

ANNIS2

Suche und Visualisierung von Mehrebenenkorpora

Amir Zeldes, Korpuslinguistik / HU Berlin

amir.zeldes@rz.hu-berlin.de



Search Form

AnnisQL: "Auto" & /.* / & /.* / & #1 .* #2 & #2 .* #3 | cat="s" & "das" & "Dorf" & #4 > * #5 & #5 . #6 (0, 0)

Match Count: 3

More Corpora	Texts	Token
<input type="checkbox"/> b4.muspill	50	1000
<input type="checkbox"/> b4.tatian	50	1000
<input type="checkbox"/> c6.hindi	50	1000
<input checked="" type="checkbox"/> d2.2samplesDEU	2	19
<input type="checkbox"/> falko.essay	133	66000
<input checked="" type="checkbox"/> pcc-11	11	1939
<input type="checkbox"/> spec	4	2823

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

Displaying Results 1 - 3 of 3

Debel hätte das Dorf jede Einnahme nötig
 dabei haben der Dorf jeder Einnahme nötig
 Nom.Sg.Fem Subj Nom.Sg.Neur Nom.Sg.Neur Acc.Sg.Fem Acc.Sg.Fem Pos

exmaralda

Select Displayed Annotation Levels

Inf-StatSeg giv-active

PSe NP NP NP NP NP NP

SentSeg s

bei hätte das Dorf jede Einnahme nötig

tiger

Show Result

Dabel hätte das Dorf jede Einnahme nötig

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)

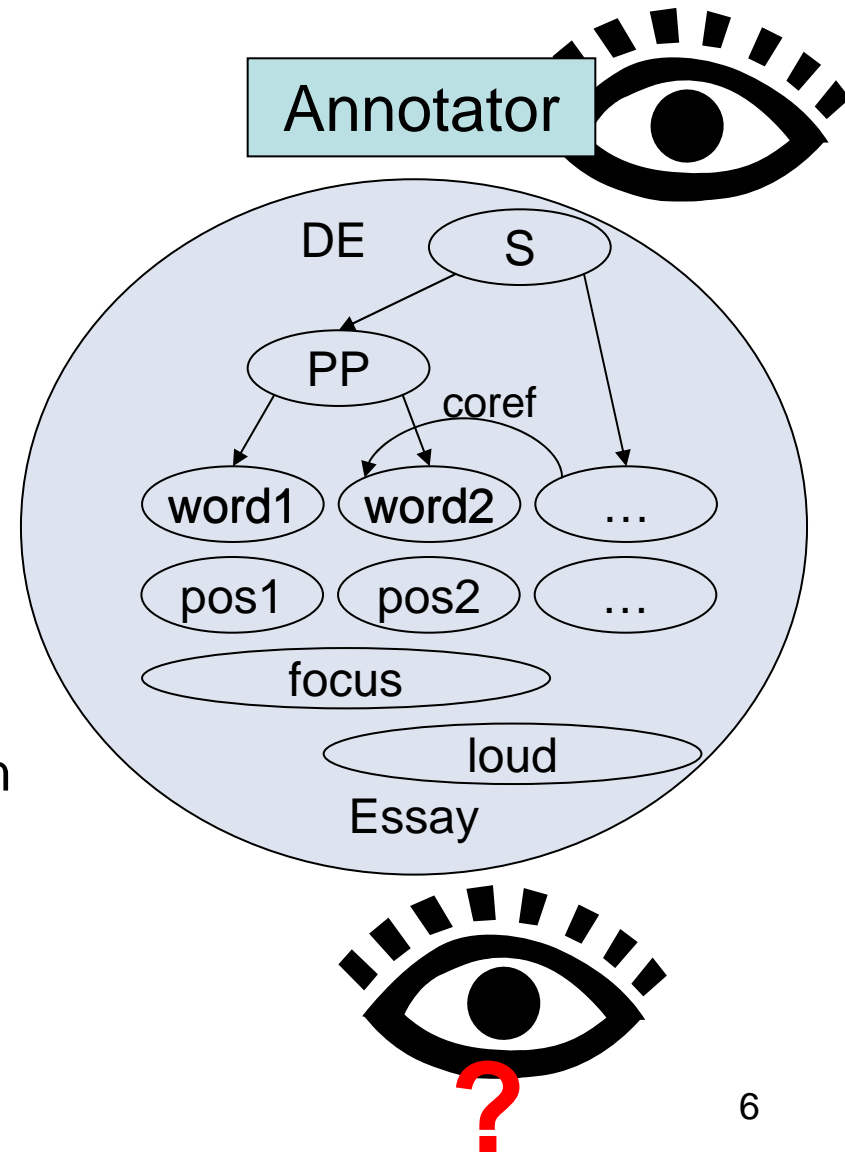
Aufgaben

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

(syntaktische Konstituenten, Abhängigkeiten, Morphologie, Phonologie, Koreferenz, multimodale Daten, ...)

