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

Maja Bärenfänger, Harald Lüngen, Mirco Hilbert & Henning Lobin

This study examines what kind of cues and constraints for discourse interpretation can be derived from the logical and generic document structure of complex texts by the example of scientific journal articles. We performed statistical analysis on a corpus of scientific articles annotated on different annotations layers within the framework of XML-based multi-layer annotation. We introduce different discourse segment types that constrain the textual domains in which to identify rhetorical relation spans, and we show how a canonical sequence of text type structure categories is derived from the corpus annotations. Finally, we demonstrate how and which text type structure categories assigned to complex discourse segments of the type “block” statistically constrain the occurrence of rhetorical relation types.

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 & 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 collection. 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 & Thompson 1988; Marcu 2000) shares three basic assumptions with other linguistic discourse theories like Segmented Discourse Representation Theory (SDRT, Asher & 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 & Thompson 1988: 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 & Thompson 1988: 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).
Table 1. Definition of the relation Concession (Mann & Thompson 1988, p.254ff.)

<table>
<thead>
<tr>
<th>relation name:</th>
<th>Concession</th>
</tr>
</thead>
<tbody>
<tr>
<td>constraints on N:</td>
<td>W (the writer) has positive regard for the situation presented in N</td>
</tr>
<tr>
<td>constraints on S:</td>
<td>W is not claiming that the situation presented in S doesn’t hold</td>
</tr>
<tr>
<td>the N + S combination:</td>
<td>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</td>
</tr>
<tr>
<td>the effect:</td>
<td>R’s (the reader’s) positive regard for the situation presented in N is increased</td>
</tr>
<tr>
<td>locus of the effect:</td>
<td>N and S</td>
</tr>
<tr>
<td>discourse markers:</td>
<td>in S: obwohl, obschon, obgleich, obzwar, zwar [although, though …]</td>
</tr>
<tr>
<td></td>
<td>in N: dennoch, doch, trotzdem [however, anyhow, nevertheless …]</td>
</tr>
</tbody>
</table>

Listing 1. Discourse marker for Concession

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 & 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: 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), parallel syntactic constructions which may

1. The examples in this article are taken from:
induce a List-relation, and punctuation marks such as ‘:’ (colon), 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 & 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. 2010; Lüngen et al. 2006), we deal with a corpus of scientific articles 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 discourse 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.

---


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.
Originally, RST provides a set of 26 rhetorical relations, which are either mono- or multinuclear (Mann & 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 & Thompson 1988: 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 & Maier 1995; and Carlson & 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 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 with approximately 172,240 words 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: 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. 2010). The physical layout structure of a 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 & 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, glossdefs 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.

---

Figure 1. RST annotation for extract in Listing 2

Listing 2. DocBook annotation (extract)


Figure 2. RST annotation for extract in Listing 3

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 top node, and SDS (sentential discourse segments)
act as the basic segments for the construction of discourse subtrees.

- CDS type = "division": Complex discourse segments of the type "division"
correspond to the lowest section level. In terms of DocBook markup it com-
prises 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 docu-
ment. 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.

- CDS type = "document": This segment type comprises all residual sect ele-
ments, 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 sec-
tions. 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 com-
parable 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 seg-
ments, 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".4 By assuming that the rhetorical structure correlates
with the logical document structure, or, as Marcu (2000: 109) says, "that sentences,

4. This constraint is of course an idealization to some degree. In the annotation process, for
example, we noticed that sometimes the last sentence(s) of one paragraph serves as a preparation
for the content of the following paragraph(s) so that it seems to be rhetorically closer to the
following paragraphs.
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. The global TTS categories for scientific articles are Problem, Background, Evidence (Framework and Method) and Answers. These categories can be further distinguished in OthersWork, ResearchTopic, Theory, Data, Results etc. In an earlier version, the schema consisted of 135 categories. When building the reduced schema with only 17 leaf categories that was used in this study, several of the original more fine-grained categories were combined to form broader categories. A suffix "_R" in the category label in Figure 4 indicates that the present category comprises several (residual) categories that were also found under the same super-category in the original schema.5

5. See Bayerl et al. (2003) for more information about the extended schema.
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 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).

Listing 4. TTS annotation (extract)
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. 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 & Muntigl 2005: 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).

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, Restatement, 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

6. 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.

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

![Figure 5](image-url)  

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

---

8. 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.
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) 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 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).

Listing 5. Global TTS annotation (extract)

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

9. 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.
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.

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**: For each article, 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.10

As a result, we obtained a contingency 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. Table 2 shows the frequencies of various TTS categories and RST relations, and the number of TTS segments and included RST segments for both. The percentages of the most frequently occurring TTS categories in terms of their share in the total number of TTS segments (121 in the PsyEngl and 361 in the LingDeu subcorpus) are *Framework* (20%), *Results* (17%) and *Measures* (11%) in PsyEngl, and *Data* (25%), *Results* (25%) and *Framework* (12%) in LingDeu.

