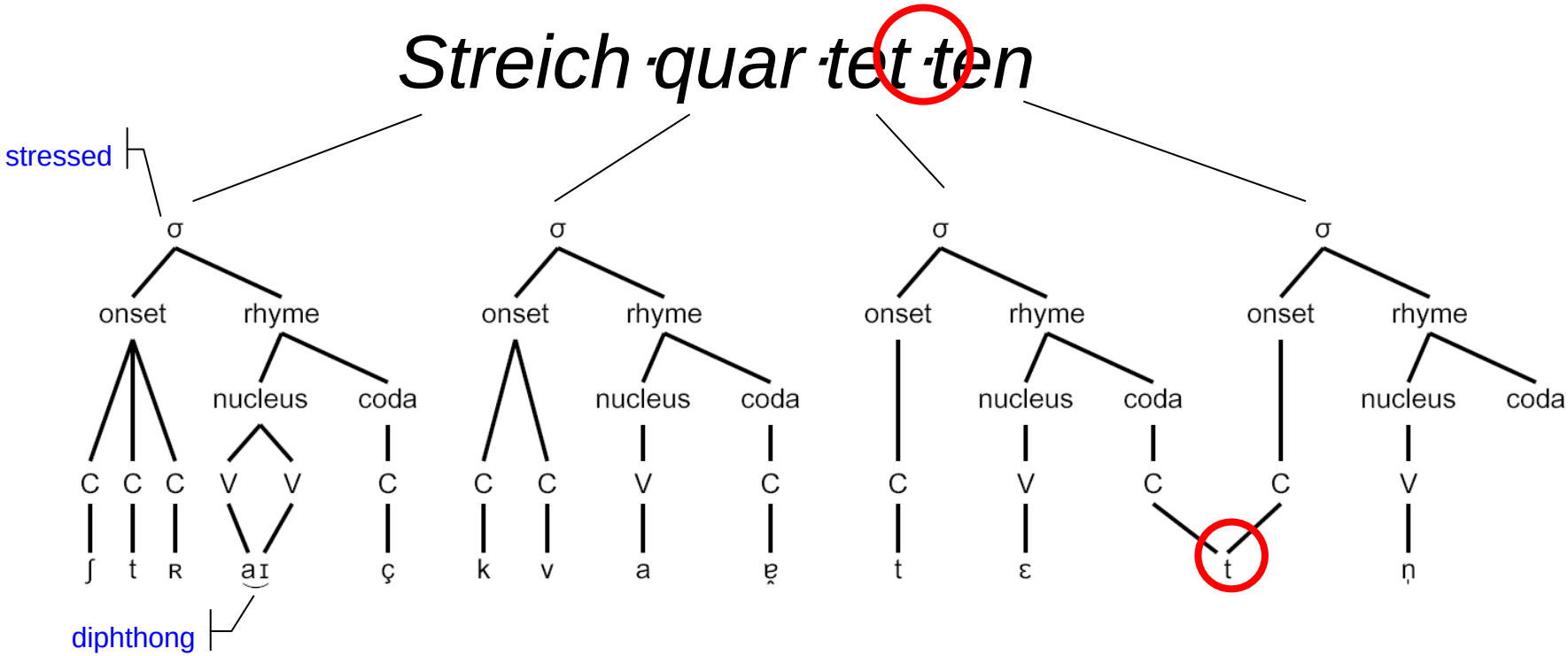
- abstraktes Datenmodel / API
- Konvertierung, Import, Export und Re-Import ausbauen
- aggregierter Export für statistische Auswertung (ARFF schon unterstützt)

Ausblick- Subtokenisierung

- Subtokenisierung (vgl. MAF, Clément & de la Clergerie 2005)
  - setzt Anspassung des Tokenkonzepts voraus:
    - „Physiche Token“ (atome) = kleinste Einheiten
    - „Referenztoken“ (bspw. Wortformen) = Referenzeinheit für die Suche (*“innerhalb von 5 token”, “adjazent”*) und den Kontext (*“10 Token links und rechts, sortiere nach dem 2. Token rechts”*)
  - Benutzer wissen nicht, ob gesuchte Wortformen Subtoken enthalten!

Ausblick- Subtokenisierung

- Subtoken-DAG, Annotationen > operatoren?
- Ambige Subtoken-Token-Alignierung



Ausblick - Parallelkorpora

- Suche momentan innerhalb eines Dokuments
- In Zukunft – mehrere Texte in einem Dokument
- Kanten unterstützen mehrfache Alignierung: paragraph/satz/wortweise… (TEI, Romary & Bonhomme 2000)
- Zirkuläre und unvollständige Alignierung
- Suche und Visualisierung

Ausblick - Anfragesprache

- Negation von:
  - Werten: pos!="ART"
  - Operatoren (mit und ohne Existenzannahme)
    #1 !> #2 (existiert #1?)

Vielen Dank!

