

Combining heterogeneous text-technological resources for anaphora resolution

Daniela Goecke
Universität Bielefeld

CoGETI Workshop
Heidelberg, 24.11.2006



1. Projekt and Research Group
2. Application Domain: Anaphora Resolution
3. Corpus Annotation
4. Sample Annotation
5. Corpus Study
6. Use of logical document structure
7. Combining heterogeneous XML resources
8. Conclusion and Outlook



- DFG Research Group 437 „Text-technological Modelling of Information“ (2002–2008)
- Projekt A2 „Sekimo“ – Secondary Information Modelling and Combination of text-technological Resources



- DFG Research Group 437 „Text-technological Modelling of Information“ (2002–2008)
- Projekt A2 „Sekimo“ – Secondary Information Modelling and Combination of text-technological Resources
 - Abstract representation to model multi-layered XML annotations
 - Architecture for the combination of heterogeneous linguistic resources
 - Markup-Unification
 - Generation of new – richer annotated – XML documents
 - Creation of a corpus of anaphoric relations
 - Application domain: resolution of definite description anaphora



- Development of a system for the automatic resolution of anaphoric relations (decision tree based)
- Subgoals
 - Annotation of a training and evaluation corpus
 - Integration of necessary knowledge (morpho-syntactic and semantic information, anaphora-antecedent distance etc.)
 - Creation of anaphora-antecedent-candidate pairs
 - Detection of the correct antecedent



- 47 German linguistic articles (collected in the C1 project, Giessen)
- 6 German newspaper articles
- Evaluation based on a subset of 2 linguistic articles, 1 newspaper article and 1 hypertext article:
 - 4196 discourse entities
 - 1971 anaphoric relations
- XML annotated corpus
- Corpus annotation is done semi-automatically



- The annotation schema
 - Is an extension of the annotation Schema developed for the B1 project of the DFG research group (Anke Holler)
 - Defines three primary semantic relation types
 - cospecLink *The man – he , city – hanseatic city*
 - bridgingLink *The room – the window*
 - corefLink as a text-world relation
 - cospecLinks and bridgingLinks hold between discourse entities (in A2 DE of type *nominal* and *namedEntity*)
 - In the XML annotation, semantic relations are modelled using ID/IDREF



- The Annotation is done in two steps:
 1. Annotation/Detection of Discourse Entities
 2. Annotation of semantic relations
- In A2 only intra-textual relations are annotated



- For each primary relation type several secondary relation types exist
- cospecLink
 - ident, synonym, hyperonym, hyponym, paraphrase, addInfo, isA
 - a man – the man*
 - Peter – he*
 - the horse – the animal*
 - Mary Baggins – the 17 year old girl*
- bridgingLink
 - possession, meronym, holonym, setMember, hasMember, association
 - Peter – his mother*
 - a room – the window*
 - two men – the younger one*



- „Lurup is a social ghetto of the hanseatic city (Hansestadt), an outskirts with single unit houses but also many apartment blocks in the west of the city (Stadt)“

```
<cnx-pi_sentence id="w826" auto="no">
```

Lurup ist ein sozialer Brennpunkt

```
<de deID="de226" headRef="w833">
```

```
<cnx-pi_token ref="w832">der</cnx-pi_token>
```

```
<cnx-pi_token ref="w833">Hansestadt</cnx-pi_token>
```

```
</de>
```

, ein Vorort mit Einzelhäusern, aber auch vielen Wohnblocks im Westen

```
<de deID="de231" headRef="w848">
```

```
<cnx-pi_token ref="w847">der</cnx-pi_token>
```

```
<cnx-pi_token ref="w848">Stadt</cnx-pi_token>
```

```
</de>.
```

```
</cnx-pi_sentence>
```



- „Lurup is a social ghetto of the hanseatic city (Hansestadt), an outskirts with single unit houses but also many apartment blocks in the west of the city (Stadt)“

```
<cnx-pi_sentence id="w826" auto="no">
```

Lurup ist ein sozialer Brennpunkt

```
<de deID="de226" headRef="w833">
```

```
<cnx-pi_token ref="w832">der</cnx-pi_token>
```

```
<cnx-pi_token ref="w833">Hansestadt</cnx-pi_token>
```

```
</de>
```

, ein Vorort mit Einzelhäusern, aber auch vielen Wohnblocks im Westen

```
<de deID="de231" headRef="w848">
```

```
<cnx-pi_token ref="w847">der</cnx-pi_token>
```

```
<cnx-pi_token ref="w848">Stadt</cnx-pi_token>
```

```
</de>.
```

```
</cnx-pi_sentence>
```



- „Lurup is a social ghetto of the hanseatic city (Hansestadt), an outskirts with single unit houses but also many apartment blocks in the west of the city (Stadt)“

```
<cnx-pi_sentence id="w826" auto="no">
```

Lurup ist ein sozialer Brennpunkt

```
<de deID="de226" headRef="w833">
```

```
<cnx-pi_token ref="w832">der</cnx-pi_token>
```

```
<cnx-pi_token ref="w833">Hansestadt</cnx-pi_token>
```

```
</de>
```

, ein Vord

```
<de deID="de226" headRef="w833">
```

```
<cnx-pi_token_ref text="Hansestadt" dependHead="w831"  
pos="N" syntax="@NH" heur="no"
```

```
< lemma="hanse#stadt" dependValue="mod" morpho="FEM SG
```

```
< GEN" id="w833" skip="no" cnx-output="correct"/>
```

```
</de>
```

```
</cnx-pi_sentence>
```

- „Lurup is a social ghetto of the hanseatic city (Hansestadt), an outskirts with single unit houses but also many apartment blocks in the west of the city (Stadt)“

```
<cnx-pi_sentence id="w826" auto="no">
```

Lurup ist ein sozialer Brennpunkt

```
<de deID="de226" headRef="w833">  
  <cnx-pi_token ref="w832">der</cnx-pi_token>  
  <cnx-pi_token ref="w833">Hansestadt</cnx-pi_token>  
</de>
```

, ein Vorort mit Einzelhäusern, aber auch vielen Wohnblocks im Westen

```
<de deID="de231" headRef="w848">  
  <cnx-pi_token ref="w847">der</cnx-pi_token>  
  <cnx-pi_token ref="w848">Stadt</cnx-pi_token>  
</de>.
```

```
</cnx-pi_sentence>
```

```
<cospecLink relType="hyperonym" phorIDRef="de231"  
  antecedentIDRefs="de226" />
```

- Automatic discourse entity detection based on the tagger output
- Annotation of semantic relations using the tool Serengeti
 - web based client-server-application
 - enables distributed work on same corpus by user accounts
 - low system requirements on client-side
 - annotation and corpus organisation in one system
 - interface for corpus analysis (inter-annotator reliability, etc.)
 - developed in the project A2 „Sekimo“



