Dialogue and Discourse 4(2) (2013) 142-173 doi: 10.5087/dad.2013.207

Linguistic Tests for Discourse Relations in the TüBa-D/Z Corpus of Written German

Yannick Versley VERSLEY@SFS.UNI-TUEBINGEN.DE
Sonderforschungsbereich 833
Universität Tübingen
72074 Tübingen, Germany

Anna Gastel ANNA.GASTEL@UNI-TUEBINGEN.DE
Sonderforschungsbereich 833
Universität Tübingen
72074 Tübingen, Germany

Editors: Stefanie Dipper, Heike Zinsmeister, Bonnie Webber

Abstract

Discourse structure and discourse relations are an important ingredient in systems for the analysis of text that go beyond the boundary of single clauses. Discourse relations often indicate important additional information about the connection between two clauses, such as causality, and are widely believed to have an influence on aspects of reference resolution. More so than for referential annotation, discourse relation annotation is rendered difficult by the absence of a general consensus on the underlying linguistic phenomena that should be targeted, as well as by a lack of strong predictions on the possible or permissible interactions between these phenomena.

While it is sometimes claimed that the structuring of discourse is only weakly constrained and as a result capturing discourse structure and discourse relations will always result in poor reproducibility of the annotation task, we want to argue in this paper that an explicit notion of the relata of discourse relations allows to delimit annotation scope and to make use of theoretical accounts of the linguistic phenomena involved without giving up the goal of theory-neutrality that is essential in making sure that a given resource stays useful to a large community of users.

In this article, we first present the general design choices that are to be made in the design of an annotation scheme for discourse structure and discourse relations. In a second part, we present the scheme used in our annotation of selected articles from the TüBa-D/Z treebank of German (Telljohann et al., 2009). The scheme used in the annotation is theory-neutral, but informed by more detailed linguistic knowledge in the way of linguistic tests that can help disambiguate between several plausible relations.

Keywords: Discourse Structure, Discourse Relations, Agreement, Linguistic Tests

1. Introduction

Discourse information has been proven useful for a number of tasks, including summarization (Schilder, 2002) and information extraction (Somasundaran et al., 2009). While coreference corpora exist for many languages, and in large and very large sizes (frequently over one million words), the annotation of discourse structure and discourse relations has only recently gained the interest of the community at large.

 Submitted 03/12; Accepted 12/12; Published online 07/13

The general idea of hierarchical discourse structure has a long history (Polanyi and Scha, 1984; Grosz and Sidner, 1986; Mann and Thompson, 1988; Webber, 1988). Mann and Thompson's Rhetorical Structure Theory (RST), being the first to aim at a descriptively adequate account of real texts, has been the basis of annotated corpora targeting the analysis and generation of discourse structure, most notably the RST discourse bank (Carlson et al., 2003) and similar corpora in other languages (cf. Stede, 2004a, van der Vliet et al., 2011), but has also drawn criticism regarding the cognitive plausibility of some of its aspects: In particular, Sanders and Spooren (1999) claim that RST does not separate between speaker intentions (which may not necessarily become shared knowledge) and coherence relations (which are instrumental for the understanding of a discourse); Wolf and Gibson (2005) take issue with the assumption that discourses are tree-structured, and propose to focus on the presence of coherence relations without any consideration of overall structure, whereas Stede (2008) levels a more focused criticism at RST's notion of nuclearity, which, as Stede claims, encompasses criteria on different linguistic levels which are not always in agreement with each other. Knott et al. (2001) propose a separation between low-level coherence relations (which can typically be signalled by conjunctions), and other means of structuring, which typically involve larger spans and make use of nominalization or discourse deixis.

More recent work has set out to take into account the aforementioned criticisms, but also to make discourse annotation more efficient and predictable. Most importantly, the authors emphasize the need to focus on a subset of the task that can be annotated reliably and that is at the same time informative with respect to a core set of discourse phenomena that are thought to be central, as is claimed by Sanders and Spooren (1999) or Knott et al. (2001) for relations that are expressible through connectives. As a result, newer approaches such as the Penn Discourse Treebank 2.0 (PDTB; Prasad et al., 2008) use these fundamental ideas — formalized in the D-LTAG formalism by Webber (2004) — to define the scope of their annotation in theory-neutral guidelines for implicit (connective-less) and explicit (connective-bearing) discourse relations.

The discourse annotation in the TüBa-D/Z treebank of German (see Telljohann et al., 2009, and Möller and Naumann, 2009, for the syntactic and referential annotation layers) has a similar scope to the PDTB, and attempts to reach a useful balance between theory-neutrality on one hand and linguistic insights on the other hand, in particular from previous work in the frameworks of Discourse Lexicalized Tree Adjoining Grammar (D-LTAG) or Segmented Discourse Representation Theory (SDRT) as well as primarily descriptive treatments of single phenomena. By making strong assumptions on the types of (semantic or formal-pragmatic) entities that can be related by various discourse relations, it is possible to derive clearer criteria for linguistic tests, and it also makes it more obvious where cases are problematic due to violations of the basic assumptions (e.g., multiple propositions for one discourse segment, presence of implications as the relata of a discourse relation instead of the stated facts; cf. Versley, 2008, and Recasens et al., 2011, for similar considerations in coreference annotation).

2. Defining the Annotation Task

When laying down the guidelines for a discourse annotation task, one defines a formal model (i.e., text annotations and their specification) which relates parts of the text (*discourse segments*: usually sentences, clauses, or larger units) to each other. We will refer to these relations between discourse segments as *discourse relations*. Such a formal definition of the task needs to answer (among others) the following two central questions:

1. Which relations between discourse segments are annotated, especially for relations that are implicit?

2. How is a given relation token described (in terms of label inventory)?

Both of these questions are inter-related in that, ideally, the relation inventory should offer a pertinent label for any pair of discourse segments that is to be annotated according to the first criterion; and conversely, if a given kind of relation between discourse segments is declared to be within the scope of our annotation, we would expect most or all of these relations to be part of the actual annotation.

Different annotations schemes start from a particular intuition to provide answers to these questions that are plausible in general, but which become problematic for some aspects of the annotation task. Individual solutions for these questions are in disagreement with each other in particular areas, and also differ in which individual aspects are problematic. Hence, a consideration of fundamental assumptions of existing annotation schemes, and the resulting answers to the above questions (which will be discussed in more detail in the following subsections), is essential.

2.1 Underlying Intuitions

With respect to the first question, RST assumes a hierarchy of discourse segments that spans the whole text. Discourse LTAG starts from the notion of a *discourse connective* as an invariant phrase that signals a relation between two (tensed) clausal arguments. Both the hierarchy assumption and the notion of connective arguments lead to discourse segments that may span multiple sentences. However, the two ideas lead to structures that are mutually incompatible in many cases. In the case of *implicit relations*, RST's hierarchy assumption is still tenable in principle, while the idea of connective arguments is of limited use.[1] In practice, existing annotation guidelines for implicit discourse relations either posit hierarchical structure (RST), ask annotators to mark all potential discourse relations that fulfill certain semantic criteria (Wolf and Gibson, 2005 and their Discourse GraphBank), or limit the annotation of implicit relations to neighbouring single sentences (PDTB). While systematic annotation difficulties cast doubt on RST's assumption that there is a full hierarchy of discourse relations (cf. Stede, 2008), neither the total absence or irrelevance of discourse structure (as postulated by Wolf and Gibson) nor the implicit assumption that implicit discourse relations never (or only very rarely) occur between single neighbouring sentences are satisfying alternatives for the question of discourse structure. Possible solutions for the structural assumptions in discourse annotation will be examined in subsections 2.4 and 2.3. The issue of relation inventory, and associated problems, will be detailed in the following subsection 2.2.

2.2 Number and Kind of Discourse Relations

One area where different descriptive proposals as well as current annotation schemes diverge consists of the choice and granularity of discourse relations.

Existing schemes reach from the most minimal model, containing just two relations (Grosz and Sidner, 1986), to one containing a taxonomy with 350 relations (Hovy and Maier, 1995). Especially for schemes with many relations, a taxonomic organisation has the advantage of reconciliating the

1. Webber, 2004, in her exposition of D-LTAG, uses the idea of coordinating *null connectives* connecting adjacent discourse subtrees, which would essentially lead to a hierarchical structure.

need for differentiated relations with the necessity to capture most pertinent behaviour in a relatively small set of relations (near the top level of the hierarchy). However, even in the presence of taxonomic organization, the level of detail of an annotation scheme should be sufficiently motivated.

Research in RST after Mann and Thompson's initial set of 20 relations seems to suggest that the exact set of relations used in a given project — whether annotation or computational generation of text — partly depends on the domain at hand and what distinctions are appropriate depends on the domain. This leads to Nicholas' (1995) claim that computational work using RST '*has arbitrarily expanded the inventory of relations used*,' creating a major problem for RST as a (linguistic) theory, or the claim of Knott and Dale (1994) that '*there often seems no motivation for introducing a new relation beyond considerations of descriptive adequacy or engineering expedience*'.

The problem of arbitrary inventory expansion can be avoided in schemes with strong restrictions on what can, or cannot, be a discourse relation. Sanders et al. (1992) do this by introducing a **relational criterion** that limits the kinds of distinctions that should be made between discourse relations. Sanders et al.'s relational criterion posits that categorization must refer to properties that concern the *informational surplus* that the coherence relation adds to the interpretation of the discourse segments in isolation, as opposed to the semantic contribution of the segments themselves. As a consequence, Sanders et al. remove temporal semantics (which is inherent in the combination of tenses, or in the semantics of a conjunction linking two segments) from consideration as useful distinctions in discourse relations.

The distinctions that Sanders et al. do make are almost fully orthogonal and organize discourse relations into a multidimensional lattice rather than a hierarchical taxonomy: The first distinction that Sanders et al. make is that of the *basic operation*, which can be *causal* (with a directed implication) or *additive* (with a — typically symmetric — conjunction between the two segments). Other, binary, distinctions are *source of coherence* (which can be *semantic*, if it relates discourse segments based on their propositional content, or *pragmatic*, if it concerns the illocutionary forces of two discourse segments), *basic* versus *non-basic order of segments* in *causal* relations, and finally the *polarity* of a relation, which is *negative* for adversative relations and *positive* otherwise.

Another proposal for theory-independent, empirically driven criteria of what is a 'psychologically real' relation comes from Knott and Dale (1994): they say that evidence for a rhetorical relation is to be found in **cue phrases** that signal this relation. Knott and Dale, and Knott (1996) give a criterion to find cue phrases, and use **substitutability** relationships for inferring taxonomical relations between cue phrases, hence constructing a fine-grained taxonomy while also avoiding the problem of arbitrariness that would otherwise arise.

Both proposals — Knott and Dale's as well as Sanders et al.'s — put limits on the coverage as well as the granularity of a scheme for discourse relations. While they are both well-motivated, it should be noted that the notion of cue phrases leads to a broader set of relations than Sanders et al.'s relational criterion.

A third proposal on criteria for discourse relation types stems from Sanders and Spooren (1999), who start from the idea that discourse relations are more or less related to either real-world entities, or to discourse-internal purposes (subject-matter vs. presentational relations: Mann and Thompson, 1988; Moore and Pollack, 1992).

