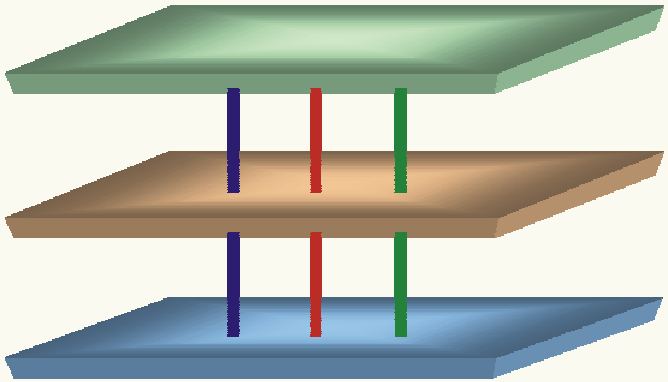
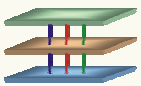
JUSTUS-LIEBIG-
UNIVERSITÄT
GIESSEN

Text parsing of a complex genre



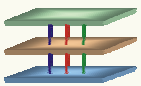
Harald Lüngen, Maja Bärenfänger, Mirco Hilbert, Henning Lobin, Csilla Puskás

Applied and Computational Linguistics
Justus Liebig University Giessen



Overview

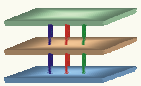
1. Introduction
2. Linguistic foundations
3. Processing
4. Status and Outlook



Introduction

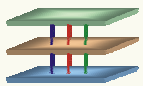
JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

- Project SemDok: *Generic document structures in linearly organised texts (*2005-2008)
- Part of DFG-Research Group *Text-technological modelling of information*
- Goals:
  - Develop a text (discourse) parser for a complex genre
  - Complex genre: Scientific articles, i.e. take into account
    - logical document structure
    - genre-specific text type structure
    - thematic structure
  - Use text-technological (XML-based) formalisms and methods



Overview

1. Introduction
2. Linguistic foundations
3. Processing
4. Status and Outlook



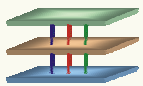
Current theories of discourse

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

ULDM	Unified Linguistic Discourse Model	Polanyi et al. 2004a,b
SDRT	Segmented Discourse Representation Theory	Asher & Lascarides 2003, Asher & Vieu 2005
RST	Rhetorical Structure Theory	Mann & Thompson 1988, Marcu 2000

Common assumptions:

1. Discourse coherence relations hold between adjacent *discourse segments.* Discourse Segments can be *elementary* or *complex*
2. Discourse is structured hierarchically
   - SDRT: graph-structured
   - RST, ULDM: tree-structured
3. Two major structural types (schemas) of discourse relations:
   - hypotactic (mononuclear, subordinating)
   - paratactic (multinuclear, coordinating)



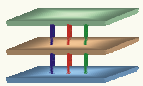
Current theories of discourse

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

ULDM	Unified Linguistic Discourse Model	Polanyi et al. 2004a,b
SDRT	Segmented Discourse Representation Theory	Asher & Lascarides 2003, Asher & Vieu 2005
RST	Rhetorical Structure Theory	Mann & Thompson 1988, Marcu 2000

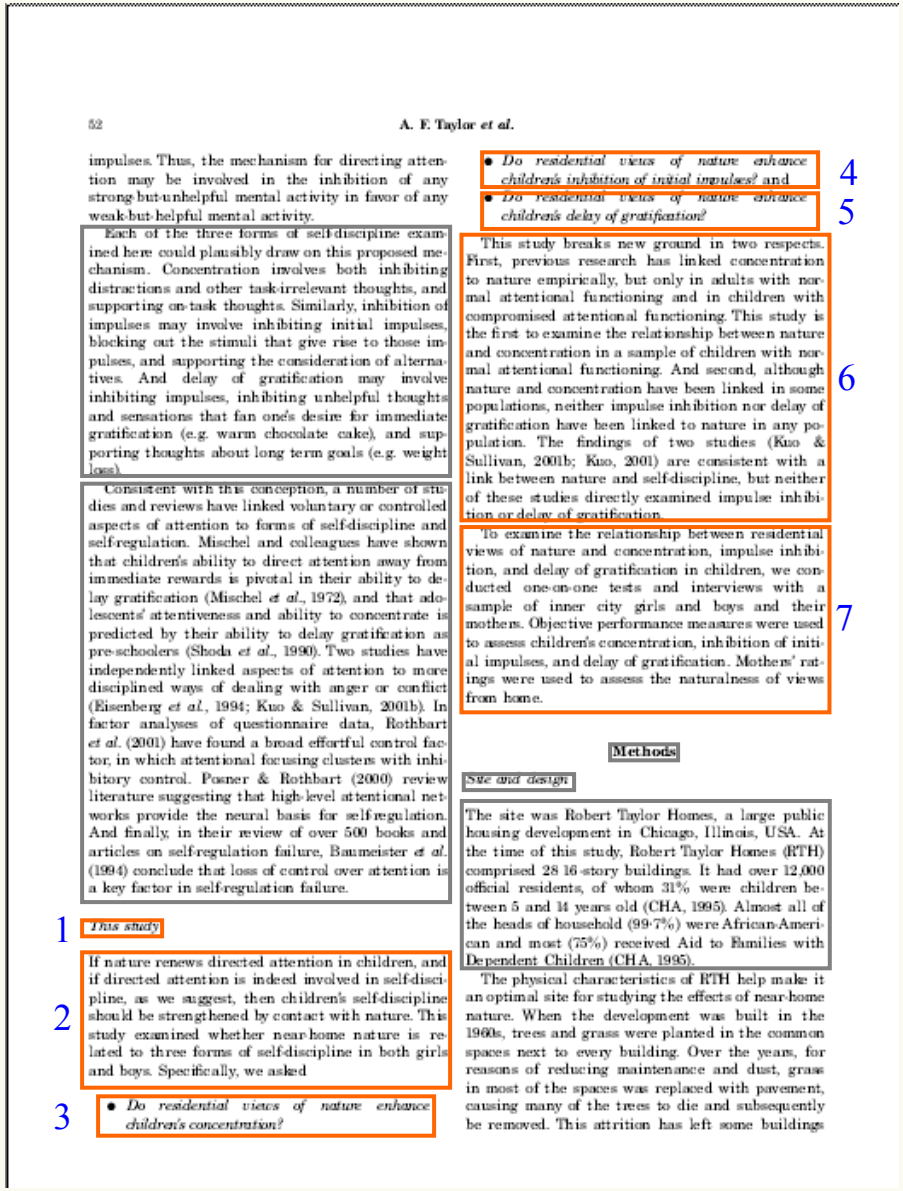
Common assumptions:

1. Discourse coherence relations hold between adjacent *discourse segments.*
   Discourse Segments can be *elementary* or *complex*
2. Discourse is structured hierarchically
   - SDRT: graph-structured
   - RST, ULDM: tree-structured
3. Two major structural types (schemas) of discourse relations:
   - hypotactic (mononuclear, subordinating)
   - paratactic (multinuclear, coordinating)



Elementary and complex segments

[This research considers potential predictors of workaholism and workaholic behaviors.]$_1$ [What are the wellsprings of workaholism?]$_2$ [Speculation and research findings suggest that individual difference characteristics and organizational factors likely serve as antecedents.]$_3$ [Individual characteristics include demographic factors]$_4$ [(e.g., gender),]$_5$ [family of origin dynamics]$_6$ [(Robinson, 1998),]$_7$ [and aspects of personality]$_8$ [(e.g., obsession-compulsion; Schwartz, 1982).]$_9$ [Organizational factors include values supporting work-personal life imbalance]$_{10}$ [(Schaef & Fassel, 1988; Killinger, 1991).]$_{11}$



52 A. F. Taylor *et al.*

impulses. Thus, the mechanism for directing attention may be involved in the inhibition of any strong-but-unhelpful mental activity in favor of any weak-but-helpful mental activity.

Each of the three forms of self-discipline examined here could plausibly draw on this proposed mechanism. Concentration involves both inhibiting distractions and other task-irrelevant thoughts, and supporting on-task thoughts. Similarly, inhibition of impulses may involve inhibiting initial impulses, blocking out the stimuli that give rise to those impulses, and supporting the consideration of alternatives. And delay of gratification may involve inhibiting impulses, inhibiting unhelpful thoughts and sensations that fan one's desire for immediate gratification (e.g. warm chocolate cake), and supporting thoughts about long term goals (e.g. weight loss).

Consistent with this conception, a number of studies and reviews have linked voluntary or controlled aspects of attention to forms of self-discipline and self-regulation. Mischel and colleagues have shown that children's ability to direct attention away from immediate rewards is pivotal in their ability to delay gratification (Mischel *et al.*, 1972), and that adolescents' attentiveness and ability to concentrate is predicted by their ability to delay gratification as pre-schoolers (Shoda *et al.*, 1990). Two studies have independently linked aspects of attention to more disciplined ways of dealing with anger or conflict (Eisenberg *et al.*, 1994; Kuo & Sullivan, 2001b). In factor analyses of questionnaire data, Rothbart *et al.* (2001) have found a broad effortful control factor, in which attentional focusing clusters with inhibitory control. Posner & Rothbart (2000) review literature suggesting that high-level attentional networks provide the neural basis for self-regulation. And finally, in their review of over 500 books and articles on self-regulation failure, Baumeister *et al.* (1994) conclude that loss of control over attention is a key factor in self-regulation failure.

1 *This study*

2 If nature renews directed attention in children, and if directed attention is indeed involved in self-discipline, as we suggest, then children's self-discipline should be strengthened by contact with nature. This study examined whether near-home nature is related to three forms of self-discipline in both girls and boys. Specifically, we asked

3 • *Do residential views of nature enhance children's concentration?*

• *Do residential views of nature enhance children's inhibition of initial impulses?* and 4

• *Do residential views of nature enhance children's delay of gratification?* 5

This study breaks new ground in two respects. First, previous research has linked concentration to nature empirically, but only in adults with normal attentional functioning and in children with compromised attentional functioning. This study is the first to examine the relationship between nature and concentration in a sample of children with normal attentional functioning. And second, although nature and concentration have been linked in some populations, neither impulse inhibition nor delay of gratification have been linked to nature in any population. The findings of two studies (Kuo & Sullivan, 2001b; Kuo, 2001) are consistent with a link between nature and self-discipline, but neither of these studies directly examined impulse inhibition or delay of gratification. 6

To examine the relationship between residential views of nature and concentration, impulse inhibition, and delay of gratification in children, we conducted one-on-one tests and interviews with a sample of inner city girls and boys and their mothers. Objective performance measures were used to assess children's concentration, inhibition of initial impulses, and delay of gratification. Mothers' ratings were used to assess the naturalness of views from home. 7

Methods

Site and design

The site was Robert Taylor Homes, a large public housing development in Chicago, Illinois, USA. At the time of this study, Robert Taylor Homes (RTH) comprised 28 16-story buildings. It had over 12,000 official residents, of whom 31% were children between 5 and 14 years old (CHA, 1995). Almost all of the heads of household (99·7%) were African-American and most (75%) received Aid to Families with Dependent Children (CHA, 1995).