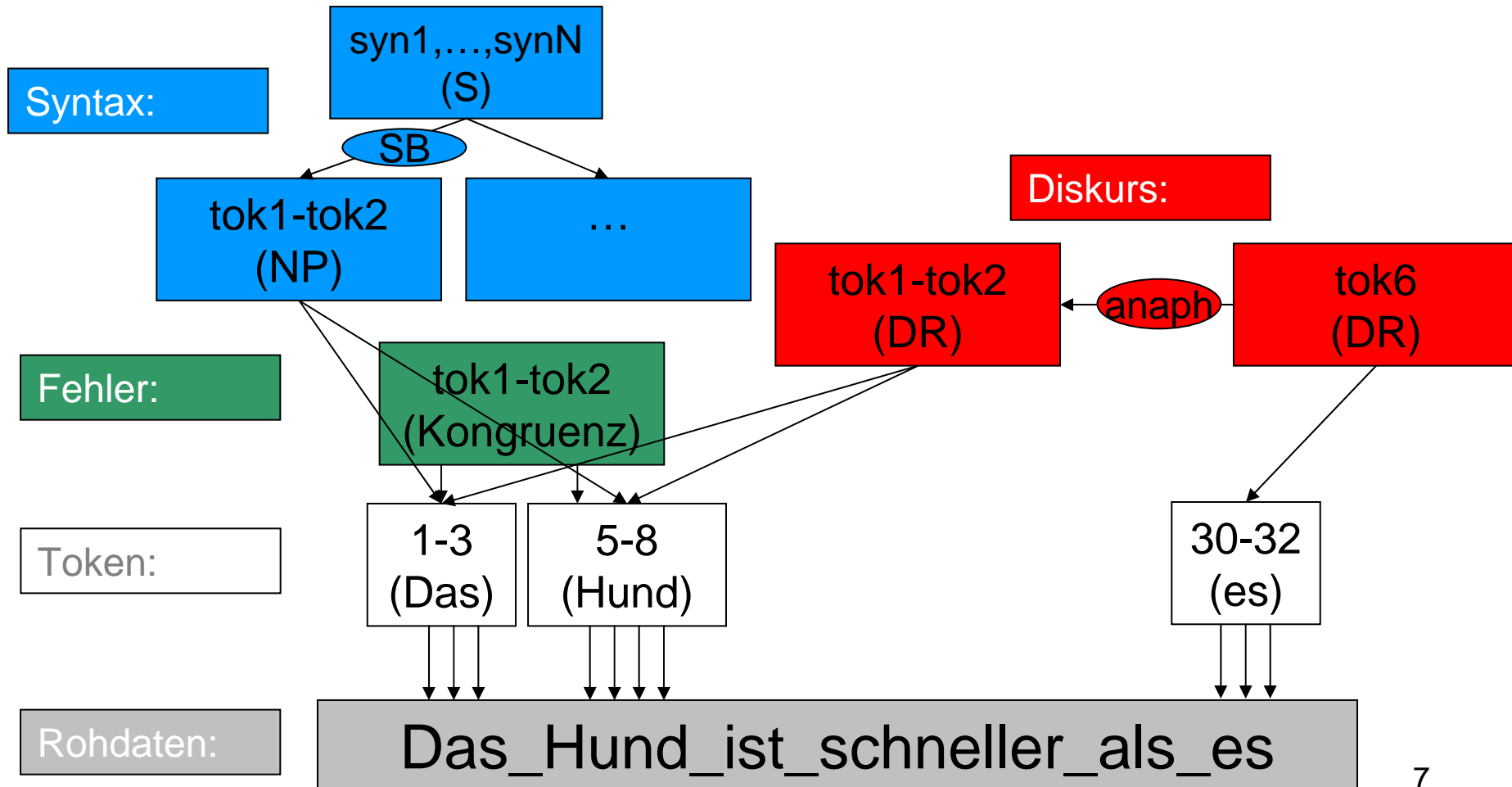
Mehrebenenendaten

- 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

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

Edge Add Clear X

Field	op	Value
cat	=	S

X

> [tiger:func="OA"]

X

>

X

> [tiger:func="SB"]

Edge Add Clear X

Field	op	Value
.		

X

Edge Add Clear X

Field	op	Value
pos	=	VVFIN

X

Edge Add Clear X

Field	op	Value
.		