---

10. 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".
# Table 2. Number and frequency of TTS categories and RST relations in the corpora

<table>
<thead>
<tr>
<th>#TTS category segments</th>
<th>#TTS Most frequent TTS segment categories</th>
<th>#RST Most frequent TTS segments included in TTSs</th>
<th>Most frequent RST relations included in TTS (as p(TTS))</th>
<th>Most frequent RST relations included in TTS (as p(Rel))</th>
</tr>
</thead>
<tbody>
<tr>
<td>PsyEngl corpus</td>
<td>17</td>
<td>121</td>
<td>Framework: 36</td>
<td>p(Framework) = 25%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Results: 20%</td>
<td>p(Measures) = 21%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Measures: 17%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Data: 11%</td>
<td>p(DataCollection) = 11%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Results: 25%</td>
<td>p(Results) = 37%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Framework: 12%</td>
<td>p(Framework) = 10%</td>
</tr>
<tr>
<td></td>
<td>36</td>
<td>801</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>17%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>11%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>12%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LingDeu corpus</td>
<td>17</td>
<td>361</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>801</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>14%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>35%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>9%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>23%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For our following analyses, we took the total number of RST segments included in a TTS segment and annotated with an RST relation (801 for PsyEngl and 297 for LingDeu) to calculate the percentages that express the estimated a priori probabilities \( p(TTS) \) (the probability of an RST segment to be included in a specific TTS) and \( p(\text{Rel}) \) (the probability of an RST Segment to be annotated with relation Rel). The highest percentages for different TTS are \( p(\text{Framework}) = (202/801) \times 100 = 25\% \), \( p(\text{Measures}) = 21\% \), and \( p(\text{DataCollection}) = 11\% \) in PsyEngl, and \( p(\text{Results}) = 37\% \), \( p(\text{Data}) = 14\% \) and \( p(\text{Framework}) = 10\% \) in LingDeu. The percentages or a priori probabilities \( p(\text{Rel}) \) for the most frequent RST relations are \( p(\text{Elaboration}) = 35\% \), \( p(\text{List}) = 9\% \), \( p(\text{Joint}) = 8\% \) and \( p(\text{Circumstance}) = 8\% \) in PsyEngl, and \( p(\text{List}) = 33\% \), \( p(\text{Elaboration}) = 23\% \) and \( p(\text{Preparation}) = 23\% \) in LingDeu.

In a subsequent step, we examined the distribution of relations in each TTS category and listed them as percentage values which correspond to the conditional probability for the occurrence of an RST Segment with relation \( \text{Rel} \) given a Text Type Structure category \( \text{TTS} \), i.e. \( p(\text{Rel} | \text{TTS}) \). For example, \( p(\text{Elaboration} | \text{Framework}) = 38\% \), \( p(\text{Joint} | \text{Framework}) = 10\% \), \( p(\text{Condition} | \text{Framework}) = 8\% \), and \( p(\text{List} | \text{Framework}) = 5\% \) in the PsyEngl corpus. For each relation type, we then calculated the average percentage of its

---

11. Due to their small size, some TTS segments do not contain any RST segments.
frequency (afp) over all TTS categories. The average frequencies of RST relations over all TTS categories are:\textsuperscript{12}

- ForPsyEngl: \textit{afp(ELABORATION)} = 30.2\%, \textit{afp(LIST)} = 13.8\%, \textit{afp(CIRCUMSTANCE)} = 9.8\%, \textit{afp(JOINT)} = 8.2\%, \textit{afp(PREPARATION)} = 7.2\%
- For LingDeu: \textit{afp(PREPARATION)} = 28.8\%, \textit{afp(LIST)} = 27.9\%, \textit{afp(ELABORATION)} = 18.1\%, \textit{afp(SUMMARY)} = 9.1\%, \textit{afp(EVIDENCE)} = 5.2\%

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 each relation \textit{Rel} for a specific TTS category by calculating the deviation factor \(D(\textit{Rel}, \textit{TTS}) = \frac{p(\textit{Rel} | \textit{TTS})}{afp(\textit{Rel})}\).

For example, \(D(\textit{CONDITION}, \textit{Framework}) = 8/1.44 = 5.6\), i.e. \textit{CONDITION} can be found 5.6 times more often in \textit{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.