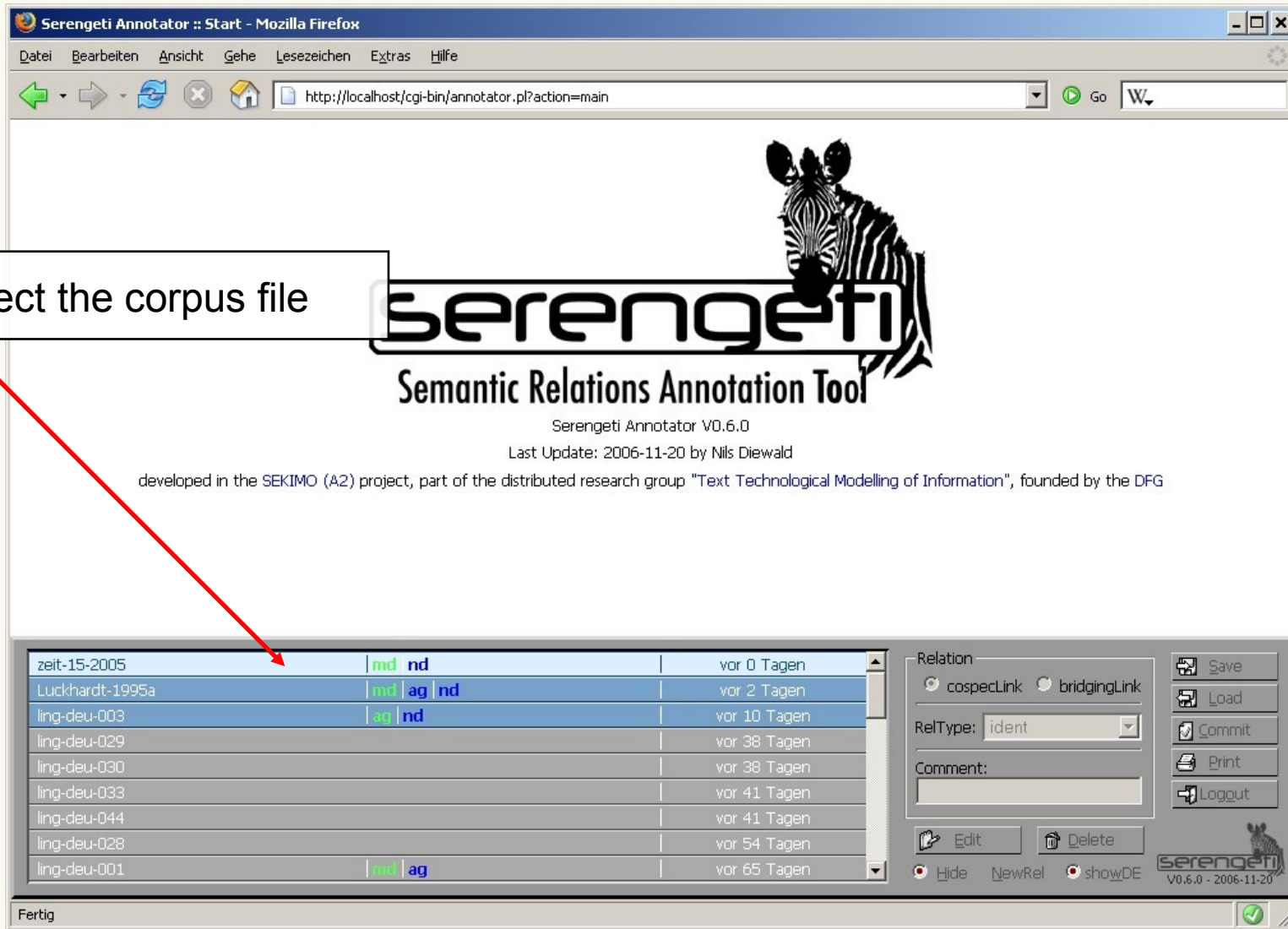
Annotation Tool



The screenshot shows the Serengeti Semantic Relations Annotation Tool interface. The browser window title is "Serengeti Annotator :: Start - Mozilla Firefox". The address bar shows the URL "http://localhost/cgi-bin/annotator.pl?action=main". The main content area features the Serengeti logo, which includes a zebra head and the text "serengeti Semantic Relations Annotation Tool". Below the logo, it states "Serengeti Annotator V0.6.0" and "Last Update: 2006-11-20 by Nils Diewald". A paragraph of text reads: "developed in the SEKIMO (A2) project, part of the distributed research group 'Text Technological Modelling of Information', founded by the DFG".

The interface includes a login section on the left with the text "Please log in" and fields for "Handle:" (containing "annotator1") and "Password:" (containing masked characters), with a "Login" button. On the right, there is a "Relation" section with radio buttons for "cospecLink" (selected) and "bridgingLink". Below this is a "RelType:" dropdown menu set to "ident" and a "Comment:" text area. A vertical column of buttons on the right includes "Save", "Load", "Commit", "Print", and "Logout". At the bottom right, there are "Edit" and "Delete" buttons, and a status bar with "Hide", "NewRel", and "showDE" options. The Serengeti logo and version information "serengeti V0.6.0 - 2006-11-20" are also present. The status bar at the bottom left of the browser window shows the word "Fertig" and a green checkmark icon.





Select the corpus file

Serengeti
Semantic Relations Annotation Tool

Serengeti Annotator V0.6.0
Last Update: 2006-11-20 by Nils Diewald
developed in the SEKIMO (A2) project, part of the distributed research group "Text Technological Modelling of Information", founded by the DFG

zeit-15-2005	md nd	vor 0 Tagen
Luckhardt-1995a	md ag nd	vor 2 Tagen
ling-deu-003	ag nd	vor 10 Tagen
ling-deu-029		vor 38 Tagen
ling-deu-030		vor 38 Tagen
ling-deu-033		vor 41 Tagen
ling-deu-044		vor 41 Tagen
ling-deu-028		vor 54 Tagen
ling-deu-001	md ag	vor 65 Tagen

Relation
 cospecLink bridgingLink

RelType: ident

Comment:

