Originally published in: Kühnlein, Peter/Benz, Anton/Sidner, Candace L. (Eds.): Constraints in Discourse 2. - Amsterdam/Philadelphia: Benjamins, 2010. Pp. 81- 104. (Pragmatics & Beyond New Series, Volume 194)
DOI: 10.1075/pbns.194.05bar

The role of logical and generic document structure in relational discourse analysis

Maja Bärenfänger, Harald Lüngen, Mirco Hilbert, Henning Lobin

Fachgebiet Angewandte Sprachwissenschaft und Computerlinguistik

Institut für Germanistik

Justus-Liebig-Universität Gießen

Otto-Behaghel-Str. 10 D

D-35394 Gießen, Germany

e-mail: {maja.baerenfaenger|harald.luengen|mirco.hilbert|henning.lobin}@uni-giessen.de

1. Introduction

The theme of this article is a corpus-based investigation of the role of logical and generic document structure in the relational discourse analysis of complex texts by the example of scientific journal articles. One aim of this is to formulate cues and constraints such that they can be used in a discourse parser for automated discourse analysis in the line of Rhetorical Structure Theory (RST, Mann and Thompson 1988). Traditionally, cues for relational discourse analysis have been derived from lexical discourse markers and syntactic features of an input text. This strategy is well-established and has yielded good results for texts of a limited size such as newspaper articles. But discourse relations between larger segments of text such as the sections and paragraphs of a research article are frequently not signalled by cues that can be identified by shallow analyses of vocabulary and grammar. We thus suggest to additionally base discourse parsing on an analysis of the logical document structure (the division of a text into textual objects such as titles, paragraphs, tables, and lists) and the generic document structure (parts of the text corresponding to text type-specific categories such as *introduction, method, results*, and *discussion* in the case of scientific articles). We aim to clarify what kind of cues and constraints for relational discourse analysis can be observed on these levels and how they can be derived from the corresponding linguistic and structural annotations of a text document. To this end, we examine a corpus of German and English scientific articles in the fields of psychology and linguistics. Its documents are annotated on the three levels in question, namely their logical document structure, their generic document structure (text type structure) and a discourse structure according to RST (which is also the target structure of an RST discourse parser).

2. Representation of relational discourse structure

Rhetorical Structure Theory (RST, Mann and Thompson 1988, Marcu 2000) shares three basic assumptions with other linguistic discourse theories like *Segmented Discourse Representation Theory* (SDRT, Asher and Lascarides 2003) and the *Unified Linguistic Discourse Model* (ULDM, Polanyi et al. 2004a, Polanyi et al. 2004b): 1. Discourse structure can be modeled as a system of discourse coherence relations which hold between parts of text, i.e. elementary or complex discourse segments. 2. Complex discourse segments are structured hierarchically and can be represented either as a graph (SDRT) or as a tree (ULDM, RST). 3. Discourse coherence relations are either hypotactic (subordinating, mononuclear) or paratactic (coordinating, multinuclear).

RST can be considered a functional theory of text structure: "It describes the relations among text parts in functional terms." (Mann and Thompson 1988, p.271). Relations between discourse segments are, amongst others, identified and described according to the effect the relational propositions have on the reader. The assignment of relations to pairs of discourse segments thus involves the recognition of the goals and beliefs of authors and readers about the meaning and function of these discourse segments. "An RST analysis always constitutes a plausible account of what the writer wanted to achieve with each part of the text. An RST analysis is thus a functional account of the text as a whole." (Mann and Thompson 1988, p.258). The definitions of the RST relations reflect this functional perspective by describing the characteristics (i.e. constraints and effects) of the relations from the points of view of author and reader, as intentions and effects (see Table 1 for an example of a definition of an RST relation).

When judgments about the intentions of the author play such a crucial role in discourse analysis according to RST, in which way can this theory be implemented in a computational approach to discourse parsing? In previous projects (Corston-Oliver 1998, Marcu 2000, Carlson and Marcu 2001, Reitter 2003), linguistic properties like syntactic or lexical features have been applied as *discourse markers* for the assignment of a discourse relation to sets of discourse segments. In this article, the term discourse marker is used to refer to a class of expressions (adverbs or conjunctions) as well as syntactic constructions which can be distinguished by their "function in discourse and the kind of meaning they encode" (Blakemore 2004, p.221). Discourse markers are treated as signals the author uses to communicate his goals and beliefs to the reader, and which, more specifically, signal a specific discourse relation. Examples of discourse markers are connectives like "jedoch" (= "however"), which marks the nucleus of a CONCESSION-relation (see Listing 1[1]), parallel syntactic constructions which may induce a LIST-relation, and punctuation marks such as ':' (co-

relation name:	CONCESSION
constraints on N:	W (the writer) has positive regard for the situation presented in N
constraints on S:	W is not claiming that the situation presented in S doesn't hold
constraints on the N + S combination:	W acknowledges a potential or apparent incompatibility between the situations presented in N and S; W regards the situations presented in N and S as compatible; recognizing that the compatibility between the situations presented in N and S increases R's positive regard for the situation presented in N
the effect:	R's (the reader's) positive regard for the situation presented in N is increased
locus of the effect:	N and S
discourse markers:	in S: obwohl, obschon, obgleich, obzwar, zwar [although, though ...] in N: dennoch, doch, trotzdem [however, anyhow, nevertheless …] in S: obwohl – in N: so in S: zwar – in N: jedoch

Table 1: Definition of the relation CONCESSION (Mann and Thompson 1988, p.254f.)

lon), which when occurring at the end of a segment may signal a PREPARATION-relation. As a consequence, an RST analysis does not, unlike an SDRT analysis, require a fully-fledged semantic representation of discourse segments.

Such a surface-oriented approach often works well for shortish text types, e.g. newspaper articles, which are characterised by a limited size and a relatively simple document and syntactic structure (Marcu 2000, Carlson and Marcu 2001, Reitter 2003). But when it comes to complex text types or longer texts with a deeply nested discourse structure, it is necessary to consider additional knowledge sources which can provide cues and constraints for the interpretation of higher levels of discourse structure, and discourse relations which are not indicated by lexical or syntactic discourse markers (such as ELABORATION or BACKGROUND). The term *cue* is used here to contrast with the term *discourse marker* to refer to more abstract signals for discourse relations like cues from the logical or generic document structure. We also distinguish cues from *constraints*: While cues can be used for bottom-up relational discourse analysis and in that respect behave like discourse markers that signal a specific discourse relation, constraints serve as top-down restrictions for discourse structures.

