Combining heterogeneous text-technological resources for anaphora resolution

Daniela Goecke
Universität Bielefeld

CoGETI Workshop
Heidelberg, 24.11.2006

Overview

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

Projekt and Research Group

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

Projekt and Research Group

- 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

The application domain

- 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

The Corpus

- 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

Annotation Schema

- 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

Annotation Schema

- 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

Annotation Schema

- 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

Sample Annotation

- „Lurup is a social ghetto of the hanseatic city (Hansestadt), an outskirt with single unit houses but also many appartment 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>
```

Sample Annotation

- „Lurup is a social ghetto of the hanseatic city (Hansestadt), an outskirt with single unit houses but also many appartment 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>
```

Sample Annotation

- „Lurup is a social ghetto of the hanseatic city (Hansestadt), an outskirt with single unit houses but also many appartment 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 Vorc
    <de deI
        <c
        <c
    </de>.
</cnx-pi_sentence>
```

```
<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"/>
```

Sample Annotation

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

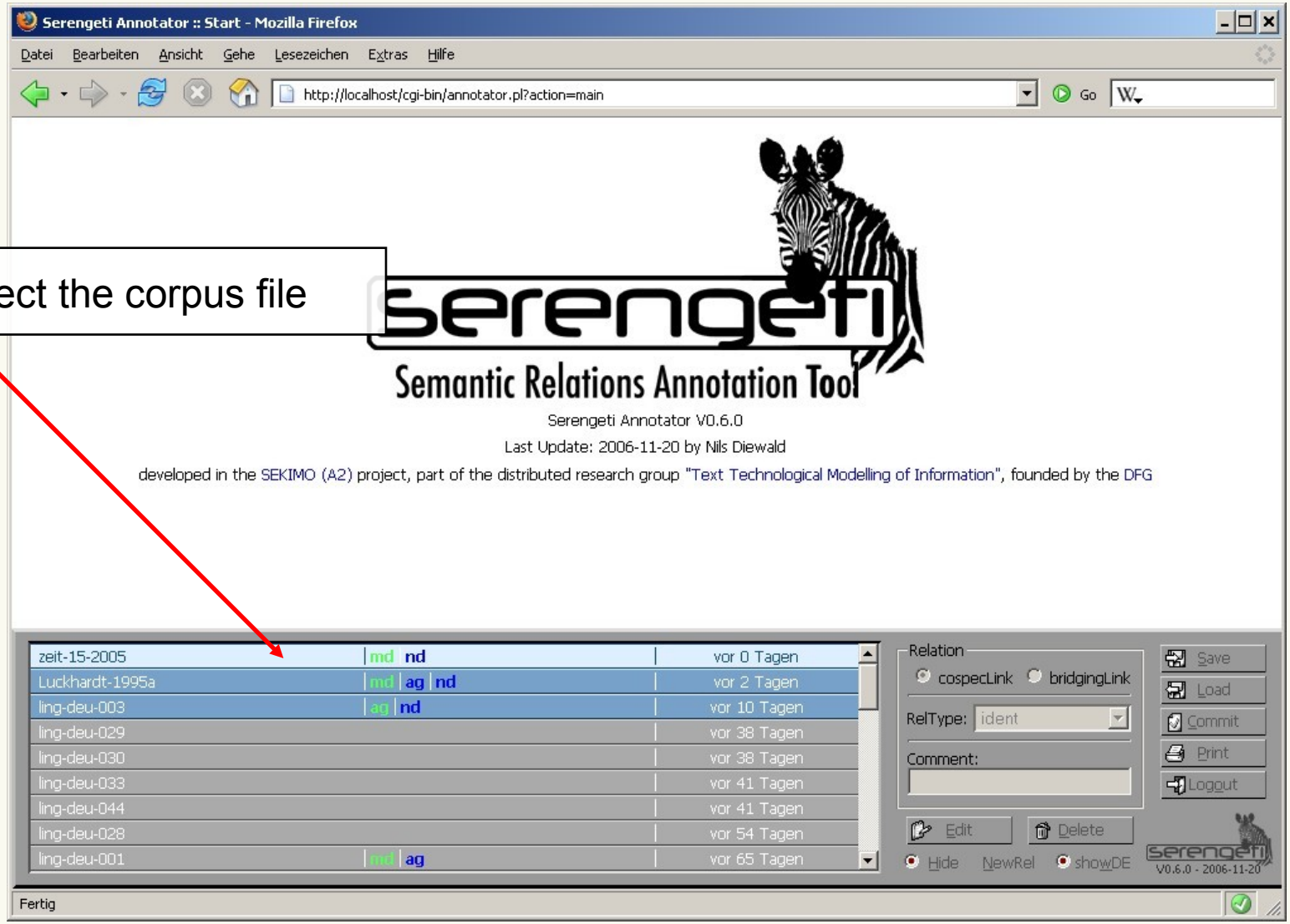
```
<cnx-pi_sentence id="w826" auto="no">
    Lurup ist ein sozialer Brennpunkt
    <de deID="de226" headRef="w833">
        <cnx-pi_token ref="w832">der</cnx-pi_token>