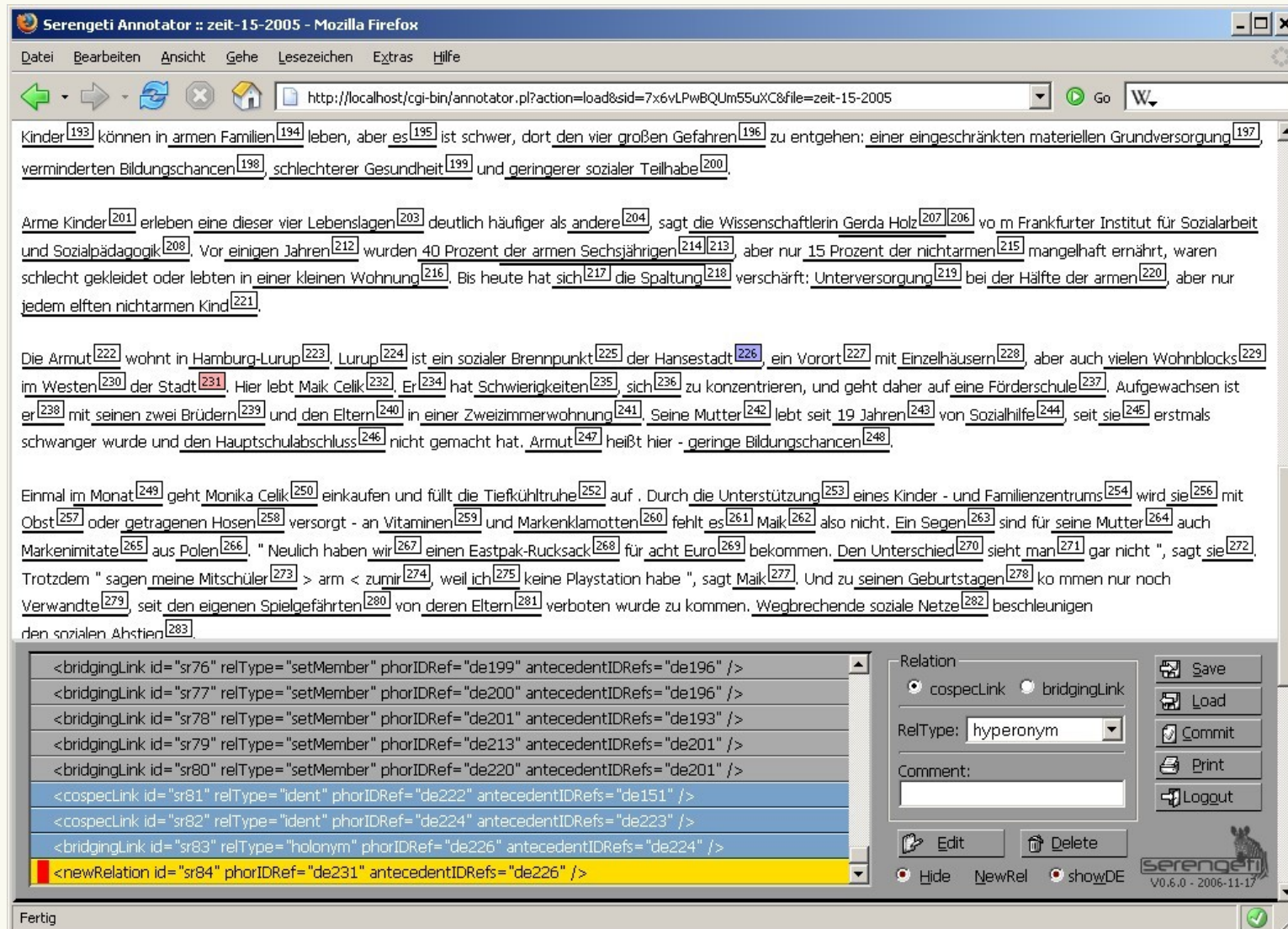
Save Load Commit Print Logout

Edit Delete

Hide NewRel showDE

Serengeti
V0.6.0 - 2006-11-20

Fertig



Serengeti Annotator :: zeit-15-2005 - Mozilla Firefox

Datei Bearbeiten Ansicht Gehe Lesezeichen Extras Hilfe

http://localhost/cgi-bin/annotator.pl?action=load&sid=7x6vLPwBQUm55uXC&file=zeit-15-2005

Kinder¹⁹³ können in armen Familien¹⁹⁴ leben, aber es¹⁹⁵ ist schwer, dort den vier großen Gefahren¹⁹⁶ zu entgehen: einer eingeschränkten materiellen Grundversorgung¹⁹⁷, verminderten Bildungschancen¹⁹⁸, schlechterer Gesundheit¹⁹⁹ und geringerer sozialer Teilhabe²⁰⁰.

Arme Kinder²⁰¹ erleben eine dieser vier Lebenslagen²⁰³ deutlich häufiger als andere²⁰⁴, sagt die Wissenschaftlerin Gerda Holz²⁰⁷²⁰⁶ vom Frankfurter Institut für Sozialarbeit und Sozialpädagogik²⁰⁸. Vor einigen Jahren²¹² wurden 40 Prozent der armen Sechsjährigen²¹⁴²¹³, aber nur 15 Prozent der nichtarmen²¹⁵ mangelhaft ernährt, waren schlecht gekleidet oder lebten in einer kleinen Wohnung²¹⁶. Bis heute hat sich²¹⁷ die Spaltung²¹⁸ verschärft: Unterversorgung²¹⁹ bei der Hälfte der armen²²⁰, aber nur jedem elften nichtarmen Kind²²¹.

Die Armut²²² wohnt in Hamburg-Lurup²²³. Lurup²²⁴ ist ein sozialer Brennpunkt²²⁵ der Hansestadt²²⁶, ein Vorort²²⁷ mit Einzelhäusern²²⁸, aber auch vielen Wohnblocks²²⁹ im Westen²³⁰ der Stadt²³¹. Hier lebt Maik Celik²³². Er²³⁴ hat Schwierigkeiten²³⁵, sich²³⁶ zu konzentrieren, und geht daher auf eine Förderschule²³⁷. Aufgewachsen ist er²³⁸ mit seinen zwei Brüdern²³⁹ und den Eltern²⁴⁰ in einer Zweizimmerwohnung²⁴¹. Seine Mutter²⁴² lebt seit 19 Jahren²⁴³ von Sozialhilfe²⁴⁴, seit sie²⁴⁵ erstmals schwanger wurde und den Hauptschulabschluss²⁴⁶ nicht gemacht hat. Armut²⁴⁷ heißt hier - geringe Bildungschancen²⁴⁸.

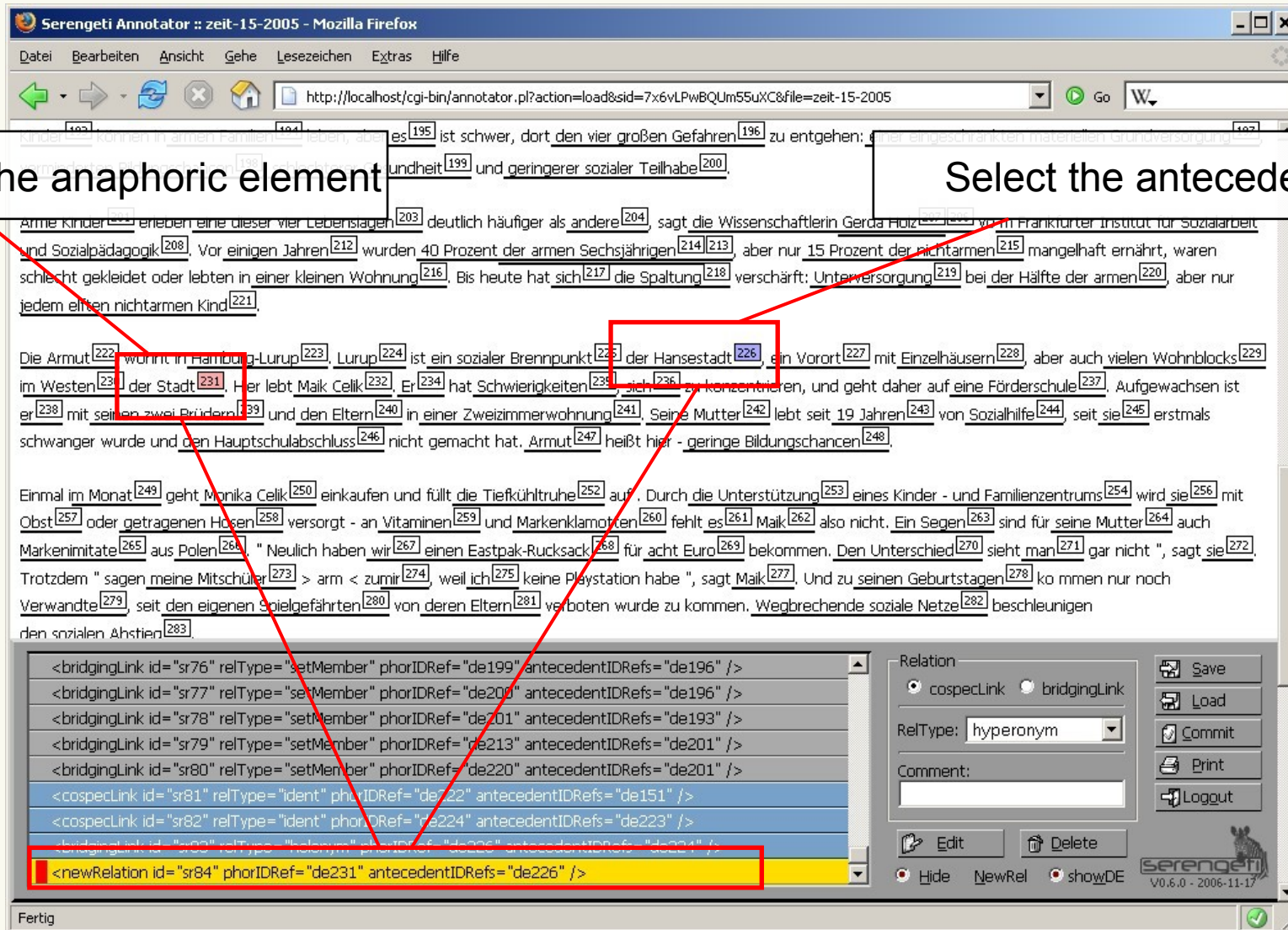
Einmal im Monat²⁴⁹ geht Monika Celik²⁵⁰ einkaufen und füllt die Tiefkühltruhe²⁵² auf. Durch die Unterstützung²⁵³ eines Kinder- und Familienzentrums²⁵⁴ wird sie²⁵⁶ mit Obst²⁵⁷ oder getragenen Hosen²⁵⁸ versorgt - an Vitaminen²⁵⁹ und Markenklamotten²⁶⁰ fehlt es²⁶¹ Maik²⁶² also nicht. Ein Segen²⁶³ sind für seine Mutter²⁶⁴ auch Markenimitate²⁶⁵ aus Polen²⁶⁶. "Neulich haben wir²⁶⁷ einen Eastpak-Rucksack²⁶⁸ für acht Euro²⁶⁹ bekommen. Den Unterschied²⁷⁰ sieht man²⁷¹ gar nicht", sagt sie²⁷². Trotzdem "sagen meine Mitschüler²⁷³ > arm < zumir²⁷⁴, weil ich²⁷⁵ keine Playstation habe", sagt Maik²⁷⁷. Und zu seinen Geburtstagen²⁷⁸ kommen nur noch Verwandte²⁷⁹, seit den eigenen Spielgefährten²⁸⁰ von deren Eltern²⁸¹ verboten wurde zu kommen. Wegbrechende soziale Netze²⁸² beschleunigen den sozialen Abstieg²⁸³.

```
<bridgingLink id="sr76" relType="setMember" phorIDRef="de199" antecedentIDRefs="de196" />
<bridgingLink id="sr77" relType="setMember" phorIDRef="de200" antecedentIDRefs="de196" />
<bridgingLink id="sr78" relType="setMember" phorIDRef="de201" antecedentIDRefs="de193" />
<bridgingLink id="sr79" relType="setMember" phorIDRef="de213" antecedentIDRefs="de201" />
<bridgingLink id="sr80" relType="setMember" phorIDRef="de220" antecedentIDRefs="de201" />
<cospecLink id="sr81" relType="ident" phorIDRef="de222" antecedentIDRefs="de151" />
<cospecLink id="sr82" relType="ident" phorIDRef="de224" antecedentIDRefs="de223" />
<bridgingLink id="sr83" relType="holonym" phorIDRef="de226" antecedentIDRefs="de224" />
<newRelation id="sr84" phorIDRef="de231" antecedentIDRefs="de226" />
```