In our project (Lobin et al. 2006; Lüngen et al. 2006), we deal with a corpus of scientific articles[2] which exhibit a highly complex document structure and a relatively large average size (~ 8600 words per article). A complex document structure implies that they are characterised by a deeply nested hierarchical structure with several levels of embedded dis-

```
<hypo relname="concession" id="i217">
<s><dm lemma="zwar" pos="ADV">Zwar</dm> wurde in der Fremdsprachenerwerbsforschung im
   Zusammenhang mit der noticing-Hypothese die Rolle der auf den Input gerichteten Auf
   merksamkeit untersucht. </s>
<n>Die Funktion der lernerseitigen Aufmerksamkeit für den Output im L2-Erwerb blieb
   bisher <dm lemma="jedoch" pos="ADV">jedoch</dm>weitgehend unberücksichtigt. </n>
</hypo>
```

Listing 1: Discourse marker for CONCESSION

course segments where the distance between the level of elementary discourse segments (EDS) and the highest level of complex discourse segments (CDS) may be five or more embeddings. As a consequence, the majority of discourse segments are not elementary, but complex – this means that lexical and syntactic discourse markers can only be applied in a limited way. Apart from their complex document (and discourse) structure, our corpus of scientific articles is characterised by a high frequency of ELABORATION relations – which are usually not indicated by lexical or syntactic discourse markers.

Originally, RST provides a set of 26 rhetorical relations, which are either mono- or multinuclear (Mann and Thompson 1988). Apart from the distinction between mono- and multinuclear relations, relations can also be subdivided into two groups based on the intentions of the author and the effects on the reader: 1. subject matter relations: "those whose intended effect is that the reader recognises the relation in question"; 2. presentational relations: "those whose intended effect is to increase some inclination in the reader, such as the desire to act or the degree of positive regard for, belief in, or acceptance of the nucleus." (Mann and Thompson 1988, p. 257).

For our corpus of scientific articles, we adapted the relation set proposed by Mann and Taboada (2005) but extended and restructured it – by defining our own relation taxonomy (as was also done by Hovy and Maier 1995, and Carlson and Marcu 2001). The motivation for all modifications of the relation set is twofold: First, the set had to be adapted to the characteristics of the text type of the documents in our corpus. Second, the set should support our application scenario[3] and therefore distinguish between relations that are mainly induced by the logical document structure or generic document structure, and relations that are mainly induced by lexical, syntactic or morphological features.

A subset of 20 articles of our corpus was annotated according to rhetorical structure using the RSTTool developed by O'Donnell (2000). Subsequently, the representations produced were converted by a Perl script into our *RST-HP* format (Lüngen et al. 2006), where, unlike in O'Donnell's representation, the basic XML tree structure of a document is also used to represent an RST tree. An RST-HP extract is displayed in Listing 1.

Besides logical and generic document structure, another level which might be called *thematic structure* plays an important role in the instantiation of certain "subject matter" relations like BACKGROUND, ELABORATION, and its subtypes (Lüngen et al. 2006). In the present article, however, we will solely concentrate on constraints and cues that can be derived from the logical and generic document structure.

3. Cues and constraints from logical document structure

According to Power et al. (2003, p.213), "document structure describes the organization of a document into graphical constituents like sections, paragraphs, sentences, bulleted lists, and figures" as well as elements like "quotation and emphasis". Such constituents can be described according to their graphical or geometric properties – they are 2D-objects which cover parts of the document area (Lobin et al. 2006). The *physical layout* structure of a

```
<glosslist>
  <glossentry>
    <glossterm>A. Dialekte: </glossterm>
    <glossdef>
       <para>Diese sind gekennzeichnet durch
          eine räumlich geringe kommunikati
          ve Reichweite aufgrund phonologi
          scher, morphosyntaktischer und le
          xikalischer Eigenheiten, die nur
          für kleine geografische Räume
         (z.B. innerhalb eines Dorfes) gel
          ten und sie von anderen regionalen
          Varietäten und von der Standard
          sprache unterscheiden.
       </para>
    </glossdef>
  </glossentry>
</glosslist>
```

Listing 2: DocBook annotation (extract)

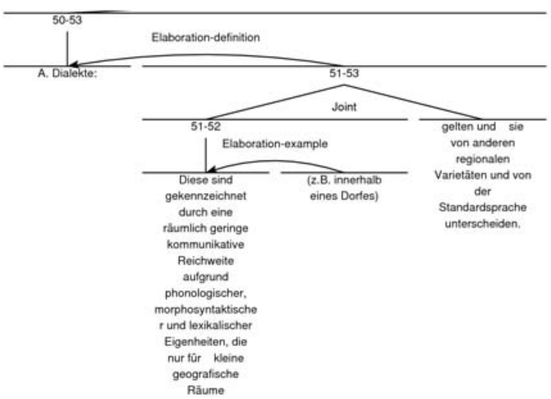

50-53
Elaboration-definition
A. Dialekte:
51-53
Joint
51-52
Elaboration-example
Diese sind gekennzeichnet durch eine räumlich geringe kommunikative Reichweite aufgrund phonologischer, morphosyntaktischer und lexikalischer Eigenheiten, die nur für kleine geografische Räume
(z.B. innerhalb eines Dorfes)
gelten und sie von anderen regionalen Varietäten und von der Standardsprache unterscheiden.

Figure 1: RST annotation for extract in Listing 2

document is a manifestation of its *logical* document structure, as in the physical layout different structural functions of parts of a text can be identified such as list, paragraph, or heading. In our corpus, logical document structure is encoded according to DocBook markup (Walsh and Muellner, 1999).

At the level of the logical document structure, we can distinguish elementary and complex constituents. The latter are combinations of adjacent elementary or (smaller) complex constituents (parts of text). This combination follows compositional principles. A document can therefore be described as structured hierarchically: complex constituents are aggregations of elementary or complex ones, e.g. an article consists of sections, a section is divided into a set of subsections or paragraphs and perhaps lists or figures, and a paragraph may contain quotations or emphasised tokens. Elements of the logical document structure indicate specific discourse substructures and can thus be used as cues for the assignment of rhetorical relations. Document structural elements which serve as cues are, for example, *listitem, glossterm, glossdef, caption* and *title*. *Listitems* indicate a LIST_DM-OTHER relation between all items of a bulleted list (as shown in Listing 3 and Figure 2), *glossterms* may induce the nucleus in an ELABORATION-DEFINITION relation, *glossdef* the satellite (shown in Listing 2 and Figure 1), *captions* often have the status of a satellite in a CIRCUMSTANCE relation with a figure or table being the nucleus, and *titles* are the satellite in a PREPARATION-TITLE relation, where the nucleus may be a section, a table or a figure.

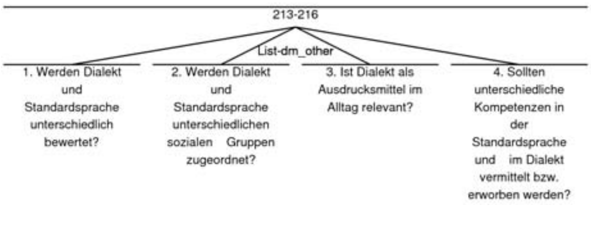

213-216
List-dm_other
1. Werden Dialekt und Standardsprache unterschiedlich bewertet?
2. Werden Dialekt und Standardsprache unterschiedlichen sozialen Gruppen zugeordnet?
3. Ist Dialekt als Ausdrucksmittel im Alltag relevant?
4. Sollten unterschiedliche Kompetenzen in der Standardsprache und im Dialekt vermittelt bzw. erworben werden?

