31. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft, Postersession der Sektion Computerlinguistik, 5.3.2009

Search and Visualization of Richly Annotated Corpora with ANNIS2



SFB 632 "Information Structure" Project D1: Linguistic Database Annotation and Retrieval

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

+ Humboldt-Universität zu Berlin, * Universität Potsdam



Background

- ANNIS2 is a versatile web browser-based search and visualization architecture for complex multilevel linguistic corpora
- Designed to provide search and visualization facilities for complex annotations such as information structure and coreference created in different subprojects of the SFB 632
- Developed for research involving diverse annotations simultaneously, allowing studies of the interaction between different phenomena (e.g. syntax or prosody and information structure)

NNIS2 🔻			
earch Form		Search Result - "Auto" & /.*/ & /.*/ & #1 .* #2 & #2 .* #3 cat="5" & "das" & "Dorf" & #4 >* #5 & #5 . #6 (0, 0)	
AnnisQL:	"Auto" & /.*/ & /.*/ & #1 .* #2	I A Page T of 1 P PI C Token Annotations * Show Citation URL	Displaying Results 1 - 3 of 3
	& #2 .* #3 cat="S" & "das" &		

Visualization Modules

- Diverse data from heterogeneous corpora requires multiple visualizations
- Simultaneous querying and visualization of different annotations enables analysis of interdependency between different types of data

ANNIS2 uses extensible Java plug-in based modules for different types of data. Currently supported are:

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	ADJA	NN
n So Masc		Nom Sa Fem	Nom Sa Fem		Dat Sq Fem	Dat So Fem	Pos Nom Se Mase	Pos Nom Se Mase	Nom So Maso

tiger:morph = Nom.Sg.Fem



- The system imports data in the PAULA XML format (Dipper 2005), which allows multiple independent standoff annotations with conflicting hierarchies, created in different tools
- Converters exist for EXMARaLDA (also from MS Excel sources), TigerXML (Annotate/Synpathy), MMAX2, RSTTool and PALinkA, and an API allowing greater import/export functionality is currently being developed
- An advanced relational database representation of the data provides infrastructure for fast searches in large and diverse hierarchical datasets

Search with ANNIS Query Language – AQL

- Based on a simple model of richly annotated nodes and labelled edges
- Similar to Tiger Query Language and NITE-QL, search nodes are declared (annotations, terminals and non-terminals) and bound by operators
- For example, the following query finds German OVS sentences in a syntactically annotated corpus:

node & pos="VVFIN" & cat="S" & node &	two no
#3 >[func="OA"] #1 &	S dom
#3 >[func="SB"] #4 &	S dom
#3 > #2 &	S dom
#1 .* #2 &	node 1
#2 .* #4	the ver

wo nodes, a verb and an S-node S dominates node 1 with label OA S dominates node 4 with label SB S dominates the verb node 1 precedes the verb the verb precedes node 4

sen	wollen	ein	Musikcafé		Das	forderten	sie	bei	der	ersten	Zosser
sen	wollen	ein	Musikcafé		der	fordern	sie	bei	der	erster	Zosser
IE	VMEIN	ART	NN	\$.	PDS	VVFIN	PPER	APPR	ART	ADJA	ADJA
g.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	Pos.*.

KWIC-style token based annotations

rs to pay movie producers for showing their films . <u>Saudi Arabia</u>, for <u>its</u> part , has and to apply the law to computer software as well as to literary works , <u>Mrs. Hills</u> inf-Stat . <u>They</u> will remain on a lower-priority list that includes 17 other countries . Those counts for some concern to <u>the U.S.</u> but are deemed to pose less-serious problems for Ame <u>PP</u>. . <u>Gary Hoffman</u>, a Washington lawyer specializing in intellectual-property cases , that protecting intellectual property is in a country 's own interest , prompted the is . `` What this tells us is that U.S. trade law is working , " <u>he</u>said . He said 0 Mexic list because of its efforts to craft a new patent law . <u>Mrs. Hills</u> said that the U.S. is ntinuing slow progress in Malaysia . " <u>She</u> did n't elaborate , although earlier U.S. trade and disregard for U.S. pharmaceutical patents in Turkey . <u>The 1988 trade act</u> requires <u>Mrs.</u> Intries by April 30 . So far , Mrs. Hills has n't deemed any cases bad enough to merit an *v* ision of the act .

t Displayed Annotation Levels 👻									
ewInf		nf-unsol							
	acc-gen		giv-active						
	NP		NP						
	PP			exmaral	da:Inf-Stal	t = giv-act	ive		
	s								
	fs		ab						
	die Ukraine	stürzte	der	1,62	Meter	große	Gennadi	Subow	

Grids for annotating continuous and discontinuous spans of tokens

Discourse view for discourse referents and coreferent expressions (anaphors/antecedents)



exmaralda

Audio/Video data aligned with token spans

Outlook

For upcoming versions of ANNIS, we will be concentrating on the following expansions:

- Support for the annotation of subtoken units
- Expanding AQL with negation, with and without implicit existence of nodes

- A graphical Query Builder makes formulating complex queries easier
- Operators define the possible overlap and adjacency relations between annotation spans, as well as recursive and labelled hierarchical relations between nodes
- ArrisQU: dt="NP"& pos="VVFIN" & ctat="5" & dt="NP" & #3 > (fare="0x1" #1 & #3 & #1 . #2 & w Match Count: Calculating... More Corpora b 4mmaphin 286 1502 C 6hind 1 2218 d 22smqlesDEU 2 19 f take assay 1 11 1939 g spec 4 2023 * Smple Search Query Builder Statistics Edge Add Clear X Field op Value Field op Value Field op Value Field op Value Field op Value
- Users can search in multiple corpora simultaneously
- Full support for Regular Expressions (RegEx) in tokens, annotations and edge label values
- Full Unicode support in both search and visualization, including RegEx

4	Page 1 of 1 🕨 🕅 🔅	Token Annotations 🔹 Show Citation URL
	नहीं मिलती। पर अब समय आ	गया है जब हम उनके
	± inline	DPDARMN

 Search result nodes and their annotations can be exported to WEKA (Witten & Frank 2005) for machine learning

- Visualization and search for parallel corpora using alignment operators
- Integration of flexible statistical functionality using aggregate functions and user defined mathematical functions
- More export / import filters to allow modification and update of corpora and subcorpora
- Supporting new data types from more annotation tools, such as Serengeti coreference data
- Development of specialized visualizations for supported data types (for example rhetorical structure document trees based on RST annotations)
- Support for annotations calling external APIs, for example lexica or cartographic resources such as Google Earth's KML for geographical annotations (e.g. to visualize distributions of dialect features)
- A shopping-cart style interface for collecting interesting results and exporting them to a file
- Enabling the embedding of specialized non-Unicode conformant fonts for historical data