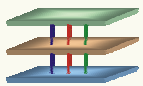
The physical characteristics of RTH help make it an optimal site for studying the effects of near-home nature. When the development was built in the 1960s, trees and grass were planted in the common spaces next to every building. Over the years, for reasons of reducing maintenance and dust, grass in most of the spaces was replaced with pavement, causing many of the trees to die and subsequently be removed. This attrition has left some buildings

Current theories of discourse

ULDM	Unified Linguistic Discourse Model	Polanyi et al. 2004a,b
SDRT	Segmented Discourse Representation Theory	Asher & Lascarides 2003, Asher & Vieu 2005
RST	Rhetorical Structure Theory	Mann & Thompson 1988, Marcu 2000

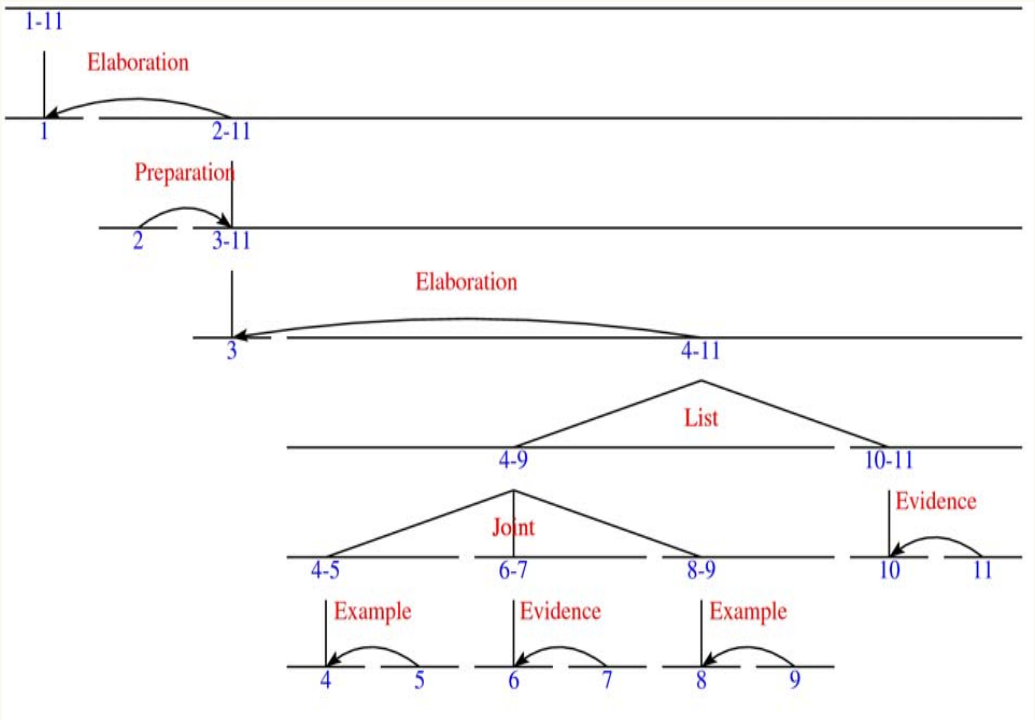
Common assumptions:

1. Discourse coherence relations hold between adjacent *discourse segments.*
   Discourse Segments can be *elementary* or *complex*
2. Discourse is structured hierarchically
   - SDRT: graph-structured
   - RST, ULDM: tree-structured
3. Two major structural types (schemas) of discourse relations:
   - hypotactic (mononuclear, subordinating)
   - paratactic (multinuclear, coordinating)

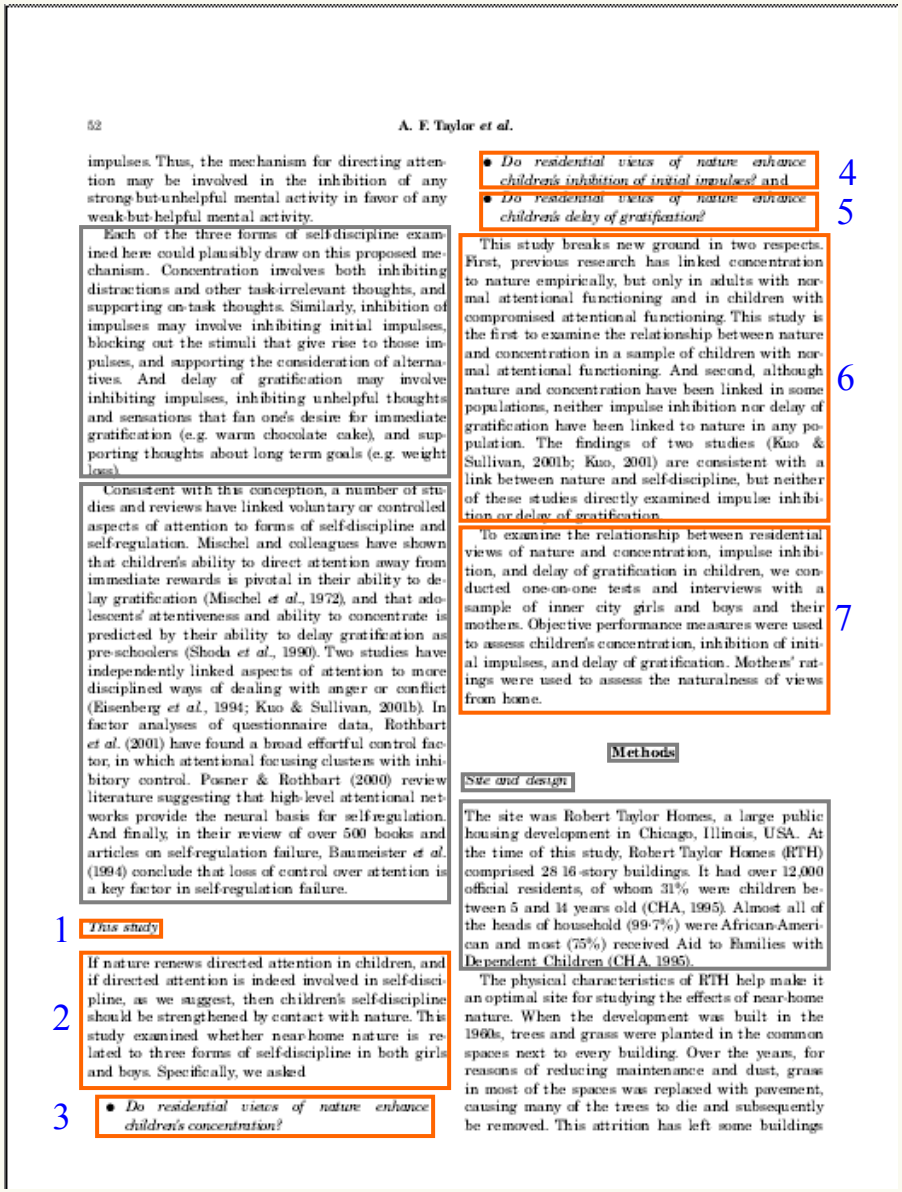


Relational structuring on the micro level

[This research considers potential predictors of workaholism and workaholic behaviors.]$_1$ [What are the wellsprings of workaholism?]$_2$ [Speculation and research findings suggest that individual difference characteristics and organizational factors likely serve as antecedents.]$_3$ [Individual characteristics include demographic factors]$_4$ [(e.g., gender),]$_5$ [family of origin dynamics]$_6$ [(Robinson, 1998),]$_7$ [and aspects of personality]$_8$ [(e.g., obsession-compulsion; Schwartz, 1982).]$_9$ [Organizational factors include values supporting work-personal life imbalance]$_{10}$ [(Schaef & Fassel, 1988; Killinger, 1991).]$_{11}$



Relational structuring on the macro level



52 A. F. Taylor *et al.*

impulses. Thus, the mechanism for directing attention may be involved in the inhibition of any strong but unhelpful mental activity in favor of any weak but helpful mental activity.

Each of the three forms of self-discipline examined here could plausibly draw on this proposed mechanism. Concentration involves both inhibiting distractions and other task-irrelevant thoughts, and supporting on-task thoughts. Similarly, inhibition of impulses may involve inhibiting initial impulses, blocking out the stimuli that give rise to those impulses, and supporting the consideration of alternatives. And delay of gratification may involve inhibiting impulses, inhibiting unhelpful thoughts and sensations that fan one's desire for immediate gratification (e.g. warm chocolate cake), and supporting thoughts about long term goals (e.g. weight loss).

Consistent with this conception, a number of studies and reviews have linked voluntary or controlled aspects of attention to forms of self-discipline and self-regulation. Mischel and colleagues have shown that children's ability to direct attention away from immediate rewards is pivotal in their ability to delay gratification (Mischel *et al.*, 1972), and that adolescents' attentiveness and ability to concentrate is predicted by their ability to delay gratification as pre-schoolers (Shoda *et al.*, 1990). Two studies have independently linked aspects of attention to more disciplined ways of dealing with anger or conflict (Eisenberg *et al.*, 1994; Kuo & Sullivan, 2001b). In factor analyses of questionnaire data, Rothbart *et al.* (2001) have found a broad effortful control factor, in which attentional focusing clusters with inhibitory control. Posner & Rothbart (2000) review literature suggesting that high-level attentional networks provide the neural basis for self-regulation. And finally, in their review of over 500 books and articles on self-regulation failure, Baumeister *et al.* (1994) conclude that loss of control over attention is a key factor in self-regulation failure.