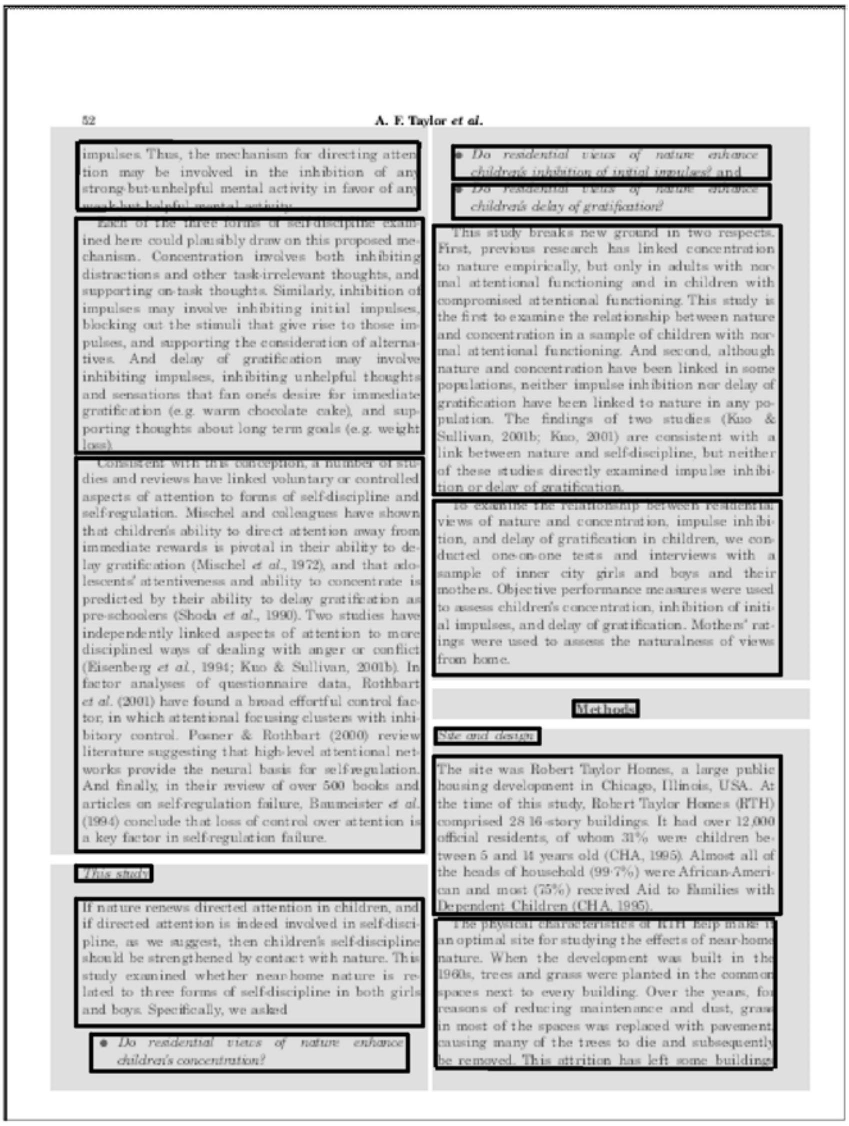
Figure 2: RST annotation for extract in Listing 3

```
<orderedlist>
    <listitem>
        <para>1. Werden Dialekt und Stan-
        dardsprache unterschiedlich bewer-
        tet?</para>
    </listitem>
    <listitem>
        <para>2. Werden Dialekt und Stan-
        dardsprache unterschiedlichen
        sozialen Gruppen zugeordnet?</para>
    </listitem>
    <listitem>...</listitem>
</orderedList>
```

Listing 3: DocBook annotation (extract)

Apart from these (and other) *cues*, the logical document structure also provides *constraints* for relational discourse analysis, insofar as the units of the logical document structure act as building blocks for discourse spans. Before we explain this statement in more detail, we shortly have to introduce our typology of discourse segments. Because of the complexity and deep nesting of discourse segments in scientific articles, we do not only distinguish elementary (EDS), sentential (SDS), and complex discourse segments (CDS), but we additionally distinguish different types of CDS (see Figure 3 for a graphical illustration):

- CDS type=”block”: This segment type corresponds to paragraphs and structural elements that are on a par with paragraphs such as titles and captions, i.e. all element types from the logical structure that contain only text or text plus inline elements. The name of the attribute “block” of this complex segment type is due to its correspondence to geometric objects that are two-dimensional blocks (rectangles) in the physical layout of a document. The segments of type “block” partition the document, i.e. every piece of text is part of exactly one CDS type=”block”. In a discourse tree for a CDS type=”block”, the block acts as an upper boundary, and SDS (sentential discourse segments) act as the basic segments for the construction of discourse subtrees.
- CDS type=”division”: Complex discourse segment of the type “division” correspond to the lowest section level. In terms of DocBook markup it comprises the smallest occurring sect1, sect2, sect3, sect4, or sect5 elements plus elements that are on a par with it, i.e. titles and paragraphs that are sisters of sect elements. The segments of type “division” also partition the whole document. In a discourse tree for a CDS type=”division”, the division acts as an upper boundary, and the CDS type=”block” acts as a basic segment type for the construction of discourse subtrees.



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.

This study

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

- *Do residential views of nature enhance children's concentration?*
- *Do residential views of nature enhance children's inhibition of initial impulses?* and
- *Do residential views of nature enhance children's delay of gratification?*

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.

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

Figure 3: Areas of text that correspond to complex discourse segments of the types *block* (black frames) and *division* (grey rectangular areas)

- CDS type=”document”: This segment type comprises all residual sect elements, i.e. those which are on a higher level than the ones described under CDS type=”division”.

Thus, the CDS type=”document” level is special in that its segments most of the time do not partition the document, depending on the depth of embedded sections. On the other hand, the segments of the “document” can be identical to those of the type “division” in a document that contains only sections of the DocBook element type sect1. In a discourse tree for a complete document, the segments of type “division” are the basic elements, and all segments of the type “document” must correspond to exactly one subtree.

This differentiation of levels or granularities of discourse segments is comparable to that proposed by Marcu (2000), who distinguishes clause, sentence, paragraph, and section level, and LeThan et al. (2004), who describe sentence-level and text-level discourse segments. In our approach, units of the logical document structure (paragraph, section, article) are used to constrain the extent to which discourse segments can be relationally combined to pairs of discourse segments, i.e. they serve as boundaries for discourse segments. This means that, for example, a CDS type=”block” can only be related to another CDS type=”block”, but not to a CDS type=”division”. By assuming that the rhetorical structure correlates with the logical document structure, or, as Marcu (2000, p.109) says, “that sentences, paragraphs, and sections correspond to hierarchical spans in the rhetorical representation of the text”, the amount of possible rhetorical interpretations can be reduced significantly.

