

Cues and constraints for the relational discourse analysis of complex text types – the role of logical and generic document structure

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

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|mirco.hilbert|henning.lobin|
harald.luengen|csilla.puskas}@uni-giessen.de

1. Introduction

Relational discourse analysis, which is based on Rhetorical Structure Theory (RST, Mann and Thompson 1988) does not, in contrast to Segmented Discourse Representation Theory (SDRT, Asher and Lascarides 2003), require a fully-fledged semantic representation. Instead, it relies on linguistic properties like syntactic or lexical features, which play the role of *discourse markers* or cues for the assignment of a discourse relation to two or more discourse segments, which are labeled either as nucleus or satellite. Examples of such discourse markers are connectives like “weil” (= “because”), which marks the satellite of a CAUSE-relation, parallel syntactic constructions which may induce a LIST-relation, and punctuation marks like a colon, which, appearing at the end of a segment, may signal a PREPARATION-relation.

This surface-oriented approach often works well for simple text types, e.g. newspaper articles, which are characterized by a limited size and a relatively simple document and syntactic structure (Marcu 2000, Carlson and Marcu 2001, Reitter and Stede 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 constraints¹ for the interpretation of higher levels of discourse structure, and discourse structures which are not indicated by lexical or syntactic discourse markers (e.g. ELABORATION or BACKGROUND).

In our project (Lobin et al. 2006; Lüngen et al. 2006), we are dealing with a corpus of scientific articles² which exhibit a highly complex document structure and a relatively large average size (~ 8600 words per article). These articles are characterized by a deeply nested hierarchical structure 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 CDS. As a consequence, the majority of discourse segments are not elementary, but complex – and this means that lexical and syntactic discourse markers can only be applied in a very limited way. Apart from their complex document (and discourse) structure, our corpus of scientific articles is characterized by a high frequency of ELABORATION relations – which are usually not indicated by lexical or syntactic discourse markers.

For the interpretation of the discourse structure of scientific articles we use the analyses of the logical and generic document structure as well as the thematic structure (lexical chains, anaphoric structure) as additional knowledge sources for the assignment of rhetorical relations. The analysis of the thematic structure is especially valuable for the instantiation of certain “subject matter” relations (Mann and Thompson 1988) like ELABORATION and its subtypes (for the analyses of our corpus, we introduced various subtypes of ELABORATION, see Section 2). An anaphor, for example, is a strong discourse marker for ELABORATION. If the anaphoric link is set through certain types of *bridging*³, an ELABORATION-DERIVATION relation is induced, while referential *identity* between an anaphor and its antecedent marks ELABORATION-IDENTITY. In the following, we will concentrate on constraints and cues that can be gained from the logical and generic document structure. These will be discussed in Section 3 and 4.

2. Set of rhetorical relations

Originally, RST provides a set of 26 rhetorical relations, which are either mono- or multinuclear (Mann and Thompson 1988). This set was intended as text type- and application-independent, as well as open and extendable. For our corpus of scientific articles, we work with the relation set proposed by Mann and Taboada (2005) but extended and restructured it – by defining our own relation taxonomy (as done by Hovy and Maier 1995, and Carlson and Marcu 2001). One modification of the relation set consists in the introduction of a comprehensive sub-taxonomy of ELABORATION relations. This extension is grounded in the fact that ELABORATION is – as corpus analyses have shown – the second most common relation in our corpus. To enable a finer grained annotation, it is therefore necessary to differentiate ELABORATION through subtypes. These subtypes are ELABORATION-DERIVATION (which comprises relations like set-member, class- instance, whole-part,

```

<glosslist>
  <glossentry>
    <glossterm>A. Dialekte: </glossterm>
    <glossdef>
      <para>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.
      </para>
    </glossdef>
  </glossentry>
</glosslist>

```

Listing 1: DocBook annotation (extract)