Sanders and Spooren argue that, among the discourse relations that are generally postulated, **propositional relations** such as *Cause* relate states of affairs, whereas **illocutionary relations** such as *Evidence* introduce a meaningful relation not between states of affairs but between *illocutions* (speech acts). They distinguish between the notion of an illocution (which is shared between speaker

and hearer) and a communicative intention (which may or may not be shared between speaker and hearer, and are mapped to illocutions by the speaker). Because **communicative intentions** may be private to the speaker, and ultimately domain-specific, they argue, the third group of discourse relations such as RST's *Preparation* which are especially concerned with communicative intentions should *not* be part of the discourse annotation.

Proposals oriented at a narrow notion of coherence relations (such as Hobbs, 1985, which inspired the annotation scheme of Wolf and Gibson's Discourse GraphBank, or the original formulation of SDRT in Asher and Lascarides, 2003), as well as Sanders et al.'s (1992) criterion, usually do not cover the group that Sanders and Spooren call *illocutionary relations*, even when these can be signaled by a cue phrase. As a result, corpora such as the Penn Discourse Treebank, while subscribing to the idea of coherence relations in principle, include relations such as *Conjunction* that would be outside the scope delined by Sanders et al. (1992).

In defining the scope of our annotation scheme, we follow the general approach of Knott and Dale, as well as Sanders and Spooren's proposal that discourse relations be limited to propositional/illocutionary relations and not include (domain-specific or non-shared) communicative intentions. For the taxonomic organization of our annotation scheme, we start from the properties of Sanders et al., but complement them with additional properties to cover other types of relations (e.g. different types of elaborating relations). We also used Knott and Dale's approach of defining discourse relations in terms of cue phrases in the construction of explicit tests. Such tests include — but are not limited to — substitution and/or insertion of selected cue phrases to improve the separability of discourse relations (section 4).

2.3 Relations and Hierarchy

As in phrasal syntax, many theories of discourse structure posit a duality of constituents on one hand and head-dependent relationships on the other; in our case, (simple or complex) discourse segments and discourse relations may be seen to hold these roles.

In Rhetorical Structure Theory, the *nucleus* and *satellite* of mononuclear relations have a similar function to heads and dependents in syntactic theory. In particular, Mann and Thompson (1988) argue that deleting the nucleus of a mononuclear relation will result in an incoherent text as the significance of the material in the satellite(s) will not be apparent, whereas it is possible to delete, or replace, satellites of a discourse relation.

In Segmented Discourse Representation Theory (Asher and Lascarides, 2003), a related but distinct set of notions is used: *coordinating discourse relations* are relations between discourse segments with the same topic, and *subordinating discourse relations* hold between a super-topic and a sub-topic. In difference to RST, coordinating and subordinating relations can link to the same discourse segment. For example, in (1), (c) is an elaboration of its super-topic (a), but it also has a *Narration* relation to its sibling (b).

(1) (a) John had a great meal.
(b) He ate salmon.
(c) He devoured lots of cheese.

Asher and Vieu (2005) concern themselves with linguistic tests for distinguishing between coordinating and subordinating discourse relations. In their tests, they apply both the *right frontier constraint* (see Webber, 1988, and later research) and a hypothesis that coordinating relations are

exactly those that are expressible within a syntactic coordination of the relation's arguments (Txurruka, 2003). Using these criteria, they can show for the causal *Result* relation as well as for *Narration* that they are coordinating in general, although they also find examples of *Result* that are clearly subordinating.

In summary, postulating a hierarchy not only of larger and smaller discourse segments but also one across more-central/less-central may well pose problems as it sometimes deviates from the predictions derived from syntactic structure, and may become a hindrance if tokens of the same relation do not always have the same structural properties. However, minimal most-important segments, similar to the assumption of (semantic or lexical) heads in syntax, provide a useful abstraction when reasoning about larger discourse segments when the underlying assumption does hold.

In our annotation scheme, we use the idea of subordination and coordination, as postulated in SRDT. For a discussion of relation types versus structural properties (coordination and subordination), see the discussion in subsection 3.2.

2.4 Delimiting the Scope of Discourse Annotation

The idea of a hierarchical discourse structure that reaches from elementary units (such as clauses) up to the complete document has been instrumental in the definition of Rhetorical Structure Theory and other approaches, but has also been claimed to be problematic for reliable annotation. Especially the larger structure of documents has been found to be a bad fit for rhetorical theories, and Taboada and Mann (2006) write that "*In general, analysis of larger units tends to be arbitrary and unintuitive*" (p. 430), noting that structures at larger levels of granularity (subsections or chapters) tend to be governed more by genre conventions than by the mechanisms that govern the small to medium level.

This is in line with the Sanders and Spooren's (1999) claim that some RST relations that link larger text units are not coherence relations, but model communicative intentions, which are not necessarily shared between speaker and hearer, and therefore can be much more arbitrary and unintuitive than the coherence relations that link smaller units (cf. subsection 2.2).

Let us examine a solution for setting the scope of discourse annotation which is based on these insights: Instead of building complete trees, one subdivides the complete discourse into segments that realize one maximally complex illocution (which we call *topic segments* in our annotation scheme, cf. section 3, but which can also be found in several other annotation schemes, as noted in subsection 2.6). Such an approach would allow multi-sentence segments as arguments of both implicit and explicit discourse relations, while maintaining a strong focus on coherence relations. In the remainder of this section, we look at existing analyses challenging (especially) the idea of hierarchy, and asking whether the respective phenomenon would lead to predictable problems for discourse annotation based on the idea of partial trees.

One criticism of full hierarchical discourse structure which supports the above idea of 'partial' discourse trees can be seen in Knott et al. (2001): Knott et al. look at cases which they call **resumption**, where one larger segment of discourse is followed by an elementary discourse unit that takes up an entity mentioned in the previous segment and makes it the central topic of the next stretch of discourse. For these cases, which typically involve a discourse relation of object-attribute-elaboration (a subtype of RST's *Elaboration* relation), Knott et al. argue that one has to choose between the overall connectivity of discourse relations on one hand and the presence of these *Elaboration* relations on the other. Knott et al. propose to segment the discourse into a sequence of

entity chains, each containing one or several discourse segments, as a means to solve the difficulties with resumption. Such entity chains each share the same global focus, and they have one *top nucleus* which is the top nucleus of one of the discourse segments of that chain.

A second source of discourse relations that would cross a hierarchical structure are **discourse adverbials** — discourse connectives that take their second argument anaphorically rather than structurally, and which also may occur crossing hierarchical structure or a complex discourse segment (in the sense of complex illocutions, or Knott et al.'s entity chains). Several researchers such as Asher and Lascarides (1998) and Webber et al. (2003) spell out discourse adverbials as deriving their second argument not from structural composition mechanisms (as is the case with conjunctions such as *but* or *because*), but from the resolution of a presupposed proposition or situation argument.

Besides cutting across structure in rare cases, discourse adverbials exhibit other properties related to presupposition resolution that can make their annotation problematic. One such problem is vagueness (when the second argument of the discourse adverbial is accommodated rather than textually specified). In other cases, multiple discourse adverbials can introduce multiple concurrent relations for the same two arguments — see example (2), from Stede, 2004b. It is also possible that the anaphoric argument of a discourse adverbial is either resolved within the same elementary discourse unit or accommodated to something that is not a discourse unit, as in examples (3) and (4), due to Webber et al. (2003):

(2) *Therefore*, we *then also* went to a bigger mountain.

(3) Every person selling "The Big Issue"
might *otherwise* be asking for spare change.

(4) John just broke his arm.
So, *for example*, he can't cycle to work now.

In the case of (4), *for example* is accommodated to be an example for all the other things that happened as a result of John breaking his arm; the latter is not at all realized in the text but is inferred by the reader.

In our annotation guidelines, we pay special attention to these two phenomena, which tended to confuse annotators in the beginning: cases of *resumption*, where referential cohesion exists between discourse segments that are otherwise unrelated, are specially marked in the annotation to make clear they behave differently. In the case of discourse adverbials, annotators are advised to disregard their contribution if the anaphoric argument is vague, part of the same discourse unit, or inaccessible structurally. For details on this, see section 3.1.

2.5 Information-theoretic Notions of Discourse Structure

To describe discourse segments, some approaches (notably, Mann and Thompson, 1988, Grosz and Sidner, 1986, or Sanders and Spooren, 1999) use intentional notions, modeling discourse segments as *complex illocutions*. An alternative approach for this problem, or possibly a group of alternative approaches, can be seen in the use of information-structural notions to describe discourse segments. Such approaches were proposed both by researchers primarily interested in information structure (Roberts, 1996; Büring, 2003) and as a means to spell out the notion of *topic* as it is postulated, e.g., in SDRT (Asher, 2004).

Assuming an *entity as a topic* potentially has useful properties – on one hand, you can postulate interaction between the topic entity and the resolution of pronouns, assuming that the topic entity (or 'global focus') is a good candidate for pronominal reference (see also Grosz and Sidner, 1986, or in part Knott et al., 2001, who claim such a global focus for larger discourse segments); on the other hand, subtopic relationships could be spelled out in a straightforward fashion in terms of semantic relations between entities (*John – John's hair – John's hair care*).

The notion of *discourse topic as a question* was formulated in more detail by van Kuppevelt (1995), who assumes questions as structure-building mechanism in discourse: Questions are licensed by previous discourse segments (or other shared perceptions, such as visually or audibly salient happenings), which he calls *feeders*; questions are then answered through new discourse segments (which can give partial or complete answers). Discourse structure arises through patterns of a discourse segment 'feeding' multiple questions, or subquestions arising from a discourse segment that in turns answers a (superordinated) question. Van Kuppevelt says a topic-constituting question has a *satisfactory* answer when it is completely answered and none of the answer segments feeds unanswered questions.

The connection between intonation and the topic-constituting question in the discourse context, which we will call simply **question under discussion** (QUD), has been investigated, among others, by Roberts (1996), Büring (2003) and in part by Asher (2004). The intonation of a sentence can have a *focus* (F, A-accent, rheme focus), which indicates that the constituent is an answers to the question under discussion, and a *contrastive topic* (CT, B-accent, theme focus) which marks constituents that may differ in a sibling QUD.[2]

Contrastive topic marking indicates the presence of other, related QUDs in the discourse model (Büring), the presence of implicit sub-questions (Büring) or may signal a partial or indirect answer to a QUD (Asher), as in example (5):

(5) Q: What did the pop stars wear?
A: The [$_{CT}$ female] pop stars wore [$_{F}$ caftans].
A′: The tuneless wannabees wore [$_{F}$ caftans].

(The alternative answer A′ is read to constitute a full direct answer, accommodating the assumption that all popstars were tuneless wannabees.)

Asher (2004) remarks that, while he generally expects CT to mark the discourse topic of an elementary discourse segment, not all relations do, or even can, receive contrastive topic marking, and that the approaches put forward by van Kuppevelt and by Büring cannot make any predictions regarding coordinated relations such as *Continuation* or *Narration*.

In our annotation guidelines, the notions of QUD and of information structural marking are used to implement tests for discourse relations where contrast pairs, or the surrounding discourse, play an important role. Section 4 discusses this in greater detail.