1 *This study*

2 If nature renews directed attention in children, and if directed attention is indeed involved in self-discipline, as we suggest, then children's self-discipline should be strengthened by contact with nature. This study examined whether near-home nature is related to three forms of self-discipline in both girls and boys. Specifically, we asked

3 • *Do residential views of nature enhance children's concentration?*

4 • *Do residential views of nature enhance children's inhibition of initial impulses?* and

5 • *Do residential views of nature enhance children's delay of gratification?*

6 This study breaks new ground in two respects. First, previous research has linked concentration to nature empirically, but only in adults with normal attentional functioning and in children with compromised attentional functioning. This study is the first to examine the relationship between nature and concentration in a sample of children with normal attentional functioning. And second, although nature and concentration have been linked in some populations, neither impulse inhibition nor delay of gratification have been linked to nature in any population. The findings of two studies (Kuo & Sullivan, 2001b; Kuo, 2001) are consistent with a link between nature and self-discipline, but neither of these studies directly examined impulse inhibition or delay of gratification.

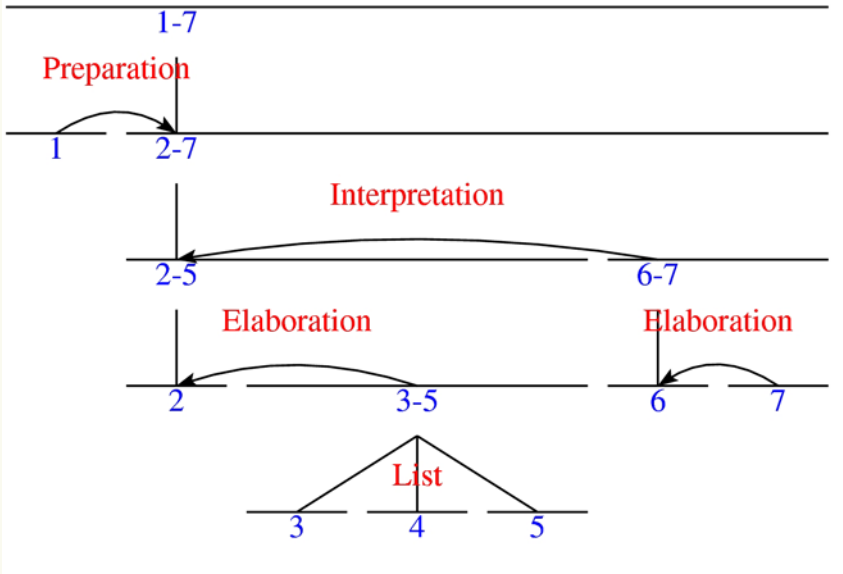
7 To examine the relationship between residential views of nature and concentration, impulse inhibition, and delay of gratification in children, we conducted one-on-one tests and interviews with a sample of inner city girls and boys and their mothers. Objective performance measures were used to assess children's concentration, inhibition of initial impulses, and delay of gratification. Mothers' ratings were used to assess the naturalness of views from home.

Methods

Site and design

The site was Robert Taylor Homes, a large public housing development in Chicago, Illinois, USA. At the time of this study, Robert Taylor Homes (RTH) comprised 28 16-story buildings. It had over 12,000 official residents, of whom 31% were children between 5 and 14 years old (CHA, 1995). Almost all of the heads of household (99·7%) were African-American and most (75%) received Aid to Families with Dependent Children (CHA, 1995).

The physical characteristics of RTH help make it an optimal site for studying the effects of near-home nature. When the development was built in the 1960s, trees and grass were planted in the common spaces next to every building. Over the years, for reasons of reducing maintenance and dust, grass in most of the spaces was replaced with pavement, causing many of the trees to die and subsequently be removed. This attrition has left some buildings



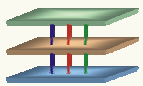
Current theories of discourse

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

ULDM	Unified Linguistic Discourse Model	Polanyi et al. 2004a,b
SDRT	Segmented Discourse Representation Theory	Asher & Lascarides 2003, Asher & Vieu 2005
RST	Rhetorical Structure Theory	Mann & Thompson 1988, Marcu 2000

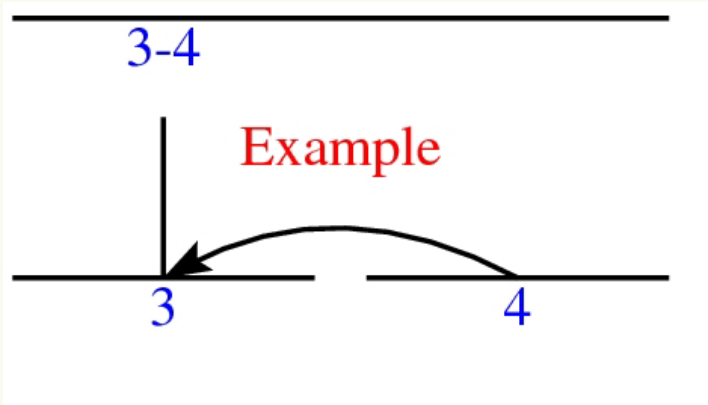
Common assumptions:

1. Discourse coherence relations hold between adjacent *discourse segments.* Discourse Segments can be *elementary* or *complex*
2. Discourse is structured hierarchically
   - SDRT: graph-structured
   - RST, ULDM: tree-structured
3. Two major structural types (schemas) of discourse relations:
   - hypotactic (mononuclear, subordinating)
   - paratactic (multinuclear, coordinating)



Hypotactic relations

1. Criterion variables included the three workaholism components identified by Spence and Robbins (1992) in their literature review and a wide range of job behaviors suggested to reflect workaholism
2. (e.g., perfectionism, hours worked, nondelegation, job stress).

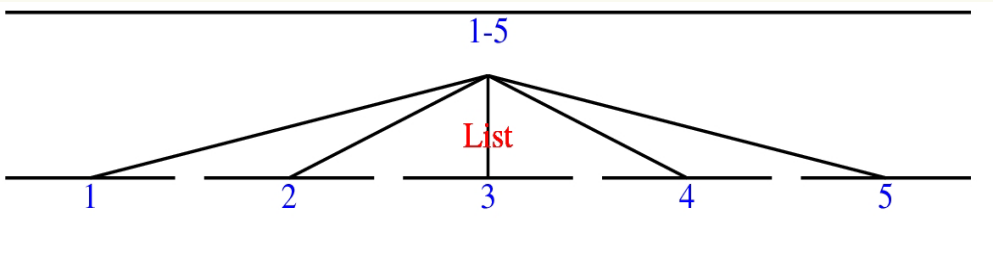


Nucleus *Satellite*

Nucleus: The more salient piece of information in a rhetorical relation

Paratactic relations

[A compelling case could be made for devoting more research attention to workaholism.]$_1$ [There has **also** been suggestions that workaholism may be increasing in North America]$_2$ [(Schor, 1991; Fassel, 1990).]$_3$ [**In addition,** it is not clear whether workaholism has positive or negative organizational consequences]$_4$ [(Machlowitz, 1980; Killinger, 1991).]$_5$ [There is **also** debate on the association of workaholic behaviors with a variety of personal well-being indicators such as psychological and physical health and self-esteem.]$_6$ [**Finally,** different types of workaholic behavior patterns likely exist, each having unique antecedents and outcomes.]$_7$



Nucleus *Nucleus* *Nucleus* *Nucleus* *Nucleus*

Discourse interpretation in SemDok

- No semantic representation…
  - … of sentences
  - … text segments
- First Source: morphology, syntax, lexikon
  - punctuation
  - lexical discourse markers
  - morphological and syntactic analysis
- Second Source: document/text structure
  - Logical document structure
  - Thematic structure (lexical cohesion and anaphoric structure)
  - Genre-specific text type structure

Scenario

What is the benefit of providing a text (a scientific article) with a discourse structure?

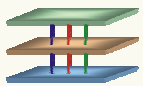
- A major processing step in many automatic text summarisation methods (Marcu 2000, Polanyi 2004)
- Preprocessing for automatic hypertextualisation to support explorative reading of scientific articles
  - Segmentation into hypertext modules
  - Generation of typed links between modules
  - Hiding of modules that contain non-salient information
  - Links leading to modules that contain salient information

Corpus

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

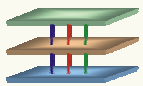
- 120 scientific articles
- English and German
- psychological and linguistic
- experimental and review

XML Annotation Layers	Annotated Articles
Logical Document Structure (DOC): extended subset of DocBook	73 + 47
Genre-specific Text Type Structure (TTS): XML Schema with 135/ 21 categories	73 + 47
Rhetorical Structure (RST): Extended subset of RST (Mann/Thompson 1988)	25 + 10
Morphological and Syntactic Structure (CNX): Machinese Syntax Tagger (Connexor Oy)	73 + 47



Overview

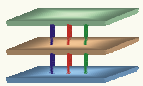
1. Introduction
2. Linguistic foundations
3. Processing
4. Status and Outlook



Processing in Version 1.0

- Auxiliary analysis components provide an input document with XML annotation layers on four levels:

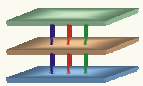
Morphology and Syntax	CNX
Logical Document Structure	DOC
Elementary Discourse Segmentation	SEG
Lexical Discourse Marker Annotation	DMS



Annotation of logical document structure

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

```
<glossentry>
   <glossterm>Confounds with problem type.
   </glossterm>
   <glossdef>
      <para>According to Johnson-Laird and Byrne (1991), valid
       syllogisms with the conclusion all or no are always single-
       model problems, syllogisms with the conclusion some are
       always single- or two-model problems, and syllogisms with
       the conclusion some ... not are always three-model
       problems. For the relational inference task, there are no
       restrictions linking the relational terms of valid
       conclusions to model counts.
      </para>
   </glossdef>
</glossentry>
```

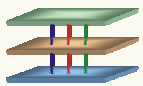



Annotation of morphology and syntax

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

```
<token id="w12250">
   <text>Confounds</text>
   <lemma value="confound"/>
   <cmp-head value="confound"/>
   <depend value="main:" head="w12249"/>
   <tags>
     <syntax value="@+FMAINV %VA"/>
     <morpho value="V PRES SG3"/>
     <pos value="V"/>
   </tags>
 </token>
 <token id="w12251">
   <text>with</text>
   <lemma value="with"/>
   <cmp-head value="with"/>
   <depend value="phr:" head="w12250"/>
   <tags>
     <syntax value="@ADVL %EH"/>
     <morpho value="PREP"/>
     <pos value="PREP"/>
   </tags>
 </token>
```

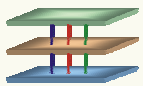
Using *Machinese Syntax* by Connexor Oy



Processing in Version 1.0

- Auxiliary analysis components provide an input document with XML annotation layers on four levels:

Morphology and Syntax	CNX
Logical Document Structure	DOC
Elementary Discourse Segmentation	SEG
Lexical Discourse Marker Annotation	DMS

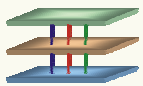


Processing in Version 1.0

- Auxiliary analysis components provide an input document with XML annotation layers on four levels:

Morphology and Syntax	CNX
Logical Document Structure	DOC
Elementary Discourse Segmentation	SEG
Lexical Discourse Marker Annotation	DMS

- Other resources defined in Version 1.0:
  - Discourse Marker Lexicon
  - Rhetorical relation taxonomy *RRSet*
  - Target representation format RST-HP



RST-HP: Basic structure

Goals:

- Representing and manipulating linguistic information using XML technology
- Employing *XML-based multi-layer annotation*
- Exploiting the XML document tree to model the hierarchical structure of discourse

Discourse structure representation RST-HP

```
<hypo relname="evidence">
  <n id="i171">
    <hypo relname="purpose">
      <n id="i55">
        <t id="ti55">The frequencies for items 1 through 3 were
           summed</t>
      </n>
      <s id="i138">
        <t id="ti138">to obtain a score for "minor" items</t>
      </s>
    </hypo>
  </n>
  <s id="i56">
    <t id="ti56">[Straus, 1979]</t>
  </s>
</hypo>
```

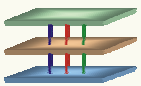
Discourse structure representation RST-HP

```
<hypo relname="evidence">
  <n id="i171">
    <hypo relname="purpose">
      <n id="i55">
        <t id="ti55">The frequencies for items 1 through 3 were
           summed</t>
      </n>
      <s id="i138">
        <t id="ti138">to obtain a score for "minor" items</t>
      </s>
    </hypo>
  </n>
  <s id="i56">
    <t id="ti56">[Straus, 1979]</t>
  </s>
</hypo>
```

Discourse structure representation RST-HP

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

```
<hypo relname="evidence">
  <n id="i171">
    <hypo relname="purpose">
      <n id="i55">
        <t id="ti55">The frequencies for items 1 through 3 were
           summed</t>
      </n>
      <s id="i138">
        <t id="ti138">to obtain a score for "minor" items</t>
      </s>
    </hypo>
  </n>
  <s id="i56">
    <t id="ti56">[Straus, 1979]</t>
  </s>
</hypo>
```

Discourse structure representation RST-HP

```
<hypo relname="evidence">
  <n id="i171">
    <hypo relname="purpose">
      <n id="i55">
        <t id="ti55">The frequencies for items 1 through 3 were
           summed</t>
      </n>
      <s id="i138">
        <t id="ti138">to obtain a score for "minor" items</t>
      </s>
    </hypo>
  </n>
  <s id="i56">
    <t id="ti56">[Straus, 1979]</t>
  </s>
</hypo>
```

Discourse structure representation RST-HP

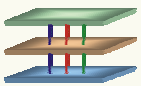
```
<hypo relname="evidence">
  <n id="i171">
    <hypo relname="purpose">
      <n id="i55">
        <t id="ti55">The frequencies for items 1 through 3 were
           summed</t>
      </n>
      <s id="i138">
        <t id="ti138">to obtain a score for "minor" items</t>
      </s>
    </hypo>
  </n>
  <s id="i56">
    <t id="ti56">[Straus, 1979]</t>
  </s>
</hypo>
```

Discourse structure representation RST-HP

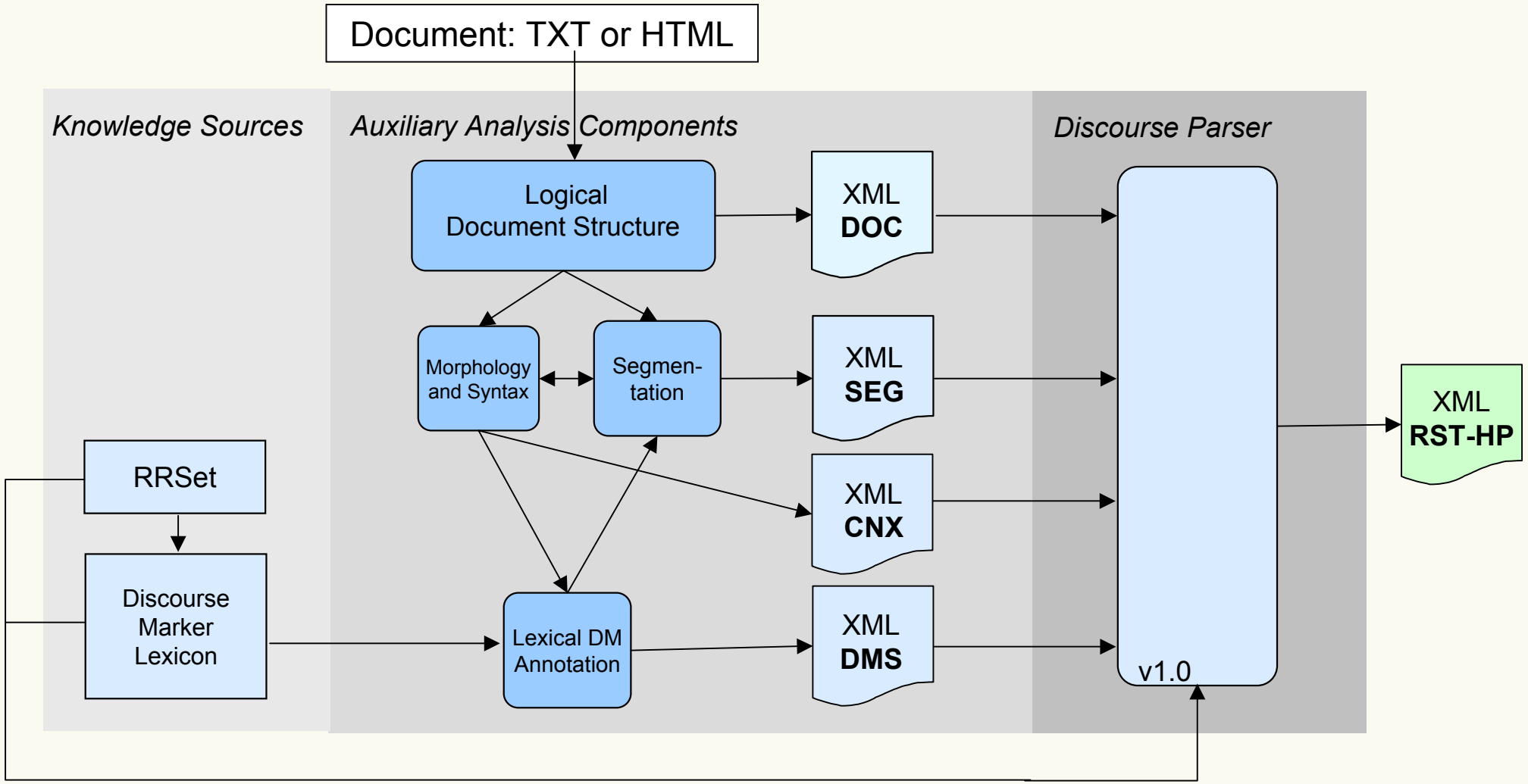
JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

Extensions to the basic tree structure

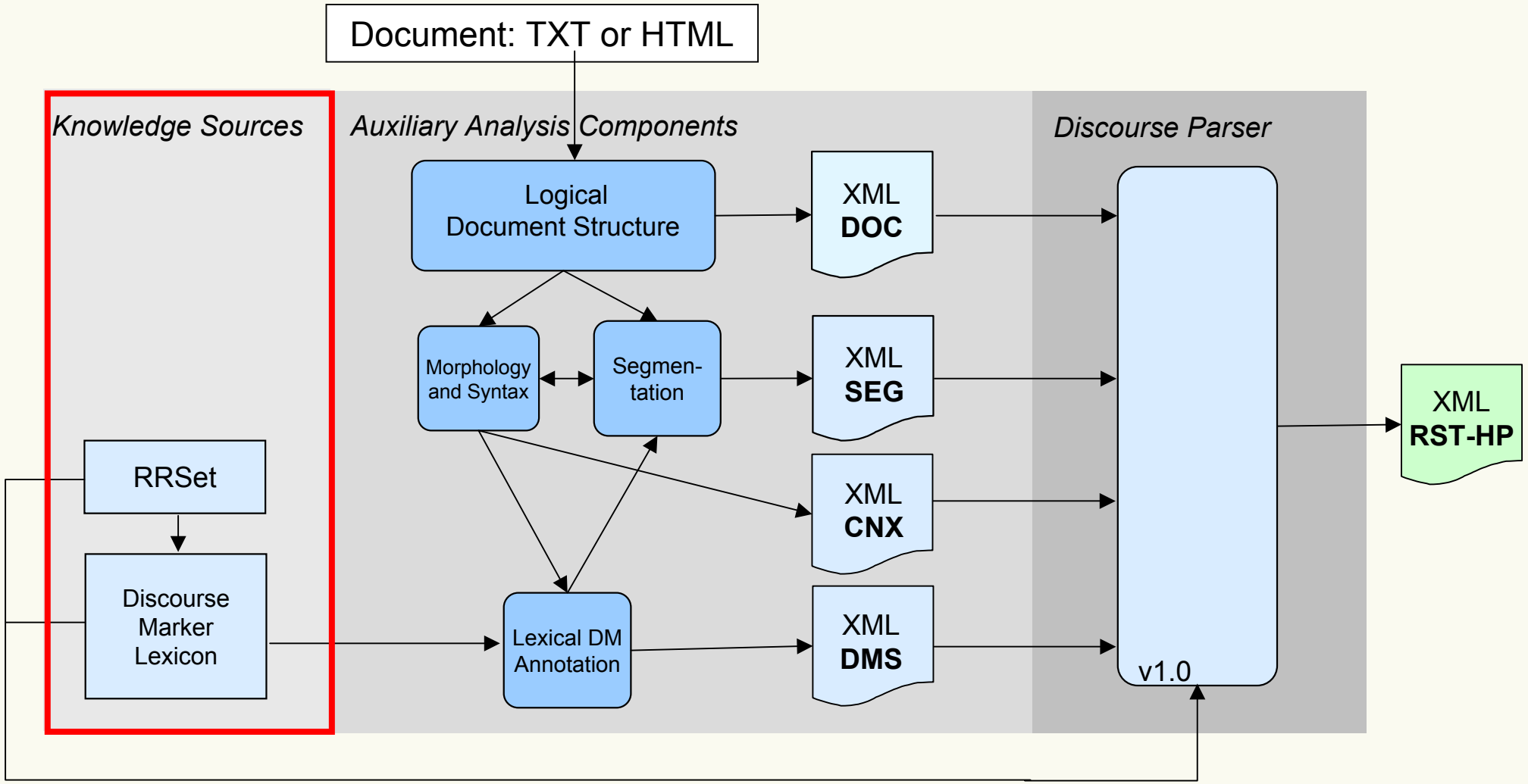
- ID/IDREF mechanism for dislocated Satellites
  (images, tables, footnotes – floating objects)
- **<embed>** elements for embedded Satellites
  (Segments disrupting other segments and marked by punctuation)
- ID/IDREF for multiple dependencies in parallel list structures