In the following sections, we will examine the generic document structure of scientific articles and how it can function as a knowledge source in discourse parsing according to RST. First, a text type structure schema for scientific articles and the corpus annotations based on it will be presented. Next, we investigate what kind of cues and constraints can be derived from the generic structure and from statistic analyses of the annotations.

4. Cues and constraints from generic document structure

4.1. Interrelations between generic document structure and relational discourse structure

Apart from the *logical* document structure, a second type of document structure exists: the *generic* document structure, or genre-specific *text type structure* (TTS), or *superstructure*

(Swales 1990, van Dijk 1980). It describes the global organisation of a document into genre-specific functional categories (or *zones*, after Teufel 1999) like, for example, *Problem, Method,* and *Results* (= categories of scientific articles). These categories represent functions of parts of a text as an instance of a specific text type, which are oriented towards the text as a whole. They can be organised hierarchically and therefore be formally described by a hierarchical schema (e.g. Kando 1999). The text type structure schema we developed for linguistic scientific articles is shown in Figure 4. The schema is based on the ones by Kando and by Teufel, but was, as a result from our corpus analyses, adapted to our corpus and to our application scenario.

For the annotation of our corpus each text was divided in TTS segments, often, but not necessarily always, consisting of a sentence. This (more or less) sentential segmentation corresponds to the segmentation realised by Kando (1999) and Teufel (1999). An example of a sentential TTS annotation is shown in Listing 4. Apart from the sentential TTS structure which we consider as the *local* text type structure, we additionally identify a macro TTS structure, or *global* text type structure. Each scientific article can be divided into functionally coherent macro sections which can be categorised according to the set of global TTS categories (Figure 4 – global categories are indicated by the grey boxes). Empirical analyses showed that the majority of scientific articles have a more or less canonical TTS

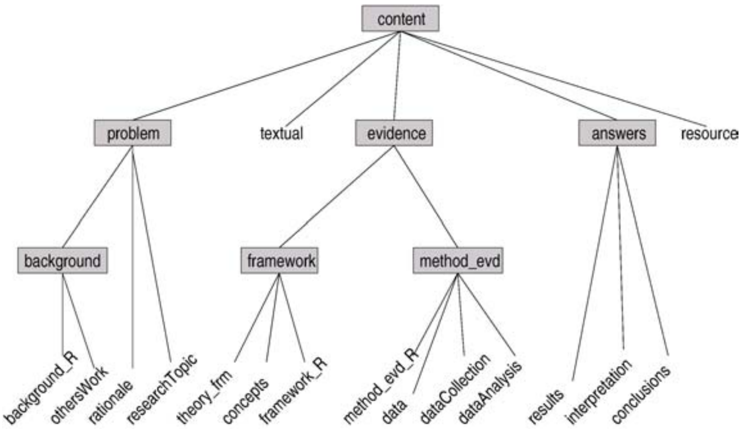


Figure 4: Text type structure (TTS) schema for scientific articles

structure which means that they are composed as a sequence of *Problem, Evidence, Answers*. Nevertheless, there are several articles which exhibit various deviations from this canonical structure. Not all global categories have to be present in scientific articles and the sequence of categories may vary, too. These articles can therefore be described as generic variations (see Section 4.3).

The role of the generic document structure for discourse analysis in the tradition of RST has lately been examined by Gruber and Muntigl (2005), and Taboada and Lavid (2003). Both approaches model the generic structure of a document as genres and stages (like *Orientation, Background, Account, Interpretation, Summary*) in the tradition of the Register and Genre Theory, i.e. as serially occurring functional stages, where each stage depends on a previous stage. Gruber and Muntigl empirically show that generic and rhetorical structure of students' academic writings coincide.[4] They found correlations between both genre dependent and independent stages, and RST relations. *Orientation*, for instance, typically occurred with PREPARATION, *Discussion* with BACKGROUND and *Summary* with SUMMARY (Gruber and Muntigl 2005, p.102). These systematic relationships between generic and rhetorical structure are differentiated for different textual levels (and generic stages), i.e. for high level textual structures (stages) as well as low levels (substages).[5]

Likewise, Taboada and Lavid provided empirical evidence for correlations between generic stages and rhetorical (and thematic) patterns in scheduling dialogues, e.g. *Opening* correlated with SOLUTIONHOOD, *Closing* with RST relations like EVALUATION, RE-

```
<segment id="s196" parent="g4" topic="results">In den Texten ist sehr oft
   nicht klar, ob ein Maskulinum nur auf Männer oder auch auf Frauen refe
   riert. </segment>
<segment id="s197" parent="g4" topic="interpretation"> Wichtige Fragen, die
   die LeserInnen an den Text haben, bleiben somit unbeantwortet. Die Politik
   wird durch den fast durchgehenden Gebrauch des generischen Maskulinums als
   "Männersache" dargestellt, Frauen werden, auch wenn sie vorhanden sind,
   selten sichtbar gemacht. Zudem wird auch mit geschlechtsspezifisch männli
   chen Wörtern wie Gründerväter der Gedanke an Männer evoziert. </segment>
```

Listing 4: TTS annotation (extract)

STATEMENT, and SUMMARY. Their intention was to use rhetorical relations as signals for a specific generic stage. Our approach works just the other way round. We intend to use the existing (so far manually assigned) generic document structure as a signal for a specific discourse structure. In our approach, three different ways of using the generic document structure as a cue or constraint for discourse interpretation are distinguished:

1. A TTS category which corresponds to an RST relation can be used as an explicit cue for a specific RST relation: As pointed out above, a text type structure category is a functional relation between a part of a text and the text as a whole, while an RST relation establishes a functional relation between two or more parts of a text (discourse segments). The category names for both types of functional relations, however, partly overlap, e.g. *Background* – BACKGROUND, *Problem* – PROBLEM-SOLUTION, *Evidence* – EVIDENCE, *Results* – RESULT, *Interpretation* – INTERPRETATION, and maybe also SUMMARY – *Conclusions*. We suspect that TTS categories are also often functions between parts of a text rather than between a part and the text as a whole. The *Answers* section of a scientific article, for example, contains an answer to what is described in the *Problem* part rather than an answer to what is described in the text as a whole. Hence, it seems reasonable to identify certain TTS categories with equivalent RST relations, e.g. an *Interpretation* TTS constituent should be an RST satellite in an INTERPRETATION relation. Empirical evidence for this hypothesis is provided by the findings of a descriptive analysis of our corpus.[6] The relation INTERPRETATION is found 17 times as frequently in TTS segments which are of type *Interpretation* than with all other TTS categories (on average), BACKGROUND occurs 9 times as frequently with *Background*, and SUMMARY 9 times as frequently with *Conclusions* (see Section 4.3).

2. Generally, a TTS category (assigned to a TTS segment) which frequently appears with RST relation A and never with relation B induces relation A with a higher probability than relation B – the TTS category can therefore be used in a statistic constraint: The quantitative analysis of our corpus – similar to the empirical research done by Gruber and Muntigl (2005) – showed high deviations from the average distribution of relations and TTS categories. Some TTS categories correlate significantly with one or two specific RST relations. The analyses and its findings will be described in greater detail in Section 4.3.