2. The most common nomenclature for these intonation curves is *focus* and *contrastive topic* as used by Büring (2003), but it is also common to find the terms *A-accent* and *B-accent* following Jackendoff (1972) or *theme focus* and *rheme focus* as suggested by Steedman (2000), respectively *theme kontrast* and *rheme kontrast* in the work of Vallduví and Vilkuna (1998). While the exact predictions on the link between semantic structures and intonations differ, the respective terms refer to the same intonation contours originally pointed out by Jackendoff.

2.6 Annotated Corpora

Besides our own annotation effort, a substantial number of corpora exists, in multiple languages. Many of these follow the general ideas of Rhetorical Structure Theory, as the English *RST Discourse Treebank* of Carlson et al. (2003), the German *Potsdam Commentary Corpus* of Stede (2004a), as well as the *RST Spanish Treebank* of da Cunha et al. (2011), multiple Brazilian Portuguese corpora (Pardo and Seno, 2005; Pardo and Nunes, 2008; Cardoso et al., 2011), and the Dutch corpus of van der Vliet et al. (2011). In terms of relation inventory, they either using the 'profligate' approach of Hovy and Maier (1995) with close to 100 relations in the case of the RST Discourse Treebank, or use a close variant of Mann and Thompson's (1988) original RST annotation scheme.

Many other corpora follow the ideas of the Penn Discourse Treebank (Prasad et al., 2008), including corpora in Arabic (Al-Saif and Markert, 2010), Czech (Mladová et al., 2008), Hindi (Kolachina et al., 2012), Italian (Tonelli et al., 2010) and Turkish (Zeyrek et al., 2010). Most of these corpora presently only cover discourse connectives and explicit discourse relations, whereas implicit discourse relations are not yet part of the annotation.

Two corpora exist which are inspired by ideas from Segmented Discourse Representation Theory: the English SDRT corpus of Reese et al. (2007), and the French AnnoDis corpus of Péry-Woodley et al. (2011).

For these three 'traditions' of discourse annotation, **annotation guidelines** are available publically (Carlson and Marcu, 2002; Reese et al., 2007; PDTB Research Group, 2008). In the contained descriptions of discourse relations, we see a tendency to move from Carlson and Marcu's mostly informal and example-based treatment of RST relations to a more formal treatment by Reese et al. or the PDTB manual, which also use additional examples where appropriate for discussing special cases and often make explicit reference to propositions (PDTB) or main eventualities (SDRT) as the relata of coherence relations.

Several corpora use the idea of **partial trees** in their annotation, with slightly differing theoretical backgrounds: The Dutch corpus of van der Vliet et al. (2011) contains subtrees for each **conversation move**. Conversation moves, in their annotation scheme, are zones in each document which realize one genre-specific communicative purpose. The definition of conversation moves is genre specific: starting from a *rhetorical purpose* that is common to all texts of one particular genre, an analyst would identify possible *rhetorical functions* for segments that constitute a discourse move. Because it depends on the analysis of a whole genre, move analysis can accurately capture admissible complex illocutions, but would be less suitable for a heterogeneous mixture of different genres such as that found in (general) newspaper text.

The SDRT-inspired AnnoDis corpus of Péry-Woodley et al. (2011) combines a macrostructure level, which uses textual cues and paragraph boundaries, with a level of microstructure which uses an inventory of coherence relations "mostly common to all discourse theories".

It is interesting to note that the idea of a text consisting of smaller segments that are mostly independent from each other has been postulated independently of discourse annotation, for example by Hearst (1997). However, only recent annotation proposals for discourse structure have addressed the question how this distinction can be made precise enough.

The two principal sources of information for a segmentation into topic segments are referential cohesion on one hand, and intentional notions on the other hand. A naïve application to referential cohesion would suggest that topic segments correspond more or less closely to the extents of referential or lexical chains, as postulated among others by Hearst (1997). However, work by Knott

et al. (2001) proposes that referential (and lexical) cohesion can normally occur across the boundaries of larger segments, and that a better indicator would be the shift of referential focus (i.e., the entity that is most salient in a discourse segment) rather than the presence or absence of mentions of a given concept or referent. From an intentional perspective, the most appropriate notion for a topic segment would be that of a **complex illocution**, which would imply that one topic segment corresponds to one specific complex speech act, such as explaining the history of a specific artifact, or describing the consequences of a planned construction.[3]

In our own annotation scheme (see subsection 3.1), we use complex illocutions as the motivating underpinning for topic segments, and introduce a special type of marking (called *transitional EDUs*) to mark cases where referential cohesion is at odds with these complex illocutions.

3. Discourse Annotation in TüBa-D/Z

In the TüBa-D/Z, a two-pronged approach has been chosen for the encoding of discourse relation information. On one hand, a sample of ambiguous discourse connectives (temporal subordinators as well as conjunctions) is disambiguated according to the discourse relation that they realize, aiming at a relatively precise account of the variability that the respective connective affords. On the other hand — and this is the part that the present article is concerned with — complete newspaper articles are annotated with discourse structure. The discourse structure consists of a segmentation of a complete discourse into topic segments (stretches of coherent texts that realize one high-level goal of a writer such as 'describing the attitude of retailers towards genetically modified products'), and organizes the text within a topic segment into a hierarchy of discourse units. This hierarchy of elementary (and composed) discourse units is realized through coordinating and subordinating relations, including relations between elementary discourse units or discourse spans that are implicit in that they are not marked by a discourse connective.

The main TüBa-D/Z corpus (Telljohann et al., 2009) contains 1.1 million word tokens with syntactic, named entity, and morphological annotation, as well as information on referential cohesion in the form of anaphora/coreference annotation, which makes it an appealing target as a text source for the additional annotation of discourse structure and discourse relations. From a purely technical point of view, the existing corpus contains rich linguistic annotation that can be exploited in conjunction with the discourse relations. From the point of view of text selection, the newspaper `die tageszeitung` offers a good balance between argumentative and descriptive writing, usually providing both reporting on current events as well as background information and commentary.

For the annotation of discourse relations, a subset of the articles was manually selected according to criteria such as length (excluding extremely short newswire-style reports as well as one or two extraordinarily long articles) and topical coherence (excluding summary articles that report a multitude of different items with only two or three sentences per item).

The current version of the discourse structure annotation, which is publically available as part of release 8 of the TüBa-D/Z, constitutes a subcorpus of 41 texts, with 21 817 word tokens containing 1 458 discourse relations, with about 28.8 sentences/article (against 20.6 sentences/article

3. A complex illocution, such as "*explain the history of a specific artefact*" or a formulated questions such as "*what happened to the artifact?*" are more clearly delimited than general topic description such as "*the history of the artifact*": the latter could also admit a discussion of other's critical views on the artifact at top-level, while the speech-act verb *explain* or the question "*What happened to X?*" would be understood not to include commenting. In comparison to the notion of conversation moves, complex illocutions do not presuppose a pre-established account of a genre's structuring conventions.

on average in the complete TüBa-D/Z, which includes brief newswire-style reports), and about 3.5 topic segments per article. Of all discourse relations, 557 (38%) are intra-sentential (between EDUs in the same sentence), whereas 367 (25%) involve at least one of Arg1 or Arg2 spanning multiple EDUs. In 182 cases (12%), at least one of Arg1 or Arg2 spans multiple sentences.[4]

3.1 Elementary Segments and Topic Segments

Elementary segments (also: elementary discourse units, EDUs) are the smallest units of text that can be the argument of a discourse relation and ideally correspond to exactly one main eventuality or one asserted proposition. Our annotation scheme is oriented at existing guidelines for English (Carlson and Marcu, 2002; Reese et al., 2007) and German (Lüngen et al., 2006).[5]

For the largest part, the guidelines for the segmentation of EDUs are defined in syntactic terms, granting EDU status to tensed clauses (matrix clauses as well as subclauses which are not center-embedded), but also to eventualities introduced by (causal, temporal or attributional) prepositional phrase adjuncts (6a), appositions introduced by right dislocation (6b), non-restrictive relative clauses as well as purpose clause adjuncts to tensed clauses. One area where semantic considerations play a role is the separation of reported proposition arguments, where verbs count as communication verbs (with discourse segment status for their arguments) if they allow direct speech or allow fronting of the reported content.[6]

(6) (a) [Nach nunmehr über sechs Wochen erfolgloser Luftangriffe]
[gibt es im Nato-Hauptquartier niemanden mehr, der das Scheitern bestreitet.]
[*After over six weeks of unsuccessful air raids,*]
[*there is no one left in Nato headquarters who disputes the demise.*]
(b) [Er sollte unseren heimischen Markt aufmischen,]
[das erste Produkt in deutschen Läden, das genmanipulierten Mais enthält.]
[*It was to stir up our domestic market,*]
[*the first product in German stores to contain GM corn.*]

As an upper boundary for discourse annotation, texts are divided into so-called **topic segments**. For every topic a title is added which summarizes that topic. All text within one topic supports an answer to the same high-level question under discussion (QUD). This QUD is the most important tool for the identification of topic segments: it should be a question that contributes to the overall topic of the article, while, for our text type of medium-length newspaper articles, a topic segment does not have a title/theme that is synonymous to that of the article. A surface cue for annotators is presented by the paragraphs of the article, since a topic boundary usually coincides with a paragraph boundary (but not vice versa – topics are usually multi-paragraph units of text). In our annotation, topic segments are therefore exactly one hierarchy level under the topic of the complete article.

4. As an example for a complex structure without multi-EDU arguments, consider a discourse graph where one superordinate EDU has relations to multiple subordinate EDUs, but where that larger segment is not related with another discourse segment. Many such simpler structures that a topic segment can have would be annotated without the involvement of relations between multi-EDU spans.
5. In comparison to the proposal of Lüngen *et al.*, we do not segment embedded/parenthetical material since this would needlessly create discontinuous EDUs. In the case of right dislocation, we make reference to the existing syntactic structure of the treebank instead of specifying punctuation heuristics.
6. TüBa-D/Z, sentence 2890; TüBa-D/Z, sentence 5736.

As an example, consider the following topic segmentation for an article about *the Nato's plans in Kosovo*:

- T0: *Opinions on the situation in Kosovo*
  QUD: What do important people say about the situation in Kosovo?
- T1: *Air strikes in Kosovo*
  QUD: What happened to the air strikes in Kosovo?
- T2: *Plans for ground troops*
  QUD: What plans are there for the deployment of ground troops?
- T3: *Role of the Red/Green minority*
  QUD: What role does the Red/Green minority in parliament play?

In the example, merging T1 with T2 would yield a very general topic about *plans of the Nato*, which corresponds to the overall topic of the article and is therefore uninformative. Conversely, splitting T0 into separate topic segments with *Naumann's opinions on the situation* and *Clinton's opinions on the situation* would clearly yield non-maximal questions under discussion, to which T0 is preferrable.

Usually, topic boundaries coincide with the absence of discourse relations, and generally less cohesion. There is, however, a notable exception to this generalization: In the annotation process for our corpus, we would frequently encounter cases where the author explicitly bridges two segments that treat different topics, with a discourse segment where one part of the sentence contains discourse-old information that explicitly creates cohesion with the previous topic segment, whereas the other part of the sentence contains novel information introducing the new topic segment. In these cases, annotator's topic boundaries would frequently differ by a single sentence. In order to allow annotators to make the disconnect between referential cohesion and topic segments explicit, we introduced **transitional discourse units** — elementary discourse segments that are at the start of a topic segment and contain referential cohesion to the previous topic segment, yet contribute to the illocutionary purpose of the new discourse segment. The marking of transitional EDUs serves the dual purpose of highlighting the common structure of such segments as well as preventing confusion over anaphoric reference of pronouns or discourse adverbials crossing the topic boundary. Additionally, the transitional EDUs can be the argument of a topic crossing relation that further serves the purpose of a cohesive transition of one topic to another.