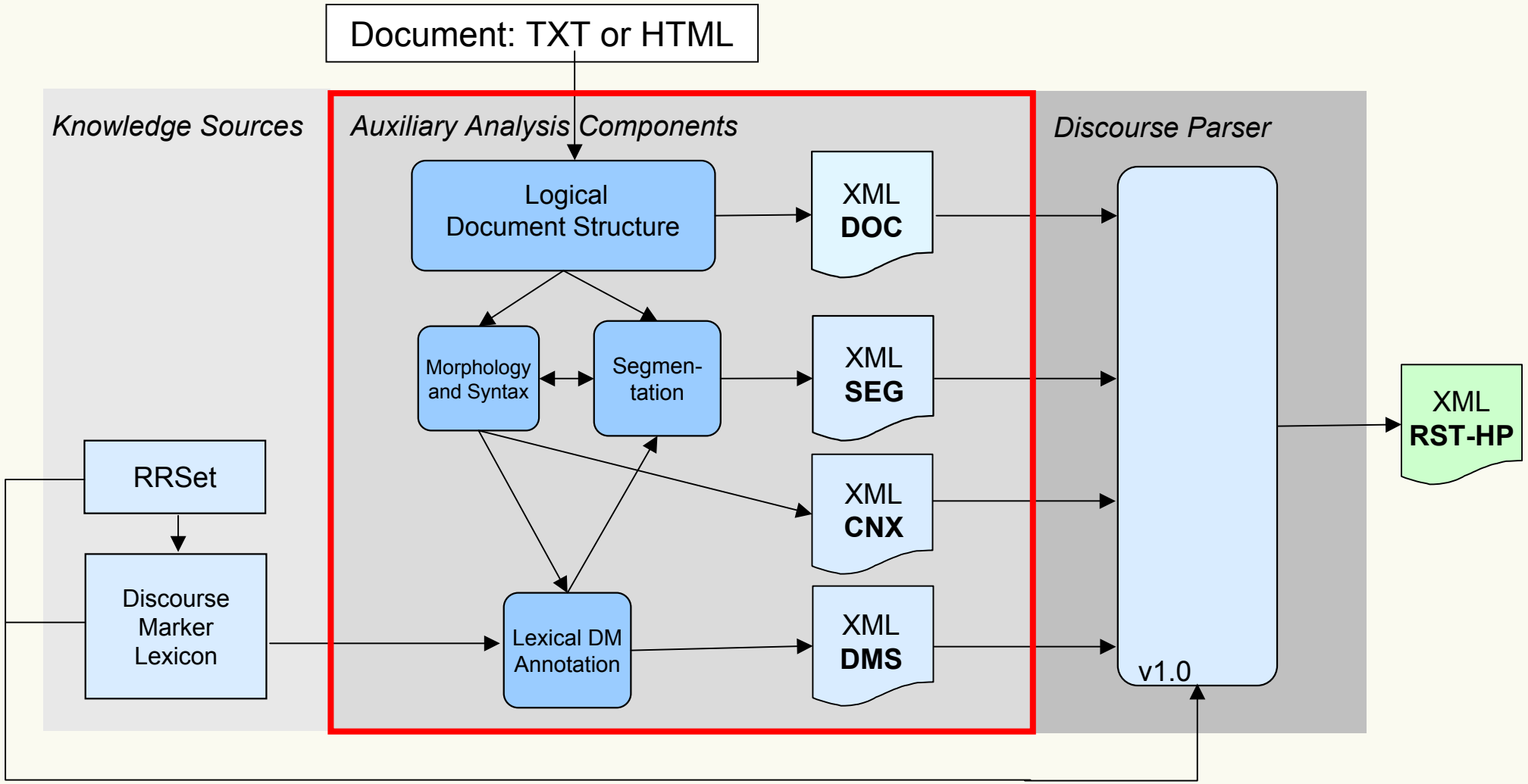
Discourse parser architecture Version 1.0



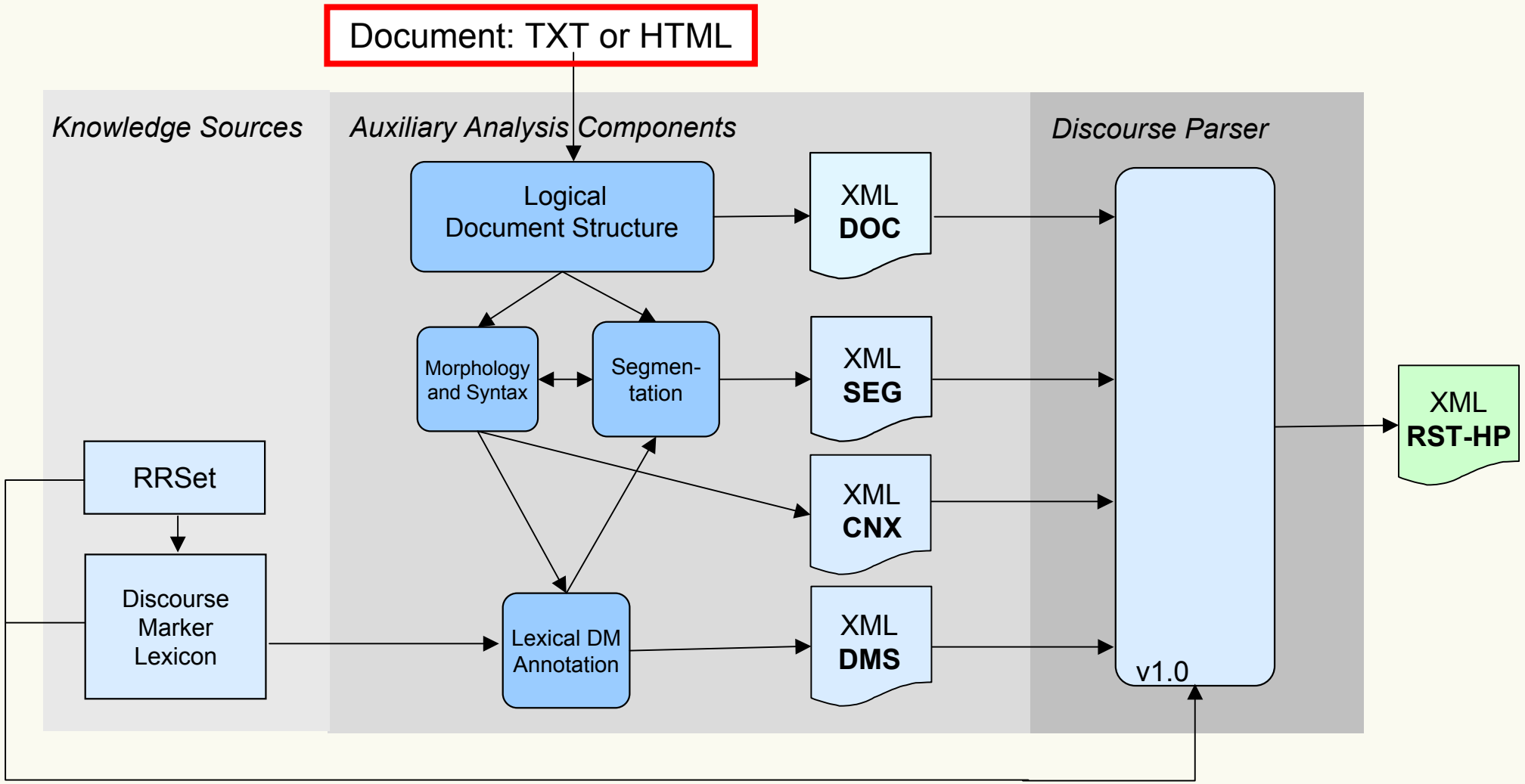
Discourse parser architecture Version 1.0



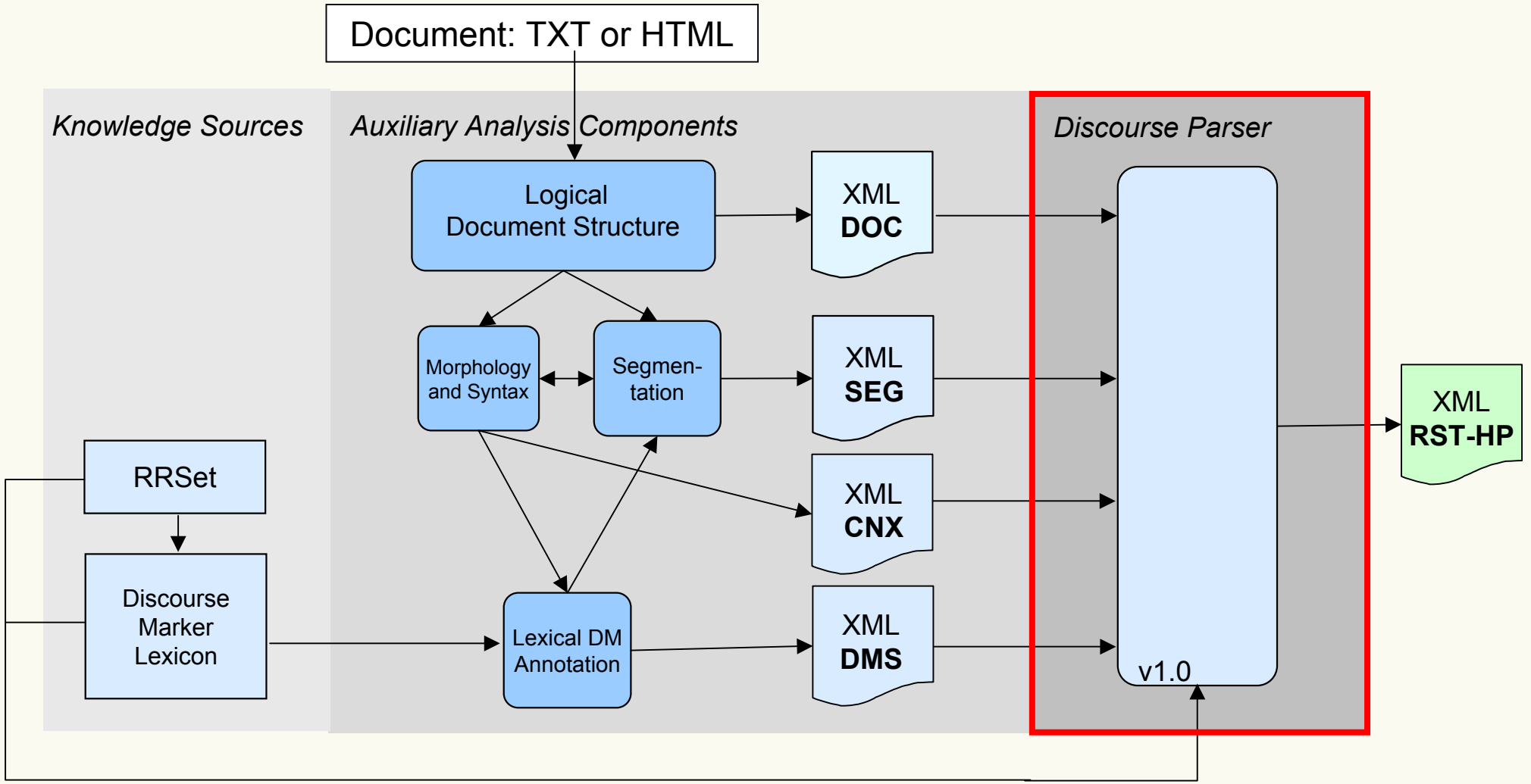
Discourse parser architecture Version 1.0