Relation
 cospecLink bridgingLink
RelType: hyperonym
Comment:
Save Load Commit Print Logout Edit Delete Hide NewRel showDE
serengeti v0.6.0 - 2006-11-17

Fertig





The screenshot shows the Serengeti Annotator interface in Mozilla Firefox. The browser address bar shows the URL: `http://localhost/cgi-bin/annotator.pl?action=load&sid=7x6vLPwBQUm55uXC&file=zeit-15-2005`. The main text area contains a German article snippet with various words and phrases annotated with IDs in brackets. Two callout boxes are present: one on the left pointing to the word "Armut" (ID 222) with the text "Select the anaphoric element", and one on the right pointing to the word "Armut" (ID 226) with the text "Select the antecedent".

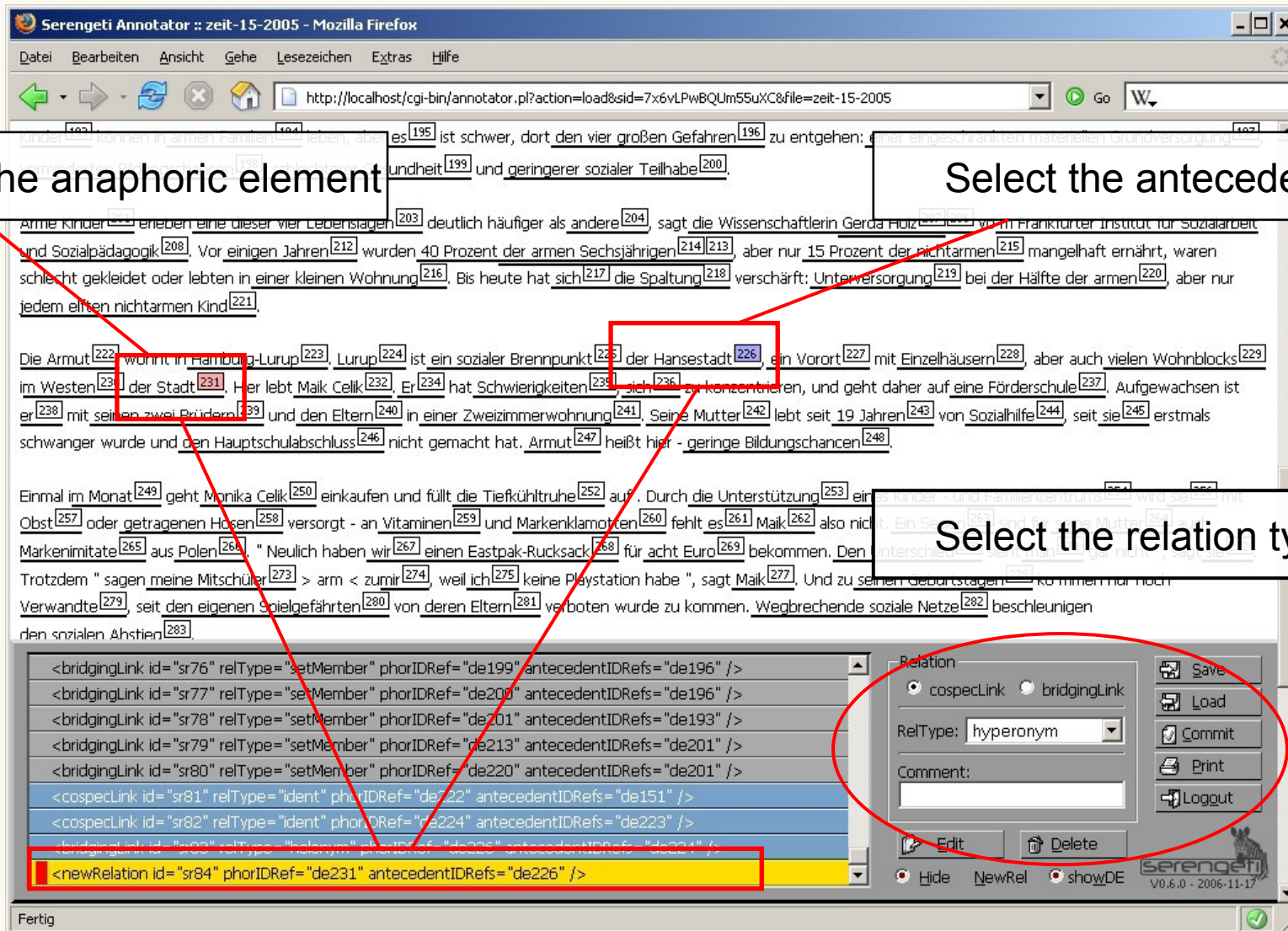
Below the text area, a list of relations is shown in a scrollable box. The last relation is highlighted in red:

```
<newRelation id="sr84" phorIDRef="de231" antecedentIDRefs="de226" />
```

To the right of the list is a control panel with the following elements:

- Relation: cospectLink bridgingLink
- RelType:
- Comment:
- Buttons: Save, Load, Commit, Print, Logout, Edit, Delete
- Options: Hide NewRel showDE
- Footer: serengeti logo, V0.6.0 - 2006-11-17





The screenshot shows the Serengeti Annotator interface in Mozilla Firefox. The browser address bar shows the URL: `http://localhost/cgi-bin/annotator.pl?action=load&sid=7x6vLPwBQUm55uXC&file=zeit-15-2005`. The main text area contains a paragraph about poverty in Hamburg-Lurup, with various words and phrases highlighted and numbered (e.g., 195, 196, 199, 200, 203, 204, 212, 214, 213, 215, 216, 217, 218, 219, 220, 221, 222, 223, 224, 225, 226, 227, 228, 229, 231, 232, 234, 235, 236, 237, 238, 239, 240, 241, 242, 243, 244, 245, 246, 247, 248, 249, 250, 252, 253, 257, 258, 259, 260, 261, 262, 265, 266, 267, 268, 269, 273, 274, 275, 277, 279, 280, 281, 282, 283).