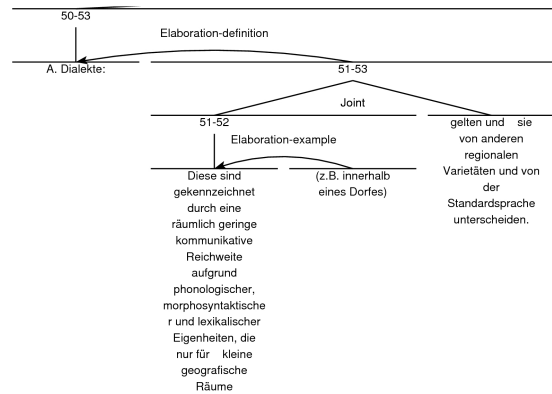


Figure 1: RST annotation for extract in Listing 1

etc.), ELABORATION-IDENTITY (referential identity between an entity in the nucleus and the satellite), ELABORATION-DEFINITION, ELABORATION-SPECIFICATION (which is similar to the well-known relation ELABORATION-OBJECT-ATTRIBUTE), and others. A second characteristic of our relation set consists of another differentiation of relations in subrelations – whenever it seemed relevant for our application scenario and when it was possible to distinguish different types of discourse markers for one relation, a new subrelation was introduced. For the relation LIST, for example, we distinguish between LIST-COORDINATION, that is induced by syntactic coordination, and LIST-DM_OTHER which may come about with discourse markers from the logical document structure, like *listitem*-elements (as shown in Listing 2 and Figure 2⁴). In a similar way, we introduced PREPARATION-TITLE (which is induced by a *title*-element in the logical document structure), PREPARATION-QUESTION, and PREPARATION-OTHER (which may come along, for example, with a colon). A third characteristic of our relation set consists in the introduction of new superordinate relation classes for relations that were often confused by human annotators because of the lack of lexical or syntactic discourse markers. These relation classes are SUPPORT-OTHER (which subsumes EVIDENCE and JUSTIFY), INTERPRETATION-EVALUATION (which acts as a superclass for INTERPRETATION and EVALUATION), and CONTRAST (which integrates CONTRAST-MULTI and ANTITHESIS). Fourthly, as in Carlson and Marcu (2001), we also distinguish subrelations on account of alternative nuclearity, for example PURPOSE-S and PURPOSE-N, or CONSEQUENCE-MULTI, CONSEQUENCE-S and CONSEQUENCE-N.

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”. These 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). For us, this *physical layout* structure encodes the *logical* document structure (according to DocBook markup, cp. Walsh and Muellner 1999), as the physical layout assigns functions to structural parts of text, e.g. a heading serves as a preparative element which orients the reader and directs his expectations for the adjacent section, whereas a figure may act as an illustration). At the level of the logical document structure, we can distinguish elementary and complex constituents. The latter are combinations of adjacent elementary constituents or parts of text. This combination follows compositional principles. A document can therefore be described as structured hierarchically: complex constituents are aggregations of elementary ones, e.g. an article consists of sections, a section is divided into a set of paragraphs and perhaps lists or figures, and a paragraph may contain quotations or emphasized tokens.

Elements of the logical document structure, which indicate a specific discourse structure or a specific rhetorical relation and which therefore serve as cues are, for example, *listitem*, *glossterm*, *caption* and *title*. *Listitem*s

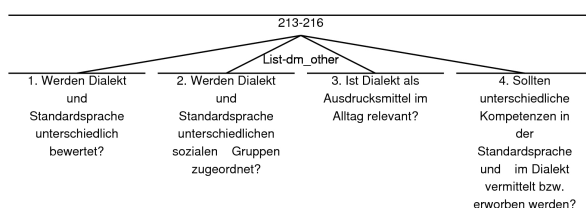


Figure 2: RST annotation for extract in Listing 2

```

<orderedlist>
  <listitem>
    <para>1. Werden Dialekt und Standard-
      sprache unterschiedlich bewertet?</para>
  </listitem>
  <listitem>
    <para>2. Werden Dialekt und Stan-
      dardsprache unterschiedlichen sozialen
      Gruppen zugeordnet?</para>
  </listitem>
  <listitem>...</listitem>
</orderedlist>

```

Listing 2: DocBook annotation (extract)