In the context of example (7),[7] the referential expression *vor diesem Hintergrund* ('in this context') creates referential cohesion to the previous topic segment even though the actual contents of both topic segments are quite different.

(7) [$_{5.0}$ Bereits seit Jahren sind sie die großen Hoffnungsträger:] [$_{5.1}$ die jungen Existenzgründer.] [$_{6.0}$ Auf ihnen ruht die erwartungsvolle Aufmerksamkeit von Politikern und Wirtschaftsexperten] [. . .]
T1: die Deutschen Existenzgründertage
[Tr-EDU$_{12.0}$ Vor diesem Hintergrund fanden an diesem Wochenende die zweiten "Deutschen Existenzgründertage" statt,] [$_{12.1}$ mit denen die Träger, Wirtschaftssenator Branoner und

7. TüBa-D/Z, sentences 1979ff.

Brandenburgs Wirtschaftsminister Dr. Dreher, Berlins "Gründungskompetenz bundesweit transportieren" wollen.]

[$_{5.0}$ *Since years they are the beacon of hope:*] [$_{5.1}$ *the young business founders.*] [$_{6.0}$ *They have the keen interest of politicians and business experts.*] [. . .]

T1: the German Founders' Days

[Tr-EDU$_{12.0}$ *In this context, the second "German Founders' Days" took place last weekend,*] [$_{12.1}$ *which the sponsors, Economics Senator Branoner and Brandenburg Economic Minister Dr. Dreher, wanted to "transport Berlin's founder competence at the national level"*]

3.2 Discourse Structure and Discourse Relations

The main content of the discourse annotation consists in discourse relations, which link two arguments that can consist either of a single EDU or a span of EDUs (generally corresponding to a complete subtree in the discourse structure). Discourse relations can be either **subordinating** (in which case the second argument is subordinated under the first), or **coordinating** (in which case the first argument comes before the second in surface order).

As every relation type is marked as either coordinating or subordinating in the annotation scheme (cf. table 1), the annotated graph of relations specifies a hierarchical discourse structure as postulated by Asher and Vieu (2005). In line with Asher and Vieu, the TüBa-D/Z discourse annotation allows a discourse unit to be linked by both a coordinating relation (to a sibling in the discourse hierarchy) and a subordinating relation (to its superordinate in the discourse hierarchy).

Because a discourse segment can possess both a link to the larger discourse segment that subordinates it and coordinating relations to its neighbouring siblings, the graph of all relations is not a tree in the graph-theoretic sense, but the subgraph of all subordinating relations would fulfill that criterion. As an informal part of the annotation, our annotation tool allows annotators to indent elementary text segments in order to reflect subordination depth.

The hierarchy of discourse segments in the annotation also provides a holistic view on discourse relations regardless whether they are syntactically mediated (through embedding or by subordinating or coordinating conjunctions), marked by discourse adverbials, or completely unmarked. Using such a coherent view, the annotation of unmarked relations between larger segments is more straightforward due to constraints found in the structural context. In converse, our view of hierarchy entails that some, but not all relations due to discourse adverbials are part of the annotation.

In example (8) below, we can see that the larger span α that occurs as the argument for the *Commentary* relation constrains the possible relation targets for the EDUs inside (EDU boundaries are marked as vertical lines).[8]

(8) [$_{\alpha}$ Die gute Nachricht: Die Weltbevölkerung wächst inzwischen langsamer als in den vergangenen Jahrzehnten.| Die schlechte Nachricht: Erreicht wird diese Entlastung der ökologischen und sozialen Systeme nicht nur durch Fortschritte bei der Geburtenkontrolle,| sondern auch durch eine Sterblichkeitsrate, die zum erstenmal seit 40 Jahren wieder ansteigt. |. . .]

[$_{\beta}$ Diese Entwicklung zeigt nach Angaben von Lester Brown, einem der Autoren der Studie, das "Versagen unserer politischen Institutionen".]

[$_{\alpha}$ *The good news: The World population is growing more slowly than in past decades.*|

8. TüBa-D/Z sentences 8429ff.; see Figure 1, in section 5, for the larger context.

The bad news: This unburdening of ecological and social systems is not only achieved by progress in family planning,| but is also due to a mortality that growing again for the first time since 40 years. | . . .]
[$_\beta$ *This development shows, according to Lester Brown, one of the study's authors, a "failure of our political institutions."*] Commentary(α,β)

The full list of relations can be seen in table 1. In some cases, different relation types have somewhat similar semantics but different properties regarding discourse structure, such as *Result* (coordinating) and *Explanation* (subordinating), or *Attribution* (subordinating, non-veridical, reported content is in Arg2) and *Source* (coordinating, veridical, reported content is in Arg1).[9]

To illustrate these distinctions, consider example (9) and (10) for the distinction between *Result* and *Explanation*:[10]

(9) [$_1$ Private Unternehmen dürfen die Telefonbücher der DeTeMedien nicht ohne deren Erlaubnis zur Herstellung einer Telefonauskunfts-CD verwenden.] [$_2$ Die beklagten Unternehmen müssen den Vertrieb der Info-CDs sofort einstellen.]
[$_1$ *Private companies may not use the telephone books by DeTeMedien without its permission for the creation of directory assistance CDs*] [$_2$ *The defendant companies must cease the distribution of their information CDs immediately.*] Result-Cause(1,2)

(10) [$_3$ Taxifahrer sind als Kolumnenthema eigentlich tabu,] [$_4$ weil sie als "weiche Angriffsziele" gelten.]
[$_3$ *Taxi drivers are normally a taboo topic for a newspaper column,*] [$_4$ *because they are considered "soft targets".*] Explanation-Cause(3,4)

In the case of (9), both arguments are necessary for coherence, so the coordinating *Result* relation is chosen. In the relation example from (10), the fact in EDU 4 is considered the cause for EDU 3, but 4 mostly contributes background information and is not important in its own right.

In addition to these distinctions, the annotation process revealed that a coordinating variant of two more relation types present a useful addition, as the corresponding relation instances both pass the coordination criterion of Txurruka (2003) and Asher and Vieu (2005) and considerations of the surrounding discourse structure. *ConcessionC* is a counterpart of the *Concession* relation and actually has a somewhat different profile from the subordinating *Concession* relation (see subsection 4.1). *RestatementC* is a counterpart of the *Restatement* relation and is typical for a restatement where the first part of the restatement is overly vague or metaphorical, such that neither part can be left out. These cases very often occur in a coordination with *und* (and), which is a further indication that the relation is coordinating rather than subordinating.

Considering the critique of Stede (2008) on RST's notion of nuclearity[11] and the large number of relations in the annotation manual of Carlson and Marcu (2002) – 23 of the 53 mononuclear relations in their annotation scheme have multinuclear counterparts – it is an interesting question whether we will forcibly end up duplicating every relation in one subordinating and one coordinating version. Indeed, we see in table 1 that this is already the case for almost all of the relations in the *Contingency*, *Temporal* and *Reporting* groups. For the symmetric relations from the *Comparison* group, as well as

9. See Hunter et al. (2006) for a rationale on distinguishing *evidential* from other relations in discourse annotation.
10. TüBa-D/Z sentences 2536ff. and 8870ff.
11. This partially also applies to the critique of Moore and Pollack, 1992, who advocate a separation of semantic relations and discourse progression.

for *Continuation* and *Conjunction* from our *Continuative* subgroup, it is clear that a subordinating counterpart is not possible.

From an information packaging point of view, having a *Summary* or *Commentary* without the commented or summarized content is not necessarily sensible. A similar argument could be made for the *Instance* and *Background* relations (where the discourse segment subordinated by these relations takes up the main event or another referent from the superordinate discourse segment).

Table 1 also contains a comparison of our relation inventory to the counterparts of the RST Discourse Treebank, the SDRT corpus of Reese et al. (2007), and the Penn Discourse Treebank. In some cases, such as our *Conjunction* relation in comparison to *List/Conjunction/Equivalence* in the Penn Discourse Treebank, or our *Commentary* relation in comparison to the relations by RST, we found it more attractive to have very few 'weak' discourse relation types, similar to the simplifications applied in the Leeds Arabic Discourse Treebank.

4. Linguistic Properties and Tests

An annotation scheme needs to make explicit the categories and criteria used in the annotation. Such an explicitation is useful in general for users of the annotated corpus to understand individual annotations, but it is also crucial for large-scale corpus annotation where multiple annotators need to reach consistent annotation decisions independently of each other.

For this operationalization of annotation guidelines, an informal description of the annotated categories is often complemented by illustrating examples, as in the manuals of Carlson et al. (2003) or Reese et al. (2007).

For difficult cases, the introduction and use of disambiguating heuristics, or **linguistic tests** is needed – usually, an insertion or substitution operation together with an estimate of which aspects of the meaning may change in the substitution, and which cannot. Using multiple linguistic tests can be helpful as long as one is clear about the **linguistic properties** that these tests are supposed to verify, which means that it is often appropriate to go beyond a pre-theoretic understanding of the testing heuristics themselves.

One of the most basic tests with respect to discourse markers and discourse relations is the insertion of different discourse markers between two units of discourse, as used at least by Sanders et al. (1992) and used as a principal tool of investigation by Knott and Dale (1994) to create their shallow taxonomy of discourse connectives. Knott (1996) discusses how **substitutability in context** can be used to abstract from cue phrases (or discourse connectives) to feature values (i.e., linguistic properties) by assuming, e.g., a feature that has different values for two discourse connectives when no context exists where they are substitutable.

In Knott's method, a pair of discourse segments is given (which occurs in a context with one original discourse marker) and a given discourse marker can be judged as either *substitutable* with the original one linking the two (in which case the intended meaning stays the same), it may be ungrammatical or incoherent using the new marker, or the result may be grammatical and coherent, but carry a new meaning, in which case the replacement is not deemed substitutable. Knott's methodology makes very few assumptions and hence is very suitable as a starting point for a theory-neutral account, and Knott notes that the resulting findings correlate with (intuition-guided) distinctions found by Sanders et al. (1992).

