

Tutorium der DGfS-Sektion Computerlinguistik

KORPUSLINGUISTIK

MIT ONLINE -

RESSOURCEN

EINE INTERAKTIVE

EINFÜHRUNG

FÜR LINGUISTEN



Annotation

Stefanie Dipper

Stefan Evert

Heike Zinsmeister

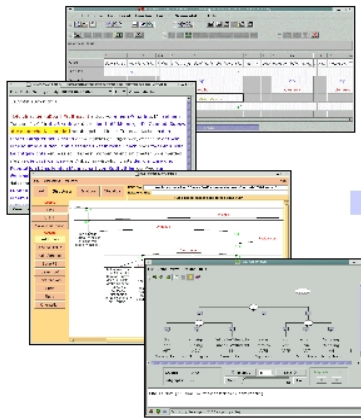
München, 28.1.2011

Manuelle Annotation

- komplexe Phänomene: manuelle Annotation
- Beispiel Informationsstruktur
 - komplexes Zusammenspiel vieler Faktoren
 - z.B. Informationsstruktur, Topik, Fokus
- Idee: Annotation der einzelnen Faktoren erlaubt Untersuchung der Interaktion

PAULA und ANNIS

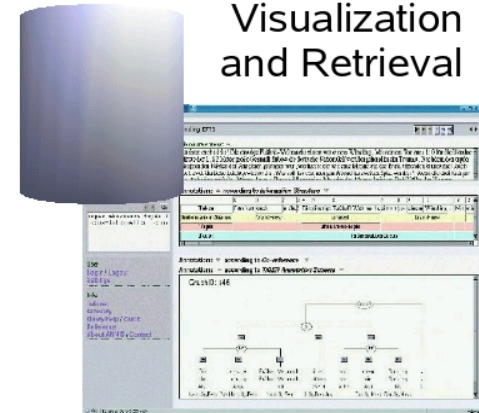
Data Annotation
*Exmaralda, MMAX,
RST Tool, annotate*



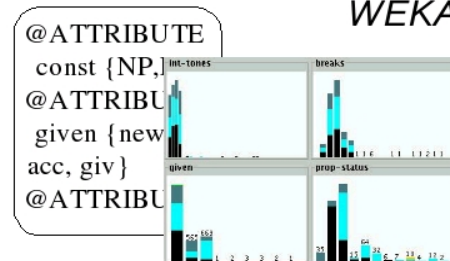
PAULA: XML-based
Standoff Interchange
Format

```
<markList  
type="tok"  
xml:base=  
"text.xml">  
...  
</markList>
```

ANNIS
Linguistic Database:
Visualization
and Retrieval



Statistical Evaluation
WEKA



SFB "Informationsstruktur"
(Potsdam/Berlin)

Exmaralda

And then you see ehm a maan in maybe

	0 [00.0]	1 [01.9]	2 [01.9]	3 [03.2]	4	5 [09.2]	6 [10.4]
		louder					
X [v]	So it starts out with: A	roo	ster crows	. ((1,3s)) ((takes breath 0,5s))	And then you see ehm a maan in maybe	his fifties. ((takes breath, 0,5s))	And so h
X [nv]	rHA on rKN, iHA on ISH	rHA up and to the right	rHA stays up		rHA back down rKN, moves iHA on ISH	rHA to the right	bHA up, rH
X [nv]			HE nods once				
X [nv]		emphasizes the crow				aproximately fifty	
Y [v]							

Done.

SFB "Mehrsprachigkeit" (Hamburg)

MMAX2



HITS (Heidelberg; früher EML)

RST Tool

The screenshot shows the RST Tool interface for a document titled 'zpgtext5.rs2'. The interface includes a menu bar (File, Options, Help, Debug) and a toolbar with buttons for Text, Structurer, Relations, Statistics, and Quit. A left sidebar contains 'Modes' (Link, Unlink, Collapse/Expand) and 'Actions' (Add Span, Add MultiNuc, Add Schema, Save PS, Save PDX, Print Canvas). The main canvas displays a document with RST annotations:

- 4-29** (green) is linked to **Evidence** (red) via a horizontal arrow.
- 14-22** (green) is linked to **23-29** (green) via a horizontal arrow.
- 18-22** (green) is linked to **volitional-result** (red) and **Nonvolitional-result** (red) via a downward arrow.
- 20-21** (green) is linked to **Joint** (red) via a downward arrow.
- 15A)** and **15B)** are linked to **Joint** (red) via a downward arrow.
- 16)** is linked to **Nonvolitional-result** (red) via a horizontal arrow.
- 17)** is linked to **Evidence** (red) via a horizontal arrow.

The document text includes: Michael O'Donnell WagSoft Linguistic Software, 15A) Our small staff is being swamped with requests for more information, 15B) and our modest resources are being stretched to the limit, 16) Your support now is critical, and 17) ZPG's 1985 Urban Stress Test may be our best opportunity ever to get the population message heard.

Michael O'Donnell, WagSoft

Annotations-Richtlinien

Für die Interpretation und Wieder-
Verwendbarkeit von annotierten Daten:

- Bedeutung der Annotation muss klar sein
 - Welche Symbole gibt es?
 - Wie sind sie definiert?
 - z.B. NN = normales Nomen
 - Kriterien für die Annotation
 - (linguistische) Tests für die Abgrenzung NN vs. NE

Inter-annotator agreement

- Problem: Annotierer können sich irren
- Methode zur Messung der Annotationsqualität:
 - Übereinstimmung zweier unabhängiger Annotatoren
 - *kappa*: Maß, das die Zufallsübereinstimmung mit berücksichtigt