3. At the highest level of discourse structure (CDS type=”document”), the global categories of the text type structure schema (*Problem, Background, Evidence, Framework, Method, Answers*) should be determined automatically for all CDS type=”division”, so that the relations between these categories can be inferred using a relational schema (see Figure 5). This approach is based on the fact that in most cases, scientific articles are organised along a specific sequence of global generic categories. For a detailed description of the procedure of instantiating global TTS categories and relations between them see Section 4.2.

4.2. Canonical sequence of global text type structure categories

The offline instantiation of global TTS categories for all top-level divisions (DocBook: sect1-n) was based on an analysis of the size (measured in the number of tokens contained)[7] of all sentential TTS categories of one section. In the analysis, it was calculated which of the global categories comprised the largest local categories, i.e. the parent category (in the tree-structured text type structure schema shown in Figure 4) of the majority of local TTS categories in the current section was looked up. Each section was then labelled

```
<segment id="i4" topic="problem" strtype="sect1">0 Einleitung ... </segment>
<segment id="i5" topic="framework" strtype="sect1">1 Positionierung des Pro
   jektes im Forschungskontext ... </segment>
<segment id="i6" topic="researchTopic" strtype="sect1">2 Erkenntnisinteressen
   und Ziele ... </segment>
```

Listing 5: Global TTS annotation (extract)

with the global category found such that the whole article is annotated as a sequence of macro section segments with a TTS category assigned (see Listing 5).

To determine whether the canonical sequence of TTS categories expressed in the text type structural schema in Figure 3 holds for the scientific articles in our corpus, we calculated the sequences of global TTS categories of each article in the corpus. We found one basic TTS sequence (Type A), which may be described as the generic prototype, and two variations of this prototype (Types B and C):

- **Type A**: This type is regarded as the prototypical one for scientific articles. Articles of this kind exhibit a sequence of all global categories *Problem – Evidence* (which may be split into *Framework* and *Method*) – *Answers* (which may be split into *Results* and *Interpretation*), (26 articles out of 47). The initial section of an article is always *Problem* and the final one is always *Answers.* By contrast, the central part of scientific articles is much less restricted than the beginning and the end. Disruptions of the canonical order occur frequently, for example through sections annotated as *OthersWork, Background* or *Framework* (7 articles of the 26). Especially the latter three categories seem to be sequentially more variable than other categories like *Conclusions* or *Method*. In one case, an article had an initial section labelled as *Answers*, because it described the findings of the study presented. It is not unusual for sentential *Answers* segments to occur in the first section of an article, but it is atypical that these segments constitute the major part of the first section. The case is therefore treated as an (atypical) variant of A.

- **Type B**: Under this type we subsume articles which exhibit the basic sequence of global categories but with one of *Problem* or *Evidence* or *Answers* missing (12 articles). Type B is the most common alternative to Type A. In most cases (8 articles), *Problem* does not occur as an initial section of its own. Instead, TTS segments with local subcategories of *Problem* are interspersed in various sections, but in none of them do they constitute the majority of sentential TTS segments.
- **Type C**: Articles whose sections are all annotated with the global category *Problem:Background* (3 articles), or with *Evidence:Framework* (1), or articles which are annotated with an arbitrary sequence of the *Problem* and *Framework* categories (3). These structures are regarded as atypical alternatives to the canonical text type structure and are mainly found in articles about theoretical work or language politics.
- **Deviations**: There is a small subset of articles (2 articles) which show various deviations from the structures introduced in Types A – C. We treat these articles as special cases.

The high number of articles in Types A and B confirms the relevance of the text type structure schema in Figure 4 and the canonical order of categories expressed in it. As a consequence, when complex discourse segments of the highest type (CDS type=”division”) are annotated with TTS categories, relations between them can be instantiated with a certain confidence. The corresponding configurations of TTS segments thus serve as discourse structural cues as illustrated in Figure 5.

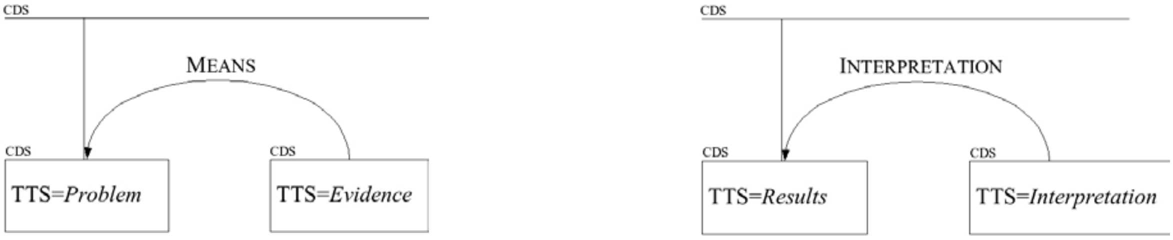


Figure 5: Rhetorical relations induced by TTS annotations of adjacent complex discourse segments

4.3. Correlations between text type structure categories and rhetorical relations

The subcorpus used in this study comprises two parts: The first consists of 10 English psychology articles with 8597 words on average. In this part (henceforth: PsyEngl), the minimal segments are elementary discourse segments (EDS), so RST relations are annotated for the EDS level and all higher levels. The second part of our corpus contains 10 German linguistic articles with 8627 words on average – this subcorpus (henceforth: LingDeu) is annotated with rhetorical relations between complex discourse segments of the type “block” and higher. On both subcorpora we examined correlations between RST relations and TTS categories. For this task, we employed the *Sekimo Tools* for the analysis of multiple XML annotations of one textual base (Witt et al., 2005). We made use of two features of the Sekimo Tools:

1. *Markup unification*: We produced a unified XML document containing both its logical document structure annotation and its TTS annotation of elementary TTS segments. By means of an XSLT style sheet, we then automatically assigned TTS categories to the *TTS block segments* of an article based on the existing TTS annotations of elementary TTS segments. The TTS category which in terms of the number of word tokens made up the largest part of a block segment was selected as the TTS category for that segment as a whole. Subsequently, adjacent block segments with identical TTS categories were joined into one TTS segment.

2. *Relation checking*: Using the Sekimo Tools, inferences about the relationship between elements on an annotation layer A and another annotation layer B of one textual base can be drawn. Possible relations between elements on different layers are *inclusion, overlapping, adjacency, independence* and others (determined in terms of the shared or unshared PCDATA element contents). Thus, taking both the newly obtained TTS block segment annotation and the RST annotation of each document, we listed the rhetorical relations between discourse segments *included* in the TTS block segment and related them to the TTS category of that segment[8].

As a result, we obtained a matrix with the relations arranged in lines and the TTS categories in columns. The matrix shows the type and the number of relations for each TTS category. The frequency of the different TTS categories and RST relations, and the number of TTS segments and included RST segments for both corpora are shown in Table 2.