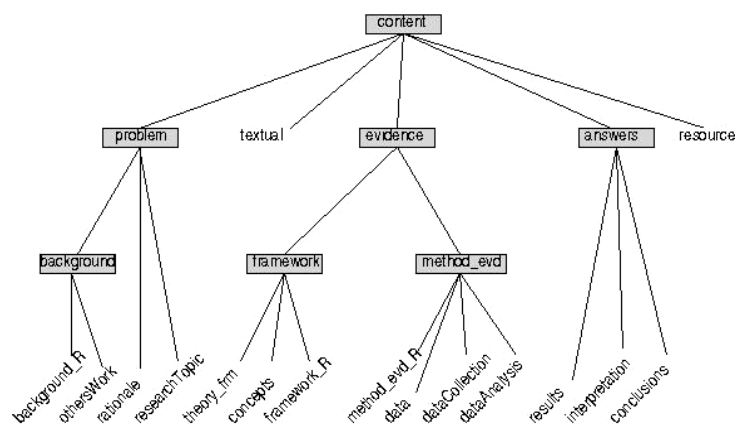


Figure 3: Text type structure (TTS) schema for scientific articles

indicate a LIST relation between all items of a bulleted list (shown in Listing 2 and Figure 2), *glossterms* may induce the nucleus in an ELABORATION-DEFINITION relation (shown in Listing 1 and Figure 1), *captions* 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, table or figure.

Apart from these (and other) *cues*, the logical document structure also serves as a *constraint* 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:

- **CDS type="para"**: A partial discourse tree which comprises one paragraph; the paragraph acts as an upper boundary for the construction of complex discourse segments; the top-nucleus of a CDS type="para" has the status of a nucleus for the whole paragraph
- **CDS type="section"**: A partial discourse tree which comprises one section (if the section consists of more than one paragraph; if not, it is a CDS type="para"); the section acts as an upper boundary for the construction of complex discourse segments; basic discourse segments are of CDS type="para"
- **CDS type="article"**: A discourse tree which comprises the whole article; basic discourse segments are of CDS type="section"

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 a CDS type="para" can only be related to other CDS type="para", but not to a CDS type="section". 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.

4. Cues and constraints from generic document structure

Apart from the *logical* document structure, a second type of document structure exists: the *generic* document structure, or, in other words, the genre-specific *text type structure* (TTS) or *superstructure* (Swales 1990, van Dijk 1980). It describes the global organization of a document into genre-specific functional categories (or *zones*, Teufel 1999) like, for example, *Introduction*, *Method*, and *Results* (= categories for 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 organized hierarchically and therefore be formally described by a hierarchical schema (e.g. Kando 1999). The text type schema we developed for scientific articles is shown in Figure 3. This schema is based on the works by Kando and Teufel, but was, as a result from our corpus analyses, adapted to our corpus as well as to our aims. As we wanted to develop an informative schema, all categories were sorted out which were considered as functional (e.g. *Reason*). Instead, we are using more thematic categories (e.g. *Method*, *Theory*). For the annotation of our corpus each text was divided in thematic segments, often, but not necessarily always, consisting of a sentence. This (more or less) sentential

```

<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 referiert. </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 "Män-
nersache" dargestellt, Frauen werden, auch wenn sie vorhanden sind, selten sichtbar ge-
macht. Zudem wird auch mit geschlechtsspezifisch männlichen Wörtern wie Gründerväter der
Gedanke an Männer evoziert. </segment>

```

Listing 3: TTS annotation (extract)

segmentation corresponds to the segmentation realized by Kando (1999) and Teufel (1999). An example of a TTS annotation is shown in Listing 3.

The role of the generic document structure for discourse analysis in the tradition of RST has lately been observed 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, depending on a previous stage. Gruber and Muntigl empirically show that generic and rhetorical structure of students academic writings coincide.⁵ They found correlations between both genre dependent and independent stages and RST relations, e.g. *Orientation* 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).⁶ Taboada and Lavid also empirically gained 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 can be distinguished:

1. A TTS category which is comparable to an RST relation can be used as an explicit cue for a specific RST relation: A text type structure category can be interpreted, as we pointed out, as a functional relation between a part of a text and the text as a whole, while RST establishes a functional relation between two or more parts of text, or 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* (a TTS category *Summary* does not exist in our schema). Hence, it seems reasonable to map those TTS categories to the equivalent RST relations, e.g. an *Interpretation* constituent may be an RST satellite in an INTERPRETATION relation. This hypothesis gains empirical evidence through the findings of a descriptive analysis we performed for our corpus.⁷ The relation INTERPRETATION can be found 17x more often in TTS segments which are of type *Interpretation* than with all other TTS categories (on average), BACKGROUND occurs 9x more often with *Background*, and SUMMARY 9x more often with *Conclusions*.
2. A TTS category (assigned to a TTS segment) which appears more often 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 as an abstract 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 the following section.
3. At the highest level of discourse structure (CDS type="article"), the global categories of the text type structure schema (*Problem*, *Background*, *Evidence*, *Framework*, *Method*, *Answers*) are instantiated automatically for all CDS type="section". The parsing procedure for this instantiation has not been implemented yet, but it shall be based on a quantitative analysis of all text type structure categories of one section where the parent element (one of the global categories) of the majority of categories of one sec-