In practical use, it is possible that features are underspecified, or that intuitively plausible discourse relations sometimes involve prototypicality or **family resemblance** effects. This means that

Our Corpus	SDRT	RST	PDTB
CONTINGENCY			
Causal			
c Result-(cause,enable)	Result	Cause	Cause:(reason,result)[b]
c Result-epistemic	Result	Evidence	Pragmatic cause:justification
c Result-speechact	Result	Rhetorical-Question	Cause:(reason,result)[b]
s Explanation-(cause,enable)	Explanation	Result	Cause:(reason,result)[b]
s Explanation-epistemic	Explanation	Explanation-argumentative	Pragmatic cause:justification
s Explanation-speechact	Explanation	Elaboration-additional	Cause:(reason,result)[b]
Conditional			
c Consequence	Consequence	Consequence	Condition
c Alternation	Alternation	Disjunction	Alternative
s Condition	Consequence	Condition	Condition
Denial			
c ConcessionC	Contrast	Antithesis/Preference	Contrast
s Concession	Contrast	Concession	Concession
s Anti-Explanation	—	Circumstance	Cause:reason
EXPANSION			
Elaboration			
s Restatement	Elaboration	Elaboration	Restatement:specification
s Instance/V	Elaboration	Example	Instantiation
s Background	Background	Circumstance	Synchronous
Interpretation			
s Summary	Elaboration	Conclusion	Restatement:generalization
s Commentary	Commentary	Comment/Evaluation/ Interpretation	Conjunction
Continuative			
c Continuation	Continuation	Joint	Conjunction
c Conjunction	Continuation/ Parallel	Joint/List	List/Conjunction/ Equivalence
TEMPORAL			
c Narration	Narration	Sequence	Asynchronous:succession
s Precondition	Precondition	Inverted-Sequence	Asynchronous:precedence
COMPARISON			
c Parallel/V	Parallel	Analogy	Conjunction
c Contrast	Contrast	Contrast	Contrast:juxtaposition
REPORTING			
s Attribution	Attribution	Attribution-N[a]	—
s Source	Source	Attribution	—

SDRT refers to the corpus annotation guidelines by Reese et al. (2007); RST refers to the annotation guidelines by Carlson et al. (2003). These annotation schemes do not necessarily reflect other schemes inspired by the same theories.

[a]) *Attribution-N* is used in Carlson et al. (2003) with the same semantics as our *Attribution* in cases where the attribution strictly cannot be veridical.

[b]) The Penn Discourse Treebank uses syntactical criteria to distinguish between *Cause:reason* and *Cause:result*, whereas we, like Reese et al. (2007), use criteria related to surface order for the distinction between *Result* and *Explanation*.

Table 1: Relation inventory in comparison with other schemes

it is useful both to reason explicitly about the connection between discourse relations and features, and to complement substitution tests with other kinds of tests where appropriate.

As an example for a non-trivial relationship between discourse connectives and properties of coherence relations, consider causal discourse relations, where the prototypical case is causation between events. In this prototypical case, causation holds between non-action events, corresponds to a temporal succession, and the second event would be predictable from the first. In such a case, both temporal connectives (*when/after*) and causal connectives (*because/since*) would be appropriate in the context. In non-prototypical cases such as piecemeal causation (in which one process influences another, while being co-temporal to it), reasons for actions, or logical causation between propositions, not all of these criteria hold, and a more detailed assessment of the participating properties is necessary to reach a crisp boundary for these relations (see subsection 4.2).

Table 2 summarizes both the grouping of discourse relations in our annotation scheme and the types of tests used for each discourse relation. Substitution/insertion of connectives (e.g., for *Alternation* or *Narration*) is complemented by paraphrase tests (e.g., the *Causal* group of relations), as well as tests that are aimed at identifying the influence of the question under discussion, such as nominalization (assuming that nominalized events/situations still realize the semantic contribution, but do not provide an answer to the QUD), explicit insertion of the linking question under discussion (for *Restatement*), or explicit marking for information-structural properties (*Parallel* and *Contrast*).

We have grouped the discourse relations into five broad categories, which each have own defining properties as well as properties typically disambiguating relations within that group:

- *Contingency* relations are defined in terms of Sanders et al.'s *causal source of coherence* – normally, such a discourse relation presupposes a rule $A \rightarrow B$ (Lagerwerf, 1998), which either has to be instantiated with real events or propositions (*Causal*), expressed as a general rule (*Conditional*), or imply a denial of an expectation from such a rule (*Denial*). Most frequently, different kinds of such relations can be distinguished with an explicit paraphrase of the rule involved (cf. subsection 4.2).

- *Expansion* relations achieve coherence through some kind of referential relation between referenced situations or entities in the arguments (*Elaboration* group, cf. subsection 4.3), making explicit the contribution of a larger group of discourse units (*Interpretation*), or link units that are only coherent because they answer a common question under discussion (*Continuative*, cf. subsection 4.4).

- *Temporal* and *Reporting* relations each express a specific (semantic) relation.

- *Comparison* relations usually involve an *overt contrast* between two objects. This overt contrast is essential for deciding between *Denial* and *Comparison* relations (subsection 4.1), but also between *Parallel* and *Conjunction* (subsection 4.4).

4.1 Adversative Relations

As an example for the linguistic tests, let us consider the domain of adversative relations, which Spenader and Lobanova (2009), based on data from the RST discourse treebank, argue to consist of three different 'discourse marker profiles', without however proposing tests which could help the annotation of discourse corpora.

Relation	Test material	Type of Test
CONTINGENCY		
Causal		
Result-cause	*Die Folge daraus war, dass* (arg2) The consequence of this was that *(arg2)*	Complex phrase
Result-enable	*Das ermöglichte es, dass* (arg2) This made it possible that *(arg2)*	Complex phrase
	Das trug dazu bei, dass (arg2) This contributed to the fact that *(arg2)*	Complex phrase
Result-epistemic	*Daraus schließe ich, dass* (arg2) I infer from this that *(arg2)*	Complex phrase
Result-speechact	*Deswegen* [speechact-verb] *ich*: (arg2) Hence I [*speechact-verb*]: *(arg2)*	Complex phrase
Explanation-cause	*Der Grund dafür ist, dass* (arg2) The reason for this is that *(arg2)*	Complex phrase
Explanation-enable	*Möglich gemacht wurde das durch* (arg2) This was possible by *(arg2)*	Complex phrase
Explanation-epistemic	*Das schließe ich daraus, dass* (arg2) I infer this from the fact that *(arg2)*	Complex phrase
Explanation-speechact	*Ich* [speechact-verb] *das, weil* (arg2) I [*speechact-verb*] this because *(arg2)*	Complex phrase
Conditional		
Consequence	(always marked)	—
Alternation	*Ansonsten* (arg2) / Otherwise, *(arg2)*	Insertion of connective
Condition	(always marked)	—
Denial		
ConcessionC	*Zwar* (arg1) *aber* (arg2) Certainly *(arg1)* although *(arg2)*	Insertion of connective
Concession	Substitution of arg1 with '*trotz* NP' ('despite *NP*')	Nominalization
Anti-Explanation	*Der Grund dafür ist nicht, dass* (arg2) The reason for this is not that *(arg2)*	Complex phrase
EXPANSION		
Elaboration		
Restatement	*Inwiefern* (arg1)? *Insofern, als* (arg2) In what respect *(arg1)*? To the extent that *(arg2)*	Question-Answer-Coherence
Instance	*Zum Beispiel* (arg2) / For example, *(arg2)*	Insertion of connective
InstanceV[a]	*Beispielsweise und vor allem* (arg2) Especially, for example, *(arg2)*	Insertion of two connectives
Background	*Was* (arg1-elab) *betrifft* (arg2) As to (arg1-elab), (arg2)	Complex phrase
Interpretation		
Summary	*Zusammenfassend gesagt,* (arg2) To summarize, *(arg2)*	Complex phrase
Commentary	*Ehrlich gesagt,* (arg2) To be honest, *(arg2)*	Insertion of adverbial
Continuative		
Continuation	—	—
Conjunction	Omitting any of (arg1) or (arg2) possible	Omission of EDU

Table 2: Summary of linguistic tests for each relation (Part I)

[a]) *InstanceV* is a variant of the *Instance* relation where an especially salient example, rather than any applicable instance, is picked out.

Relation	Test material	Type of Test
TEMPORAL		
Narration	*Dann* (arg2) / Then, *(arg2)*	Insertion of connective
Precondition	*Zuvor* (arg2) / Previously, *(arg2)*	Insertion of connective
COMPARISON		
Parallel	*auch* (arg2) / also, *(arg2)*	Insertion of connective
	CT on overt contrast	Accent placement
	switch arg1 and arg2	Symmetry
ParallelV[b]	*auch und vor allem* (arg2) / also, and especially *(arg2)*	Insertion of two connectives
Contrast	*Während* (arg1), (arg2) / While *(arg1)*, *(arg2)*	Insertion of connective
	CT/F on overt/secondary contrast	Accent placement
	switch arg1 and arg2	Symmetry
REPORTING		
Attribution	Substitution of arg2 with NP	Nominalization
Source	Speaker of arg2 can be replaced with other person	Point of view

Table 3: Summary of linguistic tests for each relation (Part II)

[b]) *ParallelV* is a variant of *Parallel* where the second relation argument is presented as an "even stronger" example for the property in question.

The connective annotation, in which the linguistic tests were developed, distinguishes between contrast relations (where two items are compared), contraexpectation (where a plausible consequence of Arg1 is denied in Arg2) and antithesis (where a question under discussion is first answered with the contents of Arg1, but the answer of Arg1 is overridden by the answer in Arg2), which is mostly identical to the three-way distinction used by RST.

(11) [QUD: Are all children equally tall?]
Peter is stubby, but Mary is tall. *contrast*

(12) Peter has bought the book, but he hasn't read it. *contraexpectation*

(13) [QUD: Should we go to the pool?]
Peter likes going to the pool, but Mary cannot swim. *antithesis*

In the discourse structure annotation, the coordinating cases of contraexpectation (12) and antithesis (13) are grouped in one relation (*ConcessionC*), against the subordinating cases ("Peter hasn't read the book although he has bought it", *Concession*) and *Contrast* relations such as (11), which helps avoid some corner cases between antithesis and contraexpectation.

Of the three, *Contrast* can be delimited most clearly: it is a characterization of two related entities (called *overt contrast* by Lang, 1984), which differ in some relevant property (the secondary contrast) while both relata talk about a common question (called *common integrator* by Lang). Using these properties, it is possible to formulate several testable assertions on the contrast relation:

Firstly, *Contrast* is symmetric: unlike with (12) and (13), a swapped version of (11), „Mary is tall, but Peter is stubby" is just as informative with respect to the QUD. Secondly, the absence of an overt contrast pair (as in (12), where both sentences share the subject) is a relatively clear sign against a contrast relation. Finally, the absence of a secondary contrast compatible with the inferred QUD, as is the case in (13), is also a sign against a *Contrast* relation. The *contraexpectation* relation is grouped among the Contingency relations in the TüBa-D/Z annotations since a single

logical relation A -> B can be realized (or rather, presupposed) in a multitude of different relations including *Consequence* ('If it is sunny, the laundry dries well'), *Result* ('Because it was sunny, the laundry dried well') or *Concession* ('Although it was not sunny, the laundry dried well'), as well as *Anti-Explanation* ('The laundry dried well, but not because it was sunny').

As a result, we can test this logical relation among the examples: In (11), Peter being stubby does not have any influence on Mary's size, and in (13), Peter's liking of the pool does not have any influence on the swimming abilities of Mary. In contrast, „Peter bought the book" raises the (plausible) expectation that he also read the book, which is directly denied in the following clause „he hasn't read it". (In other cases, such as „Peter turned the ignition but it was not his car's favorite day", a strong expectation can be implicitly denied). A less involved test for the contraexpectation relation is the possibility of an "*obwohl* (arg1), (arg2)" ('although *(arg1)*, *(arg2)*') paraphrase, which is possible in (12) but out of question in (11) and (13).