Three callout boxes with red arrows point to specific annotations:

- Select the anaphoric element:** Points to the phrase "und geringerer sozialer Teilhabe" (ID 200).
- Select the antecedent:** Points to the phrase "Die Armut" (ID 222).
- Select the relation type:** Points to the "RelType" dropdown menu in the "Relation" panel, which is set to "hyperonym".

The bottom panel shows a list of relations. The last relation is highlighted in red:

```
<newRelation id="sr84" phorIDRef="de231" antecedentIDRefs="de226" />
```

The "Relation" panel on the right includes buttons for Save, Load, Commit, Print, Logout, Edit, and Delete, along with radio buttons for Hide, NewRel, and showDE.



anaphoric element

```
<?xml version="1.0" encoding="iso-8859-1"?>
<hypo filename="zeit-15-2005-chs.xml" max_de_distance="15">
  <semRel_candidate type="cospecLink" relType="hyperonym">
    <anaphora>
      <de deID="de231" antecedentID="de226"
        GN_lookup="1" GN_lexUnitID="nOrt.1750.Stadt" GN_propernoun="nein" GN_artificial="nein" GN_wordClass="nomen" GN_sense="1"
        text="Stadt" pos="N" syntax="@NH" lemma="stadt" dependValue="mod" morpho="FEM SG GEN"
        id="w848" sentence_pos="8" position="201">der Stadt</de>
    </anaphora>
    <antecedent de_distance="15" GN_rel="" pattern="" LSA="" score="">
      <de deID="de216"
        GN_lookup="1" GN_lexUnitID="nArtefakt.3281.Wohnung" GN_propernoun="nein" GN_artificial="nein" GN_wordClass="nomen" GN_sense="1"
        text="Wohnung" dependHead="w789" pos="N" syntax="@NH" lemma="wohnung" dependValue="adv" morpho="FEM SG DAT"
        id="w793" sentence_pos="4" position="186">einer kleinen Wohnung</de>
    </antecedent>
    [...]
    <antecedent de_distance="5" GN_rel="hyperonymy" pattern="" LSA="" score="">
      <de deID="de226" anaphora="yes"
        GN_lookup="1" GN_lexUnitID="nOrt.1768.Hansestadt" GN_propernoun="nein" GN_artificial="nein" GN_wordClass="nomen" GN_sense="1"
        text="Hansestadt" dependHead="w831" pos="N" syntax="@NH" lemma="hanse#stadt" dependValue="mod" morpho="FEM SG GEN"
        id="w833" sentence_pos="3" position="196">der Hansestadt</de>
    </antecedent>
    [...]
  </semRel_candidate>
</hypo>
```

antecedent candidate



**How to define the set of antecedent candidates?
How to select the correct antecedent?**



Resolving definite description anaphora

- How to define the set of antecedent candidates?
- Definition of a *flexible* search window

	#Occurrences	MIN-distance	MAX-distance
Identity	998	1	2262
NPform=pron	688	1	19
Paraphrase	166	1	2211
Proper Name	41	1	122
Synonymy	90	2	1699
Hyperonymy	10	2	1969
Hyponymy	4	12	99
Meronymy	17	1	50
Association	102	1	1849



- Anaphora-antecedent distance varies according to the NP type of the anaphoric element (e.g. Mitkov 2002 for an overview)
- Accessibility of antecedent candidates is dependent on the hierarchical structure of the text (e.g. Vieira & Poesio 2001)



- Anaphora-antecedent distance varies according to the NP type of the anaphoric element (e.g. Mitkov 2002 for an overview)
- Accessibility of antecedent candidates is dependent on the hierarchical structure of the text (e.g. Vieira & Poesio 2001)
- Problem: Annotation of hierarchical discourse structure is needed
- Possible solution: Use knowledge of logical document structure, e.g. DocBook, LaTeX, HTML
- Logical document structure describes the organization of a text in terms of chapters, sections, paragraphs and the like



- Influence of the logical document structure on the choice of an antecedent
- Two hypotheses
 1. Influence on the discourse entity (*antecedent life span*)
DEs in <title>-elements are more accessible than DEs in <footnote>-elements
 2. Influence on the search window for a given anaphoric element (comparable to different window sizes according to the NP form)



- Influence of the logical document structure on the choice of an antecedent
- Two hypotheses
 1. Influence on the discourse entity (*antecedent life span*) : DEs in <title>-elements are more accessible than DEs in <footnote>-elements
 2. Influence on the search window for a given anaphoric element (comparable to different window sizes according to the NP form)
- Corpus analysis to check hypothesis
 - 3490 discourse entities (anaphora:antecedent 1,36:1)
 - 1675 anaphoric expressions
 - 1234 antecedents



Distribution of Discourse Entities

	ling-deu-003	ling-deu-010	zeit-15-2005
<article>	1,35 : 1	1,3 : 1	1,6 : 1
<sect1>	1,3 : 1	1,31 : 1	1,6 : 1
<sect2>	1,3 : 1	no elements	no elements
<para>	1,35 : 1	1,35 : 1	1,6 : 1
<title>	0,84 : 1	0,25 : 1	(0)
<subtitle>	(0:2)	no elements	1,3 : 1
<listitem>	1,5 : 1	2,72 : 1	no elements
<footnote>	2,4 : 1	no elements	no elements
<glossentry>	1,4 : 1	1,15 : 1	no elements
<glossterm>	0,66 : 1	0,6 : 1	no elements
<glossdef>	1,55 : 1	1,2 : 1	no elements



Distribution of Discourse Entities

	ling-deu-003	ling-deu-010	zeit-15-2005
<article>	1,35 : 1	1,3 : 1	1,6 : 1
<sect1>	1,3 : 1	1,31 : 1	1,6 : 1
<sect2>	1,3 : 1	no elements	no elements
<para>	1,35 : 1	1,35 : 1	1,6 : 1
<title>	0,84 : 1	0,25 : 1	(0)
<subtitle>	(0:2)	no elements	1,3 : 1
<listitem>	1,5 : 1	2,72 : 1	no elements
<footnote>	2,4 : 1	no elements	no elements
<glossentry>	1,4 : 1	1,15 : 1	no elements
<glossterm>	0,66 : 1	0,6 : 1	no elements
<glossdef>	1,55 : 1	1,2 : 1	no elements



- The relationship of anaphoric and antecedent DEs is quite stable throughout a document (sect1, sect2, para)
- There are elements that tend to contain more likely antecedent elements (title, subtitle, glossterm introduce the text's topics)
- There are elements that tend to contain more likely anaphoric elements (footnote, listitem)



- A closer look at footnotes and listitems
 - For 60% of the antecedents within a listitem element, the anaphoric element is within the same element. For 66% the distance between antecedent and anaphora is smaller than 10 DEs.
 - All antecedents within a footnote element have their anaphora within the same element, for 56% of the antecedents, the anaphoric element is only one or two DEs away (often pronouns), less than 1% are more than 10 DEs away
- The parent element of a DE serves as a clue for the antecedent life span.