        <cnx-pi_token ref="w833">Hansestadt</cnx-pi_token>
    </de>
    , ein Vorort mit Einzelhäusern, aber auch vielen Wohnblocks 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" />
```

Corpus Annotation

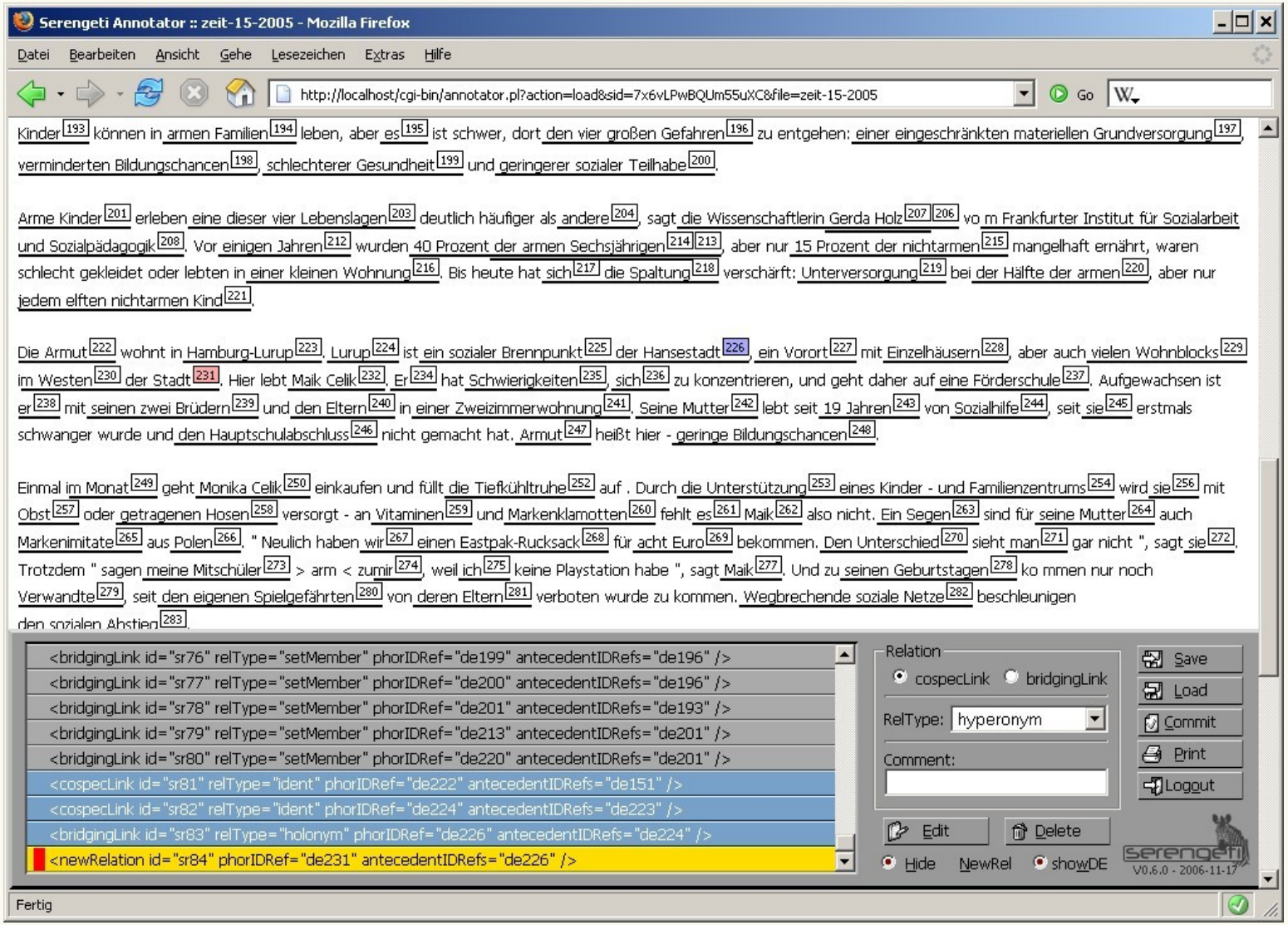
- Automatic discourse entity detection based on the tagger output
- Annotation of semantic relations using the tool Serengeti
  - web based client-server-application
  - enables distributed work on same corpus by user accounts
  - low system requirements on client-side
  - annotation and corpus organisation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

Annotation Tool

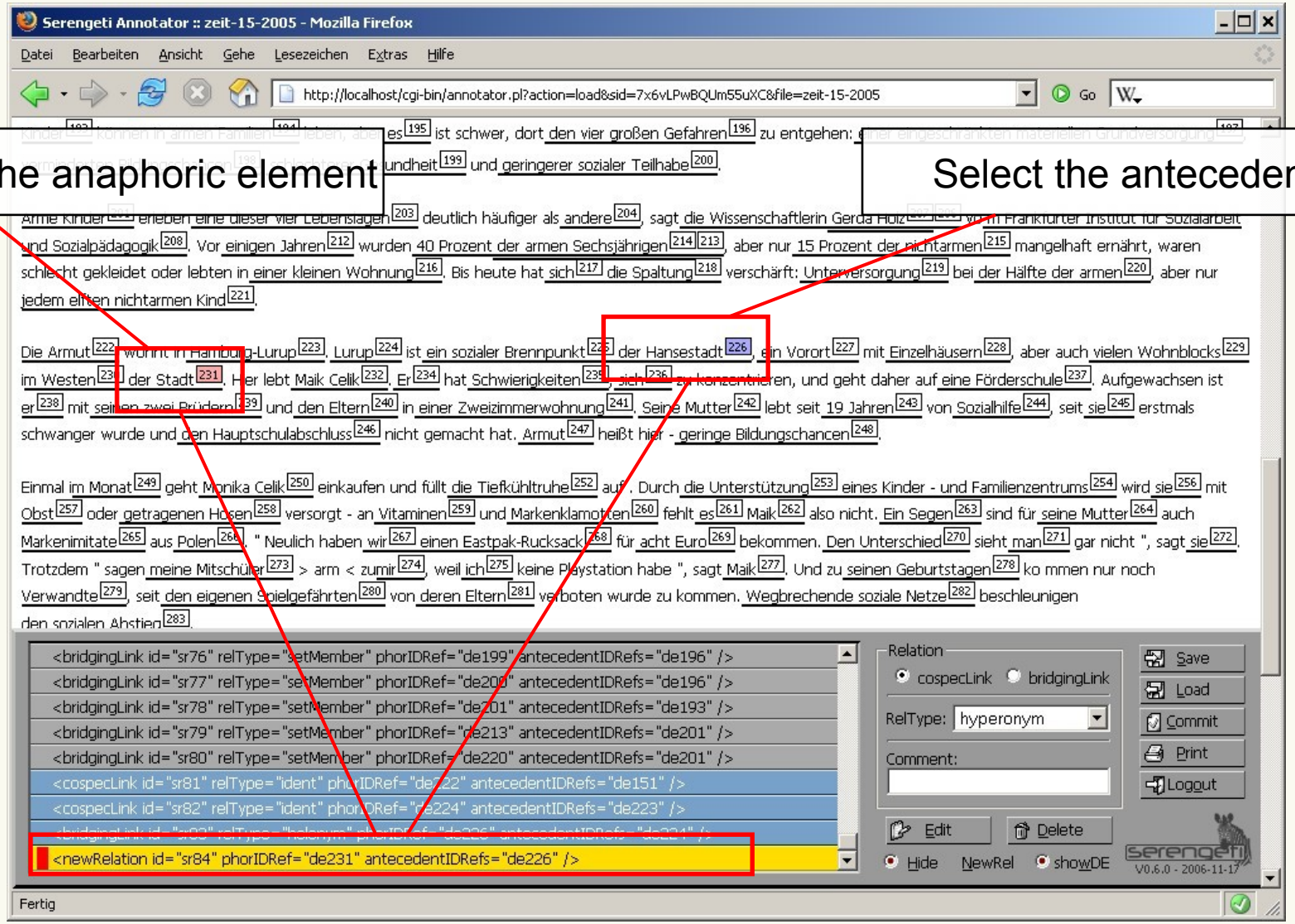
Annotation Tool



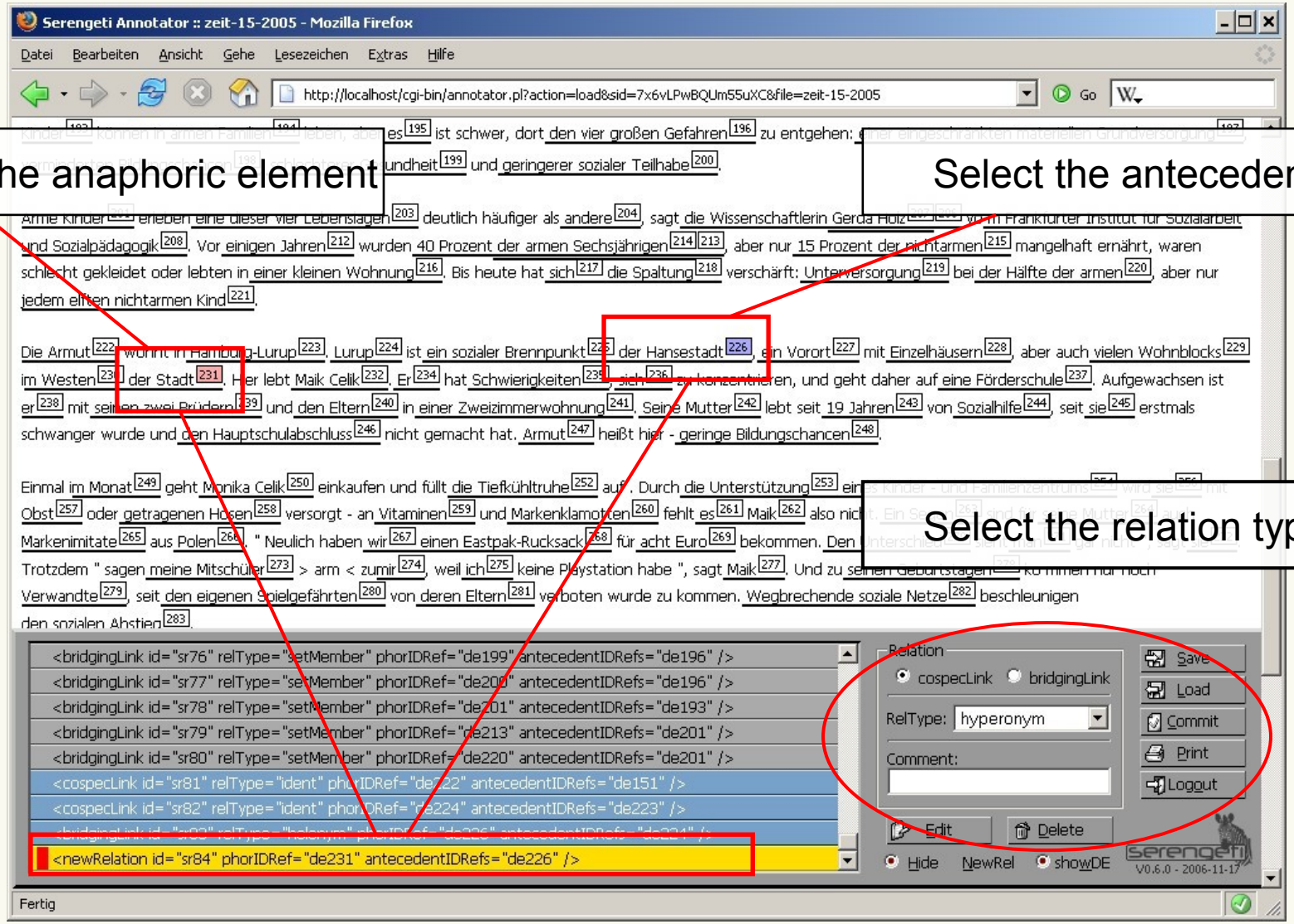
Annotation Tool



Annotation Tool



Annotation Tool



Training and Evaluation Set

Text Technological Modelling of Information

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="advl" 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

Issues

Text Technological Modelling of Information

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

	#Occurences	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

Definition of a flexible search window

- 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)

Definition of a flexible search window

- 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

Logical Document Structure

- 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)

Logical Document Structure

- 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

Interpretation of results

- 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)

Interpretation of results

- 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

Combination of resources

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…>
```