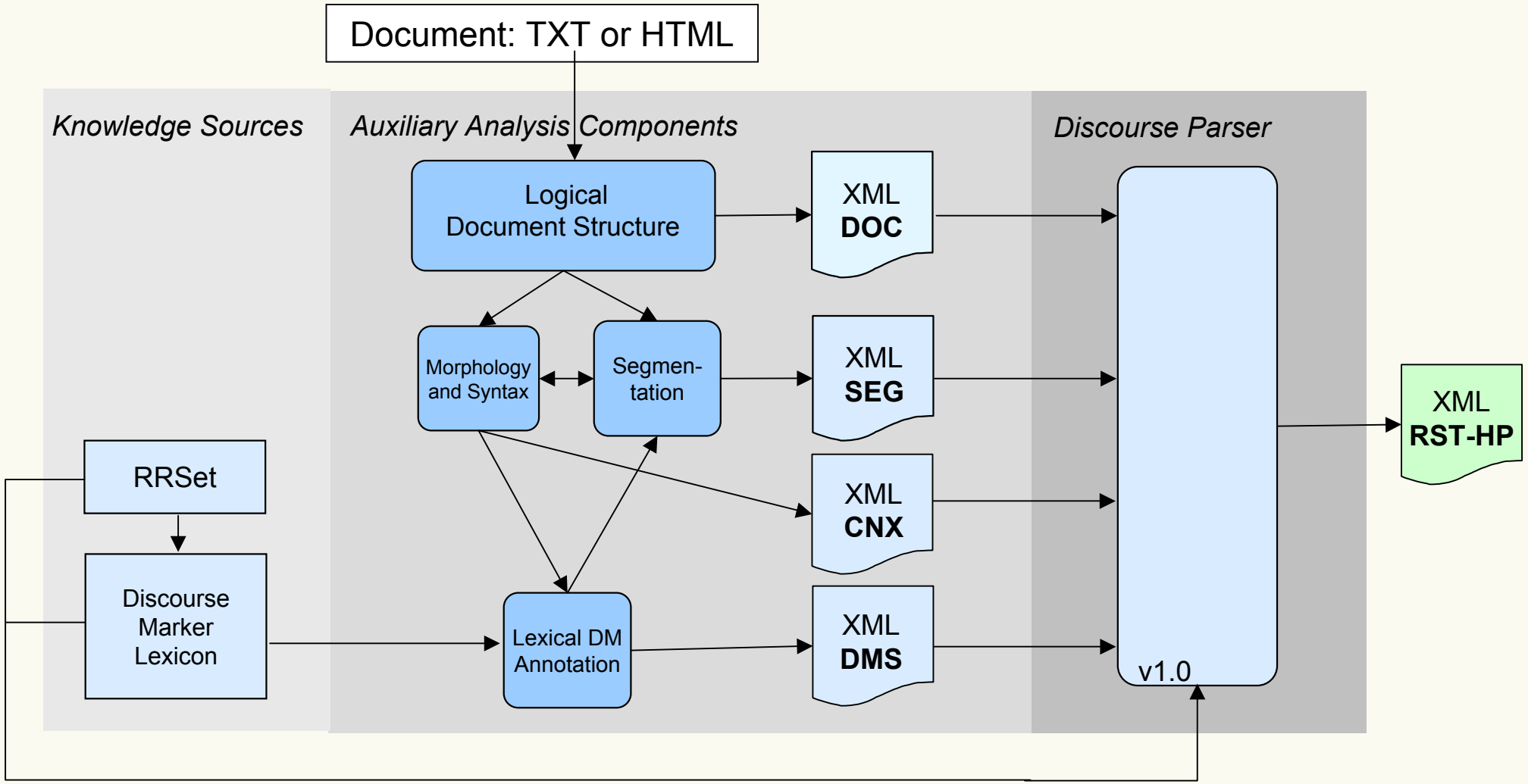
Discourse parser architecture Version 1.0



Discourse parser architecture Version 1.0



Discourse parser architecture Version 1.0

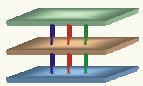


- Shift-Reduce Parser similar to Marcu (2000)
- Controlled by reduce operation rules derived from linguistic constraints

Non-terminal reduce rule introducing a hypotactic relation

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

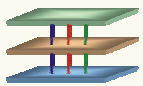
```
<reduce source="mhi" scope="sds">
  <in>
    <hpx:undefined id="$idN">$contentN</hpx:undefined>
    <hpx:undefined id="$idS">$contentS</hpx:undefined>
  </in>
  <out>
    <hpx:undefined id="generate-id()">
      <hp:hypo relname="Concession">
        <hp:n id="$idN">$contentN</hp:n>
        <hp:s id="$idS">$contentS</hp:s>
      </hp:hypo>
    </hpx:undefined>
  </out>
  <constraint test="text-inclusion($contentS, dms:dm[.='obwohl'])"/>
</reduce>
```

Non-terminal reduce rule introducing a hypotactic relation

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

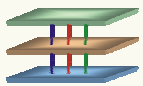
```
<reduce source="mhi" scope="sds">
  <in>
    <hpx:undefined id="$idN">$contentN</hpx:undefined>
    <hpx:undefined id="$idS">$contentS</hpx:undefined>
  </in>
  <out>
    <hpx:undefined id="generate-id()">
      <hp:hypo relname="Concession">
        <hp:n id="$idN">$contentN</hp:n>
        <hp:s id="$idS">$contentS</hp:s>
      </hp:hypo>
    </hpx:undefined>
  </out>
  <constraint test="text-inclusion($contentS, dms:dm[.='obwohl'])"/>
</reduce>
```

Non-terminal reduce rule introducing a hypotactic relation

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

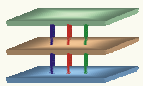
```
<reduce source="mhi" scope="sds">
  <in>
    <hpx:undefined id="$idN">$contentN</hpx:undefined>
    <hpx:undefined id="$idS">$contentS</hpx:undefined>
  </in>
  <out>
    <hpx:undefined id="generate-id()">
      <hp:hypo relname="Concession">
        <hp:n id="$idN">$contentN</hp:n>
        <hp:s id="$idS">$contentS</hp:s>
      </hp:hypo>
    </hpx:undefined>
  </out>
  <constraint test="text-inclusion($contentS, dms:dm[.='obwohl'])"/>
</reduce>
```

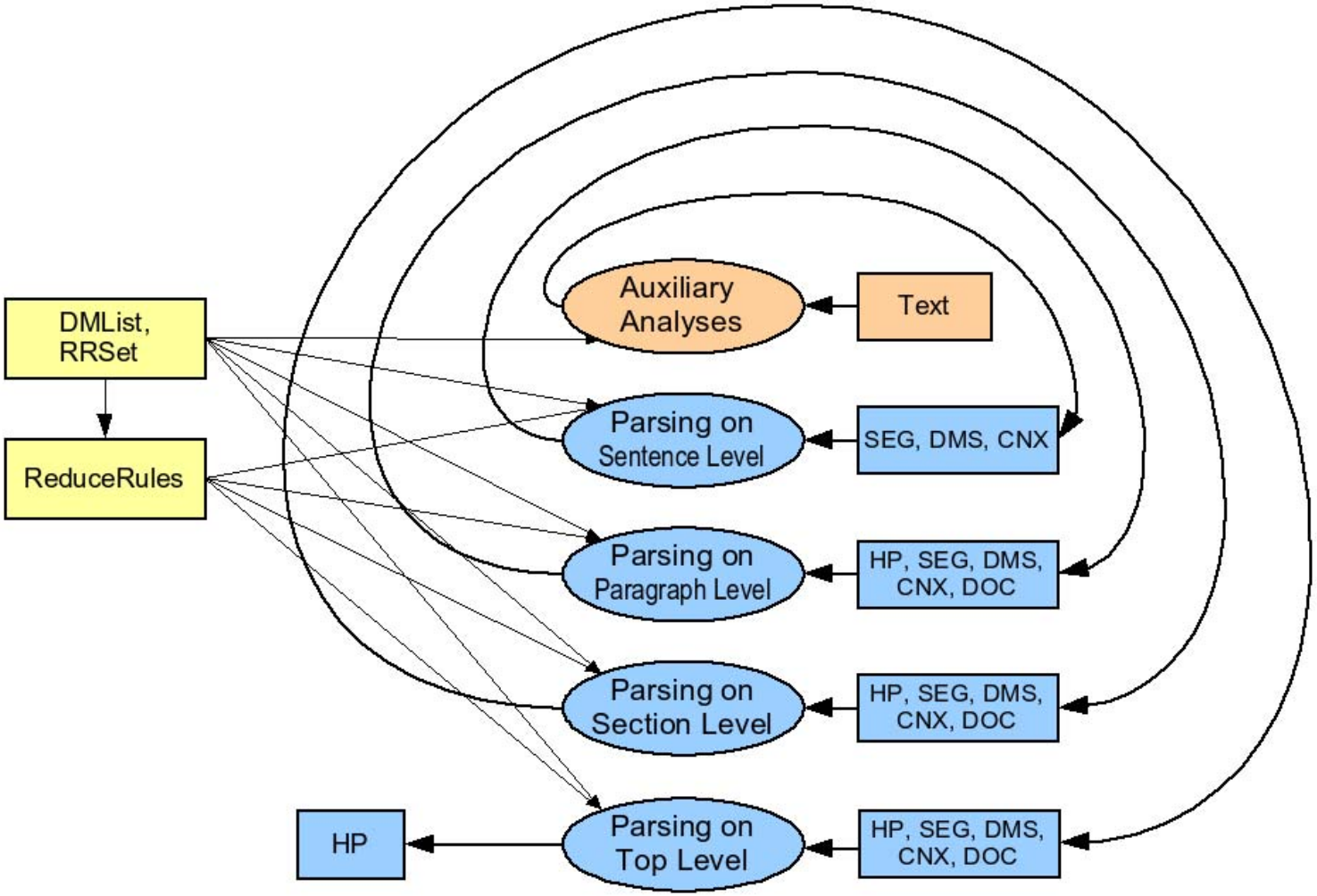
Non-terminal reduce rule introducing a hypotactic relation