How to integrate the document structure? → Markup Unification



Logical document structure vs. Semantic relations

```
layer doc:<para>... im Westen der Stad ..</para>
```

```
layer de :<analysis>... im Westen <de>der Stadt</de>...<analysis>
```

#char	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18					
	i	m		W	e	s	t	e	n		d	e	r		S	t	a	d	t					
	<para>	i	m		W	e	s	t	e	n		d	e	r		S	t	a	d	t	</para>			
	<analysis>											<de>	d	e	r		S	t	a	d	t	</de>	...	</...>

```
node( doc,          0, 18, [1], element(para) ).  
node( analysis,    0, 18, [1], element(analysis) ).  
node( analysis,    10,18, [1,1], element(de) ).
```

Markup Unification:

```
<analysis><para>... im Westen </de>der Stadt</de>...</para></an...>
```

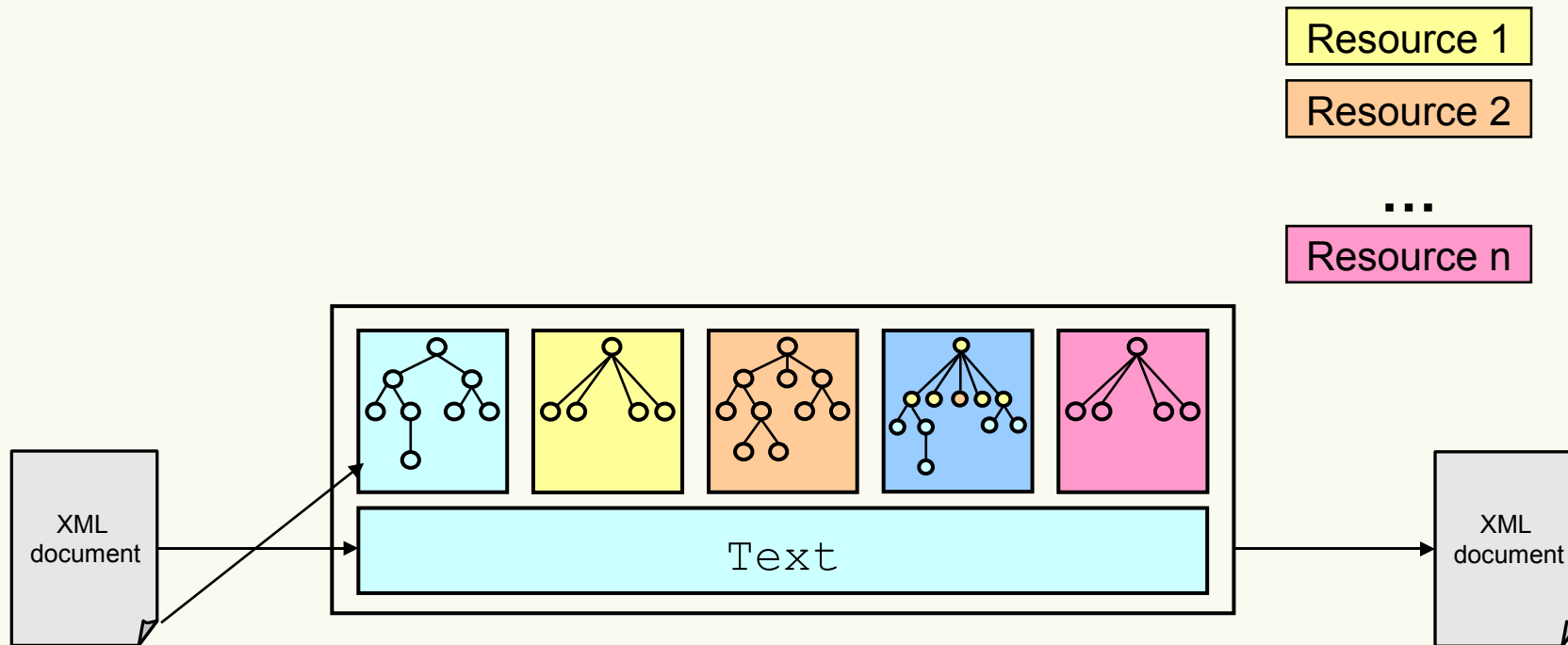


Benefit of Markup Unification


- Add new informationen from heterogeneous resources
- Resources can be developed independently
- Flexible use of resources: No conversion necessary
- Extraction of relevant parts: Application of resources only for relevant text parts
- Reuse of existing resources