In the typical case of antithesis, the relation between two clauses is only clear with respect to the question under discussion, whereas overt/secondary contrast pair or a denied expectation would be interpretable regardless of context. While it could be argued that antithesis should be seen as subsuming the other two relations (as argued by Umbach and Stede, 1999, as well as Spenader and Maier, 2005, who both use the term Contrast for a relation), such an analysis would fail to explain cases ambiguous between parallel and contrast, and would have difficulties accounting for constructions that can only occur with contraexpectation, such as "*trotz* NP" ('despite *NP*') paraphrases and "*obwohl* S" ('although *S*').

4.2 Temporal/Causal Markers

Markers with a primary temporal meaning, such as *nachdem* ('after'/'since'/'as'), *als* ('when'/'as'), or *während* ('while') often convey causal or contrastive relations. In the case of temporal relations between events (for *nachdem* and *als*), Herweg (1991) and Bäuerle (1995) claim that their core meaning is not a relation between temporal intervals, but one of situational location – in the case of *nachdem*, in the post-phase of the subclause's event, in the subclause's process phase for *während* or in the general (situational) proximity of the subclause's event for *als*.

This situational connection can also impart causal information by the intermediary of the construction of a post-state (e.g. from *going to the supermarket* to *being in the supermarket*), or other semantic relations in the case of (event) redescriptions with *als* as in example (14) by Bäuerle, which would correspond to our *Restatement*:

(14) Als Fritz den Hund fütterte, gab er ihm Schappi.
When Fritz fed the dog, he gave him Chappi.

In contrast to causal readings and redescriptions, contrastive readings seem to occur parasitically on the situations present in temporal *nachdem* or *während* connectives. Often, a primarily contrastive or parallel reading with these connectives contains an explicit temporal adverbial in addition to the *nachdem* or *während* clause:[12]

(15) [Diese Kuppel war im Juni 1998 zum zweiten Mal eingestürzt,]
[nachdem sie im Februar 1997 erstmals eingebrochen war.]

12. TüBa-D/Z, sentence 680

[*This cupola had collapsed for the second time in June 1998,*]
[*after it had caved in for the first time in February 1997*]. *Narration+Parallel*

Cases where a temporal coherence relation co-occurs with a causal or contrastive one can often be detected by substituting a synchronous marker (*als* or *während*) with an asynchronous one (*nachdem*) or vice-versa, since the non-temporal information would keep the coherence independently of the temporal relation.

Comparing the results of a substitution with *weil* ('because') with those that one gets by substituting with *kurz nachdem* ('shortly after'), which forces a purely temporal reading, reveals some cases that are not purely temporal, but also do not lend themselves to substitution with *weil*. Such cases of contributing causes are distinguished from (strong) causes, and are presented as a *Weak-Result* relation by Bras et al. (2009). Such enabling conditions are annotated with the *Result-enable* and *Explanation-enable* relations in our scheme, and can be distinguished from their stronger counterparts *Result-cause* and *Explanation-cause* by the fact that they allow the addition of a 'main' cause with a *weil* adjunct:[13]

(16) [Nachdem die Fraktionsvorsitzende Künast sich anders entschieden hat,]
[gilt Bundestagsvizepräsidentin Vollmer als aussichtsreichste Kandidatin.]
(*weil sie als kompetent gilt*)
[*After the parliamentary leader Künast has decided otherwise,*]
[*Bundestag vice president Vollmer is considered the most promising candidate.*]
(because she is considered competent) *Result-enable*

(17) [Im vergangen Jahr war es an der Kastanienallee zu Ausschreitungen gekommen,]
[nachdem die Polizei ein Transparent aus dem Demozug entfernen wollte.]
(??*weil die Demonstranten unzufrieden waren*)
[*Last year riots occurred at the Kastanienallee*]
[*after police wanted to remove a transparent from the demonstration train.*]
(??because the demonstrators were unhappy) *Explanation-cause*

Note that the distinction between main and contributing causes often hinges on the conceptualization of the events (and their manipulability) by the speaker rather than on the events themselves: A spontaneous eruption of violence (*riots occurred*) may be *caused* by actions of the police whereas conscious action (*the demonstrators threw stones*) may be better explained by an independent motivation rather than surrounding events alone.

In addition, there are non-temporal uses of *während* (as a general marker of *Contrast*) and *nachdem* (for the justification of claims, called *pragmatic cause* or *evidence* in other annotation schemes). In the first case, one can add temporal adverbials (such as *in the morning* and *in the evening*) to both arguments of the connective so as to exclude the temporal reading of *während*. In the second case, the difference is between a connection between situations (*cause* and *enable*) and a connection between claims (our *Result-epistemic*, cf. also Sweetser, 1990; Sanders, 2005), which can, but does not have to, occur with an order inverse to normal causality between events, such as example (18), due to Sweetser:

(18) John loved her, because he came back.

13. TüBa-D/Z, sentence 2551; TüBa-D/Z, sentence 18884. Note how example (16) would still be acceptable if the *nachdem*-clause were postposed in order to match the subclause ordering in (17).

In sentences where a prediction or other claim is supported by a *Result-epistemic* or *Explanation-epistemic* relation, the claim is non-factive (Lang, 2000, explains the distinction as *assuming* rather than *asserting* a proposition) and can be amended or retracted, as in example (19):[14]

(19) Somit ist Mala, eine Einödgemeinde, wohl jetzt der Wunschort der Atomindustrie.
[Nachdem aber schon vier Kommunen dem SKB bewiesen haben, daß sie trotz hoher Arbeitslosigkeit nein zum Atomklo sagten,] [sieht es für den Strahlenmüll zappenduster aus.]
(*. . . oder die Atomindustrie hat noch einen Trumpf im Ärmel*)

Hence, the solitary borough of Mala seems to be the desired locality for the atomic industry.
[*As already four municipalities proved to the SKB that they rejected an atomic disposal site despite high unemployment,*] [*the outlook is dim for the radiation waste.*]
(. . . or the atomic industry has another ace in their sleeve) *Result-epistemic*

4.3 The Expansion Group

In the process of discourse annotation, we also noticed that a group of discourse relations that typically occur with unmarked instances was posing some difficulties. In older proposals, many of these would have been called *elaboration*. Indeed, this specific set of subordinating relations serves the purpose of giving further information and elaborating the first argument in a way that is not strictly necessary for the coherence and understanding of the main text.

The criteria for these relations — the *Expansion.Elaboration* group containing the *Instance*, *Restatement* and *Background* and the *Expansion.Interpretation* group containing *Summary* and *Commentary* — often have to take into account the surrounding discourse. As they do not convey a strong semantic relation, substitution with prepositional phrases or even insertion of subordinating conjunctions is not the most straightforward test. Instead, the most indicative tests for this group of relations pick out properties such as the perspectivization of each connected discourse unit, or the relation of the themes/questions under discussion of Arg1 and Arg2.

We use the **Instance** for relations where Arg2 concerns a proper subset of the events described in Arg1. These cases allow the insertion of phrases such as "*zum Beispiel*" ('for example') in Arg2.

(20) Max did well in school this year. In biology [*for example*] he had an 'A'.

In this case, Max's doing well in biology this year is conceptualized as a proper subset of Max doing well in school this year.

In case of **Restatement**, the main event of Arg2 elaborates the main event of Arg1 as a whole. Jasinskaja (2007) claims that the Restatement relation works similar to an apposition in the case of nominal phrases in sentence syntax. The test of asking "*Inwiefern* (arg1)" ('In what respect (*arg1*)') and reading Arg2 as the answer to this makes the question under discussion – further qualification of Arg1 by a redescription as Arg2 – explicit. If the question is asked with sentential (i.e. largest possible) focus and Arg2 is still a good and coherent answer, the relation must be a *Restatement*.[15]

(21) Max did well in school this year. [*In what respect did Max do well in school this year?*]
He had an 'A' in every subject.

14. TüBa-D/Z, sentence 39373
15. Many *Restatement* relations can be marked with *insofern, als/dass* (in that), which however often feels awkward for stylistic reasons or due to the added syntactic complexity.

In the discourse relation **Background**, Arg2 elaborates a non-central part of Arg1, typically one phrase. This can be tested with an addition such as "*Was* [...] *angeht,*" ('As to [*elaborated phrase*]'). This frame-setting topic makes explicit that it is not the main event of Arg1 that is elaborated, but the specific phrase.[16]

(22) Max did well in school this year.
As to this year, this was his second last high school year.

In summary, *Restatement*, *Background* and *Instance* are different in which part of Arg1 they elaborate. The two other relations, *Commentary* and *Summary* differ from the former three because Commentary introduces information that differs in perspectivization and Summary does not introduce any new facts at all.

In the words of Reese et al. (2007), **Commentary** is a relation where Arg2 provides an opinion or evaluation of the content associated with Arg1. A test has to establish that the perspective of the commentary in Arg2 is in fact the author's, whereas Arg1 is usually an objective fact.

We can test for author perspectivization using commentary adverbials such as the German "*ehrlich gesagt*" ('to tell the truth, . . . ', 'frankly, . . . ') which are utterance modifiers and cannot occur under other perspectives than the author's (Bellert, 1977; Potts, 2005).

In our case, we want to go beyond just taking the presence of utterance modifiers in the original text as an indicator for a Commentary relation (as, e.g., Reese et al. do). Using the addition of utterance modifiers as a linguistic test by manipulating the corpus sentences means that annotators have to take into account possible shifts of meaning, as in (24):

(23) Max did well in school this year. [*Frankly,*] this is exactly what I expected.

(24) # Max did well in school this year. Frankly, he had an 'A' in every subject.

In (24), the utterance would sound odd in a neutral context.

For **Summary** relations, Arg2 is a reformulated and condensed version of (the most important part of) the content of Arg1.[17] Inserting a summative adverbial such as "*zusammenfassend gesagt*" ('in summary') is only possible if Arg2 does not contribute any new information for the reader, which provides us with a plausible linguistic test. In (25) and (26), the summative adverbial excludes further elaboration.

(25) Max had an 'A' in biology, maths, physics, and history.
[*In summary,*] he did well in school this year.

(26) Max did well in school this year.
In summary, he had an 'A' in every subject.

An acceptable reading of (26) would entail an additional implication such as "I could tell you more facts about this, but in summary [of these untold facts], he had an 'A' in every subject".

16. One reviewer pointed at the possibility of pronominalizing the elaborated part in *arg2*, which often sounds more natural. Using *as to* as an explicit frame-setting construction is more independent of other factors influencing pronominalizability, especially the salience of the entity in arg1 or the presence of confounders. In addition, it gives the correct result when there is a dissociation between pronominalized entity and topic entity such as in *that letter from him*.

17. Jasinskaja (2007) does *not* differentiate between the cases that we would annotate as *Summary* and those that would be a *Restatement* in our annotation scheme, subsuming both cases under her Restatement label.

4.4 Weak Coherence

More frequently than one might expect, two discourse segments that belong together do not show a discernable semantic or referential relation such as those discussed in the previous subsections. Such relations occur as *Continuation* or *Parallel* in the scheme of Reese et al. (2007), as *Joint* or *List* relations in that of the RST Discourse Treebank (Carlson and Marcu, 2002) or as *Conjunction* and other relations in the Penn Discourse Treebank (Prasad et al., 2008).

While their presence is often obvious from the surrounding context, the lack of semantic distinctions can make these relations somewhat elusive. However, using an explicit notion of the context's contribution we can hope to get a firmer hold of these relations.