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

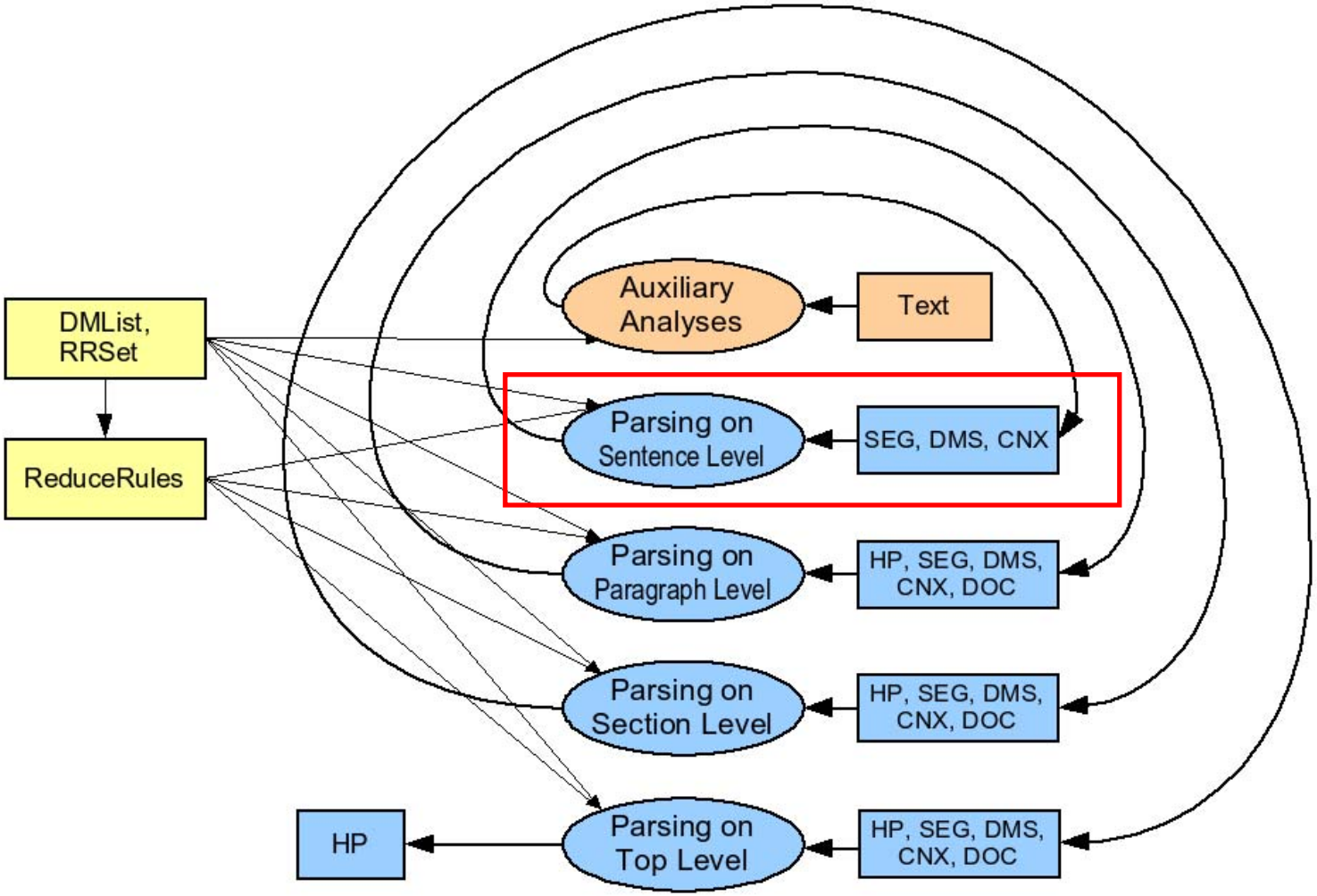
```
<reduce source="mhi" scope="sds">
  <in>
    <hpx:undefined id="$idN">$contentN</hpx:undefined>
    <hpx:undefined id="$idS">$contentS</hpx:undefined>
  </in>
  <out>
    <hpx:undefined id="generate-id()">
      <hp:hypo relname="Concession">
        <hp:n id="$idN">$contentN</hp:n>
        <hp:s id="$idS">$contentS</hp:s>
      </hp:hypo>
    </hpx:undefined>
  </out>
  <constraint test="text-inclusion($contentS, dms:dm[.='obwohl'])"/>
</reduce>
```

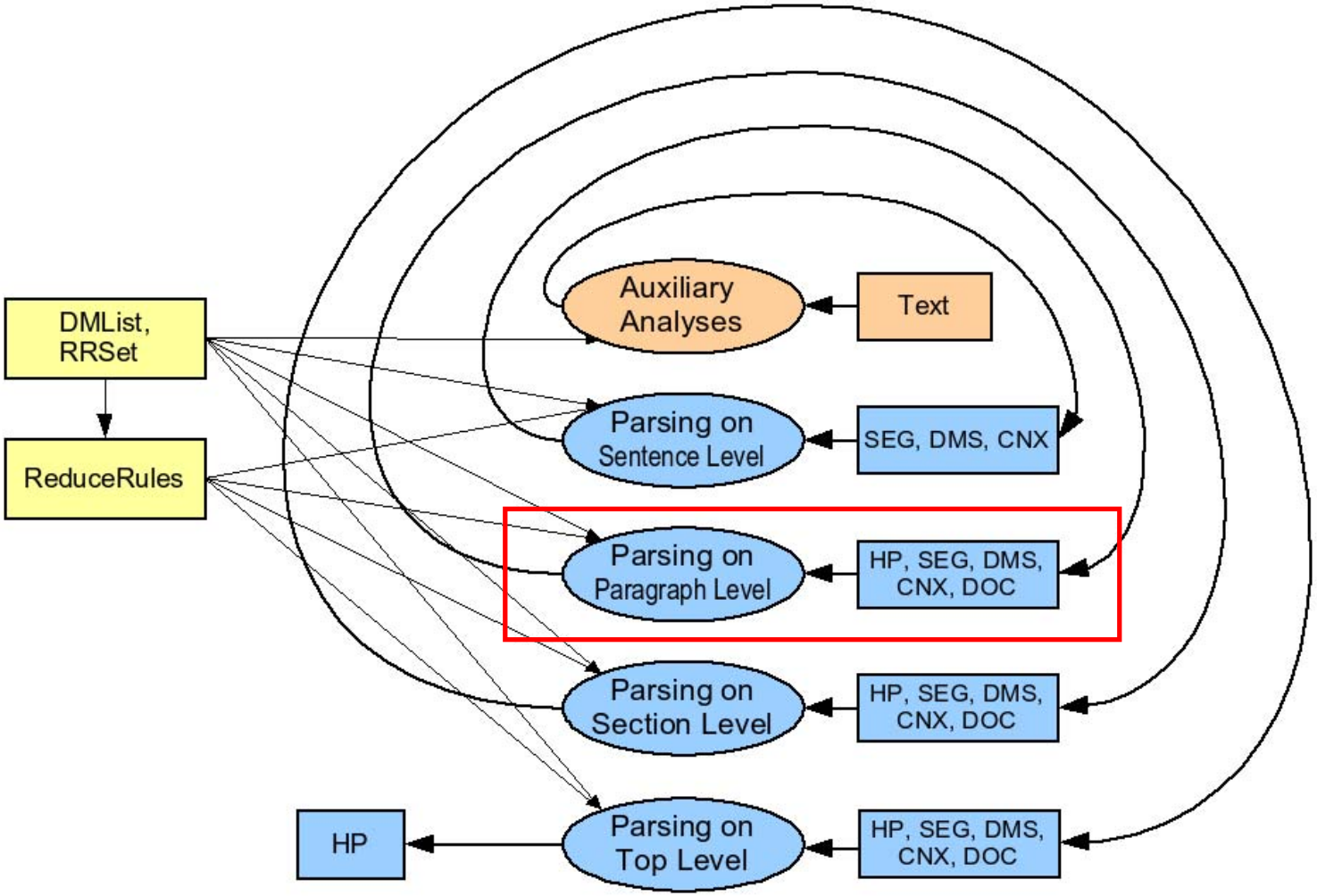
Parse loops


DMList, RRSet
ReduceRules
Auxiliary Analyses
Text
Parsing on Sentence Level
SEG, DMS, CNX
Parsing on Paragraph Level
HP, SEG, DMS, CNX, DOC
Parsing on Section Level
HP, SEG, DMS, CNX, DOC
Parsing on Top Level
HP, SEG, DMS, CNX, DOC
HP
1