In a subsequent step, we examined the distribution of relations in each TTS category and listed them as percentage values. In the TTS category *Framework*, for example, ELABORATION takes 38% of all relations in that category, JOINT 10%, CONDITION 8%, LIST 5%, etc.[9] For each relation type, we calculated the average percentage of its frequency over all TTS categories. The average frequencies of RST relations over all TTS categories are:[10]

- For PsyEngl: 30.2% ELABORATION, 13.8% LIST, 9.8% CIRCUMSTANCE, 8.2% JOINT, 7.2% PREPARATION
- For LingDeu: 28.8% PREPARATION, 27.9% LIST, 18.1% ELABORATION, 9.1% SUMMARY, 5.2% EVIDENCE

	#TTS categories used	**#TTS Segments**[11]	**Most frequent TTS categories**	**#RST relation types used**	**#RST Segments included in TTSs**	**Most frequent RST relations included in TTS**
PsyEngl corpus	17	121	Framework: 20% Results: 17% Measures: 11%	36	801	ELABORATION: 35% LIST: 9% JOINT: 8% CIRCUMSTANCE: 8%
LingDeu corpus	17	361	Data: 25% Results: 25% Framework: 12%	17	297	LIST: 33% ELABORATION: 23% PREPARATION: 23%

Table 2: Number and frequency of TTS categories and RST relations in the corpora

To find out which RST relations are more prominent in TTS category A than in TTS category B, we compared the average percentage with the actual percentage of a relation for a specific TTS category. For example, the number of CONDITION relations (or more precisely: RST segments related through CONDITION) amounts on average to 1.44 % of all relations occurring within a TTS category. In *Framework*, the percentage of CONDITION relations amounts to 8%. To calculate the difference between the average distribution of a relation (over all TTS categories) and the actual distribution of a relation for a specific category (in the current example, *Framework*), we divide the actual percentage by the average percentage, e.g. 8 / 1.44. The result of this calculation is a factor (in the following: Factor D), which describes the deviation of the frequency of a relation for a specific TTS category from the average distribution of this relation over all TTS categories. To return to our example, CONDITION can be found 5.6 times more often in *Framework* than (on average) in all other TTS categories. The results of all these calculations are given in Figure 6. The diagram shows the different distributions of RST relations. The peaks in the graphs indicate that a TTS category can be clearly distinguished by a different distribution of RST relations. All TTS categories have special characteristics with respect to the occurrences and frequencies of RST relations – some relations are found up to 17 times more often in a specific TTS category than in any other one.

To verify the results, we additionally calculated the deviation of the expected frequency of a relation at a specific TTS category from its actual frequency. We assumed that if a TTS category takes the percentage X of all included RST segments/ relations of a corpus, it could be expected that this category would take the same percentage X of all included instances of relation R. For example, the amount of RST segments included in TTS segments assigned as *Framework* is 202 (in the PsyEngl corpus). As the whole corpus comprises 801 included RST relation instances (i.e. segments with an RST relation assigned), Framework holds 201 / 801 (= 25%) of all included RST relation instances. One could therefore expect that 25% of all included ELABORATION instances in the corpus would be found in TTS segments categorised as *Framework*, i.e. 25% of 282 (in the PsyEngl corpus) = 71. As before, the difference between the expected number of relation instances and the actual number can be described by a factor (henceforth: Factor EA) which is the result of the division of the actual number (e.g. 76) by the expected number (e.g. 71), in this case 1.1.

Some of the factors we obtained were extremely high. Very often this was due to only a small number of included relation instances of a type in the corpus. For this reason, we ignored those relations which have less than 10 included instances in both subcorpora. Due to the different number of RST instances in the corpora, the number of different RST relations with 10 or more included instances varies across the corpora. In the LingDeu corpus, only 10 RST relations have more than 10 included instances, whereas 19 relations have more than 10 included instances in the PsyEngl corpus.

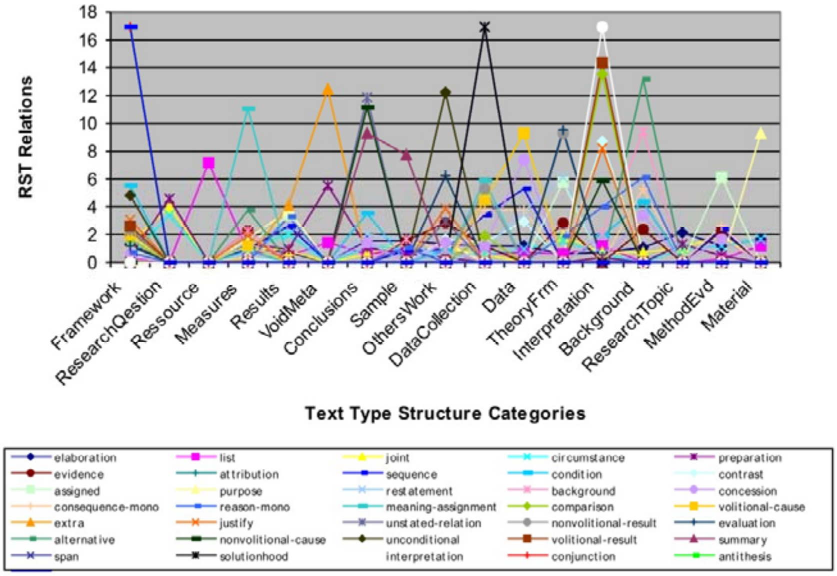


Figure 6: Distribution of RST relations over the different TTS categories for the PsyEngl corpus

LINGDEU/ **PSYENGL**	**FW**	**MES**	**RES**	**CON**	**SMP**	**OW**	**DC**	**DT**	**INT**	**BCK**	**RT**	**ME**	**TF**
ELABORATION	2.2/ ...										**5.5 / 2.1** **4.4 / …**		
LIST								2.2 / …					
EVIDENCE					4.8 / … 5.3 / …	…/ 2.8 …/ 4.9							…/ 2.8
SUMMARY		**11.0/…** **16.5/…**		**… / 9.3** **…/ 10.3**	**2.8/ 7.7** **4.1/ 8.5**		… / … 2.8 / …						
CONTRAST							7.5 / … 4.9 / …	**7.6 / 2.9** **5.1 / 2.1**	**… / 8.7** **…/ 6.4**				
CONSEQUENCE-MONO				**6.7 /...** **5.5 /…**			... / 3.7 … / 3.2	... / 3.8 …/ 3.3		**9.2 / 5.2** **7.6 / 4.6**		…/ 2.5 … / 2.2	
REASON-MONO			… / 3.3 …/ 4.9						... / 3.9 …/ 5.9	... / 6.1 …/ 9.1			… / … … / 2.8
CONCESSION						**8.0 / …** **7.6 / …**		**… / 7.4** **…/ 9.1**		**8.0 / 3.4** **7.6 / 4.2**			
BACKGROUND				4.5 / … 5.5 / …		**6.2 / ...** **7.6 / …**				**6.2 / 9.4** **7.6 / 8.3**			
ASSIGNED			…/ 3.4 …/ 3.8									**…/ 6.1** **…/ 6.9**	**…/ 5.8** **…/ 6.6**
MEANING-ASSIGNMENT		... / 11.1 …/ 3.7					... / 5.9						