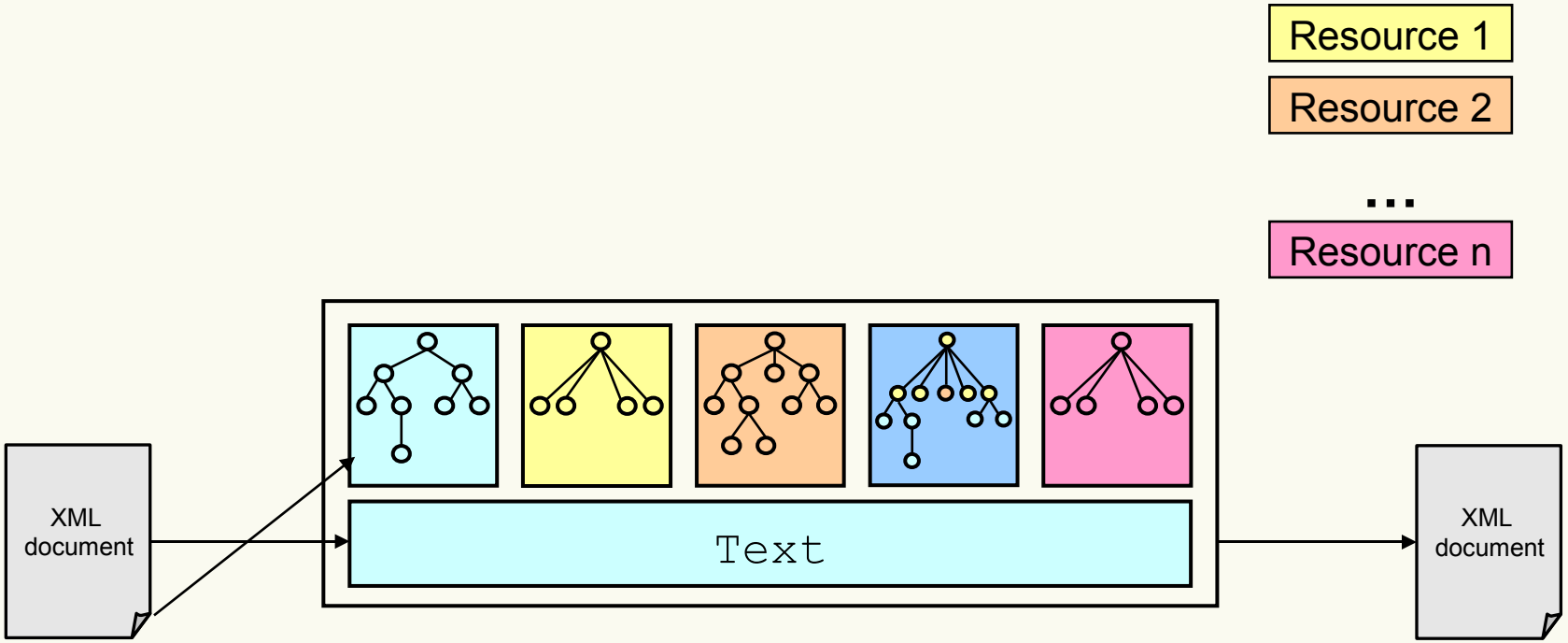
Combination of resources

Text Technological Modelling of Information

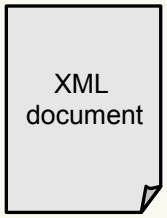
Benefit of Markup Unification

- 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

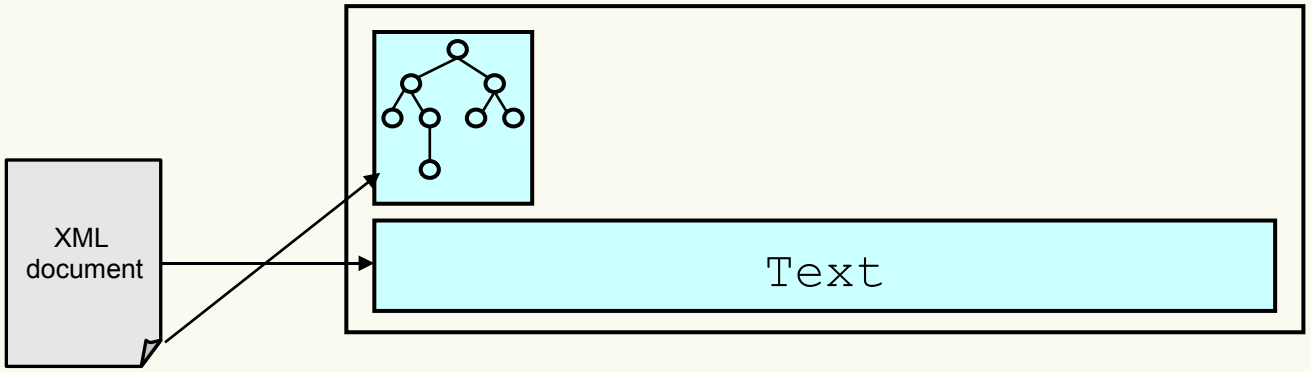