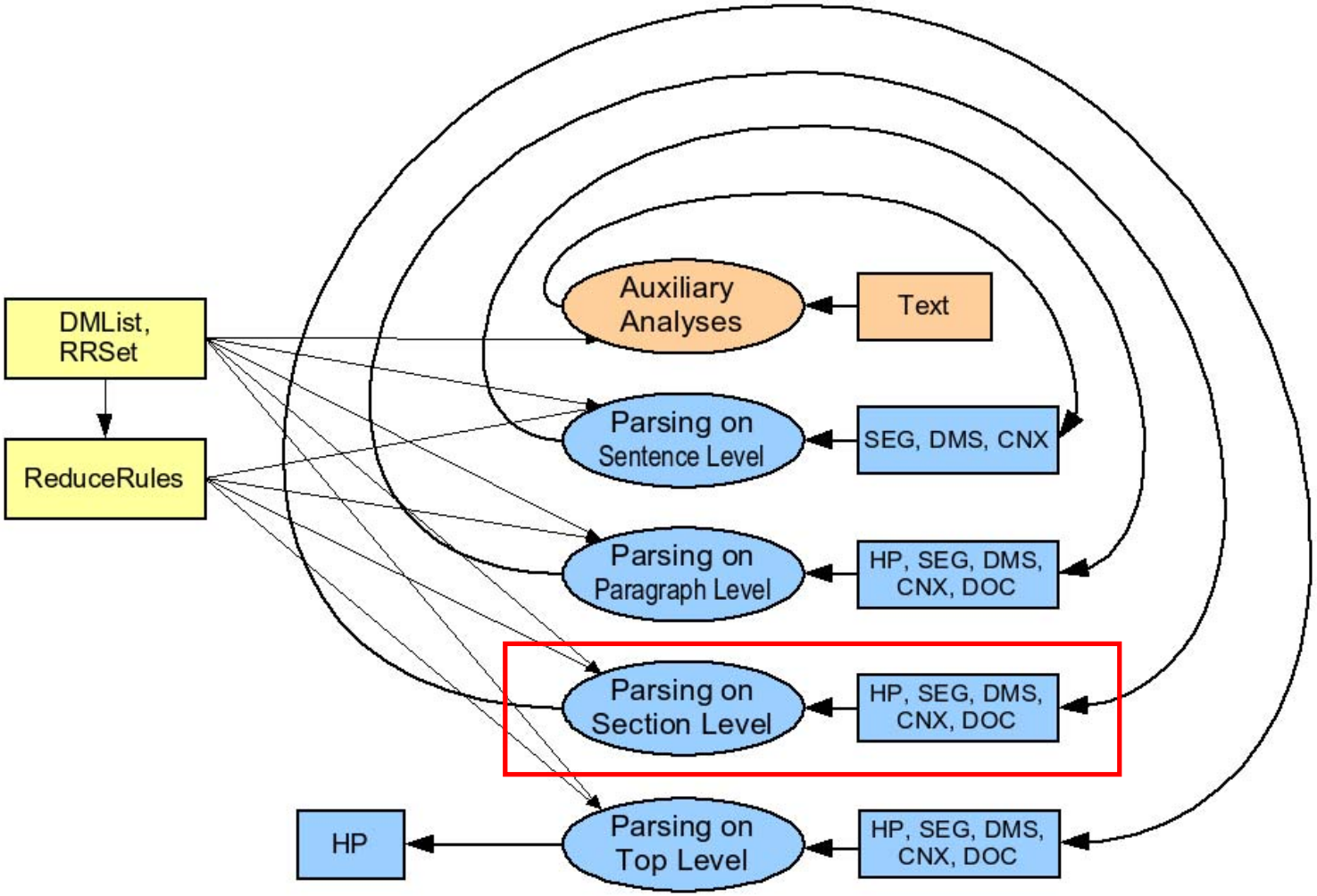
Parse loops


DMList, RRSet
ReduceRules
Auxiliary Analyses
Text
Parsing on Sentence Level
SEG, DMS, CNX
Parsing on Paragraph Level
HP, SEG, DMS, CNX, DOC
Parsing on Section Level
HP, SEG, DMS, CNX, DOC
Parsing on Top Level
HP, SEG, DMS, CNX, DOC
HP
1

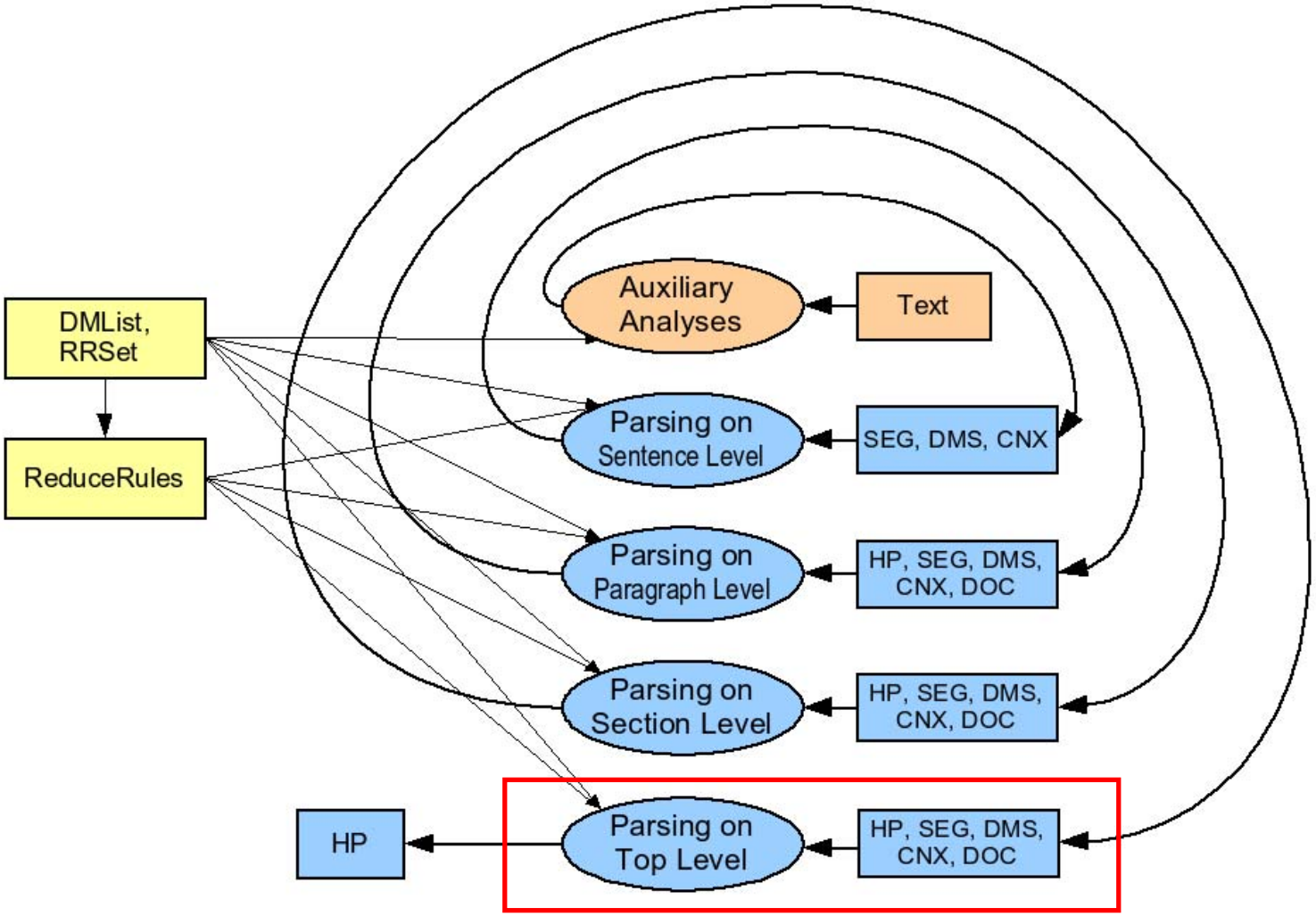
Parse loops


DMList, RRSet
ReduceRules
Auxiliary Analyses
Text
Parsing on Sentence Level
SEG, DMS, CNX
Parsing on Paragraph Level
HP, SEG, DMS, CNX, DOC
Parsing on Section Level
HP, SEG, DMS, CNX, DOC
Parsing on Top Level
HP, SEG, DMS, CNX, DOC
HP
1

Parse loops


DMList, RRSet
ReduceRules
Auxiliary Analyses
Text
Parsing on Sentence Level
SEG, DMS, CNX
Parsing on Paragraph Level
HP, SEG, DMS, CNX, DOC
Parsing on Section Level
HP, SEG, DMS, CNX, DOC
Parsing on Top Level
HP, SEG, DMS, CNX, DOC
HP
1

Parse loops

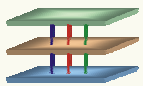

DMList, RRSet
ReduceRules
Auxiliary Analyses
Text
Parsing on Sentence Level
SEG, DMS, CNX
Parsing on Paragraph Level
HP, SEG, DMS, CNX, DOC
Parsing on Section Level
HP, SEG, DMS, CNX, DOC
Parsing on Top Level
HP, SEG, DMS, CNX, DOC
HP
1

Parsing the input annotations

- Input annotations are loaded as a Prolog fact base according to Bayerl et al. (2003)
- XML-based multi-layer annotation: The primary textual data are annotated several times, still in the internal Prolog representation they appear only once and serve as a link between annotation layers (Bayerl et al. 2003)
- SEG annotation layer is the annotation layer that controls the parse loops
- Method: Shift-Reduce Parsing similar to Marcu (2000)
- If two of more different relations are proposed to hold between two segments, the taxonomy of relations defined in the RRSet can be consulted to find a common parent relation

Overview

1. Introduction
2. Linguistic foundations
3. Processing
4. Status and Outlook



Status

- Project is in its first year
  - Resources are built
    - Corpus and annotations
    - Lexical discourse marker lexicon
    - RRSet
    - Target format
    - Reduce rules
  - Automatic segmentation into elementary and sentential discourse segments is implemented
  - Parser up to paragraph level is (almost) implemented, no evaluations yet

Outlook

- Parsing of *Elaboration* relations
  - Not signalled by lexical discourse markers
  - Via ontological relations between discourse referents (cf. Polanyi 2004a)
  - Use GermaNet and a domain ontology
  - Special Interest: Event anaphora
- Identification of complex discourse segments
  - Find thematically coherent text spans via Lexical Cohesion Analysis (Hirst 1991)
  - Identify mismatches between discourse structure and logical document structure
  - Text type structure category detection
- Evaluation
  - Comparison with the manually constructed annotations of our corpus using standard methods
  - External evaluation in cooperation with project partner: Does the discourse structure improve automatic hypertextualisation?