Integration of linguistic resources

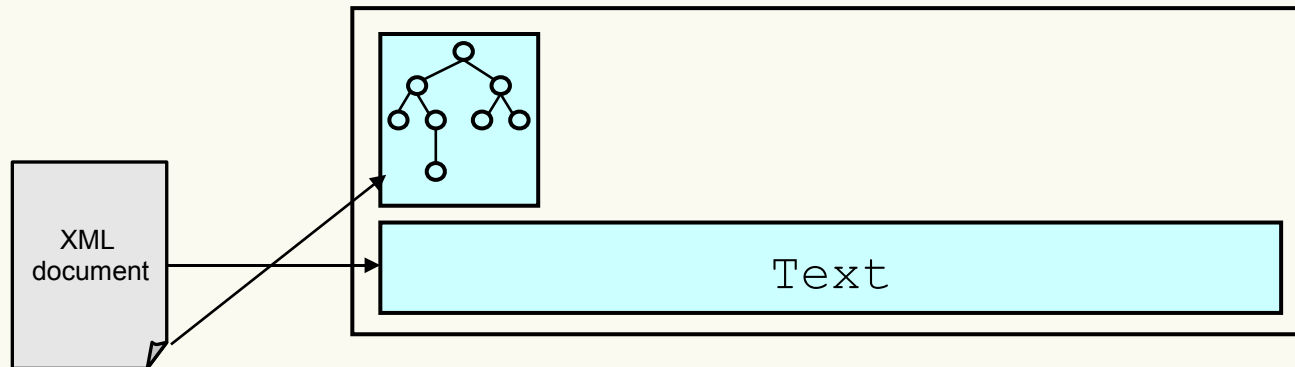


Integration of linguistic resources

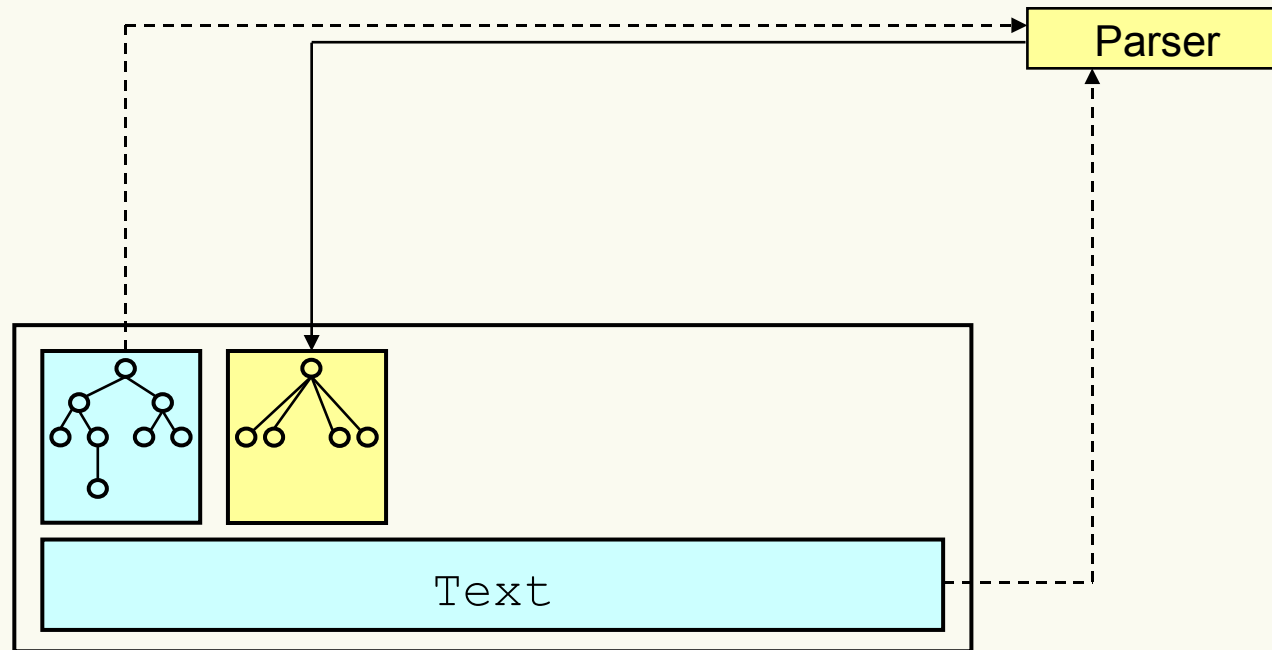


XML
document

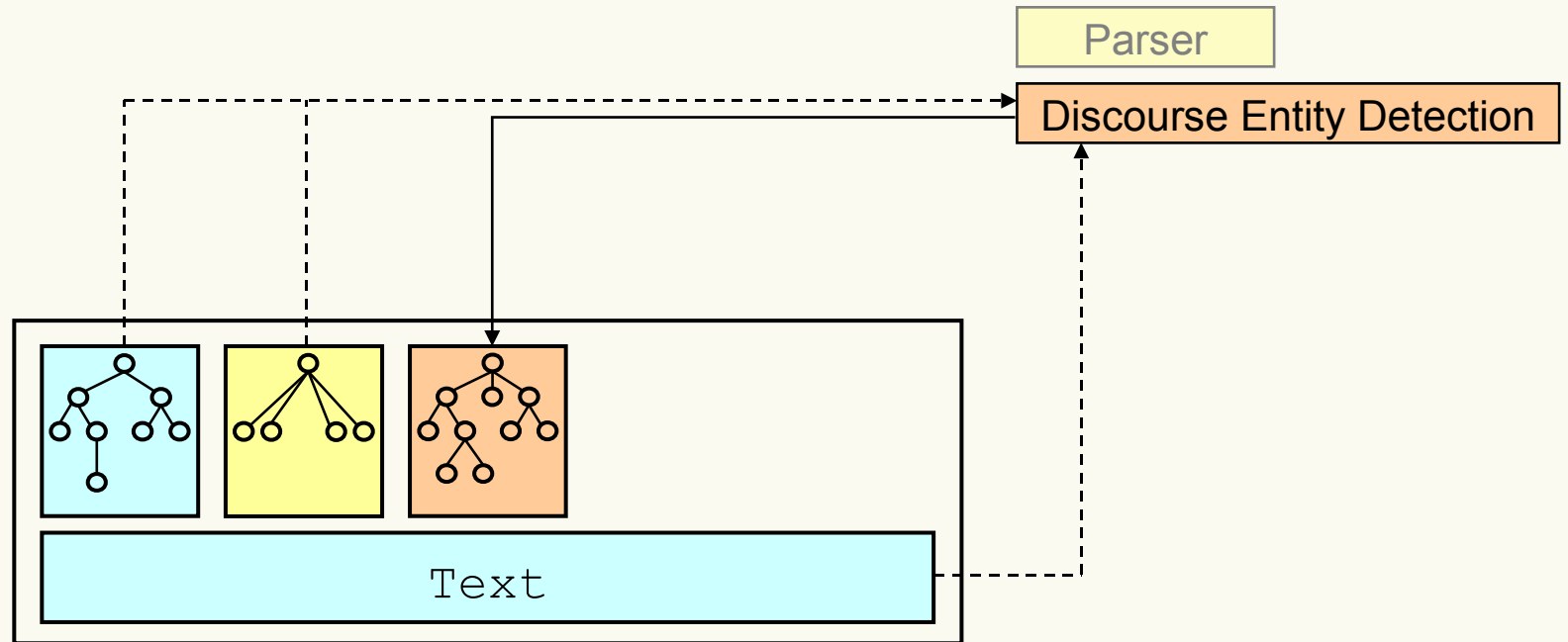




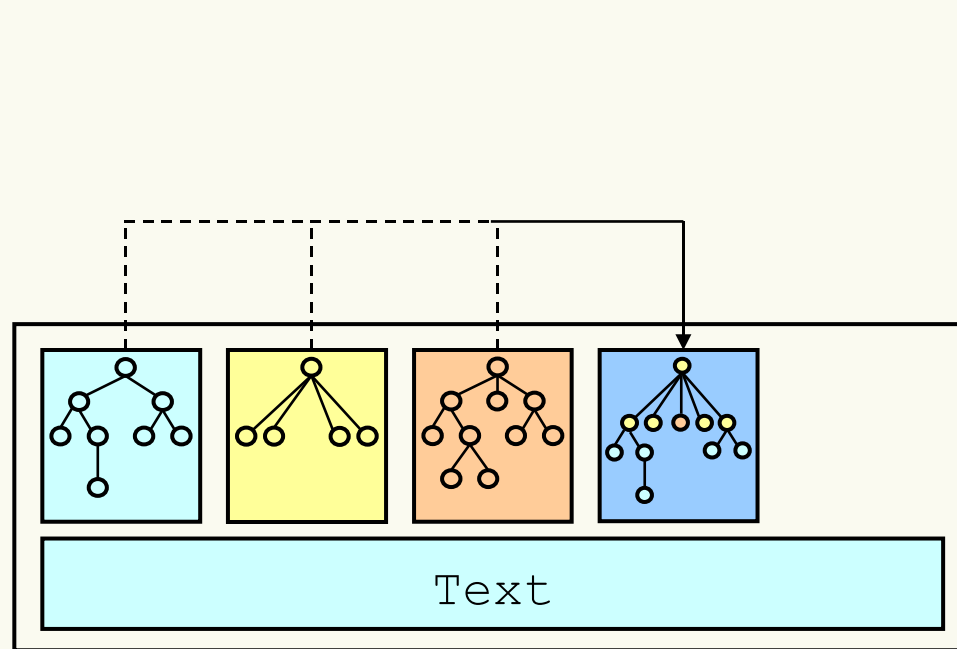
Integration of linguistic resources



Integration of linguistic resources



Integration of linguistic resources



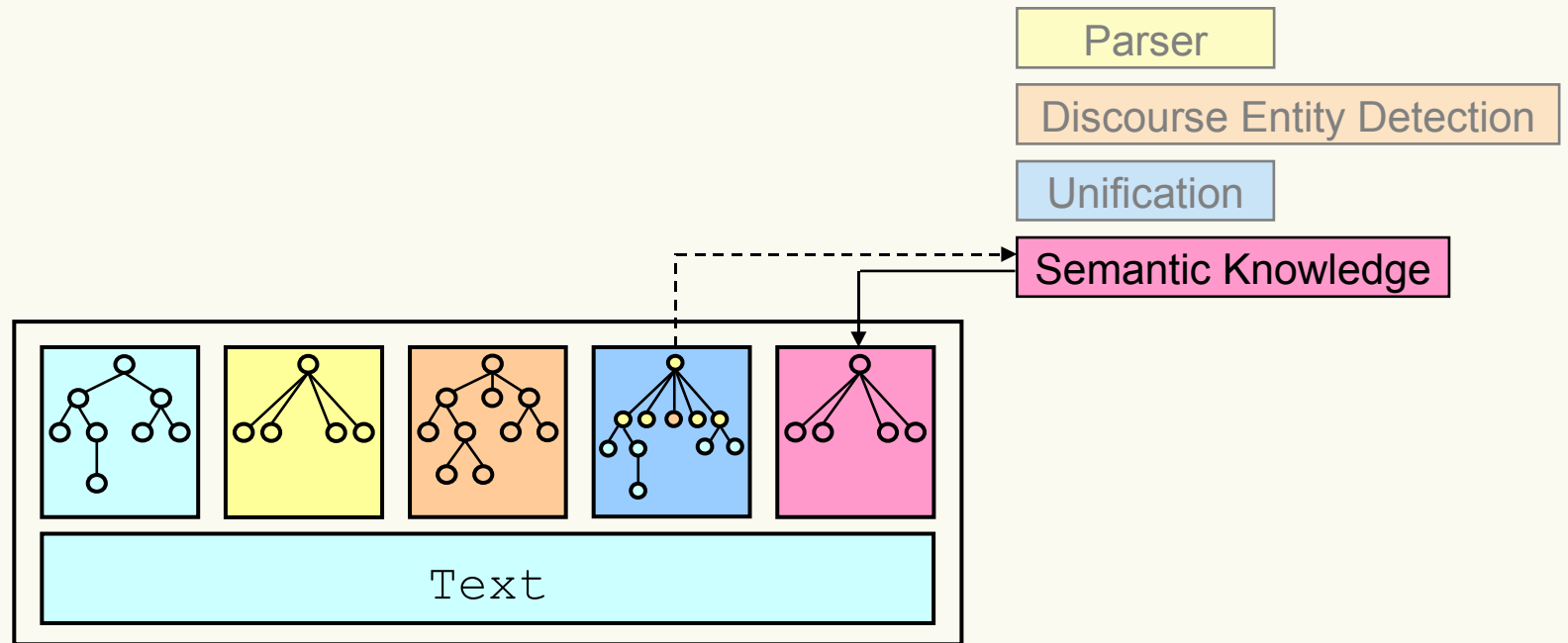
Parser

Discourse Entity Detection