Integration of linguistic resources



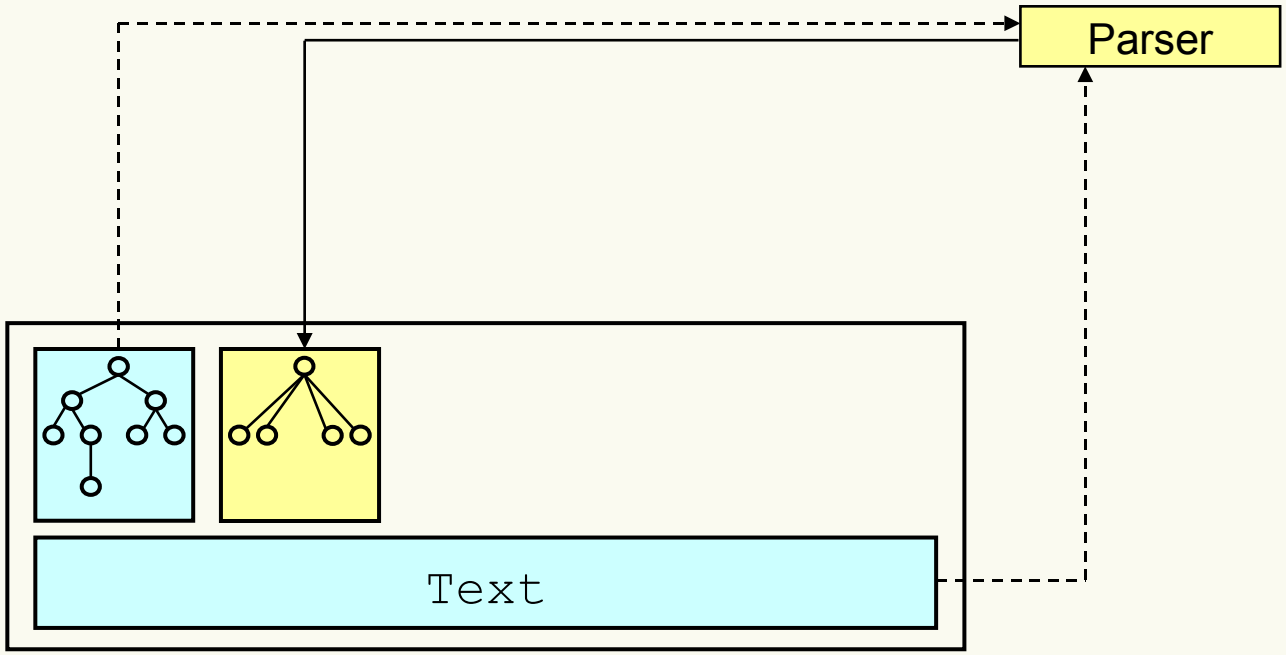
Integration of linguistic resources



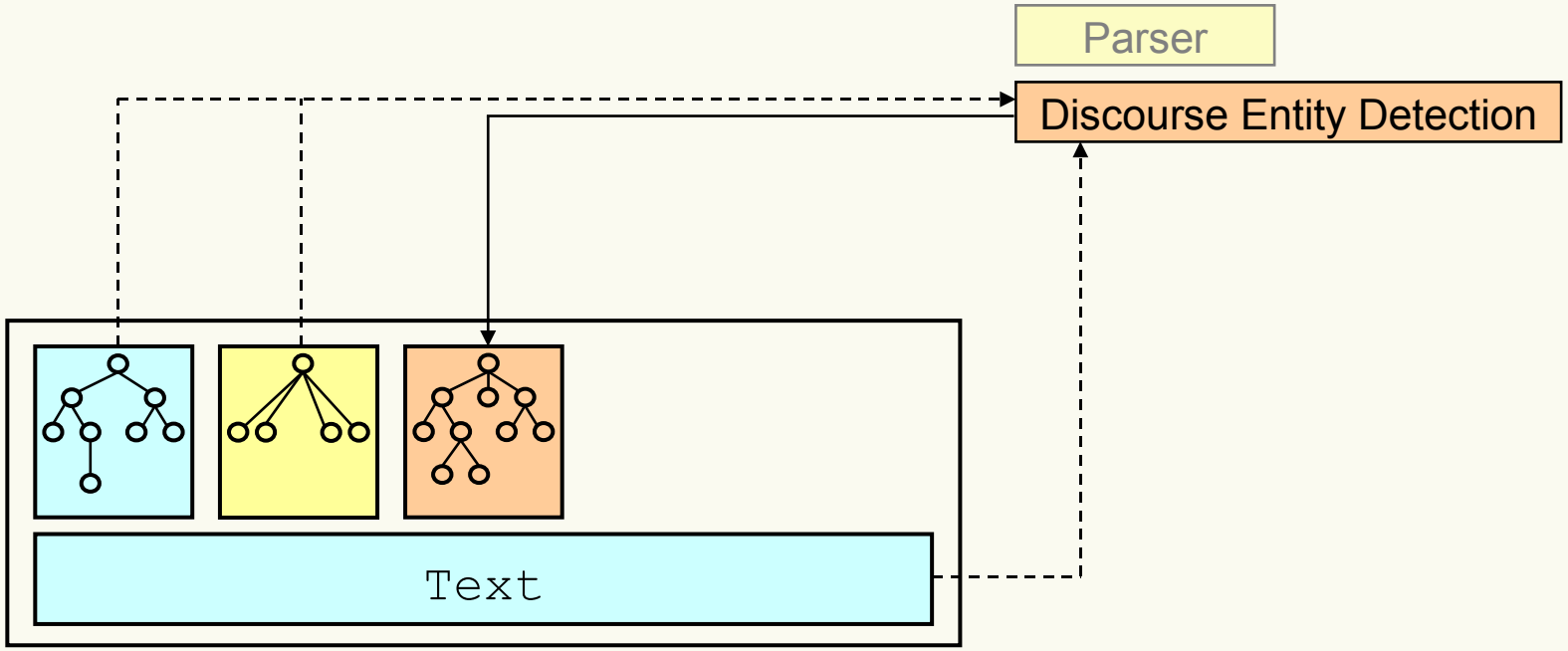
Integration of linguistic resources



Integration of linguistic resources



Integration of linguistic resources

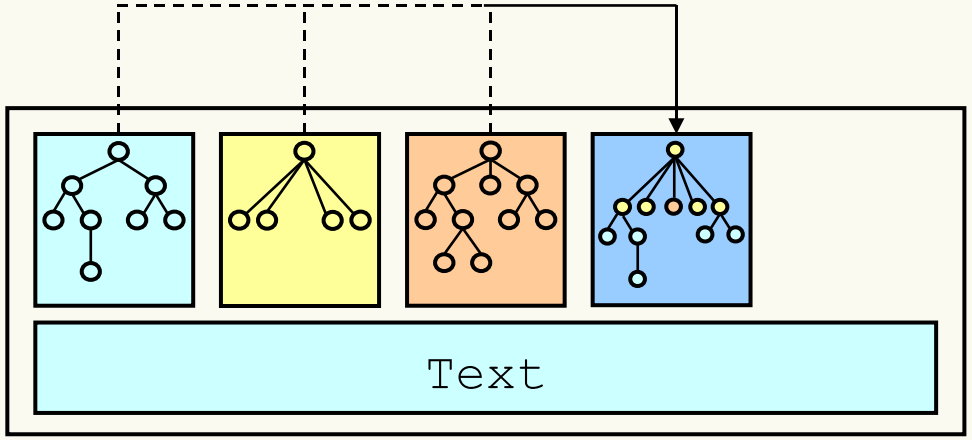