SEQUENCE			.../ 2.6 …/ 2.2				.. / 3.5 …/ 2.9	… / 5.3 …/ 4.6					
ATTRIBUTION						…/ 2.8	…/ 5.0 …/ 3.4				…/ 3.4 … / 2.3		
CONDITION	…/ 5.5 …/ 2.8			…/ 3.6					…/ 4.3				
RESTATEMENT								... / 3.6 …/ 3.0					**…/ 6.1** **… / 5.1**

Legend:						
	FW = *Framework*	RES = *Results*	SMP = *Sample*	DC = *DataCollection*	INT = *Interpretation*	ME = *MethodEvd*
	MES = *Measures*	CON = *Conclusions*	OW = *OthersWork*	DT = *Data*	BCK = *Background*	TF = *TheoryFrm*

Table 3: Correlations between TTS categories and RST relations

In Table 3, the correlations between the highly frequent RST relations and TTS categories are shown. In the first line of each cell, the deviation of the frequency of a relation for a specific TTS category from its average distribution is represented by Factor D introduced above (only Ds which are higher than 2.0 are shown). In the second line of each cell, the difference between the expected number of a relation at a specific TTS category and its actual number is indicated by the Factor EA (only EAs higher than 2.0). The first alternative always refers to the LingDeu corpus, the second one to the PsyEngl corpus. Numbers in bold type indicate the most significant results.

It is remarkable that the findings for the two corpora are only partly overlapping. The reason for the differences could be either the different sizes of the minimal discourse segments (EDS for PsyEngl, CDS type=”block” for LingDeu), or the domain, or even a language-specific style of discourse organisation. One or all of these factors seem to influence the prominence of TTS categories and RST relations, e.g. English psychology articles contain many more segments of the TTS category *Measures* than German linguistic articles, whereas e.g. the RST relation PREPARATION is much more common in German linguistic articles. Therefore, it may be problematic to transfer the findings to scientific articles from other languages, domains and/ or of different segment granularity. However, some of the correlations of TTS categories and RST relations are similar for both corpora and have therefore high empirical evidence. These are *ResearchTopic* – ELABORATION, *Sample* – SUMMARY, *Data* – CONTRAST, *Background* – CONSEQUENCE-MONO, *Background* – CONCESSION, and *Background* – BACKGROUND, cf. Table 3.

Apart from the correlations that both corpora exhibit, correlations that hold only in one corpus can be found. The clearest correlations of RST relations and TTS categories are those where both factors (Factor EA and Factor D) are higher than 5.0. For the LingDeu corpus, these are *Measures* – SUMMARY, *Conclusions* – CONSEQUENCE-MONO, *OthersWork* – CONCESSION, and *OthersWork* – BACKGROUND. For PsyEngl, the most prominent correlations are *Conclusions* – SUMMARY, *Data* – CONCESSION, *Interpretation* – CONTRAST, *MethodEvd* – ASSIGNED, *TheoryFrm* – ASSIGNED, and *TheoryFrm* – RESTATEMENT.

Provided that the assignment of TTS categories to CDS segments of a scientific article can be done automatically (as has so far been tested and reported in Langer et al. 2004), the factors D calculated from our corpus can be employed as statistic constraints for the assignment or disambiguation of RST relations to segments pairs included in the CDS. Conversely, the factors EA could be used if TTS assignment were the main goal of analysis using RST annotation as an auxiliary analysis. In our discourse parser, the former strategy will be pursued.

5. Conclusion and outlook

The aim of the study presented was to examine what kind of cues and constraints for discourse interpretation can be derived from the logical and generic document structure of complex texts such as scientific journal articles. Consequently, we performed several analyses on a corpus of scientific articles that is annotated on different XML annotation layers: The XML annotation of the logical document structure is realised by using an (extended) subset of the DocBook DTD (Walsh and Muellner 1999). The generic document structure (text type structure) is encoded using an XML schema based on the text type structure schema for scientific articles shown in Figure 4. Moreover, XML annotations of RST analyses of several articles were provided. So far, the texts of the corpus are annotated manually and semi-automatically.[12] The XML-based multi-layer annotation approach (Witt et al. 2005) was used to examine dependencies between XML elements on different annotation layers.

In Section 3, we argued that logical structure elements like *title, listitem, glossterm* can serve as cues in automated discourse analysis just like traditionally lexical discourse markers such as conjunctions and sentential adverbs. On the other hand, we introduced the discourse segment types EDS, SDS, and CDS (with its subtypes) and pointed out how they can be used as constraints to narrow down the textual domains inside which to identify relation spans. With regard to generic document structure, in Section 4.2 we showed that in our corpus a canonical sequence of text type structure categories occurring in the majority of articles can be established. Moreover, most deviations from this sequence could be grouped into two types. With a certain confidence, such sequences or partial sequences can be used as cues to assign discourse relations along relational schemas as presented in Figure 5. Finally, in 4.3, we demonstrated in a corpus analysis how and which TTS categories assigned to complex discourse segments of type “block” statistically constrain the occurrence of rhetorical relation types.

In the future, we will work on integrating the cues and constraints described in this study into a discourse parser that takes several XML annotation layers of the same text as its input and provides a new XML annotation layer containing the discourse analysis as its output (Lüngen, et al. 2006). So far, the parser uses lexical discourse markers and several grammatical features as cues for relation assignment. The parser consists of cascaded iterative applications of a bottom-up chart parser and is realised in Prolog, and the discourse cues are encoded in the form of its *reduce rules*. These are derived from a discourse marker lexicon and also make reference to a syntax and morphology XML annotation layer which is generated using the commercial *Machinese Syntax* tagger software from Connexor Oy. Using this architecture, further reduce rules that make reference to the logical and the generic structure annotation layers of a document will be generated from the representations of the results of this study and integrated into the parser's rule files. Evaluations as to the contribution of this type of rules to the overall results will be provided in due course.

Notes

[1] The examples in this article are taken from:
Baßler, H. and H. Spiekermann (2001). Dialekt und Standardsprache im DaF-Unterricht. Wie Schüler urteilen – wie Lehrer urteilen. In: *Linguistik Online*, 9; Bühlmann, R. (2002). Ehefrau Vreni haucht ihm ins Ohr... Untersuchungen zur geschlechtergerechten Sprache und zur Darstellung von Frauen in Deutschschweizer Tageszeitungen. In: *Linguistik Online*, 11; Bärenfänger, O. and S. Beyer (2001). Zur Funktion der mündlichen L2-Produktion und zu den damit verbundene kognitiven Prozessen für den Erwerb der fremdsprachlichen Sprechfertigkeit. Linguistik Online, 8.