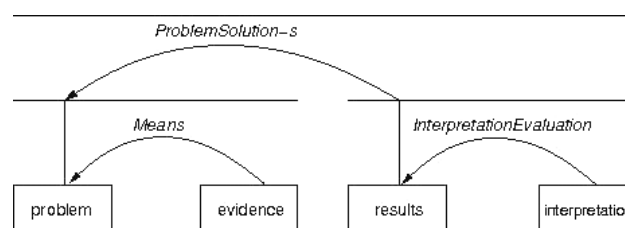


Figure 4: Possible relations between TTS-categories

tion should become the status of the section’s global generic function. The relations between the global generic categories shall then be annotated by using an abstract relational article schema like the one shown in Figure 4. Deviations from this article schema are not a priori ruled out, i.e. global categories could be omitted in a specific instantiation of the schema. Nonetheless, we base our approach to the global generic structure on the assumption that scientific articles are almost always organized along the outlined global generic categories, and that strong deviations may indicate that a document belongs to a different text type.

5. Correlations between text type structure categories and rhetorical relations

Our test corpus comprises two parts: The first consists of 10 English psychology articles with 8597 words on average. In this part of the corpus (in the following: PsyEngl corpus) the discourse segments are elementary (EDS). The second part of our corpus contains 10 German linguistic articles with 8627 words on average – this corpus (in the following: LingDeu corpus) contains rhetorical relations between complex discourse segments on paragraph level. In the analyses of the corpora we examined the correlations between RST relations and TTS categories. For each *TTS segment*⁸ we listed the rhetorical relations between discourse segments included in the TTS segment and attached them to the TTS category of that segment. As a result, we obtained a matrix with the relations arranged in lines and the TTS categories in columns. This matrix shows the type and number of used 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 1.

	TTS categories used	Number of TTS Segments ⁹	Most frequent TTS categories	RST relations used	Number of 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 1: Number and frequency of TTS categories and RST relations in the corpora

In a second 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.¹⁰ For each relation, we calculated the average percentage of their frequency over all TTS categories. The average frequency of RST relations over all TTS categories is:¹¹

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

To find out which RST relations are more prominent in TTS category A than B, we compared the average percentages 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 in average (for all TTS categories) to 1.44 % of the number of all relations for 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 this case *Framework*), we divide the actual percentage through the average percentage, e.g. $10 / 1.44$. The result of this procedure 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 stay with our example, CONDITION can be found 5.6 times more often in *Framework* than (in average) in all other TTS categories. The results of this calculation procedure are illustrated in Figure 5. The figure shows the different distributions of RST relations. The peaks in the graphs point out that the TTS categories can be clearly distinguished by a different distribution of relations, and that they all have special characteristics with regard to the occurrence and frequency 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 segments (each having an assigned RST relation), Framework holds $202 / 801 (= 25\%)$ of all