Integration of linguistic resources

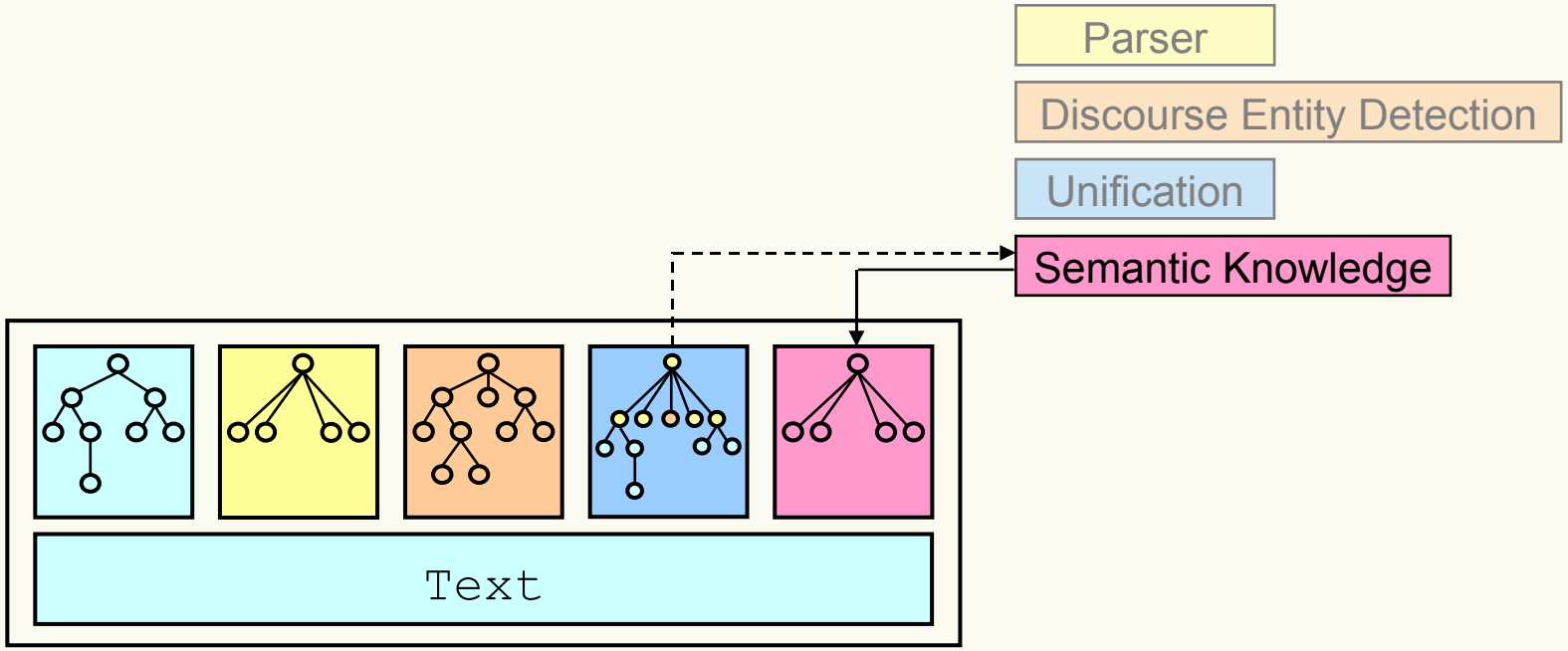
Parser

Discourse Entity Detection

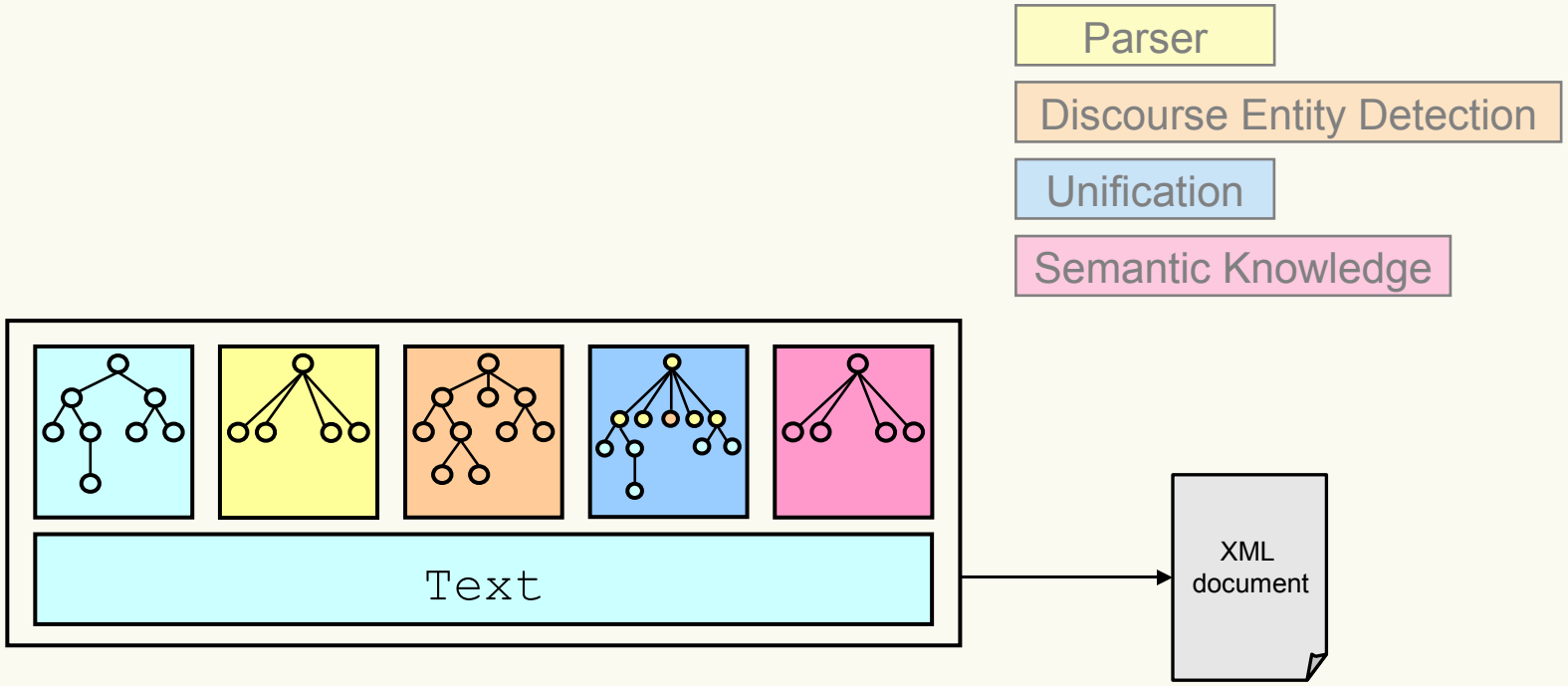
Unification



Integration of linguistic resources



Integration of linguistic resources

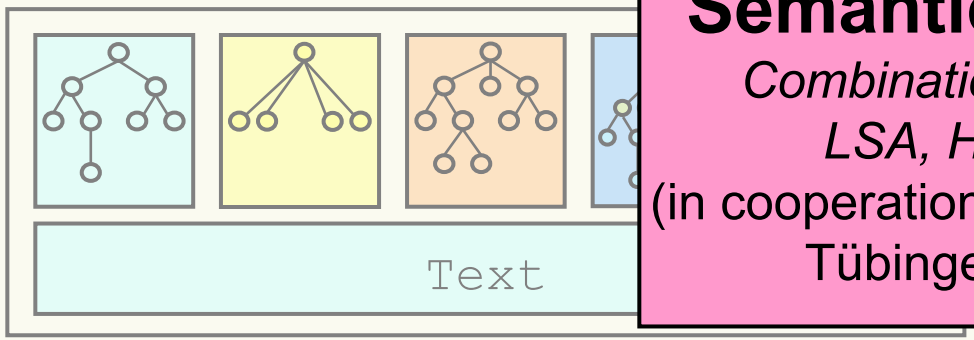


Integration of linguistic resources

Parser

Discourse Entity Detection

Unification



Semantic Knowledge

Combination of GermaNet, LSA, Hearst Pattern

(in cooperation with the C2 project, Tübingen/Osnabrück)

Conclusion and outlook

- 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