An ANNIS2 arbeiten:

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

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

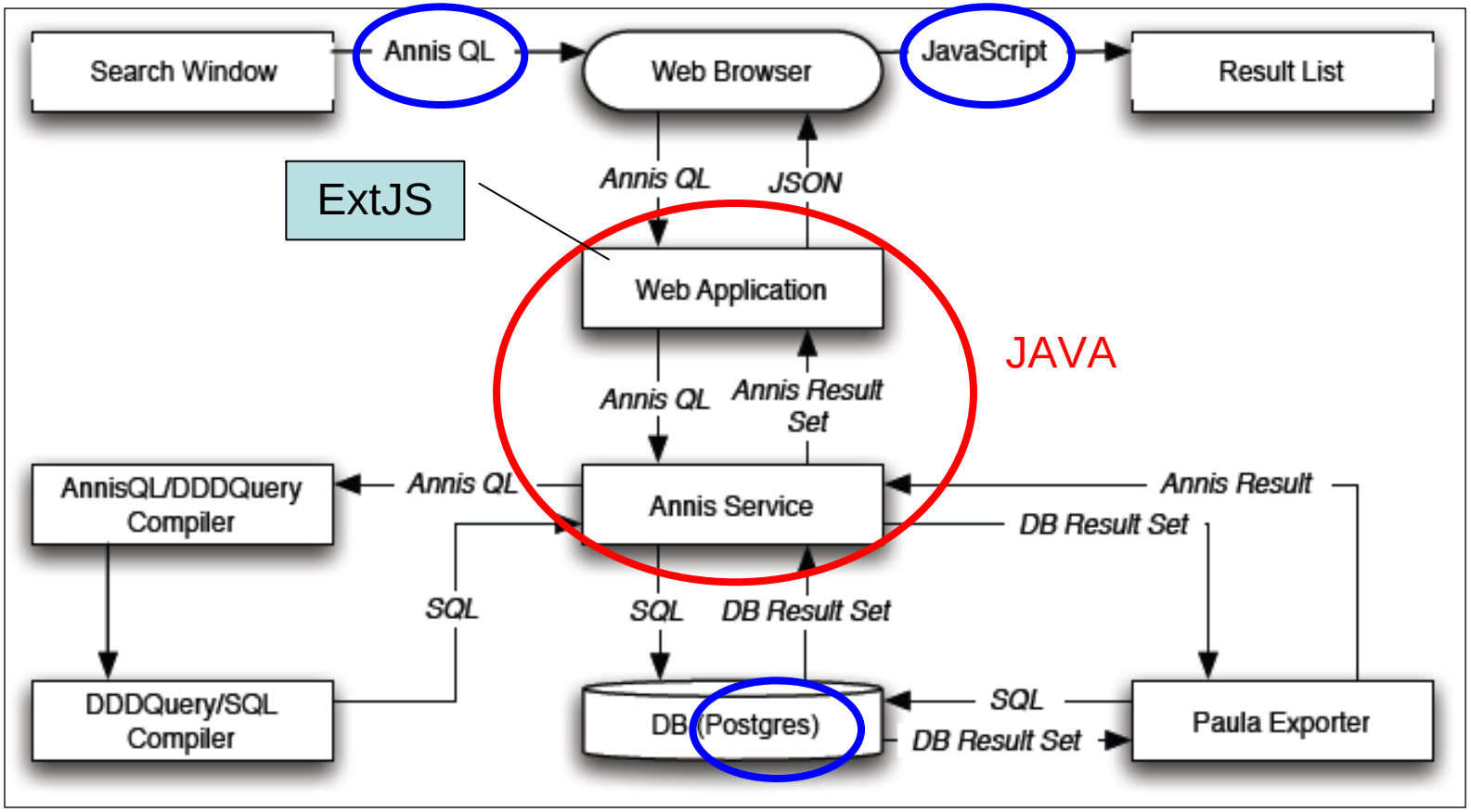
Literatur

- 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.
- 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: *Proceedings of the 2nd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, 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.

Literatur II

- Ide, N./Romary, L. (2004a) International Standard for a Linguistic Annotation Framework. *Natural Language Engineering*, 10 (3-4), 211-225.
- Ide, N./Romary, L. (2004b), A Registry of Standard Data Categories for Linguistic Annotation. *Proc. 4th International Conference on Language Resources and Evaluation*, 135-138.
- 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.
- 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.
- Stuart, R./Aumann, G./Bird. S. (2007) Managing Fieldwork Data with Toolbox and the Natural Language Toolkit. *Language Documentation & Conservation* 1(1), 44–57.
- Wittenburg, P./Brugman, H./Russel, A./Klassmann, A./Sloetjes, H. (2006) ELAN: a Professional Framework for Multimodality Research. In: *Proceedings of LREC 2006*.
- 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.

Struktur – Annis2



- Lizenz: Apache 2.0

RelAnnis (vereinfacht)

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