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 ↙ ↘ B A	
->LABEL*	Labeled pointing path	LABEL LABEL ↙ ↘ ↙ ↘ B x y z A	
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

der	wie	eine	Mumie	auf	der	Bank	sitzende	ukrainische	Coach
der	wie	ein	Mumie	auf	der	Bank	sitzend	ukrainisch	Coach
ART	KOKOM	ART	NN	APPR	ART	NN	ADJA		
n.Sg.Masc	--	Nom.Sg.Fem	Nom.Sg.Fem	--	Dat.Sg.Fem	Dat.Sg.Fem	Pos.Nom.Sg.Masc	Pos.Nom	

tiger:morph = Nom.Sg.Fem

exmaralda
 Select Displayed Annotation Levels ▾

Focus_newInf		nf-unsol		
Inf-Stat	acc-gen		giv-active	
NP	NP		NP	
PP	PP		exmaralda:Inf-Stat = giv-active	
Sent	s			
Topic	fs		ab	
tok	die	Ukraine	stürzte	der 1,62 Meter große Gennadi Subov

VP
 OA
 NP
 RC
 HD
 NK
 S
 OC
 SB
 VP
 DA
 MO
 PP
 AC
 NK
 NK

alles glauben, was uns heute über den Kampf

rs to pay movie pro

and to apply the law to computer software as well as to literary works , Mrs. Hills said 0 .

. They will remain on a lower-priority list that includes 17 other countries . Those countries --

of some concern to the U.S. but are deemed to pose less-serious problems for American

Gary Hoffman , a Washington lawyer specializing in intellectual-property cases , said 0 the

n that protecting intellectual property is in a country 's own interest , prompted the

ia . " What this tells us is that U.S. trade law is working , " he said . He said 0 Mexico could

list because of its efforts to craft a new patent law . Mrs. Hills said that the U.S. is still

ntinuing slow progress in Malaysia . " She did n't elaborate , although earlier U.S. trade

and disregard for U.S. pharmaceutical patents in Turkey . The 1988 trade act requires Mrs.

entries by April 30 . So far , Mrs. Hills has n't deemed any cases bad enough to merit an

vision of the act .

▶

00:00

00:00

Ausblick - Datenverarbeitung

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

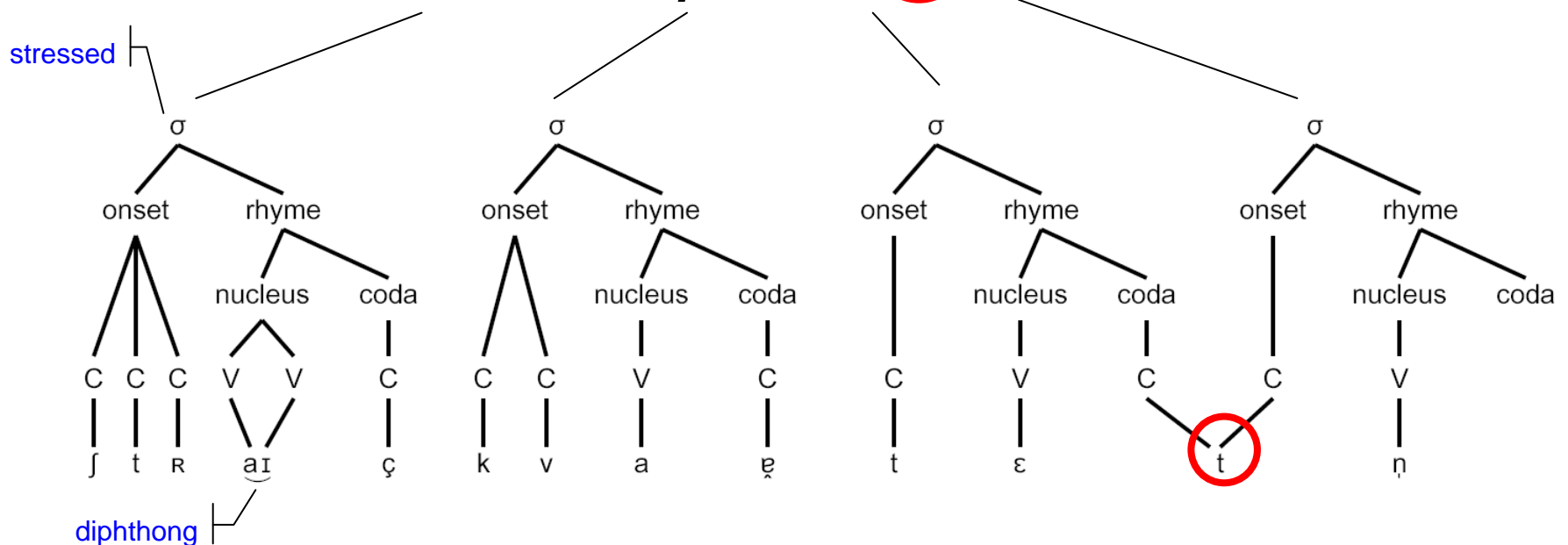
Ausblick- Subtokenisierung

- Subtokenisierung (vgl. MAF, Clément & de la Clergerie 2005)
 - setzt Anpassung des Tokenkonzepts voraus:
 - „Physische 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