[2] The whole corpus comprises 120 scientific articles in different languages (English and German), domains (psychology and linguistics) and sub-genres (experimental and review); this corpus is split in two subcorpora: PsyEngl (English psychology articles) and LingDeu (German linguistic articles). Part of the work described in this article (e.g. the canonical sequence of global text type structure categories) is based on the smaller LingDeu corpus of 47 linguistic articles.

[3] The application scenario of our project is an online-system which supports novice learners (first or second year students) in selective and efficient reading of scientific articles, and which can furthermore be used as a learning environment where students may learn something about the structural and argumentative characteristics of the genre "scientific article". In order to personalise the system and to allow students to upload scientific articles, a discourse parser is being developed which implements the task of automatically adding discourse structure annotations.

[4] Their corpus consists of 19 student academic term papers (lengths ranging between 1865 and 7271 words). For the annotation, 35 RST relations were used and 46 genre stage categories.

[5] Gruber and Muntigl's relational discourse analysis based on RST was restricted to the level of subchapters; clauses were not annotated.

[6] 2 x 10 texts of our corpus were analysed: For the 10 German linguistic articles, RST annotations were done on the level of paragraphs (CDS type="block") as minimal units, for the 10 English psychology articles RST annotations were done with clauses (EDS) as minimal units.

[7] We would like to thank Mario Klapper who did most of the manual XML annotations and numeric analyses for the work presented in this section of the paper.

[8] The RST annotation also contains numerous relation instances *not* included in a TTS segment, i.e. relations between segments of larger types than CDS type=”block”.

[9] This distribution of RST relations can be found in the PsyEngl corpus.

[10] These percentages do *not* refer to the number of RST relations in the corpus, but to the average distribution of relations in each TTS category.

[11] Due to their small size, some TTS segments do not contain any RST segments.

[12] The DocBook-annotation was produced semi-automatically, the annotation of the text type structure manually. A first approach to the automatic assignment of TTS categories to text segments in our corpus yielded mixed results: The recall and precision figures for some TTS categories were good, while recall for other categories was poor (Langer et al. 2004).

References

1 Asher, N. and A. Lascarides (2003). *Logics of Conversation*. Cambridge, U.K.: Cambridge University Press.

2 Blakemore, Diane (2004): Discourse Markers. In: Horn, L. R. and G. Ward (Ed.): *The Handbook of Pragmatics*. Oxford: Blackwell, p. 221-240.

3 Carlson, L. and D. Marcu (2001). *Discourse tagging reference manual.* Technical Report ISI-TR-545. Marina del Rey CA: Information Science Institute.

4 Corston-Oliver, Simon (1998). *Computing of Representations of the Structure of Written Discourse*. PhD thesis, University of California, Santa Barbara.

5 Gruber, H. and P. Muntigl (2005). Generic and Rhetorical Structures of Texts: Two Sides of the Same Coin? In: *Folia Linguistica* XXXIX (1-2). Special Issue: Approaches to Genre. Berlin: Mouton de Gruyter, 75-114.

6 Kando, N. (1999). Text structure analysis as a tool to make retrieved documents usable. In: *Proceedings of the 4th International Workshop on Information Retrieval with Asian Languages*, Taipei, Taiwan, 26-135.

7 Holler, A. (2003). *Spezifikation für ein Annotationsschema für Koreferenzphänomene im Hinblick auf Hypertextualisierungsstrategien*. [http://www.hytex.uni-dortmund.de/hytex/publikationen.html#Dokus]

8 Langer, H., H. Lüngen, and P. S. Bayerl (2004). Towards automatic annotation of text type structure: Experiments using an XML-annotated corpus and automatic text classification methods. In: *Proceedings of the workshop on XML-based richly annotated corpora (XBRAC) at the LREC 2004*. Lissabon, 8-14.

9 LeThanh, H., G. Abeysinghe, and C. Huyck (2004). Generating Discourse Structures for Written Texts. In: *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland.

10 Lobin, H., M. Bärenfänger, M. Hilbert, H. Lüngen, and C. Puskas (2006, to appear). Discourse relations and document structure. In: Metzing, D. and A. Witt (Ed.): *Linguistic modeling of information and Markup Languages, Contributions to language technology.* (Series Text, Speech and Language Technology). Dordrecht: Springer.

11 Lüngen, H., M. Bärenfänger, M. Hilbert, H. Lobin, and C. Puskas (2006). Text parsing of a complex genre. In: *Proceedings of the Conference on Electronic Publishing (ELPUB),* Bansko, Bulgarien.

12 Mann, W. C. and S. A. Thompson (1988). Rhetorical Structure Theory: Toward a functional theory of text organisation. In: *Text* 8(3), 243-281.

13 Mann, W. C. and Taboada, M. (2005). *Rhetorical Structure Theory. Relation Definitions*. [http://www.sfu.ca/rst/01intro/definitions.html]

14 Marcu, D. (2000). *The theory and practice of discourse parsing and summarization*. Cambridge, MA: MIT Press.

15 O'Donnell, M. (2000): RSTTool 2.4 .A markup tool for Rhetorical Structure Theory. In: *Proceedings of the International Natural Language Generation Conference (INLG'2000).* Mitzpe Ramon, Israel, 253-256.

16 Polanyi, L., C. Culy, M. van den Berg, G. L. Thione, and D. Ahn (2004a). A rule based approach to discourse parsing. In: *Proceedings of the 5th Workshop in Discourse and Dialogue*, Cambridge, MA., 108-117.

17 Polanyi, L., C. Culy, M. van den Berg, G. L. Thione, and D. Ahn (2004b). Sentential structure and discourse parsing. In: *Proceedings of the ACL 2004 Workshop on Discourse Annotation*, Barcelona, 49-56.

18 Power, R., D. Scott, and N. Bouayad-Agha (2003). Document structure. In: *Computational Linguistics,* 29(2), 211-260.

19 Reitter, D. (2003): Simple signals for complex rhetorics: On rhetorical analysis with rich-feature support vector models. In: Uta Seewald-Heeg (ed.): *Sprachtechnologie für die multilinguale Kommunikation. Textproduktion, Recherche, Übersetzung, Lokalisierung. Beiträge der GLDV-Frühjahrstagung 2003*. Volume 18 of LDV-Forum, p. 38-52.

20 Swales, J. M. (1990). *Genre Analysis. English in academic and research settings.* Cambridge, UK: Cambridge University Press.

21 Taboada, M. and J. Lavid (2003). Rhetorical and thematic patterns in scheduling dialogues. In: *Functions of Language* 10(2), 147-148.

22 Teufel, S. (1999). *Argumentative Zoning: Information Extraction from Scientific Text.* Ph. D. thesis, University of Edinburgh.

23 van Dijk, T. A. (1980). *Macrostructures: An interdisciplinary study of global structures in discourse, interaction, and cognition.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

24 Walsh, N. and L. Muellner (1999). *DocBook: The Definitive Guide,* O'Reilly.

25 Witt, A., H. Lüngen, H., D. Goecke, and F. Sasaki (2005). Unification of XML documents with concurrent markup. In: *Literary and Linguistic Computing*, 20(1), 103-116.