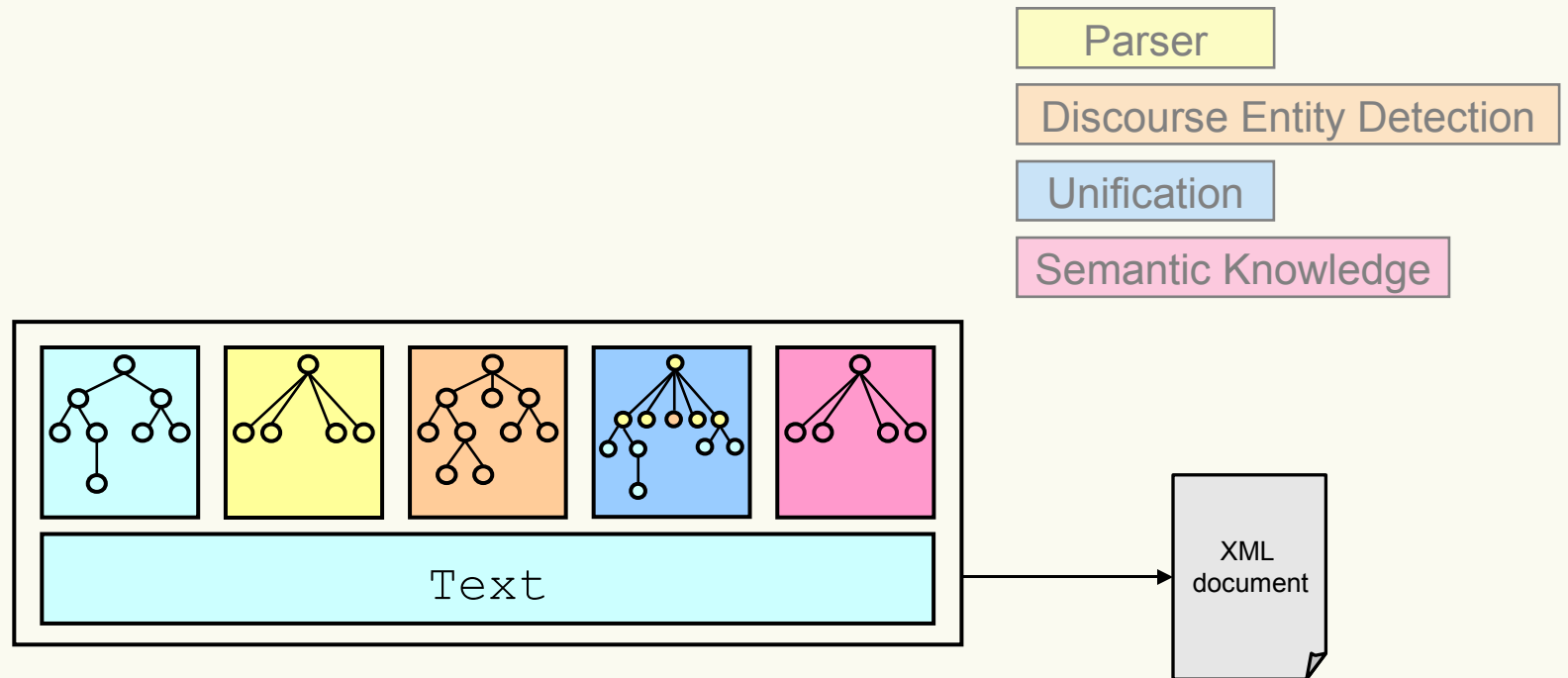
Unification

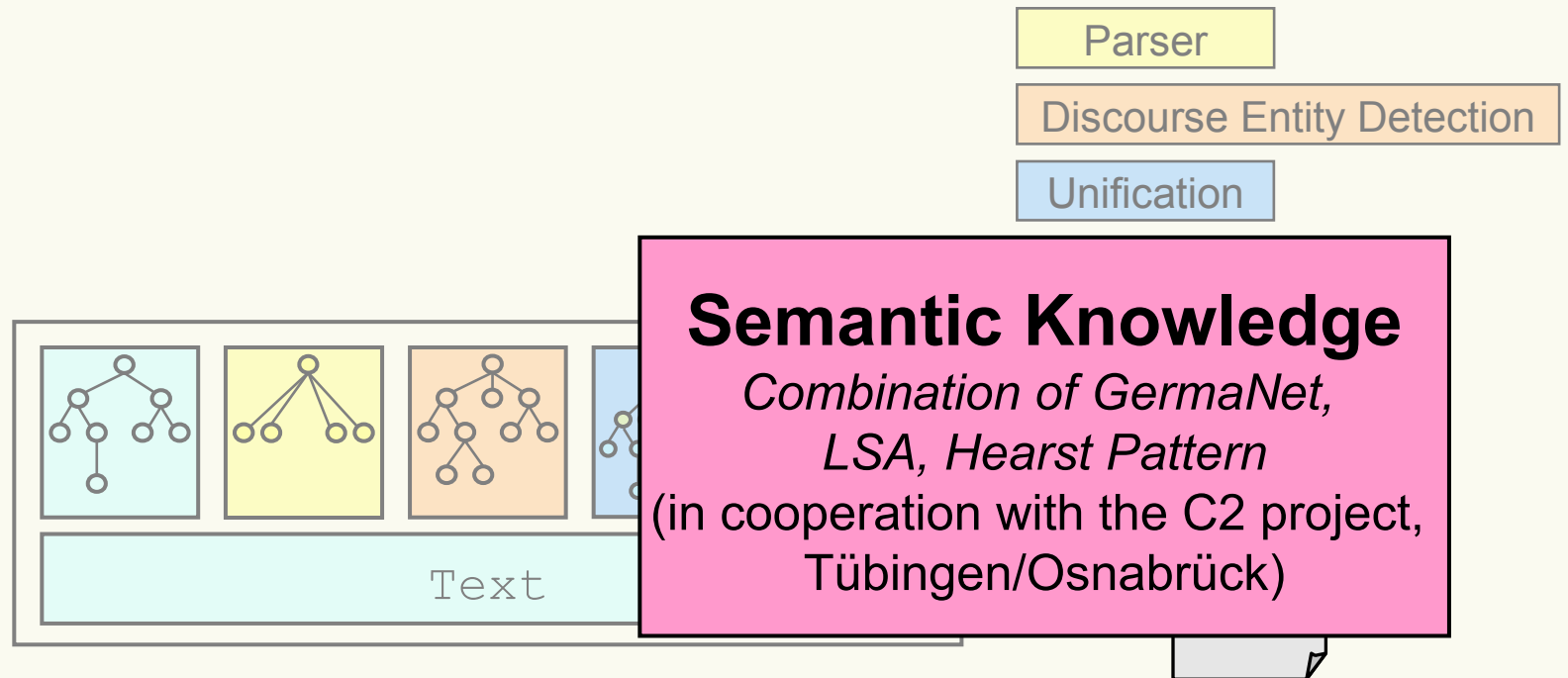


Integration of linguistic resources



Integration of linguistic resources





- Extend the corpus analysis
- Translate corpus findings into suitable features and XML attributes
- Translate XML annotations into feature vectors
- Train decision trees to resolve definite description anaphora automatically



Thank you!

Daniela Goecke : daniela.goecke@uni-bielefeld.de