Streich·quar·tet·ten



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/>



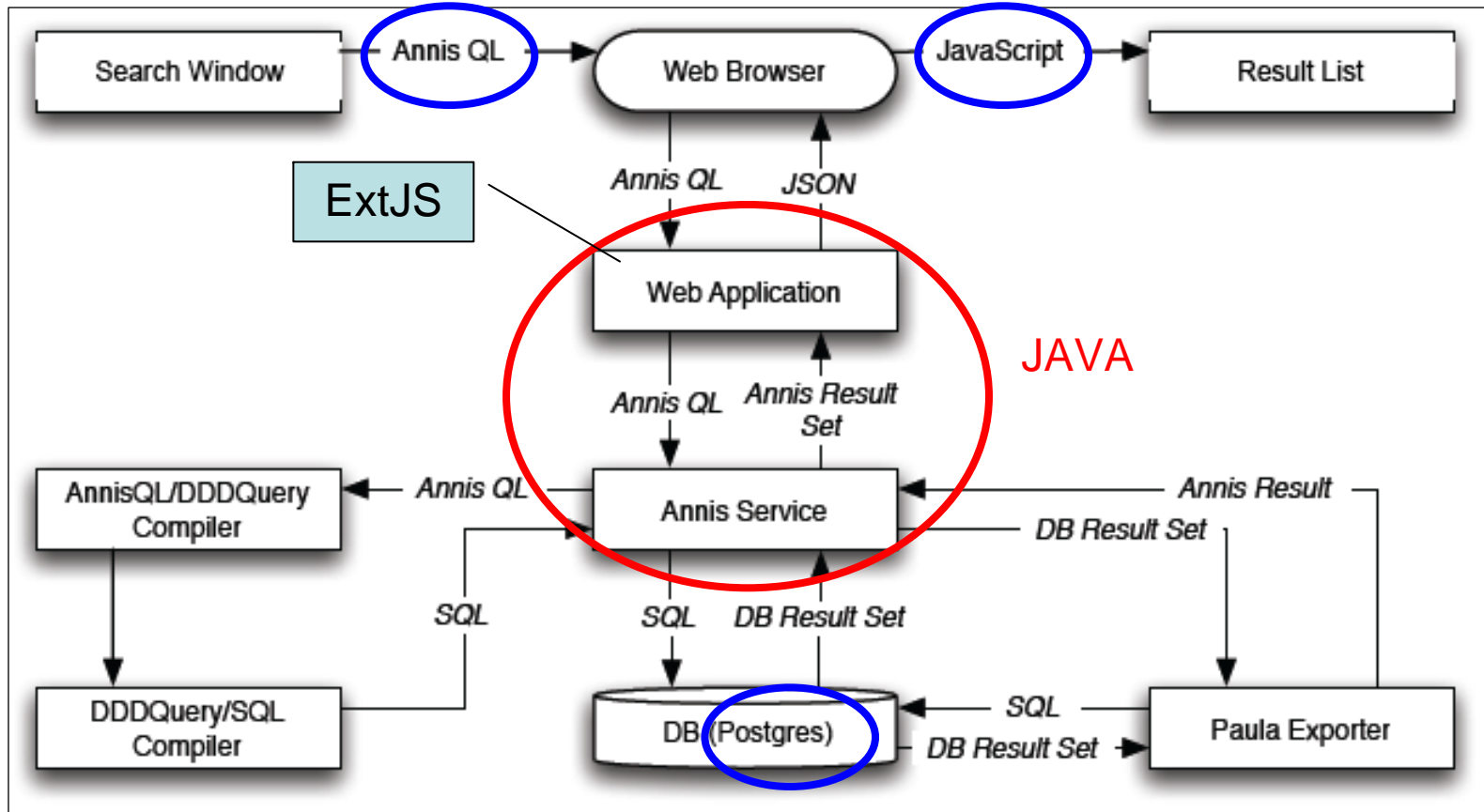
Literatur

- 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)