In classical examples of weak coherence, two discourse segments together provide an answer to a question under discussion where the single segments do not. This can be illustrated with this example from Blakemore and Carston's (2005) discussion of the possible relations expressible by the conjunction *and*:

(27) A: Shall we start without Jane?
B: [Well, she did say to start if she was late,]
[and we have been waiting for half an hour now.] *Continuation*

Using such an accumulation-of-evidence refinement to the *Continuation* relation with an inventory close to that of Reese et al. (2007) led to situations where annotators would use the *Parallel* relation label for cases that do not have an overt contrast between two entities. While such examples do receive a *Parallel* or *Continuation* label in Reese's scheme, we found it preferable to introduce a third category *Conjunction* as a coordination relation between discourse segments that provide independent (satisfactory) answers to the same question under discussion while not incorporating a comparison between different entities.[18]

Example (28)[19] shows a typical example of *Conjunction*: *being calm* (inferred from 'Sauer's voice sounds austere') and *recounting precise details* both contribute to the credibility/reliability of the witness independently, which excludes *Continuation*. Furthermore, there is no overt contrast which would allow a *Parallel* relation.

(28) Tatsächlich lässt der zweite Zeuge den Mandanten kaum Chancen auf eine milde Strafe.
(QUD: *Warum ist Sauer ein verlässlicherer Zeuge als Emrich?)*
[Im Gegensatz zu Emrich klingt Sauers Stimme nüchtern,]
[und er kann sich an genaue Details erinnern.]
Indeed the second witness leaves the clients scarcely a chance for a mild penalty.
(QUD: Why is Sauer a more reliable witness than Emrich?)
[*In contrast to Emrich, Sauer's voice sounds austere,*]
[*and he can remember precise details.*] *Conjunction*

18. As our notion of discourse structure follows SDRT in that it postulates a structure that simultaneously includes a *coordinating* relation between dependants of a superordinate segment, such as *Conjunction*, and a *subordinating* relation between the superordinate segment and its subordinated segment, e.g., *Explanation-epistemic*, we need to postulate a conjunction relation where RST just annotates the subordinating relations. In the typical case of a conjunction co-occurring with multiple instances of the same subordinating relation, we consider the *Conjunction* to be implied if no other relations is annotated and the result looks very similar to RST's "multiple satellites" construction.
19. Tüba-D/Z, sentence 2469

5. Inter-annotator and Inter-adjudicator Agreement

As in any annotation project, the annotation scheme and its description in terms of annotation guidelines is not the sole determinant of the final quality but needs to be seen in conjunction with the training and abilities of actual annotators. Numerical measures of agreement target this notion of final quality by asking whether the annotation task is repeatable when done by two annotators independently of each other.

In this section, we report quantitative agreement figures from two studies involving a sample of documents annotated, respectively, by the same two annotators: one pilot study with three documents, and one using a larger sample of seven documents. We also give a more detailed appraisal of agreements and disagreements based on a larger sample that however involves a larger group of annotators. Among the aspects of our annotation scheme, EDU segmentation shows invariably high agreement ($\kappa > 0.9$ for all articles).

Among the disagreements in EDU segmentation, non-typical communication verbs (*think*, *contend*, *write*) are correlated with the presence or absence of a particular type of relation (esp. the *Reporting* group), and non-restrictive relative clauses, usually correlate with the presence or absence of a *Background* relation. However, these are not among the most frequent nor the most interesting groups of disagreements (most of these are resolved in the first pass). The most interesting aspect of an agreement study therefore consists in discourse relations, as well as the placement of topic segments and transitional EDUs.

In previous studies on annotator agreement, Marcu et al. (1999) determined κ values between $\kappa = 0.54$ (Brown corpus) and $\kappa = 0.62$ (MUC) for fine-grained RST relations and between $\kappa = 0.59$ (Brown) and $\kappa = 0.66$ (MUC) for coarser-grained relations. In their reliability study with the Penn Discourse Treebank, Prasad et al. (2008) determined agreement values between 80% (finest level) and 94% (coarsest level with 4 relation types), but did not report any chance-corrected values. Al-Saif and Markert (2010) report values of $\kappa = 0.57$ for their PDTB-inspired connective scheme, saying that most disagreements are due to highly ambiguous connectives such as '*w*' (Arabic counterpart of *and*), which can receive one of several relations. In a study on their Dutch RST corpus, van der Vliet et al. (2011) found an inter-annotator agreement of $\kappa = 0.57$.

To the best of our knowledge, no agreement figures have been published on the RST-based Potsdam Commentary Corpus (Stede, 2004a) or any other German corpus with discourse relation annotation.

5.1 Quantitative Agreement Figures

In the regular annotation process, two annotators create EDU segmentation, topic segments, and discourse relations independently from each other; in a second step, the results from both annotators are compared (cf. figure 1) and a coherent gold-standard annotation is created after discussing the goodness-of-fit of respective partial analyses with the text and the applicability of linguistic tests. In order to account for the complete annotation process including the revision step, we follow Burchardt et al. (2006) and separately report inter-annotator agreement, which is determined after the initial annotation, and inter-adjudicator agreement, which is determined after an additional adjudication step, which is carried out by two adjudicators based on the original set of annotations but working independently from each other.

In the case where multiple relations were annotated between the same EDU ranges (for example, a temporal *Narration* relation in addition to a *Result-Cause* relation from the *Contingency* group),

Figure 1: Display of annotation differences

Restatement	Background	8
Result-cause	Result-enable	7
Restatement	Explanation-cause	6
Restatement	Instance	5
Continuation	Narration	4
Conjunction	Parallel	4
Narration	Result-cause	4
Narration	Parallel	4
Restatement	Explanation-epistemic	3
Explanation-cause	Explanation-enable	3

Table 4: Most frequent disagreements (in 18 documents annotated by different annotators)

we counted the annotations as matching whenever the complete set of relations (i.e. Narration, Result-Cause in the example) is the same across annotators.

For the agreement among relations, we performed a pilot study on a sample of three documents, where annotators agreed on 49 relation spans (about 40% of all spans); later, we performed a second study on a sample of seven documents, where annotators agreed on 108 relation spans (about half of all spans). The pilot study did not make an effort to match relations among spans with segmentation differences, which can agree due to differences in EDU segmentation or differences in the exact span of a complex discourse unit. In the second study, we accounted for some differences in EDU segmentation by matching two EDUs if they have the same span or if their clauses have the same (syntactic dependency) head terminals. In both studies, the agreement was only evaluated among relations with compatible spans (evaluating the labeling of relations but not the discourse structure). The percentage of multi-EDU spans in both samples is about the same as in the corpus as a whole (27.6% and 29.8%, compared to 25.1% for the whole corpus).

In the pilot study, annotator agreement yielded $\kappa = 0.55$ for individual relations, and $\kappa = 0.65$ for the middle level of the taxonomy (nine relation types), whereas in the second study, we found $\kappa = 0.59$ for individual relations and $\kappa = 0.69$ for the middle level of the taxonomy.

For the inter-adjudicator task, we found an agreement on 82 relation spans (about two thirds of all annotated spans), among which relation agreement was at $\kappa = 0.83$ for individual relations, and $\kappa = 0.85$ for the middle level of the taxonomy, or a reduction of disagreements of about 57%. For the second study, with 167 matching spans (=78% of all annotated spans), we found $\kappa = 0.90$ for individual relations and $\kappa = 0.91$ for the middle level of the taxonomy, with, again, a reduction of disagreements by about two thirds.

To measure quantitative agreement on the topic segmentation, we used the same sample of seven newspaper articles, which were marked for topic boundaries independently and without considering the paragraph boundaries in the printed newspaper, yielding 21 (annotator 1) or 22 (annotator 2) topic segments, of which 16 matches. Modeling the annotation process as tagging each sentence as topic start or as nonboundary yields a value of $\kappa = 0.71$. Positive specific agreement on nontrivial boundaries (i.e., those not at the start of the main text of the article) is at $F_1 = 0.62$.

Agreement on transitional EDUs is very good: among the seven texts we used for our study, six articles showed perfect agreement, with four matching transitional EDUs altogether and one unmatched. Modeling the annotation process as tagging each inter-topic boundary with or without a transitional EDU yields a value of $\kappa = 0.85$, or a positive specific agreement of $F_1 = 0.88$.

5.2 Identifying Problematic Areas

Table 4 shows the most frequent confusions between discourse relations in cases where different annotators could agree on a span, but chose different relations, collected over a larger sample of documents annotated by at least two trained annotators. Surprisingly, one of the greatest sources of confusion is disagreement between the *Restatement* and *Explanation-cause* or *Explanation-epistemic* relations. Such disagreements can sometimes arise when situation identity is difficult to assess, as in example (29):[20]

(29) [Die Deklarationsbestimmungen der EU sind bis heute schwammig:]
[Hingewiesen werden muß nur auf im Endprodukt nachweisbare Genmanipulationen.]
[*The declaration provisions of the EU are vague even today:*]
[*Only genetic manipulations that are demonstrable in the final product have to be indicated.*] *Restatement*

In this case, insertion of a causal marker is possible (which would indicate an *Explanation-cause* relation), but the stronger test of inserting a '*in what regard?*' subquestion shows that (29) is indeed a case of *Restatement*.

In other cases, annotated relations vary in the degree of causality (between purely temporal and an *enable* relation, or between *enable* and *cause*). In some cases, modal embedding or other modifiers make the causality judgement more difficult, as in example (30),[21] where a non-modal "if there is a need for it, we destroy the GM crops" would receive a *Result-Enable* relation (since there is a causal link, but the knowledge does not actually force the administration to destroy crops), but the modal embedding means that the discourse relation must be a *Result-Cause* relation relating the knowledge of the seed locations to the fact that it is possible to find and destroy the crops.

(30) ["Wir wissen aber, wo das Saatgut hingegangen ist", so Fluhme,]
["wenn wirklich die Notwendigkeit bestehen sollte, könnten wir die betroffenen Felder ausfindig machen und die Vernichtung der gentechnisch veränderten Pflanzen sicherstellen.]
[*"But we know where the seed has gone", says Fluhme,*]
[*"if there was a need for it, we could locate the concerned fields and ensure the destruction of the genetically modified crops.*] *Result-Cause*

6. Summary

In this article, we have presented the most important design choices that an annotation scheme for discourse structure and discourse relation faces, and presented the particular solutions taken in the annotation scheme for discourse annotation in the TüBa-D/Z corpus. We have explicitly linked the categories of the annotation scheme to linguistic categories that have been described in the literature and demonstrated how this can be harnessed for the creation of linguistic tests that help in the annotation of discourse relations. Finally, we have presented figures for inter-annotator and inter-adjudicator agreement in our corpus and have shown how the remaining ambiguities in the annotation scheme relate to problems such as event identity or parthood that also pose difficulties for other annotation schemes (such as event relation annotation in TimeML).

20. TüBa-D/Z, sentences 5717ff.
21. TüBa-D/Z, sentence 1083.

Acknowledgements Yannick Versley and Anna Gastel were supported by the Deutsche Forschungsgemeinschaft (DFG) as part of SFB 833. We would like to thank the three anonymous reviewers for their helpful suggestions.

References

Amar Al-Saif and Katja Markert. The Leeds Arabic discourse treebank: Annotating discourse connectives for Arabic. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, 2010.

Nicholas Asher. Discourse topic. *Theoretical Linguistics*, 30(2-3):163–201, 2004.