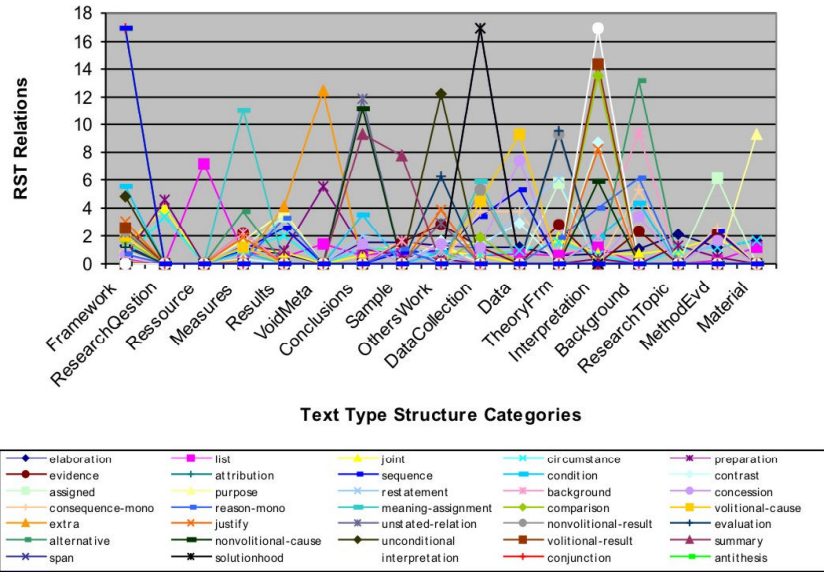


Figure 5: Distribution of RST relations over the different TTS categories for the PsyEngl corpus

included RST segments/ relations. One could therefore expect that 25% of all included ELABORATION relations in the corpus would be found in TTS segments assigned as *Framework*, i.e. 25% of 282 (in the PsyEngl corpus) = 71. Like before, the difference between the expected number of relations/ RST segments and the actual number can be described by a factor (in the following: Factor EA), which is the result of the division of the expected number (i.e. 71) and the actual number (i.e. 76), in this case 1.1.

Some of the factors we obtained were extremely high (up to 17.0). Very often this was due to a small amount of included relations of that type in the corpus. For this reason, we ignored those relations which have less than 10 included instances in each of the corpora. Due to the different number of RST segments 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 different RST relations have more than 10 included instances, whereas 19 relations have more than 10 included instances in 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		
list								2.2 / ...			4.4 / ...		
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 2: Correlations between TTS categories and RST relations

In Table 2, the correlations between more common 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 the average distribution is represented through the above mentioned Factor D (only factors which are higher than 2.0). In the second line of each cell, the difference of the expected number of a relation at a specific TTS category and its actual number are represented through the Factor EA (only factors which are higher than 2.0). The first number in the cells of the table always refer to the LingDeu corpus, the second 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 size of discourse segments (EDS for PsyEngl, CDS type="para" for LingDeu), or the domain and maybe the language specific style of discourse organization. One or all of these factors seem to influence the prominence of TTS categories and RST relations, e.g. English psychology articles contain a lot more segments of the TTS category *Measures* than German linguistic articles, whereas 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 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.

Apart from these correlations, which are identical for both corpora, we can also describe correlations for each of the corpora alone. 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 pronounced correlations are *Conclusions* – SUMMARY, *Data* – CONCESSION, *Interpretation* – CONTRAST, *MethodEvd* – ASSIGNED, *TheoryFrm* – ASSIGNED, and *TheoryFrm* – RESTATEMENT.

All of the empirically gained correlations, which are represented by the Factors EA and D can be used as statistic constraints and cues for the assignment of RST relations to discourse segments which have an assigned TTS category. In this function, the factors will be integrated in the discourse parser, which is being developed in our research project.

6. Representation of linguistic analyses

To make the additional knowledge sources accessible for automatic discourse parsing, logical and generic document structure are represented as XML annotation layers, which are stored in independent files. Apart from these two annotation layers we additionally provide a syntax/morphology annotation layer, which is created automatically by the commercial *Machinese Syntax* tagger software from Connexor Oy, and a rhetorical structure (RST) layer, which will serve as training and evaluation material for the discourse parser. 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, which is based on the text type structure schema for scientific articles shown in Figure 3. So far, the texts of the corpus are annotated manually or semi-automatically.¹² The XML-based multi-layer annotation approach (Witt et al. 2005) is used to examine dependencies between different document levels and to provide a rich input for the discourse parser. Moreover, the discourse parser has access to a discourse marker lexicon (which is not fully available yet as it is work in progress). This lexicon shall not only contain lexical discourse markers, but also abstract discourse markers like syntactic cues and cues from the logical and generic document structure, gained through the empirical analysis of the relationship between the relational discourse structure and the logical and generic document structure.

Notes

¹ We distinguish between cues and constraints: While cues can be used for bottom-up relational discourse analysis and therefore act as discourse markers for the instantiation of specific discourse relations (RST relations) between two or more discourse segments, constraints serve as top-down restrictions for discourse structures.