\textbf{Figure 6.} Distribution of RST relations over the different TTS categories for the PsyEngl corpus

\textsuperscript{12} These percentages do \textit{not} refer to the number of RST relations in the corpus, but to the average distribution of relations in each TTS category.
Similarly, we calculated the deviation of the expected frequency of a TTS category at a specific RST segment from its actual frequency by calculating the factor \( \text{E}(TTS, \text{Rel}) : = p(TTS|\text{Rel})/p(TTS) \). For example, the percentage of \text{ELABORATION} segments under \text{Framework} w.r.t. the total number of \text{ELABORATION} segments is \( p(\text{Framework}|\text{ELABORATION}) = (76/282) \times 100 = 27\% \). The factor \( \text{E}(\text{Framework}, \text{ELABORATION}) = 27/25 = 1.1 \) describes the deviation between the expected number and the actual number of segments under \text{Framework} that are annotated with \text{ELABORATION}.

Some of the factors we obtained were extremely high. 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 two 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.

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 \text{Measures} than German linguistic articles, whereas e.g. the RST relation \text{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 \text{ResearchTopic} – \text{ELABORATION}, \text{Sample} – \text{SUMMARY}, \text{Data} – \text{CONTRAST}, \text{Background} – \text{CONSEQUENCE-MONO}, \text{Background} – \text{CONCESSION}, and \text{Background} – \text{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 E and Factor D) are higher than 5.0. For the LingDeu corpus, these are \text{Measures} – \text{SUMMARY}, \text{Conclusions} – \text{CONSEQUENCE-MONO}, \text{OthersWork} – \text{CONCESSION}, and \text{OthersWork} – \text{BACKGROUND}. For PsyEngl, the most prominent correlations are \text{Conclusions} – \text{SUMMARY}, \text{Data} – \text{CONCESSION}, \text{Interpretation} – \text{CONTRAST}, \text{MethodEvd} – \text{ASSIGNED}, \text{TheoryFrm} – \text{ASSIGNED}, and \text{TheoryFrm} – \text{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
Table 3. Correlations between TTS categories and RST relations

<table>
<thead>
<tr>
<th>LINGDEU/PSYENGL</th>
<th>FW</th>
<th>MES</th>
<th>RES</th>
<th>CON</th>
<th>SMP</th>
<th>OW</th>
<th>DC</th>
<th>DT</th>
<th>INT</th>
<th>BCK</th>
<th>RT</th>
<th>ME</th>
<th>TF</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Elaboration</strong></td>
<td>2.2/...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5.5/2.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>List</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.4/...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Evidence</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.8/...</td>
<td>2.8/2.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5.3/...</td>
<td>4.9/2.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Summary</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>11.0/...</td>
<td>9.3/7.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>16.5/...</td>
<td>10.3/8.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Contrast</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7.5/...</td>
<td>7.6/2.9</td>
<td>8.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.9/...</td>
<td>5.1/2.1</td>
<td>6.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Consequence-mono</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6.7/...</td>
<td>3.7/3.8</td>
<td>9.2/5.2</td>
<td>2.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5.5/...</td>
<td>3.2/3.3</td>
<td>7.6/4.6</td>
<td>2.2</td>
<td></td>
</tr>
<tr>
<td><strong>Reason-mono</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.../3.3</td>
<td>.../3.9</td>
<td>6.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.../4.9</td>
<td>.../5.9</td>
<td>9.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Concession</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8.0/...</td>
<td>7.4/3.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7.6/...</td>
<td>9.1/4.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BACKGROUND</td>
<td>4.5/...</td>
<td>6.2/...</td>
<td>6.2/9.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5.5/...</td>
<td>7.6/...</td>
<td>7.6/8.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assigned</td>
<td>.../3.4</td>
<td>.../6.1</td>
<td>.../5.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.../3.8</td>
<td>.../6.9</td>
<td>.../6.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Meaning-assignment</td>
<td>.../11.1</td>
<td>.../5.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.../3.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sequence</td>
<td>.../2.6</td>
<td>.../3.5</td>
<td>.../5.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.../2.2</td>
<td>.../2.9</td>
<td>.../4.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attribution</td>
<td>.../2.8</td>
<td>.../5.0</td>
<td>.../3.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.../3.4</td>
<td></td>
<td>.../2.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Condition</td>
<td>.../5.5</td>
<td>.../3.6</td>
<td>.../4.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.../2.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Restatement</td>
<td>.../3.6</td>
<td></td>
<td>.../6.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.../3.0</td>
<td></td>
<td>.../5.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Legend:

FW = Framework  RES = Results  SMP = Sample  DC = Data-Collection  INT = Interpretation  ME = Method-Evd
MES = Measures  CON = Conclusions  OW = Others-Work  DT = Data  BCK = Background  TF = Theory-Frm
constraints for the assignment or disambiguation of RST relations to segments pairs included in the CDS. Conversely, the factors E 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 & 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. 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

---

13. 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).
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.

References