Nicholas Asher and Alex Lascarides. The semantics and pragmatics of presupposition. *Journal of Semantics*, 15(3):239–299, 1998.

Nicholas Asher and Alex Lascarides. *Logics of Conversation*. Cambridge University Press, 2003.

Nicholas Asher and Laure Vieu. Subordinating and coordinating discourse relations. *Lingua*, 115 (4):591–610, 2005.

Rainer Bäuerle. Temporalsätze und Bezugspunktsetzung im Deutschen. In Brigitte Handwerker, editor, *Fremde Sprache Deutsch*, pages 155–176. Narr, 1995.

Irena Bellert. On semantic and distributional properties of sentential adverbs. *Linguistic Inquiry*, 8 (2):337–351, 1977.

Diane Blakemore and Robyn Carston. The pragmatics of sentential coordination with "and". *Lingua*, 115:569–589, 2005.

Myriam Bras, Anne Le Draoulec, and Nicholas Asher. A formal analysis of the French connective *alors*. *Oslo Studies in Language*, 1(1):149–170, 2009.

Daniel Büring. On d-trees, beans, and b-accents. *Linguistics and Philosophy*, 26(5):511–545, 2003.

Paula Christina Figueira Cardoso, Erick Galani Maziero, Maria Lucía del Rosario Castro Jorge, Eloize Rossi Marques Seno, Ariano Di Felippo, Lucia Helena Machado Rino, Maria das Graças Volpe Nunes, and Thiago Alexandro Salgueiro Pardo. CSTNews — a discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese. In *3rd RST Brazilian Meeting*, 2011.

Lynn Carlson and Daniel Marcu. Discourse tagging reference manual. Technical report, Information Sciences Institute, University of Southern California, 2002.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Current Directions in Discourse and Dialogue*. Kluwer, 2003.

Iria da Cunha, Juan Manuel Torres-Moreno, and Gerardo Sierra. On the development of the RST Spanish treebank. In *Proceedings of the 5th Linguistic Annotation Workshop (LAW V)*, 2011.

Barbara Grosz and Candice Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986.

Marti A. Hearst. TextTiling: segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, 1997.

Michael Herweg. Temporale Konjunktionen und Aspekt. *Kognitionswissenschaft*, 2(3–4):51–90, 1991.

Jerry Hobbs. On the coherence and structure of discourse. Technical Report 85-37, Center for the Study of Language and Information (CSLI), 1985.

Eduard H. Hovy and Elisabeth Maier. Parsimonious or profligate: How many and which discourse structure relations? Unpublished manuscript, 1995.

Julie Hunter, Nicholas Asher, Brian Reese, and Pascal Denis. Evidentiality and intensionality: Two uses of reportative constructions in discourse. In *Constraints in Discourse 2006*, 2006.

Ray Jackendoff. *Semantic Interpretation in Generative Grammar*. MIT Press, Cambridge, MA, 1972.

Ekaterina Jasinskaja. *Pragmatics and Prosody of Implicit Discourse Relations: The Case of Restatement*. PhD thesis, Universität Tübingen, 2007.

Alistair Knott. *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. PhD thesis, University of Edinburgh, 1996.

Alistair Knott and Robert Dale. Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes*, 18(1):35–62, 1994.

Alistair Knott, John Oberlander, Michael O'Donnell, and Chris Mellish. Beyond elaboration: The interaction of relations and focus in coherent text. In *Text representation: linguistic and psycholinguistic aspects*. John Benjamins, 2001.

Sudheer Kolachina, Rashmi Prasad, Dipti Misra Sharma, and Aravind Joshi. Evaluation of discourse relation annotation in the Hindi Discourse Treebank. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, 2012.

Jan van Kuppevelt. Discourse structure, topicality and questioning. *Journal of Linguistics*, 31(1): 109–147, 1995.

Luuk Lagerwerf. *Causal Connectives Have Presuppositions*. Holland Academic Graphics, The Hague, 1998.

Ewald Lang. Adversative connectors on distinct levels of discourse: A re-examination of Eve Sweetser's three-level approach. In Elizabeth Couper-Kuhlen and Bernd Kortmann, editors, *Cause-Condition-Concession-Contrast*, number 33 in Topics in English Linguistics. Mouton de Gruyter, 2000.

Harald Lüngen, Csilla Puskàs, Maja Bärenfänger, Mirco Hilbert, and Henning Lobin. Discourse segmentation of German written text. In *Proceedings of the 5th International Conference on Natural Language Processing (FinTAL 2006)*, 2006.

William C. Mann and Sandra A. Thompson. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.

Daniel Marcu, Estebaliz Amorrortu, and Magdalena Romera. Experiments in constructing a corpus of discourse trees. In *ACL Workshop on Standards and Tools for Discourse Tagging*, 1999.

Lucie Mladová, Šárka Zikánová, and Eva Hajičova. From sentence to discourse: Building an annotation scheme for discourse based on Prague Dependency Treebank. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, 2008.

Vera Möller and Karin Naumann. Manual for the annotation of in-document referential relations. Technical report, Seminar für Sprachwissenschaft, Universität Tübingen, 2009.

Johanna D. Moore and Martha E. Pollack. A problem for RST: The need for multi-level discourse analysis. *Computational Linguistics*, 18(2):537–544, 1992.

Nick Nicholas. Parameters for an ontology of Rhetorical Structure Theory. *University of Melbourne Working Paper in Linguistics*, 15(15):77–93, 1995.

Thiago Alexandre Salgueiro Pardo and Maria das Graças Volpe Nunes. On the development and evaluation of a Brazilian Portuguese discourse parser. *Journal of Theoretical and Applied Computing*, 15(2):43–64, 2008.

Thiago Alexandre Salgueiro Pardo and Eloize Rossi Marques Seno. Rhetalho: um corpus de referência anotado retoricamente. In *Anais do V Encontro de Corpora*, 2005.

The PDTB Research Group. The Penn Discourse Treebank 2.0 annotation manual. Technical Report IRCS-08-01, Institute for Research in Cognitive Science, University of Pennsylvania, 2008. `http://www.seas.upenn.edu/ pdtb/PDTBAPI/pdtb-annotation-manual.pdf`.

Marie-Paule Péry-Woodley, Stergos D. Afantenos, Lydia-Mai Ho-Dac, and Nicholas Asher. La ressource AnnoDis, un corpus enrichi d'annotation discursives. *Traitement Automatique des Langues*, 52(3):71–101, 2011.

Livia Polanyi and Remko Scha. A syntactic approach to discourse semantics. In *Proceedings of COLING 6*, 1984.

Christopher Potts. *The Logic of Conventional Implicatures*. Oxford University Press, 2005.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, 2008.

Marta Recasens, Eduard Hovy, and M. Antonìa Martí. Identity, non-identity and near-identity: Adressing the complexity of coreference. *Lingua*, 121(6):1138–1152, 2011.

Brian Reese, Julie Hunter, Nicholas Asher, Pascal Denis, and Jason Baldridge. Reference manual for the analysis and annotation of rhetorical structure. Technical report, University of Texas at Austin, 2007.

Craige Roberts. Information structure in discourse: Towards an integrated formal theory of pragmatics. OSU Working Papers in Linguistics 49, Ohio State University, 1996.

Ted Sanders. Coherence, causality and cognitive complexity in discourse. In *Symposium on the Exploration and Modelling of Meaning (SEM-05)*, 2005.

Ted Sanders and Wilbert Spooren. Communicative intentions and coherence relations. In Wolfram Bublitz, Uta Lenk, and Eija Ventola, editors, *Coherence in Text and Discourse*, pages 235–250. John Benjamins, 1999.

Ted J. M. Sanders, Wilbert P. M. Spooren, and Leo G. M. Noordman. Toward a taxonomy of coherence relations. *Discourse Processes*, 15(1):1–35, 1992.

Frank Schilder. Robust discourse parsing via discourse markers, topicality and position. *Natural Language Engineering*, 8(3):235–255, 2002.

Swapna Somasundaran, Galileo Namata, Janyce Wiebe, and Lise Getoor. Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, 2009.

Jennifer Spenader and Emar Maier. Contrast as denial in multi-dimensional semantics. *Journal of Pragmatics*, 41(9):1707–1726, 2005.

Manfred Stede. The Potsdam Commentary Corpus. In *ACL'04 Workshop on Discourse Annotation*, 2004a.

Manfred Stede. Does discourse processing need discourse topics? *Theoretical Linguistics*, 30(2-3): 241–253, 2004b.

Manfred Stede. RST revisited: Disentangling nuclearity. In Catherine Fabricius-Hansen and Wiebke Ramm, editors, *'Subordination' versus 'coordination' in sentence and text - A cross-linguistic perspective*. John Benjamins, Amsterdam, 2008.

Mark Steedman. Information structure and the syntax-phonology interface. *Linguistic Inquiry*, 31 (4):649–689, 2000.

Eve Sweetser. *From etymology to pragmatics*. Cambridge University Press, Cambridge, 1990.

Maite Taboada and William C. Mann. Rhetorical Structure Theory: Looking back and moving ahead. *Discourse Studies*, 8(3):423–459, 2006.

Heike Telljohann, Erhard W. Hinrichs, Sandra Kübler, Heike Zinsmeister, and Kathrin Beck. Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). Technical report, Seminar für Sprachwissenschaft, Universität Tübingen, 2009.

Sarah Tonelli, Giuseppe Riccardi, Rashmi Prasad, and Aravind Joshi. Annotation of discourse relations for conversational spoken dialogs. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, 2010.

Isabel Gomez Txurruka. The natural language conjunction 'and'. *Linguistics and Philosophy*, 26 (3):255–285, 2003.

Carla Umbach and Manfred Stede. Kohärenzrelationen: Ein Vergleich von Kontrast und Konzession. KIT-Report 148, Technische Universität Berlin, 1999.

Enric Vallduví and Maria Vilkuna. On Rheme and Kontrast. In *The Limits of Syntax*. Academic Press, 1998.

Yannick Versley. Vagueness and referential ambiguity in a large-scale annotated corpus. *Research on Language and Computation*, 6(3–4):333–353, 2008.

Ninke van der Vliet, Ildiko Berzlanovich, Gosse Bouma, Gisela Redeker, and Markus Egg. Building a discourse-annotated dutch text corpus. In Stefanie Dipper and Heike Zinsmeister, editors, *Proc. Beyond Semantics (DGfS workshop)*, 2011.

Bonnie Webber. D-LTAG: extending lexicalized TAG to discourse. *Cognitive Science*, 28(5):751–779, 2004.

Bonnie Webber, Matthew Stone, Aravind Joshi, and Alistair Knott. Anaphora and discourse structure. *Computational Linguistics*, 29(4):545–587, 2003.

Bonnie Lynn Webber. Discourse deixis: Reference to discourse segments. In *Proceedings of the 26th annual meeting on Association for Computational Linguistics*, 1988.

Florian Wolf and Edward Gibson. Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31(2):249–287, 2005.

Deniz Zeyrek, Işin Demirşahin, Ayışığı Sevdik-Çallı, Hale Ögel Balaban, İhsan Yalçınkaya, and Ümit Deniz Turan. The annotation scheme of the Turkish Discourse Bank and an evaluation of inconsistent annotation. In *Proceedings of the Fourth Linguistic Annotation Workshop (LAW-IV)*, 2010.