² The corpus comprises 120 scientific articles, which are mixed in language (English and German), domain (psychology and linguistics) and sub-genre (experimental and review).

³ The term bridging for us subsumes the different types of indirect anaphoric relations like nominal hyperonymy and meronymy (cf. Holler 2003).

⁴ 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.

-
- ⁵ 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.
- ⁶ The relational discourse analysis based on RST was restricted to the level of subchapters; clauses were not annotated.
- ⁷ 2 x 10 texts of our corpus were analyzed: 10 German linguistic articles whose RST annotations were done on the level of paragraphs (CDS), 10 English psychology articles whose RST annotations were done on the clause level (EDS).
- ⁸ As a preparation of our empirical analyses, we automatically assigned TTS categories to *TTS segments*. This assignment was based on the TTS annotation of elementary thematic segments – the TTS category, which was quantitatively most prominent in a paragraph (or an element on the same level such as title or table) was chosen as the TTS category for the paragraph as a whole. Subsequently, adjacent paragraphs with identical TTS categories were joined into one TTS segment.
- ⁹ Due to their size, some TTS segments do not contain any RST segments.
- ¹⁰ This distribution of RST relations can be found in the PsyEngl corpus.
- ¹¹ These percentages are *not* referring to the number of RST relations in the corpus, but to the average distribution of relations in each TTS category.
- ¹² The DocBook-annotation was done 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 produced very different results: The recall and precision figures for some TTS categories were good, while recall for other categories was very poor (Langer et al. 2004).

References

- 1 Asher, N. and A. Lascarides (2003). *Logics of Conversation*. Cambridge, U.K.: Cambridge University Press.
- 2 Carlson, L. and D. Marcu (2001). *Discourse tagging reference manual*. Technical Report ISI-TR-545. Marina del Rey CA: Information Science Institute.
- 3 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.
- 4 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.
- 5 Holler, A. (2003). *Spezifikation für ein Annotationsschema für Koreferenzphänomene im Hinblick auf Hypertextualisierungsstrategien*. [<http://www.hytext.uni-dortmund.de/hytext/publikationen.html#Dokus>]
- 6 Langer, H., H. Lungen, 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.
- 7 LeThanh, H., G. Abeyasinghe, and C. Huyck (2004). Generating Discourse Structures for Written Texts. In: *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland.
- 8 Lobin, H., M. Bärenfänger, M. Hilbert, H. Lungen, and C. Puskas (2006, to appear). Discourse relations and document structure. In: Metzging, D. and A. Witt (Ed.): *Linguistic modeling of information and Markup Languages, Contributions to language technology*. (Series Text, Speech and Language Technology). Dordrecht: Springer.
- 9 Lungen, 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.
- 10 Mann, W. C. and S. A. Thompson (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. In: *Text* 8(3), 243-281.
- 11 Mann, W. C. and Taboada, M. (2005). *Rhetorical Structure Theory. Relation Definitions*. [<http://www.sfu.ca/rst/01intro/definitions.html>]
- 12 Marcu, D. (2000). *The theory and practice of discourse parsing and summarization*. Cambridge, MA: MIT Press.
- 13 Power, R., D. Scott, and N. Bouayad-Agha (2003). Document structure. In: *Computational Linguistics*, 29(2), 211-260.
- 14 Reitter, D. and M. Stede (2003). Step by step: Underspecified markup in incremental rhetorical analysis. In: *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03) at the EACL*, Budapest.
- 15 Swales, J. M. (1990). *Genre Analysis. English in academic and research settings*. Cambridge, UK: Cambridge University Press.
- 16 Taboada, M. and J. Lavid (2003). Rhetorical and thematic patterns in scheduling dialogues. In: *Functions of Language* 10(2), 147-148.
- 17 Teufel, S. (1999). *Argumentative Zoning: Information Extraction from Scientific Text*. Ph. D. thesis, University of Edinburgh.
- 18 van Dijk, T. A. (1980). *Macrostructures: An interdisciplinary study of global structures in discourse, interaction, and cognition*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
- 19 Walsh, N. and L. Muellner (1999). *DocBook: The Definitive Guide*, O'Reilly.
- 20 Witt, A., H. Lungen, H., D. Goecke, and F. Sasaki (2005). Unification of XML documents with concurrent markup. In: *Literary and Linguistic Computing*, 20(1), 103-116.